technische universität
dortmund

Reduktion hochdimensionaler Datensätze für die logische Regression unter der Verwendung von Leverage Scores mit besonderer Berücksichtigung von SNP-Daten

Masterarbeit

Fakultät Statistik

Autor: Alexander Wollenberg

Dozentin: Prof. Dr. Katja Ickstadt

13.12.2016

Inhaltsverzeichnis

1 Einleitung

Heutzutage ist es keine Seltenheit mehr, dass Datensätze potentiell sehr hoch dimensioniert sind und sehr viele Datenpunkte oder Variablen enthalten. Moderne Rechner sind häufig noch nicht in der Lage solche Datensätze in einer angemessenen Zeitspanne zu verarbeiten, wenn dies überhaupt möglich ist. Das Mooresche Gesetz besagt, dass sich die Rechenleistung in etwa alle zwei Jahre verdoppelt. Je besser und vor allem kleiner die Prozessoren werden, umso schwerer ist dies zu erreichen. In dem Artikel „More than Moore“ (Waldrop 2016) gibt der Autor an, dass sich diese Verbesserungen in den nächsten Jahren, in aller Voraussicht, auf die bisherige Weise nicht mehr erreichen lassen werden. Das Wachstum von Datenmengen wird laut Gantz und Reinsel (2012) diese Entwicklung hingegen einhalten und das digitale Datenvolumen wird sich bis zum Jahre 2020 alle zwei Jahre verdoppeln. Daher ist es nicht möglich, große Datenmengen nur über technische Entwicklung zu bewältigen, sondern es bedarf neuen Verfahren um dies zu tun. In der Informatik gibt es für Regressionsverfahren den Ansatz, für die Reduktion von Datensätzen eine Stichprobe der Beobachtungen aus den Daten zu entnehmen und die Beobachtungen dabei proportional zu den Leverage Scores in die Stichprobe aufzunehmen.

Ein Anwendungsgebiet in dem regelmäßig beträchtliche Datenmengen anfallen ist die Genetik. Das menschliche Genom ist sehr umfangreich und daher können genetische Datensätze schnell zehntausende von Datenpunkten enthalten. Ein Einzelnukleotid-Polymorphismus, im Englischen Single Nucleotide Polymorphism, abgekürzt mit SNP (Snips gesprochen) ist die Veränderung eines einzelnen Basenpaares in einem DNA-Strang. Diese stehen in Verdacht, vor allem seltene Krankheiten auszulösen. Eine Besonderheit von SNP-Daten ist, dass ein SNP alleine meistens nicht für das Bilden einer Krankheit verantwortlich ist, sondern erst wenn mehrere SNPs in gewissen Wechselwirkungen vorliegen. Ein Verfahren, das speziell für die Suche von komplexen Wechselwirkungen gedacht ist, ist die logische Regression. In dieser Arbeit geht es darum, Datensätze für die logische Regression zu reduzieren, entsprechend den Rechenaufwand bzw. die Rechenzeit zu verringern und trotzdem noch gute Ergebnisse zu erzielen.

Es wird sich zeigen, dass die Leverage Scores grundsätzlich beim selektieren von Beobachtungen und Variablen helfen können. In jeder Datensituation ist es möglich, eine Gewichtung der Leverage Scores zu finden, die dazu führt, dass es eine Verbesserung bei den angepassten Modellen gibt, im Vergleich dazu die Beobachtungen oder Variablen durch eine einfache Zufallsauswahl zu wählen. Weiter helfen die Leverage Scores dabei, in der Gruppe der Erkrankten weniger stark besetzte Untergruppen besser zu finden. Im Kontext der Variablenselektion sind vor allem die Cross Leverage Scores mit der abhängigen Variable dazu geeignet, die wichtigen Einflussvariablen zu identifizieren.

Diese Arbeit ist wie folgt strukturiert: Im nächsten Kapitel werden die Problemstellung und die Ziele dieser Arbeit genauer vorgestellt. Im dritten Kapitel wird die logische Regression eingeführt. Dazu werden die Grundlagen wie Logikausdrücke und Logikbäume eingeführt, das logische Regressionsmodell formal definiert und der logicFS-Algorithmus

vorgestellt, um das Modell an die SNP-Daten anzupassen. Das vierte Kapitel beschäftigt sich mit SNP-Daten, deren Besonderheiten und welche Daten für diese Arbeit verwendet werden. Das fünfte Kapitel bildet den Hauptteil dieser Arbeit und beschäftigt sich mit der Reduktion der Datensätze. Darin wird einführend das Vorgehen zum Reduzieren der Datensätze vorgestellt und in diesem Zusammenhang die Leverage Scores erklärt. Das Kapitel ist weiter unterteilt in die beiden Fälle, die bei SNP-Daten auftreten können: einmal die Situation, dass mehr Beobachtungen als Variablen vorliegen und dass es mehr Variablen als Beobachtungen gibt.

2 Problemstellung und Zielsetzung

Es kommt häufig vor, dass Datensätze von enormen Umfang vorliegen. Durch immer bessere Techniken und stärkere Vernetzungen, nicht zuletzt durch das Internet ist es inzwischen möglich, Datensätze von überwältigenden Ausmaßen in nur wenigen Sekunden zur Verfügung zu haben. Insbesondere gilt dies für die Genetik. Mit immer feineren und besseren Verfahren ist es möglich, große Teile des menschlichen Erbgutes zu entschlüsseln. Spezielle Projekte wie das HapMap-Projekt (siehe etwa Graw 2015, S. 512) und das 1000 Genomes Projekt (The 1000 Genomes Project Consortium 2015) haben sich genau diesem Ziel verschrieben. Die Analyse und der Umgang mit solchen großen Datenmengen stellt immer noch eine Herausforderung dar. Ziel ist es, die Daten in einer angemessenen Zeit „gut genug“ zu analysieren. Was dies bedeutet und wie sich dies eventuell erreichen lässt, ist Aufgabe der Forschung. Es geht darum, Mittel und Verfahren zu finden bzw. zu untersuchen, die dazu geeignet sind.

Diese Arbeit beschäftigt sich mit dem Thema, hochdimensionale Datensätze für die logische Regression zu reduzieren. Ein häufiges Anwendungsgebiet für die logische Regression findet sich in der Genetik bei sogenannten SNP-Daten (siehe **Kapitel 4**). Dabei handelt es sich um die Änderung eines einzelnen Basenpaares in einem DNA-Strang. Diese Änderungen kommen relativ häufig vor und stehen im Verdacht, seltene Erkrankungen wie etwa Krebs auszulösen. Da SNPs sich häufig gegenseitig beeinflussen und somit in Wechselwirkung zueinander stehen, bietet sich zu deren Auswertung die logische Regression an, da dieses Regressionsverfahren speziell nach Wechselwirkungen hoher Ordnung sucht (siehe **Kapitel 3**). Bei der logischen Regression geht es darum, Kombinationen von binären Variablen zu finden, die einen Einfluss auf die Zielgröße haben. Bei der Zielgröße handelt es sich häufig um eine Art von Klassifikation wie der Status: „die Krankheit liegt vor“ bzw. „die Krankheit liegt nicht vor“.

Als grundlegende Situation geht es um Datensätze mit n Personen (Beobachtungen). An jeder Person werden insgesamt d SNPs (Variablen) bestimmt. Aufgrund der besonderen Gegebenheit von SNP-Datensätzen sind zwei Situationen möglich: es gibt mindestens so viele Personen wie SNPs und somit ist $n \geq d$ oder es gibt mehr SNPs als Personen und somit ist $n < d$. Im zweiten Fall wird es sogar häufig der Fall sein, dass es deut-

lich mehr SNPs als Personen gibt. Grund für diese zwei Situationen ist der große Umfang des menschlichen Genoms, der weit in die zehntausende reicht. Somit kann es etwa sein, dass eine Auswahl von SNPs, von denen vermutet wird Auslöser einer bestimmten Krankheit zu sein, an einer Personengruppe bestimmt wurde, etwa im Rahmen einer Fall-Kontroll-Studie. Im nächsten Schritt geht es dann darum, zu identifizieren welche Wechselwirkungen zwischen den SNPs potentiell für die Krankheit verantwortlich sind. Andersherum kann es etwa darum gehen (wie bei dem HapMap-Projekt) eine Sammlung des menschlichen Genoms anzulegen, ohne eine a priori Vermutung über die Einflüsse der SNPs aufzustellen. Diese Dichotomie des Datenmaterials führt dazu, dass es in diesen beiden Fällen um unterschiedliche Problemstellungen geht.

Der Grund für die Unterscheidung der beiden Fälle ist die Verwendung der Leverage Scores. Die logische Regression selbst ist nicht darauf angewiesen, dass es mehr Beobachtungen n als Variablen d gibt (anders als etwa in der linearen Regression). Für die Berechnung der Leverage Scores ist es jedoch notwendig, dass $n \geq d$ ist (siehe **Kapitel 5.1**).

In dem Fall $n \geq d$ geht es darum, aus einem gegebenen Datensatz eine Stichprobe vom Umfang n' ($n' < n$) zu ziehen, anhand derer ein logisches Regressionsmodell angepasst wird. Untersucht werden soll, ob es einen Einfluss hat, die Beobachtungen proportional zu ihren Leverage Scores in die Stichprobe aufzunehmen, im Vergleich dazu jede Person mit der selben Gewichtung zu wählen. Als Referenz für die Güte des Modells wird die Klassifikationsrate k des Modells verwendet. Das Modell wird an der sogenannten Lernstichprobe vom Umfang n' angepasst und die restlichen $n'' = n - n'$ Beobachtungen des Datensatzes die nicht Teil der Lernstichprobe sind, mit dem Modell evaluiert und entsprechend klassifiziert. Der Anteil der richtig klassifizierten Personen bildet die Klassifikationsrate. Je höher diese ist, desto besser ist die Güte des Modells. Die Klassifikationsrate liegt zwischen 0 und 1 bzw. bildet einen Prozentanteil, wobei ein Wert von 1 bedeutet, dass 100% der Beobachtungen richtig klassifiziert werden und ein Wert von 0, dass 0% der Beobachtungen richtig klassifiziert werden. Der Referenzwert für die Klassifikationsrate ist der Wert 0,5 bzw. 50%. Bei einer Klassifikationsrate von 50% ist das Modell in dem Sinne ungeeignet, dass es genauso gut möglich wäre, die Personen mit der selben Wahrscheinlichkeit zufällig in die Klassen einzuteilen. Bei einer komplett zufälligen Klassifikation würden bei zwei gleichstark besetzten Klassen, erwartet etwa 50% der Personen richtig klassifiziert. Bei einer Rate von unter 50% wäre es bei der Klassifikation in zwei Klassen möglich, die Vorhersage des Modells umzukehren und somit eine Richtigklassifizierung von über 50% zu erhalten. Mit K wird die absolute Anzahl von richtig klassifizierten Beobachtungen bezeichnet.

Bei dem Fall $n < d$ geht es nicht mehr darum Personen für eine Stichprobe proportional zu ihren Leverage Scores zufällig in die Stichprobe aufzunehmen, sondern eine Stichprobe von Variablen vom Umfang d' ($d' < d$) zu entnehmen und anhand derer, ein logisches Regressionsmodell anzupassen. Es geht somit effektiv um eine Variablenselek-

tion mit Hilfe der Leverage Scores. Die Anzahl der Personen in dem Datensatz bleibt in diesem Fall unverändert. Weiterhin soll als Gütemaß die Klassifikationsrate verwendet werden. Da es jedoch nicht möglich ist, wie im obigen Fall beschrieben, für die Stichprobe die übriggebliebenen Beobachtungen zu verwenden, ist es nötig, die Lern- bzw. Teststichprobe vorher zufällig zuzuteilen. Nachdem die Personen zufällig in diese zwei Stichproben eingeteilt wurden, wird anhand der d' selektierten Variablen ein logisches Regressionsmodell mit den Personen der Lernstichprobe angepasst und die Personen der Teststichprobe anschließend klassifiziert.

In beiden Fällen wird im Rahmen dieser Arbeit untersucht, ob die Leverage Scores einen positiven Einfluss auf die Wahl der Beobachtungen bzw. Variablen im dem Sinne haben, dass sich die Klassifikationsrate durch die Wahl anhand der Leverage Scores verbessert und welche Modelle sich durch diese Auswahl ergeben.

Entstanden ist diese Arbeit im Rahmen des Sonderforschungsbereiches 876 „Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung“[1], in dem Teilprojekt C4 „Regressionsverfahren für sehr große, hochdimensionale Daten“ an der Technischen Universität Dortmund, im Zeitraum vom Juli bis Dezember 2016.

Sämtliche Berechnungen und Grafiken in dieser Arbeit sind mit der Statistik Software R in der Version 3.3.1 erstellt worden (R Core Team 2016).

3 Die logische Regression

Die logische Regression (logic regression) ist eine Form der Regressionsanalyse mit binären Einflussvariablen (Ruczinski et al. 2003, S. 475ff). Ziel der logischen Regression ist, die abhängige Variable des Regressionsmodells durch Kombinationen von binären Einflussvariablen auszudrücken. Bei vielen Regressionsverfahren ist es häufig der Fall, dass die Zielvariable als eine einfache Kombination (linear oder durch eine Link-Funktion) der Einflussvariablen ausgedrückt wird und Wechselwirkungen zwischen den Variablen niedrig gehalten werden. Jedoch kann es vorkommen, dass besonders Wechselwirkungen hoher Ordnung für den Einfluss der Variablen auf die Zielvariable verantwortlich sind. Dies kann etwa bei binären Einflussvariablen der Fall sein. Ein gutes Beispiel dafür und daher auch besonderes Augenmerk dieser Arbeit sind SNP-Daten (siehe **Kapitel 4**). Bei genetischen Daten sind häufig die Einflüsse einzelner Gene und vor allem deren Wechselwirkungen ausschlaggebend für das Bilden (seltener) Krankheiten. Eine Analyse sämtlicher potentieller Wechselwirkungen und Kombinationen von Wechselwirkungen stellt eine große Herausforderung dar, vor allem bei sehr großen Datensätzen.

Formal geht es darum, für eine Anzahl d binärer Einflussvariablen $X_1, ..., X_d$, eine bessere Vorhersage für die abhängige Variable Y durch die Einflussvariablen zu erhalten, indem neue Einflussvariablen $L_1, .., L_m$ konstruiert werden, die Kombinationen der binären Einflussvariablen bilden, wie beispielsweise: „X_1, X_2, X_3 UND X_4 sind WAHR (gleich 1)“

[1] http://sfb876.tu-dortmund.de/index.html

oder „X_5 ODER X_6 ABER NICHT X_7 sind WAHR (gleich 1)“ (vgl. Ruczinski et al. 2003, S. 475). Die Beziehung des Verfahrens zu solchen Logikausdrücken führt zu ihrer Namensgebung.

Ein Überblick anderer Verfahren, binäre Daten zu modellieren findet sich bei Ruczinski et al. (2003, S. 479f). Diese Verfahren verwenden häufig Entscheidungsbäume bzw. Entscheidungsregeln basierend auf Booleschen Funktionen.

Dieses Kapitel stellt die logische Regression vor. Im ersten Unterkapitel werden die Grundlagen für die logische Regression gelegt, in erster Linie Logikausdrücke und damit verbunden deren Darstellung als Logikbäume. Das zweite Unterkapitel stellt das logische Regressionsmodell formal vor. Abschließend geht es um das Anpassen logischer Regressionsmodelle und in diesem Zuge um das Simulated-Annealing und im Kontext von SNP-Daten den logicFS-Ansatz.

3.1 Logikausdrücke

Bei der logischen Regression geht es darum, Kombinationen von Binärvariablen zu finden, die eine gute Vorhersage für die Zielvariable geben. Der nachfolgende Abschnitt über Logikausdrücke basiert auf Ruczinski et al. (2003, S. 477ff). Kombinationen binärer Variablen sind Verknüpfungen Boolescher Logikausdrücken der Form $L = (X_1 \wedge X_2) \vee X_3^{\mathrm{C}}$ (vgl. Ruczinski et al. 2003, S. 477). Boolesche Variablen und Ausdrücke sind solche, die nur zwei mögliche Zustände (wie bei Binärvariablen der Fall) annehmen können. Der Ausdruck L liest sich: „X_1 UND X_2 sind WAHR ODER X_3 ist NICHT WAHR“. Als Kombination binärer Variablen ist der Ausdruck entsprechend wieder binär. Dies bedeutet, dass der Ausdruck den Wert 1 (Wahr) annimmt, wenn die Bedingungen erfüllt sind und den Wert 0 (Falsch) andernfalls. Nachfolgend findet sich eine Einführung in die Boolesche Algebra basierend auf der Formulierung von Ruczinski et al. (siehe Ruczinski et al. 2003, S. 477):

- Die einzigen Werte, die angenommen werden können, sind 1 und 0, sinngemäß als Wahr und Falsch, Ja und Nein oder An und Aus zu verstehen.
- Die Variablen werden mit X_i ausgedrückt. Der Index i gibt an, um welche Variable $i = 1, ..., d$ des Modells es sich handelt. Jede Realisation der Variable nimmt einen der beiden möglichen Werte an, weshalb diese als Binärvariablen bezeichnet werden.
- Es gibt drei logische Operatoren mit denen sich die Binärvariablen kombinieren und in einen Zusammenhang setzen bzw. verknüpfen lassen: $\wedge$ (UND), $\vee$ (ODER), $^{\mathrm{C}}$ (NICHT). Dabei bezeichnet X_i^{C} das Komplement von X_i, $i = 1, ..., d$. Der Wert der Binärvariable wird durch diesen Operator umgekehrt.
- Ein logischer Ausdruck ist ein Ausdruck der Form $X_1 \wedge X_2^{\mathrm{C}}$, also entsprechend eine Kombination der Binärvariablen durch die Operatoren.
- Als Gleichungen werden Logikausdrücke mit einem Namen L der Form $L = X_1 \wedge X_2^{\mathrm{C}}$ bezeichnet, um auf diesen Logikausdruck zu verweisen.

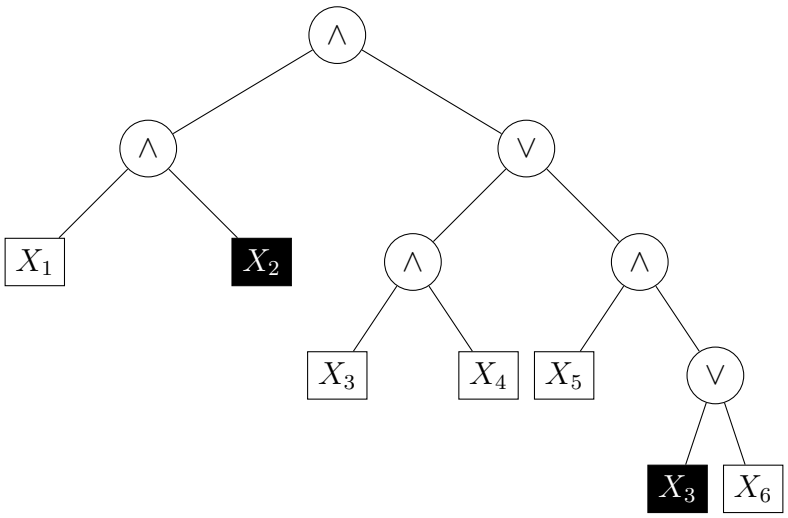

Abbildung 1: Repräsentation eines Logikbaumes für den Logikausdruck $L = (X_1 \wedge X_2^{\mathrm{C}}) \wedge [(X_3 \wedge X_4) \vee (X_5 \wedge (X_3^{\mathrm{C}} \vee X_6))]$. Mit schwarz hinterlegte Blätter stehen für das Komplement der entsprechenden Binärvariable.

Durch das Verwenden von Klammern lassen sich beliebige Boolesche Ausdrücke erzeugen, indem nacheinander zwei Binärvariablen, eine Binärvariable und ein anderer Boolescher Ausdruck oder zwei Boolesche Ausdrücke miteinander verknüpft werden. Nach Ruczinski et al. (2003, S. 477) ist

$$L = (X_1 \wedge X_2^{\mathrm{C}}) \wedge [(X_3 \wedge X_4) \vee (X_5 \wedge (X_3^{\mathrm{C}} \vee X_6))]$$

Beispiel eines solchen Logikausdruckes. Derartige Ausdrücke lassen sich als Logikbäume (binäre Bäume) darstellen. Der zu L gehörende Logikbaum ist in **Abbildung 1** dargestellt. Die Auswertung eines solchen Baumes als Logikausdruck beginnt von unten nach oben. Die Vorhersage des Baumes (ob er den Wert 1 oder 0 zugewiesen bekommt) entscheidet sich nach der Evaluation von unten an der Wurzel (dem obersten Knoten, siehe weiter unten) des Baumes.

Für Logikbäume gelten die nachfolgenden Beschreibungen und Regeln:

- Jedes Element eines Logikausdruckes (Binärvariablen, Komplemente der Binärvariablen, Operatoren) wird durch einen Knoten repräsentiert.
- Jeder Knoten besitzt entweder genau zwei oder keinen Nachfolgeknoten.
- Ein Nachfolgeknoten wird als Kind bezeichnet, der Knoten selbst als Mutter des Nachfolgeknotens. Zwei Nachfolgeknoten des selben Mutterknotens heißen Geschwister.
- Der Knoten ohne Mutterknoten ist die Wurzel des Baumes.
- Knoten ohne Kinder heißen Blätter.

- In den Blättern des Baumes befinden sich nur Binärvariablen bzw. deren Komplemente. In allen anderen Knoten befinden sich $\wedge$-/$\vee$-Operatoren.

Der in **Abbildung 1** dargestellte Logikbaum erfüllt alle diese Regeln. Da ein Logikausdruck nicht eindeutig ist, ist entsprechend auch ein Logikbaum nicht eindeutig (vgl. Ruczinski et al. 2003, S. 478). Da der Ausdruck uneindeutig ist, es aber vorteilhafter ist eine eindeutige Schreibweise zu verwenden, gibt es als eine eindeutige Schreibweise die sogenannte Disjunktive Normalform (DNF). Diese ist eine Art der Notation für Logikausdrücke, in dem Boolesche Ausdrücke als $\vee$-Kombinationen von $\wedge$-Ausdrücken geschrieben werden (siehe Ruczinski et al. 2003, S. 505). Der Logikausdruck

$$L_{DNF} = (X_1 \wedge X_2^{\mathrm{C}} \wedge X_3 \wedge X_4) \vee (X_1 \wedge X_2^{\mathrm{C}} \wedge X_3^{\mathrm{C}} \wedge X_5) \vee (X_1 \wedge X_2^{\mathrm{C}} \wedge X_5 \wedge X_6)$$

ist die zu L gehörende DNF. Der Vorteil der DNF ist, dass sich die Wechselwirkungen zwischen den Variablen direkt anhand der $\wedge$-Kombinationen ablesen lassen. Ein Logikbaum dieses Ausdruckes wäre komplexer und würde mehr Äste enthalten. Eine Evaluation des Baumes mit gegebenen Werten würde aber zum selben Ergebnis führen (vgl. Ruczinski et al. 2003, S. 478). Der in **Abbildung 1** dargestellte Logikbaum würde somit beispielsweise den Wert 1 (Wahr) erhalten, wenn die Variablen die Bedingung erfüllen: „X_1 UND NICHT X_2 UND X_3 UND X_4 sind WAHR (gleich 1)“.

Als nächstes wird das logische Regressionsmodell formal vorgestellt.

3.2 Das logische Regressionsmodell

Basierend auf der Einführung von Logikausdrücken (siehe **Kapitel 3.1**) wird nun das logische Regressionsmodell definiert. Es sind $X_1, ..., X_d$ die binären Einflussvariablen des Modells und Y die Zielvariable (nicht notwendigerweise binär). Das logische Regressionsmodell hat nach Ruczinski et al. (2003, S. 479) die folgende Form:

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j.$$

Dabei sind β_j, $j = 0, ..., t$ die Parameter des Modells, g eine Link-Funktion und L_j, $j = 1, ..., t$ Logikausdrücke bestehend aus Booleschen Kombinationen der Einflussvariablen. Dieses Modell berücksichtigt gleichzeitig mehrere Bäume, wobei jeder Logikausdruck L_j zu einem eigenen Logikbaum korrespondiert. Die Link-Funktion umfasst zum Beispiel die lineare Regression $g(E[Y]) = E[Y]$ und die logistische Regression mit der Logit-Funktion als Link-Funktion. Da es in dieser Arbeit primär um Klassifikation geht, ist die Link-Funktion für diese Arbeit die Logit-Funktion:

$$g(E[Y]) = logit(E[Y]) = ln\left(\frac{E[Y]}{1 - E[Y]}\right).$$

Für jeden Typ von Modell ist es nötig eine Score-Funktion zu definieren, welche die Güte des gewählten Modells misst. Für die lineare Regression ist dies die Fehlerquadratsumme und für die logistische Regression die binomial Deviance.

Eine Anpassung des Modells entspricht dem Finden der Logikausdrücke L_j, $j = 1, ..., t$ und gleichzeitig den Parametern β_j, $j = 0, ..., t$, welche die Score-Funktion minimieren. Nachfolgend geht es um das Anpassen logischer Regressionsmodelle.

3.3 Anpassen von logischen Regressionsmodellen

Für einen binären Datensatz stellt sich die Herausforderung, das beste bzw. für die Daten geeignetste logische Regressionsmodell zu finden. Die Anzahl möglicher Bäume ist sehr groß und es gibt kein eindeutiges Kriterium, alle möglichen Bäume zu berücksichtigen (vgl. Ruczinski et al. 2003, S. 481). Auch ist es nicht möglich bzw. praktikabel sämtliche Logikbäume zu bestimmen.

Da eine solche globale Betrachtung nicht praktikabel ist, geht es nun darum, wie vorgegangen werden kann um gut passende Modelle zu finden. Grundsätzlich geht es darum, auf der Oberfläche der Daten mögliche Bäume zu konstruieren. Dies passiert durch eine Anzahl von Bewegungen, die von einem Baum in einen möglichen Nachbarbaum übergehen. Ein Nachbarbaum ist dabei der Baum, der sich durch eine einzelne Bewegung erreichen bzw. konstruieren lässt. Jede Bewegung hat eine Gegenbewegung mit der sich die Veränderung eindeutig umkehren lässt. Für den Übergang von einen Baum zu seinem Nachbarn schlagen Ruczinski et al. (2003, S. 481f) die nachfolgenden Bewegungen vor. Illustriert werden die Bewegungen anhand des Beispieles von Ruczinski et al. (2003, S. 481), um eine Veranschaulichung der möglichen Bewegungen zu geben. In **Abbildung 2** ist beispielhaft ein Logikbaum dargestellt, der durch die Bewegungen entsprechend in einen Nachbarbaum verändert wird. Der zu dem Baum gehörende Logikausdruck lautet $L = X_1^{\mathrm{C}} \wedge (X_2 \vee X_3)$ und in eindeutiger Schreibweise $L_{DNF} = (X_1^{\mathrm{C}} \wedge X_2) \vee (X_1^{\mathrm{C}} \wedge X_3)$. Die schwarz hinterlegten Blätter repräsentieren das Komplement der entsprechenden Binärvariable.

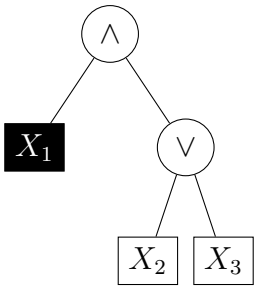

Abbildung 2: Ursprünglicher Baum des Logikausdruckes $L = X_1^{\mathrm{C}} \wedge (X_2 \vee X_3)$ zur Illustration der möglichen Bewegungen, um in einen Nachbarbaum zu gelangen.

- **Ändern eines Blattes**: Ein Blatt des Baumes wird ausgewählt und durch ein anderes Blatt an derselben Position ersetzt. Um Dopplungen zu vermeiden, ist es nicht möglich, ein Blatt mit seinem Geschwisterknoten oder dessen Komplement zu ersetzen, wenn dieser Knoten ebenfalls ein Blatt ist. Die Gegenbewegung ist es, das Blatt in seinen ursprünglichen Zustand zurück zu bringen. In **Abbildung 3** ist das Blatt mit der Binärvariable X_2 durch ein Blatt mit der Binärvariable X_4^{C} ersetzt worden.

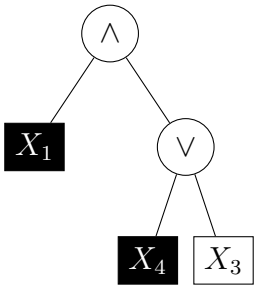

Abbildung 3: Resultierender Nachbarbaum durch das Ändern eines Blattes.

- **Ändern eines Operators**: Jeder Logikoperator $\wedge$ auf einem Knoten kann durch den Operator $\vee$ ersetzt werden und umgekehrt. Diese Bewegung ist gleichzeitig ihre Gegenbewegung. In **Abbildung 4** ist der $\wedge$-Operator in der Wurzel des ursprünglichen Baumes durch den $\vee$-Operator ersetzt worden.

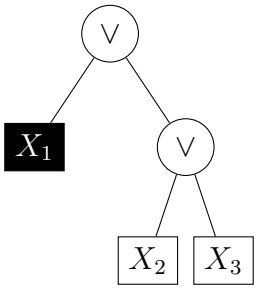

Abbildung 4: Resultierender Nachbarbaum durch das Ändern eines Logikoperators.

- **Wachsen und Stutzen**: Aus jedem Knoten, der kein Blatt ist, kann ein neuer Arm wachsen. Ein Arm ist eine Verbindung von einem Knoten zu einem anderen Knoten oder Unterbaum. Ein Unterbaum ist selbst wieder ein Baum, der durch einen Arm mit einem anderen Baum verbunden ist. Zum Wachsen wird der Rest des Baumes an diesem Knoten zum rechten Arm des Knotens und der linke Arm wird zu einem Blatt mit einer beliebigen Binärvariablen. Diese beiden Unterbäume werden durch einen Knoten mit einem Logikoperator verbunden. In **Abbildung 5 (a)** ist aus dem $\vee$-Operator des ursprünglichen Baumes aus **Abbildung 2** ein

neuer Arm gewachsen. Die Gegenbewegung vom Wachsen ist das Stutzen. Dazu wird ein Blatt aus dem Baum entfernt und der Unterbaum am Geschwisterknoten des entfernten Blattes wird an den Mutterknoten gerückt. In **Abbildung 5 (b)** wurde der ursprüngliche Baum an dem ∧-Operator gestutzt.

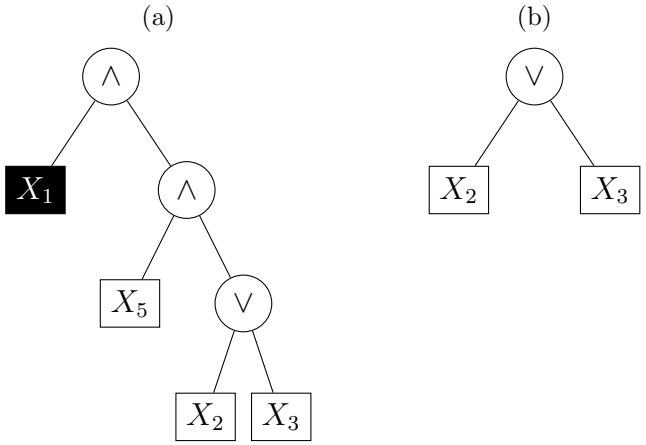

Abbildung 5: Resultierender Nachbarbaum aus: **(a)** dem Wachsen eines neuen Armes, **(b)** dem Stutzen eines Armes.

- **Aufteilen und Löschen**: Jedes Blatt eines Baumes kann aufgeteilt werden, indem ein Geschwisterknoten bzw. Blatt hinzugefügt wird und ein gemeinsamer Mutterknoten festgelegt wird. In **Abbildung 6 (a)** wurde das Blatt mit der Binärvariable X_3 aufgeteilt und das Geschwisterblatt mit der Binärvariable X_6^{C} hinzugefügt. Die Gegenbewegung ist das Löschen eines Blattes. An einem Mutterknoten mit zwei Blättern wird eines dieser Blätter entfernt. In **Abbildung 6 (b)** wurde das Blatt mit der Binärvariable X_3 aus dem ursprünglichen Baum entfernt.

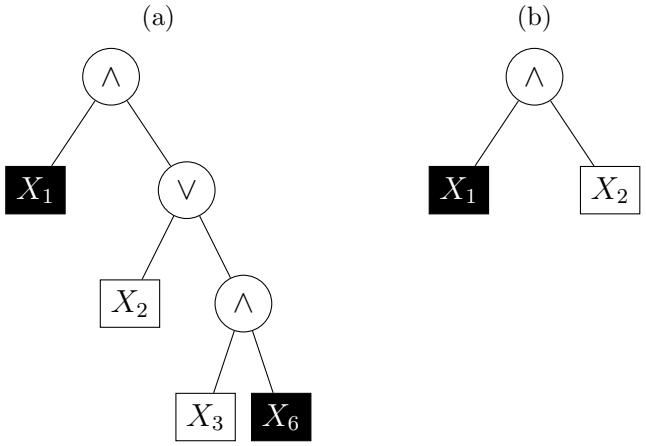

Abbildung 6: Resultierender Nachbarbaum aus: **(a)** Aufteilen eines Blattes, **(b)** Löschen eines Blattes.

Mit diesen Bewegungen kann jeder Logikbaum von jedem anderen Logikbaum in einer endlichen Anzahl von Schritten erreicht werden (vgl. Ruczinski et al. 2003, S. 482). Dies ist wichtig für die zugrundeliegende liegende Theorie der Markov-Ketten. Es wäre sogar möglich, die Anzahl der Bewegungen zu reduzieren. So sind nach Ruczinski et al. (2003, S. 482) etwa das Stutzen und Wachsen nicht notwendig für die Markov-Ketten Theorie, jedoch verbessert sich die Anpassung der Modelle durch diese Bewegungen. Die maximale Anzahl der Bäume wird dabei als fest angenommen.

Der von Ruczinski et al. (2003) vorgestellte Ansatz zum Finden eines Baumes ist das auf Monte-Carlo-Simulation basierende „Simulated-Annealing“ (siehe Ruczinski et al. 2003, S. 481ff). Für diese Arbeit wird der erweiterte logicFS-Ansatz von Schwender und Ickstadt verwendet (vgl. **Kapitel 3.3.2**, siehe Schwender und Ickstadt 2008), da dieser speziell für SNP-Daten gedacht ist. Nachfolgend werden die Ansätze zum Anpassen der Modelle vorgestellt.

3.3.1 Simulated-Annealing

Das sogenannte Simulated-Annealing ist ein auf Markov-Ketten basierender Suchalgorithmus und ist dazu geeignet, logische Regressionsmodelle anzupassen. Definiert ist der Algorithmus nach Ruczinski et al. (2003, S. 485) auf einem Zustandsraum $\mathcal{S}$, der eine Ansammlung von möglichen Zuständen s der Modelle ist. Jeder dieser Zustände repräsentiert die möglichen Konfigurationen des zu bestimmenden Problems. Die Zustände sind verbunden durch ein Nachbarschafts-System und die Menge der paarweisen Nachbarzustände bildet eine Teilmengenalgebra M in $\mathcal{S} \times \mathcal{S}$. Die Elemente in M heißen Bewegungen. Zwei Zustände s, s' sind benachbart, falls sie jeweils mit einer Bewegung zu erreichen sind. Entsprechend sind zwei Zustände durch i-Schritte benachbart, falls sie mit i Bewegungen zu erreichen sind. Für die Anwendung der logischen Regression ist der Zustandsraum endlich.

Grundsätzlich ist die Idee des Algorithmus, von einem Zustand durch zugelassene Bewegungen in Nachbarzustände überzugehen und zu überprüft, ob sich der durch eine Score-Funktion definierte Score ϵ verbessert oder verschlechtert. Wenn der neue Score besser ist als der vorherige, wird die Bewegung durchgeführt. Ist der neue Score nicht besser, wird die Bewegung mit einer gewissen Wahrscheinlichkeit durchgeführt. Diese Wahrscheinlichkeit hängt von dem Score der beiden verglichenen Zuständen ab und einem Parameter, der den Zeitpunkt der Markov-Kette berücksichtigt. Je weiter die Kette voranschreitet umso unwahrscheinlicher wird ein Übergang, wenn der Score des Nachfolgezustandes unter dem des ursprünglichen Zustandes liegt. Nach Ruczinski et al. (2003, S. 485) liefert dieser Algorithmus gute Werte der Score-Funktion. Zusätzlich zu der Score-Funktion und den möglichen Bewegungen ist es noch nötig, die Auswahlwahrscheinlichkeiten für eine entsprechende Bewegung festzulegen, eine Akzeptanzfunktion und einen Abkühlungsmechanismus für die Markov-Kette, der die Akzeptanzwahrscheinlichkeit mit beeinflusst. Grundsätzlich besitzt jeder Schritt des Algorithmus eine Temperatur T die

angibt, wie weit die Markov-Kette vorangeschritten ist und entsprechend, wie viele Iterationen der Algorithmus ausgeführt hat. Zur Definition der Akzeptanzwahrscheinlichkeit einer neuer Bewegung bezeichnet ϵ_{alt} den Score des aktuellen Zustandes und ϵ_{neu} den Score des möglichen neuen Zustandes nach einer Bewegung, wobei niedrigere Werte einen besseren Score bedeuten. Der neue Zustand wird dann mit der Wahrscheinlichkeit

$$\theta = \min\{1, exp([\epsilon_{alt} - \epsilon_{neu}]/T)\}$$

akzeptiert. Die Akzeptanzwahrscheinlichkeit beeinflusst nach Ruczinski et al. (2003, S. 507) die Güte des Algorithmus und die Autoren geben an, dass die besten Resultate durch das Ändern eines Blattes (vgl. **Abbildung 3**) erzielt werden. Zudem beeinflusst das Ändern eines Logikoperators (vgl. **Abbildung 4**) die Vorhersage des Baumes beträchtlich und erhält dadurch nur eine geringere Akzeptanzwahrscheinlichkeit. Die Temperatur kann auf zwei fundamental unterschiedliche Weisen gehandhabt werden. Zum einen kann die Temperatur nach jedem Schritt weiter reduziert werden oder für eine gewisse Anzahl Schritte konstant gehalten werden und erst dann in größerem Umfang reduziert werden. Die Autoren verweisen darauf, dass es keine eindeutig bessere Wahl für das Reduzieren der Temperatur gibt, sich jedoch etwas bessere Ergebnisse mit dem temporären Konstanthalten der Temperatur erzielen lassen.

Zu Beginn des Algorithmus und zu dessen Ende werden durch eine hohe bzw. niedrige Temperatur fast alle Bewegungen akzeptiert bzw. abgelehnt. Um den Algorithmus zu beschleunigen und nicht zu viel Zeit zu Beginn bzw. zum Ende zu verbringen, gibt es zwei Vorgehensweisen. Bei fester Temperatur wird nach jedem akzeptierten Schritt die Anzahl der Schritte gezählt und wenn eine Schwelle überschritten ist, wird die Temperatur gesenkt. Dies geschieht, damit der Algorithmus nicht zu viel Zeit in zufälligen Modellen verbringt. Die Schwelle liegt bei 1% oder 10% der Iterationsschritte des Algorithmus. Zu späteren Zeitpunkten des Algorithmus wird diese Schwelle nicht erreicht und der Algorithmus erreicht das Ende der Markov-Kette.

Jeder Logikbaum hat eine endliche Anzahl an Nachbarbäumen. Gegen Ende des Algorithmus mit einer niedrigen Temperatur werden nur noch wenige Bewegungen akzeptiert und da das Verfahren zufallsbasiert ist kann es passieren, dass eine Bewegung mehrmals abgelehnt wird, bevor sie akzeptiert wird. Der größte zeitliche Aufwand in Sinne der Berechnungszeit ist das Auswerten der Logikbäume und das Bestimmen des Scores. Um die Rechenzeit des Algorithmus weiter zu beschleunigen, wird zusätzlich ein zweiter Prozess ausgeführt, der die Scores aller bereits besuchten Bäume speichert. Entsprechend muss jeder mögliche Status nur einmal bestimmt werden und danach nur noch der Score abgerufen werden. Dies beschleunigt nach Ruczinski et al. (2003, S. 509) den Algorithmus deutlich bei niedriger Temperatur.

Das Vorgehen des Algorithmus für die logische Regression ist nach Ruczinski et al. (2003, S. 486), alle Bäume des Regressionsmodells parallel zu bestimmen, einen dieser Bäume zufällig auszuwählen und dann zufällig (gemäß einer vorher festgelegten Vertei-

lung) eine der vorgestellten Bewegungen (siehe **Kapitel 3.3**) zu nehmen, einen Nachbarbaum zu konstruieren und dann das logische Regressionsmodell erneut anzupassen. Der Score des neuen Modells wird dann mit dem Score des Modells (Zustandes) vor der Bewegung verglichen. Für dieses Vorgehen ist es nötig, dass die Anzahl der Bäume t zwar beliebig groß, jedoch fest ist. Theoretisch können die Logikbäume beliebig groß werden und erreichen irgendwann den bestmöglichen Score. Dies ist bei endlicher Rechenzeit jedoch nicht unbedingt möglich. Damit die Modelle gut interpretierbar bleiben, empfehlen Ruczinski et al. (2003, S. 509), die Anzahl der Blätter auf 8 oder 16 zu beschränken und maximal ein bis drei Bäume wachsen zu lassen, obwohl es auch möglich ist, den Algorithmus mit fünf Bäumen starten zu lassen.

Als nächstes geht es um den logicFS-Ansatz. Dieser ist eine alternativer Ansatz der für das Anpassen logischer Regressionsmodelle auf SNP-Daten gedacht ist. In diesem Ansatz wird das Simulated-Annealing auf Teilmengen der Daten angewandt.

3.3.2 Der logicFS-Ansatz

Der logicFS-Ansatz ist ein Ansatz für das Anpassen logischer Regressionsmodelle für SNP-Daten. Vorgestellt wurde der Ansatz von Schwender und Ickstadt (2008, S. 187ff) in dem Artikel „Identification of SNP interactions using logic regression“. Wie in dem Artikel angemerkt, liegt eine besondere Herausforderung bei SNP-Daten darin, Klassifikationsregeln der nachfolgenden Art zu bestimmen:

> „Wenn SNP A ein heterozygoter Genotyp ist UND SNP B ist ein homozygoter Genotyp ODER SNP C UND SNP D sind NICHT homozygot, hat eine Person ein höheres Risiko an einer bestimmten Krankheit zu erkranken.“

(vgl. Schwender und Ickstadt 2008, S. 188). Solche Klassifikationsregeln lassen sich mit der logischen Regression bestimmen. Jeder SNP lässt sich mit jeweils zwei Binärvariablen codieren, die den Typus der Base beschreiben (vgl. **Kapitel 4**). Im Kontext einer Fall-Kontroll-Studie wird mit Hilfe der logischen Regression nach einem Logikausdruck L gesucht, der den Krankheitsstatus am besten beschreibt. So könnte eine neue, nicht an der Studie beteiligte Person als krank eingestuft werden, wenn L für diese Person wahr ist. Bei der Analyse von SNP-Daten geht es somit darum, Interaktionen zwischen den einzelnen SNPs zu identifizieren und zu untersuchen, ob diese Interaktionen potentiell zu einem höheren Risiko für eine Erkrankung führen. Um die Logikausdrücke und somit die Wechselwirkungen der SNPs einfacher interpretieren zu können, werden für den logicFS-Ansatz alle Logikausdrücke in ihre DNF konvertiert.

Der logicFS-Ansatz ist im Gegensatz zum sogenannten Markov-Ketten-Monte-Carlo Ansatz von Kooperberg und Ruczinski (2005, S. 157ff) ein auf Teilmengen basiertes Verfahren, in dem das Simulated-Annealing (siehe **Kapitel 3.3.1**) auf unterschiedliche

Teilmengen der Daten angewandt wird. Der Markov-Ketten-Monte-Carlo-Ansatz ist eine Kombination der logischen Regression mit Markov-Ketten-Monte-Carlo-Simulationen. Ziel dabei ist es nicht das „beste“ Modell für SNP-Daten zu finden, sondern möglichst viele Modelle und dabei solche Wechselwirkungen zu identifizieren, die in einer Vielzahl der Modelle auftreten. Die Kritik von Schwender und Ickstadt an diesem Vorgehen ist, dass es eine Besonderheit bei SNP-Daten ist, dass manche Wechselwirkungen nur für einen Teil der Beobachtungen erklärend für die Krankheit sind. Diese Wechselwirkungen werden potentiell durch den Markov-Ketten-Monte-Carlo-Ansatz nur schwer gefunden.

Der Algorithmus für den logicFS-Ansatz ist nach Schwender und Ickstadt (2008, S. 190):

1. Ziehe eine Bootstrap-Lernstichprobe von Umfang n aus den n Beobachtungen des Datensatzes.

2. Passe ein logisches Regressionsmodell an die Daten der Lernstichprobe an.

3. Konvertiere jeden Logikausdruck in seine DNF bestehend aus Primimplikanten.

4. Wiederhole Schritte $1. - 3.$ insgesamt B mal.

Eine Bootstrap-Stichprobe ist eine Stichprobe aus den Daten, die mit Zurücklegen gezogen wird. Unter einem Primimplikanten ist eine Wechselwirkung bzw. Einzelvariable zu verstehen, deren DNF des Logikausdruckes sich nicht weiter vereinfacht lässt. Es können somit keine Binärvariablen oder Kombinationen von Binärvariablen entfernt werden, ohne die Bedeutung des Logikausdruckes zu verändern. Als Beispiel ist in dem Logikausdruck

$$L = (X_1 \wedge X_2 \wedge X_3) \vee (X_1 \wedge X_2^{\mathrm{C}} \wedge X_3) \vee (X_4 \wedge X_5)$$

die Binärvariable X_2 überflüssig und der Ausdruck lässt sich zu

$$L_{DNF} = (X_1 \wedge X_3) \vee (X_4 \wedge X_5)$$

vereinfachen. Entsprechend ist die DNF des Logikausdruckes die minimale $\vee$-Kombination von $\wedge$-Ausdrücken.

Bei SNP-Daten geht es nicht nur darum, potentielle Wechselwirkungen aufzufinden, sondern diese auch auf ihre Wichtigkeit hin zu bewerten. Einige der durch den Algorithmus identifizierten Wechselwirkungen sind potentiell sehr wichtig für die Klassifikation. Um die Wichtigkeit der Interaktionen quantifizieren zu können, muss ein Wichtigkeitsmaß definiert werden. Dabei gibt es jeweils ein Maß für den Fall, dass es nur einen einzelnen Baum im Modell gibt bzw. ein Maß für mehrere Bäume.

Die beiden Maße basieren auf der Klassifikation der sogenannten Out-Of-Bag (OOB) Beobachtungen. Dies sind in einem Bootstrap-Verfahren diejenigen Beobachtungen, die im b-ten Schritt, $b = 1, ..., B$, nicht Teil der Lernstichprobe sind. Der OOB-Fehler bezeichnet entsprechend den Anteil der Beobachtungen, die falsch klassifiziert werden.

Nach Schwender und Ickstadt (2008, S. 190) ist

$$VIM_{single} = \frac{1}{B}\left(\sum_{b:\mathcal{P}\in L_b}(N_b - N_b^-) + \sum_{b:\mathcal{P}\notin L_b}(N_b^+ - N_b)\right) \tag{1}$$

das Maß für einen einzelnen Baum. Dabei ist L_b die Menge der Primimplikanten identifiziert im b-ten Schritt des Algorithmus, $b = 1, ..., B$. N_b ist die Anzahl der korrekt klassifizierten OOB-Beobachtungen im b-ten Schritt, durch das in diesem Schritt konstruierte logische Regressionsmodell. Entsprechend sind N_b^- bzw. N_b^+ die Anzahl der OOB-Beobachtungen die richtig Klassifizierten werden, nachdem der $\mathcal{P}$-te Primimplikant aus dem Modell entfernt bzw. dem Modell hinzugefügt wird. Das Wichtigkeitsmaß misst somit den Einfluss des $\mathcal{P}$-ten Primimplikanten auf die Klassifikationsrate indem verglichen wird, wie gut das Modell die OOB-Beobachtungen klassifiziert, wenn $\mathcal{P}$ Teil des logischen Regressionsmodells ist bzw. nicht Teil des Modells ist.

Für den Fall, dass es mehrere Bäume gibt, ist es nicht eindeutig möglich, eine Wechselwirkungen einem speziellen Baum hinzuzufügen, da es nicht eindeutig ist, an welchem Baum dies geschehen soll. Aus diesem Grund wird der Primimplakant $\mathcal{P}$ nur aus den Modellen entfernt und nicht hinzugefügt. Zum Berechnen des Wichtigkeitsmaßes $VIM_{Multiple}$ werden zuerst die Anzahl der richtig klassifizierten OOB-Beobachtungen N_b für jeden der B Schritten bestimmt, dann wird $\mathcal{P}$ aus allen Modellen entfernt und abschließend die richtig klassifizierten OOB-Beobachtungen $N_b^\star$ berechnet. Das Wichtigkeitsmaß für mehrere Bäume ist nach Schwender und Ickstadt (2008, S. 191) durch

$$VIM_{Multiple} = \frac{1}{B}\sum_{b=1}^{B}(N_b - N_b^\star) = \frac{1}{B}\sum_{b:\mathcal{P}\in L_b}(N_b - N_b^\star) \tag{2}$$

gegeben. Große Werte durch diese zwei Wichtigkeitsmaße implizieren eine erhöhte Wichtigkeit für die Klassifikation, wohingegen Werte gegen 0 für eine geringe Wichtigkeit sprechen. Die Maße können negative Werte annehmen, was bedeutet, dass ein Aufnehmen der Wechselwirkung die Klassifikation verschlechtert.

Neben der Identifikation von Wechselwirkungen ist es möglich, den logicFS-Ansatz zur Klassifikation zu verwenden, da es eine Bagging-Variante der logischen Regression ist (vgl. Schwender und Ickstadt 2008, S. 197). Der Ausdruck Bagging steht für Bootstrap-Aggregating und ist eine Methode, um die Vorhersagen von Regressions- bzw. Klassifkationsverfahren miteinander zu kombinieren. Bei dem logicFS-Ansatz wird der Krankheitsstatus einer neuen Person durch die Wahl der Mehrheit (majority voting) bestimmt. Die Variablen der Beobachtung werden an den B Bäumen des Algorithmus ausgewertet. Die Beobachtung wird daraufhin in die Gruppe klassifiziert, welche durch die Mehrheit der Bäume prognostiziert wird.

Die Rechenzeit des Algorithmus wird durch mehrere Faktoren bestimmt: die Anzahl der Wiederholungen B, die Anzahl der Iterationen des Simulated-Annealing-Algorithmus, der maximalen Anzahl der Variablen im Modell und der maximalen Anzahl an Bäumen.

Diese Faktoren bestimmen ebenfalls die Genauigkeit der Modelle. Angepasst werden die Modelle in R mit der Funktion „logic.bagging“ aus dem R-Paket „logicFS“ (Schwender 2013).

4 SNP-Daten

In diesem Abschnitt geht es um sogenannte Einzelnukleotid-Polymorphismen, englisch Single Nucleotide Polymorphisms (SNPs gesprochen wie Snips). Es soll darum gehen, dem Leser einen Überblick über die Art des Datenmaterials zu geben und einen Einblick in die dahinterstehenden Problematik. Der Abschnitt basiert auf dem Buch „Genetik“ von Graw (2015).

Bei SNPs handelt es sich um einen Begriff aus der Genetik. In der Genetik geht es grundlegend laut Graw (2015, S. 2) um die „Aufklärung der Regeln und Mechanismen der Vererbung, [...] darüber hinaus auch [...] die Unterschiede in der genetischen Ausstattung verschiedener Organismen funktionell zu erklären“. Es geht somit darum, gewisse Eigenschaften von Organismen über ihre Vererbung zu erklären, wie etwa Erkrankungen und wie sich diese durch Vererbung bzw. durch eine Veränderung nach der Vererbung in den Genen potentiell erkennen lassen. Die Aufgabe der Statistik ist es in diesem Zusammenhang, das erhobene genetische Datenmaterial auszuwerten und dabei zu helfen, solche Wechselwirkungen zu erkennen und die Erhöhung des Krankheitsrisikos zu bestimmen. Weiter steht die Genetik nach Graw (2015, S. 2) „im Schnittpunkt anderer biologischer Disziplinen [...] und beeinflusst mit ihren methodischen Ansätzen diese Bereiche“. Entsprechend ist die Genetik ein sehr modernes und aktuelles Forschungsgebiet mit immer weiteren Entdeckungen und Fortschritten, vor allem auch im technologischen Bereich zum Aufschlüsseln der Gene.

Die Gesamtheit der vererbbaren Informationen eines Organismus befindet sich in dessen Genen. Die verschiedenen Formen eines Gens werden mit Allel bezeichnet, darunter fallen Kategorien wie normal oder mutiert. Der Träger der Erbanlagen ist das Chromosom. Bei höher entwickelten Organismen wie dem Menschen ist der Chromosomensatz diploid, also aus zwei Teilen bestehend. Chromosomen selbst bestehen aus Desoxyribonukleinsäure (DNA). Als Hauptkomponenten besteht die DNA aus vier heterozyklischen, organischen Basen: Adenin, Guanin, Cytosin und Thymin. Die Struktur der DNA ist eine Doppelhelix. Jeweils zwei Basen sitzen in der Struktur gegenüber, werden durch eine Wasserstoffbrücke zusammengehalten und bilden ein Basenpaar (bp). Sind die beiden Basen gleich, wird dieser Zustand als homozygot bezeichnet, unterscheiden sich die Basen wird dieser Zustand als heterozygot bezeichnet.

Ein SNP ist der Austausch eines einzelnen Nukleotids (Single Nucleotide, den Bausteinen der DNA) nach der Vererbung in einem DNA-Strang. Sie erlauben es, Unterschiede im Genom einer Spezies zu untersuchen. Nach Graw (2015, S. 496) kann davon ausgegangen werden, dass es sich bei etwa jedem tausendsten Basenpaar um ein SNP handelt.

Dies bedeutet, dass die Anzahl der SNPs im menschlichen Genom sehr hoch ist. Etwa 10 Mio. SNPs sind schon im menschlichen Genom bekannt (siehe Graw 2015, S. 511). Eine wichtige Erkenntnis von SNPs ist es, dass diese nicht unabhängig sind, sondern laut Graw (2015, S. 511) in komplexen Abhängigkeiten zueinander stehen. Daher ist die Analyse von SNP-Daten auf ihre Abhängigkeiten und Wechselwirkungen von einem besonderen Interesse. Ein Polymorphismus bedeutet, dass die Änderung häufig genug in einer Population vorkommt, um nicht als eine einfache Mutation zu gelten. Da ein SNP eine einzelne Änderung in einem Gen ist, wird als Referenz das entsprechende Gen ohne Veränderungen herangezogen, wie es bei dem Großteil der Population auftritt. Daher kann ein SNP von drei Arten sein:

- „Homozygoter Referenz-Genotyp“: beide Basen, die den SNP beschreiben, sind die häufiger auftretende Variante.
- „Heterozygoter Variant-Genotyp“: eine der beiden Basen, die den SNP beschreiben, ist die häufiger auftretende Variante und eine ist der nicht häufiger auftretende Variante.
- „Homozygoter Variant-Genotyp“: beide Basen, die den SNP beschreiben, sind die nicht häufiger auftretende Variante.

Da es sich bei SNP-Daten um kategoriale Daten handelt, ist es für eine Analyse notwendig, diese drei Ausprägungsmöglichkeiten zu codieren. Dies geschieht entweder auf einer Skala von 0 bis 2 oder auf einer Skala von 1 bis 3. Gängige Konventionen ist es, den homozygoten Referenz-Genotypen mit 0 bzw. 1 zu codieren, den heterozygoten Variant-Genotyp mit 1 bzw. 2 und den homozygoten Variant-Genotypen mit 2 bzw. 3. Bei der logischen Regression (vgl. **Kapitel 3**) geht es im Kontext von SNP-Daten darum, basierend auf Booleschen Kombinationen der Einflussvariablen Vorhersagen für die abhängige Variable zu treffen. Die SNPs können als Binärvariablen (vgl. **Kapitel 3.1**) der Form

$$S_1 : \text{SNP } S \text{ ist nicht der homozygote Referenz-Genotyp}$$

oder

$$S_2 : \text{SNP } S \text{ ist der homozygote Referenz-Genotyp}$$

interpretiert werden. Diese Variablen können mit Logikoperatoren wie dem Kompliment $^{\mathrm{C}}$ negiert werden (z.B. S_2^{C}: SNP S ist NICHT der homozygote Referenz-Genotyp) und durch die Operatoren $\wedge/\vee$ zu Logikausdrücken verknüpft werden (siehe **Kapitel 3.1**).

Für die Analyse mit der logischen Regression (vgl. **Kapitel 3**) ist es nötig, diese drei Ausprägungsmöglichkeiten in Binärdaten zu codieren. Jeder SNP S_i einer einzelnen Beobachtung wird dazu aufgespalten in zwei Binärvariablen S_{i_1} und S_{i_2}, $i = 1, ..., d$, die

den Typus der Base beschreiben:

S_{i_1} : Mindestens einer der Basen, die den SNP S_i beschreiben, ist die weniger häufig auftretende Variante.

S_{i_2} : Beide Basen, die den SNP S_i beschreiben, sind die weniger häufig auftretende Variante.

Diese Dummy-Variablen werden anstelle des entsprechenden SNP verwendet. Die Anzahl d der Einflussvariablen verdoppelt sich somit und es liegen $2d$ Einflussvariablen vor. Es lassen sich alle drei Ausprägungsmöglichkeiten durch diese Binärvariablen darstellen, indem diese den Status Wahr oder Falsch annehmen und entsprechend den Typus der Base angeben. Dabei beschreibt S_{i_1} die dominante Variation und S_{i_2} einen rezessiven Effekt.

Im nächsten Unterkapitel werden die für diese Arbeit verwendeten SNP-Datensätze erläutert.

4.1 SNP-Datensätze

Für diese Arbeit liegen mehrere SNP-Datensätze vor. Der größte und umfangreichste Datensatz ist der sogenannte HapMap-Datensatz. Das Ziel des HapMap-Projektes war die Erstellung einer Haplotyp-Karte des menschlichen Genoms. Ein Haplotyp bezeichnet eine Gruppe von Allelen benachbarter Gene, die durch eine Person getragen und vererbt werden. Dazu wurden in der ersten Phase 270 Proben von vier Populationen genommen, die aus unterschiedlichen geographischen Regionen stammten. Das Projekt wurde noch ausgeweitet und in 1184 DNA-Proben von elf Populationen 1,6 Mio. SNPs untersucht (vgl. Graw 2015, S. 512). Mit der Hinzunahme anderer Projekte sind über 10 Mio. vererbte Veränderungen der DNA bekannt, wobei es sich bei den meisten dieser Veränderungen um SNPs handelt. Aufbauend auf diesen Projekten entstehen dann genomweite Assoziationsstudien (GWAS). Bei GWAS geht es darum SNPs mit gewissen Krankheiten in Verbindung zu bringen (siehe Graw 2015, S. 512).

Für diese Arbeit wird ein Teil der HapMap-Daten verwendet. Der Datensatz stammt aus dem R-Paket „SNPassoc“ (González et al. 2014). Das Paket ist Teil des Bioconductor-Projekts (Huber et al. 2015). Das Projekt bietet Software-Pakete zur Analyse von genetischem Datenmaterial an. Enthalten sind in den HapMap-Daten $d = 9307$ SNPs (Variablen) von $n = 120$ Personen. Die Personen gehören einer von zwei ethnischen Gruppen an: insgesamt $n_1 = 60$ Beobachtungen der europäischen Population (mit CEU gekennzeichnet) und $n_2 = 60$ Beobachtungen zu den westafrikanischen Yoruba (mit YRI gekennzeichnet). Die SNP-Daten liegen nicht in codierter Form vor, sondern enthalten die Information der entsprechenden Basen des SNPs. Mit der Funktion „additive“ aus dem R-Paket „SNPassoc“ können die Daten in 0/1/2 Daten codiert werden. Von den 9307 SNPs sind bei 1657 alle Ausprägungen des entsprechenden SNPs gleich. Diese werden daraufhin aus dem Datensatz entfernt, da diese SNPs keine Informationen über Unterschiede zwischen den Populationen enthalten. Somit bleiben noch $d = 7648$ SNPs über. Etwa 4% der

verbleibenden Daten sind fehlende Werte. Um diese zu ersetzen, werden aus den Randverteilungen des entsprechenden SNP zufällig Ausprägungen bestimmt, um diese Werte zu ersetzen. Dazu werden die relativen Häufigkeiten der Ausprägungen des entsprechenden SNP als Wahrscheinlichkeiten für das Ziehen verwendet. Dieses Vorgehen basiert auf dem Vorgehen von Schwender und Ickstadt bei der Bereinigung des Genica-Datensatzes (vgl. Schwender und Ickstadt 2008, S. 194).

Ein weiterer Datensatz ist der Datensatz „data.logicfs“ aus dem R-Paket „logicFS“ (Schwender 2013). Dieses Paket ist ebenfalls Teil des Bioconductor Projekts. Bei dem Datensatz handelt es sich um einen Datensatz von $n = 400$ Personen in den Zeilen der Datenmatrix und $d = 15$ SNPs S in den Spalten der Datenmatrix. Die Datenmatrix enthält die Ausprägungen der SNPs, codiert von 1 bis 3 für den entsprechenden Typus des SNP. Dazu liegt ein Vektor vor, der den Fall-/Kontrollstatus der i-ten Beobachtung angibt, $i = 1, ..., 400$. Der Datensatz ist ein simulierter Datensatz und dient als Inspiration für die hier durchgeführte Simulationsstudie. Enthalten sind $n_1 = n_2 = 200$ Fälle bzw. Kontrollen, wobei $n = n_1 + n_2$. Eine Person ist dabei ein Fall, wenn mindestens eine der folgenden Bedingungen erfüllt ist:

$$
\begin{aligned}
&L_{FS_1} : S_1 = 3 \\
&L_{FS_2} : S_2 = 1 \wedge S_4 = 3 \\
&L_{FS_3} : S_3 = 3 \wedge S_5 = 3 \wedge S_6 = 1.
\end{aligned}
$$

Die Wechselwirkungen befinden sich auf den jeweils ersten sechs Variablen des Datensatzes. Für diese Arbeit erhält der Datensatz die Bezeichnung logicFS-Datensatz.

Im nachfolgenden Unterkapitel wird die durchgeführte Simulationsstudie beschrieben.

4.2 Simulation der SNP-Daten

Für diese Arbeit sollen in erster Linie Daten simuliert werden. Ziel der Simulation ist es, unter kontrollierten Bedingungen und mit bekannten Wechselwirkungen in den Daten die Methodiken zu überprüfen und zu evaluieren. Der Vorteil von simulierten Daten ist ihre Reproduzierbarkeit und die a priori bekannten Verhältnisse. Ein Nachteil von Simulationen ist es, dass diese a priori getroffenen Verhältnisse möglicherweise nicht der Realität entsprechen und sich die Ergebnisse auf echten Daten anders verhalten. In der Simulation werden unterschiedliche Datensätze erzeugt, mit unterschiedlichen Grundvoraussetzungen. Als Inspiration für die Simulation der Datensätze dient der logicFS-Datensatz. Beide möglichen Fälle bei SNP-Daten $n \geq d$ und $n < d$ sollen berücksichtigt werden. Zusätzlich werden die beiden Parameter n und d variiert, um ihren Einfluss auf die Methode zu untersuchen. Durchgeführt wird die Simulation mit der R-Funktion „simulateSNPs“ aus dem Paket „scrime“ (Schwender und Fritsch 2013). Simuliert werden die SNPs von n Personen, die einer Kategorie zugewiesen werden. Im Fall einer klinischen Studie könnten dies etwa die zwei Kategorien „Krankheit liegt vor“ und „Krankheit liegt nicht vor“ sein.

Im Falle der HapMap-Daten (siehe **Kapitel 4.1**) entsprechen die Kategorien der ethnischen Herkunft der jeweiligen Person. In jedem Fall wird die Kategorie der Person mit 1 für einen Erfolg (die Krankheit liegt vor, die Person entstammt der ethnischen Gruppe) und 0 für einen Misserfolg (die Krankheit liegt nicht vor, die Person entstammt nicht der ethnischen Gruppe) codiert. Dies entspricht der abhängigen Variable des Modells. Der Einfachheit halber wird für die Simulationen von einer Fall-Kontroll-Studie ausgegangen und die Personen der Gruppe 1 als krank eingestuft, wobei die SNPs (mögliche) Ursachen dieser Krankheit sind. Entsprechend werden die Personen der Gruppe 0 als gesund eingestuft.

Es liegen n_1 kranke Personen vor, n_2 gesunde Personen und es ist $n = n_1 + n_2$. Jede Person besitzt d SNPs S. Somit liegen pro Datensatz $n \cdot d$ Datenpunkte vor. Die Daten werden so simuliert, dass eine Person in der Regel als krank eingestuft wird, wenn bei ihr eine gewisse Kombination von SNPs gemeinsam auftritt. Anders kann es von Interesse sein, dass nicht alle Personen die krank sind auch eine solche Wechselwirkung aufweisen. So könnten zum Beispiel andere Faktoren zum Bilden der Krankheit beitragen, als die durch die SNPs gegebenen genetischen Einflussfaktoren. Dies könnten etwa Umwelteinflüsse, wie eine Feinstaubbelastung oder ungesunde Lebensweisen wie das Rauchen sein. Mit einem gewissen Anteil $\rho \in (0,5; 1]$ weisen die kranken Personen die durch die SNPs ausgelösten genetischen Faktoren auf, so dass insgesamt $\lceil n_1 \cdot \rho \rceil$ der kranken Personen mindestens eine der Wechselwirkungen besitzen.

In der Simulationen wird die zweite gängige Kodierung der SNP-Daten verwendet, auf der Skala von 0 bis 2, für den entsprechenden Typus des SNPs. Einer Person wird der Status krank zugewiesen, wenn (mindestens) eine der Bedingungen

$$L_1 : S_1 = 2 \wedge S_2 \neq 0 \wedge S_3 = 1 \tag{3}$$

$$L_2 : S_4 \neq 0 \wedge S_5 = 2 \tag{4}$$

$$L_3 : S_6 \neq 1 \tag{5}$$

erfüllt ist. Da es sich bei SNP-Daten um kategoriale Daten handelt und diese somit keine feste Anordnung besitzen, befinden sich die interessierenden Wechselwirkungen auf den ersten sechs Variablen. Für unterschiedliche Kombinationen von n und d werden Datensätze erzeugt. Die Anzahl der Personen mit den Wechselwirkungen sind zusätzlich unterschiedlich stark besetzt. Von den n_1 Fällen besitzen $n_{1_1} = \lfloor n_1/2 \rfloor$ Fälle die Wechselwirkung L_1, $n_{1_2} = \lfloor n_1/3 \rfloor$ die Wechselwirkung L_2 und $n_{1_3} = \lfloor n_1/6 \rfloor$ die Einzelvariable L_3. Die fehlenden Beobachtungen, damit die Gleichung $n_1 = n_{1_1} + n_{1_2} + n_{1_3}$ erfüllt ist, werden abwechselnd nacheinander den Gruppen n_{1_1} und n_{1_3} zugeteilt, bis alle Fälle einer Gruppe zugeteilt sind. Die Idee für die unterschiedliche Besetzungsstärke in den Untergruppen ist, dass unterschiedliche Wechselwirkungen bei unterschiedlich vielen Personen zu einer Erkrankung führen sollen, da es in der Methodik um eine zufallsbasierte Reduktion der Datensätze geht. Dabei ist es interessant zu untersuchen, ob die geringer besetzten Gruppen durch die Zufallsauswahl ebenfalls repräsentiert werden oder eine eventuelle

Nichtaufnahme zu schlechteren Klassifikationen durch die Modelle führt.

Die angegebenen Wechselwirkungen übertragen sich wie folgt in binäre Schreibweise. Zuerst wird jeder SNP aufgespalten in zwei binäre Dummy-Variablen S_{i_1} und S_{i_2}, $i = 1, .., d$ (vgl. **Kapitel 4**). Für die Wechselwirkung L_1 entstehen somit sechs neue Binärvariablen, die diese beschreiben: S_{1_1}, S_{1_2}, S_{2_1}, S_{2_2}, S_{3_1} und S_{3_2}. Da die angegebenen Modelle und Wechselwirkungen der R-Funktion „logic.bagging“ als Notation X_i, $i = 1, ..., 2d$ für die Binärvariablen verwenden, wird für diese Arbeit ebenfalls diese Schreibweise verwendet. Somit schreiben sich die binären Dummy-Variablen statt mit doppelten Indizes mit nur einem. Die erste Wechselwirkung enthält die Binärvariablen $X_1, ..., X_6$, die zweite die Binärvariablen $X_7, ..., X_{10}$ und L_3 die Binärvariablen X_{11}, X_{12}. Mit der gängigen Kodierung $S_{i_1} = 0 \wedge S_{i_2} = 0$ für den homozygoten Referenz-Genotypen, $S_{i_1} = 1 \wedge S_{i_2} = 0$ für den heterozygoten Variant-Genotypen und $S_{i_1} = 1 \wedge S_{i_2} = 1$ für den homozygoten Variant-Genotypen, $i = 1, ..., d$, ergeben sich mit

$$L_{1_{DNF}} = (X_1 \wedge X_2 \wedge X_3 \wedge X_5 \wedge X_6^{\mathrm{C}}) \tag{6}$$

$$L_{2_{DNF}} = (X_7 \wedge X_9 \wedge X_{10}) \tag{7}$$

$$L_{3_{DNF}} = (X_{11}^{\mathrm{C}} \wedge X_{12}^{\mathrm{C}}) \vee (X_{11} \wedge X_{12}) \tag{8}$$

die DNF der gewählten Wechselwirkungen. Bei der DNF der Wechselwirkung L_1 ist die Binärvariable X_4 nicht enthalten, da diese sowohl den Wert 1 als auch 0 erhalten kann und somit für die Wechselwirkung überflüssig ist. Selbes gilt für die Binärvariable X_8. Da die Wechselwirkung L_3 eine Ungleichung für die Ausprägung des SNPs mit 1 ist, wird durch die DNF der Fall ausgeschlossen, dass X_{11} und X_{12}^{C} bzw. X_{11}^{C} und X_{12} gemeinsam auftreten.

Erzeugt werden in jeder Simulation für jede Situation $B_{sim} = 100$ Datensätze. Diese Anzahl wurde gewählt, um verschiedene Datensätze vorliegen zu haben, die Rechenzeit jedoch noch in Grenzen zu halten. Ein jeder Datensatz enthält eine Matrix $X \in \{0, 1, 2\}^{n \times d}$, in der jeweils n Ausprägungen von d SNPs enthalten sind und einen Vektor $y \in \{0, 1\}^n$, in dem der Klassifikationsstatus der i-ten Person angegeben ist. Insgesamt werden vier verschiedene Simulationen durchgeführt.

Simulation 1: Die erste Simulation besitzt $n = 400$ Beobachtungen mit jeweils $n_1 = n_2 = 200$ Beobachtungen pro Kategorie. Aus der Gruppe $n_1 = n_{1_1} + n_{1_2} + n_{1_3}$ der Fälle besitzen 100 der gesamten Beobachtungen ($n_{1_1} = 100$) die Wechselwirkung L_1, 66 der gesamten Beobachtungen die Wechselwirkung L_2 und 34 die Einzelvariable L_3 als genetische Ursache für ihre Erkrankung. Die angegebene Anzahl der Personen in den jeweiligen Untergruppen besitzen fest die Wechselwirkungen. Die Werte der restlichen SNPs, sowie die SNPs der Kontrollgruppe werden zufällig gleichverteilt mit einer Auswahlwahrscheinlichkeit von jeweils 1/3 gezogen. Die Anzahl der SNPs d wird zusätzlich variiert mit jeweils $d \in \{10, 20, 30, 40, 50\}$, wobei die Wechselwirkungen beibehalten werden. Dies soll die Situation darstellen, dass im Vorhinein nicht bekannt ist, welche der SNPs einen Einfluss auf die Krankheit haben und nur vermutet wird, welche dies sein

könnten. Entsprechend enthält der Datensatz mit steigendem d immer mehr unwichtige Informationen und es soll darum gehen, wie sich dies auf die Analyse auswirkt. Für Simulation 1 werden somit 500 Datensätze simuliert.

Simulation 2: Die zweite Simulation ist vom Grundaufbau her wie Simulation 1. Dies bedeutet, es gibt dieselbe Anzahl an Personen n mit derselben Anzahl für jede der Untergruppen der Wechselwirkungen und die selbe Anzahl an Variablen d. Der Unterschied liegt darin, dass in jeder Untergruppe der Erkrankten der Anteil der Personen, deren Erkrankung durch die genetischen Einflüsse der SNPs verursacht ist, bei $\rho = 0,8$ liegt. Insgesamt werden somit bei 80% der Erkrankten die Krankheit durch die SNPs ausgelöst und bei 20% durch andere Faktoren. Der Grund dafür ist, dass es interessant ist zu untersuchen, wie die Methodik auf ein solches „Rauschen“ in den Daten reagiert. Rauschen ist dabei nicht wie im klassischen statistischen Sinne gemeint. An dieser Stelle ist damit die Situation gemeint, dass es Beobachtungen gibt, die nicht durch die genetischen Einflüsse erklärt sind. Die Ausprägungen der SNPs dieser speziellen Fälle werden ebenfalls gleichverteilt mit 1/3 Wahrscheinlichkeit gezogen. Für diese Simulation werden erneut 500 Datensätze simuliert.

Simulation 3: Bei der dritten Simulation geht es darum zu untersuchen, wie die Methodik sich auf sehr große Datenumfänge anwenden lässt. Dazu werden jeweils $B_{sim} = 100$ Datensätze erzeugt mit $n = n_1 + n_2 = 5000$ wobei $n_1 = n_2 = 2500$ ist. Erneut liegen $d \in \{10, 20, 30, 40, 50\}$ SNPs pro Person vor und die Erkrankten besitzen fest die angegeben Wechselwirkungen. Das Verhältnis in der Gruppe der Fälle ist dasselbe wie bei Simulation 1 und 2, die absolute Anzahl pro Untergruppe ist entsprechend größer mit $n_{1_1} = 1250$, $n_{1_2} = 833$ und $n_{1_3} = 417$. Ziel bei dieser Simulation wird es sein, nur einen sehr geringen Anteil der Daten zu verwenden um ein logisches Regressionsmodell anzupassen. Für Simulation 3 werden erneut 500 Datensätze simuliert.

Simulation 4: In der vierten Simulation soll der zweite Spezialfall der SNP-Daten abgedeckt werden. Somit ist diese Simulation darauf ausgelegt, dass es mehr SNPs d als Beobachtungen n gibt. Anders als in den vorhergegangenen Simulationen geht es für diese Simulation nicht darum Personen für die logische Regression zu selektieren, sondern Variablen aus dem Datensatz. Insgesamt werden vier verschiedene Situationen mit $d \in \{250, 300, 400, 500\}$ und jeweils $n = n_1 + n_2 = 200$ Personen abgedeckt, wobei $n_1 = n_2 = 100$ ist. Auch diese Datensätze besitzen die beiden Wechselwirkungen und die Einzelvariable in demselben Verhältnissen wie in Simulation 1, mit $n_{1_1} = 50$, $n_{1_2} = 33$ und $n_{1_3} = 17$. Für diese Situation werden entsprechend 400 Datensätze simuliert.

Bei etwa 20% bis 40% der Datensätze ist der Fall eingetreten, dass eine der Wechselwirkungen bzw. Teile der Wechselwirkung schon durch die anderen Wechselwirkungen erklärt wird und daher aus dem Datensatz entfernt wird. Diese Datensätze werden trotzdem behalten, da es bei genetischen Daten durchaus passieren kann, dass es zu Änderungen in den Wechselwirkungen kommt bzw. deren Wichtigkeit und Umfang variiert.

Tabelle 1: Überblick der durchgeführten Simulationen.

Simulation	n	n_1	n_{1_1}	n_{1_2}	n_{1_3}	ρ	d
1	400	200	100	66	34	1	10/20/30/40/50
2	400	200	80	53	27	0,8	10/20/30/40/50
3	5000	2500	1250	833	417	1	10/20/30/40/50
4	200	100	50	33	17	1	250/300/400/500
Wechselwirkung			L_1	L_2	L_3		

In **Tabelle 1** ist eine Übersicht der durchgeführten Simulationen dargestellt. Die erste Spalte enthält die für diese Arbeit gewählte Nummerierung der entsprechenden Simulation. Die zweite Spalte enthält die Anzahl der pro Datensatz generierten Beobachtungen und die dritte Spalte den Anteil der Fälle. In den nächsten drei Spalten ist eine Übersicht über die Anzahl der pro Wechselwirkung vertretenen Beobachtungen gegeben und zusätzlich, welche der Wechselwirkungen für den Krankheitsstatus verantwortlich ist. Die siebte Spalte gibt an, welcher Anteil der Fälle ihren Krankheitsstatus durch die genetischen Einflüsse der SNPs erklärt bekommen. Die letzte Spalte enthält eine Übersicht über die Anzahl der SNPs, die in den Datensätzen vorkommen.

5 Reduktion hochdimensionaler Datensätze für die logische Regression

Nun soll es konkret um die Reduktion hochdimensionaler Datensätze für die logische Regression gehen. Es gibt zwei große Teilbereiche. Zum einen liegt die Situation vor, dass es mehr Beobachtungen n als Variablen d gibt. Dies ist der klassische Fall und findet in vielen anderen statistischen Regressionsproblemen seine Anwendung. Bei SNP-Daten ist es jedoch sehr häufig möglich, dass deutlich mehr Variablen als Beobachtungen vorliegen. Dies ist etwa bei den HapMap-Daten der Fall (vgl. **Kapitel 4.1**). Daher ist es von Interesse beide Fälle getrennt zu betrachten und zu untersuchen wie sich die gewählte Methodik auf diese beiden Fälle auswirkt.

Im nächsten Unterkapitel wird allgemein das Vorgehen zum Reduzieren der Datensätze vorgestellt. In diesem Zusammenhang werden die Leverage Scores vorgestellt und erläutert wie diese dazu verwendet werden, die Datensätze zu reduzieren. Danach geht es getrennt nach den beiden Fällen um die Reduktion der Datensätze anhand der in **Kapitel 4** vorgestellten Daten.

5.1 Das Vorgehen zum Reduzieren der Datensätze

Gegenstand dieser Arbeit ist es, Datensätze für die logische Regression zu reduzieren. Um die Daten gezielt reduzieren zu können, bedarf es Informationen die nach Möglichkeit aus den Daten selbst stammen. Als Methodik zur Reduktion der Datensätze werden in dieser Arbeit die sogenannten Leverage Scores (LS) verwendet. Die LS stehen im Kontext der Regressionsanalyse. Die Beschreibung basiert auf Hoaglin und Welsch (1978).

Ein lineares Regressionsmodell ist gegeben (formuliert nach Hoaglin und Welsch 1978, S. 17) durch

$$Y = X\beta + u,$$

wobei $Y \in \mathbb{R}^n$ die abhängige Variable ist, $X \in \mathbb{R}^{n\times d}$ die Matrix der Einflussvariablen $X_1, ..., X_d$ mit $X_i \in \mathbb{R}^n$, $i = 1, .., d$, $\beta \in \mathbb{R}^d$ der Parametervektor des Modells und $u \in \mathbb{R}^n$ der zufällige Störterm. Die Realisationen der abhängigen Variable sind $y = (y_1, ..., y_n)'$ und die der Einflussvariablen $x_i = (x_{i1}, ..., x_{id})'$, $i = 1, ..., n$. Bei der linearen Regression wird zur Anpassung des Modells die Kleinste-Quadrate-Methode verwendet. Dazu werden die Annahmen getroffen, dass $rang(X) = d$ ist, $E(u) = 0$ und $Var(u) = \sigma^2 \mathrm{I}_n$.

Die prognostizierten Werte des Modells sind gegeben durch $\hat{y} = X\hat{\beta}$, wobei $\hat{\beta} = (X'X)^{-1}X'y$ ist. Damit ergibt sich

$$\hat{y} = X(X'X)^{-1}X'y = Hy.$$

Die sogenannte Hat-Matrix $H \in \mathbb{R}^{n\times n}$ ist somit die Matrix

$$H = X(X'X)^{-1}X'.$$

Die Hat-Matrix H bildet den Vektor y ab auf $\hat{y}$. Es kann untersucht werden, ob ein Datenpunkt y_i eine besondere Hebelwirkung (leverage) auf seinen prognostizierten Wert $\hat{y}_i$, $i = 1, ..., n$ hat. Dadurch lassen sich Ausreißer in den Realisationen der Einflussvariablen entdecken. Diese Informationen befinden sich in der Hat-Matrix. Geometrisch können y und die Spalten von X als Punkte im $\mathbb{R}^n$ angesehen werden. Die Punkte $X\beta$ bilden einen Teilraum im $\mathbb{R}^d$ als Linearkombination des Spaltenraumes von X. Die prognostizierten Werte im Vektor $\hat{y}$ bilden den Punkt in dem Teilraum, welcher den geringsten Abstand zu y hat. Der Vektor $\hat{y}$ ist also die orthogonale Projektion auf y in den Teilraum. Somit ist die Hat-Matrix ein Projektor. Der Einfluss des Punktes y_i auf $\hat{y}_i$ ist repräsentiert durch das i-te Diagonalelement h_{ii} der Hat-Matrix H. Somit ist der i-te Leverage Score gegeben durch

$$l_i = h_{ii}.$$

Neben den Leverage Scores gibt es die Cross Leverage Scores (CLS) c_{ij}. Diese beschreiben den Einfluss der Beobachtung y_i auf $\hat{y}_j$, $i, j = 1, ..., n$. Die CLS finden sich in den Spalten bzw. Zeilen der Hat-Matrix. Die Hebelwirkung von y_i auf $\hat{y}_j$ ist somit durch

$$c_{ij} = h_{ij}$$

gegeben. Dabei ist h_{ij} der Eintrag der Hat-Matrix in der i-ten Zeile und j-ten Spalte, $i, j = 1, ..., n$. Wenn $i = j$ ist, dann ist $c_{ij} = l_i$ für $i = 1, ..., n$.

In dieser Arbeit werden Leverage Scores mit LS und die Cross Leverage Scores mit CLS abgekürzt. Die Abkürzung (C)LS wird an den Stellen verwendet, an denen gleichzeitig auf die Leverage bzw. die Cross Leverage Scores hingewiesen wird.

Für die LS gilt $0 \leq l_i \leq 1$ und es ist $\sum_{i=1}^{n} l_i = d$. Ein Wert von $l_i = 0$ bedeutet, dass $\hat{y}_i$ vom Design her auf den Wert 0 fixiert ist und es nicht von y_i bzw. generell von einem beliebigen y_j beeinflusst wird. Falls $l_i = 1$ ist bedeutet dies, dass $\hat{y}_i = y_i$ ist und das Modell perfekt zu diesem Punkt passt. Diese beiden Randfälle bedeuten zusätzlich, dass $c_{ij} = 0$ ist für $j = 1, ..., n$, $j \neq i$.

Numerisch ist die Bestimmung der (C)LS durch die oben formulierte Hat-Matrix nicht effizient. Nach Hoaglin und Welsch (1987, S. 21) gibt es numerisch bessere Verfahren. Bei Drineas et al. (2012, S. 2) ist angemerkt, dass es genügt, um die LS der Matrix X zu bestimmen, eine Orthonormalbasis des Spaltenraumes zu bestimmen und dann die euklidische Norm der Zeilen dieser Basis zu verwenden. Bestimmen lässt sich die Orthonormalbasis per QR-Zerlegung. Für eine Matrix $X \in \mathbb{R}^{n \times d}$, $n \geq d$ ist

$$X = QR$$

die QR-Zerlegung, wobei $Q \in \mathbb{R}^{n \times n}$ eine orthogonale Transformation ist und R eine obere Dreiecksmatrix. Die Hat-Matrix kann bei vollem Rang von X mit

$$H = QQ'$$

ausgedrückt werden. Bestimmt werden die (C)LS in dieser Arbeit per QR-Zerlegung mit der R-Funktion „qr“ aus dem Basis-Paket.

Es sollen die (C)LS der Daten dazu verwendet werden, Beobachtungen bzw. Variablen aus den Daten zu selektieren. Dies geschieht durch eine zufallsbasierte Methode, bei der eine Beobachtung bzw. Variable abhängig von ihrem entsprechenden (C)LS Wert in die Stichprobe aufgenommen wird. Dieser Ansatz entstammt der Informatik und wird auch in dem Teilprojekt C4 (in dessen Rahmen diese Arbeit entsteht) mit angewandt. Die Idee ist es, die Datenmatrix und die abhängige Variable gemeinsam zu betrachten. Im Kontext des linearen Regressionsmodells geht es nach Geppert et al. (2015) darum, auf dem gesamten Datensatz $[X, Y] \in \mathbb{R}^{n\times(d+1)}$ eine lineare Abbildung $\Pi \in \mathbb{R}^{r\times n}$ anzuwenden ($r < n$), so dass der gesamte Datensatz mit einer sogenannten Skizze $[\Pi X, \Pi Y] \in \mathbb{R}^{r\times(d+1)}$ ersetzt wird, die deutlich kleiner ist als der ursprüngliche Datensatz. Dies führt dazu, dass die Likehood-Funktion deutlich schneller angepasst werden kann. Nach Geppert et al. (2015) ist es möglich, die Likehood-Funktion bis auf einen Fehler ε sehr gut anzunähern und gleichzeitig die Dimension der Beobachtungen n auf die Dimension r zu reduzieren. Die Berechnung der Likehood-Funktion hängt dann nicht mehr von n ab, was dazu führt, dass die Anpassung der Modelle deutlich beschleunigt wird.

Im Kontext dieser Arbeit geht es um die logische Regression. Die Skizze Π besteht darin, die Realisationen der Zielvariable $Y \in \{0, 1\}^n$ mit einer Link-Funktion $g(\cdot)$ zu transformieren und zur Reduktion der Dimension der Beobachtungen bzw. Variablen diese proportional zu ihren (C)LS in die Stichprobe aufzunehmen. Jede Realisation y_i, $i = 1, ..., n$ des Vektors der abhängigen Variable Y wird vorher transformiert

$$\tilde{y}_i = g(y_i + (-1)^{y_i} \cdot \varepsilon),$$

dabei ist g eine geeignet gewählte Link-Funktion und $\varepsilon > 0$. Die Wahl von ε spielt keine große Rolle, solange es nur klein genug ist. Für diese Arbeit wird $\varepsilon = 10^{-4}$ gewählt und als Link-Funktion wird die logit-Funktion

$$g(y) = log\left(\frac{y}{1-y}\right)$$

gewählt. Da die y_i nur die Werte 0 und 1 annehmen können, ist es notwendig, die Transformation vorzunehmen um den Wertebereich der Realisationen in die reellen Zahlen $\mathbb{R}$ zu bringen und die Werte linearer als Binärdaten zu machen. Zusätzlich gilt für die logit-Funktion $\lim\limits_{y\to 1} g(y) = \infty$ und $\lim\limits_{y\to 0} g(y) = -\infty$. Daher wird die Transformation in einer ε-Umgebung um die Werte der Realisierungen der abhängigen Variable vorgenommen. Zum Bestimmen der (C)LS wird die Datenmatrix X mit dem transformierten Vektor der Realisationen der abhängigen Variable $\tilde{y} \in \mathbb{R}^n$ vereinigt und bildet die Matrix

$$\widetilde{X} = (X|\tilde{y}) \in \mathbb{R}^{n\times(d+1)}.$$

Anschließend wird die QR-Zerlegung der Matrix $\widetilde{X}$ bestimmt, um eine Orthonormalbasis Q des Spaltenraumes von $\widetilde{X}$ zu erhalten. Abhängig von $m = \max\{n, d\}$ geschieht dies mit $\widetilde{X}$ oder mit $\widetilde{X}^T$. Die Zahl m beschreibt die Situation $n \geq d$ oder $n < d$. Dementsprechend

ist die reduzierte Dimension r ebenfalls von der Situation abhängig. Die (C)LS finden sich in der Hat-Matrix

$$H = QQ'.$$

Die Dimension der Hat-Matrix hängt auch von m ab. Im Falle $m = n$ ist $H \in \mathbb{R}^{m \times m}$ und im Falle $m = d$ ist $H \in \mathbb{R}^{(m+1)\times(m+1)}$. Die LS bilden die Hauptdiagonale der Hat-Matrix, die CLS befinden sich in den Spalten bzw. Zeilen von H.

Da SNPs kategoriale Daten sind und die beeinflussenden Wechselwirkungen a priori festgelegt auf den ersten sechs Variablen liegen, stellt sich die Frage, ob die (C)LS abhängig von der Reihenfolge der Variablen sind. Dieselbe Frage stellt sich für die LS der Beobachtungen, da es sich bei den ersten n_1 Beobachtungen um Fälle und den zweiten n_2 Beobachtungen um Kontrollen handelt.

Die Bestimmung der QR-Zerlegung geschieht über die sogenannte Householder-Matrix bzw. -Transformation. Nach Golub und Van Loan (1996, S. 209) ist eine Householder-Matrix eine $(n \times n)$-Matrix der Form

$$\mathbb{H} = \mathrm{I}_n - \frac{2}{v^T v} v v^T,$$

wobei $v \in \mathbb{R}^n \setminus \{0\}$ ein sogenannter Householder-Vektor (ungleich dem Nullvektor) ist und I_n die $(n \times n)$-Einheitsmatrix. Die Matrix $\mathbb{H}$ ist symmetrisch und orthogonal, was bedeutet, dass $\mathbb{H}^T = \mathbb{H}$ und $\mathbb{H}^{-1} = \mathbb{H}^T$ ist. Die QR-Zerlegung einer Matrix $X \in \mathbb{R}^{n \times d}$ mit $X = QR$ geschieht durch das wiederholte Ausführen von Housholder-Transformationen, so dass $Q = \mathbb{H}_1 \cdot \mathbb{H}_2 \cdot ... \cdot \mathbb{H}_n$ das Produkt mehrerer Househoulder-Matrizen ist (siehe Golub und Van Loan 1996, S. 224). In der Householder-Matrix $\mathbb{H}_i$ ist der Vektor v so zu bestimmen, dass gilt $\mathbb{H}_n \cdot ... \cdot \mathbb{H}_1 \cdot X = R$ wobei R eine obere Dreiecksmatrix ist. Es kann argumentiert werden, dass die Householder-Transformation und damit die QR-Zerlegung nicht von der Reihenfolge der Zeilen und Spalten abhängt, da eine Permutation sich durch jeden Schritt der Bestimmung hindurchzieht. So ändert eine Permutation in dem Vektor v nicht dessen Länge sondern lediglich die Reihenfolge der Einträge. Werden bei der Bestimmung der Hat-Matrix alle Spalten und Zeilen gleichmäßig permutiert und vor allem der Vektor der abhängigen Variable im Einklang mit den Zeilen der Datenmatrix X, ergeben sich dieselben Werte für die (C)LS nur in anderer Reihenfolge. Es kann somit davon ausgegangen werden, dass die Bestimmung der (C)LS unabhängig von der Anordnung der Datenmatrix und den Realisationen der abhängigen Variable ist.

Wie bereits angemerkt sollen die bestimmten (C)LS unter anderem als Wahrscheinlichkeiten für das Ziehen der Beobachtungen bzw. der Variablen verwendet werden, um logische Regressionsmodelle anzupassen. Dafür sollen die (C)LS auf unterschiedliche Art und Weise gewichtet werden, um Wahrscheinlichkeiten zu bilden, proportional zu denen die Beobachtungen oder Variablen in die Stichprobe aufgenommen werden. Nachfolgend sind elf Gewichtsfunktionen formuliert mit denen die (C)LS gewichtet werden, um die

entsprechenden Wahrscheinlichkeiten zu bilden. Es sind

$$f_1(\tilde{l}_i) = \tilde{l}_i^2 \tag{9}$$

$$f_2(l_i) = l_i \tag{10}$$

$$f_3(l_i) = exp(-l_i) \tag{11}$$

$$f_4(\tilde{l}_i) = \begin{cases} (\tilde{l}_i + \widetilde{IQ})^2 & \text{, falls } \tilde{l}_i \geq \widetilde{Q}_{0,75} \\ (\tilde{l}_i - \widetilde{IQ})^2 & \text{, falls } \tilde{l}_i \leq \widetilde{Q}_{0,25} \\ \tilde{l}_i^2 & \text{, sonst} \end{cases} \tag{12}$$

$$f_5(\tilde{l}_i) = \begin{cases} \tilde{l}_i^2 & \text{, falls } \tilde{l}_i \geq \widetilde{Q}_{0,875} \text{ oder } \tilde{l}_i \leq \widetilde{Q}_{0,125} \\ 0 & \text{, sonst} \end{cases} \tag{13}$$

$$f_6(\tilde{l}_i) = \tilde{l}_i^{-2} \tag{14}$$

$$f_7(\tilde{l}_i) = \begin{cases} (\tilde{l}_i + \widetilde{IQ})^2 & \text{, falls } \tilde{l}_i \geq \widetilde{Q}_{0,8} \\ (\tilde{l}_i + \widetilde{IQ})^2 & \text{, falls } \tilde{l}_i \geq \widetilde{Q}_{0,4} \text{ oder } \tilde{l}_i \leq \widetilde{Q}_{0,6} \\ (\tilde{l}_i - \widetilde{IQ})^2 & \text{, falls } \tilde{l}_i \leq \widetilde{Q}_{0,2} \\ \tilde{l}_i & \text{, sonst} \end{cases} \tag{15}$$

$$f_8(\tilde{l}_i) = \begin{cases} \tilde{l}_i^2 & \text{, falls } \tilde{l}_i \leq \widetilde{Q}_{0,25} \\ 0 & \text{, sonst} \end{cases} \tag{16}$$

$$f_9(\tilde{l}_i) = \begin{cases} \tilde{l}_i^2 & \text{, falls } \tilde{l}_i \geq \widetilde{Q}_{0,75} \\ 0 & \text{, sonst} \end{cases} \tag{17}$$

$$f_{10}(\tilde{l}_i) = \phi(\tilde{l}_i) \tag{18}$$

$$f_{11}(\tilde{l}_i) = 1 - |\tilde{l}_i| \tag{19}$$

die gewählten Gewichtsfunktionen. Für gegebene (C)LS l_i, $i = 1, ..., m$, bezeichnet $\tilde{l}_i = l_i - \bar{l}$ die mittelwertbereinigten (C)LS. Die gewichteten (C)LS werden mit $\dot{l}_i$ bzw. $\dot{\tilde{l}}_i$, $i = 1, ..., m$, bezeichnet. Weiter bezeichnet $\widetilde{IQ}$ den Interquartilsabstand von $\tilde{l}_1, ..., \tilde{l}_m$, $\widetilde{Q}_p$ das entsprechende p-Quantil von $\tilde{l}_1, ..., \tilde{l}_m$ und ϕ die Dichtefunktion der Standardnormalverteilung.

Da die gewichteten (C)LS als Wahrscheinlichkeiten verwendet werden um entweder Beobachtungen oder Variablen auszuwählen, werden die nach der Gewichtung resultierenden Werte $\dot{\tilde{l}}_1, ..., \dot{\tilde{l}}_m$ normiert durch $\dot{\tilde{l}}_i / \sum_{i=1}^m \dot{\tilde{l}}_i$, $i = 1, ..., m$. Somit gilt $\sum_{i=1}^m \dot{\tilde{l}}_i = 1$.

Der Grund für die Mittelwertbereinigung ist, die (C)LS, die nahe im Zentrum der Werte liegen, näher an 0 heranzubringen. Zusätzlich sind alle Gewichtsfunktionen so konstruiert worden, dass die resultierenden Werte nicht negativ werden. Dies ist nötig, da die Werte als Wahrscheinlichkeiten verwendet werden. Für die CLS ist es möglich, dass die gewichteten Werte negativ werden können. In diesem Fall werden die Werte $\dot{\tilde{l}}_1, ..., \dot{\tilde{l}}_m$ ersetzt durch $|\dot{\tilde{l}}_1|, ..., |\dot{\tilde{l}}_m|$.

Die Gewichtsfunktionen sind so gewählt worden, um drei unterschiedliche Bereiche

der (C)LS stärker hervorzuheben. Diese drei Bereiche sind die niedrigen (C)LS, die hohen (C)LS und die mittleren (C)LS. Die niedrigen (C)LS sind die Werte, die eine geringe Hebelwirkung (leverage) besitzen und die hohen (C)LS entsprechend die Werte, die eine große Hebelwirkung besitzen. Untersucht werden soll, ob sich eventuell eine hohe Hebelwirkung stärker auf die Selektion auswirkt als eine niedrige oder umgekehrt, bzw. ob eine mittlere Hebelwirkung besser geeignet ist oder eine Kombination der Bereiche. Jede gewählte Gewichtsfunktion korrespondiert mit einem dieser drei Bereiche bzw. einer Kombination dieser. Die Ausnahme bildet die Gewichtsfunktion f_2. Bei dieser Gewichtsfunktion geht es darum, die (C)LS nicht zu gewichten, sondern zu untersuchen ob die (C)LS auch ohne Gewichtung dazu geeignet sind Beobachtungen oder Variablen zu selektieren. Da die Gewichtsfunktionen über einen laufenden Prozess gewählt wurden, sind ähnlich gewichtete Bereiche teilweise nicht kontinuierlich nummeriert (etwa bei den Gewichtsfunktionen f_6 und f_{10}). Um jedoch Verwirrungen bei der bereits begonnenen Auswertung der Daten zu vermeiden, wird die Nummerierung beibehalten.

Die Gewichtsfunktionen f_1, f_4 und f_5 gewichten stärker die Randbereiche und somit gleichzeitig die hohen und niedrigen (C)LS. Die Gewichtsfunktionen werden dabei immer restriktiver. Die Gewichtung von f_1 legt zwar schon eine starke Gewichtung auf die Ränder, diskriminiert aber nicht gegen die mittleren Werte. Bei den Gewichtsfunktionen f_4 und f_5 ist dies der Fall. Die Idee dabei ist es, immer stärker nur die Randbereiche der Werte hervorzuheben und die Mitte zu vernachlässigen. Durch f_5 werden sogar nur die Ränder berücksichtigt. Bei der Gewichtsfunktion f_4 wird eine in den Daten vorhandene Schwelle (in Form des Interquartilsabstandes) eingebaut, um die Randbereiche gegen den mittleren Bereich abzusetzen.

Bei der Gewichtung durch f_6, f_{10} und f_{11} geht es im Gegensatz darum die mittleren (C)LS stärker hervorzuheben. Die Gewichtsfunktionen f_{10} und f_{11} diskriminieren nicht so streng gegen die Randbereiche. Dahingegen diskriminiert f_6 sehr stark gegen die Ränder. Dies führt teilweise dazu, dass nur sehr ausgewählte Beobachtungen bzw. Variablen durch die (C)LS selektiert werden und zwar solche deren (C)LS Werte sehr stark im Zentrum der Werte liegen.

Die Gewichtsfunktionen f_3, f_8 und f_9 heben entsprechend einen der Randbereiche hervor. Bei f_3 und f_8 sind dies die Werte mit niedriger Hebelwirkung und bei f_9 die Werte mit hoher Hebelwirkung. Erneut berücksichtigen f_8 und f_9 keine Werte, die nicht in dem Bereich liegen. Die Idee dahinter ist es, zu untersuchen ob der entsprechende Wertebereich dazu geeignet ist eine gute Selektion durchzuführen. Als Grenzen für den Bereich werden dafür das 0,25- bzw. das 0,75-Quantil gewählt, um Zugriff auf die oberen bzw. unteren 25% der Daten zu haben. Die Gewichtsfunktion f_3 wird als nicht so strenger Kontrast zu f_8 gewählt. Eine nicht so strenge Alternative zu f_9 ist die Gewichtung durch f_2. Da f_2 die Werte unverändert lässt, führt dies automatisch dazu, dass Beobachtungen bzw. Variablen mit hohen Werten eine höhere Wahrscheinlichkeit besitzen ausgewählt zu werden.

Tabelle 2: Übersicht der durch die Gewichtsfunktionen stärker gewichteten Bereiche der Leverage bzw. Cross Leverage Scores.

Gewichteter Bereich	Gewichtsfunktionen	Anteil der verfügbaren Daten
Hohe und niedrige (C)LS	f_1, f_4, f_5	100%/100%/25%
Niedrige (C)LS	f_3, f_8	100%/25%
Hohe (C)LS	f_9	25%
Mittlere (C)LS	f_6, f_{10}, f_{11}	100%/100%/100%
Mittlere, hohe und niedrige (C)LS	f_7	100%
(C)LS als Kriterium	f_2	100%

Als letztes ist f_7 eine Kombination der drei Bereiche, wobei die Randbereiche erneut stärker gewichtet sind. Es werden wieder Informationen in den Daten verwendet, um die Bereiche hervorzuheben. Dazu wird erneut der Interquartilsabstand verwendet.

Eine Übersicht der durch die Gewichtsfunktionen besonders hervorgehobenen Bereiche findet sich in **Tabelle 2**. In der ersten Spalte ist der entsprechende Bereich der (C)LS angegeben, der durch die Gewichtsfunktion berücksichtigt wird. In der zweiten Spalte finden sich die korrespondierenden Gewichtsfunktionen. In der dritten Spalte ist der Anteil der gesamten Daten aufgetragen, der durch die Gewichtsfunktion noch zur Verfügung steht. Dies ist von Interesse, da einige der Gewichtsfunktionen gegen gewisse Bereiche der Daten diskriminieren und es für diese Gewichtsfunktionen nicht möglich ist, dass alle (C)LS verwendet werden und somit ist es nicht möglich, dass alle Beobachtungen in die Stichprobe gelangen können.

In **Abbildung 7** ist eine graphische Repräsentation der Gewichtsfunktionen abgebildet. Diese verdeutlicht, wie stark die einzelnen Gewichtsfunktionen die (C)LS gewichten. Bei der Darstellung ist die Normierung bereits berücksichtigt. Eindeutig lassen sich gewisse Merkmale der Gewichtsfunktionen erkennen. So ist etwa in **Abbildung 7 (f)** zu erkennen, dass f_6 sehr stark gegen die Randbereiche der Werte der (C)LS diskriminiert. Die Diskriminierung der nicht berücksichtigten Bereiche lassen sich auch in den **Abbildung 7 (e)**, **(h)** und **(i)** erkennen.

Um aus den Daten Beobachtungen zu selektieren, werden in erster Linie die LS der Beobachtungen verwendet. Die CLS ergeben in diesem Kontext keinen Sinn, da diese den Einfluss der i-ten Beobachtung auf die j-te Beobachtung beschreiben. Diese Information ist jedoch nicht geeignet für die Wahl der Beobachtungen. Bei der Wahl der Variablen werden sowohl die LS der Variablen, als auch die CLS der Variablen berücksichtigt, dabei ganz speziell die CLS der Variablen mit der abhängigen Variable. Zur Bestimmung der Hat-Matrix wird wie beschrieben der transformierte Vektor der abhängigen Variable mit der Datenmatrix vereinigt und entsprechend mit der transformierten Datenmatrix $\widetilde{X}^T \in \mathbb{R}^{(d+1)\times n}$ die QR-Zerlegung durchgeführt. Die resultierende Hat-Matrix $H \in \mathbb{R}^{(d+1)\times(d+1)}$ enthält dann in der $(d+1)$-ten Zeile bzw. Spalte, den Einfluss (leverage) jeder einzelnen

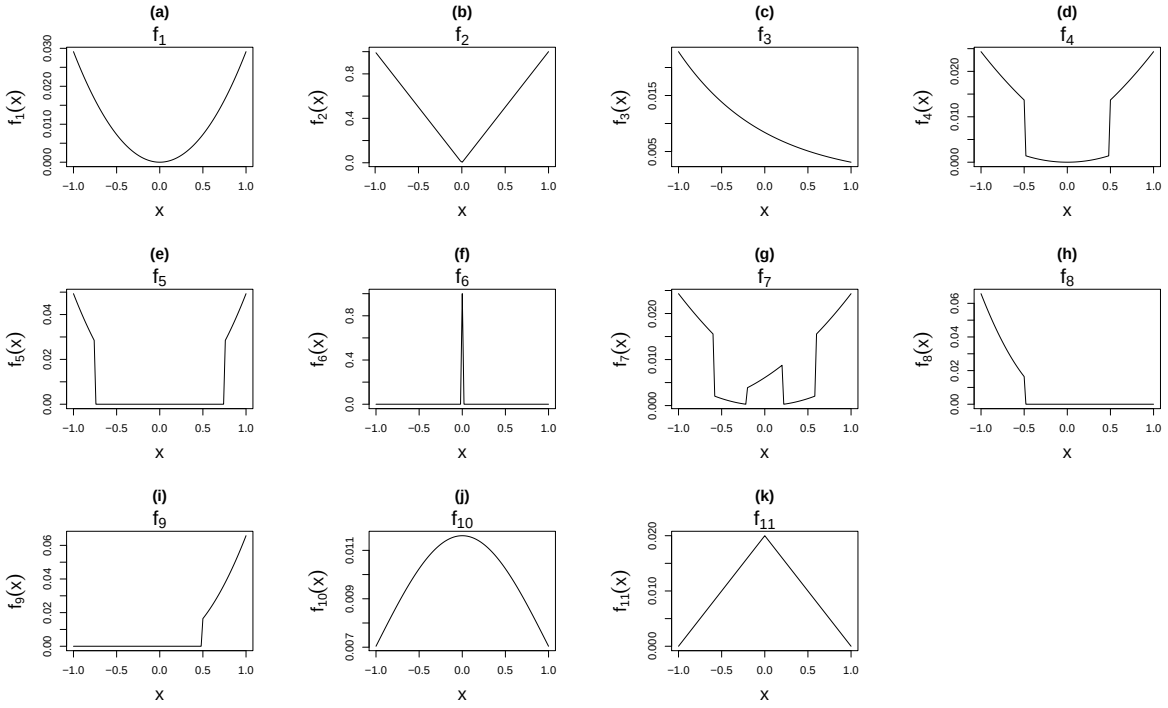

Abbildung 7: Graphische Repräsentation der gewählten Gewichtsfunktionen.

Variable auf die abhängige Variable.

Jede Beobachtung bzw. Variable eines Datensatzes besitzt einen Index i, $i = 1, ..., m$. Für das zufällige Ziehen der Stichprobe vom Umfang r wird der Index der Beobachtung bzw. Variable (abhängig von der Situation) mit einer gewissen Wahrscheinlichkeit p_i gezogen und die entsprechende Beobachtung bzw. Variable dementsprechend der Stichprobe hinzugefügt. Dazu werden für die hinzugefügte Beobachtung i sämtliche Ausprägungen der SNPs mit in die Stichprobe aufgenommen, wohingegen für eine Variable i die Ausprägungen des entsprechenden SNPs jeder Person für diesen SNP in die Stichprobe mitaufgenommen wird. Die nächste Beobachtung bzw. Variable gelangt mit einer Wahrscheinlichkeit $\tilde{p}_j = \frac{p_j}{\sum_{l:l\neq i} p_l}$ in die Stichprobe, $j = 1, ..., m$, $j \neq i$, usw. Dies geschieht bis die vorher festgelegt Anzahl r an Beobachtungen bzw. Variablen gezogen sind. Die Auswahlwahrscheinlichkeiten ergeben sich aus den gewichteten (C)LS: $p_i = \dot{l}_i$ bzw. $p_i = \dot{\tilde{l}}_i$. Als Vergleich dazu wird eine einfache Zufallsauswahl (Simple Random Sample, SRS) verwendet. Bei der SRS ist die Auswahlwahrscheinlichkeit für jede Beobachtung bzw. Variable gleich $p_i = \frac{1}{m}$, $i = 1, ..., m$. Der Vergleich soll zeigen, ob es einen Unterschied macht, die (C)LS bei der Wahl für die Stichprobe zu berücksichtigen oder ob es keinen Unterschied macht bzw. einen negativen Einfluss hat.

Als zweiter Ansatz werden die Beobachtungen bzw. Variablen nicht zufällig in die Stichprobe aufgenommen, sondern fest anhand des entsprechenden (C)LS-Wertes ausgewählt. Dazu werden die (C)LS mit den Gewichtsfunktionen gewichtet und entsprechend die r Beobachtungen bzw. Variablen in die Stichprobe aufgenommen, die nach der Gewichtung den höchsten Wert besitzen. Dies hat den Sinn, dass z.B. durch die Gewichtung

mit f_9 die r Beobachtungen in die Stichprobe gelangen, die den höchsten (C)LS-Wert besitzen.

Sowohl bei der festen Auswahl als auch bei dem Ziehen proportional zu den (C)LS gibt es Einschränkungen. Beim zufälligen Ziehen der Beobachtungen kann es passieren, dass in der Stichprobe keine Fälle oder Kontrollen vorhanden sind. Sollte dies passieren, ist es nicht möglich ein logisches Regressionsmodell anzupassen. Daher wird das Ziehen der Stichprobe solange wiederholt, bis mindestens ein Fall bzw. mindestens eine Kontrolle Teil der Stichprobe sind. Bei der festen Auswahl der Beobachtungen anhand der LS kann der Fall auftreten, dass es in der fest gewählten Stichprobe keine Fälle bzw. keine Kontrollen gibt. Wenn dies auftritt, erhält die Klassifikationsrate der Stichprobe durch die feste Auswahl den Wert 0, da keine Klassifikation durchgeführt werden kann. Beim festen Auswählen der Variablen kann es passieren, dass es in den SNPs weniger als drei Ausprägungsmöglichkeiten gibt. Dies ist im Kontext der Analyse von SNP-Daten nicht sinnvoll, da es in der Codierung der SNP-Daten drei Ausprägungsmöglichkeiten gibt und das Fehlen von Ausprägungen zu Informationsverlust führt und dazu, dass die Klassifikation für eine neue Person potentiell nicht möglich ist, sollte diese eine der fehlenden Ausprägungen besitzen. Entsprechend erhält die Klassifikationsrate ebenfalls den Wert 0, da kein Modell angepasst werden kann. Aus demselben Grund wird beim Ziehen der Variablen proportional zu den (C)LS die Ziehung solange wiederholt, bis alle drei Ausprägungsmöglichkeiten vorhanden sind.

Für das Anwenden des logicFS-Ansatzes (vgl. **Kapitel 3.3.2**) werden in jeder Situation dieselben Einstellungen verwendet. Als erstes werden die SNPs des entsprechenden Datensatzes in Binärvariablen codiert mit der R-Funktion „make.snp.dummy“ aus dem R-Paket „logicFS“ (Schwender 2013). In $B = 100$ Wiederholungen des Algorithmus wird in jedem Schritt ein einzelner Baum mit einer maximalen Anzahl von 8 Blättern angepasst. Als Algorithmus zum Anpassen der Bäume wird das Simulated-Annealing verwendet. Die Klassifikationen werden per Bagging bestimmt, indem die zu klassifizierende Beobachtung durch alle B Bäume klassifiziert wird und in die Klasse eingeteilt wird, die durch die Mehrheit der Bäume bestimmt wird. Anschließend wird der klassifizierte Status mit dem wahren Status der Beobachtung verglichen. Dies bildet die Klassifikationsrate k. In dem nächsten Unterkapitel geht es um den ersten Spezialfall bei SNP-Daten. Wie bei vielen anderen Regressionsproblemen liegen in dieser Situation mindestens so viele Beobachtungen wie Variablen vor. An den Daten der Simulationen 1 bis 3 sowie an den Daten des logicFS-Datensatzes wird untersucht, ob es einen Einfluss hat, die Beobachtungen proportional zu den LS in die Stichprobe aufzunehmen.

5.2 Der Fall $n \geq d$

Im Fall $n \geq d$ geht es darum, aus den n Beobachtungen des Datensatzes eine Stichprobe $n' < n$ zu ziehen, um daran ein logisches Regressionsmodell anzupassen. Bewertet werden soll, ob es genügt, einen gewissen Teil der Daten zu verwenden und ob sich bessere Ergebnisse erzielen lassen, wenn diese Beobachtungen proportional zu ihren LS in die Stichprobe aufgenommen werden. Von jedem Datensatz wird ein Prozentsatz P an Beobachtungen in die Stichprobe aufgenommen, welcher vom Umfang des Datensatzes abhängt. An diese Lernstichprobe vom Umfang $n' = \lceil P \cdot n \rceil$ werden logische Regressionsmodelle nach dem logicFS-Ansatz angepasst. Jede der $n'' = n - n'$ Beobachtungen des Testdatensatzes wird anhand dieser logischen Regressionsmodelle über die entsprechenden SNPs ausgewertet und in den Status krank oder gesund klassifiziert. Diese Prognose wird mit dem wahren Status verglichen. Der Anteil der Beobachtungen, die aufgrund ihrer Prognose richtig klassifiziert werden, bildet die Klassifikationsrate k. Zusätzlich werden in ausgewählten Fällen die gefundenen Wechselwirkungen verglichen. Bei der Bestimmung der Modelle wird das Wichtigkeitsmaß in **Gleichung (1)** aus **Kapitel 3.3.2** bestimmt und die gefundenen Wechselwirkungen somit bewertet.

Einführend wird die vorgestellte Methodik an dem logicFS-Datensatz angewandt, danach an den drei Simulationen 1 bis 3.

5.2.1 Auswertung der logicFS-Daten

Als erstes wird das Vorgehen an den logicFS-Daten (siehe **Kapitel 4.1**) angewandt. Auf dem ganzen Datensatz wird einführend ein logisches Regressionsmodell mit dem logicFS-Ansatz angepasst. In **Abbildung 8 (a)** sind die LS der Beobachtungen dargestellt, bestimmt wie in **Kapitel 5.1** beschrieben. Die LS der Fälle sind in rot besonders gekennzeichnet. Als horizontale Linien sind das 0,25-, das 0,5- und das 0,75-Quantil der LS dargestellt. Zu erkennen ist, dass die LS rechtsschief sind. Es gibt mehr Werte die nach oben ausreißen und somit eine hohe Hebelwirkung besitzen. Weiter ist zu erkennen, dass die LS bei den Fällen eine stärker ausgeprägte Schiefe besitzen als bei den Kontrollen, da es mehr extreme Werte gibt.

In **Abbildung 8 (b)** ist die Wichtigkeit der 15 wichtigsten Wechselwirkungen aufgetragen (vgl. **Gleichung (1)**). Die Wechselwirkungen sind in DNF und können somit direkt abgelesen werden. Im Kontext dieser Grafiken bezeichnet $!X_i$ in der gesamten Arbeit das Kompliment der entsprechenden Binärvariable X_i, $i = 1, ..., 30$. Die in **Kapitel 4.1** angegebenen Wechselwirkungen des logicFS-Datensatzes werden durch

$$
\begin{aligned}
L_{FS_{1_{DNF}}} &= (X_1 \wedge X_2) \\
L_{FS_{2_{DNF}}} &= (X_3^{\mathrm{C}} \wedge X_4^{\mathrm{C}} \wedge X_7 \wedge X_8) \\
L_{FS_{3_{DNF}}} &= (X_5 \wedge X_6 \wedge X_9 \wedge X_{10} \wedge X_{11}^{\mathrm{C}} \wedge X_{12}^{\mathrm{C}})
\end{aligned}
$$

als DNF ausgedrückt. Alle der 9 wichtigsten Wechselwirkungen enthalten Binärvariablen, die nach Konstruktion der Daten zum Bilden des Krankheitsstatus beitragen. Die einzige

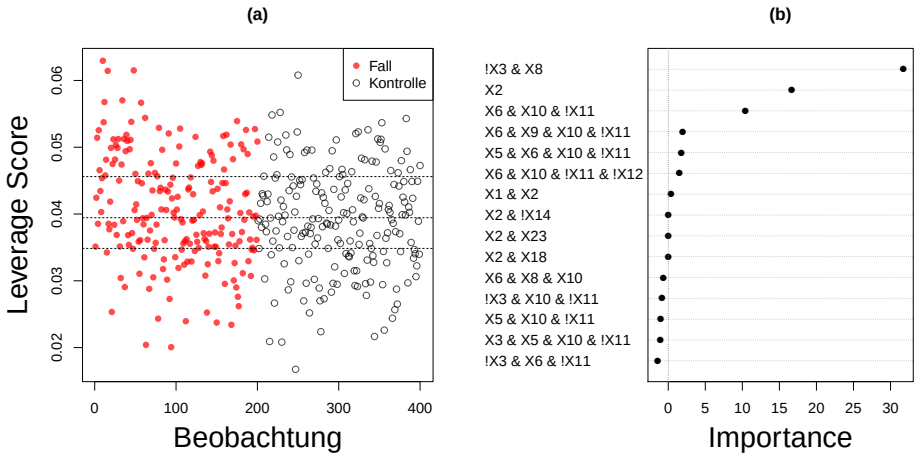

Abbildung 8: Überblick einiger Merkmale des logicFS-Datensatzes: **(a)** Leverage Scores der Beobachtungen, **(b)** Wichtigkeitsmaß der gefundenen Wechselwirkungen beim Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz auf den gesamten Datensatz. Das Komplement einer Binärvariable X_i ist mit $!X_i$ gekennzeichnet.

Wechselwirkung, die vollends identifiziert wird, ist L_{FS_1} und dies mit einer vergleichsweise geringen Wichtigkeit. Negative Werte des Wichtigkeitsmaßes treten auf, wenn es Überschneidungen von Wechselwirkungen gibt, z.B. bei der Wechselwirkung $(X_3^{\mathrm{C}} \wedge X_{10} \wedge X_{11}^{\mathrm{C}})$. Das angepasste Modell besitzt einen OOB-Fehler von 0% was bedeutet, dass jede der Beobachtungen, die nicht Teil der Bootstrap-Stichprobe ist, richtig klassifiziert wird. Nach Schwender und Ickstadt (2008, S. 193) gelangen im Erwartungswert etwa zwei Drittel der Beobachtungen in die Bootstrap-Stichprobe. Dies bedeutet, dass das Modell an etwa einem Drittel der Daten evaluiert wird und jede dieser Beobachtungen richtig klassifiziert wird. Es kann somit davon ausgegangen werden, dass es in dieser Datensituation möglich ist, mit dem richtigen Modell jede Beobachtung korrekt zu klassifizieren.

Nun soll aus dem Datensatz eine Stichprobe vom Umfang n' entnommen werden und daran der logicFS-Ansatz angewendet werden. Als Skizze für die Auswahl der Beobachtungen wird das in **Kapitel 5.1** beschriebene Vorgehen verwendet. Der Umfang der Stichprobe n' entspricht der reduzierten Dimension r. Dem Vorgehen entsprechend wird der Vektor der Realisationen der abhängigen Variable transformiert und mit der Datenmatrix vereinigt, um die LS zu bestimmen. Als Prozentanteil P der Daten werden jeweils 5%, 10%, 15%, 20% und 25% verwendet und die Stichproben besitzen somit einen Umfang von $n' \in \{20, 40, 60, 80, 100\}$ Beobachtungen. In dem Fall $n' = 20$ besitzt die verbleibende Datenmatrix keinen vollen Spaltenrang mehr, da es nach der Transformation der SNP-Daten in Binärdaten mehr Variablen als Beobachtungen gibt. Dieser Anteil ist der geringste Anteil an dem es möglich ist ein logisches Regressionsmodell anzupassen, da sonst nicht genug Beobachtungen zur Verfügung stehen. Als Obergrenze wird 25% gewählt, da es um die Reduktion von Datensätzen geht und daher kein zu großer Anteil der Daten verwendet werden soll.

Die Beobachtungen, welche in die Stichprobe aufgenommen werden, kommen auf un-

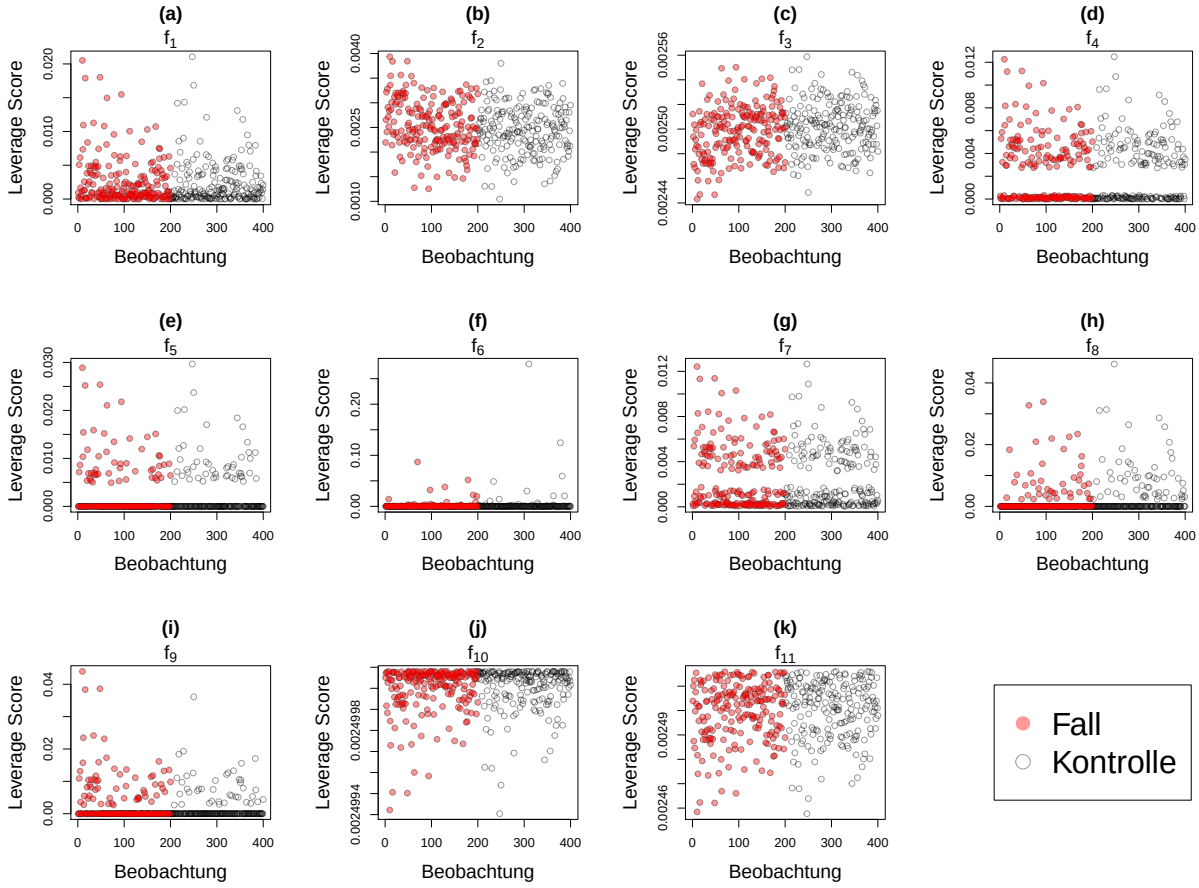

Abbildung 9: Darstellung der Leverage Scores der 400 Beobachtungen des logicFS-Datensatzes nach der Gewichtung durch die Gewichtsfunktionen f_1 bis f_{11}.

terschiedliche Weise in die Stichprobe. Als Referenz wird immer die SRS genommen. Jede Beobachtung besitzt bei der SRS dieselbe Wahrscheinlichkeit in die Stichprobe zu gelangen. Im Vergleich dazu wird jeder Beobachtung eine Wahrscheinlichkeit proportional zu ihrem LS-Wert zugewiesen. Die in **Abbildung 8 (a)** dargestellten LS des Datensatzes werden mit den elf in **Gleichung (9)** bis **(19)** in **Kapitel 5.1** vorgestellten und in **Abbildung 7** dargestellten Gewichtsfunktionen gewichtet und bilden nach der Gewichtung die Auswahlwahrscheinlichkeiten. Dies ist in **Abbildung 9** dargestellt.

Deutlich lassen sich Unterschiede in den resultierenden Werten der LS nach der Gewichtung erkennen. Diese teilweise sehr unterschiedlichen Gewichtungen führen dazu, dass sehr unterschiedliche Beobachtungen mit einer höheren Wahrscheinlichkeit aufgenommen werden.

Neben der zufälligen Auswahl proportional zu den LS wird der Ansatz verfolgt, die Beobachtungen fest anhand ihrer LS in die Stichprobe aufzunehmen. Dazu werden die LS entsprechend der jeweiligen Gewichtsfunktion gewichtet und dann die n' Beobachtungen in die Stichprobe aufgenommen, die nach der Gewichtung die höchsten Werte besitzen. Dies bedeutet z.B. bei der Gewichtung durch f_9, dass die n' Beobachtungen mit den höchsten LS in die Stichprobe aufgenommen werden, wohingegen bei der Gewichtung mit

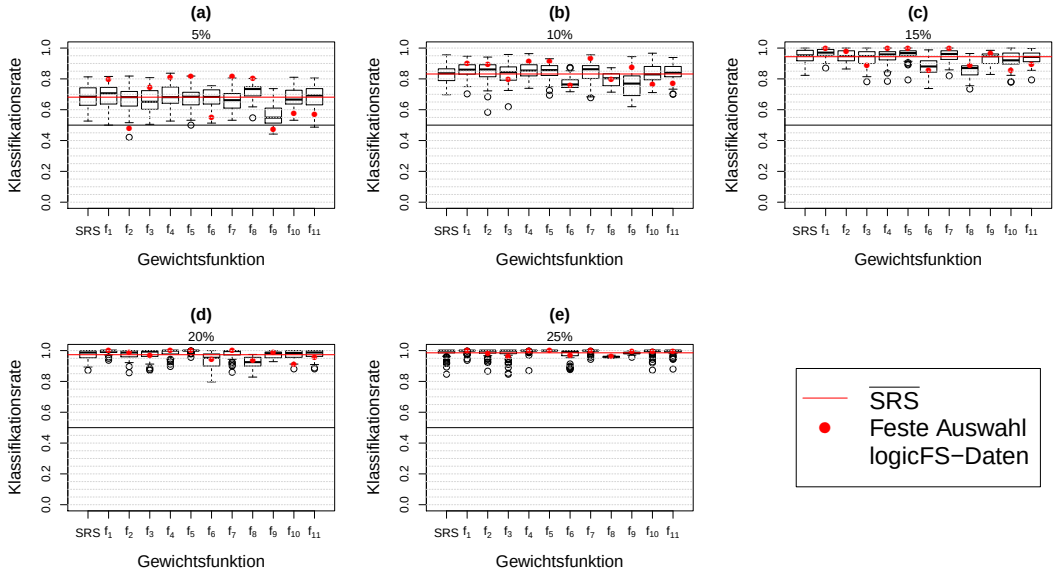

Abbildung 10: Boxplots der Klassifikationsraten bei wiederholter Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf unterschiedlichen Prozentsätzen der logicFS-Daten.

der Gewichtsfunktion f_{10} die Beobachtungen mit LS Werten direkt aus dem Zentrum der Werte aufgenommen werden. Diese Methodik ist sehr starr und es können auf jedem Datensatz immer nur dieselben Beobachtungen in die Stichprobe gelangen. Weiter führt es bei vielen Gewichtsfunktionen dazu, dass dieselben Beobachtungen durch diese Gewichtsfunktionen ausgewählt werden. Da es in dem Datensatz nur eine Realisation der LS gibt, kann die Klassifikation, die durch diese Auswahl durchgeführt wird, nicht weiterführend überprüft werden. Dies wird erst in der anschließenden Simulationsstudie möglich.

Pro Gewichtsfunktion werden 100-mal für jeden der gewählten Anteile der Daten $n' \in \{20, 40, 60, 80, 100\}$ die Beobachtungen zufällig gezogen, jeweils per SRS und proportional zu den gewichteten LS. Mit diesen Daten werden logische Regressionsmodelle nach dem logicFS-Ansatz angepasst und die Klassifikationsrate anhand der restlichen $n'' \in \{380, 360, 340, 320, 300\}$ Beobachtungen bestimmt. Gleichzeitig werden für jede Gewichtung die Beobachtungen fest gewählt und ein entsprechendes Modell angepasst.

In **Abbildung 10** sind Boxplots der so entstehenden Klassifikationsraten dargestellt. Als horizontale Linie ist das arithmetische Mittel der SRS in rot dargestellt. Der Richtwert 0,5 für die Klassifikation ist als schwarze, horizontale Linie dargestellt. Als besonders hervorgehobene Punkte sind jeweils die Klassifikationsraten bei fester Auswahl der Beobachtungen anhand der LS dargestellt.

Wie sich leicht vermuten lässt, ist die Klassifikationsrate von der Anzahl der verwendeten Daten abhängig und je größer der verwendete Anteil ist, umso besser wird die Klassifikation. In **Abbildung 10 (a)** ist auffällig, dass die feste Auswahl der Beobachtungen anhand der LS bei etwa der Hälfte der Gewichtsfunktionen zu einer deutlich besseren Klassifikationsrate führt als die zufällige Auswahl. Dies sind die Gewichtsfunktionen die

zu den niedrigen LS korrespondieren und diese die gegen die mittleren LS diskriminieren. Die besten Klassifikationen lassen sich im arithmetischen Mittel bei 5% der ursprünglichen Daten durch die Gewichtung von f_8 und damit durch die niedrigen LS erzielen. In dieser Arbeit ist damit gemeint, wenn sich Klassifikationen durch eine Gewichtung erzielen lassen, dass die Klassifikationen durch die Stichprobe entstehen die anhand der Gewichtung durch diese Gewichtsfunktion gewählt sind. Wie in **Abbildung 10 (a)** zu erkennen ist, liegt die Box des Boxplots bereits über dem Mittelwert der einfachen Zufallsauswahl $\overline{SRS}_{0,05} = 0,6738$ und somit sind drei Viertel der Klassifikationsraten besser als dieser Wert. Das arithmetische Mittel der Klassifikationsraten bei der Gewichtung durch f_8 liegt bei 0,7182 und der Median bei 0,7342. Dies deutet auf eine leicht ausgeprägte Linksschiefe der Klassifikationsraten hin. Entsprechend gibt es mehr hohe Werte der Klassifikationsraten als niedrige. Weiter bedeutet es, dass bei einer im arithmetischem Mittel ungefähr 4% höheren Klassifikationsrate ungefähr $n'' \cdot 0,04 = 380 \cdot 0,04 \approx 15$ Personen im Mittel zusätzlich häufiger richtig klassifiziert werden als bei der SRS. Im Kontext der Klassifikation des Krankheitsstatus ist dies schon ein Zugewinn. Demgegenüber liegen die Klassifikationsraten bei der festen Auswahl für die Gewichtsfunktionen f_1, f_4, f_5, f_7 und f_8 bei 0,8 oder leicht höher. Dies lässt die erste Vermutung zu, dass die feste Auswahl bei einer Stichprobe von nur 5% der Gesamtdaten besser geeignet ist als die zufällige Auswahl. Dies wird in der folgenden Simulationsstudie näher betrachtet.

Bei dem Übergang von 5% auf 10% des Datensatzes besitzt die Datenmatrix wieder vollen Spaltenrang, da es mehr Beobachtungen als Binärvariablen gibt. In **Abbildung 10 (b)** fällt auf, dass sich die Klassifikationsraten bei allen Gewichtungen verbessern. Dies lässt die Vermutung zu, dass es für gewisse Gewichtungen von Vorteil ist, dass die verwendete Datenmatrix einen vollen Spaltenrang besitzt. Dies gilt besonders für die Gewichtsfunktionen f_6, f_{10} und f_{11} die zu den mittleren LS korrespondieren. Sehr interessant ist zudem, dass die Klassifikation durch die Gewichtsfunktion f_8 nun zum Großteil unter dem arithmetischen Mittel der einfachen Zufallsauswahl $\overline{SRS}_{0,1} = 0,8403$ liegt, mit einem Wert von 0,7978 für das arithmetische Mittel und 0,8042 für den Median der Klassifikationsraten. Dahingegen hebt sich die Gewichtung durch f_1 am besten ab. Bei dieser Gewichtung liegen die Klassifikationsraten vermehrt über dem arithmetischen Mittel der SRS mit einem Wert von 0,8614 für das arithmetische Mittel und 0,8597 für den Median. Weiter lässt sich bei 10% der Originaldaten festhalten, dass die Gewichtungen der mittleren und besonders der hohen LS (durch f_9) im arithmetischen Mittel zu schlechteren Klassifikationsraten führt. Weiter liegt die Klassifikation durch die feste Auswahl bei der Hälfte der Gewichtungen im oberen Bereich der Klassifikationsraten, hebt sich aber nicht mehr so eindeutig ab.

Bei größeren Umfängen der Daten ändert sich die Eignung der Gewichtung. Besonders auffällig ist, dass die Klassifikationen durch die Gewichtsfunktion f_8 bei allen größeren Stichprobenumfängen unter dem mittleren Niveau der SRS bleibt und sich sogar einpendelt. Mit dem mittleren Niveau ist das arithmetische Mittel der Klassifikationsraten ge-

meint. Dieser Wert wird als Richtwert für den Vergleich herangezogen, da er im Gegensatz zu dem Median nicht robust ist. Der Median gibt zwar einen besseren Richtwert für das Zentrum der Klassifikationsraten, jedoch gibt es bei fast allen Gewichtungen Ausreißer in den Klassifikationsraten (siehe etwa **Abbildung 10 (e)**) und es ist von Interesse, wie sich diese Ausreißer auf den mittleren Wert auswirken. Daher wird bewusst ein nicht-robustes Maß für den Vergleich herangezogen.

Die Gewichtungen durch f_1 und f_5 heben sich weiter bei größeren Umfängen der Stichprobe hervor. Die Varianz der Klassifikationsraten ist kleiner als bei den anderen Gewichtungen und das mittlere Niveau liegt zu großen Teilen über dem mittleren Niveau der SRS. Auffällig ist in **Abbildung 10 (e)** zu erkennen, dass bei den meisten Gewichtsfunktionen die Klassifikationsrate bei nahezu 1 liegt, aber es deutliche Ausreißer nach unten gibt. Zusätzlich ist weiter festzuhalten, dass bei den Gewichtsfunktionen, die nur 25% der Daten verwenden, bei dieser Auswahl zwar immer die selben Beobachtungen verwendet werden um das Modell anzupassen, es jedoch Schwankungen in den Klassifikationsraten gibt. Dies liegt daran, dass der logicFS-Ansatz selbst ein zufallsbasiertes Verfahren ist und die Beobachtungen zufällig in die Bootstrap-Stichprobe gelangen. Offensichtlich spielt es eine Rolle welche der Beobachtungen in die Bootstrap-Stichprobe aufgenommen werden.

Die restriktive Gewichtung durch f_6 scheint sehr schlecht für die Auswahl der Beobachtungen zu sein. Sowohl bei der festen Auswahl als auch bei der zufälligen Auswahl proportional zu den LS liegen die Klassifikationsraten in der Regel deutlich unter dem mittleren Niveau der SRS. Als Grund dafür lässt sich überlegen, dass immer dieselben wenigen Beobachtungen mit sehr hoher Wahrscheinlichkeit in die Stichprobe aufgenommen werden und diese nicht Träger einiger der Wechselwirkungen sind. Interessant ist auch der Kontrast zu den beiden Gewichtsfunktionen f_{10} und f_{11}, die ebenfalls zu den mittleren LS korrespondieren (vgl. **Tabelle 2**). Das mittlere Niveau der Klassifikationen durch f_6 liegt bei allen fünf Stichprobenumfängen unter dem Niveau dieser beiden anderen Gewichtungen. Es scheint bei dieser Datensituation nicht sehr vorteilhaft zu sein, so stark gegen die Randbereiche der LS zu diskriminieren.

Die Gewichtsfunktionen f_1 und f_5 heben sich ab 15% der Daten besonders hervor. So liegen in **Abbildung 10 (c)** die Boxen der Boxplots über dem mittleren Niveau der SRS. Dies setzt sich auch bei 20% bzw. 25% der Daten fort und gilt ebenfalls für f_4 ab 20% der Originaldaten. Es treten bei dieser Gewichtung jedoch häufiger Ausreißer nach unten auf. Dies deutet stark darauf hin, dass es günstig ist, die Randbereiche der LS zu berücksichtigen und gegen die mittleren Werte zu diskriminieren. Bei 25% der Daten werden durch f_5 alle zur Verfügung stehenden Beobachtungen verwendet. Dies führt zu einer konstanten Klassifikationsrate von 100%. Nicht einmal durch die Bootstrap-Stichprobe entsteht Variation in den Klassifikationsraten.

Insgesamt ist der Zugewinn in den Klassifikationsraten durch die Auswahl proportional zu den LS eher gering und liegt im Bereich von etwa zwei bis vier Prozentpunkten. Im Kontext des Auffindens einer Krankheit können solche Zugewinne jedoch kritisch sein.

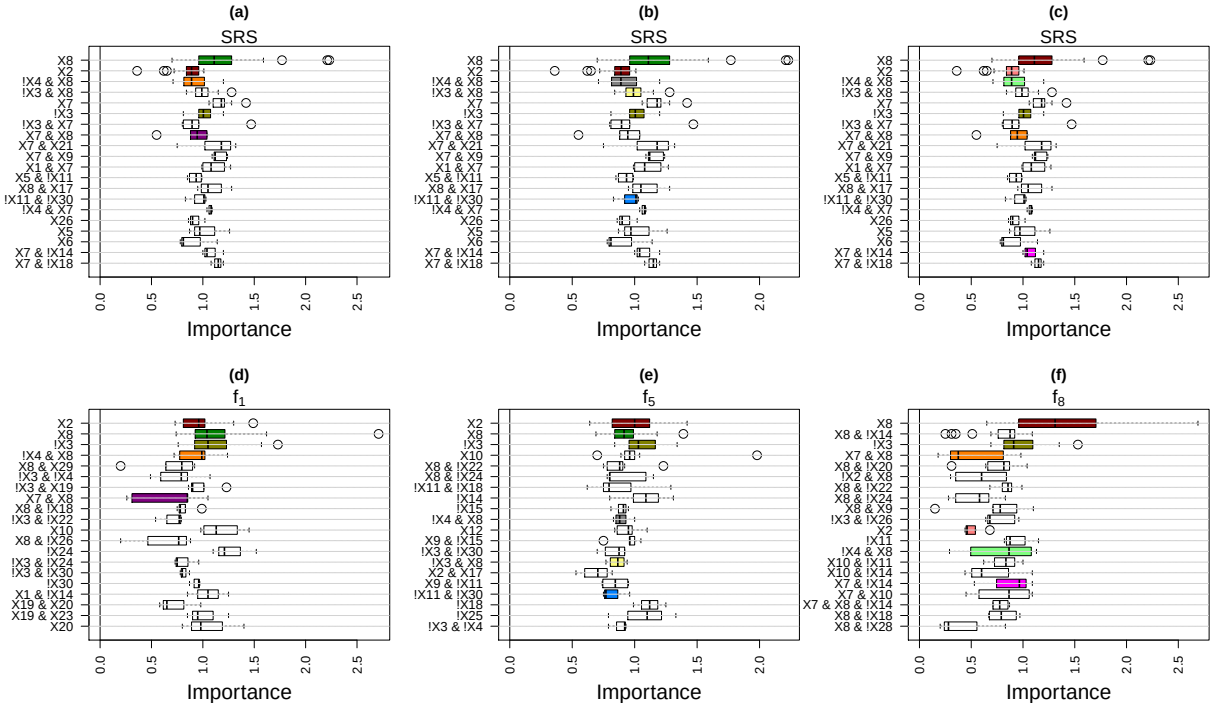

Abbildung 11: Boxplots der Wichtigkeitsmaße der gefundenen Wechselwirkungen des logicFS-Datensatzes bei unterschiedlicher Auswahl einer Stichprobe bestehend aus 5% des Originaldatensatzes, sortiert danach, wie häufig diese gefunden werden. Eingefärbt sind die Überschneidungen der Wechselwirkungen zwischen der einfachen Zufallsauswahl und der zufälligen Auswahl proportional zu den Leverage Scores gewichtet nach: **(d)** f_1, **(e)** f_5, **(f)** f_8.

Daher geht es als nächstes darum, zu evaluieren, welche Wechselwirkungen durch die Modelle gefunden werden und welche Wichtigkeit diese zugewiesen bekommen. Dies ist ein Kernpunkt bei SNP-Daten da es von besonderem Interesse ist, die Wechselwirkungen zu identifizieren, die zuständig für das Auslösen der Krankheit sind und auf ihre Wichtigkeit hin zu beurteilen. Da sich bei dem logicFS-Datensatz besonders die Gewichtsfunktionen f_1, f_5 und f_8 zumindest bei gewissen Anteilen der Daten hervorheben, werden die Modelle dieser Gewichtungen gesondert betrachtet. Um die Gesamtanzahl der Wechselwirkungen überschaubar zu halten, werden aus jedem angepassten Modell jeweils die fünf wichtigsten Wechselwirkungen und das entsprechende Maß herangezogen.

In **Abbildung 11** sind Boxplots der Wichtigkeitsmaße (aus **Gleichung (1)**) der am häufigsten auftretenden Wechselwirkungen beim Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz bei einer Verwendung von 5% der logicFS-Daten dargestellt. Angeordnet sind die Boxplots danach, wie häufig die entsprechende Wechselwirkung durch die Modelle gefunden werden, sortiert von oben nach unten. Dies bedeutet, dass die oberste Wechselwirkung in den Modellen am häufigsten vorkommt, die zweite Wechselwirkung am zweithäufigsten, usw. In **Abbildung 11 (a), (b)** und **(c)** sind die gefundenen Wechselwirkungen durch die Auswahl per SRS dargestellt und darunter in **Abbildung 11 (d),**

(e) und **(f)** jeweils die Wechselwirkungen durch die Auswahl proportional zu den LS mit den Gewichtsfunktionen f_1, f_5 und f_8. Eingefärbt sind die Boxplots der Wechselwirkungen, bei denen es zwischen der SRS und der Wahl proportional zu den LS Überschneidungen gibt.

Als erstes ist festzuhalten, dass bei 5% der verwendeten Daten die Wechselwirkungen recht simpel sind. Unter den häufigsten Wechselwirkungen bei jeder Auswahl ist nur eine Dreifachwechselwirkung vertreten: $(X_7 \wedge X_8 \wedge X_{14}^{\mathrm{C}})$ (vgl. **Abbildung 11 (f)**). Weiter ist anzumerken, dass bei der SRS und der zufälligen Auswahl proportional zu den LS zum Großteil andere Wechselwirkungen gefunden werden. Für jede der drei Gewichtsfunktionen gibt es nur jeweils sechs Wechselwirkungen, die sich mit der SRS decken. Insgesamt besitzt das Wichtigkeitsmaß geringe Werte und für einige der Wechselwirkungen eine große Varianz. Durch die SRS wird nur eine einzelne Binärvariable gefunden, die nach Konstruktion unbedeutend für die Erklärung des Krankheitsstatus ist. In allen anderen Wechselwirkungen ist mindestens eine nach Konstruktion wichtige Binärvariable vertreten.

Auffällig ist, dass die Binärvariable X_8 besonders häufig durch die Gewichtung von f_8 in den Wechselwirkungen vorkommt, bei insgesamt 11 der 20 Wechselwirkungen. Bei der Anwendung des logicFS-Ansatzes auf den vollen Datensatz ist diese Binärvariable in 4 der 15 wichtigsten Wechselwirkungen vertreten (vgl. **Abbildung 8 (b)**) dazu noch in der Wechselwirkung mit der höchsten Wichtigkeit: $(X_3^{\mathrm{C}} \wedge X_8)$. Die Binärvariable X_3^{C} wird am dritthäufigsten durch die Gewichtung mit f_8 gefunden. Dies legt erstens die Vermutung nahe, dass die Träger dieser Binärvariable niedrige LS besitzen und zweitens, dass das Auffinden dieser Binärvariable möglicherweise zuständig ist für die im arithmetischen Mittel besseren Klassifikationsraten durch diese Gewichtung. Als Kontrast dazu ist die bei den anderen Gewichtungen am häufigsten gefundene Binärvariable X_2 im vollen Modell erst an elfter Stelle, mit einer vergleichsweise geringen Wichtigkeit. Die Gewichtung durch f_1 findet viele unwichtige Wechselwirkungen. Fünf der am häufigsten gefundenen Wechselwirkungen besitzen nach Konstruktion der Daten keinen Einfluss auf den Krankheitsstatus. Bei der Gewichtung durch f_5 werden ebenfalls viele einzelne Binärvariablen als wichtig erkannt, die nach Konstruktion nicht bedeutend für den Krankheitsstatus sind. Bei diesem Stichprobenumfang liegen vermutlich aus diesem Grund die Klassifikationsraten im mittleren Niveau nicht über denen der SRS.

Bei der festen Auswahl der Beobachtungen anhand der LS liegt bei 5% der Daten die Klassifikationsrate von drei Gewichtsfunktionen über denen der zufälligen Auswahl (vgl. **Abbildung 10 (a)**). Die Gewichtung durch f_1 findet in den Wechselwirkungen vermehrt die Binärvariable X_2. Diese ist in 7 der 20 wichtigsten Wechselwirkungen vertreten. Weiter wird die Binärvariable X_8 als wichtigste Binärvariable mit einer Wichtigkeit von 1,4 bewertet. Die Binärvariable X_{29}^{C} wird als einziger Einfluss gefunden, der nach Konstruktion nicht wichtig ist. Dies steht im Kontrast zu der zufälligen Auswahl. Dies könnte eventuell der Grund für die deutlich bessere Klassifikationsrate der festen Auswahl sein.

Bei der festen Auswahl mit f_5 ist die Binärvariable X_2 noch häufiger in den Wechselwirkungen vertreten, in 9 der 20 wichtigsten Wechselwirkungen. Die Binärvariable X_8 wird ebenfalls als am wichtigsten bewertet. Interessanterweise sind 4 Wechselwirkungen dabei, die nach Konstruktion keinen Einfluss besitzen. Dies scheint sich jedoch nicht auf die Klassifikationsrate auszuwirken. Bei der Gewichtung durch f_8 werden erstens komplexere Wechselwirkungen bei diesem Prozentanteil gefunden und erneut sehr zuverlässig die Binärvariable X_8. Diese kommt in 15 der 20 wichtigsten Wechselwirkungen vor. Zudem wird keine nach Konstruktion nicht einflussreiche Binärvariable alleine selektiert. Die Binärvariable X_2 findet sich dagegen gar nicht unter den Wechselwirkungen. Anscheinend genügt es, für eine vergleichsweise hohe Klassifikationsrate, eine der beiden Binärvariablen selektiert zu haben. Dies kann jedoch auch eine Erklärung für die etwa 20% falschklassifizierten Beobachtungen sein. Dies legt stark die Vermutung nahe, dass die Träger der entsprechenden Wechselwirkungen andere LS besitzen.

Eine Erhöhung des Stichprobenumfangs P führt dazu, dass die gefundenen Wechselwirkungen komplexer werden. Dies geht mit einer Verbesserung der Klassifikationsraten einher. Weiter verstärken sich die Überschneidung zwischen den gefundenen Wechselwirkungen durch die SRS und durch die Auswahl proportional zu den LS. Interessanterweise nimmt dies bei einem Stichprobenumfang von 25% der Originaldaten wieder ab. Dies liegt vermutlich daran, dass die Wechselwirkungen immer komplexer werden und somit die Überschneidungen der eindeutig selben Wechselwirkungen wieder abnehmen. Wie bereits angemerkt und in **Abbildung 10** zu erkennen, ist die Verbesserung der Klassifikationsrate durch eine Erhöhung des Stichprobenumfangs bei der Gewichtsfunktion f_8 weniger deutlich als bei den anderen Gewichtsfunktionen. Dazu nehmen die Überschneidungen mit den Wechselwirkungen der SRS auch nicht so deutlich zu und liegen hinter den anderen Gewichtungen zurück. Da f_8 nur niedrige LS berücksichtigt und demgegenüber f_1 und f_5 auch hohe und mittlere LS ist dies ein weiteres Indiz dafür, dass Beobachtungen mit niedrigen LS andere Wechselwirkungen besitzen als Beobachtungen mit mittleren und hohen LS. So bedarf es einem Stichprobenumfang von 20% der Originaldaten bis die Binärvariable X_2 durch f_8 unter die fünf am häufigsten erkannten Wechselwirkungen aufgenommen wird und dazu mit einer vergleichsweise geringen Wichtigkeit. Demgegenüber wird X_2 durch die beiden anderen Gewichtungen als deutlich wichtiger bewertet und befindet sich immer unter den am häufigsten gefundenen Wechselwirkungen. Schon bei 10% der Originaldaten treten keine Wechselwirkungen mehr auf, die nicht wenigsten eine nach Konstruktion wichtige Binärvariable enthalten. Die drei wichtigsten Wechselwirkungen beim Anpassen der Modelle auf dem gesamten Datensatz X_2, $(X_3^{\mathrm{C}} \wedge X_8)$ und $(X_6 \wedge X_{10} \wedge X_{11}^{\mathrm{C}})$ werden durch die SRS, f_1 und f_5 bereits bei 10% der Originaldaten erkannt und finden sich unter den wichtigsten Wechselwirkungen. Bei f_8 findet sich $(X_6 \wedge X_{10} \wedge X_{11}^{\mathrm{C}})$ erst ab 15% der Originaldaten unter den am häufigsten gefundenen Wechselwirkungen, wohingegen $(X_3^{\mathrm{C}} \wedge X_8)$ die höchste Wichtigkeit erhält.

In **Abbildung 12** sind Boxplots der Wichtigkeitsmaße der am häufigsten gefunde-

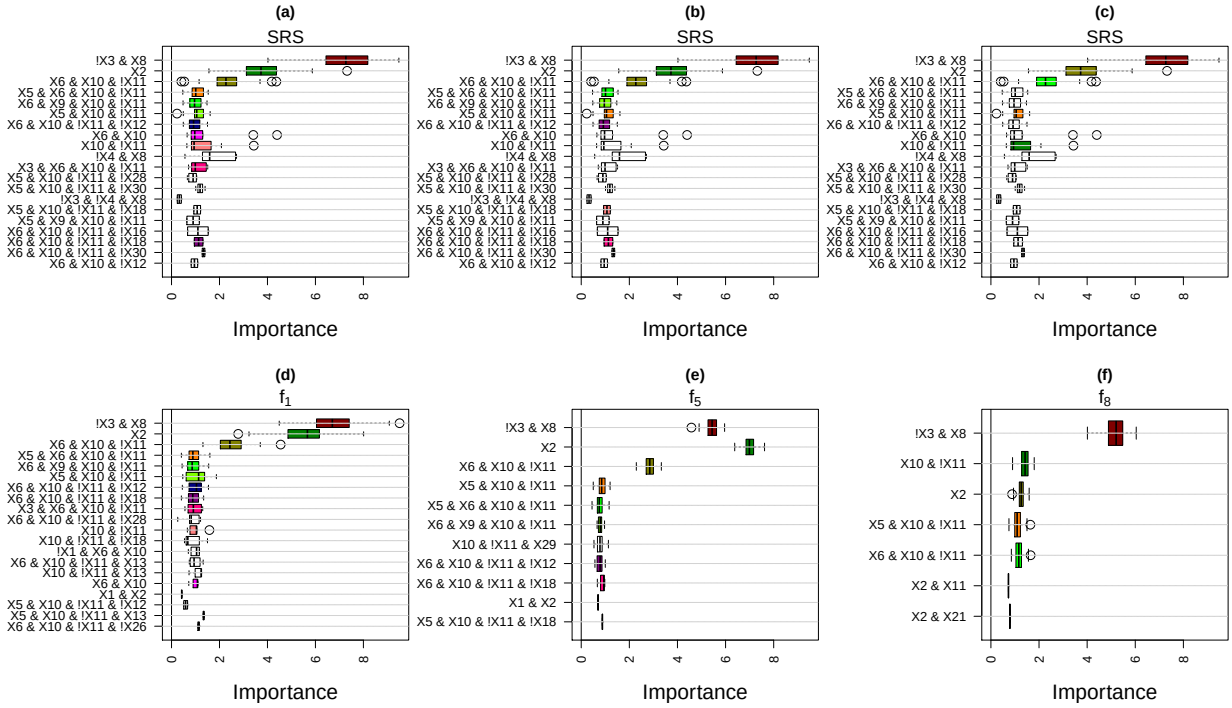

Abbildung 12: Boxplots der Wichtigkeitsmaße der gefundenen Wechselwirkungen des logicFS-Datensatzes bei unterschiedlicher Auswahl einer Stichprobe bestehend aus 25% des Originaldatensatzes, sortiert danach, wie häufig diese gefunden werden. Eingefärbt sind die Überschneidungen der Wechselwirkungen zwischen der einfachen Zufallsauswahl und der zufälligen Auswahl proportional zu den Leverage Scores gewichtet nach: **(d)** f_1, **(e)** f_5, **(f)** f_8.

nen Wechselwirkungen bei der Verwendung von 25% der logicFS-Daten dargestellt. Im Vergleich zu **Abbildung 11** ist deutlich zu erkennen, dass die Wechselwirkungen komplexer sind. Weiter ist festzuhalten, dass von den beiden Gewichtsfunktionen f_5 und f_8 nur 25% der Gesamtdaten verwendet werden (siehe **Tabelle 2**). Damit ist das zufällige Ziehen proportional zu den LS bei diesem Prozentanteil gleich mit der festen Auswahl. Die Varianz der Wichtigkeit in **Abbildung 12 (e)** und **(f)** liegt nicht an der Auswahl der Beobachtungen, sondern nur daran, welche der ausgewählten Beobachtungen in die Bootstrap-Stichprobe gelangen. Bei der Gewichtung mit f_5 führt dies zu einer konstanten Klassifikationsrate von 100%. Wie in **Abbildung 10 (e)** zu erkennen ist, gibt es teilweise deutliche Ausreißer bei den Klassifikationsraten der anderen Gewichtsfunktionen. Die beiden Randbereiche der LS scheinen somit in dieser Datensituation sehr gut dazu geeignet zu sein, die Beobachtungen für das Anpassen der Modelle zu wählen. Nur zwei Binärvariablen in allen 11 gefundenen Wechselwirkungen sind nach Konstruktion nicht von Bedeutung. Bei der SRS sind dies dreimal so viele Binärvariablen wobei zu berücksichtigen ist, dass erstens mehr Wechselwirkungen gefunden werden und diese nur mit geringer Gewichtung einfließen. Dies kann jedoch ein Grund für die Ausreißer in den Klassifikationsraten bei der SRS sein. Ähnliches gilt für die Gewichtung durch f_1. Weiter

ist festzuhalten, dass die Gewichtung durch f_8 bei diesem Prozentanteil zum schlechtesten mittleren Niveau bei den Klassifikationsraten führt (siehe **Abbildung 10 (e)**). Dafür scheint verantwortlich zu sein, dass nur eingeschränkt Wechselwirkungen gefunden werden und diese eine vergleichsweise geringe Wichtigkeit erhalten. Im Vergleich dazu liegt die Klassifikationsrate im arithmetischen Mittel bei der Gewichtung durch f_9 um etwa 3 Prozentpunkte höher. Die Gewichtsfunktion hebt dabei die hohen LS hervor. Von den Gewichtungen, die nur einen Teil der Daten zulassen, schneidet jedoch die Vereinigung der beiden Bereiche (durch f_5) am besten ab. Dies weist erneut darauf hin, dass beide Randbereiche zu berücksichtigen sind. Weiter steht im Vergleich zu f_8 die Gewichtsfunktion f_3 die ebenfalls die niedrigen LS hervorhebt. Die Gewichtung durch diese Gewichtsfunktion liegt bei größeren Stichprobenumfängen zwar ebenfalls hinter den anderen Gewichtungen zurück, dafür jedoch im mittleren Niveau über f_8. Dies lässt die Vermutung zu, dass es nicht sinnvoll ist, nur die 25% der Daten mit den niedrigsten LS zu verwenden.

Bereits 15% der Originaldaten sind genug, um mit der Gewichtung durch f_1 bei fester Auswahl der Beobachtungen eine Klassifikationsrate von nahezu 100% zu erhalten. Mit höheren Prozentsätzen der verwendeten Daten liegt die Rate konstant bei 100%. Die gefundenen Wechselwirkungen nähern sich dabei schon denen aus **Abbildung 8 (b)** an. Bei allen drei Prozentsätzen und im vollen Modell werden die Wechselwirkungen $(X_3^{\mathrm{C}} \wedge X_8)$, X_2 und $(X_6 \wedge X_{10} \wedge X_{11}^{\mathrm{C}})$ als die drei wichtigsten Wechselwirkungen identifiziert, wobei das Wichtigkeitsmaß mit dem Stichprobenumfang wächst. Bei der Gewichtung durch f_5 werden ebenfalls diese drei Wechselwirkung gefunden und das Wichtigkeitsmaß ist höher. Zudem liegt die Klassifikationsrate bereits bei 15% der Originaldaten bei 100%. Bei der festen Auswahl der Beobachtungen durch die Gewichtung mit f_8 ist die Binärvariable X_2 erst bei 25% der Daten unter den wichtigsten Wechselwirkungen. Erneut lässt dies die Vermutung zu, dass dies der Grund für das schlechtere Abschneiden dieser Gewichtung bei mehr Prozent der Originaldaten ist.

Die Auswertung des logicFS-Datensatzes legt den Verdacht nahe, dass bei sehr kleinen Stichprobenumfängen die Gewichtung der niedrigen LS einen positiven Einfluss auf die Klassifikationsrate hat, bei größeren Umfängen die Gewichtung beider Randbereiche der LS und dass die feste Auswahl der Beobachtungen besser geeignet ist als das zufällige Auswählen proportional zu den LS. Da es sich bei dem logicFS-Datensatz um einen simulierten Datensatz handelt, ist es jedoch nicht so einfach möglich mit nur einer Realisation dieses Datensatzes die Ergebnisse zu verallgemeinern. Aus diesem Grund wird die in **Kapitel 4.2** beschriebene Simulationsstudie durchgeführt, um deutlich mehr Daten zur Verfügung zu haben und gleichzeitig verschiedene Datensituationen abzudecken. Simulation 1 ist direkt von dem logicFS-Datensatz inspiriert und diesem nachempfunden. Der Unterschied liegt in der Anzahl der Variablen. Absichtlich wird eine steigende Anzahl Variablen gewählt, um zu untersuchen ob es einen Einfluss hat wie viele uninformative Variablen in den Datensätzen sind.

In Simulation 2 geht es darum, zu beurteilen wie es sich auf die Daten auswirkt wenn

nicht alle Fälle den Krankheitsstatus durch die genetischen Einflüsse der SNPs erklärt haben. Bei den Datensätzen dieser Simulation ist entsprechend nur ein Teil der Fälle durch die SNPs erklärt und die restlichen Fälle durch andere Faktoren.

Bei Simulation 3 geht es darum, einen hochdimensionierten Datensatz vorliegen zu haben, mit sehr viel mehr Beobachtungen als Variablen ($n \gg d$). Da es im Endeffekt um so hochdimensionierte Datensätze geht, ist es von besonderem Interesse solche Daten zu untersuchen, um herauszufinden ob die LS einen positiven Einfluss auf die Auswahl der Beobachtungen haben.

5.2.2 Auswertung von Simulation 1

Nun geht es um die Auswertung der Simulation 1. Für jedes $d \in \{10, 20, 30, 40, 50\}$ werden jeweils 100 Datensätze simuliert. Entsprechend besteht Simulation 1 aus 500 Datensätzen. Auf jedem der 500 Datensätze wird zunächst der logicFS-Ansatz angewandt, mit den beschriebenen Einstellungen. Es ergibt sich, dass der OOB-Fehler unabhängig von der Anzahl der Variablen ist. Für jede Anzahl d liegt der Median des OOB-Fehlers bei 0,25%, das arithmetische Mittel bei etwa 0,3% und das Maximum bei 1,25%. Bei jeweils 400 Beobachtung pro Datensatz bedeutet dies, dass höchstens fünf Beobachtungen nicht richtig klassifiziert werden. Diese Fehlklassifikationen lassen sich vermutlich dadurch erklären, dass der logicFS-Ansatz ein zufallsbasiertes Verfahren ist. Es kann bei dieser Datensituation davon ausgegangen werden, dass es möglich ist ein Modell anzupassen das nahezu jede Beobachtung richtig klassifiziert und das dies unabhängig von der Anzahl der Variablen d ist.

In **Abbildung 13** sind Boxplot der Wichtigkeitsmaße (aus **Gleichung (1)**) der 15 auf allen Datensätzen am häufigsten gefundenen Wechselwirkungen dargestellt, getrennt nach Anzahl der Variablen im Datensatz. Sortiert sind die Wechselwirkungen von oben nach unten, wobei die oberste Wechselwirkung am häufigsten gefunden wird, die zweite am zweithäufigsten usw. Die nach Konstruktion wahren Wechselwirkungen in DNF finden sich in **Gleichung (6)** bis **(8)**. Auffällig ist, dass die drei Wechselwirkungen bzw. Binärvariablen X_{11}^{C}, X_{12} und $(X_7 \wedge X_{10})$ in allen Datensituationen in der selben Reihenfolge gefunden werden. Der Umstand, dass die Binärvariablen X_{11}^{C} und X_{12} am häufigsten selektiert werden ist etwas verwunderlich. Diese gehören zu Trägern von L_3 und diese Gruppe ist die nach Konstruktion der Datensätze am schwächsten besetzte Gruppe (vgl. **Tabelle 1**). Jedoch zeigt sich beim Wichtigkeitsmaß, dass es Einflüsse anderer Wechselwirkungen gibt, die im arithmetischem Mittel eine höhere Wichtigkeit besitzen. Weiter fällt auf, dass einige der Wechselwirkungen eine sehr hohe Varianz bei dem Wichtigkeitsmaß besitzen. Diese gehören in erster Linie zu L_1. Eine Erklärung dafür könnte sein, dass bei einigen der Datensätze diese Wechselwirkung verkürzt wurde, da der Krankheitsstatus bereits aus Teilen der Wechselwirkung erklärt wird. Keine der am häufigsten gefundenen Wechselwirkungen enthält Binärvariablen die nach Konstruktion nicht von Bedeutung sind. Negative Werte treten in erster Linie dann auf, wenn Überschneidungen zwischen

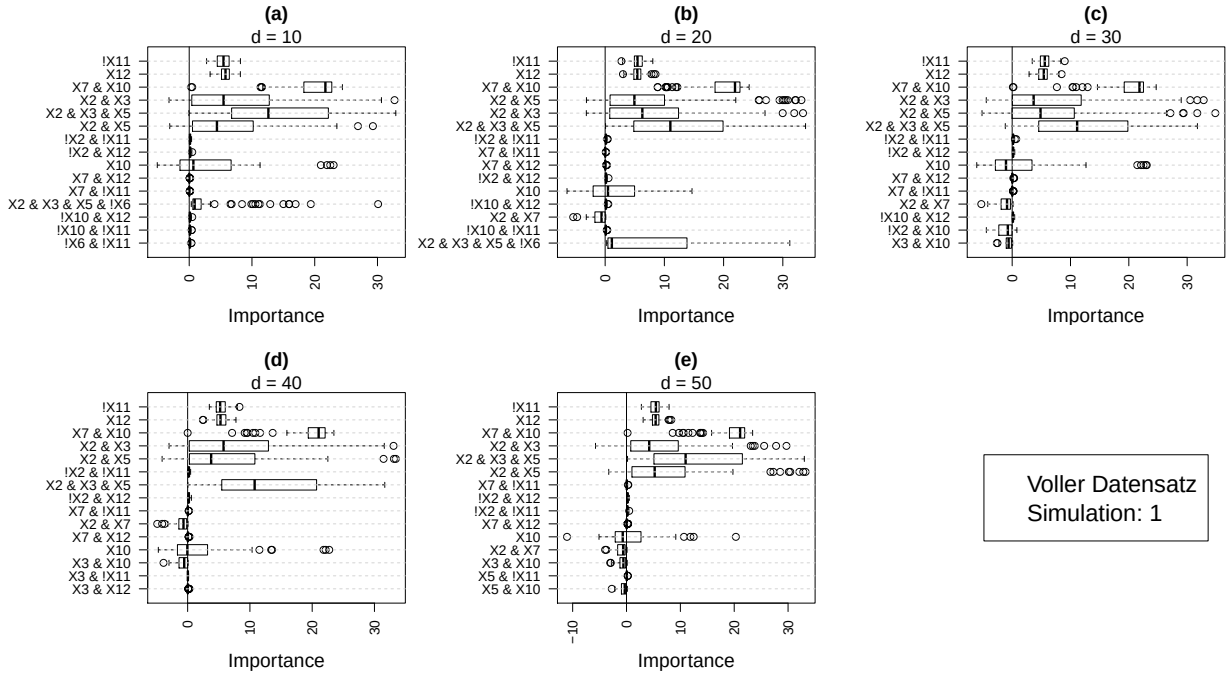

Abbildung 13: Boxplots der Wichtigkeitsmaße der am häufigsten gefundenen Wechselwirkungen beim Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf die vollen Datensätze der Simulation 1.

Wechselwirkungen gefunden werden, ähnlich wie bei dem logicFS-Datensatz. Interessanterweise erhält die Binärvariable X_{10} häufiger einen negativen Wert. Insgesamt sind die Wechselwirkungen nicht sehr komplex, was an der maximalen Anzahl von 8 Blättern pro Baum liegen kann. Eine Erhöhung der Blätter führt jedoch erstens dazu, dass die Berechnung aufwändiger wird und damit länger dauert und zweitens, dass zudem die Gefahr besteht, dass Binärvariablen in die Wechselwirkungen aufgenommen werden, die von der Konstruktion der Daten her nicht für den Krankheitsstatus verantwortlich sind. Insgesamt scheinen die Modelle nur bedingt von der Anzahl der Variablen abzuhängen. Den größten Einfluss hat die Anzahl d auf die benötigte Zeit zum Anpassen der Modelle, die mit steigendem d teilweise deutlich ansteigt.

Das Finden und Bewerten der Wechselwirkungen scheint bei dieser Datensituation vergleichsweise schwer zu sein. Es lässt sich vermuten, dass sich dies potentiell negativ auf die Stichproben auswirkt, da es mit dem vollen Datensatz schon schwer ist, einige der Wechselwirkungen zuverlässig zu finden. Dies überträgt sich aber offensichtlich nicht auf den OOB-Fehler, weshalb es gut möglich ist, dass die Klassifikationsraten nicht dadurch beeinflusst werden.

Erneut sollen Stichproben vom Umfang n' aus den Datensätzen entnommen werden, um daran den logicFS-Ansatz anzuwenden. Für das Bestimmen der Beobachtungen werden für jeden Datensatz die LS nach dem in **Kapitel 5.1** vorgestellten Vorgehen bestimmt, mit den Gewichtsfunktionen gewichtet und diese Bilden für den entsprechenden Datensatz die Auswahlwahrscheinlichkeiten bzw. das Kriterium für die feste Auswahl der Beobach-

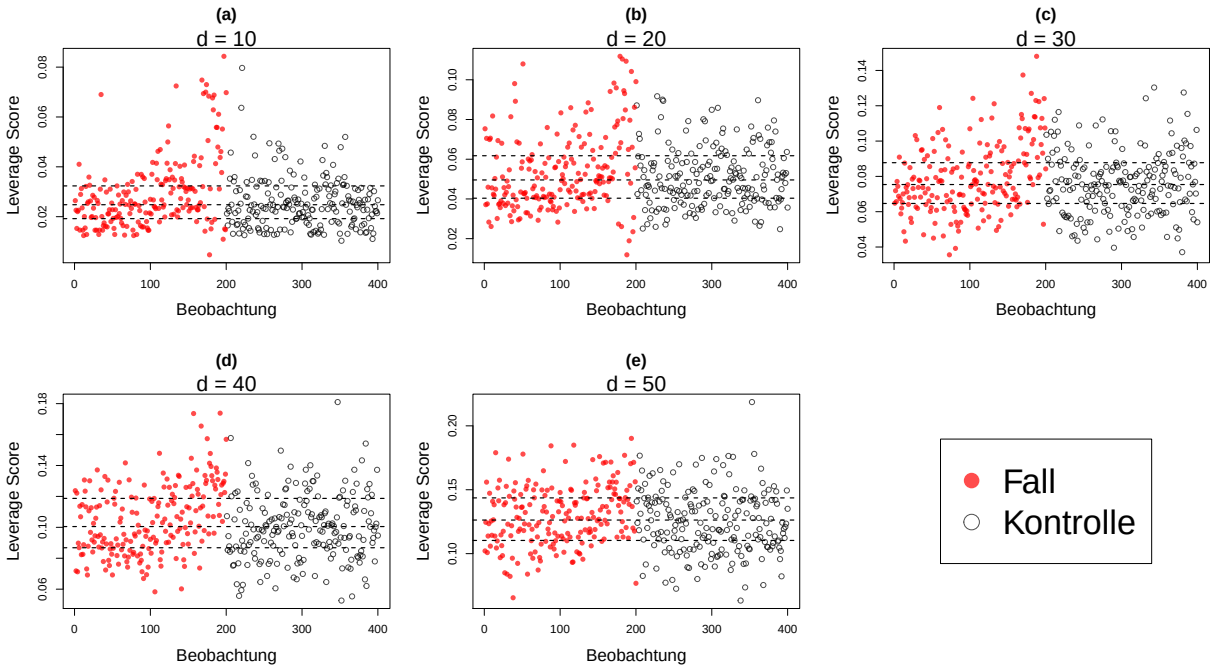

Abbildung 14: Beispiel der Leverage Scores von jeweils 400 Beobachtungen unterschiedlicher Datensätze mit Anzahl Variablen d aus Simulation 1.

tungen. Da das Ziehen der Beobachtungen proportional zu den LS ein zufallsbasiertes Verfahren ist, ist es wie bei dem logicFS-Datensatz nötig, das Verfahren auf jedem Datensatz zu wiederholen. Aufgrund des großen Umfangs der vorliegenden Daten geschieht dies auf jedem Datensatz 20-mal. Für jede Anzahl d der Variablen wird das Verfahren somit für jede Gewichtsfunktion 2000-mal angewandt. Ein weiterer Vorteil der Simulationsstudie ist, dass die feste Auswahl der Beobachtungen anhand der LS vergleichbar wird. Bei dem logicFS-Datensatz schneidet diese Methode teilweise besonders gut ab und es ist interessant, zu untersuchen, ob dies auf mehreren Realisationen der Datensätze ähnlich gut funktioniert. Entsprechend ist dies für jede Anzahl d bei jeder Gewichtsfunktion 100-mal möglich.

In **Abbildung 14** sind beispielhaft die LS der jeweils 400 Beobachtungen verschiedener Datensätze aus Simulation 1 dargestellt. Für jede Anzahl d gibt es eine Darstellung. Es ist zu erkennen, dass die LS mit steigender Anzahl an Variablen größere Werte annehmen. Dies liegt daran, dass $\sum_{i=1}^{n} l_i = d$ ist (vgl. **Kapitel 5.1**). Weiter nehmen mit größerem d die Ausreißer in den Werten ab und es gibt nicht mehr so extreme Werte. Wie bei dem logicFS-Datensatz gibt es mehr höhere als niedrigere Werte, entsprechend sind die LS rechtsschief. Dies setzt sich über die LS aller Datensätze fort. Die in **Abbildung 14 (a)** zu erkennenden Beobachtungen mit sehr hohen Werte im Zentrum der Daten sind die Beobachtungen mit der Wechselwirkung L_3. Der Umstand, dass sich viele hohe Werte bei diesen Beobachtungen ergeben, ist bei allen 100 Datensätzen mit $d = 10$ zu erkennen. Ansonsten ist immer zu erkennen, dass die LS der Fälle dazu neigen, höhere Werte anzunehmen als die LS der Kontrollen, wobei dies mit mehr Variablen nicht mehr ganz so

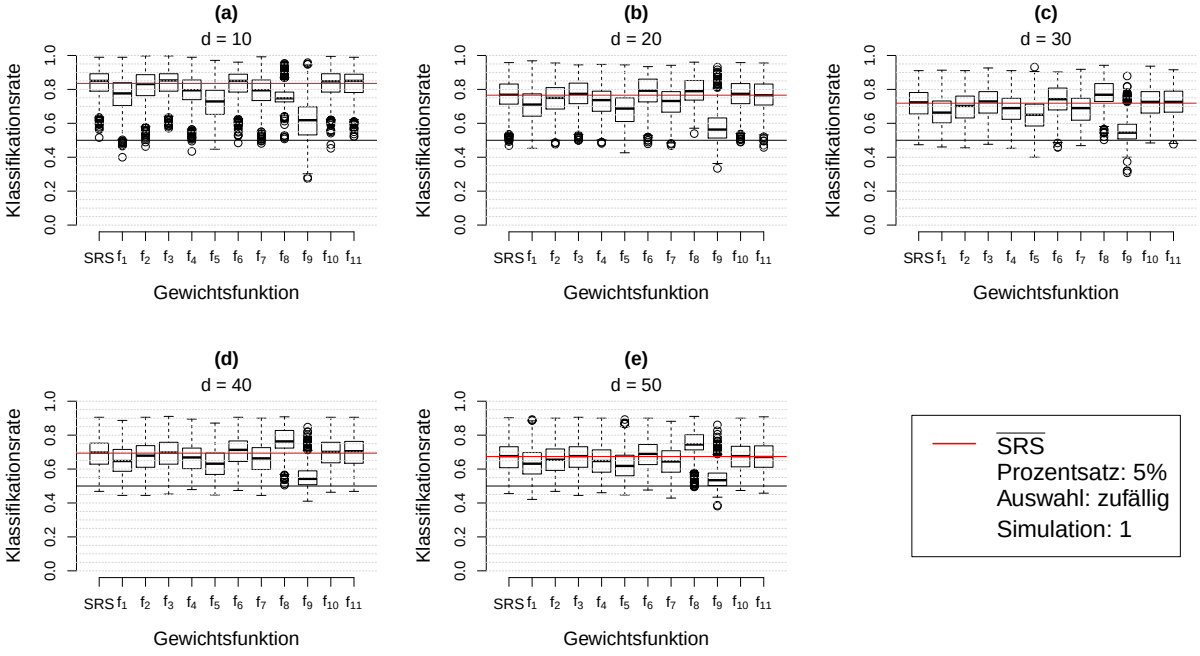

Abbildung 15: Boxplots der Klassifikationsraten bei wiederholter Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 5% der Daten aus Simulation 1, bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

stark zu sein scheint. Vereinzelt gibt es Datensätze in denen die LS stärker streuen und die Kontrollen ebenfalls extremere LS-Werte besitzen. Die unterschiedlichen Ausprägungen der LS in den einzelnen Gruppen der Wechselwirkungen legen die Vermutung nahe, dass eine Gewichtung nur dieser speziellen Bereiche (z.B. die Gewichtung der hohen LS durch f_9) dazu führt, dass dieser Einfluss durch die Gewichtung bevorzugt erkannt wird.

Erneut werden als Prozentanteile P der Daten 5%,10%,15%,20% und 25% der Originaldaten verwendet und somit Stichproben der Umfänge $n' \in \{20, 40, 60, 80, 100\}$. Angepasst werden die Modelle mit denselben Einstellungen wie bei dem gesamten Datensatz. In **Abbildung 15** sind Boxplots der Klassifikationsraten bei einem Stichprobenumfang von 5% der Originaldaten dargestellt. Getrennt sind die Darstellungen nach der Anzahl der Variablen d in den entsprechenden Datensätzen. Als rote horizontale Linie ist das arithmetische Mittel der SRS dargestellt und als schwarze horizontale Linie der Richtwert 0,5. Grundsätzlich lässt sich festhalten, dass das mittlere Niveau der SRS mit einer steigenden Anzahl an Variablen immer mehr abnimmt. Weiter ist zu erkennen, dass die Klassifikationsraten durch die Auswahl proportional zu den LS ebenfalls durch die Anzahl der Variablen beeinflusst wird. Alle Gewichtungen verlieren im mittleren Niveau durch die steigende Anzahl Variablen. Jedoch unterscheidet sich die Intensität, in der dies geschieht.

In **Tabelle 3** sind die arithmetischen Mittel der Klassifikationsraten beim Anpassen der logischen Regressionsmodelle nach dem logicFS-Ansatz auf 5% der Originaldaten in Prozent angegeben. In der ersten Spalte ist angegeben, wie viele Variablen sich jeweils in dem Datensatz befinden. Die restlichen Spalten enthalten das zu der jeweiligen Gewich-

Tabelle 3: Arithmetisches Mittel der Klassifikationsraten in Prozent bei wiederholter Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 5% der Daten aus Simulation 1, bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

d	SRS	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}
10	83,54	76,87	81,87	**83,61**	79,20	72,24	82,73	79,06	76,24	62,23	83,54	83,26
20	76,59	70,77	74,52	76,91	72,80	68,31	78,42	72,43	**79,33**	57,85	76,81	76,42
30	71,87	66,56	69,75	72,45	68,63	65,15	74,04	68,36	**77,74**	55,73	72,24	72,45
40	69,37	65,07	67,68	69,59	66,38	63,37	70,52	66,13	**77,03**	55,37	69,75	69,99
50	67,40	63,65	65,69	67,37	64,85	62,38	68,82	64,66	**75,12**	54,65	67,45	67,29

tung gehörende arithmetische Mittel. Für jede Anzahl an Variablen ist der höchste Wert besonders hervorgehoben.

Festzuhalten ist, dass die Gewichtung durch f_8 bereits ab $d = 20$ den höchsten Wert des arithmetischen Mittels besitzt und sich dies für jedes folgende d fortsetzt. Auffällig ist weiterhin der Sprung im mittleren Niveau von $d = 10$ auf $d = 20$ für diese Gewichtung, da dies die einzige Gewichtung ist, bei der das mittlere Niveau ansteigt anstelle zu sinken. Dies deckt sich mit den Ergebnissen der Auswertung des logicFS-Datensatzes, da bei diesem Datensatz bei 5% der Originaldaten ebenfalls f_8 die besten Klassifikationsraten erzielt. Weiter bleibt das mittlere Niveau von f_8 viel konstanter als bei den anderen Gewichtungen. Bereits ab $d = 30$ liegt in **Abbildung 15** die Box der Klassifikationsraten durch die Gewichtsfunktion f_8 über dem mittleren Niveau der SRS. Somit sind 75% der Klassifikationsraten besser als dieser Wert. Für die Wahl der Beobachtungen bei nur 5% der Originaldaten sind die niedrigen LS somit wie bei den logicFS-Daten am besten geeignet. Besonders fällt dies bei $d = 50$ auf. Hier liegt die mittlere Klassifikationsrate mindestens 6 Prozentpunkte höher als bei allen anderen Gewichtungen. Dies entspricht mindestens 30 Personen, die im Mittel zusätzlich häufiger richtig klassifiziert werden. Darüber hinaus ist die Varianz der Klassifikationsraten am geringsten. Interessant dabei ist erneut der Kontrast zwischen f_3 und f_8, die beide zu den niedrigen LS korrespondieren. Jedoch diskriminiert f_3 nicht gegen die anderen Wertebereiche. Dies scheint jedoch ausschlaggebend für die Unterschiede im Niveau der Klassifikationsraten zu sein. Dies deckt sich erneut mit den Ergebnissen der Auswertung des logicFS-Datensatzes.

Die Varianz scheint für alle Gewichtungen nicht von d abhängig zu sein, mit Ausnahme der Gewichtsfunktion f_9 bei der sich die Varianz nahezu halbiert. Weiter ist auffällig bei der Gewichtung durch f_9, dass es ab $d = 20$ deutlich Ausreißer nach oben gibt (vgl. **Abbildung 15**). Die schlechte Klassifikation durch f_9 lässt sich vermutlich durch die in **Abbildung 14** gewonnene Erkenntnis begründen, dass vor allem Träger der Wechselwirkung L_3 hohe LS besitzen, die entsprechend bevorzugt durch diese Gewichtung selektiert werden.

Neben der zufälligen Auswahl proportional zu den LS kann nun wiederholt die fes-

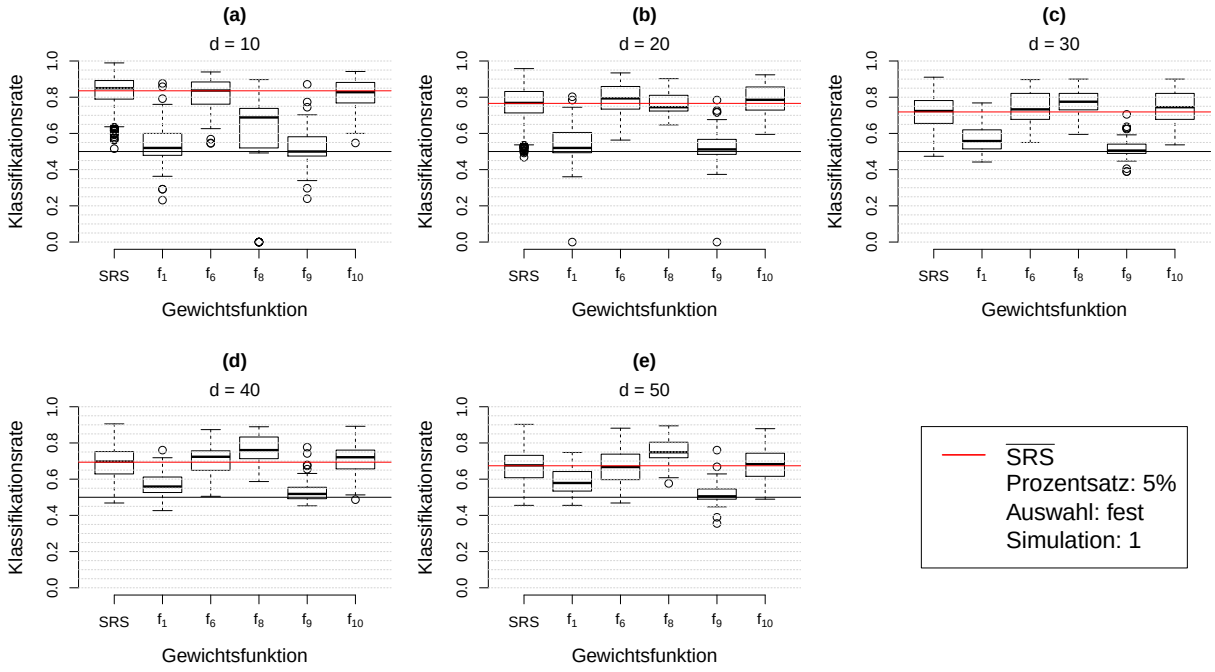

Abbildung 16: Boxplots der Klassifikationsraten beim Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 5% der Daten aus Simulation 1, bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

te Auswahl der Beobachtungen anhand der LS durchgeführt werden. Die Boxplots der sich ergebenden Klassifikationsraten für 5% der Originaldaten sind in **Abbildung 16** dargestellt. Bei der festen Auswahl führen viele Gewichtsfunktionen dazu, dass dieselben Beobachtungen in die Stichprobe aufgenommen werden. So gelangen etwa durch f_3 und f_8 die Beobachtungen mit den niedrigsten LS in die Stichprobe, durch f_1, f_4 und f_5 die Beobachtungen mit den allerniedrigsten bzw. allerhöchsten LS usw. Aufgrund dessen wird auf eine zusätzliche Darstellung der Klassifikationsraten dieser Gewichtungen verzichtet. Es gibt zwar teilweise leichte Schwankungen in den Klassifikationsraten bei den Gewichtsfunktionen die fest dieselben Beobachtung selektieren, dies liegt jedoch daran, dass der logicFS-Ansatz selbst zufallsbasiert ist.

Sehr auffällig ist, dass bei $d = 10$ drei der fünf Gewichtsfunktionen deutlich unter dem mittleren Niveau der SRS liegen. Nur die Gewichtung der mittleren LS durch die Gewichtsfunktionen f_6 und f_{10} liegt in etwa auf diesem mittleren Niveau. Bei einer niedrigen Anzahl Variablen und somit bei wenig überflüssigen Informationen in den Daten scheint weder bei der zufälligen Auswahl proportional zu den LS, noch bei der festen Auswahl der Beobachtungen eine Verbesserung der Klassifikationsraten durch die LS zu entstehen. Dies ändert sich jedoch deutlich, je mehr Variablen sich im Datensatz befinden. Ab $d = 30$ liegt die Klassifikationsrate bei der festen Auswahl der Beobachtungen mit niedrigen LS durch f_8 zu einem Großteil über dem mittleren Niveau der SRS. Die Klassifikation durch die Wahl der mittleren LS liegt im arithmetischen Mittel ab $d = 30$ etwa 5 bis 6 Prozentpunkte über dem mittleren Niveau der SRS. Dies bedeutet, dass durchschnittlich

etwa 20 Personen zusätzlich häufiger richtig klassifiziert werden. Die Wahl der Beobachtung anhand der mittleren LS verhält sich bei den Klassifikationsraten sehr ähnlich zu der SRS, wobei die Varianz etwas geringer ist. Dabei ist zu beachten, dass für die feste Auswahl jeweils 100 Klassifikationsraten für jede Datensituation vorliegen, für die SRS entsprechend 2000. Das mittlere Niveau ist für diese Gewichtsfunktionen bei einer geringen Anzahl von Variablen nahezu identisch mit dem mittleren Niveau durch die zufällige Auswahl proportional zu den LS. Bemerkenswert ist weiter, dass die Gewichtsfunktionen, welche die hohen LS mitberücksichtigen, durchgehend deutlich schlechter abschneiden. Dies zeigt sich vor allem bei der Gewichtsfunktion f_9. Offensichtlich sind die hohen LS bei diesem Prozentanteil ungeeignet dazu, die Beobachtungen fest zu selektieren. Da die Gewichtsfunktion f_1 beide Randbereiche mitberücksichtigen und besser abschneidet als f_9, kann davon ausgegangen werden, dass die hohen LS für die schlechten Klassifikationsraten verantwortlich sind.

Grundsätzlich lässt sich wie bei dem logicFS-Datensatz erkennen, dass die Klassifikationsraten mit steigendem Prozentanteil zunehmend besser und die Varianzen kleiner werden. Interessant dabei ist, dass die zufällige Auswahl bei jedem Prozentanteil und für jede Gewichtung deutliche Ausreißer nach unten aufweist. Verglichen damit weisen die Klassifikationen durch die feste Auswahl deutlich weniger extreme Ausreißer auf. Dabei ist jedoch erneut zu berücksichtigen, dass für die zufällige Auswahl pro Prozentsatz und Gewichtsfunktion jeweils 2000 Klassifikationsraten vorliegen, für die feste Auswahl jeweils nur 100. Der Median der SRS liegt immer über dem arithmetischen Mittel, was mit den starken Ausreißern nach unten auf eine Linksschiefe der Klassifikationsraten hinweist.

Die Gewichtung der Randbereiche durch f_1 schneidet bei der festen Auswahl in den Klassifikationsraten in erster Linie dann schlecht ab, wenn entweder wenige Prozent der Originaldaten verwendet werden bzw. wenn wenig Variablen im Datensatz sind. Bei höheren Prozentanteilen und mehr Variablen im Datensatz verbessern sich die Klassifikationsraten jedoch deutlich. Bei einem Anteil von 10% der Daten liegt sowohl die zufällige Auswahl proportional zu den LS als auch die feste Auswahl der Beobachtungen konstant unter dem mittleren Niveau der SRS. Dabei wird der Abstand kleiner, je mehr Variablen sich im Datensatz befinden. Auffällig ist der Unterschied zwischen der zufälligen Auswahl und der festen Auswahl. Die zufällige Auswahl ist im mittleren Niveau nahezu gleichauf mit der SRS, wobei es stärkere Ausreißer nach unten gibt. Die feste Auswahl liegt im arithmetischen Mittel teilweise 15 Prozentpunkte unter dem mittleren Niveau der SRS. Dies bessert sich erst wieder mit mehr Variablen in den Datensätzen. In **Abbildung A3** und **Abbildung A4** im Anhang sind die Boxplots der Klassifikationsraten beim Verwenden von 15% der Originaldaten dargestellt, in **Abbildung A3** mit der zufälligen Auswahl proportional zu den LS und in **Abbildung A4** mit der festen Auswahl der Beobachtungen anhand der LS. Grafiken im Anhang sind mit einem anführenden **A** gekennzeichnet. Bei dieser Darstellungen der Boxplots ist angegeben, von welcher Simulation die Daten stammen, ob die Beobachtungen zufällig oder fest anhand der LS gezogen werden und

welcher Prozentanteil der Originaldaten verwendet wird. In **Abbildung A3** und **Abbildung A4** ist zu erkennen, dass erstens bei $d = 10$ Variablen und der festen Auswahl der Beobachtungen die Klassifikationsraten sehr schlecht im Vergleich zu der SRS sind und diese eine sehr hohe Varianz besitzen. Gleichzeitig ist in **Abbildung A4 (a)** zu erkennen, dass die Klassifikationsraten durch die Gewichtung f_8 insgesamt am schlechtesten sind. Dies deutet darauf hin, dass die niedrigen LS für diese Datensituation und diesen Prozentanteil ungeeignet für die Wahl der Beobachtungen sind und dies erklärt vermutlich das schlechte Abschneiden bei der Berücksichtigung beider Randbereiche. Für keinen der gewählten Prozentsätze hebt sich die Gewichtung der LS durch f_1 von der SRS ab. Die Unterschiede im mittleren Niveau liegen bei etwa 1 Prozentpunkt Unterschied in den Klassifikationsraten. Alleine bei 15% der Originaldaten liefert die Gewichtung von f_1 ab $d = 30$ bei der zufälligen Auswahl der Beobachtungen proportional zu den LS, im arithmetischen Mittel die besten Klassifikationsraten (vgl. **Abbildung A3**). Ein Zugewinn lässt sich bei großen Stichprobenumfängen und vielen Variablen bei der Varianz verzeichnen. Dies gilt besonders bei der festen Auswahl. In **Abbildung A8 (e)** lässt sich etwa erkennen, dass es bei der festen Auswahl deutlich weniger Ausreißer nach unten gibt als bei der SRS. Hier liegt die Standardabweichung der Klassifikationsraten bei etwa 0,5 Prozentpunkten unter der Standardabweichung der SRS. Anders als bei den logicFS-Daten lässt sich durch die Gewichtung der LS mit f_1 keine Klassifikationsrate von fast 100% erreichen und die Varianz wird auch nicht stabil.

Die Gewichtung der niedrigen LS durch f_8 erweist sich bei einem Prozentanteil von 5% als am erfolgreichsten im Hinblick auf das arithmetische Mittel der Klassifikationsraten. Wie bereits angemerkt und in **Abbildung A4 (a)** zu erkennen, ändert sich dies mit steigenden Stichprobenumfängen. Dies ist schon bei den logicFS-Daten zu beobachten. Weiterhin ist das mittlere Niveau ab $d = 20$ relativ konstant und nicht weiter von der Anzahl der Variablen abhängig. Die feste Auswahl schneidet bei allen steigenden Prozentanteilen am schlechtesten ab und insgesamt ist der Zugewinn in den Klassifikationsraten deutlich geringer als bei den anderen Gewichtsfunktionen. Es scheint somit in dieser Datensituation für geringe Stichprobenumfänge von Vorteil zu sein, nur niedrige LS zu berücksichtigen, mit steigendem Stichprobenumfängen wird dies jedoch zum Nachteil. Interessant ist der Kontrast bei der zufälligen Auswahl der Beobachtungen proportional zu den LS zwischen f_3 und f_8. Da die Gewichtsfunktion f_8 nur Werte aus dem Randbereich berücksichtigt und f_3 auch andere Werte, führt dies zu einem Unterschied in den Klassifikationsraten bei der zufälligen Auswahl und steigt mit größeren Stichprobenumfängen immer deutlicher an (vgl. etwa **Abbildung A7**). Dies verstärkt erneut die Annahme, dass die Wahl von Beobachtungen nur mit niedrigen LS bei steigenden Stichprobenumfängen nachteilhaft wird.

Konträr dazu verhält sich die Gewichtung der hohen LS durch die Gewichtsfunktion f_9. Bei niedrigen Stichprobenumfängen ist die feste Auswahl anhand der hohen LS besonders schlecht und das mittlere Niveau liegt deutlich unter der SRS, was sich mit steigender

Anzahl Variablen sogar noch verdeutlicht. Mit steigenden Stichprobenumfängen ändert sich dies jedoch. Bei Verwendung von 15% der Originaldaten liegt das mittlere Niveau bei der zufälligen Auswahl der Beobachtungen proportional zu den LS bei der Gewichtung durch f_9 konstant über dem mittleren Niveau der SRS, wobei ein größeres d dies weiter begünstigt. Der Unterschied im mittleren Niveau liegt bei etwa 1 bis 1,4 Prozentpunkten. Weiter ist das mittlere Niveau auch konstant höher als das mittlere Niveau der SRS. Im Vergleich dazu schwankt bei den Gewichtungen durch f_1, f_4 und f_5 das mittlere Niveau. Die Verbesserung der Klassifikationsraten durch die Gewichtung mit f_9 bei größeren Stichprobenumfängen verstärkt den Verdacht, dass bei kleinen Stichprobenumfängen in erster Linie Träger von L_3 in die Stichprobe selektiert werden, wohingegen bei steigenden Stichprobenumfängen auch vermehrt Träger der anderen beiden Wechselwirkungen aufgenommen werden. Jedoch ist die sehr hohe Varianz in den Klassifikationsraten zu berücksichtigen. Zudem gibt es teilweise extrem viele Ausreißer nach unten, vor allem bei der zufälligen Auswahl proportional zu den LS (vgl. etwa **Abbildung A5**). Demgegenüber ist die feste Auswahl deutlich stabiler. Dabei ist jedoch erneut die geringere Anzahl der Klassifikationsraten bei dieser Auswahl zu berücksichtigen. Die besten Ergebnisse werden bei einem Prozentanteil von 25% und einer festen Auswahl der Beobachtung durch diese Gewichtung erzielt. Dies ist in **Abbildung A8** zu erkennen. Dabei ist das mittlere Niveau bereits ab $d = 20$ über dem mittleren Niveau der SRS und die Standardabweichung der Klassifikationsraten ist in etwa halb so groß, wobei es sich um etwa ein Prozentpunkt Unterschied handelt. In dieser Datensituation ist es anscheinend von Vorteil, bei wenigen zur Verfügung stehenden Daten die niedrigen LS zu berücksichtigen und bei mehr zur Verfügung stehenden Daten die hohen LS. Im Vergleich dazu bringt es keine große Verbesserung, beide Bereiche zu berücksichtigen.

Die Gewichtung der mittleren LS durch f_6, f_{10} und f_{11} hebt sich bei der zufälligen Auswahl proportional zu den LS bei keinem Stichprobenumfang von der SRS ab. Die restriktive Gewichtung durch f_6 führt eher dazu, dass die Klassifikationsraten bei steigendem Stichprobenumfang im mittleren Niveau hinter den anderen beiden Gewichtungen liegen. Ähnlich zu dem Argument bei den logicFS-Daten lässt sich vermuten, dass die restriktive Gewichtung dazu führt, dass immer dieselben Beobachtungen mit einer sehr hohen Wahrscheinlichkeit in die Stichprobe gelangen und diese nicht unbedingt zum Anpassen der Modelle geeignet sind. Nicht so restriktiv gegen die Randbereiche zu sein, erscheint erfolgversprechender. Dies deckt sich mit dem Ergebnis, dass die Konzentration auf die hohen bzw. niedrigen LS bei unterschiedlichen Stichprobenumfängen besser geeignet ist. Ansonsten ist die zufällige Auswahl anhand dieser Gewichtsfunktionen nahezu identisch mit der SRS. Weder im mittleren Niveau (außer bei f_6) noch bei der Varianz gibt es merkliche Unterschiede. Aus diesem Grund kann davon ausgegangen werden, dass es in dieser Datensituation keinen Unterschied macht, die Beobachtungen komplett zufällig zu ziehen oder proportional zu den mittleren LS. Bei der festen Auswahl der Beobachtungen anhand der mittleren LS verschlechtern sich die Klassifikationsraten bei steigendem

Prozentanteil und einer geringen Anzahl Variablen d (siehe etwa **Abbildung A6 (a), (b)** und **(c)**). Die feste Auswahl liegt zudem bei jedem Prozentanteil unter dem mittleren Niveau der SRS.

Die Vereinigung der drei Bereiche der LS bildet die Gewichtsfunktion f_7. Bei der zufälligen Auswahl der Beobachtungen proportional zu den LS liegt das mittlere Niveau ab einem Prozentanteil von 15% der Originaldaten leicht über dem mittleren Niveau der SRS. Die Diskrepanz wird mit steigender Anzahl Variablen d größer. Es handelt sich jedoch nur um etwa 0,5 Prozentpunkte Zugewinn in den Klassifikationsraten. Insgesamt hebt sich diese Gewichtung in keiner Situation stark von den anderen Gewichtungen ab und in keiner Situation ist das mittlere Niveau besser als bei einer anderen Gewichtung. Alle drei Bereiche der LS auf diese Weise zu berücksichtigen, führt somit nicht zu Verbesserungen in den Klassifikationsraten.

Als Referenz steht die Gewichtsfunktion f_2, mit der die LS direkt zur Auswahl der Beobachtungen genutzt werden. Dies korrespondiert dazu, dass Beobachtungen mit hohen LS eine erhöhte Wahrscheinlichkeit besitzen, in die Stichprobe aufgenommen zu werden. Bei der zufälligen Auswahl der Beobachtungen proportional zu f_2 sieht es im Vergleich zu f_9 jedoch bei den Klassifikationsraten anders aus. Grade bei den Prozentsätzen bei denen f_9 besonders schlecht abschneidet, gibt es im mittleren Niveau fast keinen Unterschied zwischen der zufälligen Wahl proportional zu f_2 und der SRS. Interessanterweise liegt die mittlere Klassifikationsrate vor allem bei $d = 10$ über dem mittleren Niveau der SRS. Dies gilt ab einem Prozentanteil von 10% der Originaldaten. Bei 10% der Originaldaten liegt das mittlere Niveau für $d = 10$, $d = 20$ und $d = 30$ über dem mittleren Niveau der SRS und am besten unter allen anderen Gewichtungen (vgl. **Abbildung A1**). Das mittlere Niveau liegt bei höchsten 1,19 Prozentpunkten über dem mittleren Niveau der SRS. Dies entspricht etwa 5 Beobachtungen die im Mittel zusätzlich häufiger richtig klassifiziert werden.

In **Tabelle 4** sind sowohl die Differenzen im mittleren Niveau $\Delta\overline{SRS}_P$ der Klassifikationsraten zwischen der SRS und der zufälligen Auswahl proportional zu den LS in Prozentpunkten angegeben, als auch die Differenzen im mittleren Niveau der absoluten Klassifikationen $\Delta\overline{K}_P = \Delta\overline{SRS}_P \cdot (n - n')$ zwischen der SRS und der zufälligen Auswahl proportional zu den LS abgetragen. Ein positiver Wert bedeutet, dass die Klassifikationsrate bzw. die absolute Anzahl der Klassifikationen durch die zufällige Auswahl der Beobachtungen proportional zu den LS höher liegt. Da die Klassifikationsraten insgesamt von der Datensituation abhängen und sich den Boxplots entnehmen lassen, ist es von einem größeren Interesse den Unterschied zwischen der Auswahl der Beobachtungen proportional zu den LS und per SRS zu betrachten. In **Tabelle 4** wird sich auf die Auswahl der Gewichtsfunktionen f_1, f_2, f_8 und f_9 beschränkt, da diese sich in gewissen Datensituationen hervorheben. Die erste Spalte enthält den Prozentanteil P, die zweite Spalte die Anzahl der Variablen d. In den restlichen Spalten sind für die entsprechende Gewichtsfunktion die Differenzen im mittleren Niveau der Klassifikationsraten und die

Tabelle 4: Differenzen im mittleren Niveau der Klassifikationsraten und mittlere Anzahl Beobachtungen die zusätzlich häufiger richtig klassifiziert werden bei der zufälligen Wahl der Beobachtungen anhand der einfachen Zufallsauswahl und proportional gewichtet nach den Leverage Scores mit den Daten der Simulation 1. Die Spalten mit $\Delta\overline{SRS}_P$ enthalten die Differenzen im mittleren Niveau der Klassifikationsraten in Prozentpunkten und die Spalten mit $\Delta\overline{K}_P$ die Differenzen der im Mittel häufiger richtig klassifizierten Beobachtungen. Hervorgehoben ist in jeder Spalte jeweils der größte positive Zugewinn.

		f_1		f_2		f_8		f_9	
P	d	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$
	10	-6,66	-26,39	-1,67	-6,60	-7,30	-28,89	-21,31	-84,39
	20	-5,82	-23,06	-2,07	-8,21	**2,74**	**10,83**	-18,74	-74,21
5	30	-5,30	-21,01	-2,12	-8,40	**5,87**	**23,26**	-16,14	-63,91
	40	-4,30	-17,04	-1,69	-6,70	**7,66**	**30,32**	-14,00	-55,44
	50	-3,75	-14,85	-1,71	-6,75	**7,72**	**30,58**	-12,75	-50,47
	10	-1,41	-5,53	**1,19**	**4,65**	-13,17	-51,61	-10,64	-41,71
	20	-0,34	-1,33	**0,64**	**2,52**	-4,00	-15,67	-8,38	-32,86
10	30	-0,24	-0,95	**0,11**	**0,45**	-1,51	-5,90	-10,82	-42,42
	40	-1,17	-4,59	-0,42	-1,64	-0,68	-2,67	-12,81	-50,20
	50	-1,69	-6,62	-0,69	-2,69	**0,23**	**0,92**	-14,87	-58,29
	10	-0,05	-0,20	**1,15**	**4,47**	-14,68	-56,95	-4,07	-15,79
	20	0,45	1,74	**1,09**	**4,21**	-6,31	-24,48	-0,12	-0,45
15	30	**1,25**	**4,84**	0,93	3,60	-5,46	-21,18	-0,37	-1,45
	40	**1,13**	**4,37**	0,65	2,53	-5,04	-19,57	0,11	0,42
	50	**0,90**	**3,50**	0,48	1,85	-4,96	-19,26	-0,92	-3,57
	10	0,22	0,83	**0,68**	**2,62**	-15,48	-59,45	-0,93	-3,57
	20	0,12	0,47	0,83	3,18	-7,39	-28,36	**1,32**	**5,08**
20	30	0,44	1,68	0,86	3,29	-6,75	-25,93	**0,99**	**3,82**
	40	0,72	2,77	0,70	2,68	-6,69	-25,69	**1,41**	**5,40**
	50	0,86	3,29	0,74	2,85	-6,77	-26,01	**1,55**	**5,94**
	10	0,15	0,55	**0,53**	**2,02**	-15,58	-59,21	0,27	1,01
	20	-0,04	-0,14	0,44	1,67	-8,11	-30,80	**1,29**	**4,91**
25	30	-0,10	-0,40	0,47	1,79	-7,53	-28,61	**1,00**	**3,82**
	40	0,21	0,79	0,54	2,05	-7,34	-27,89	**1,19**	**4,54**
	50	0,34	1,29	0,35	1,32	-7,51	-28,54	**1,32**	**5,03**

Differenz der absoluten Klassifikationen abgetragen, in Abhängigkeit von P und d. Der jeweils größte positive Zugewinn ist in jeder Zeile besonders hervorgehoben. Den stärksten Zugewinn zwischen der SRS und der Gewichtung der LS gibt es bei einem Prozentanteil von 5% der Originaldaten und der Gewichtung durch f_8. Dies nimmt mit einer steigenden Anzahl an Variablen zu. Auffällig ist das Abfallen in den Differenzen ab 10% der Originaldaten. Zwar gibt es fast immer eine Gewichtung der LS, die im mittleren Niveau über der SRS liegt, jedoch bewegen sich die Zugewinne im Bereich von etwa einem Prozentpunkt. Die zweite Gewichtung, die im mittleren Niveau auffällig besser als die SRS liegt, ist die Gewichtung durch f_9 ab einem Prozentanteil 20% der Originaldaten. Das mittlere Niveau

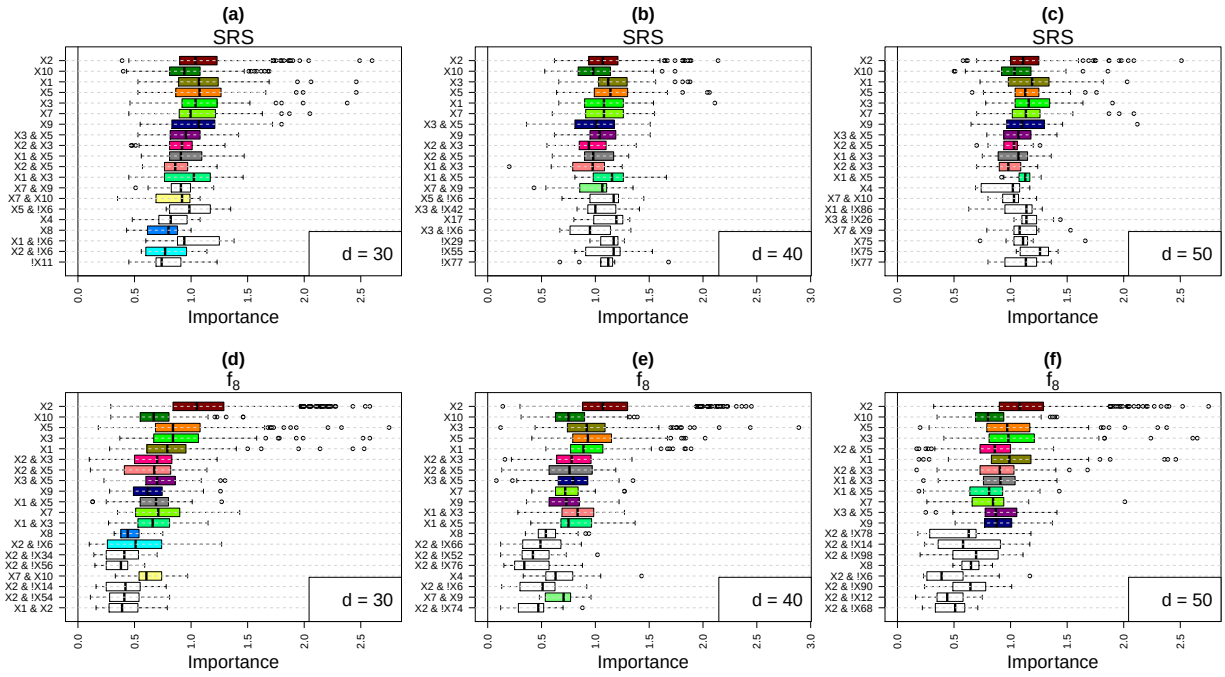

Abbildung 17: Boxplots der Wichtigkeitsmaße der 20 am häufigsten gefundenen Wechselwirkungen bei unterschiedlicher Auswahl einer Stichprobe bestehend aus 5% des Originaldatensatzes der Daten aus Simulation 1 für $d = 30, d = 40$ und $d = 50$, sortiert danach wie häufig diese gefunden werden. Eingefärbt sind die Überschneidungen der Wechselwirkungen zwischen der einfachen Zufallsauswahl und der Auswahl proportional zu den Leverage Scores gewichtet durch f_8.

liegt in diesem Fall bei über einem Prozentpunkt höher als bei der SRS und dies führt dazu, dass durchschnittlich etwa 4 Beobachtungen zusätzlich häufiger richtig klassifiziert werden. Am stärksten heben sich somit in Simulation 1 die Gewichtung durch f_8 bei 5% der Originaldaten und die Gewichtung durch f_9 bei 20% und 25% der Originaldaten ab. Darum werden für diese Gewichtungen der LS noch die gefundenen Wechselwirkungen genauer betrachtet.

In **Abbildung 17** sind Boxplots der Wichtigkeitsmaße aus **Gleichung (1)** der 20 am häufigsten gefundenen Wechselwirkungen durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf jeweils 5% der Originaldaten dargestellt, für die Datensätze mit $d = 30, d = 40$ und $d = 50$, beim zufälligen Ziehen der Beobachtungen per SRS und proportional zu den LS mit der Gewichtung durch f_8. Sortiert sind die Wechselwirkungen nach der Häufigkeit in denen diese durch die Modelle gefunden werden, von oben nach unten. Für jedes d sind die Wechselwirkungen der SRS und der Gewichtung mit f_8 übereinander dargestellt. Eingefärbt sind die Boxplots der Wechselwirkungen danach, ob diese sich zwischen den Auswahlmethoden überschneiden. Grundsätzlich sind die gefundenen Wechselwirkungen simpel und bestehen aus höchstens zwei Binärvariablen. Weiter ist anzumerken, dass die durch das Anpassen der Modelle auf den gesamten Datensätzen am häufigsten gefundene Binärvariablen X_{11}^{C} und X_{12} sich nicht unter den

15 am häufigsten gefundenen Wechselwirkungen befinden. Je nach Datensituation sind diese Binärvariablen nicht einmal unter den 20 am häufigsten gefundenen Wechselwirkungen vertreten. Weiter ist festzuhalten, dass es sehr viele Überschneidungen bei den Wechselwirkungen gibt, die mit steigender Anzahl Variablen abnehmen. Bei der zufälligen Auswahl der Beobachtungen proportional zu f_8 befindet sich in jeder Wechselwirkung wenigstens eine Binärvariable, die nach Konstruktion einen Einfluss besitzt. Die Ausnahme dazu bildet die Binärvariable X_8, die bei allen drei Datensituationen unter den 20 am häufigsten gefundenen Wechselwirkungen vertreten ist.

Bei der SRS sind ab $d = 40$ gleich mehrere Binärvariablen selektiert, die nach Konstruktion keinen Einfluss auf den Krankheitsstatus besitzen. Die beiden Binärvariablen X_2 und X_{10} werden durch jede Gewichtung am häufigsten gefunden und ähnlich stark gewertet. Der Grund für das bessere Abschneiden der Gewichtung durch f_8 muss somit durch andere Faktoren begründet sein. Bei der Auswahl der Beobachtungen durch f_8 beziehen sich die meisten der gefundenen Wechselwirkungen auf L_1. Diese ist erstens die komplexeste der drei Wechselwirkungen und zweitens bilden die Träger dieser Wechselwirkung die größte Gruppe unter der Gruppe der erkrankten Personen (vgl. **Tabelle 1**). Es ergibt Sinn, dass durch diese Wechselwirkungen viele Personen aus dieser Gruppe richtig klassifiziert werden und somit die Klassifikationsrate höher liegt. Aus der Gruppe der Erkrankten mit der Wechselwirkung L_2 werden in etwa dieselben Wechselwirkungen gefunden. Die Binärvariablen X_{11} und X_{12} aus L_3 sind bei der Gewichtung durch f_8 gar nicht unter den 20 am häufigsten gefundenen Wechselwirkungen vertreten. Dies dürfte der Grund dafür sein, dass die Klassifikationsraten bei steigender Anzahl Variablen nicht bei über 90% liegen (vgl. etwa **Abbildung 15 (d)** und **(e)**), da diese Gruppe der Erkrankten nicht erfasst wird. Dies deckt sich mit der Erkenntnis, dass die Träger von L_3 vermehrt hohe LS besitzen und damit durch die Wahl der Beobachtungen nach niedrigen LS nicht in die Stichprobe aufgenommen werden. Zusätzlich handelt es sich bei dieser Gruppe um die am schwächsten besetzte. Dies reduziert die Möglichkeit, dass dieser Einfluss bei einem geringen Stichprobenumfang erfolgreich erfasst wird. Interessant ist, dass die Gewichtung der beiden Randbereiche durch die Gewichtsfunktionen f_1, f_4 und f_5 nicht besser abschneidet, da diese die beiden Randbereiche und somit die Träger von mehr Wechselwirkungen berücksichtigen. Jedoch ist vermutlich dafür der Stichprobenumfang zu gering.

Die feste Auswahl der Beobachtung anhand der LS führt hingegen zu deutlich anderen Wechselwirkungen. Symptomatisch für die Wechselwirkungen der festen Auswahl ist es, dass wichtige binäre Einflussvariablen wie etwa X_2 in Kombination mit anderen Binärvariablen gefunden werden, die nach Konstruktion der Daten keinen Einfluss auf den Krankheitsstatus besitzen. Bei dem Fall $d = 30$ befinden sich unter den 20 am häufigsten gefundenen Wechselwirkungen durch die Gewichtsfunktion f_8, 13 solcher Wechselwirkungen mit der Binärvariable X_2. Dadurch lässt sich Kritik an der festen Auswahl üben. Diese führt zu einer Auswahlverzerrung in dem Sinne, dass immer die Beobachtungen mit den entsprechend niedrigsten (bzw. von der Gewichtsfunktion abhängig extremste) LS in die

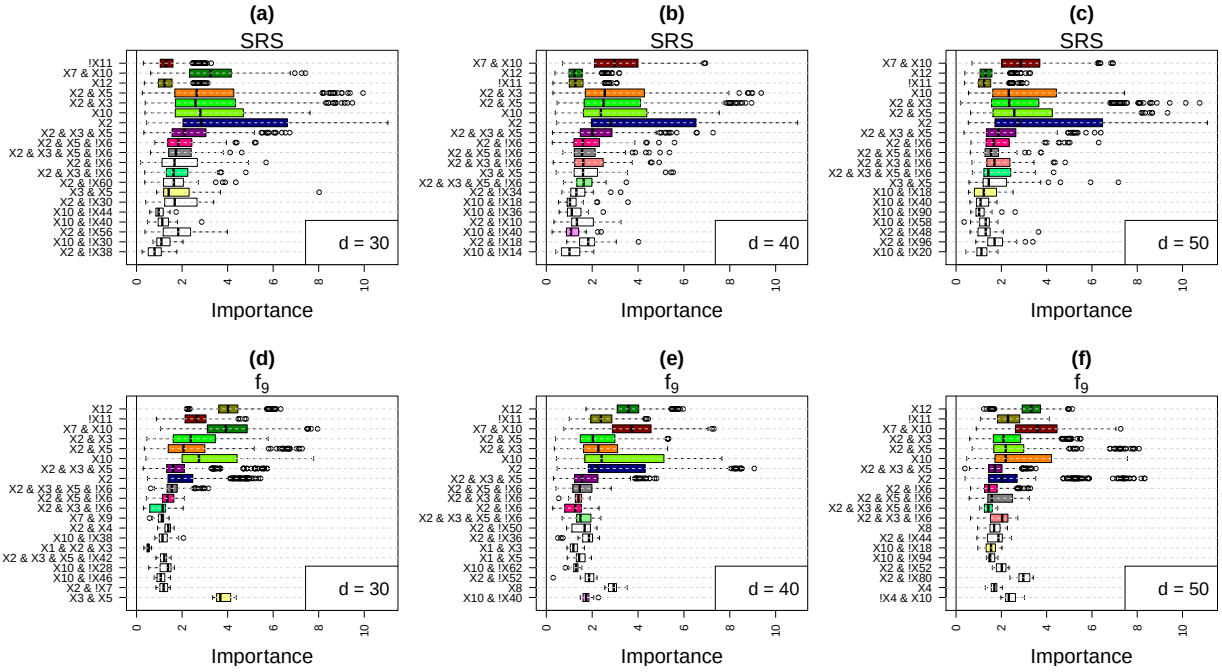

Abbildung 18: Boxplots der Wichtigkeitsmaße der 20 am häufigsten gefundenen Wechselwirkungen bei unterschiedlicher Auswahl einer Stichprobe bestehend aus 25% des Originaldatensatzes der Daten aus Simulation 1 für $d = 30, d = 40$ und $d = 50$, sortiert danach wie häufig diese gefunden werden. Eingefärbt sind die Überschneidungen der Wechselwirkungen zwischen der einfachen Zufallsauswahl und der Auswahl proportional zu den Leverage Scores gewichtet durch f_9.

Stichprobe aufgenommen werden und deren Merkmale entscheidend für das Finden der Wechselwirkungen sind. Dies könnte als Grund dafür dienen, dass die Binärvariable X_2 im Zusammenhang mit so vielen verschiedenen anderen nicht wichtigen Binärvariablen gefunden wird. Dies ändert sich auch bei einer größeren Anzahl Variablen d nur wenig. Für $d = 40$ und $d = 50$ sind immer noch 10 der 20 am häufigsten gefundenen Wechselwirkungen von dieser Art. Bei den restlichen 10 Wechselwirkungen handelt es sich immerhin um andere nach Konstruktion wichtige Binärvariablen und es gelangen keine völlig überflüssigen Binärvariablen in die Modelle. Dies kann als Erklärung dafür dienen, dass die Klassifikationsraten vergleichbar sind mit denen der zufälligen Auswahl proportional zu den LS.

In **Abbildung 18** sind (wie in **Abbildung 17**) die Wichtigkeitsmaße der gefundenen Wechselwirkungen durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf jeweils 25% der Daten dargestellt, nun mit der zufälligen Auswahl der Beobachtung gewichtet durch f_9. Durch den größeren Stichprobenumfang werden die Wechselwirkungen komplexer. Dabei fällt vor allem bei einer höheren Anzahl Variablen d im Vergleich zu **Abbildung 13** auf, dass die Wechselwirkungen komplexer und näher an den DNF der wahren Wechselwirkungen liegen (siehe **Gleichung (6)** bis **(8)**). Anders als bei dem Prozentsatz von 5% befinden sich bei den Wechselwirkungen, die durch die SRS

gefunden werden, keine einzelnen Binärvariablen, die nach Konstruktion keinen Einfluss haben. Auffällig ist es, dass die beiden Binärvariablen X_{11}^{C} und X_{12} durch die Gewichtung mit f_9 am häufigsten gefunden werden und zudem noch ein relativ hohes Wichtigkeitsmaß erhalten. Zwar sind diese beiden Binärvariablen beim Anpassen der Modelle auf den ganzen Datensatz ebenfalls am häufigsten vertreten (siehe **Abbildung 13**), jedoch besitzen andere Wechselwirkungen eine deutlich höhere Wichtigkeit. Erneut zeigt sich darin, dass Träger von L_3 höhere LS besitzen und dadurch bevorzugt durch f_9 in die Stichprobe gelangen. Dies ist jedoch nicht automatisch etwas Negatives. So werden zuverlässig für eine Untergruppe der Erkrankten die entsprechenden Binärvariablen gefunden und als wichtig bewertet. Wäre diese Gruppe in den Daten stärker vertreten, würde dies potentiell zu einer besseren Klassifikationsrate führen. Die Gewichtung der beiden Randbereiche durch f_8 bzw. f_9 scheint davon abzuhängen, welche Gruppe der Erkrankten stärker vertreten ist und welche LS diese Gruppen besitzen. Durch die Berücksichtigung beider Randbereiche, können auf diese Weise eventuell mehrere Untergruppen gefunden werden und dadurch die genetischen Faktoren für einen insgesamt größeren Teil der Erkrankten. Bei der Auswahl durch die SRS gibt es große Schwankungen in den Wechselwirkungen, die am häufigsten gefunden werden. Zwar handelt es sich bei den am häufigsten gefundenen Wechselwirkungen um dieselben, die Häufigkeit in der diese gefunden werden unterscheidet sich jedoch teilweise sehr. Durch die SRS werden vermehrt Wechselwirkungen der Träger von L_1 gefunden. Dies ist nicht verwunderlich, da diese die größte Gruppe unter den Erkrankten stellen. Bei der Auswahl durch f_9 ist dies nicht so eindeutig der Fall und es werden viele Wechselwirkungen der Träger von L_1 und L_2 gefunden, wobei für L_2 vermehrt die Binärvariable X_{10} vertreten ist.

Es lässt sich durch die gefundenen Wechselwirkungen argumentieren, dass es generell von Vorteil sein kann, die Daten zu reduzieren, anstatt den gesamten Datensatz zu analysieren. Zwar ist das Wichtigkeitsmaß der gefundenen Wechselwirkungen geringer, dies liegt jedoch daran, dass sich weniger Beobachtungen in der Stichprobe befinden. Da das Wichtigkeitsmaß auf der Anzahl der richtig klassifizierten Beobachtungen basiert, sind die Werte bei weniger Beobachtungen geringer. Weiter werden für mindestens eine der Gruppen der Erkrankten die Binärvariablen sehr zuverlässig gefunden, die Wechselwirkungen sind komplexer und näher an den DNF der wahren Einflüsse und zusätzlich ist die Varianz der Wichtigkeitsmaße geringer. Dabei sollte berücksichtigt werden, dass die Anpassung des Modells auf den vollen Daten jeweils nur einmal durchgeführt wird, die Anpassung durch die zufällige Wahl der Beobachtungen proportional zu den LS jeweils 20-mal pro Datensatz. Eventuell ist es vorteilhafter, öfter auf Teilmengen des Datensatzes die Anpassung durchzuführen und die gefundenen Wechselwirkungen der Modelle, wie hier geschehen, zusammenzulegen. Der logicFS-Ansatz basiert selbst auf dieser Vorgehensweise.

Bei der festen Auswahl von 25% der Beobachtungen anhand der LS gewichtet durch f_9 ergeben sich ähnlich wie bei f_8 deutlich andere Wechselwirkungen. Sehr auffällig dabei ist, dass es bei den gefundenen Wechselwirkungen deutlich Überschneidungen zwischen

den einzelnen Einflüssen gibt. Viele der 20 am häufigsten gefundenen Wechselwirkungen enthalten Binärvariablen, die zu zwei der DNF der Wechselwirkungen gehören. Dazu sind die Wechselwirkungen deutlich simpler und enthalten eine oder zwei Binärvariablen. Alleine die Dreifachwechselwirkung ($X_2 \wedge X_3 \wedge X_5$) findet sich bei allen Datensituationen unter den 20 am häufigsten gefundenen Wechselwirkungen. Die hohen Klassifikationsraten widersprechen der Vermutung, dass Überschneidungen zwischen den Wechselwirkungen dazu führen, dass das Wichtigkeitsmaß negativ wird und dürfte dies für zumindest simple Wechselwirkungen widerlegen. Falsche Wechselwirkungen einer höheren Ordnung würden für eine neue Beobachtung vermutlich zu einer falschen Klassifikation führen, woraufhin sich die Klassifikationsrate senkt. Zum Finden wichtiger Einflüsse dürfte die feste Auswahl der Beobachtungen anhand der LS somit in dieser Datensituation eher ungeeignet sein.

Insgesamt lässt sich bei der Auswertung der Simulation 1 festhalten, dass es bei steigenden Prozentsätzen der Daten vergleichsweise wenig Zugewinn im mittleren Niveau durch die Gewichtung der LS gibt. Die größten Zugewinne lassen sich wie bei dem logicFS-Datensatz bei 5% der Originaldaten und der Gewichtung durch f_8 erzielen. Im Kontrast zu den logicFS-Daten zeigt sich, dass die feste Auswahl der Beobachtungen nicht zu einer 100% Richtigklassifikation führt und dass sich in den am häufigsten gefundenen Wechselwirkungen durch die feste Auswahl Verzerrungen in dem Sinne finden, dass viele nicht wichtige Binärvariablen in Kombination mit wenigen wichtigen Binärvariablen gefunden werden. Auch ergibt sich im Vergleich zu den logicFS-Daten, dass bei größeren Prozentsätzen die Gewichtung der hohen LS am besten geeignet ist und dass sich durch die Gewichtung unterschiedlicherer Bereiche der LS potentiell unterschiedliche Untergruppen in der Gruppe der Erkrankten selektieren lassen. Darin liegt potentiell die größte Stärke, durch die Auswahl proportional zu den LS.

Im nächsten Unterabschnitt geht es darum, wie sich die Methode verhält, wenn es in den Daten Fälle gibt, deren Krankheitsstatus nicht durch die genetischen Einflüsse der SNPs erklärt sind.

5.2.3 Auswertung von Simulation 2

Angelehnt an Simulation 1 ist die Simulation 2. Die Grundvoraussetzungen sind dabei dieselben, d.h. es werden gleich viele Datensätze simuliert, mit der selben Anzahl an Beobachtungen und Variablen. Der entscheidende Unterschied ist, dass nicht alle Fälle den Krankheitsstatus durch die SNPs erklärt haben (vgl. **Kapitel 4.2**). Von Interesse ist es daher, wie sich dieses „Rauschen“ auf die Auswahl der Beobachtungen und das Anpassen der Modelle auswirkt.

Einführend werden erneut auf jedem der Datensätze logische Regressionsmodelle mit dem logicFS-Ansatz angepasst. Das konstruierte Rauschen in den Datensätzen lässt sich in dem OOB-Fehler wiederfinden. Dieser liegt in allen Datensituationen im arithmetischen Mittel bei 13%, wobei der Fehler mit steigender Anzahl Variablen d ansteigt. Dies gilt auch für den Median der OOB-Fehler, der anfangs bei 13% liegt, ab $d = 40$ bei 13,5%.

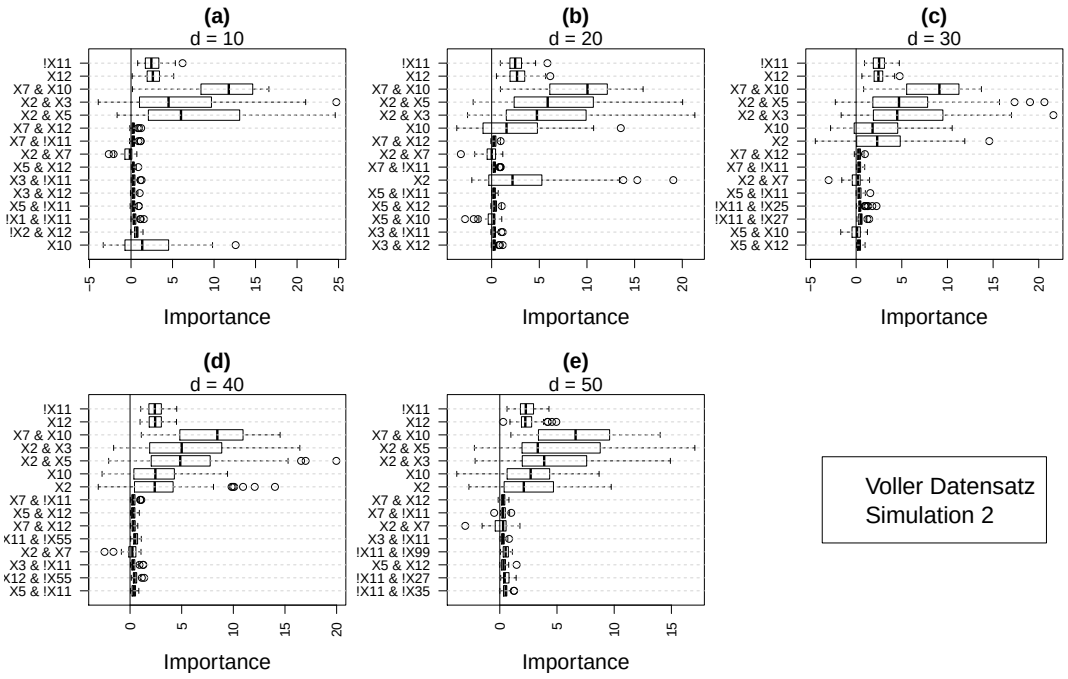

Abbildung 19: Boxplots der Wichtigkeitsmaße der gefundenen Wechselwirkungen beim Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf die vollen Datensätze der Simulation 2.

Das Minimum beträgt in allen Datensituationen 12,5%. Es kann also davon ausgegangen werden, dass sich im optimalen Fall eine Klassifikationsrate von 87,5% erreichen lässt.

In **Abbildung 19** sind Boxplots der resultierenden Wichtigkeitsmaße aus **Gleichung (1)** der 15 in allen Datensätzen am häufigsten gefundenen Wechselwirkungen dargestellt, getrennt nach der Anzahl der Variablen im Datensatz. Im Vergleich zu **Abbildung 13** ist zu erkennen, dass die Wichtigkeit der gefundenen Wechselwirkungen abnimmt. Dies liegt daran, dass das Wichtigkeitsmaß auf der Klassifikation der Beobachtungen basiert. Wenn es in dem Datensatz schwerer ist, die Daten korrekt zu klassifizieren, nimmt das Maß entsprechend ab. Weiter zeigt sich, dass die gefundenen Wechselwirkungen im Vergleich zu Simulation 1 simpler sind und höchstens aus Zweifachwechselwirkungen bestehen. Zusätzlich kommen bei steigender Anzahl Variablen im Datensatz Binärvariablen in den Wechselwirkungen vor, die nach Konstruktion keinen Einfluss auf den Krankheitsstatus besitzen. Dies legt die Vermutung nahe, dass sich diese Eigenschaften auf die Stichproben übertragen und entsprechend die Klassifikation erschweren.

In den LS der Datensätze lassen sich keine großen Unterschiede zu denen von Simulation 1 erkennen, weshalb auf eine gesonderte Darstellung verzichtet wird. Das Vorgehen wird parallel zu dem Vorgehen aus Simulation 1 auf die Daten der Simulation 2 angewandt, mit der selben Anzahl Wiederholungen pro Datensatz und den selben Einstellungen für den logicFS-Ansatz.

In **Abbildung 20** sind Boxplots der Klassifikationsraten beim wiederholten Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz auf 5% der Originaldaten darge-

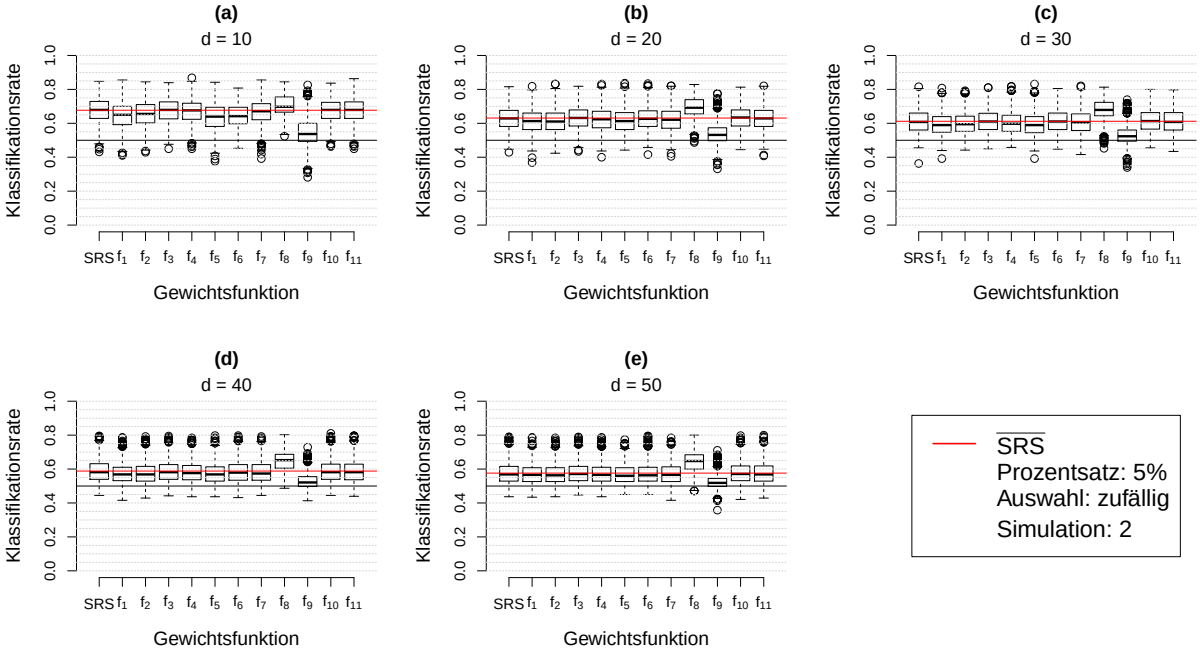

Abbildung 20: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 5% der Daten aus Simulation 2 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

stellt, wobei die Beobachtungen zufällig proportional zu den LS in die Stichprobe aufgenommen werden. Insgesamt liegt das mittlere Niveau der Klassifikationsraten bereits bei $d = 10$ deutlich niedriger als bei Simulation 1. Zusätzlich nimmt das mittlere Niveau mit steigendem d weiter ab. Analog zu Simulation 1 mit dem selben Prozentanteil ist zu erkennen, dass die Gewichtung durch f_8 zu den besten Klassifikationsraten führt und die Gewichtung durch f_9 zu den schlechtesten. Abhängig von d liegt das mittlere Niveau der Klassifikationsraten von f_8 bei mehr als 6 Prozentpunkten über dem mittleren Niveau der SRS. Dies entspricht mehr als 22 Personen die im Mittel zusätzlich häufiger richtig klassifiziert werden. Alle anderen Gewichtungen sind nahezu gleichauf mit der SRS. Die Differenz im mittleren Niveau bei diesem Prozentanteil liegt jedoch nicht auf der Höhe des OOB-Fehlers und hängen von der Gewichtung und der Anzahl der Variablen ab.

In **Tabelle 5** sind die Differenzen im mittleren Niveau der Klassifikationsraten zwischen den vergleichbaren Daten aus Simulation 1 und Simulation 2 beim zufälligen Auswählen der Beobachtungen proportional zu den LS mit 5% der Originaldaten dargestellt. Die erste Spalte enthält die Anzahl der Variablen im Datensatz, die anderen Spalten für die entsprechende Gewichtsfunktion die Differenz der Prozentpunkte im mittleren Niveau der Klassifikationsraten. Ein positiver Wert bedeutet, dass das mittlere Niveau gegenüber Simulation 1 gefallen ist. Als erstes ist anzumerken, dass das mittlere Niveau für alle Gewichtungen gefallen ist. Das Rauschen in den Daten überträgt sich somit auf die Stichprobe bzw. die Klassifikation durch die Modelle basierend auf der Stichprobe. Jedoch zeigt sich dies in unterschiedlichem Maße. Den geringsten Verlust im mittleren Niveau

Tabelle 5: Differenzen im mittleren Niveau der Klassifikationsraten beim Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz zwischen vergleichbaren Daten der Simulation 1 und Simulation 2 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores mit 5% der Originaldaten. Die Differenzen sind in Prozentpunkten angegeben.

d	SRS	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}
10	15,83	12,21	16,35	15,94	12,07	8,56	18,05	12,29	5,52	7,75	15,95	15,57
20	13,49	9,51	13,20	13,53	10,45	7,06	15,33	10,18	9,92	4,08	13,57	13,40
30	10,66	7,05	9,66	11,05	8,41	5,81	12,65	7,67	9,65	2,63	10,70	11,14
40	10,55	7,60	10,09	10,90	8,29	5,90	12,07	7,95	12,25	2,57	11,08	11,38
50	9,79	6,57	8,56	9,68	7,61	5,36	11,63	7,31	10,68	2,26	9,62	9,65

besitzt ab $d = 20$ die Gewichtsfunktion f_9. Das mittlere Niveau liegt jedoch konstant unter dem mittleren Niveau der SRS und befindet sich sehr nahe an dem Richtwert 0,5. Somit ist der geringere Verlust unbedeutend. Nahezu alle Gewichtungen besitzen den geringsten Verlust bei $d = 50$. Die Ausnahme dazu bildet die Gewichtsfunktion f_8. Dies bedeutet, dass die Gewichtung durch f_8 bei Rauschen in den Daten von weniger Variablen profitiert. Selbst der Verlust im mittleren Niveau der SRS liegt nicht auf der Höhe des OOB-Fehlers, sondern hängt von d ab. Ein Grund dafür könnte sein, dass bei dem geringen Stichprobenumfang von $n' = 20$ die Klassifikationsraten generell niedriger liegen und das Rauschen sich daher nicht im vollen Ausmaß überträgt, da generell mehr Beobachtungen falsch klassifiziert werden und in diese Gruppe auch solche Beobachtungen fallen, die den Krankheitsstatus nicht durch die SNPs erklärt haben. Dies würde erklären, warum die ohnehin schlechte Klassifikation durch f_9 am geringsten fällt.

Rauschen in den Daten bedeutet, dass die Klassifikation der Beobachtungen schwerer wird. Dies lässt die Vermutung zu, dass die Varianz der Klassifikationsraten steigt, da es zu Situationen kommen kann, in denen mehr bzw. weniger Fälle in der Stichprobe vorkommen, die ihren Krankheitsstatus durch die SNPs erklärt haben und somit die Modelle stärker schwanken. Weiter gibt es nun unter den Klassifikationsraten vermehrt Ausreißer nach oben und der Median der Klassifikationsraten liegt unter dem arithmetischen Mittel. Dies bedeutet, dass die Klassifikationsraten in dieser Situation rechtsschief sind. Bei einem Prozentsatz von 5% der Originaldaten und der zufälligen Auswahl der Beobachtungen proportional zu den LS ist dies nicht der Fall. Abhängig von der Gewichtsfunktion reduziert sich die Varianz und die Reduktion liegt bei 2 bis 3 Prozentpunkten im Vergleich zu Simulation 1 und nimmt mit steigenden d zu. Erneut bildet f_8 dazu die Ausnahme. Unabhängig von d liegt für diese Gewichtsfunktion die Reduktion der Standardabweichung bei etwa 0,8 Prozentpunkten. Bei diesem Prozentanteil der Daten scheint es somit für die Gewichtung durch f_8 relativ unbedeutend zu sein, dass es Rauschen in den Daten gibt und die Klassifikationsraten bleiben relativ stabil unter dem erwarteten Verlust von 13,5 Prozentpunkten im mittleren Niveau. Mit steigendem Prozentsatz der ausgewähl-

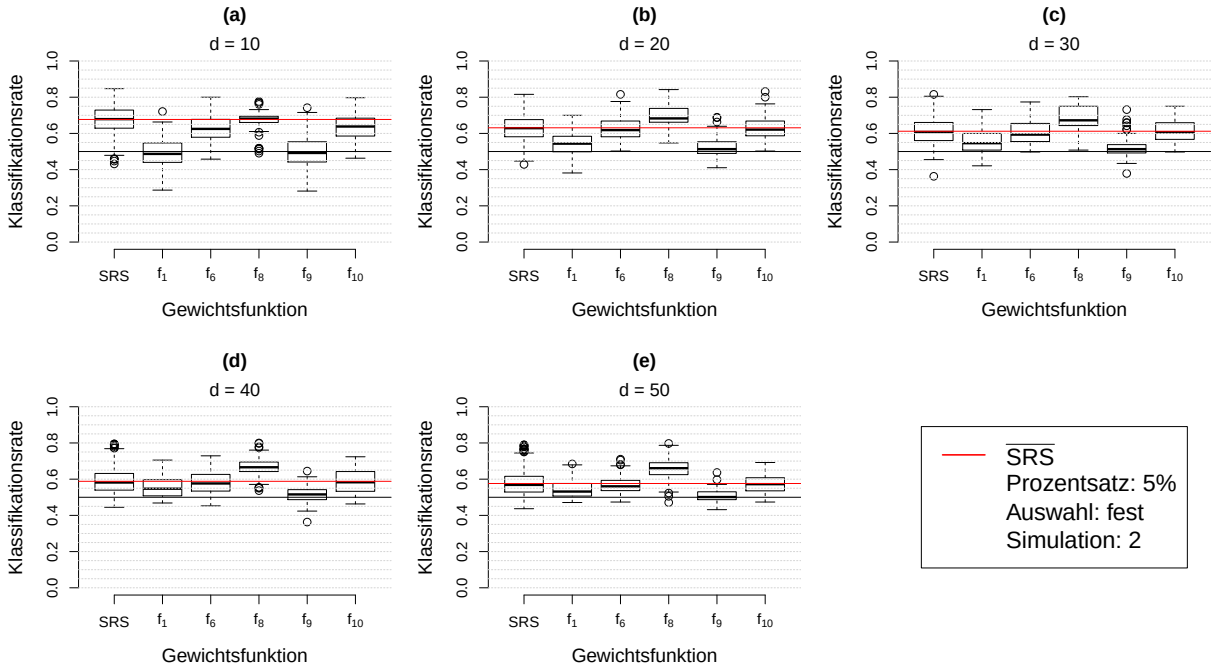

Abbildung 21: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 5% der Daten aus Simulation 2, bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

ten Daten bleibt es nicht dabei, dass die Varianz der Klassifikationsraten sinkt, sondern hängt von der Gewichtung und der Anzahl der Variablen ab. Die Reduktion der Varianz bei diesem geringen Prozentanteil liegt vermutlich ebenfalls daran, dass das Niveau der Klassifikationsraten generell niedriger liegt und es somit weniger Schwankung bei der Klassifikation gibt, da ohnehin mehr Beobachtungen falsch klassifiziert werden. Mit steigenden Prozentsätzen und einhergehendem steigenden mittleren Niveau ändert sich dies.

Bei einem Prozentsatz von 25% der Originaldaten liegt die Varianz der Klassifikationsraten bei allen Gewichtungen über den Varianzen aus Simulation 1 und die Standardabweichungen sind um 1 bis 2 Prozentpunkte höher. Die Ausnahme dazu bilden erneut die Gewichtung durch f_8, bei der die Standardabweichung der Klassifikationsraten nahezu gleich zu der aus Simulation 1 ist und die Gewichtung durch f_9 bei der die Varianz sich deutlich erhöht und die Standardabweichungen um 5 bis 6 Prozentpunkte höher liegen.

Die **Abbildung 21** enthält Boxplots der Klassifikationsraten bei fester Auswahl der Beobachtungen anhand der LS mit 5% der Originaldaten der Simulation 2. Ähnlich zu **Abbildung 16** ist zu erkennen, dass es die Gewichtungen der niedrigen LS durch f_8 bzw. die Gewichtung der mittleren LS durch f_6 und f_{10} sind, die in etwa gleichauf bzw. besser als das mittlere Niveau der SRS liegen.

Die **Tabelle 6** enthält die Differenzen im mittleren Niveau der Klassifikationsraten zwischen den Daten der Simulation 1 und der Simulation 2 bei fester Auswahl der Beobachtungen anhand der LS mit 5% der Originaldaten. Die Differenzen sind in Prozentpunkten angegeben. Ein negativer Wert bedeutet eine Verbesserung im mittleren Niveau

Tabelle 6: Differenzen im mittleren Niveau der Klassifikationsraten beim Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz zwischen vergleichbaren Daten der Simulation 1 und Simulation 2 bei fester Auswahl der Beobachtungen anhand der Leverage Scores mit 5% der Originaldaten. Die Differenzen sind in Prozentpunkten angegeben.

d	SRS	f_1	f_6	f_8	f_9	f_{10}
10	15,83	4,51	17,92	-5,26	2,48	17,25
20	13,49	0,69	16,51	6,62	0,39	15,45
30	10,66	1,23	13,60	9,06	-0,58	13,61
40	10,55	1,90	12,11	9,73	1,37	11,76
50	9,79	4,25	10,27	9,88	1,13	10,45

gegenüber der Simulation 1. Es fällt auf, dass die feste Auswahl bei dieser Datensituation bei $d = 10$ für die Gewichtung durch f_8 im mittleren Niveau besser abschneidet als in der Simulation 1 und dass der Verlust im mittleren Niveau für die Gewichtung der mittleren LS höher liegt als bei der zufälligen Auswahl proportional zu den LS. Verglichen mit **Tabelle 5** ist zu erkennen, dass für die Gewichtsfunktion f_8 das mittlere Niveau weniger gefallen ist als bei der zufälligen Auswahl proportional zu den LS. Dabei ist anzumerken, dass das mittlere Niveau der festen Auswahl bei Simulation 1 niedriger liegt als das mittlere Niveau der zufälligen Auswahl proportional zu den LS.

Besonders auffällig ist, dass die Gewichtung durch f_9 bei der zufälligen Auswahl proportional zu den LS konstant unter dem mittleren Niveau der SRS liegt und die Varianz deutlich größer ist, als bei allen anderen Gewichtungen. Auch bessert sich dies nicht bei Verwendung von 25% der Originaldaten. In Simulation 1 schneidet die Gewichtung durch f_9 bei diesem Prozentanteil am besten ab, liegt in dieser Datensituation jedoch abhängig von der Anzahl der Variablen d teilweise deutlich unter dem mittleren Niveau der SRS. Dies gilt ebenfalls für die feste Auswahl der Beobachtungen anhand der LS. Erst bei einem Prozentanteil von 25% der Originaldaten schneidet die Gewichtung der hohen LS nicht am schlechtesten ab und nähert sich dem mittleren Niveau der SRS an. Jedoch ist die Varianz größer (siehe etwa **Abbildung A16**). Die Verwendung der hohen LS ist somit für diese Datensituation ungeeignet zur Wahl der Beobachtungen für die Stichprobe. Eine Vermutung dafür ist, dass die Fälle, die keine Träger einer Wechselwirkungen sind hohe LS besitzen, dadurch bevorzugt in die Stichprobe gelangen und entsprechend Wechselwirkungen in die Modelle aufgenommen werden, die keinen Einfluss besitzen.

Die Gewichtung der niedrigen LS durch f_8 hebt sich ähnlich wie in Simulation 1 bei niedrigen Prozentanteilen der Daten hervor. Bei 10% der Originaldaten liegt das mittlere Niveau ab $d = 20$ bei der zufälligen Auswahl proportional zu den LS deutlich über dem mittleren Niveau der SRS. Mit mehr Variablen im Datensatz steigt die Differenz und liegt bei $d = 50$ mit 6,6 Prozentpunkten über dem arithmetischen Mittel der SRS. Dies entspricht etwa 24 Personen, die im Mittel zusätzlich häufiger richtig klassifiziert werden

(siehe **Abbildung A9**). Der Unterschied in den Klassifikationsraten zu Simulation 1 liegt nahe bei dem erwarteten Verlust durch das Rauschen von etwa 12 bis 13 Prozentpunkten. Bei der festen Auswahl der Beobachtungen anhand der LS liegt das mittlere Niveau für diese Gewichtung ab $d = 20$ nahezu gleichauf mit der zufälligen Auswahl. Die Varianz der festen Auswahl ist zwar geringer als die Varianz der zufälligen Auswahl, jedoch erreicht die zufällige Auswahl im arithmetischem Mittel höhere Klassifikationsraten. Interessant ist erneut der Kontrast zwischen den Klassifikationsraten durch f_3 und f_8. Bei niedrigen Prozentanteilen der Originaldaten liegt das mittlere Niveau durch die Gewichtung von f_8 deutlich über dem mittleren Niveau von f_3 (siehe **Abbildung 20** und **Abbildung A9**). Mit steigendem Prozentanteil kehrt sich dies um und das durch f_3 erreichte mittlere Niveau liegt deutlich über dem mittleren Niveau von f_8 (siehe **Abbildung A13** und **Abbildung A15**). Dies unterstützt erneut die Vermutung, dass es bei geringen Prozentanteilen vorteilhaft ist, nur niedrige LS zu berücksichtigen, dies jedoch bei steigenden Prozentanteilen zum Nachteil wird. Ab einem Prozentanteil von 15% zeigt sich bei der Gewichtung durch f_8 erneut, dass die Klassifikationsraten relativ konstant bleiben und sich nicht mehr so stark verbessern wie bei den anderen Gewichtungen und entsprechend unter dem mittleren Niveau der SRS zurückbleiben.

Mit steigendem Prozentsatz heben sich andere Gewichtungen der LS hervor als in Simulation 1. In erster Linie sind dies bei der zufälligen Auswahl proportional zu den LS die Gewichtungen, welche die Randbereiche (mit-) berücksichtigen: f_1, f_4, f_5 und f_7. Ab einem Prozentsatz von 15% der Originaldaten liegt das mittlere Niveau der Klassifikationsraten durch diese Gewichtsfunktionen konstant über dem mittlerem Niveau der SRS, mit je nach Gewichtung und Anzahl der Variablen 0,6 bis 1,7 Prozentpunkte über dem mittleren Niveau der SRS. Am deutlichsten hebt sich dabei die Gewichtung durch f_5 hervor, die gegen die mittleren LS diskriminiert und nur die beiden Randbereiche der Werte berücksichtigt (vgl. **Tabelle 2**, siehe **Abbildung A11**). Dies gilt bei steigendem Prozentsatz vor allem bei größerer Anzahl Variablen d im Datensatz (vgl. **Abbildung A13 (c)**, **(d)** und **(e)**). Es scheint in dieser Datensituation bei größeren Prozentsätzen der verwendeten Daten am besten für die Wahl der Beobachtungen zu sein, gegen die mittleren LS zu diskriminieren. Die Varianz der Klassifikationsraten ist bei diesen Gewichtungen in etwa auf demselben Niveau, wobei die Standardabweichung der Klassifikationsraten für die Gewichtung durch f_5 leicht höher ist. Dies liegt vermutlich daran, dass es für diese Gewichtung etwas stärkere und mehr negative Ausreißer in den Klassifikationsraten gibt (siehe etwa **Abbildung A13 (d)**). Der Verlust im mittleren Niveau bei diesen Gewichtsfunktionen (sowie bei der SRS) liegt im Vergleich zur Simulation 1 ab einem Prozentsatz von 10% der Originaldaten deutlich über dem OOB-Fehler und eher in Bereichen von 15 bis 17 Prozentpunkten statt der erwarteten 13 Prozentpunkte.

In **Tabelle 7** sind die Differenzen im mittleren Niveau in Prozentpunkten zwischen der Simulation 1 und der Simulation 2 bei 25% der Originaldaten für diese Gewichtsfunktionen angegeben. Für fast alle Gewichtungen und Prozentsätze ist der Verlust im mittleren

Tabelle 7: Differenzen im mittleren Niveau der Klassifikationsraten beim Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz zwischen vergleichbaren Daten der Simulation 1 und Simulation 2 bei zufälliger Auswahl der Beobachtungen anhand der Leverage Scores mit 25% der Originaldaten. Die Differenzen sind in Prozentpunkten angegeben.

d	SRS	f_1	f_4	f_5	f_7
10	15,78	19,91	19,73	19,90	19,59
20	16,48	15,94	15,94	15,69	16,06
30	16,97	15,36	15,44	15,54	15,31
40	17,22	16,35	16,44	16,17	16,18
50	17,23	16,62	16,40	16,38	16,59

Niveau gegenüber Simulation 1 bei $d = 10$ am größten, fällt dann ab und steigt mit steigendem d wieder an. Das Rauschen in den Daten verstärkt sich somit bei einer aus den Daten entnommenen Stichprobe. Eine mögliche Erklärung dafür kann dieselbe Argumentation wie für die Erhöhung der Varianz sein. Wenn es passiert, dass viele Fälle in die Stichprobe aufgenommen werden, die nicht Träger einer Wechselwirkung sind, verschlechtern sich dadurch die Modelle und entsprechend die Klassifikationsraten. Bei kleinen Stichproben ist die Klassifikationsrate von vorneherein geringer und dieser Effekt verstärkt sich durch die verrauschten Beobachtungen.

Interessanterweise ist die Klassifikation durch die Gewichtungen, welche die Randbereiche mitberücksichtigen, von der Anzahl der Variablen d in dem Sinne abhängig, dass die Klassifikationsraten bei unterschiedlichen Prozentsätzen bei jeweils anderen Werten von d am besten abschneiden. Bei einem Prozentanteil von 15% liegt das höchste mittlere Niveau bei $d = 20$ bei der Gewichtung durch f_5 mit einem Wert von 78,20% und fällt danach wieder ab. Bei 20% bzw. 25% verhält es sich ähnlich. Die Klassifikationsraten liegen bei $d = 20$ und $d = 30$ im mittleren Niveau am höchsten und fallen danach wieder, wobei dieses Abfallen bei 25% der Originaldaten nicht so deutlich ist wie bei 20% der Originaldaten (vgl. **Abbildung A13** und **A15**). Die feste Auswahl der Beobachtungen anhand der LS ist für diese Gewichtsfunktionen durch f_1 vertreten und verhält sich ähnlich wie die zufällige Auswahl. Ab einem Prozentanteil von 15% der Originaldaten und mit steigendem d liegt das mittlere Niveau über dem mittleren Niveau der SRS. Am stärksten zeigt sich dies bei 20% der Originaldaten und nimmt mit 25% wieder ab. Bei einem Prozentanteil von 20% liegt das mittlere Niveau ab $d = 30$ bei etwa 1,6 Prozentpunkten über dem mittleren Niveau der SRS. Dies entspricht etwa 5 Beobachtungen, die im Mittel zusätzlich häufiger richtig klassifiziert werden. Der Verlust im mittleren Niveau gegenüber der Simulation 1 liegt auf derselben Höhe wie bei der zufälligen Auswahl proportional zu den LS, bei mehr Variablen im Datensatz liegt der Verlust etwa ein Prozentpunkt höher. Generell liegt das mittlere Niveau der festen Auswahl unter dem mittleren Niveau der zufälligen Auswahl bzw. nahezu gleichauf (siehe etwa **Abbildung A15** und **Abbildung A16**). Der Vorteil

der festen Auswahl liegt darin, dass es weniger negative Ausreißer gibt und entsprechend die Varianz der Klassifikationsraten geringer ist. Dabei bleibt zu berücksichtigen, dass für jede Datensituation bei der festen Auswahl 100 Klassifikationsraten vorliegen, für die zufällige Auswahl proportional zu den LS jeweils 2000.

Die Gewichtung der mittleren LS durch f_6, f_{10} und f_{11} hebt sich nicht besonders hervor, weder bei der zufälligen noch der festen Auswahl der Beobachtungen und liegt konstant auf dem mittleren Niveau der SRS. Es kann somit davon ausgegangen werden, dass die mittleren LS in dieser Datensituation keinen Einfluss auf die Wahl der Beobachtungen haben.

Bei $d = 10$ Variablen gibt es weder bei der zufälligen Auswahl proportional zu den LS noch bei der festen Auswahl eine Gewichtung der LS, die sich merklich gegen die SRS abhebt. Die einzige Ausnahme dazu bildet die Gewichtung durch f_2 bei einem Prozentanteil von 20% bzw. 25% der Originaldaten. In diesen beiden Fällen liegt das mittlere Niveau minimal über dem mittleren Niveau der SRS. Dabei sollte erneut berücksichtigt werden, dass solche kleinen Unterschiede eventuell zufallsbedingt sind. Weiter unterstützt dies die Vermutung, dass die LS nützlicher bei einer steigenden Anzahl nicht-informativer Variablen wird.

In **Tabelle 8** sind die Differenzen im mittleren Niveau der Klassifikationsraten $\Delta\overline{SRS}_P$ bei der zufälligen Auswahl der Beobachtungen anhand der SRS und proportional gewichtet zu den LS, sowie die Differenz der Anzahl der Beobachtungen, die im Mittel zusätzlich häufiger richtig klassifiziert werden, zwischen der SRS und der zufälligen Auswahl proportional zu den LS $\Delta\overline{K}_P$ dargestellt. Die Unterschiede in den Klassifikationsraten sind in Prozentpunkten angegeben. Ein positiver Wert bedeutet, dass die Gewichtung der LS im Mittel zu einem besseren Wert führt. Die erste Spalte enthält den Prozentsatz P, die zweite Spalte die Anzahl der Variablen d im Datensatz. Die restlichen Spalten enthalten entsprechend die Differenzen für die jeweilige Gewichtsfunktion. In jeder Zeile ist der jeweils beste Wert besonders hervorgehoben. Wie in **Tabelle 4** ist zu bemerken, dass es bis auf bei $d = 10$ immer eine Gewichtung der LS gibt, die im mittleren Niveau über der SRS liegt. Auch fällt auf, dass die größten Zugewinne bei einem Prozentsatz von 5% bzw. 10% der Originaldaten durch die Gewichtung mit f_8 erzielt werden. Dabei ist der Verlust im mittleren Niveau durch den Sprung von 5% auf 10% der Originaldaten viel geringer als in Simulation 1 und der Nutzen durch die Gewichtung der LS entsprechend größer. Weiter ist festzuhalten, dass bei steigendem Prozentanteilen abhängig von der Anzahl der Variablen sehr unterschiedliche Gewichtungen am besten abschneiden. Dafür liegen die Zugewinne eher bei über einem Prozentpunkt im mittleren Niveau und sind leicht größer als bei Simulation 1. Dies unterstützt weiter die Vermutung, dass die LS besser für die Wahl der Beobachtungen geeignet sind je mehr überflüssige bzw. ungeeignete Informationen (in Form von überzähligen Variablen bzw. Rauschen) sich in den Daten befinden.

Bei den beiden Prozentsätzen von 5% bzw. 10% der Originaldaten hebt sich somit erneut die Gewichtung durch f_8 hervor. Bei einem Prozentsatz von 5% der Originalda-

Tabelle 8: Differenzen im mittleren Niveau der Klassifikationsraten und mittlere Anzahl Beobachtungen die zusätzlich häufiger richtig klassifiziert werden bei der zufälligen Wahl der Beobachtungen anhand der einfachen Zufallsauswahl und proportional gewichtet nach den Leverage Scores mit den Daten der Simulation 2. Die Spalten mit $\Delta\overline{SRS}_P$ enthalten die Differenzen im mittleren Niveau der Klassifikationsraten in Prozentpunkten und die Spalten mit $\Delta\overline{K}_P$ die Differenzen der im Mittel häufiger richtig klassifizierten Beobachtungen. Hervorgehoben ist in jeder Spalte jeweils der größte positive Zugewinn.

		f_1		f_2		f_5		f_8	
P	d	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$
5	10	-3,04	-12,06	-2,19	-8,68	-4,03	-15,96	**3,01**	**11,91**
	20	-1,85	-7,34	-1,79	-7,07	-1,85	-7,33	**6,31**	**24,97**
	30	-1,69	-6,69	-1,12	-4,42	-1,87	-7,39	**6,88**	**27,24**
	40	-1,35	-5,35	-1,23	-4,87	-1,35	-5,35	**5,96**	**23,62**
	50	-0,53	-2,11	-0,47	-1,88	-0,59	-2,33	**6,83**	**27,04**
10	10	-0,08	-0,30	-0,92	-3,59	-1,84	-7,22	-0,61	-2,39
	20	-0,06	-0,25	-1,77	-6,95	0,04	0,15	**3,08**	**12,06**
	30	-0,57	-2,24	-1,33	-5,20	-0,28	-1,09	**4,71**	**18,46**
	40	-0,42	-1,64	-0,84	-3,28	-0,52	-2,06	**5,72**	**22,41**
	50	-0,65	-2,54	-0,77	-3,03	-0,43	-1,67	**6,67**	**26,15**
15	10	-0,12	-0,48	-0,43	-1,67	-2,60	-10,08	-3,69	-14,31
	20	1,47	5,70	-0,36	-1,38	**1,72**	**6,66**	-0,11	-0,44
	30	0,97	3,77	-0,56	-2,18	**1,66**	**6,44**	0,78	3,04
	40	0,63	2,46	-0,76	-2,94	1,01	3,92	**1,66**	**6,46**
	50	0,54	2,09	-0,62	-2,42	1,05	4,07	**2,15**	**8,32**
20	10	-0,64	-2,45	**0,04**	**0,17**	-3,86	-14,83	-6,38	-24,49
	20	**1,23**	**4,71**	0,15	0,59	0,95	3,64	-3,16	-12,12
	30	1,39	5,34	-0,04	-0,15	**1,70**	**6,54**	-2,49	-9,57
	40	0,77	2,97	-0,03	-0,13	**1,37**	**5,24**	-1,63	-6,28
	50	0,91	3,49	-0,18	-0,69	**1,64**	**6,28**	-1,40	-5,37
25	10	-0,84	-3,21	**0,62**	**2,36**	-4,95	-18,81	-8,21	-31,21
	20	**0,75**	**2,87**	0,41	1,56	-0,11	-0,42	-5,38	-20,45
	30	**1,27**	**4,82**	0,54	2,04	0,92	3,48	-4,39	-16,66
	40	0,93	3,53	0,24	0,93	**1,32**	**5,01**	-3,51	-13,34
	50	0,99	3,76	0,09	0,35	**1,32**	**5,01**	-3,62	-13,75

ten entsprechen die am häufigsten gefundenen Wechselwirkungen sehr stark denen aus Simulation 1 (siehe **Abbildung 17**). Vor allem liegen die Wichtigkeitsmaße in ähnlichen Größenverhältnissen und die Varianz ist in etwa gleich. Bei der SRS zeigt sich hingegen, dass bei diesem Prozentanteil viele nach Konstruktion unwichtige Binärvariablen gefunden und als wichtig bewertet werden. Bei $d = 30$ sind 6 der 20 am häufigsten gefunden Wechselwirkungen Binärvariablen, die keinen Einfluss auf den Krankheitsstatus besitzen. Mit steigender Anzahl Variablen im Datensatz nimmt dies wieder etwas ab, so sind es bei $d = 50$ vier Binärvariablen. Bei der Gewichtung durch f_8 tritt unter den 20 am häufigsten gefundenen Wechselwirkungen keine Binärvariable alleine auf, die nach Konstruktion

unwichtig wäre. Dafür zeigt sich, dass die Binärvariable X_2 sehr häufig in Kombination mit anderen, nach Konstruktion unwichtigen Binärvariablen auftritt. Dies verstärkt sich deutlich bei einem Prozentanteil von 10% der Originaldaten. Bei diesem Prozentanteil tritt die Binärvariable X_2 in 15 der 20 am häufigsten gefundenen Wechselwirkungen auf, in 11 davon in Kombination mit unwichtigen Binärvariablen. Dies verstärkt sich mit steigender Anzahl Variablen im Datensatz. Bei $d = 40$ sind es 13 von 16 Wechselwirkungen die X_2 und unwichtige Binärvariablen enthalten, bei $d = 50$ sind es 12 von 15. Die gefundenen Wechselwirkungen werden eindeutig durch die Binärvariable X_2 dominiert. Im Vergleich zu **Abbildung 19** ist dies sehr auffällig, da die Binärvariable X_2 sich höchstens viermal unter den am häufigsten gefundenen Wechselwirkungen befindet. Die Ähnlichkeit der gefundenen Wechselwirkungen bei der Gewichtung durch f_8 mit 5% der Originaldaten dürfte der Grund dafür sein, dass das mittlere Niveau für diese Gewichtung besser liegt und niedriger gefallen ist als bei den meisten anderen Gewichtungen (vgl. **Tabelle 5**). Die feste Auswahl der Beobachtungen anhand der LS führt bei dem Prozentsatz von 5% der Originaldaten dazu, dass viele Zweifachwechselwirkungen aufgenommen werden, die eine nach Konstruktion einflussreiche und eine unwichtige Binärvariable enthalten. Dabei liegt der Fokus nicht auf der Binärvariable X_2, sondern ist über die Wechselwirkungen verteilt. Auch werden Zweifachwechselwirkungen gefunden, die nur aus unwichtigen Binärvariablen bestehen, bei denen das Wichtigkeitsmaß vergleichsweise hohe Werte annimmt. Dies dürfte eine Erklärung dafür sein, dass das mittlere Niveau der zufälligen Auswahl proportional zu den LS höher liegt als bei der festen Auswahl der Beobachtungen. Zum Finden wichtiger Wechselwirkungen ist die feste Auswahl somit eher ungeeignet, auch wenn sich die unwichtigen Wechselwirkungen nicht so stark auf die Klassifikationsraten auszuwirken scheinen. Anzumerken ist dabei, dass für die feste Auswahl der Beobachtungen der logicFS-Ansatz auf den jeweiligen Datensatz nur einmal angewandt wird, für die zufällige Auswahl entsprechend 20-mal pro Datensatz. Dies unterstützt die Vermutung, dass es für das Finden wichtiger Wechselwirkungen besser ist, die Modelle öfter anzupassen und die Wechselwirkungen zu verwenden, die durch die Modelle öfter gefunden werden.

Mit steigenden Prozentanteilen heben sich die Gewichtsfunktionen hervor, die sowohl die niedrigen als auch die hohen LS berücksichtigen. In **Abbildung 22** sind Boxplots der Wichtigkeitsmaße (aus **Gleichung (1)**) der gefundenen Wechselwirkungen bei wiederholtem Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz bei zufälliger Wahl der Beobachtungen per SRS und proportional zu den LS gewichtet durch f_5 mit 25% der Originaldaten dargestellt. Sortiert sind die gefundenen Wechselwirkungen danach wie häufig diese in allen Modellen vorkommen, von oben nach unten. Eingefärbt sind für jedes d jeweils die Überschneidungen in den Wechselwirkungen zwischen der Wahl per SRS und der Gewichtung durch f_5. Sofort fällt auf, dass der Großteil der Wechselwirkungen ebenfalls durch die Binärvariable X_2 dominiert ist und viele Zweifachwechselwirkungen auftreten, die X_2 und eine andere nach Konstruktion unwichtige Binärvariable enthalten. Bei kleinerem d gibt es sehr viele Überschneidungen in den am häufigsten gefundenen

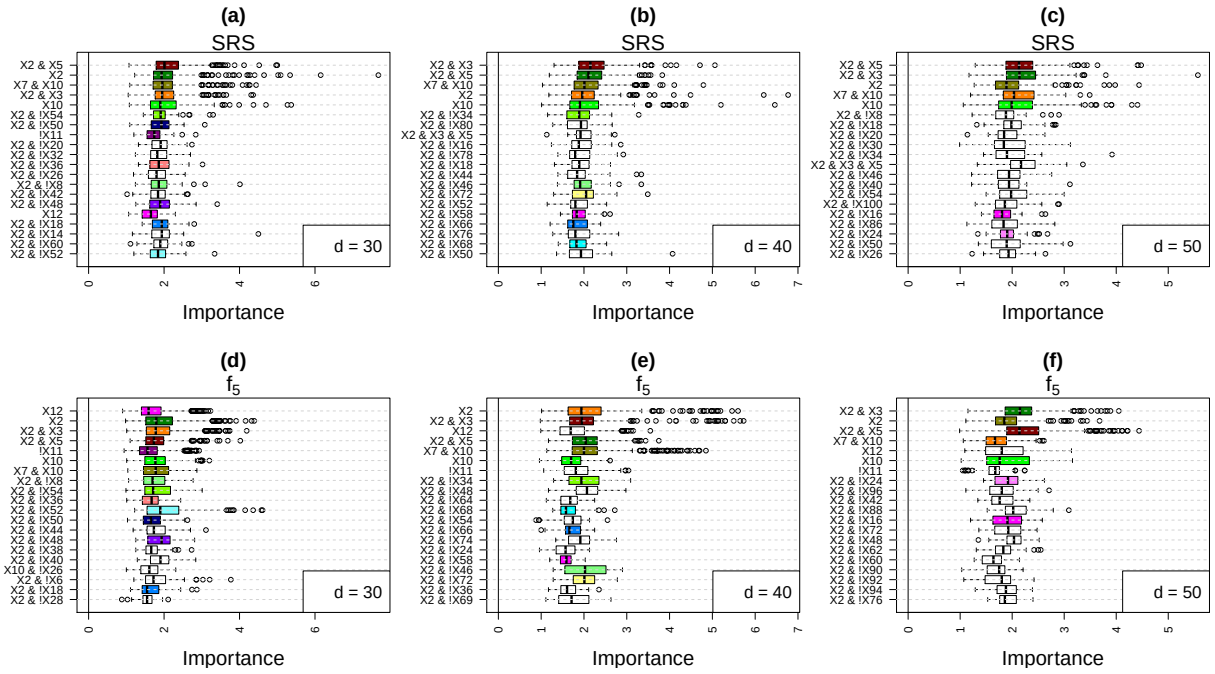

Abbildung 22: Boxplots der Wichtigkeitsmaße der 20 am häufigsten gefundenen Wechselwirkungen bei unterschiedlicher Auswahl einer Stichprobe bestehend aus 25% des Gesamtdatensatzes der Daten aus Simulation 2 für $d = 30, d = 40$ und $d = 50$, sortiert danach, wie häufig diese gefunden werden. Eingefärbt sind die Überschneidungen der Wechselwirkungen zwischen der einfachen Zufallsauswahl und der Auswahl proportional zu den Leverage Scores gewichtet durch f_5.

Wechselwirkungen zwischen der SRS und der Gewichtung durch f_5. Mit steigendem d werden die Überschneidungen weniger. Der Grund für das bessere Abschneiden der Gewichtung durch f_5 ist nicht einfach ersichtlich. Durch die Dominanz der Binärvariable X_2 sind nur wenige andere Wechselwirkungen vertreten. Zu erkennen ist jedoch in **Abbildung 22 (d)**, **(e)** und **(f)**, dass die Binärvariablen X_{11}^{C} und X_{12} durch die Gewichtung mit f_5 häufiger gefunden werden und als wichtiger eingestuft werden. So wird bei $d = 30$ die Binärvariable X_{12} bei der Gewichtung durch f_5 am häufigsten gefunden, bei der SRS erst an sechzehnter Stelle. Bei $d = 40$ und $d = 50$ werden die Binärvariablen X_{11}^{C} und X_{12} durch die SRS gar nicht unter den 20 häufigsten Wechselwirkungen gefunden. Selbst wenn der Zugewinn im mittleren Niveau durch die Gewichtung der LS nur im Bereich von einem Prozentpunkt bei den Klassifikationsraten liegt, scheint es doch in dieser Datensituation erneut der Fall zu sein, dass die zufällige Wahl der Beobachtungen proportional zu den LS dabei helfen kann, Wechselwirkungen in den Untergruppe der Fälle zu finden. Da die Gewichtung durch f_5 nur insgesamt 25% der Werte zulässt (vgl. **Tabelle 2**), führt dies bei jedem einzelnen Datensatz dazu, dass in jeder Wiederholung die selben Beobachtungen in die Stichprobe aufgenommen werden. Trotzdem verbessert sich das Finden der Wechselwirkungen dadurch, den logicFS-Ansatz auf den selben Daten zu wiederholen, da die gefundenen Wechselwirkungen stabiler werden und die Dominanz durch die Binärvariable

X_2 sich reduziert. Auch hier zeigt sich, dass es von Vorteil zu sein scheint, die Beobachtungen proportional zu den LS in die Stichprobe aufzunehmen und das Anpassen der Modelle zu wiederholen.

Bei der Gewichtung durch f_1 werden die gefundenen Wechselwirkungen bei der zufälligen Auswahl und 25% der Originaldaten ebenfalls von X_2 dominiert. Tendenziell werden die beiden Binärvariablen X_{11}^{C} und X_{12} durch diese Gewichtung ebenfalls öfter gefunden und mit einer höheren Wichtigkeit bewertet als durch die SRS. Dies bildet den Hauptunterschied zwischen den beiden Selektionsmethoden. Die Überschneidungen in den gefundenen Wechselwirkungen ist sehr groß und das Wichtigkeitsmaß liegt in ähnlichen Größenverhältnissen. Durch die Gewichtung mit f_1 werden die mittleren LS mitberücksichtigt, im Gegensatz zu der Gewichtung durch f_5. Bei weniger Variablen im Datensatz scheint dies besser dazu geeignet zu sein, die Beobachtungen auszuwählen, wohingegen es bei mehr Variablen besser zu sein scheint, gegen die mittleren Bereiche der LS zu diskriminieren.

Die feste Auswahl der Beobachtungen durch f_1 ist teilweise auch durch die Binärvariable X_2 dominiert, allerdings nicht ganz so stark. Dafür treten ebenfalls andere wichtige binäre Einflussvariablen in Kombination mit nach Konstruktion nicht einflussreichen Binärvariablen auf. So lassen sich etwas andere binäre Einflussvariablen finden. Der Schwachpunkt, dass die feste Auswahl auf jedem Datensatz nur einmal durchgeführt werden kann, bleibt jedoch bestehen und es ist nicht garantiert, dass interessante Wechselwirkungen sich in der festen Auswahl finden.

Insgesamt ist die Auswahl der Beobachtungen anhand der LS in dieser Datensituation nicht so eindeutig wie bei Simulation 1. Parallel ist, dass bei einem geringen Prozentsatz der Originaldaten die Wahl anhand der niedrigen LS besser geeignet ist und zwar teilweise sehr deutlich. Der Verlust im mittleren Niveau ist geringer als bei anderen Gewichtsfunktionen und die gefundenen Wechselwirkungen ähneln denen aus Simulation 1. Erneut zeigt sich, dass die Wahl der Beobachtungen anhand der niedrigen LS bei einer geringen Anzahl an Daten bzw. vieler unwichtiger Variablen im Datensatz am besten zur Wahl der Beobachtungen geeignet ist. Mit steigendem Prozentanteil ergibt sich im Unterschied zur Simulation 1, dass die Wahl der Beobachtungen anhand der niedrigen und hohen LS am besten geeignet ist und dass die Gewichtung durch die LS dabei helfen kann, in einer erschwerten Situation Wechselwirkungen für Untergruppen der Fälle zu finden.

Im nächsten Abschnitt geht es um die Datensituation, wenn es deutlich mehr Beobachtungen als Variablen gibt.

5.2.4 Auswertung von Simulation 3

Mit Simulation 3 liegen die für diese Situation am höchsten dimensionierten Datensätze vor. Bei dieser Dimensionierung handelt es sich um die realistischste, wenn es um die Reduktion von hochdimensionierten Datensätzen geht. Von Interesse ist, wie sich die Ergebnisse der Auswertung von Simulation 1 verhalten, wenn deutlich mehr Beobachtungen

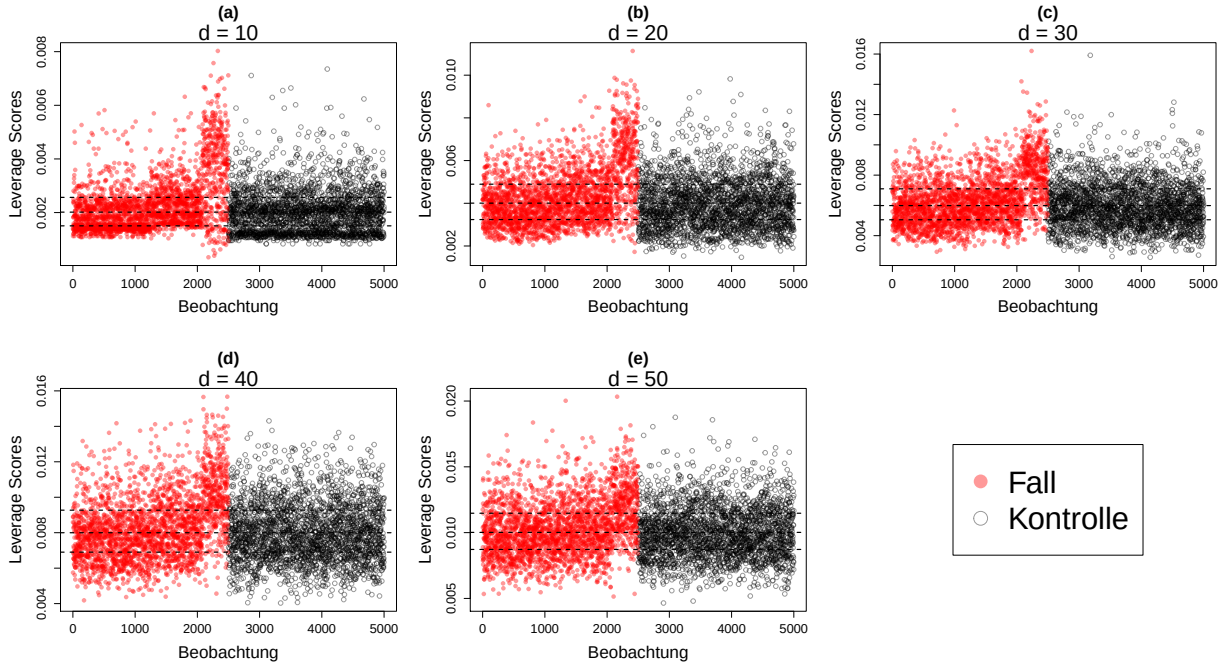

Abbildung 23: Beispiel der Leverage Scores von jeweils 5000 Beobachtungen unterschiedlicher Datensätze mit Anzahl Variablen d aus Simulation 3.

vorliegen. Jeder Datensatz der Simulation 3 besitzt nun $n = 5000$ Beobachtungen. Die Anzahl der Variablen d ist parallel zu den anderen beiden Simulationen und auch der relative Anteil der Fälle der Untergruppen ist im selben Verhältnis wie bei den anderen vorhergegangenen Simulationen, wobei der absolute Anteil entsprechend größer ist. Da die Datensätze hochdimensioniert sind, wird es nun darum gehen, Stichproben von sehr kleinem Umfang zu ziehen. Die Prozentanteile P, die dafür gewählt werden, sind 1%, 2%, 3%, 4% und 5% der Originaldaten. Dies führt zu Stichprobenumfängen von $n' \in \{50, 100, 150, 200, 250\}$. Entsprechend stehen ab 2% der Gesamtdaten so viele Beobachtungen zur Verfügung, wie bei den vorherigen beiden Simulationen bei 25% der Originaldaten. Wie zuvor sollen die Beobachtungen anhand der LS zufällig bzw. fest in die Stichprobe aufgenommen werden. In **Abbildung 23** sind beispielhaft die LS der jeweils 5000 Beobachtungen pro Datensatz für unterschiedliche Anzahl der Variablen d aufgetragen.

Als horizontale Linien sind jeweils das 0,25-, das 0,5- und das 0,75-Quantil eingetragen. Deutlich lässt sich in **Abbildung 23 (a)** erkennen, dass die LS der Träger von L_3 sehr hohe bzw. auch niedrige LS-Werte besitzen. Mit einer steigenden Anzahl an Variablen nimmt dies ab, diese Beobachtungen heben sich jedoch weiter hervor. Die Struktur der LS weist Ähnlichkeiten zu der Struktur in **Abbildung 14** der Simulation 1 auf. Dies lässt die Vermutung zu, dass sich für die Gewichtsfunktionen, welche die Randbereiche gewichten (wie etwa f_9), ähnliche Erkenntnisse ergeben werden.

Einführend werden auf den gesamten Datensätzen logische Regressionsmodelle nach dem logicFS-Ansatz angepasst. Es ergibt sich für alle Anpassungen ein OOB-Fehler von

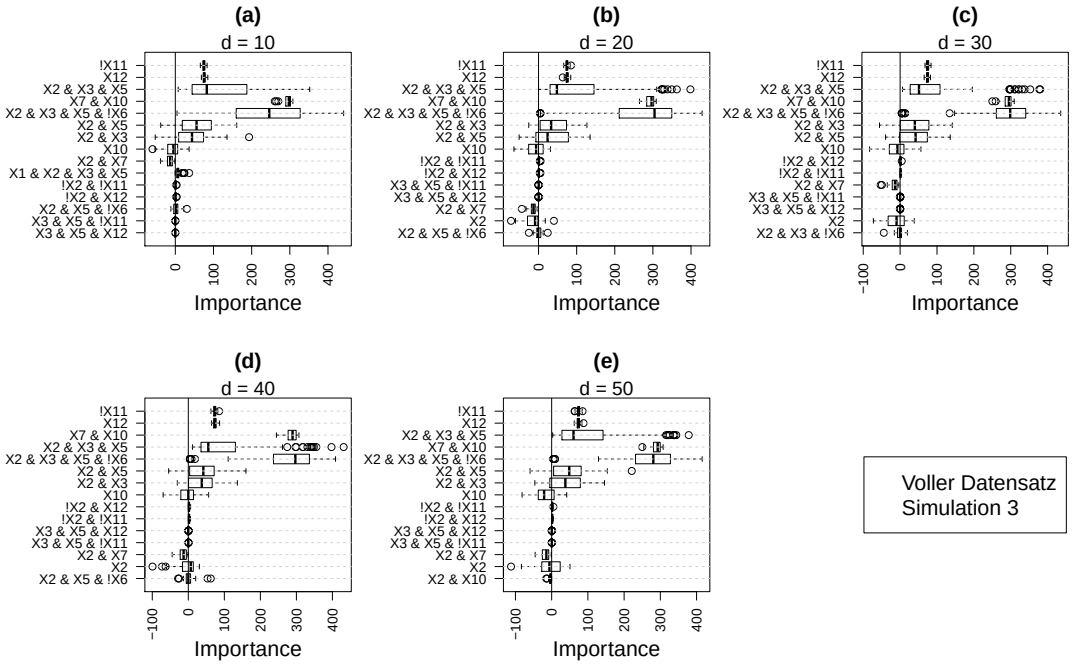

Abbildung 24: Boxplots der Wichtigkeitsmaße der gefundenen Wechselwirkungen beim Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf die vollen Datensätze der Simulation 3.

nahezu 0%, unabhängig von der Anzahl Variablen d im Datensatz. Maximal beträgt der OOB-Fehler 0,1%, was bedeutet, dass maximal 5 Beobachtungen falsch klassifiziert werden. Eine steigende Anzahl der Beobachtungen verbessert somit die Anpassung der Modelle und es kann davon ausgegangen werden, dass es möglich ist, ein für die Daten sehr gut passendes Modell aufzustellen.

In **Abbildung 24** sind Boxplots der Wichtigkeitsmaße (aus **Gleichung (1)**) der gefundenen Wechselwirkungen dargestellt, geordnet nach den am häufigsten gefundenen Wechselwirkungen, von oben nach unten. Im Vergleich zu **Abbildung 13** lässt sich erkennen, dass die gefundenen Wechselwirkungen komplexer werden und sich den DNF (aus **Gleichung (6)** bis **(8)**) der nach Konstruktion wahren Einflüssen annähern. Auch wird die Varianz des Wichtigkeitsmaßes für die beiden Binärvariablen X_{11}^{C} und X_{12} stabiler. Das Ausmaß der Wichtigkeit hängt, wie bereits angemerkt, von der Anzahl der Beobachtungen ab. Bei steigender Anzahl der Beobachtungen nimmt das Maß entsprechend deutlich zu. Teilweise werden noch Überschneidungen von binären Einflussvariablen gefunden, die nach Konstruktion keinen Einfluss besitzen, wie z.B. $(X_2^{\mathrm{C}} \wedge X_{12})$. Jedoch besitzen diese Wechselwirkungen eine vergleichsweise geringe Wichtigkeit.

Wie im bisherigen Vorgehen werden Stichproben aus den jeweiligen Datensätzen sowohl zufällig proportional zu den LS und fest anhand der LS entnommen und nach dem logicFS-Ansatz logische Regressionsmodelle angepasst.

In **Abbildung 25** sind die Klassifikationsraten durch das zufällige Auswählen der Beobachtungen proportional zu den LS bei 1% der Originaldaten dargestellt. Wie bei

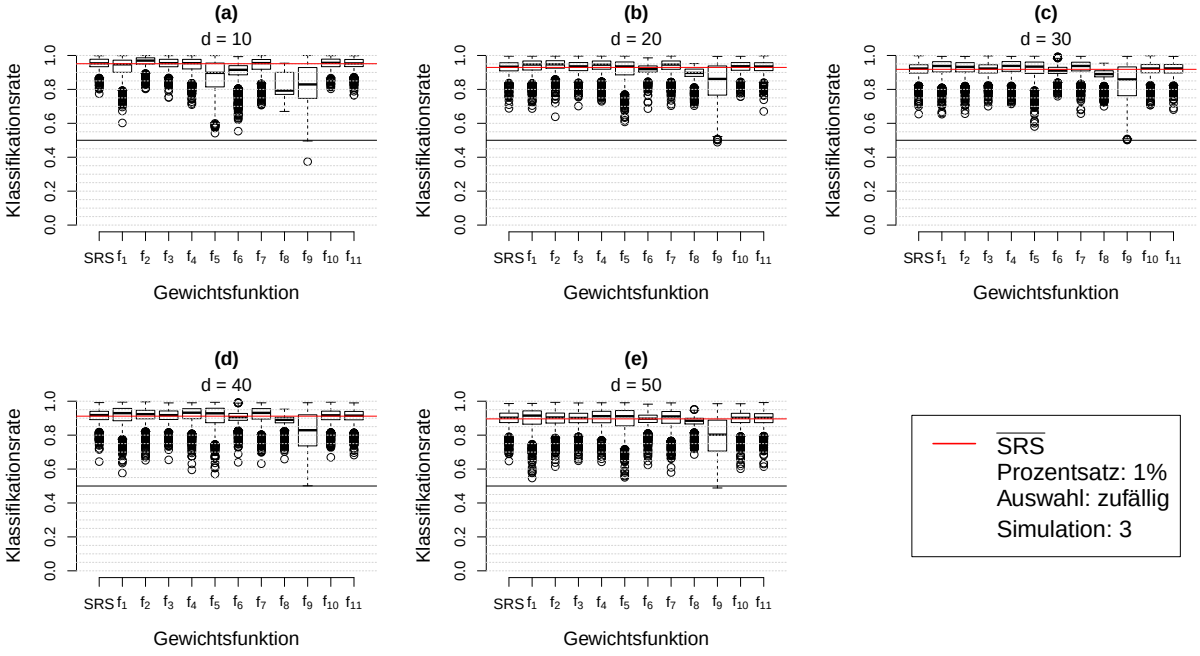

Abbildung 25: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 1% der Daten aus Simulation 3, bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

Simulation 1 sind die Klassifikationsraten linksschief mit relativ vielen Ausreißern nach unten. Im Vergleich zu **Abbildung A3** fällt auf, dass das mittlere Niveau in etwa 1 bzw. 2 Prozentpunkte niedriger liegt. In Simulation 1 stehen in dieser Situation jeweils 60 Beobachtungen zur Verfügung, für Simulation 3 dagegen 50 Beobachtungen. Vermutlich liegt die Verschlechterung daran, dass in Simulation 3 mehr Beobachtungen für die Klassifikation zur Verfügung stehen und daher auch mehr Beobachtungen potentiell falsch klassifiziert werden können. Wie in **Abbildung 25 (a)**, **(b)** und **(c)** bereits zu erkennen, zeigt sich in dieser Datensituation im Vergleich zu Simulation 1 und 2 ein deutliches Muster. Für jeden Prozentsatz der Daten und jede Anzahl Variablen d liegen bei der zufälligen Auswahl proportional zu den LS konstant diejenigen Gewichtungen über dem mittleren Niveau der SRS, welche die hohen LS berücksichtigen. Bei steigenden Prozentsätzen der verwendeten Daten sind dies in erster Linie die beiden Gewichtsfunkionen f_2 und f_9. Da die Gewichtsfunktion f_2 die LS unverändert lässt, führt dies automatisch dazu, dass Beobachtungen mit einer hohen Hebelwirkungen mit größerer Wahrscheinlichkeit in die Stichprobe aufgenommen werden.

In **Tabelle 9** sind die Differenzen $\Delta\overline{SRS}_P$ im mittleren Niveau der Klassifikationsraten und die Anzahl der im Mittel zusätzlich häufiger richtig klassifizierten Beobachtungen $\Delta\overline{K}_P$ zwischen der SRS und der zufälligen Auswahl der Beobachtungen proportional gewichtet nach den LS dargestellt. Die erste Spalte enthält den entsprechenden Prozentanteil P und die zweite Spalte die Anzahl d der Variablen im Datensatz. Die restlichen Spalten enthalten die Differenzen der entsprechende Gewichtsfunktion aufgeteilt nach der

Tabelle 9: Differenzen im mittleren Niveau der Klassifikationsraten und mittlere Anzahl Beobachtungen die zusätzlich häufiger richtig klassifiziert werden bei der zufälligen Wahl der Beobachtungen anhand der einfachen Zufallsauswahl und proportional gewichtet nach den Leverage Scores mit den Daten der Simulation 3. Die Spalten mit $\Delta\overline{SRS}_P$ enthalten die Differenzen im mittleren Niveau der Klassifikationsraten in Prozentpunkten und die Spalten mit $\Delta\overline{K}_P$ die Differenzen der im Mittel häufiger richtig klassifizierten Beobachtungen. Hervorgehoben ist in jeder Spalte jeweils der größte positive Zugewinn.

		f_1		f_2		f_4		f_7		f_9	
P	d	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$	$\Delta\overline{SRS}_P$	$\Delta\overline{K}_P$
1	10	-2,23	-110,44	**1,19**	**59,14**	-1,07	-53,07	-1,19	-58,99	-12,15	-601,31
	20	0,28	14,07	**1,27**	**62,65**	0,63	31,31	0,69	34,20	-8,21	-406,23
	30	0,67	33,39	0,67	33,17	**0,98**	**48,54**	0,96	47,65	-7,79	-385,74
	40	0,15	7,34	0,35	17,47	**0,82**	**40,44**	0,77	38,16	-9,52	-471,18
	50	0,13	6,30	-0,10	-4,84	**0,46**	**22,90**	0,23	11,55	-10,59	-523,99
2	10	-0,10	-4,68	**0,50**	**24,49**	0,03	1,30	0,07	3,54	-5,47	-267,93
	20	-0,02	-0,78	**0,50**	**24,31**	-0,08	-3,94	-0,08	-3,68	-0,75	-36,71
	30	0,01	0,41	**0,42**	**20,73**	0,01	0,64	0,08	3,90	0,15	7,27
	40	0,12	5,96	0,38	18,60	0,07	3,21	0,25	12,48	**0,62**	**30,35**
	50	0,35	16,91	0,34	16,66	0,24	11,73	0,32	15,83	**0,67**	**32,62**
3	10	0,11	5,25	**0,25**	**11,97**	0,16	7,80	0,19	9,22	-2,78	-134,98
	20	-0,02	-1,19	**0,25**	**12,15**	-0,12	-5,73	-0,01	-0,54	0,20	9,70
	30	-0,18	-8,65	0,15	7,25	-0,18	-8,80	-0,17	-8,36	**0,39**	**18,86**
	40	-0,03	-1,42	0,24	11,69	-0,01	-0,33	0,05	2,23	**0,50**	**24,38**
	50	-0,09	-4,49	0,15	7,43	-0,19	-9,08	-0,10	-4,98	**0,45**	**21,83**
4	10	0,11	5,14	**0,19**	**8,95**	0,15	6,98	0,13	6,23	-1,80	-86,57
	20	0,00	-0,06	0,16	7,88	-0,05	-2,41	-0,02	-0,91	**0,30**	**14,45**
	30	-0,10	-4,64	0,12	5,77	-0,08	-3,70	-0,13	-6,17	**0,34**	**16,35**
	40	-0,03	-1,60	0,11	5,24	-0,09	-4,41	-0,07	-3,16	**0,33**	**15,64**
	50	-0,15	-7,16	0,05	2,57	-0,12	-5,80	-0,15	-7,08	**0,26**	**12,64**
5	10	0,10	4,61	**0,16**	**7,50**	0,15	6,90	0,11	5,05	-0,81	-38,48
	20	0,02	0,79	0,11	5,17	0,01	0,26	0,00	0,14	**0,29**	**13,63**
	30	0,00	0,24	0,10	4,95	-0,03	-1,56	-0,05	-2,16	**0,29**	**13,70**
	40	0,00	0,10	0,04	1,90	-0,04	-2,11	-0,03	-1,35	**0,28**	**13,11**
	50	-0,05	-2,27	0,04	2,07	-0,05	-2,31	-0,05	-2,44	**0,26**	**12,13**

Differenz der Prozentpunkte im mittleren Niveau der Klassifikationsrate und wie viele Beobachtungen entsprechend im Mittel zusätzlich häufiger richtig klassifiziert werden. Ein positiver Wert bedeutet eine Verbesserung gegenüber der SRS, ein negativer Wert entsprechend eine Verschlechterung. Besonders hervorgehoben ist in jeder Zeile jeweils der Wert der im Vergleich zur SRS die größte positive Abweichung besitzt. Es werden diejenigen Gewichtsfunktionen dargestellt, welche die hohen LS mitberücksichtigen, da diese sich unter den anderen Gewichtungen am deutlichsten hervorheben. Als erstes fällt auf, dass (wie bei den vorherigen Simulationen) die größten Differenzen im mittleren Niveau bei kleinen Prozentsätzen liegen, am deutlichsten bei den Gewichtungen durch f_2 und f_4. Weiter scheint es von der Anzahl der Variablen abhängig zu sein, welche der Gewichtsfunktionen sich am stärksten abhebt. Bei $d = 10$ ist dies konstant die Gewichtung durch f_2. Weiter ist die Gewichtung durch f_4 nur bei einem Prozentsatz von 1% der Originaldaten unter den besten Gewichtungen. Die Verbesserung gegenüber der SRS liegt

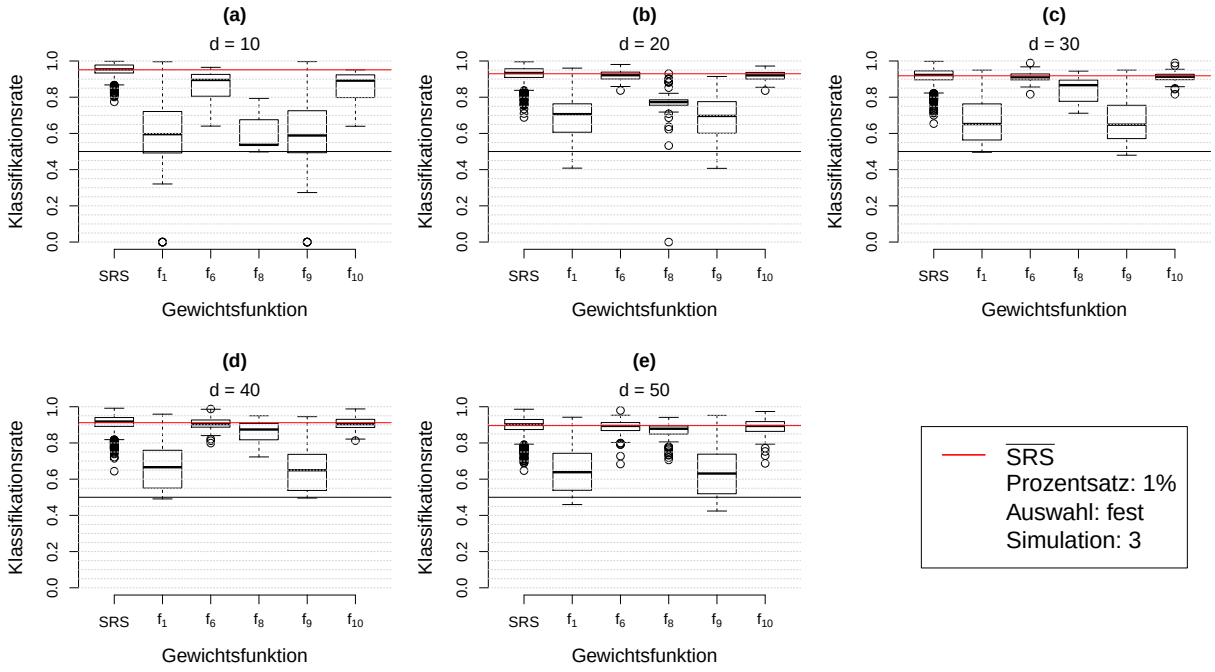

Abbildung 26: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 1% der Daten aus Simulation 3 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

bei fast einem Prozentpunkt im mittleren Niveau und aufgrund der großen Umfänge der Datensätze bedeutet dies, dass teilweise über 40 oder sogar 60 Beobachtungen zusätzlich im Mittel häufiger richtig klassifiziert werden.

Die geringsten Zugewinne gibt es bei einem Prozentsatz von 5% der Originaldaten. Ein Grund dafür kann sein, dass die Klassifikationsraten generell sehr hoch liegen (vgl. **Abbildung A23**) und entsprechend generell nur wenig Beobachtungen falsch klassifiziert werden, was wiederum wenig Spielraum für Verbesserungen in den Klassifikationsraten zulässt. Dennoch hebt sich die Gewichtung durch f_9 ab $d = 20$ gegenüber der SRS ab und es werden etwa 13 Beobachtungen zusätzlich im Mittel häufiger richtig klassifiziert.

Interessant ist, dass die beiden Gewichtsfunktionen f_1 und f_7 sich in keiner Situation gegenüber den anderen drei Gewichtungen abheben. Diese Gewichtungen diskriminieren am schwächsten gegen die mittleren LS. Eventuell sind die mittleren LS nicht gut für die Wahl der Beobachtungen geeignet. Weiter verstärkt wird dies dadurch, dass die Gewichtung der mittleren LS durch f_6, f_{10} und f_{11} sich in keiner Situation gegenüber der SRS abheben. Das mittlere Niveau liegt nahezu immer gleichauf bzw. teilweise unter dem mittleren Niveau der SRS. Dies gilt sowohl bei der zufälligen als auch der festen Auswahl der Beobachtungen. Es kann somit davon ausgegangen werden, dass die mittleren LS in dieser Datensituation ebenfalls keinen Einfluss auf die Auswahl der Beobachtungen besitzen.

In **Abbildung 26** sind die Klassifikationsraten resultierend durch die feste Auswahl der Beobachtungen anhand der LS bei 1% der Originaldaten dargestellt. Deutlich ist zu erkennen, dass bei der festen Auswahl eine Berücksichtigung der hohen LS zu vergleichs-

weise sehr schlechten Ergebnissen führt. Dahingegen liegt die Gewichtung der mittleren LS sehr stark auf demselben mittleren Niveau wie die SRS. In den Grafiken **Abbildung A20 (e)**, **Abbildung A22 (c)**, **(d)**, **(e)** und **Abbildung A24 (c)**, **(d)**, **(e)** im Anhang ist zu erkennen, dass die feste Auswahl der Beobachtungen anhand der Gewichtung durch f_1 und f_9 bei steigenden Prozentsätzen auf nahezu dem gleichen mittleren Niveau der SRS liegt. Es bedarf entsprechend vieler Variablen und eines relativ großen Prozentanteils der Daten, bis die feste Auswahl überhaupt auf vergleichbare Ergebnisse kommt. Der einzige Vorteil liegt in der geringeren Varianz der Klassifikationsraten. Erneut ist dabei anzumerken, dass für die feste Auswahl weniger Datenpunkte als für die SRS vorliegen. Die feste Auswahl der Beobachtungen anhand der LS ist in dieser Datensituation somit nicht zu empfehlen.

Die Wahl der Beobachtungen anhand der niedrigen LS, vor allem mit starker Diskriminierung gegen die anderen Bereiche durch f_8, ist in dieser Datensituation absolut ungeeignet. Das mittlere Niveau liegt konstant unter dem mittleren Niveau der SRS, sowohl bei der zufälligen Auswahl als auch der festen Auswahl (siehe **Abbildung A17** bis **A24**). Erneut zeigt sich der Kontrast zu f_3. Die Gewichtung durch diese Gewichtsfunktion führt erneut im Vergleich zu f_8 zu deutlich besseren Ergebnissen. Es ist somit nur von Vorteil gegen die mittleren und hohen LS zu diskriminieren, wenn sehr wenige Beobachtungen aus den Daten vorliegen. Die geringste Anzahl $n' = 50$ bei 1% der Originaldaten scheint bereits genug zu sein, damit andere Gewichtungen bessere Ergebnisse erzielen.

In **Abbildung 27** sind Boxplots der Wichtigkeitsmaße (aus **Gleichung (1)**) der 20 am häufigsten gefundenen Wechselwirkungen durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz und dem zufälligen Ziehen der Beobachtungen proportional zu den LS gewichtet durch f_2 auf 1% der Originaldaten dargestellt. Angeordnet sind die Boxplots danach, wie häufig die Wechselwirkung in den Modellen gefunden werden, von oben nach unten. Eingefärbt sind die Boxplots der Wechselwirkungen für jede Anzahl der Variablen d, die sich zwischen der SRS und der Gewichtung durch f_2 überschneiden. Die Gewichtung durch f_2 erzielt bei 1% der Originaldaten den größten positiven Abstand zur SRS (vgl. **Tabelle 9**). In **Abbildung 27** ist zu erkennen, dass die Überschneidung der gefundenen Wechselwirkungen mit der SRS sehr groß sind. Verglichen mit **Abbildung 24** ist zu erkennen, dass etwa die Hälfte der gefundenen Wechselwirkungen auf dem gesamten Datensatz sich in denen, basierend auf den Stichproben, wiederfinden. Im Vergleich dazu besitzt keine der Wechselwirkungen ein negatives Wichtigkeitsmaß. Die Größe der Maße ist deutlich geringer, da diese mit deutlich weniger Daten bestimmt werden. Der Großteil der gefundenen Wechselwirkungen korrespondiert zu L_1. Dies ist nicht verwunderlich, da die Träger von L_1 die größte Gruppe unter den Erkrankten stellt.

In **Abbildung 27 (d)** ist zu erkennen, dass die Wechselwirkung $(X_2 \wedge X_3 \wedge X_6^{\mathrm{C}})$ nicht unter den am häufigsten gefundenen Wechselwirkungen bei der SRS ist. Dies könnte ein Grund für die bessere Klassifikation durch f_2 sein. Ein weiterer Grund könnte sein, dass unterschiedliche Binärvariablen öfter gefunden werden. Die Binärvariable X_{12} wird durch

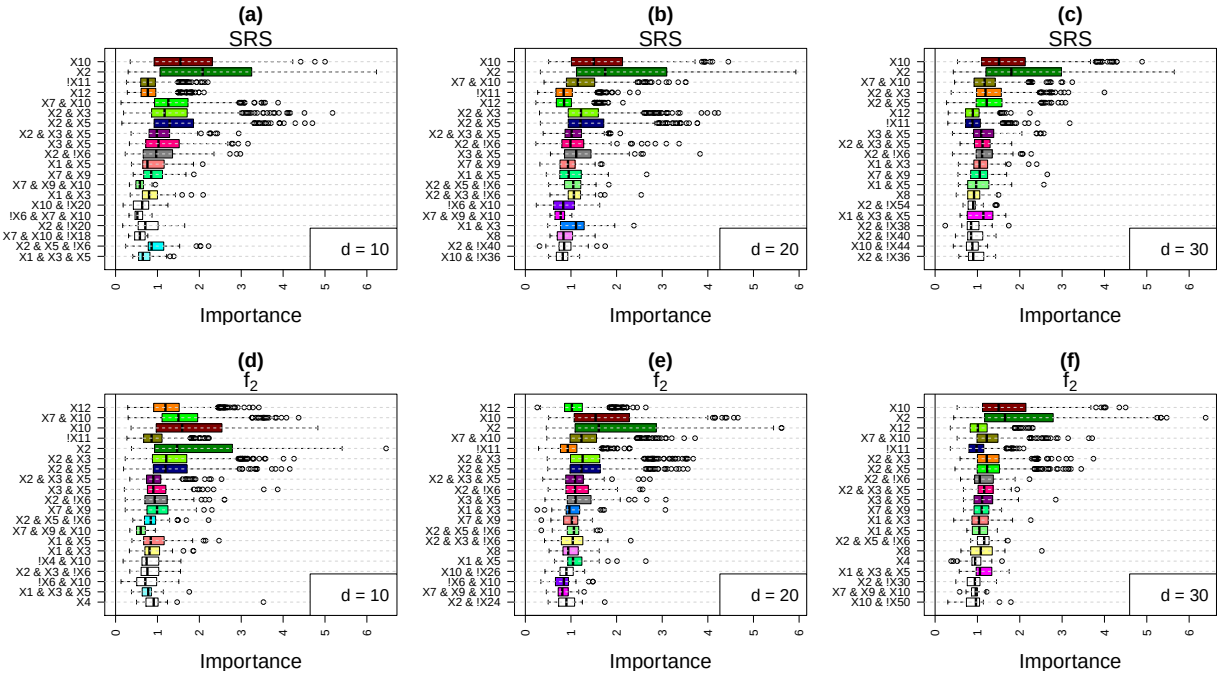

Abbildung 27: Boxplots der Wichtigkeitsmaße der 20 am häufigsten gefundenen Wechselwirkungen bei unterschiedlicher Auswahl einer Stichprobe bestehend aus 1% des Gesamtdatensatzes der Daten aus Simulation 3 für $d = 10, d = 20$ und $d = 30$, sortiert danach, wie häufig diese gefunden werden. Eingefärbt sind die Überschneidungen der Wechselwirkungen zwischen der einfachen Zufallsauswahl und der Auswahl proportional zu den Leverage Scores gewichtet durch f_2.

die Gewichtung mit f_2 bei $d = 10$ und $d = 20$ am häufigsten gefunden und im arithmetischen Mittel als wichtiger eingeschätzt als durch die SRS. Bei $d = 30$ befindet sich die Binärvariable bei der Gewichtung durch f_2 unter den drei am häufigsten gefundenen Wechselwirkungen, bei der SRS hingegen erst an der sechsten Stelle. Da es die wenigsten Unterschiede in den Wechselwirkungen gibt, die zu L_1 korrespondieren, legt dies die Vermutung nahe, dass die bessere Klassifikationsrate an den Trägern von L_3 liegt, die durch X_{12} mit einer wichtigen Binärvariable häufiger vertreten sind. Die Gewichtung durch f_4 führt bei dem Prozentsatz von 1% der Originaldaten ab $d = 30$ zu den besten Ergebnissen im Vergleich zu der SRS. Bei dieser Gewichtung wird die Binärvariable X_{12} ebenfalls häufiger gefunden und besitzt höhere Werte für das Wichtigkeitsmaß als bei der SRS. Dies legt ebenfalls die Vermutung nahe, dass der höhere Anteil in den Klassifikationsraten an den Trägern von L_3 liegt. Diese besitzen verglichen mit den anderen Beobachtungen höhere bzw. niedrigere LS-Werte (vgl. **Abbildung 23**). Dies könnte erklären, warum die Binärvariable häufiger gefunden wird und entsprechend die Beobachtungen aus dieser Gruppe häufiger richtig klassifiziert werden. Dies bedeutet, dass die Wahl der Beobachtungen proportional zu den LS in dieser Datensituation ebenfalls dazu führt, dass gewisse Untergruppen in den Beobachtungen besser gefunden werden.

Mit steigendem Prozentanteil der verwendeten Daten führt die Gewichtung durch

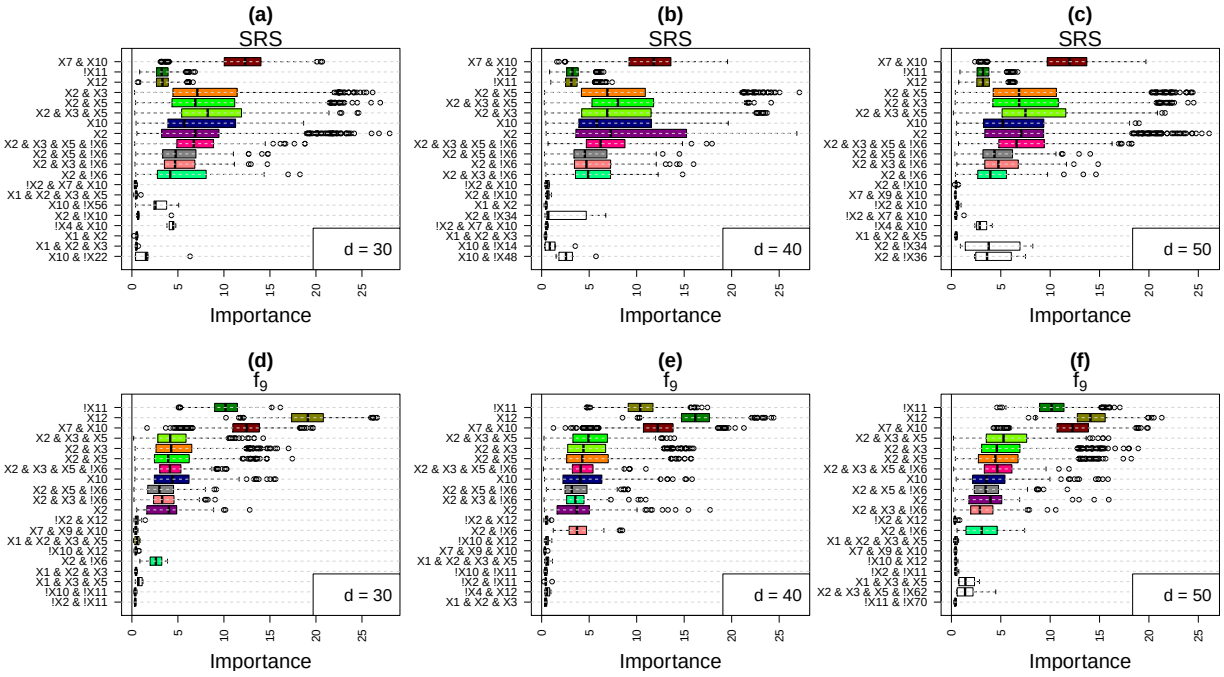

Abbildung 28: Boxplots der Wichtigkeitsmaße der 20 am häufigsten gefundenen Wechselwirkungen bei unterschiedlicher Auswahl einer Stichprobe bestehend aus 5% des Gesamtdatensatzes der Daten aus Simulation 3 für $d = 30, d = 40$ und $d = 50$, sortiert danach, wie häufig diese gefunden werden. Eingefärbt sind die Überschneidungen der Wechselwirkungen zwischen der einfachen Zufallsauswahl und der Auswahl proportional zu den Leverage Scores gewichtet durch f_9.

f_9 eindeutig zu den besten Ergebnissen. Die **Abbildung 28** zeigt erneut Boxplots der Wichtigkeitsmaße der 20 am häufigsten gefundenen Wechselwirkungen durch die zufällige Auswahl proportional zu den LS, dieses Mal gewichtet nach f_9 und mit 5% der Originaldaten. Insgesamt ist zu erkennen, dass nahezu alle gefundenen Wechselwirkungen bei der zufälligen Wahl der Beobachtungen proportional gewichtet nach f_9 nur aus nach Konstruktion wichtigen binären Einflussvariablen bestehen. Nur bei $d = 50$ treten in zwei der 20 am häufigsten gefunden Wechselwirkungen unwichtige Binärvariablen auf. Jedoch gibt es erneut Überschneidungen zwischen den wahren Wechselwirkungen. Bei der Wahl durch die SRS treten dagegen vermehrt unwichtige Binärvariablen in den Wechselwirkungen auf. Sehr stark ist zu erkennen, dass es Unterschiede in den Gewichtungen und der Häufigkeit der beiden Binärvariablen X_{11}^{C} und X_{12} gibt. Durch die Gewichtung mit f_9 sind diese beiden Binärvariablen als deutlich wichtiger gewertet und als die beiden am häufigsten gefundenen Wechselwirkungen vertreten. Dies unterstützt die Vermutung, dass durch die hohen LS sich in diesen Daten gewisse Untergruppen (in diesem Fall die Träger von L_3) finden lassen. Dies würde auch gut dazu passen, dass die Klassifikationsraten im mittleren Niveau höher liegen als bei der SRS, da die Träger von L_3 besser gefunden und entsprechend klassifiziert werden. Insgesamt deckt sich dieses Ergebnis mit der Simulation 1. In Simulation 1 werden durch die Gewichtung der hohen LS ebenfalls die beiden Binärvaria-

blen X_{11}^{C} und X_{12} häufiger gefunden und entsprechend liegen die Klassifikationsraten im mittleren Niveau über dem mittleren Niveau der SRS.

Die feste Auswahl durch f_9 führt zu sehr ähnlichen Ergebnissen und die am häufigsten gefundenen Wechselwirkungen überschneiden sich sehr stark und erhalten eine ähnliche Maße für die Wichtigkeit. Trotzdem tritt bei der festen Auswahl erneut vereinzelt auf, dass nach Konstruktion wichtige binäre Einflussvariablen in Kombination mit unwichtigen Binärvariablen gefunden werden. Allerdings deutlich seltener als bei Simulation 1 und mit einem geringeren Wichtigkeitsmaß. Es besteht somit fast kein Unterschied zwischen der zufälligen Auswahl proportional zu den LS und der festen Auswahl außer, dass das mittlere Niveau durch die zufällige Auswahl minimal höher liegt.

Wenn die Möglichkeit besteht, die Beobachtungen nach ihren LS auszuwählen, sollte dies nach Möglichkeiten auf jeden Fall anhand der hohen LS geschehen. Das Niveau und die Varianz der Klassifikationsraten liegt mit der richtigen Gewichtung auf jeden Fall nicht schlechter als bei der SRS und in vielen Datensituationen sogar besser. Im Kontext von medizinischen Daten ist jede Beobachtung deren Erkrankung zusätzlich erkannt wird ein Zugewinn, für den vergleichsweise geringen zeitlichen und rechnerischen Aufwand zum Bestimmen der LS.

5.2.5 Fazit und Anmerkungen zu dem Fall $n \geq d$

Durch die Simulationsstudie ergibt sich, dass es in fast jeder Datensituation eine Gewichtung der LS gibt, die dazu führt, dass wenn die Beobachtungen zu diesen Werten proportional in die Stichprobe aufgenommen werden, im arithmetischen Mittel ein Ergebnis der Klassifikationsraten erzielt wird, das wenigstens nicht schlechter bzw. in vielen Datensituationen besser ist als die einfache Zufallsauswahl.

Wenn nur wenige Daten verwendet werden bzw. es Fälle gibt, die nicht durch genetische Einflüsse erklärt sind, ist die Gewichtung der niedrigen LS durch f_8 am besten geeignet, die Beobachtungen für die Stichprobe auszuwählen. Für die beiden Simulationen ohne Rauschen (Simulation 1 und 3) ist es bei einem steigenden Stichprobenumfang hingegen bei fast allen Datensituationen von Vorteil, die Beobachtungen anhand der hohen LS mit der Gewichtung durch f_9 für die Stichprobe zu wählen. Demgegenüber ist es bei Rauschen in den Daten bei steigendem Prozentanteil besser, beide Randbereiche der LS zur Wahl der Beobachtungen zu verwenden. In **Tabelle 10** sind für jede Simulation für den entsprechenden Prozentanteil und Anzahl Variablen im Datensatz die Gewichtsfunktion abgetragen, die im arithmetischen Mittel zu den besten Klassifikationsraten führen. In der ersten Zeile ist die entsprechende Simulation angegeben, die zweite Zeile enthält die Prozentanteile P, die den Daten entnommen werden. In den restlichen Zeilen ist für die jeweilige Anzahl an Variablen d die Gewichtung der LS aufgetragen, die im arithmetischen Mittel bei der zufälligen Auswahl proportional zu den LS zu der besten Klassifikationsrate führt.

Bei einer geringen Anzahl an Variablen ist es in Simulation 1 und 3 am besten, die

Tabelle 10: Übersicht der im mittleren Niveau am besten geeigneten Gewichtungen für die zufällige Auswahl der Beobachtungen proportional zu den Leverage Scores bei den Daten der Simulationen 1 bis 3.

Sim.		1					2					3				
P		5%	10%	15%	20%	25%	5%	10%	15%	20%	25%	1%	2%	3%	4%	5%
	10	f_3	f_2	f_2	f_2	f_2	f_8	f_4	f_3	f_2	f_2	f_2	f_2	f_2	f_2	f_2
	20	f_8	f_2	f_2	f_9	f_9	f_8	f_8	f_5	f_1	f_1	f_2	f_2	f_2	f_9	f_9
d	30	f_8	f_{10}	f_1	f_9	f_9	f_8	f_8	f_5	f_5	f_1	f_4	f_2	f_9	f_9	f_9
	40	f_8	f_{11}	f_1	f_9	f_9	f_8	f_8	f_8	f_5	f_5	f_4	f_9	f_9	f_9	f_9
	50	f_8	f_6	f_5	f_9	f_9	f_8	f_8	f_8	f_5	f_5	f_4	f_9	f_9	f_9	f_9

LS direkt als Auswahlwahrscheinlichkeiten zu verwenden. Auffällig ist es, dass die mittleren LS in nur sehr vereinzelten Fällen zum besten mittleren Niveau führen (vgl. **Tabelle 10**). Es kann somit davon ausgegangen werden, dass die mittleren LS keinen positiven Einfluss auf die Wahl der Beobachtungen haben. Auch alle drei Bereiche durch die Gewichtsfunktion f_7 zu Berücksichtigen, bringt keine Verbesserung gegenüber dem Fall, nur die Randbereiche zu berücksichtigen.

Weiter ergibt sich, dass die zufällige Auswahl der Beobachtungen proportional zu den LS tendenziell besser ist als die feste Auswahl der Beobachtungen. Die feste Auswahl der Beobachtungen führt zwar teilweise zu leicht höheren Klassifikationsraten und einer Reduktion in der Varianz. Es zeigt sich jedoch, dass durch die feste Auswahl teilweise Verzerrungen in den Wechselwirkungen in dem Sinne auftreten, dass Binärvariablen als einflussreich erkannt werden, die nach Konstruktion keinen Einfluss besitzen und dass die feste Auswahl davon abhängt, welche Beobachtungen die entsprechenden LS-Werte aufweisen.

Als ein großer Punkt ergibt sich, dass die Gewichtung der LS dabei helfen kann, Untergruppen in den Fällen zu erkennen. In allen Datensituationen führt die (Mit-)Berücksichtigung der hohen LS bei der Wahl der Beobachtungen dazu, dass die einflussreichen Binärvariablen für die am schwächsten besetzte Untergruppe der Träger der Wechselwirkung L_3 besser aufgefunden werden. Es sollte dabei berücksichtigt werden, dass die Gewichtung der LS von der hier gewählten Datensituation abhängt. Wenn es andere Wechselwirkungen in den Daten gibt, könnte es sein, dass die Träger dieser Wechselwirkungen andere LS besitzen und entsprechend für das Auffinden dieser Untergruppen andere Gewichtungen der LS geeignet sind. Es empfiehlt sich, die Beobachtungen sowohl anhand der niedrigen als auch der hohen LS zufällig in die Stichprobe aufzunehmen und dieses Vorgehen mehrmals zu wiederholen, um potentiell geeignete Wechselwirkungen in den Daten aufzufinden.

Eine Vermutung ist es, dass es einen Unterschied macht, welches Verhältnis zwischen den Fällen und Kontrollen in der Stichprobe herrscht. Die einzige Einschränkung, die es für die zufällige Auswahl gibt, ist die, dass es mindestens einen Fall und mindestens ei-

ne Kontrolle geben muss, da es sonst nicht möglich ist, ein logisches Regressionsmodell anzupassen. Somit gibt es keine Einschränkungen zum Verhältnis. Dazu werden für diese Arbeit verschiedene Untersuchungen durchgeführt. In diesen Untersuchungen zeigt sich, dass das Verhältnis zwischen den beiden Gruppen in der Stichprobe keine Rolle spielt. Das Verhältnis auf gleich große Teile zu zwingen, bringt keine merkliche Verbesserung der Klassifikationsraten. Eine Untersuchung der besten Klassifikationsraten und dem herrschenden Verhältnis in diesen Stichproben gibt keinen Anlass zu der Vermutung, dass es vom Vorteil ist ein festgelegtes Verhältnis zu erzwingen.

Weiter führen eine Erhöhung der Anzahl der Wiederholungen B für den logicFS-Ansatz, eine Erhöhung der Anzahl der Bäume auf drei und die Erhöhung der maximalen Anzahl der Blätter bei dem logicFS-Datensatz nicht zu einer bemerkbaren Verbesserung der Klassifikationsraten. Eine Erhöhung der maximalen Anzahl an Blättern führt nur zu komplexeren Wechselwirkungen, welche jedoch nicht zu einer besseren Klassifikation beitragen. Aufgrund dieser Versuche bleibt es bei den vorher gewählten Anzahlen.

Alternativ zu den beiden Ansätzen, die Beobachtungen zufällig proportional zu den LS in die Stichprobe aufzunehmen und die Beobachtungen fest anhand ihrer LS zu wählen, wird ein dritter Ansatz verfolgt. Die Idee rührte daher, dass es am zeitaufwendigsten ist, ein logisches Regressionsmodell anzupassen, wohingegen es nur sehr wenig Zeit kostet, die Beobachtungen für die Stichprobe anhand der LS zu bestimmen. Dies führt zu dem Ansatz, den Index der entsprechenden Beobachtung mehrmals proportional zu den LS zu ziehen und die Beobachtungen zu verwenden, die am häufigsten gezogen werden. Dieser Ansatz soll einen Kompromiss zwischen der zufälligen und der festen Auswahl bilden, da es sehr wahrscheinlich ist, dass eine Beobachtung mit einer hohen Gewichtung in die Stichprobe gelangt aber nicht auf jeden Fall aufgenommen wird. Es zeigt sich jedoch, dass dieses Verfahren zu keiner merklichen Verbesserung der Klassifikationsraten führt, sondern im Gegenteil bei vielen Gewichtsfunktionen zu einer deutlichen Verschlechterung. Auf jeden Fall ist dieser Ansatz der festen Auswahl unterlegen, weshalb nicht weiter darauf eingegangen wird. Im nächsten Unterkapitel geht es um den zweiten Spezialfall bei SNP-Daten, wenn mehr Variablen d als Beobachtungen n vorliegen.

5.3 Der Fall $n < d$

Im Folgenden geht es um die zweite vorkommende Situation bei SNP-Daten. Wie bereits in **Kapitel 4** angemerkt, liegen in sogenannten genomweiten Assoziationsstudien häufig sehr hochdimensionierte Datensätze vor, in denen es deutlich mehr Variablen als Beobachtungen gibt ($d \gg n$). Da es für die Bestimmung der (C)LS der Matrix $\widetilde{X} \in \mathbb{R}^{n \times d}$ notwendig ist, dass $n \geq d$ ist (vgl. **Kapitel 5.1**) lassen sich anhand der (C)LS nicht mehr Beobachtungen selektieren, sondern es werden mithilfe der transponierten Matrix $\widetilde{X}^T \in \mathbb{R}^{d \times n}$ entsprechend Variablen durch die (C)LS selektiert. Dies stellt eine andere Herausforderung dar und entsprechend gibt es andere Bedingungen. Es wird zwar weiterhin die Klassifikationsrate zur Messung der Güte herangezogen. Dies ist jedoch nicht der Fokus dieser Untersuchung sondern eher wie und welche Variablen sich durch die LS und besonders die CLS selektieren lassen. Als Daten liegen die Datensätze der Simulation 4 und ein Teil der HapMap-Daten vor. Bei diesen Daten sollen die Variablen anhand der (C)LS selektiert werden. Einführend wird dies bei den Daten der Simulation 4 durchgeführt.

5.3.1 Auswertung von Simulation 4

Als erstes geht es in diesem Fall um die Daten der Simulation 4. Diese Datensätze besitzen jeweils die SNPs von 200 Beobachtungen, von denen 100 Fälle und 100 Kontrollen sind. Die Anzahl der SNPs d variiert: $d \in \{250, 300, 400, 500\}$. Jeder Datensatz besitzt somit d SNPs S_i, $i = 1, ..., d$. Für jedes d werden erneut jeweils 100 Datensätze simuliert. Somit besteht Simulation 4 aus insgesamt 400 Datensätzen. Jeder dieser Datensätze besitzt die drei Wechselwirkungen L_1, L_2 und L_3 aus **Gleichung (3)**, **(4)** und **(5)**. Entsprechend enthalten die Datensätze deutlich mehr, nach Konstruktion unwichtige Variablen. Um die Klassifikationsraten bestimmen zu können, müssen die Beobachtungen der Datensätze in Lern- und Testdaten eingeteilt werden. Mithilfe der Lerndaten werden die logischen Regressionsmodelle nach dem logicFS-Ansatz angepasst und die Beobachtungen der Testdaten durch die angepassten Modelle klassifiziert. Dazu werden bei jedem Datensatz die Hälfte der Fälle und die Hälfte der Kontrollen zufällig, mit der selben Wahrscheinlichkeit, in die Lernstichprobe aufgenommen und die restlichen Beobachtungen bilden die Teststichprobe. Somit bestehen sowohl die Lernstichprobe als auch die Teststichprobe aus jeweils 100 Beobachtungen, wobei jeweils 50 Fälle und 50 Kontrollen vertreten sind.

Einführend wird auf jedem der Datensätze der Simulation 4 der logicFS-Ansatz mit den beschriebenen Einstellungen angewandt. Interessanterweise ist der OOB-Fehler von der Anzahl der Variablen d abhängig. Mit steigender Anzahl der Variablen im Datensatz steigt der OOB-Fehler leicht an. Bei $d = 250$ liegt der OOB-Fehler im arithmetischen Mittel bei 1,8% und im Median bei 1,5%. Der maximale Wert beträgt 6%. Dahingegen liegt der OOB-Fehler bei $d = 500$ im arithmetischen Mittel bei 2,6% und im Median bei 2,5%. Das Maximum beträgt 7%. Entsprechend erschwert sich die Anpassung der

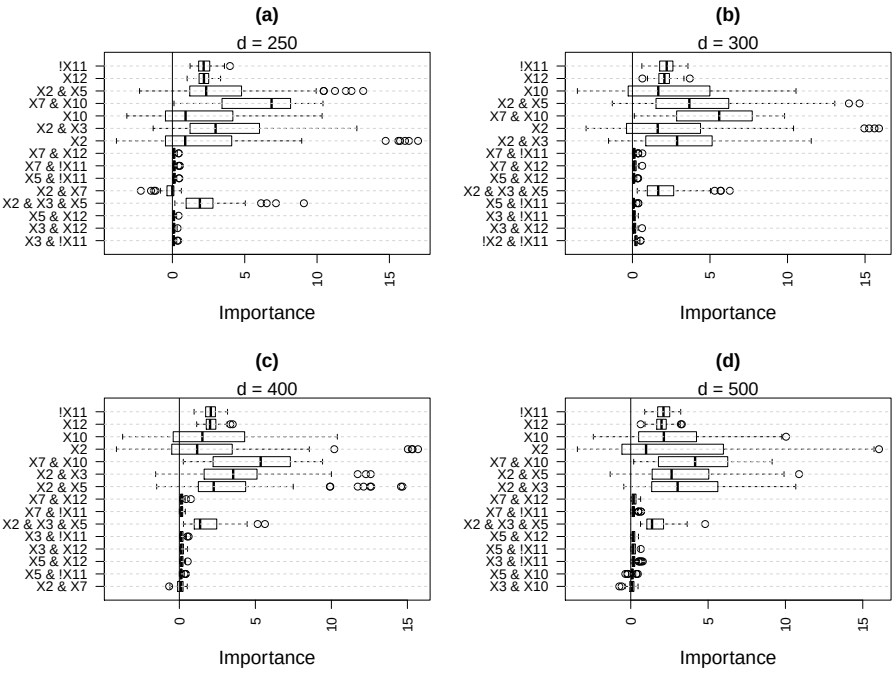

Abbildung 29: Boxplots der Wichtigkeitsmaße der gefundenen Wechselwirkungen beim Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf die vollen Datensätze der Simulation 4.

Modelle durch eine erhöhte Anzahl von unnötigen Variablen im Datensatz. Dabei ist weiter zu berücksichtigen, dass der OOB-Fehler gestiegen ist, obwohl es insgesamt weniger Beobachtungen zu klassifizieren gibt als in Simulation 1 und 3.

In **Abbildung 29** sind Boxplots der Wichtigkeitsmaße (aus **Gleichung (1)**) der 15 am häufigsten gefundenen Wechselwirkungen beim Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz mit den Daten der Simulation 4 dargestellt. Verglichen mit **Abbildung 13** und **Abbildung 24** sind Parallelen zu erkennen. Die beiden Binärvariablen X_{11}^{C} und X_{12} werden in jeder Datensituation am häufigsten gefunden. Im Kontrast werden bei den Daten der Simulation 4 die beiden Binärvariablen X_2 und X_{10} häufig alleine selektiert. Dies ist bei Simulation 1 und 3 unter den 15 am häufigsten gefundenen Wechselwirkungen nicht der Fall. Zusätzlich tritt bei den Daten der Simulation 4 die Binärvariable X_6^{C} in keiner der 15 am häufigsten gefundenen Wechselwirkungen auf. Die Komplexität der Wechselwirkungen ist in großen Teilen auf Zweifachwechselwirkungen beschränkt und ähneln denen aus Simulation 1. Ein Grund dafür kann die geringere Anzahl an Beobachtungen sein.

Grundsätzlich sollen die LS und nun auch die CLS herangezogen werden, um Variablen aus den Datensätzen zu entnehmen. Im Kontext der Selektion der Beobachtungen haben die CLS keinen Sinn für die Wahl der Beobachtungen ergeben. Hier spielen diese nun eine zentrale Rolle, vor allem die CLS der Variablen mit der abhängigen Variable. Wie in **Kapitel 5.1** beschrieben, wird zum Bestimmen der Hat-Matrix erneut der transformierte Vektor der Realisationen der abhängigen Variable mit der Datenmatrix vereinigt und

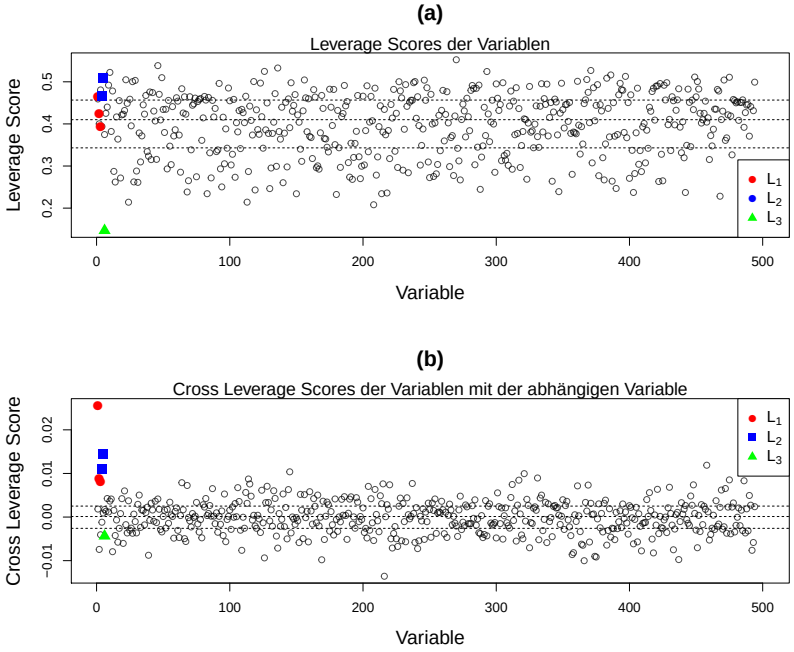

Abbildung 30: Darstellung der Leverage Scores der Variablen und der Cross Leverage Scores der Variablen mit der abhängigen Variable des ersten Datensatzes der Simulation 4 mit $d = 500$ SNPs: **(a)** Leverage Scores der Variablen, **(b)** Cross Leverage Scores der Variablen mit der abhängigen Variable. Die Werte der nach Konstruktion für den Krankheitsstatus erklärenden SNPs sind besonders hervorgehoben.

entsprechend mit der transformierten neu entstandenen Datenmatrix $\widetilde{X}^T \in \mathbb{R}^{(d+1)\times n}$ die QR-Zerlegung durchgeführt. Die resultierende Hat-Matrix $H \in \mathbb{R}^{(d+1)\times(d+1)}$ enthält dann die $(d+1)$-ste Zeile bzw. Spalte, die den Einfluss jeder Variable auf die abhängige Variable beschreiben. Dies sind die CLS der Variablen mit der abhängigen Variable. Der Eintrag $h_{(d+1),(d+1)}$ der Hat-Matrix enthält den Einfluss der abhängigen Variable auf sich selbst. Dieser Wert ist für die vorliegende Situation nicht informativ. Daher wird dieser Wert aus den (C)LS entfernt.

In **Abbildung 30 (a)** sind beispielhaft die LS der Variablen und in **Abbildung 30 (b)** die CLS der Variablen mit der abhängigen Variable des ersten Datensatzes der Simulation 4 mit $d = 500$ SNPs dargestellt. Die Werte der nach Konstruktion für den Krankheitsstatus verantwortlichen Variablen sind besonders hervorgehoben. Es zeigt sich in **Abbildung 30 (b)**, dass die CLS negative Werte annehmen. Die Gewichtsfunktionen (aus **Gleichung (9)** bis **(19)**) sind so konstruiert, dass die resultierenden Gewichtungen nicht negativ werden bzw. im Falle von f_2 werden die gewichteten Werte im Absolutbetrag verwendet. Weiter ist ein deutlicher Kontrast zwischen den LS und den CLS der Variablen mit der anhängigen Variable der nach Konstruktion wichtigen SNPs zu erkennen. Die CLS heben sich deutlich ab und besitzen mit die höchsten Werte. Nur die Einzelvariable S_6 von L_3 bildet eine Ausnahme dazu. Demgegenüber sind die LS der nach Konstruktion wichtigen SNPs viel stärker in den anderen Werten eingebettet. Die auffällige Ausnahme

bildet erneut die Einzelvariable S_6, die den allerniedrigsten LS Wert besitzt. Dargestellt ist ein Datensatz mit $d = 500$ SNPs, da sich der gezeigte Effekt bei einer großen Anzahl unwichtiger Variablen am besten erkennen lässt.

Diese Werte werden nun verwendet, um die Variablen aus den Datensätzen zu selektieren. Anders als in **Kapitel 5.2** wird nicht ein Prozentsatz der Variablen selektiert, sondern ein fester Anteil $d' \in \{10, 20, 30, 40, 50\}$. Dies geschieht, um vergleichen zu können welche Variablen fest durch die (C)LS selektiert werden, unabhängig von der Gesamtanzahl der Variablen d. Es werden sowohl die LS als auch die CLS der Variablen mit der abhängigen Variable verwendet, um die d' der d Variablen zu selektieren. Nach der Selektion der Variablen werden die d' selektierten SNPs der Lerndaten in Binärvariablen transformiert, so dass $2d'$ Binärvariablen vorliegen. Auf diese Daten wird der logicFS-Ansatz angewandt. Anschließend werden für die Testdaten dieselben d' selektierten Variablen aus dem Datensatz entnommen, in Binärvariablen transformiert, die Beobachtungen mit dem angepassten Modell per Bagging klassifiziert und mit dem wahren Krankheitsstatus verglichen. So entstehen erneut die Klassifikationsraten k.

Wie sich überlegen lässt, ist es nicht sinnvoll bzw. zielführend die Variablen zufällig proportional zu den (C)LS in die Stichprobe aufzunehmen. Zwar erhalten die wichtigen Einflussvariablen durch die Gewichtung höhere Auswahlwahrscheinlichkeiten, es ist jedoch immer noch vom Zufall abhängig, ob die nach Konstruktion wichtigen Variablen in die Stichprobe aufgenommen werden. Da es von Interesse ist, die wichtigen Variablen mit Sicherheit in die Stichprobe aufzunehmen, lässt sich vermuten, dass die feste Auswahl der Variablen anhand der (C)LS besser geeignet ist. Dies bestätigt sich in der Auswertung, weshalb für diese Arbeit nicht weiter auf die zufällige Auswahl der Variablen proportional zu den (C)LS eingegangen wird. Bei der festen Auswahl der Variablen werden die LS der Variablen und CLS der Variablen mit der abhängigen Variable auf die in **Kapitel 5.1** beschriebene Weise bestimmt, mit den Gewichtsfunktionen f_1, f_6, f_8, f_9 und f_{10} gewichtet und entsprechend die d' Variablen selektiert, welche nach der Gewichtung die höchsten Werte besitzen. Zum Vergleich werden d' Variablen per SRS gezogen und mit diesen Variablen die Klassifikation durchgeführt. Aufgrund des großen Umfangs der Daten wird die Wahl der Variablen per SRS pro Datensatz nur 10-mal durchgeführt. Es liegen pro Datensatz somit eine Klassifikationsrate für die feste Auswahl vor und 10 Klassifikationsraten für die SRS.

In **Abbildung 31** sind Boxplots der Klassifikationsraten bei der festen Auswahl der Variablen mit $d' = 10$ Variablen aus den Datensätzen der Simulation 4, mit $d = 250$ und $d = 500$ Variablen im Originaldatensatz dargestellt, ausgewählt mit den LS der Variablen und den CLS der Variablen mit der abhängigen Variable. Deutlich ist der Unterschied in den Klassifikationsraten zwischen der Auswahl der Variablen durch die LS und die CLS zu erkennen. Die Klassifikationsraten liegen bei der Auswahl durch die LS deutlich schlechter als bei der Wahl durch die CLS. Bei der Wahl der Variablen durch die CLS liegen die Klassifikationsraten bei den Gewichtsfunktionen deutlich höher, welche die hohen

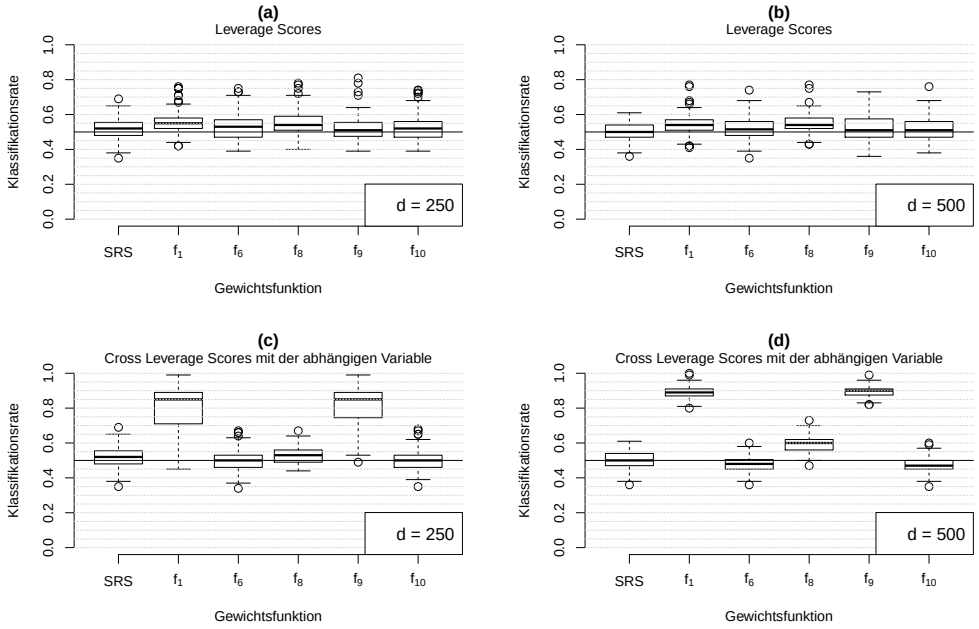

Abbildung 31: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz bei fester Selektion von $d' = 10$ Variablen der Daten der Simulation 4, mit $d = 250$ und $d = 500$ Variablen im Originaldatensatz, anhand der: **(a)**,**(b)** Leverage Scores der Variablen, **(c)**,**(d)** Cross Leverage Scores der Variablen mit der abhängigen Variable.

Werte mitberücksichtigen, was sich mit einer steigenden Anzahl d an Gesamtvariablen deutlich verbessert. Je mehr uninformative Variablen sich in den Datensätzen befinden, umso besser wird die Wahl der Variablen durch die CLS der Variablen mit der abhängigen Variable, wenn nicht gegen die hohen Werte diskriminiert wird. Die Wahl der Variablen anhand der mittleren bzw. niedrigen Werte ist entsprechend ungeeignet. Diese Erkenntnisse decken sich mit **Abbildung 30**, da die LS der wichtigen Einflussvariablen sich nicht besonders gegenüber den anderen Werten abheben (mit Ausnahme des SNPs S_6) und die CLS der Variablen mit der abhängigen Variable nahezu nur sehr hohe Werte besitzen. An der SRS ist zu erkennen, dass die zufällige Auswahl der Variablen nicht sinnvoll ist und die Klassifikationsraten entsprechend sehr nahe bei dem Richtwert von 0,5 liegen. Da die Klassifikationsraten bei der festen Auswahl eine relativ große Varianz besitzen, scheint es vom entsprechenden Datensatz abhängig zu sein, wie viele der wichtigen Einflussvariablen sich unter den $d' = 10$ der selektierten Variablen befinden. Offensichtlich ist es jedoch möglich mithilfe der CLS der Variablen mit der abhängigen Variable wichtige Einflussvariablen zu finden.

Eine Erhöhung der zur Verfügung stehenden Anzahl an Variablen d' verbessert bei einer großen Anzahl Variablen d im Datensatz, nur minimal die Klassifikationsrate, bei der Wahl der Variablen durch die CLS mit der abhängigen Variable. Dafür dauert es überproportional länger die Modelle anzupassen. Auf eine gesonderte Darstellung der Boxplots der Klassifikationsraten wird daher verzichtet. Das Anpassen der logischen Regressions-

modelle ist zeitlich weniger von der Anzahl der Beobachtungen n abhängig, als vielmehr von der Anzahl der Variablen d. Auch bei der Wahl der Variablen anhand der LS führt die Erhöhung der Anzahl der selektierter Variablen d', nur zu minimalen Verbesserungen der Klassifikationsraten und ist durch die Erhöhung der Rechenzeit nicht zu rechtfertigen.

Tabelle 11: Mittleres Niveau der Klassifikationsraten in Prozent durch das Anpassen logischer Regressionsmodelle nach dem logicFS-Ansatz für die feste Auswahl von d' Variablen anhand der Cross Leverage Scores der Variablen mit der abhängigen Variable für die Daten der Simulation 4.

d'	d	SRS	f_1	f_9
10	250	51,49	79,34	82,00
	300	50,37	87,30	88,04
	400	50,39	89,20	89,10
	500	50,26	89,11	89,03
20	250	57,02	82,98	86,14
	300	53,65	89,07	89,31
	400	58,07	89,21	89,10
	500	50,61	89,63	89,41
30	250	57,86	85,75	87,57
	300	53,48	89,81	89,68
	400	58,97	89,33	89,11
	500	51,68	89,81	89,44
40	250	57,04	88,13	88,53
	300	52,19	90,28	89,88
	400	59,84	89,80	89,33
	500	51,25	89,79	89,52
50	250	56,88	88,80	89,46
	300	53,38	90,41	90,01
	400	56,07	89,94	89,23
	500	50,40	90,01	89,66

In **Tabelle 11** ist das arithmetische Mittel der Klassifikationsraten bei fester Auswahl von d' Variablen anhand der CLS mit der abhängigen Variable bzw. per Wahl durch die SRS dargestellt. Die erste Spalte enthält die Anzahl der gewählten Variablen d', die zweite Spalte die Anzahl der Gesamtvariablen d im Datensatz. Die restlichen Spalten enthalten das mittlere Niveau der Klassifikationsraten bei fester Auswahl der Variablen anhand der CLS der Variablen mit der abhängigen Variable durch die beiden Gewichtsfunktionen f_1 und f_9 bzw. per SRS in Prozent. Es werden nur die Gewichtsfunktionen berücksichtigt, welche die hohen CLS mitberücksichtigen, da andere Bereiche der CLS für die Wahl der Variablen ungeeignet sind. Als erstes lässt sich festhalten, dass die SRS für die Wahl der Variablen absolut ungeeignet ist und das mittlere Niveau der Klassifikationsraten

Tabelle 12: Anzahl der Datensätze der jeweils 100 Datensätze der Simulation 4, in denen sich der entsprechende SNP S_i, $i = 1, ..., 6$, in den $d' = 10$ niedrigsten bzw. höchsten Leverage Scores der Variablen oder den Cross Leverage Scores der Variablen mit der abhängigen Variable wiederfinden lässt.

		f_9						f_8					
		SNP						SNP					
	d	S_1	S_2	S_3	S_4	S_5	S_6	S_1	S_2	S_3	S_4	S_5	S_6
LS	250	4	2	1	6	9	0	0	0	9	0	0	99
	300	4	1	0	1	4	0	0	0	7	0	0	100
	400	6	3	0	2	12	0	0	0	2	0	0	100
	500	17	3	0	2	6	0	0	0	2	0	0	100
CLS	250	78	33	22	40	68	8	0	0	1	0	0	15
	300	94	62	29	56	94	4	0	0	0	0	0	7
	400	100	76	41	66	100	1	0	0	0	0	0	4
	500	100	83	47	80	100	1	0	0	0	0	0	5

entsprechend sehr nahe an dem Richtwert von 50% liegt. Bei einer geringeren Anzahl an Gesamtvariablen d liegt das mittlere Niveau leicht höher, jedoch immer noch deutlich unter dem mittleren Niveau der Gewichtsfunktionen. Der Unterschied im mittleren Niveau zwischen der Gewichtung der CLS durch f_1 und f_9 ist verschwindend gering. Nur bei $d = 250$ und einer geringeren Anzahl selektierter Variablen d' liegt das mittlere Niveau durch die Gewichtung mit f_9 leicht über dem durch f_1. Auch lässt sich festhalten, dass eine Erhöhung von d' nur bei $d = 250$ zu einer Verbesserung in den Klassifikationsraten führt. In allen anderen Datensituationen ist das mittlere Niveau nahezu unverändert. Da eine Erhöhung von d' mit einer deutlichen Erhöhung der Rechenzeit einhergeht, ist davon abzuraten, eine zu große Anzahl d' zu wählen. Erneut zeigt sich, dass die LS bzw. in diesem Fall die CLS in den Situationen besonders geeignet sind, wenn viele uninformative Informationen in den Daten vorhanden sind, was in dieser Situation durch viele unwichtige Variablen repräsentiert wird.

Da sich die wichtigen Einflussvariablen in den CLS der Variablen mit der abhängigen Variable wiederfinden lassen, ist es von Interesse, zu untersuchen, welche dieser Variablen sich wie häufig wiederfinden. Wie in **Abbildung 30 (b)** zu erkennen ist, lassen sich die meisten Variablen in den hohen Werten der CLS mit der abhängigen Variable wiederfinden bzw. im Falle von S_6 in **Abbildung 30 (a)** zu erkennen, in den niedrigen Werten der LS der Variablen. Repräsentiert sind die hohen Werte durch die Gewichtsfunktion f_9, da diese Gewichtung dazu führt, dass bei der festen Auswahl die Variablen mit den höchsten (C)LS Werten selektiert werden. Umgekehrt ist die Gewichtsfunktion f_8 dafür verantwortlich, dass die Variablen mit den niedrigsten Werten selektiert werden.

In **Tabelle 12** sind die Häufigkeiten dargestellt, die sich der SNP S_i, $i = 1, ..., 6$, in den $d' = 10$ niedrigsten oder höchsten LS der Variablen bzw. CLS der Variablen mit der abhängigen Variable in den jeweils 100 Datensätzen der Simulation 4 wiederfinden lässt.

Ein Eintrag in der Tabelle bedeutet, dass der entsprechende SNP sich diese Anzahl in den 100 Datensätzen, in den entsprechenden (C)LS wiederfinden lässt. Da es sich um 100 Datensätze insgesamt handelt, kann die Anzahl als Prozentzahl interpretiert werden. Die Tabelle ist aufgeteilt nach den niedrigen und den hohen Bereichen der LS bzw. CLS mit der abhängigen Variable, repräsentiert durch die Gewichtsfunktionen f_8 bzw. f_9. Die erste Spalte enthält die Information, ob es sich um die LS der Variablen oder die CLS der Variablen mit der abhängigen Variable handelt. Die zweite Spalte enthält die Gesamtanzahl der Variablen d im Datensatz. In den nächsten Spalten ist für jeden SNP aufgetragen, wie häufig sich dieser in den $d' = 10$ niedrigsten bzw. höchsten LS oder CLS befindet. Wie in **Abbildung 30 (a)** ganz eindeutig zu erkennen, besitzt S_6 den allerniedrigsten LS Wert. Dies gilt für nahezu alle Datensätze der Simulation 4. Bei mindestens 90% aller Datensätze besitzt diese Variable den niedrigsten LS-Wert und wie in **Tabelle 12** zu erkennen, befindet sich diese Variable bei fast allen Datensätzen unter den 10 Variablen mit den niedrigsten LS-Wert. Weiter ist anzumerken, dass die restlichen Variablen sich nur sehr selten mit den höchsten LS wiederfinden lassen und nahezu gar nicht mit den niedrigen LS. Im starken Kontrast dazu lassen sich S_1 und S_5 ab $d = 300$ bei fast allen Datensätzen in den hohen CLS der Variablen mit der abhängigen Variable wiederfinden. Für die Variablen S_2 und S_4 steigt die Anzahl mit steigendem d ebenfalls an und diese Variablen lassen sich entsprechend bei einem Großteil der Datensätze in den hohen CLS der Variablen mit der abhängigen Variable wiederfinden. Am schwersten ist die Variable S_3 zu finden und erst ab $d = 500$ befindet sich diese Variable bei etwas unter der Hälfte der Datensätze, unter den 10 Variablen mit den höchsten CLS der Variablen mit der abhängigen Variable. Dies erklärt, warum die feste Auswahl der Variablen anhand der hohen CLS der Variablen mit der abhängigen Variable zu so guten Klassifikationsraten führt. Mit den beiden Variablen S_1 und S_5 sind zwei der Untergruppen der Fälle, mit mindestens einer nach Konstruktion wichtigen Variable, in nahezu jeder Datensituation vertreten. Dies führt dazu, dass mindestens eine wichtige Variable für den Krankheitsstatus dieser beiden Gruppen erkannt wird und die Klassifikation anhand dieser Variable durchgeführt werden kann. Entsprechend werden die meisten Fälle aus diesen Gruppen richtig klassifiziert. Die Träger von L_3 dürften der Grund dafür sein, dass die Klassifikationsraten im arithmetischen Mittel unter 100% liegen, da diese Untergruppe durch die CLS der Variablen mit der abhängigen Variable nur sehr selten gefunden wird.

Die hohen CLS der Variablen mit der abhängigen Variable sind auf jeden Fall dazu geeignet, potentiell wichtige Einflussvariablen in den SNP-Daten zu identifizieren und sollten in einer Variablenselektion unbedingt berücksichtigt werden. Da sich die Variable S_6 nahezu immer in den niedrigen LS wiederfindet, könnten sich Entscheidungsregeln der Art konstruieren lassen, dass immer eine gewisse Anzahl der Variablen mit hohen CLS der Variablen mit der abhängigen Variable zu berücksichtigen sind und eine gewisse Anzahl Variablen mit niedrigen LS. Aus zeitlichen Gründen wird eine solche Analyse nicht mehr durchgeführt und kann Bestandteil zukünftiger Forschung sein.

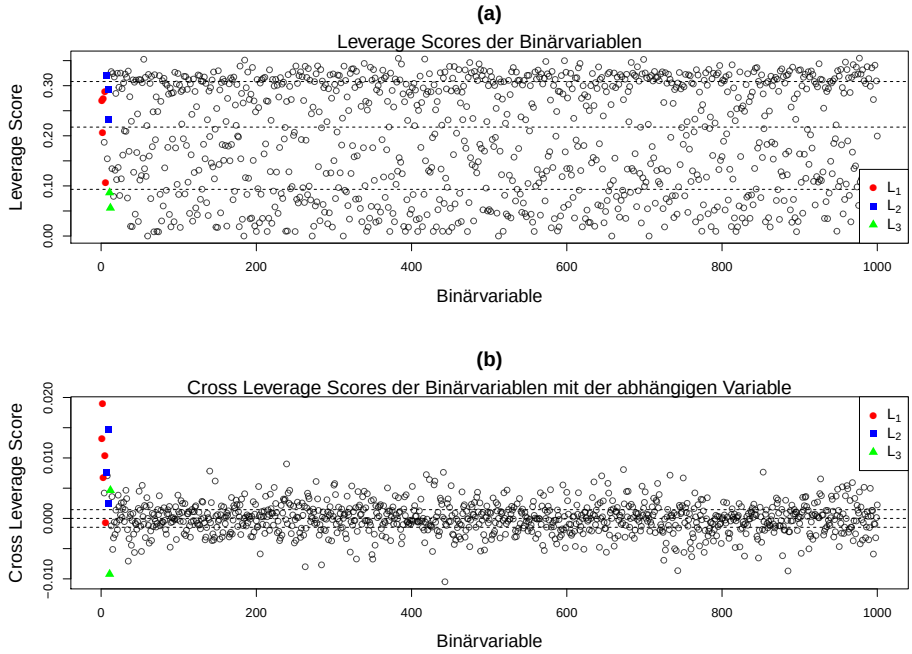

Abbildung 32: Darstellung der Leverage Scores der Binärvariablen und Cross Leverage Scores der Binärvariablen mit der abhängigen Variable der $2d = 1000$ Binärvariablen des ersten Datensatzes der Simulation 4 mit $d = 500$ SNPs: **(a)** Leverage Scores der Binärvariablen, **(b)** Cross Leverage Scores der Binärvariablen mit der abhängigen Variable. Die Werte der nach Konstruktion für den Krankheitsstatus erklärenden Binärvariablen sind farblich besonders hervorgehoben.

Anstatt die SNPs direkt mithilfe der (C)LS zu selektieren, ist es in diesem Fall möglich, die binärcodierten Variablen zur Variablenselektion heranzuziehen. In diesem Fall sind die einflussgebenden Variablen nicht mehr die ersten sechs Variablen, sondern befinden sich unter den ersten zwölf Binärvariablen, wobei die Binärvariablen X_4 und X_8 nach Konstruktion keinen Einfluss besitzen (siehe **Gleichung (6)** bis **(8)**). Um die (C)LS auf diese Weise zu bestimmen, werden die d SNPs vor der Vereinigung des Datensatzes mit den transformierten Realisationen der abhängigen Variable in Binärvariablen codiert (anstelle danach) und die Hat-Matrix $H \in \mathbb{R}^{(2d+1)\times(2d+1)}$ mit Hilfe dieser so entstandenen $2d$ Binärvariablen bestimmt. In **Abbildung 32** sind die $2d = 1000$ LS der Binärvariablen bzw. die CLS der Binärvariablen mit der abhängigen Variable des ersten Datensatzes aus Simulation 4 mit $d = 500$ SNPs dargestellt. Im Vergleich zu **Abbildung 30** ist zu erkennen, dass sich die nach Konstruktion wichtigen Binärvariablen nicht in den Randbereichen der LS der Binärvariablen wiederfinden lassen und die Werte entsprechend sehr verteilt im Zentrum der Werte liegen. Demgegenüber ist in **Abbildung 32 (b)** zu erkennen, dass erneut viele der nach Konstruktion wichtigen Binärvariablen einen hohen CLS-Wert mit der abhängigen Variable besitzen und vereinzelt (zu erkennen an der Binärvariable X_{11}) einen niedrigen CLS-Wert mit der abhängigen Variable.

Die Häufigkeit mit der sich jede Binärvariable X_i, $i = 1, 2, 3, 5, 6, 7, 9, 10, 11$, in den

Tabelle 13: Anzahl der Datensätze der jeweils 100 Datensätzen der Simulation 4, in denen sich die entsprechende Binärvariable X_i, $i = 1, 2, 3, 5, 6, 7, 9, 10, 11$, in den $2d' = 20$ niedrigsten bzw. höchsten Cross Leverage Scores der Binärvariablen mit der abhängigen Variable wiederfinden lässt.

	d	X_1	X_2	X_3	X_5	X_6	X_7	X_9	X_{10}	X_{11}	X_{12}
f_9	250	50	100	69	73	1	65	53	100	0	93
	300	66	100	75	75	0	72	59	100	0	92
	400	81	100	77	83	0	67	59	100	0	76
	500	79	100	88	92	0	78	66	100	0	61
f_8	250	0	0	0	0	12	0	0	0	100	0
	300	0	0	0	0	10	0	0	0	95	0
	400	0	0	0	0	10	0	0	0	90	0
	500	0	0	0	0	6	0	0	0	78	0

$2d' = 20$ niedrigsten bzw. höchsten CLS der Binärvariablen mit der abhängigen Variable wiederfinden lässt, ist in **Tabelle 13** abgetragen. Da die Gesamtanzahl der Variablen d sich verdoppelt, wird die Anzahl der betrachteten Variablen d' ebenfalls verdoppelt. Durch den Umstand, dass die LS der Binärvariablen zum Finden der nach Konstruktion wichtigen Binärvariablen ungeeignet sind, wird auf eine gesonderte Darstellung verzichtet. Auffällig ist der Unterschied zu **Tabelle 12**. Die zu S_6 korrespondierenden Binärvariablen X_{11} und X_{12} finden sich nicht in den niedrigen LS der Binärvariablen sondern in den niedrigen bzw. hohen CLS der Binärvariablen mit der abhängigen Variable. Interessant ist festzuhalten, dass die beiden Binärvariablen X_6 und X_{11} (und eigentlich auch X_{12}) als ihr Kompliment in den wahren Wechselwirkungen vertreten sind und diese beiden Binärvariablen sich eher in den niedrigen CLS der Binärvariablen mit der abhängigen Variable finden. Es scheint so zu sein, dass die nach Konstruktion wichtigen Binärvariablen eine hohe Hebelwirkung auf die abhängige Variable besitzen, wohingegen die Komplemente der Binärvariablen eine niedrige Hebelwirkung auf die abhängige Variable besitzen. Vergleichen mit **Abbildung 29** fällt auf, dass die Binärvariablen die in der Anpassung der logischen Regressionsmodelle nach dem logicFS-Ansatz mit den vollen Datensätzen am häufigsten als einzelne Binärvariable gefunden werden, sich am häufigsten in den CLS der Binärvariablen mit der abhängigen Variable finden lassen: X_2, X_{10}, X_{11}^{C} und X_{12}, wobei X_{12} mit steigender Anzahl an Gesamtvariablen d weniger häufig in den $2d' = 20$ höchsten CLS der Binärvariablen mit der abhängigen Variable vertreten ist. Die übrigen Binärvariablen, die in erster Linie in Kombination mit anderen Binärvariablen gefunden werden, treten ebenfalls seltener in den $2d' = 20$ höchsten CLS der Binärvariablen mit der abhängigen Variable auf. Sehr interessant ist es, dass die Binärvariable X_6 bzw. X_6^{C} weder in den Modellen auf den gesamten Datensätzen unter den 15 häufigsten Wechselwirkungen vertreten ist (vgl. **Abbildung 29**) noch häufig in den hohen bzw. niedrigen CLS der Binärvariablen mit der abhängigen Variable. Für eine angestrebte Variablenselektion sollten die CLS der Binärvariablen mit der abhängigen Variable ebenfalls wenigstens

Tabelle 14: Anzahl der Datensätze der jeweils 100 Datensätzen der Simulation 4, in denen sich ein SNP S_i, $i = 1, .., 6$, in den $d' = 10$ niedrigsten bzw. höchsten Cross Leverage Scores eines anderen SNPs S_j, $j = 1, ...6$, in den Daten wiederfinden lässt.

		$d = 250$						$d = 500$					
		S_1	S_2	S_3	S_4	S_5	S_6	S_1	S_2	S_3	S_4	S_5	S_6
f_9	S_1	100	54	25	0	1	15	100	89	54	0	0	2
	S_2	51	100	17	1	0	6	89	100	21	0	0	1
	S_3	18	10	100	2	1	12	55	26	100	0	0	1
	S_4	0	1	3	100	40	9	0	0	0	100	72	0
	S_5	0	0	3	44	100	14	0	0	0	73	100	1
	S_6	7	2	6	5	7	100	6	3	1	4	10	100
f_8	S_1	0	0	2	22	37	1	0	0	0	52	85	2
	S_2	0	0	1	12	12	7	0	0	1	21	64	0
	S_3	2	1	0	5	5	7	0	1	0	11	26	0
	S_4	21	11	8	0	0	6	55	23	9	0	0	1
	S_5	36	15	6	0	0	8	85	63	27	0	0	1
	S_6	1	5	4	4	3	0	7	2	4	4	4	0

mitberücksichtigt werden, da sich in dieser Datensituation einige der nach Konstruktion wichtigen Binärvariablen bereits erkennen lassen.

Eine weitere interessante Fragestellung ist, ob sich eine Variable aus einer Wechselwirkung eventuell in den CLS der anderen, an der selben Wechselwirkung beteiligten Variablen erkennen lässt. Es lässt sich vermuten, dass die Variablen, die in einer Wechselwirkung miteinander stehen, diese Beziehung durch ihre CLS ausdrücken. Sollte dies der Fall sein, wäre es möglich, gezielt nach Wechselwirkungen in den CLS zu suchen.

In **Tabelle 14** sind die Anzahl der Datensätze aufgetragen, in denen sich einer der nach Konstruktion einflussreichen SNPs S_i, $i = 1, ..., 6$, in den $d' = 10$ niedrigsten oder höchsten CLS eines anderen nach Konstruktion einflussreichen SNPs S_j, $j = 1, ..., 6$, wiederfinden lässt. Getrennt ist die Tabelle nach der Anzahl Gesamtvariablen $d = 250$ und $d = 500$ um den Einfluss der steigenden Anzahl Gesamtvariablen d zu verdeutlichen. Die Spalten stehen dafür, dass es sich um die CLS des entsprechenden SNP handelt. In den Zeilen steht die Anzahl der Datensätze in denen sich der entsprechende SNP in den $d' = 10$ höchsten oder niedrigsten CLS des SNPs, um dessen Spalte es sich handelt, finden lässt, repräsentiert durch die Gewichtsfunktionen f_9 und f_8. Dabei fällt auf, dass jeder SNP mit sich selbst den höchsten CLS Wert besitzt und entsprechend in jedem der 100 Datensätze der Simulation 4 in den höchsten CLS auftritt. Als zweites ist anzumerken, dass die Anzahl nicht symmetrisch ist. Die SNPs die sich zusammen in einer Wechselwirkung befinden, finden sich in erster Linie unter den $d' = 10$ höchsten CLS. In den niedrigsten CLS befinden sich in erster Linie SNPs anderer Wechselwirkungen. Mit einer steigenden Anzahl Gesamtvariablen im Datensatz nimmt die Anzahl der Datensätze zu, in denen sich die SNPs in den CLS der anderen an der selben Wechselwirkung beteiligten SNPs wiederfinden lassen. Erneut zeigt sich, dass es einen Zugewinn an Information durch die

CLS gibt, wenn viele uninformative Variablen in den Datensätzen vorliegen. Die stärksten Zusammenhänge besitzen jeweils die beiden Variablen S_1 und S_2 aus der Wechselwirkung L_1 und die beiden Variablen S_4 und S_5 aus der Wechselwirkung L_2. Bei $d = 500$ finden sich diese Variablen bei einem Großteil der Datensätze in den CLS der jeweils anderen Variable. Die Variable S_6 findet sich nur sehr vereinzelt in den niedrigsten bzw. höchsten CLS einer anderen Variable, interessanterweise öfter bei weniger Gesamtvariablen. Eventuell ist es zum Auffinden einer Einzelvariable von Vorteil, wenn es insgesamt weniger unwichtige Informationen in den Datensätzen gibt.

In **Tabelle B1** im Anhang dieser Arbeit ist angelehnt an **Tabelle 14** die entsprechende Tabelle der Binärvariablen dargestellt. Es zeigt sich, dass der Zusammenhang der Binärvariablen in Form der CLS teilweise stärker ist als bei den SNPs. Bereits bei $d = 250$ finden sich die Binärvariablen X_1 und X_2 bzw. X_9 und X_{10} bei fast jedem Datensatz unter den $2d' = 20$ höchsten CLS der jeweils anderen Binärvariable. Aus der Wechselwirkung $L_{1_{DNF}}$ finden sich die Binärvariablen X_3 und X_5 ebenfalls in relativ vielen der Datensätzen in den höchsten CLS der Binärvariable X_2. Auch in dieser Situation sind es die Binärvariablen, die zu L_3 bzw. $L_{3_{DNF}}$ korrespondieren, die nur wenig bzw. gar nicht in den hohen oder niedrigen CLS der anderen Binärvariablen vertreten sind. Die Ausnahme dazu bildet erneut die Binärvariable X_2, bei der X_{11} und X_{12} relativ häufig unter den 20 niedrigsten oder höchsten CLS vertreten sind. Anders als bei den (C)LS-Werten der SNPs ist es dabei jedoch nicht der Fall, dass das Komplement einer Binärvariable sich in den niedrigen Werten der CLS findet, da die Binärvariable X_{11} bei der Anpassung der Modelle auf den vollen Datensätzen immer durch das Komplement X_{11}^{C} vertreten ist, sich in den CLS der Binärvariable X_2 jedoch unter den hohen Werten findet. Auffällig ist noch, dass die beiden Binärvariablen X_2 und X_{10} sich bei fast allen Datensätzen unter den $2d' = 20$ niedrigsten CLS der jeweils anderen Binärvariable finden, obwohl diese aus unterschiedlichen Wechselwirkungen stammen. Wie in **Abbildung 29** zu erkennen, wird die Binärvariable X_{10} zwar vermehrt in Kombination mit anderen Binärvariablen gefunden, jedoch nicht mit X_2. Diese beiden Binärvariablen sind dafür als einzelne Binärvariablen unter den am häufigsten gefundenen Wechselwirkungen vertreten und besitzen mit am häufigsten negative Werte in den Wichtigkeitsmaßen.

Die nach Konstruktion wichtigen Variablen bzw. Binärvariablen lassen sich teilweise in den CLS der jeweils anderen Variablen wiederfinden. Wenn es vorher jedoch nicht bekannt ist welche Variablen einen Einfluss auf den Krankheitsstatus besitzen, gestaltet sich die gezielte Suche nach den Variablen als sehr schwierig. Der Grund dafür ist, dass viele andere und uninformative Variablen ebenfalls hohe CLS mit den nach Konstruktion wichtigen SNPs besitzen und sich die wichtigen Variablen in ihren Werten nicht so eindeutig hervorheben, wie es in den CLS der Variablen mit der abhängigen Variable der Fall ist. Als Empfehlung kann nur gegeben werden, dass wenn ein SNP bekannt dafür ist an dem Auslösen einer Krankheit beteiligt zu sein, es potentiell von Vorteil ist, die anderen SNPs mit hohen CLS Werten genauer zu betrachten.

Im nächsten Unterkapitel geht es um die HapMap-Daten, bei denen ebenfalls Variablen mit Hilfe der (C)LS selektiert werden sollen.

5.3.2 Auswertung der HapMap-Daten

Als letzter und am höchsten dimensionierter Datensatz liegt ein Teil der HapMap-Daten vor (siehe **Kapitel 4.1**). Von Interesse ist, ob sich wichtige Variablen ebenfalls durch die (C)LS selektieren lassen. Bei den HapMap-Daten ist es unmöglich auf dem gesamten Datensatz den logicFS-Ansatz anzuwenden.

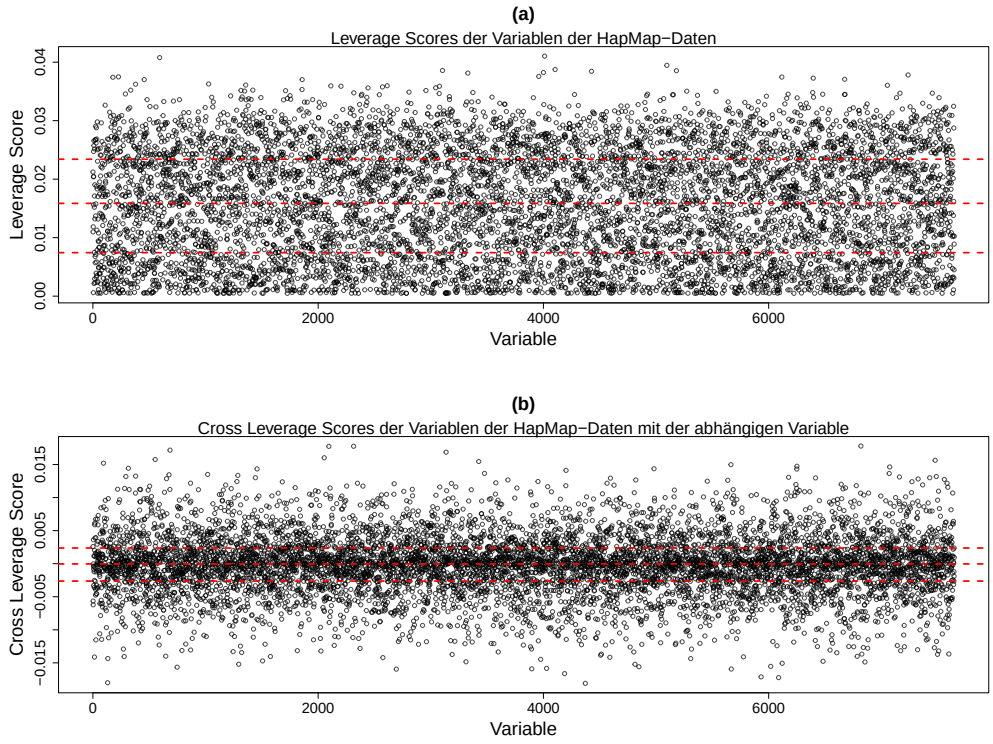

Abbildung 33: Darstellung der Leverage Scores der Variablen bzw. Cross Leverage Scores der Variablen mit der abhängigen Variable des HapMap-Datensatzes: **(a)** Leverage Scores der Variablen, **(b)** Cross Leverage Scores der Variablen mit der abhängigen Variable.

In **Abbildung 33 (a)** sind die LS der Variablen und in **Abbildung 33 (b)** die CLS der Variablen mit der abhängigen Variable der HapMap-Daten dargestellt. Die Hat-Matrix bzw. entsprechend die (C)LS werden wie bei Simulation 4 bestimmt (siehe **Kapitel 5.3.1**). Als rote horizontale Linie sind jeweils das 0,25-, das 0,5- und das 0,75-Quantil der Werte dargestellt. Das Zentrum der Werte der CLS der Variablen mit der abhängigen Variable ist relativ stark geballt. Es gibt jedoch einige Werte, die stark nach unten bzw. oben abweichen. Bei den LS der Variablen ist das Zentrum der Werte viel breiter, trotzdem gibt es einige Variablen mit sehr hohen LS-Werten und sehr viele Variablen mit LS-Werten nahe an 0.

Aus diesem Datensatz sollen nun ebenfalls Variablen mithilfe der (C)LS entnommen werden. Um die Klassifikationsraten bestimmen zu können, werden die $n = 120$ Beobach-

tungen zu Beginn zufällig in gleichgroße Lern- und Testdaten eingeteilt, mit der selben Anzahl von Fällen und Kontrollen. Die Klassifikation erfolgt nun nicht mehr in die Kategorien krank und gesund, sondern in die ethnische Zugehörigkeit der entsprechenden Beobachtung (vgl. **Kapitel 4.1**). Dabei wird die ethnische Zugehörigkeit zu den Europäern mit 1 codiert und die ethnische Zugehörigkeit zu den westafrikanischen Yoruba mit 0. Es werden zufällig jeweils 30 der Europäer und 30 der Yoruba in die Lern- bzw. Teststichprobe aufgenommen.

Die in **Abbildung 33** dargestellten (C)LS werden mit den Gewichtsfunktionen f_1, f_6, f_8, f_9 und f_{10} gewichtet und eine Anzahl $d' \in \{10, 20, 30, 40, 50\}$ von Variablen fest anhand der nach der Gewichtung am höchsten entstandenen Werte entnommen. Diese Variablen werden verwendet, um die logischen Regressionsmodelle nach dem logicFS-Ansatz anzupassen und die Beobachtungen der Teststichprobe zu klassifizieren. Diese Klassifikation wird mit der wahren ethnischen Zugehörigkeit der Beobachtung verglichen und so entstehen erneut die Klassifikationsraten k.

Es zeigt sich, dass es sehr einfach ist, bei den HapMap-Daten eine Klassifikation durchzuführen. Bereits $d' = 10$ der Variablen sind genug, um bei der festen Auswahl der Variablen anhand der CLS mit der abhängigen Variable eine Klassifikationsrate von 100% zu erreichen. Dabei spielt es keine Rolle, ob die niedrigen, hohen oder eine Kombination der beiden Bereiche der CLS durch die Gewichtsfunktionen f_8, f_9 und f_1 verwendet werden. Nur gegen diese Bereiche zu diskriminieren (mit den Gewichtsfunktionen f_6 und f_{10}) führt zu Klassifikationsraten von nahezu 50%. Sogar bei der Auswahl der Variablen durch die SRS ergibt sich im arithmetischem Mittel ein Wert der Klassifikationsraten von 85,77%. Dabei wird die Ziehung der Variablen per SRS 100-mal. Da es bei einer komplett zufälligen Auswahl von $d' = 10$ Variablen schon möglich ist, so hohe Klassifikationsraten zu erzielen, lässt sich vermuten, dass viele der SNPs in dem Datensatz den Unterschied zwischen der ethnischen Zugehörigkeit erklären. Jedoch ist es durch die Gewichtung der CLS mit der abhängigen Variable möglich, gezielt SNPs zu entnehmen, die diesen Unterschied mit einer festen Sicherheit erklären. Interessanterweise sind es bei Wahl von $d' = 10$ Variablen anhand der LS der Variablen, die Gewichtsfunktionen f_6 und f_{10} die zu einer Klassifikationsrate von 100% führen. Dahingegen liegt die Klassifikationsrate bei der Wahl der Variablen durch f_9 bei 68,33%, bei der Wahl durch f_1 bei 55% und mit der Wahl durch f_8 ist keine Klassifikation möglich, da die durch diese Gewichtung selektierten SNPs nicht alle Ausprägungen besitzen. Es scheint somit der Fall zu sein, dass bei einem so hochdimensionierten Datensatz die LS der Variablen und die CLS der Variablen mit der abhängigen Variable sich konträr bei der Wahl der Variablen verhalten.

Eine Erhöhung der selektierten Variablen d' führt bei der Wahl der Variablen per SRS zu einer Verbesserung der Klassifikationsraten. Mit $d' = 50$ liegt die Klassifikationsrate im arithmetischen Mittel bei 94,67% bei 100 wiederholten Ziehungen. Bei den anderen Gewichtungen führt eine Erhöhung der Anzahl der Variablen d' in der Stichprobe erneut nicht zu einer Verbesserung der Klassifikationsraten.

Tabelle 15: Die $d' = 10$ SNPs der HapMap-Daten, die durch die Gewichtsfunktionen f_1, f_8 und f_9 anhand der festen Auswahl der Variablen durch die Cross Leverage Scores der Variablen mit der abhängigen Variable selektiert werden. Diejenigen SNPs, die durch mehr als eine Gewichtsfunktion selektiert werden, sind besonders hervorgehoben.

	S_1	S_2	S_3	S_4	S_5
f_1	**rs6670842**	**rs311992**	**rs6814827**	**rs1485768**	**rs7752055**
f_8	**rs6670842**	rs13420968	rs809039	rs10504132	rs7833862
f_9	rs3767067	rs619228	**rs311992**	rs10805068	**rs6814827**
	S_6	S_7	S_8	S_9	S_{10}
f_1	**rs368297**	**rs10868791**	**rs1373013**	**rs2370893**	**rs9909962**
f_8	**rs368297**	**rs10868791**	rs9534610	**rs1373013**	**rs2370893**
f_9	**rs1485768**	**rs7752055**	rs2034510	**rs9909962**	rs2833795

In **Tabelle 15** sind die $d' = 10$ SNPs der HapMap-Daten dargestellt, die durch die feste Auswahl der Variablen durch die CLS der Variablen mit der abhängigen Variable durch die Gewichtung der Gewichtsfunktionen f_1, f_8 und f_9 selektiert werden. Diejenigen SNPs, die durch mehr als eine Gewichtsfunktion selektiert werden, sind besonders hervorgehoben. Deutlich ist zu erkennen, dass es viele Überschneidungen zwischen der Selektion durch f_1 und den anderen beiden Gewichtsfunktionen gibt. Anhand dieser SNPs lässt sich eine Klassifikationsrate von 100% erreichen und der genetische Unterschied zwischen den beiden Populationen erklären. Abschließend zu dem Fall $n < d$ wird im nächsten Unterkapitel ein Fazit gezogen.

5.3.3 Fazit zu dem Fall $n < d$

In diesem zweiten großen Fall bei SNP-Datensätzen zeigen sich vor allem die CLS der Variablen mit der abhängigen Variable als sehr nützlich. Der Einfluss der nach Konstruktion wichtigen Variablen ist sehr deutlich in den CLS der Variablen mit der abhängigen Variable zu erkennen und kann dabei helfen, wichtige Einflussvariablen zu selektieren. Je mehr uninformative Variablen in den Datensätzen vorliegen, umso stärker lassen sich die relevanten Informationen erkennen. Anders als bei der Selektion von Beobachtungen ist es in diesem Fall nur sinnvoll, die Variablen fest anhand der CLS der Variablen mit der abhängigen Variable zu wählen. Es zeigt sich, dass die CLS erneut dabei helfen aus uninformativen Informationen in den Daten die wichtigen herauszufiltern. Sogar bei einem so hochdimensionierten Datensatz wie den HapMap-Daten ist es mit Hilfe der CLS der Variablen mit der abhängigen Variable möglich, mit nur einem Bruchteil der Gesamtvariablen jede Beobachtungen korrekt zu klassifizieren. In dem Kontext der Variablenselektion bei SNP-Daten sollten die hohen CLS der Variablen mit der abhängigen Variable auf jeden Fall berücksichtigt werden, da diese potentiell sehr deutlich auf wichtige Einflussvariablen hindeuten. Weiter ist es potentiell möglich in den niedrigen LS der Variablen, einflussreiche Einzelvariablen zu finden.

6 Zusammenfassung und Ausblick

Eine Aufgabe der Statistik im Kontext genetischer Daten ist es, dabei zu helfen nach Auslösern von Krankheiten zu suchen. Datensätze sogenannter genomweiter Assoziationsstudien oder Datensätze wie die HapMap-Daten enthalten SNP-Daten von enormem Umfang. Auch andere Datensätze nehmen schnell sehr große Umfänge an. Forschungsgegenstand ist es, diese Datensätze zu reduzieren und besser handhabbar zu machen. Dabei gibt es zwei Fälle zu berücksichtigen. Einmal gibt es mindestens so viele Beobachtungen wie Variablen und als Zweites gibt es mehr Variablen als Beobachtungen.

In der Informatik gibt es im Kontext von Regressionsverfahren den Ansatz anstelle der Betrachtung des gesamten Datensatzes, gezielt Beobachtungen anhand der Leverage Scores zu entnehmen und die Modelle an eine Stichprobe der Daten anzupassen. In dieser Arbeit werden die Leverage Scores dazu verwendet, Stichproben aus SNP-Datensätze zu entnehmen. In einer Simulationsstudie werden mit Hilfe der logischen Regression nach potentiellen Wechselwirkungen für das Auslösen des Krankheitsstatus gesucht und anhand der Klassifikationsraten der Modelle die Güte der Stichprobenauswahl bewertet. In einem zweiten Fall werden anhand der Leverage bzw. der Cross Leverage Scores Variablen aus den SNP-Datensätzen selektiert.

Bei der Auswahl der Beobachtungen durch die Leverage Scores gibt es in jeder Datensituation eine Gewichtung der Leverage Scores, die dazu führt, dass die resultierenden Klassifikationsraten wenigstens nicht schlechter und in vielen Datensituationen sogar besser sind, im Vergleich dazu jede Beobachtung mit einer gleichen Auswahlwahrscheinlichkeit in die Stichprobe aufzunehmen. Bei besonders kleinen Stichprobenumfängen sind es die niedrigen Leverage Scores und bei größeren Stichprobenumfängen und vielen Variablen in den Datensätzen sind es die hohen Leverage Scores, die besonders für die Wahl der Beobachtungen geeignet sind. Wenn es in den Daten Fälle gibt deren Krankheitsstatus nicht durch die genetischen Einflüsse erklärt sind, eignen sich die hohen und niedrigen Leverage Scores für die Wahl der Beobachtungen. Weiter können die Leverage Scores dabei helfen, Untergruppen in der Gruppe der Erkrankten zu finden und deren genetische Einflüsse auf den Krankheitsstatus besser zu finden.

Für die Selektion der Variablen sind es vor allem die Cross Leverage Scores der Variablen mit der abhängigen Variable, die dazu geeignet sind, die wichtigen Einflussvariablen zu selektieren. Viele der wichtigen Einflussvariablen besitzen eine hohe Hebelwirkung auf die abhängige Variable und lassen sich dementsprechend potentiell in diesen Werten wiederfinden. Dies führt dazu, dass eine gute Anpassung der Modelle mit nur einem Bruchteil der Gesamtvariablen möglich ist. Bei den HapMap-Daten sind 10 der insgesamt 7648 SNPs ausgewählt durch die Cross Leverage Scores der Variablen mit der abhängigen Variable bereits genug, um Modelle anzupassen, die eine hundertprozentige Richtigklassifikation durchführen.

Im Rahmen der Arbeit ergeben sich einige interessante Ansatzpunkte für zukünftige Forschung. Die zwei großen Fälle, die bei SNP-Datensätzen auftreten können, sorgen

dafür, dass sich diese Ansätze an unterschiedlichen Stellen ergeben. In dem Fall, dass es mehr Beobachtungen als Variablen gibt wäre es weiterführend interessant zu untersuchen, wie sich die Wahl durch die Leverage Scores verhält, wenn es sehr viele nicht durch genetische Einflüsse erklärte Fälle in sehr hochdimensionierten Datensätzen gibt. In diesem Rahmen wäre es von besonderem Interesse zu untersuchen, ob die Leverage Scores in dieser Situation ebenfalls dabei helfen, gezielter Untergruppen aufzufinden. Die Leverage Scores können ebenfalls dabei Helfen, Ausreißer in den Daten zu identifizieren (vgl. Hoaglin und Welsch, 1978). Daher bleibt noch die Fragestellung zu klären (vor allem bei der Wahl der Beobachtung anhand der hohen Leverage Scores), ob es bei SNP-Daten entsprechende Ausreißer gibt und ob diese Ausreißer sich potentiell negativ auf die Anpassung der Modelle auswirken. Weiter sollte in diesem Kontext untersucht werden, was passiert, wenn die SNPs sehr unsauber bzw. falsch gemessen werden und wie sich dies auf die Wahl anhand der Leverage Scores und auf die angepassten Modelle auswirkt.

In dem Fall, dass mehr Variablen als Beobachtungen vorliegen, ergeben sich einige weitere Ansatzpunkte. Grundsätzlich sollte ein alternatives Variablenselektionsverfahren herangezogen werden, um dieses zum Vergleich zu nutzen, anstelle der einfache Zufallsauswahl. Mögliche Verfahren sind etwa der Random Forest (vgl. Breiman 2001) oder die sogenannte Spike-and-Slab-Variablenselektion (vgl. Ishwaran und Rao 2005). In diesem Kontext sollten die Wechselwirkungen untersucht werden, die sich bei der Auswahl der Variablen durch die logische Regression ergeben. Als zweiter Ansatzpunkt kann die Selektion der Variablen durch die Cross Leverage Scores der Variablen mit der abhängigen Variable genauer betrachtet werden, wenn sehr viele Beobachtungen vorliegen, die den Krankheitsstatus nicht durch die genetischen Einflüsse erklärt haben. Da die wichtigen Einflussvariablen sich teilweise sehr deutlich in den Cross Leverage Scores der Variablen mit der abhängigen Variable wiederfinden lassen, ist es von besonderem Interesse, wie nicht erklärte Fälle sich darauf auswirken. Weiter kann noch untersucht werden, ob Einzelvariablen bei SNP-Daten generell sehr niedrige Leverage Scores besitzen und ob es daher sinnvoll ist, bei einer möglichen Variablenselektion eine Kombination niedriger Leverage Scores und hoher Cross Levereage Scores mit der abhängigen Variable zu berücksichtigen. Als letzter Ansatzpunkt kann das 1000 Genomes Projekt herangezogen werden. Als Nachfolgeprojekt zu dem HapMap-Projekt können potentiell die Leverage Scores dieser SNP-Daten untersucht werden und damit verglichen werden, ob es sich bei den selektierten SNPs um bereits bekannte wichtige Einflüsse handelt.

Literatur

[1] Breiman L. (2001): Random Forest, Machine Learning, Vol. 45, No. 1, S. 5-32

[2] Drineas P., Magdon-Ismail M., Mahoney M., Woodruff D. (2012): Fast approximation of matrix coherence and statistical leverage, The Journal of Machine Learning Research, Vol. 13, No. 1, S. 3475-3506

[3] Gantz J., Reinsel D. (2012): The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, IDC Research Inc., Framingham, USA. Online im Internet: `https://www.emc-technology.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf`. Stand: 13.12.2016

[4] Geppert L., Ickstadt K., Munteanu A., Quedenfeld J. (2015): Random projections for Bayesian regression, Statistics and Computing. doi:10.1007/s11222-015-9608-z

[5] Golub H., Van Loan C. (1996): Matrix Computation, 3. Auflage, Johns Hopkins University Press, London

[6] González J., Armengol L., Guinó E., Solé X., Moreno V. (2014). SNPassoc: SNPs-based whole genome association studies. R package version 1.9-2. Online im Internet: `https://CRAN.R-project.org/package=SNPassoc`. Stand: 13.12.2016

[7] Graw J. (2015): Genetik, 6. Auflage, Berlin, Springer

[8] Hoaglin D., Welsch R. (1978): The Hat Matrix in Regression and ANOVA, The American Statistican, Vol. 32, No. 1, S. 17-22

[9] Huber W., Carey V., Gentleman R., Anders S., Carlson M., Carvalho B., Bravo H., Davis S., Gatto L., Girke T., Gottardo R., Hahne F., Hansen K., Irizarry R, Lawrence M., Love M, MacDonald J., Obenchain V., Oles A., Pages H., Reyes A., Shannon P., Smyth G., Tenenbaum D., Waldron L., Morgan M. (2015): Orchestrating high-throughput genomic analysis with Bioconductor, Nature Methods, Vol. 12, No. 2, S. 115-121. Online im Internet: `http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html`. Stand: 13.12.2016

[10] Ishwaran H., Rao J. (2005): Spike and slab variable selection: Frequentist and Bayesian strategies, The Annals of Statistics, Vol. 33, No. 2, 730–773. doi:10.1214/009053604000001147

[11] Kooperberg C., Ruczinski I. (2005): Identifying Interacting SNPs Using Monte Carlo Logic Regression, Genetic Epidemiology, Vol. 28, No. 2, S. 157-170

[12] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Online im Internet: `https://www.R-project.org`. Stand: 13.12.2016

[13] Ruczinski I., Kooperberg C., LeBlanc M. (2003): Logic Regression, Journal of Computational and Graphical Statistics, Vol. 12, No. 3, S. 475-511

[14] Schwender H. (2013). logicFS: Identification of SNP Interactions. R package version 1.42.0.

[15] Schwender H., Fritsch A. (2013). scrime: Analysis of High-Dimensional Categorical Data such as SNP Data. R package version 1.3.3. Online im Internet: `https://CRAN.R-project.org/package=scrime`. Stand: 13.12.2016

[16] Schwender H., Ickstadt K. (2008): Identification of SNP interactions using logic regression, Biostatistics, Vol. 9, No. 1, S.187-198

[17] The 1000 Genomes Project Consortium (2015): A global reference for human genetic variation, Nature, Vol. 526, S. 68–74. doi:10.1038/nature15393

[18] Waldrop M. (2016): More than Moore, Nature, Vol. 530, S. 145-147

A Graphiken

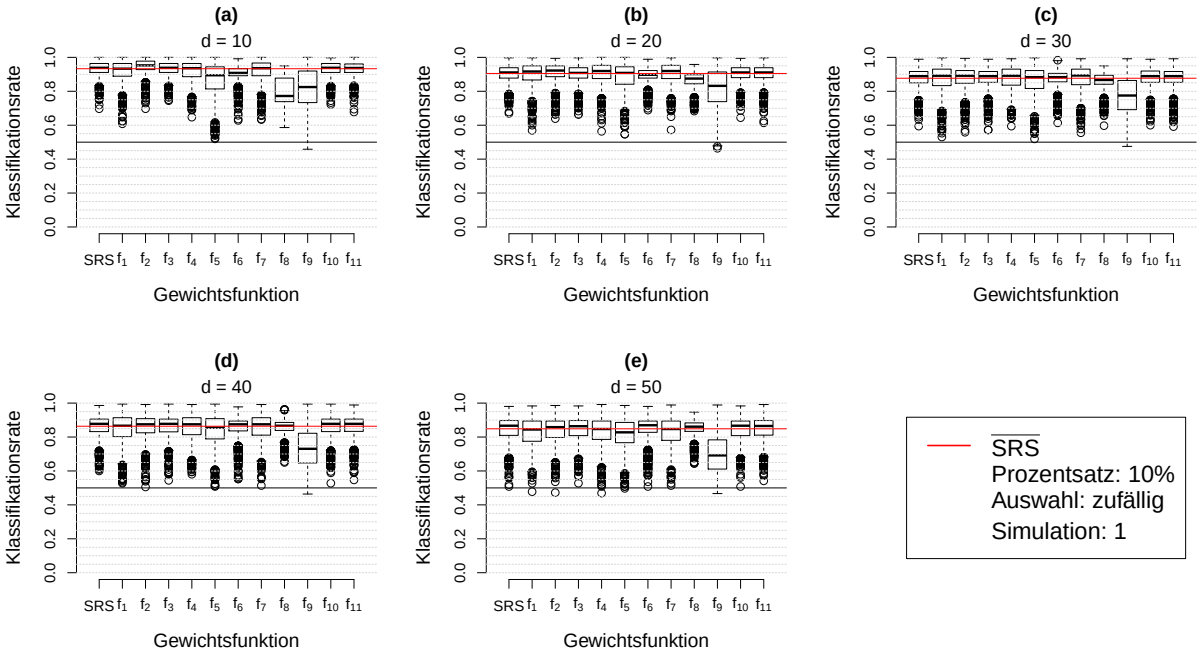

Abbildung A1: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 10% der Daten aus Simulation 1 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

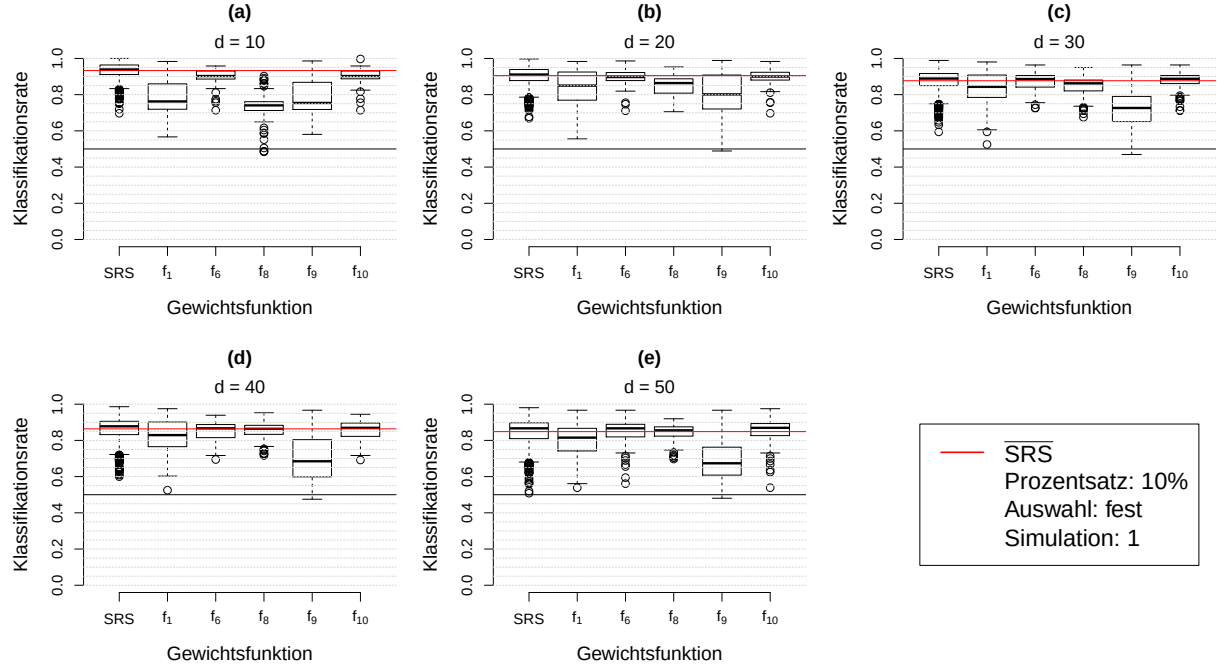

Abbildung A2: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 10% der Daten aus Simulation 1 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

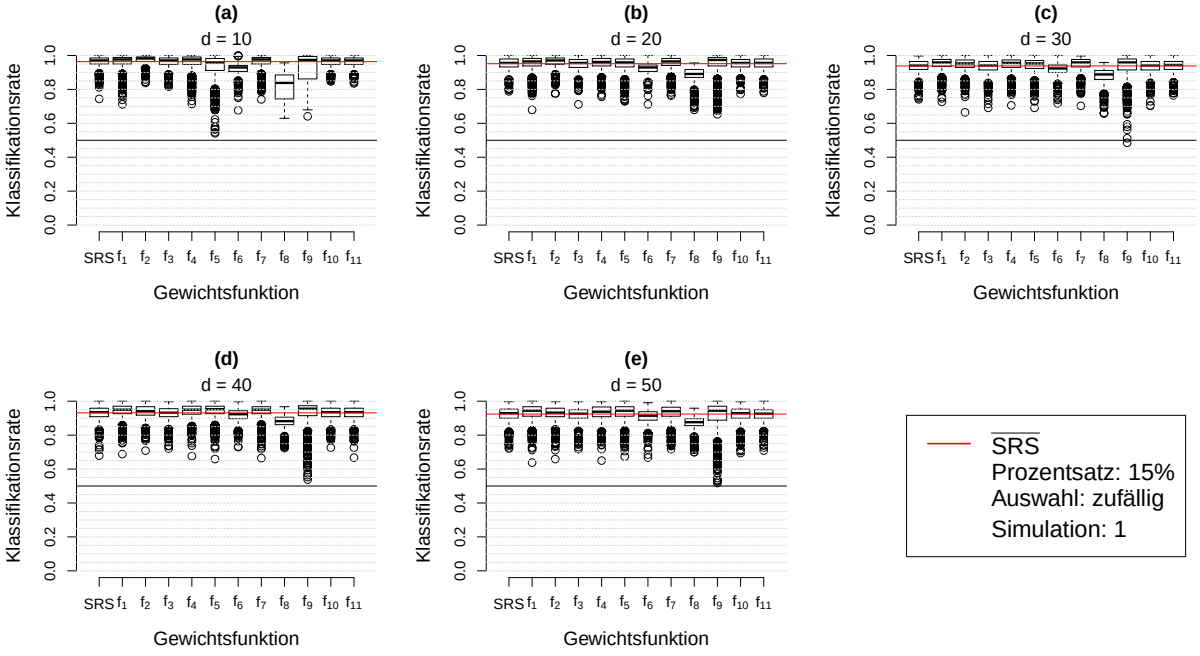

Abbildung A3: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 15% der Daten aus Simulation 1 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

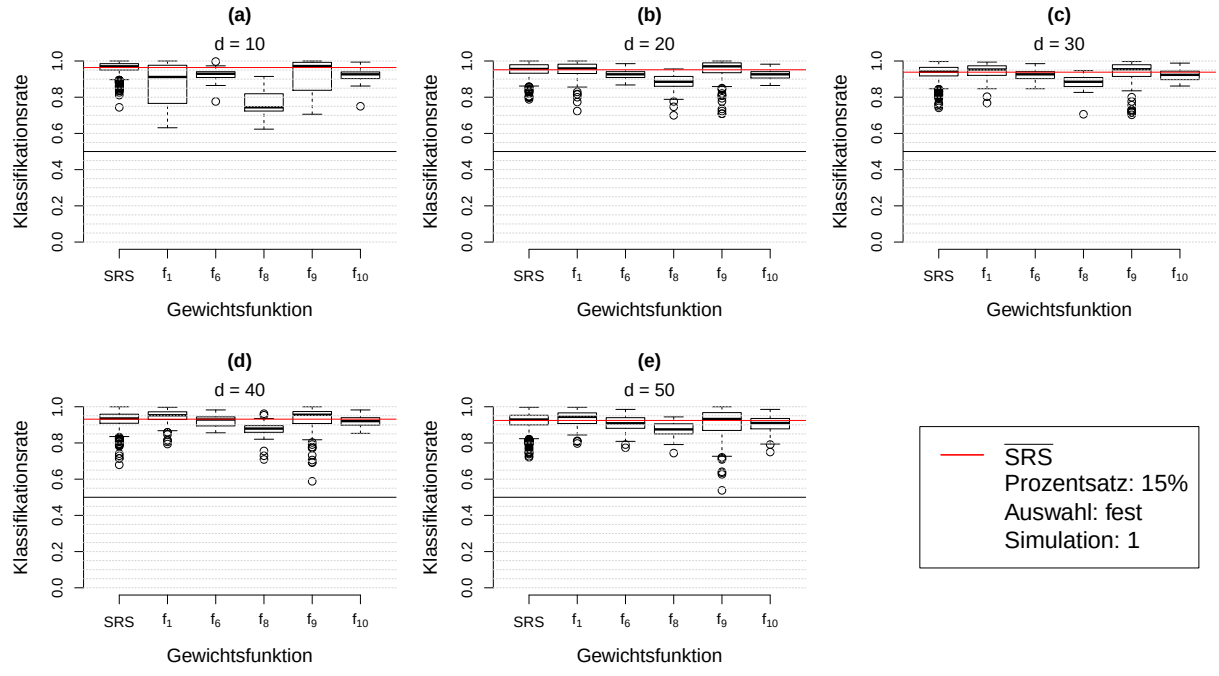

Abbildung A4: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 15% der Daten aus Simulation 1 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

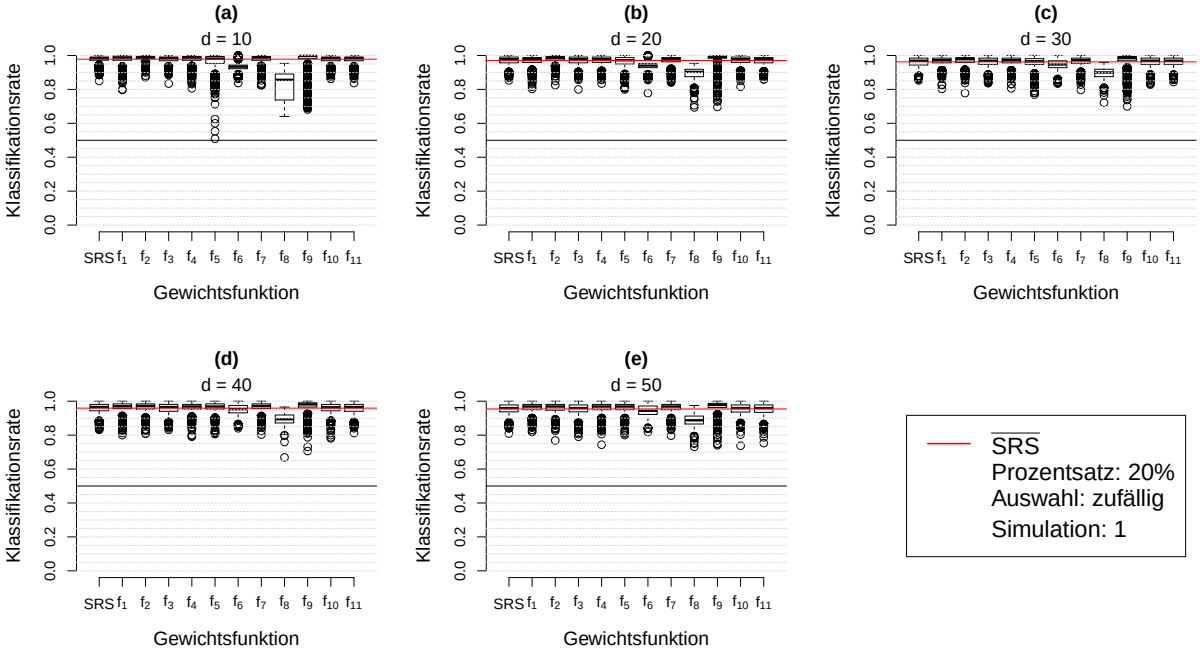

Abbildung A5: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 20% der Daten aus Simulation 1 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

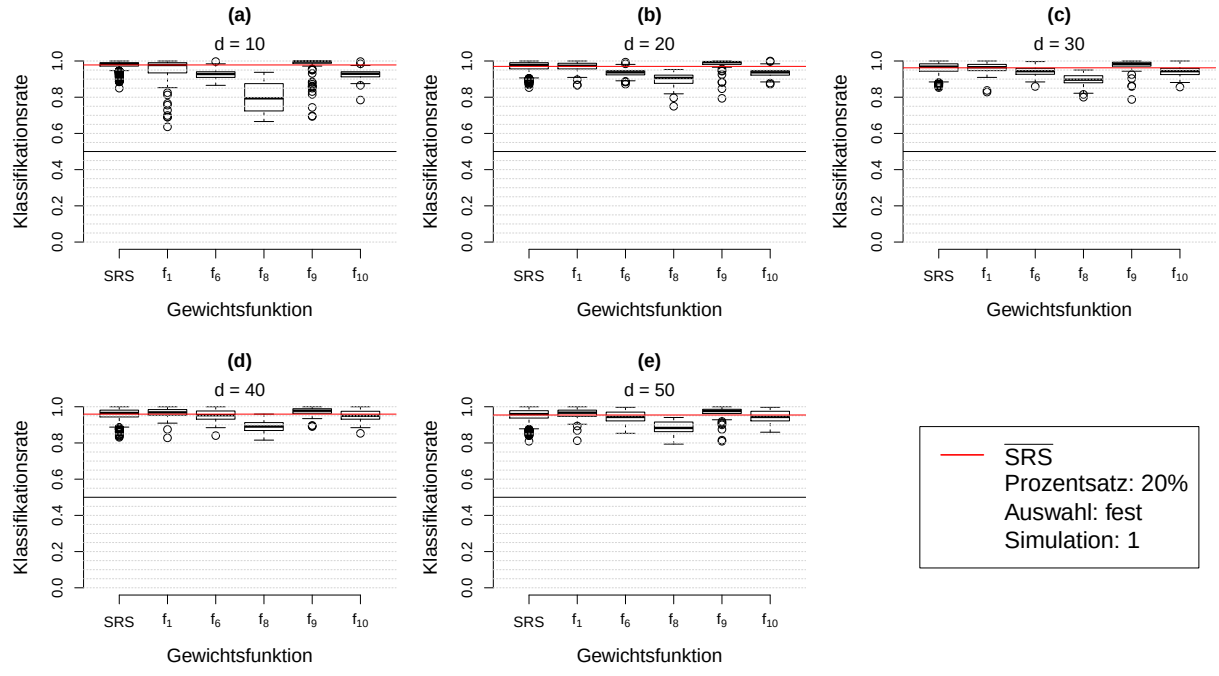

Abbildung A6: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 20% der Daten aus Simulation 1 bei fester Auswahl der Beobachtungen anhand der Leverage Scores

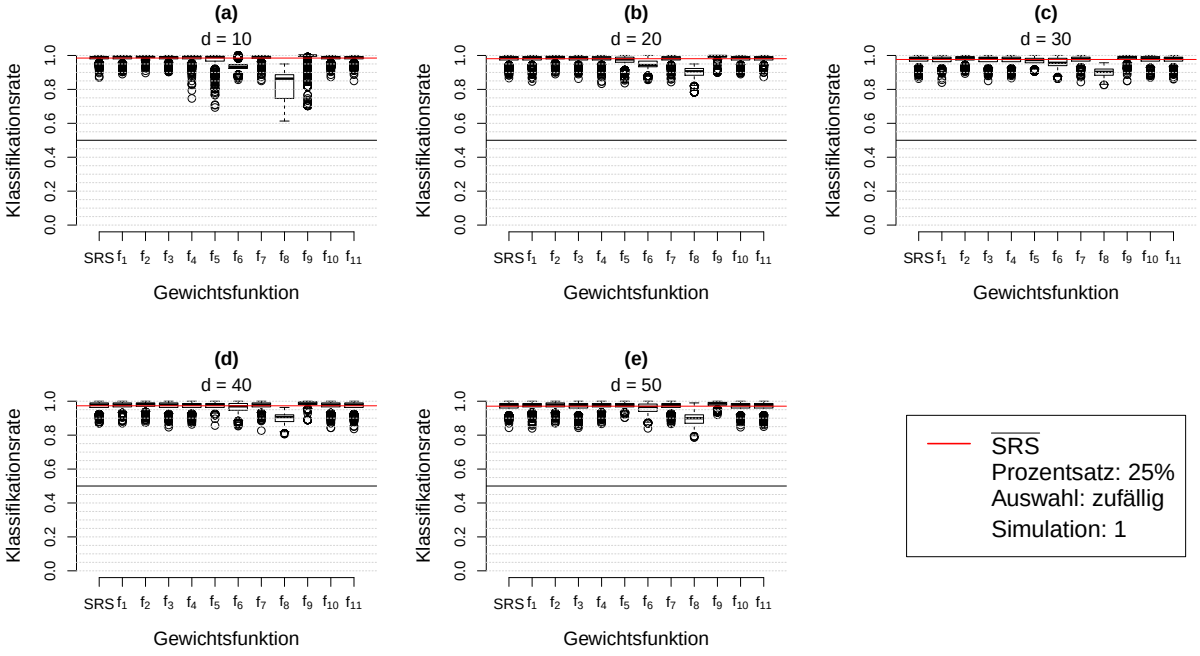

Abbildung A7: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 25% der Daten aus Simulation 1 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

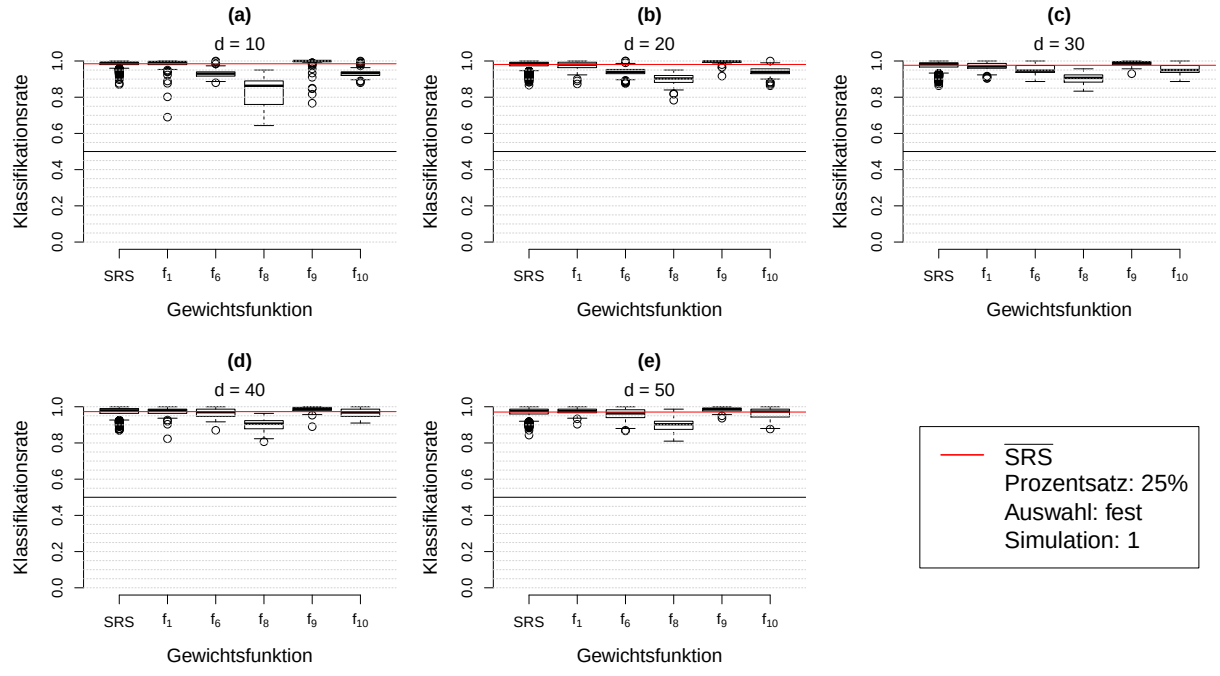

Abbildung A8: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 25% der Daten aus Simulation 1 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

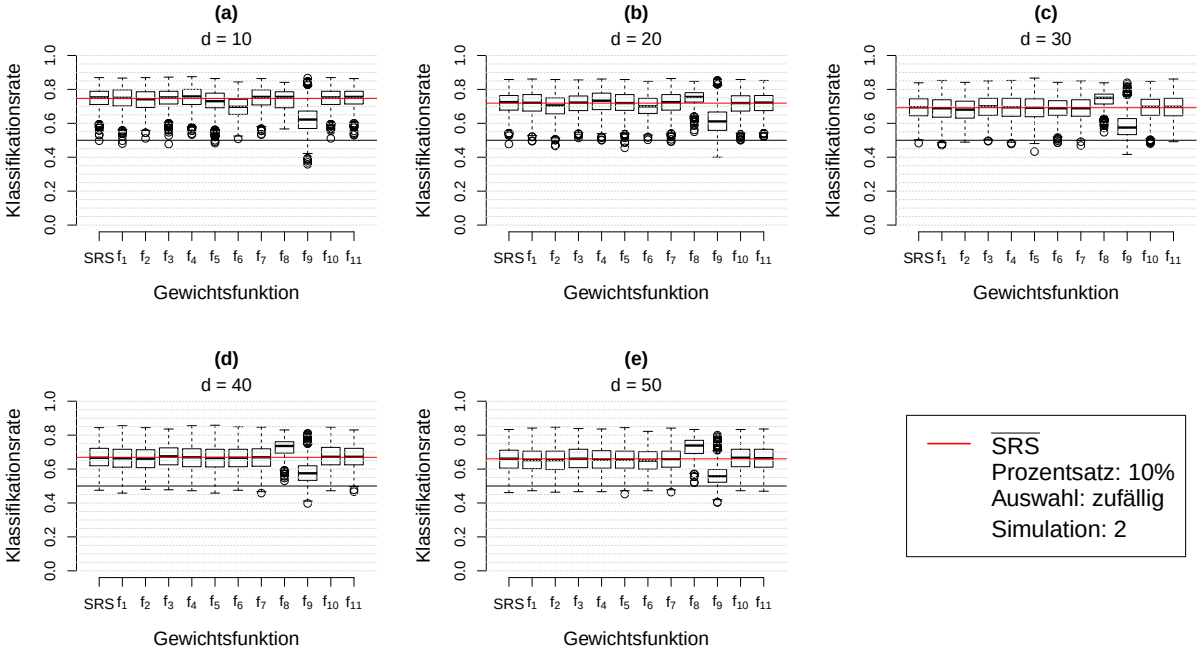

Abbildung A9: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 10% der Daten aus Simulation 2 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

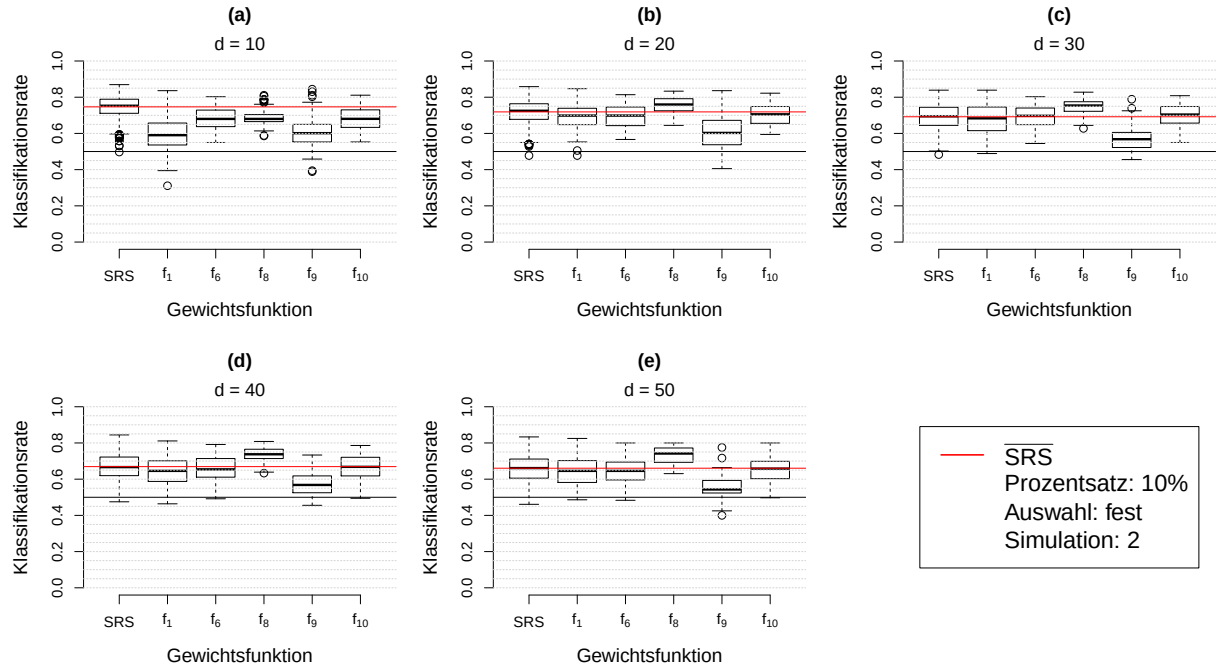

Abbildung A10: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 10% der Daten aus Simulation 2 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

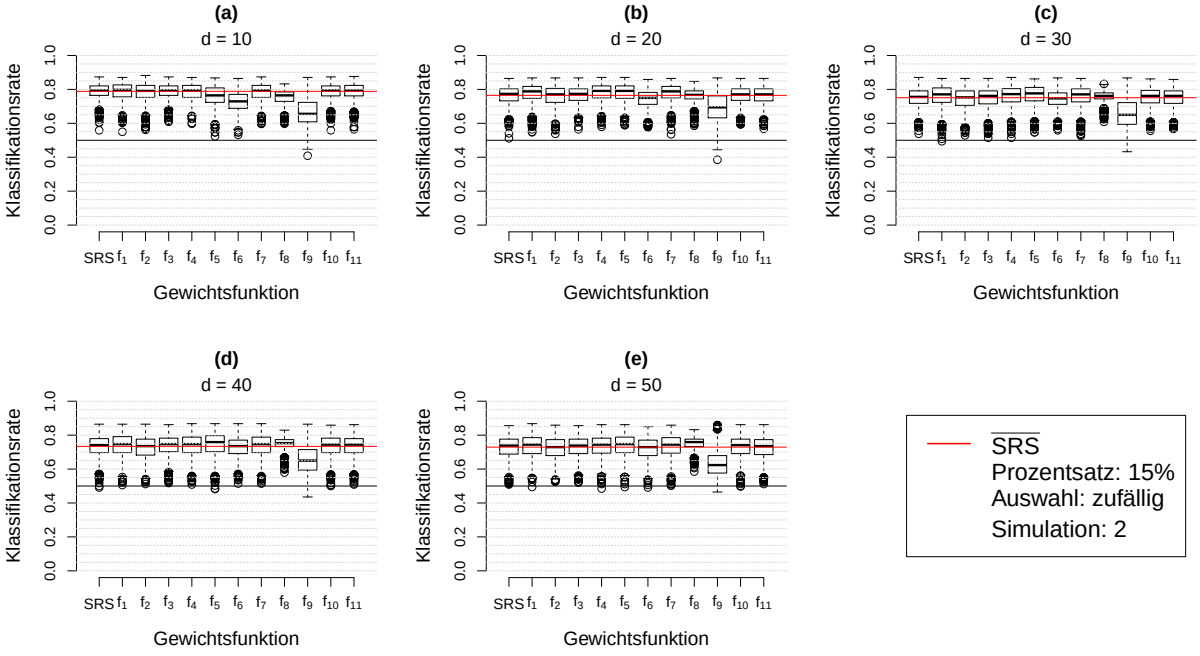

Abbildung A11: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 15% der Daten aus Simulation 2 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

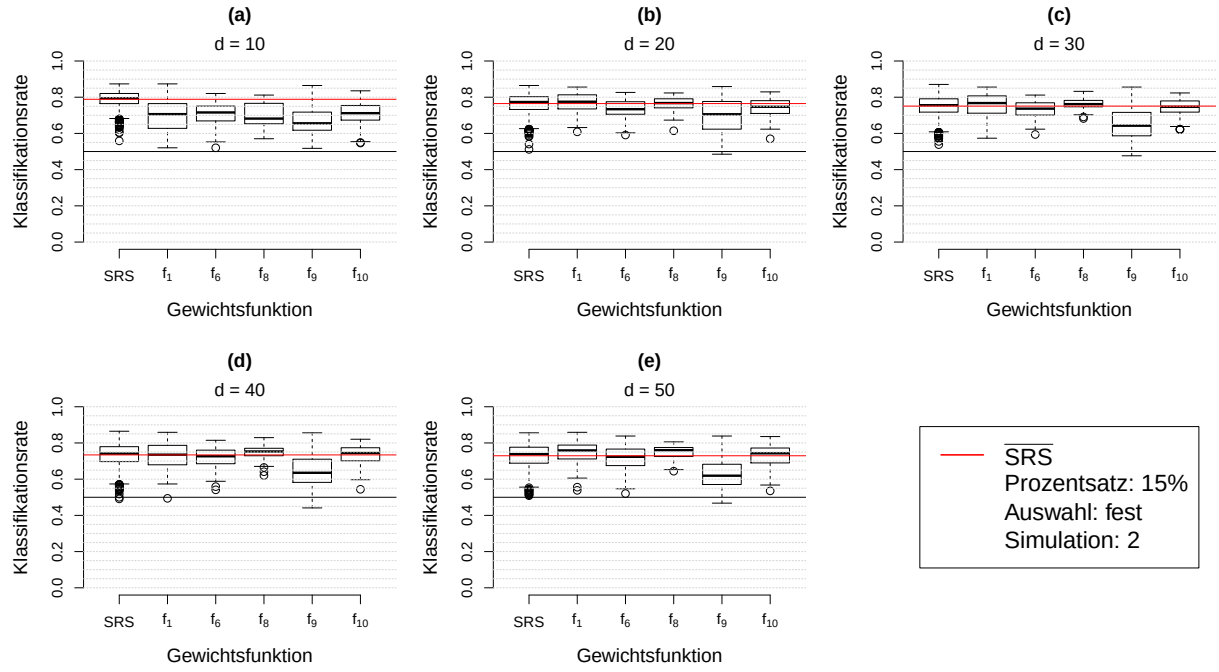

Abbildung A12: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 15% der Daten aus Simulation 2 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

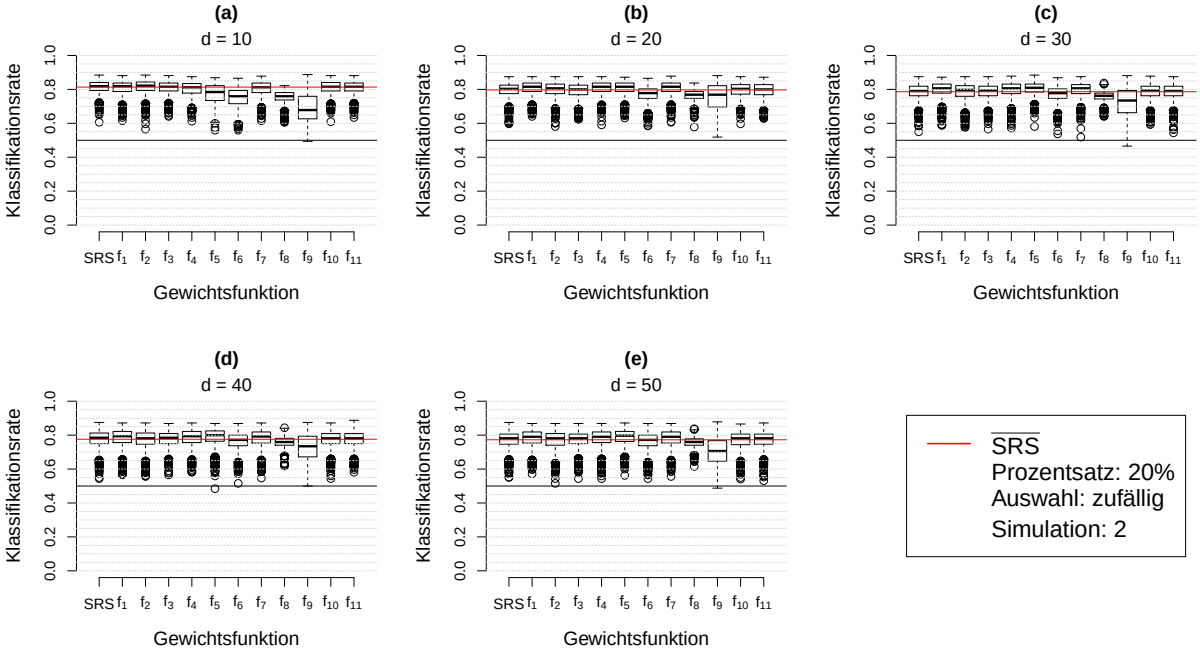

Abbildung A13: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 20% der Daten aus Simulation 2 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

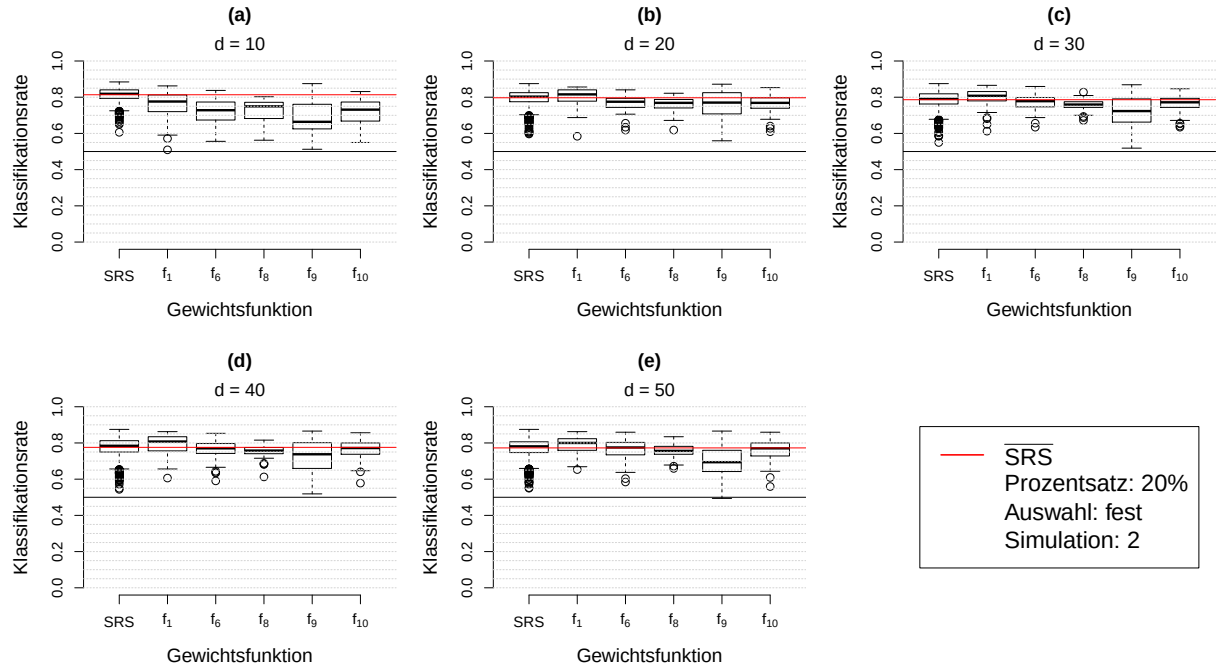

Abbildung A14: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 20% der Daten aus Simulation 2 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

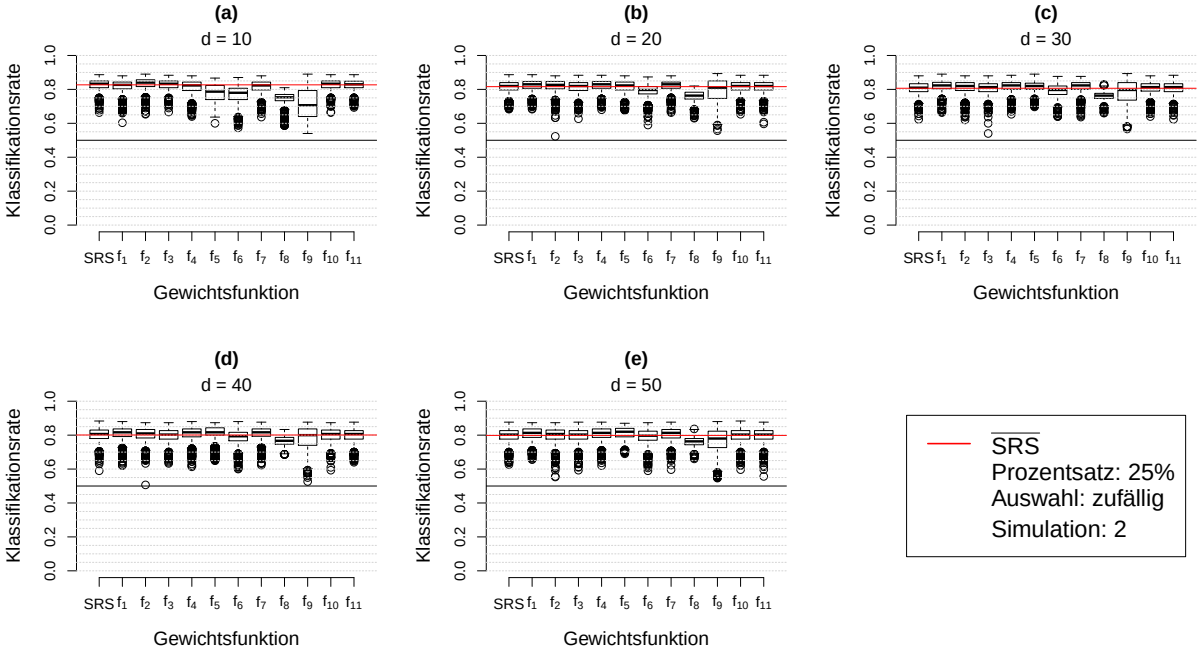

Abbildung A15: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 25% der Daten aus Simulation 2 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

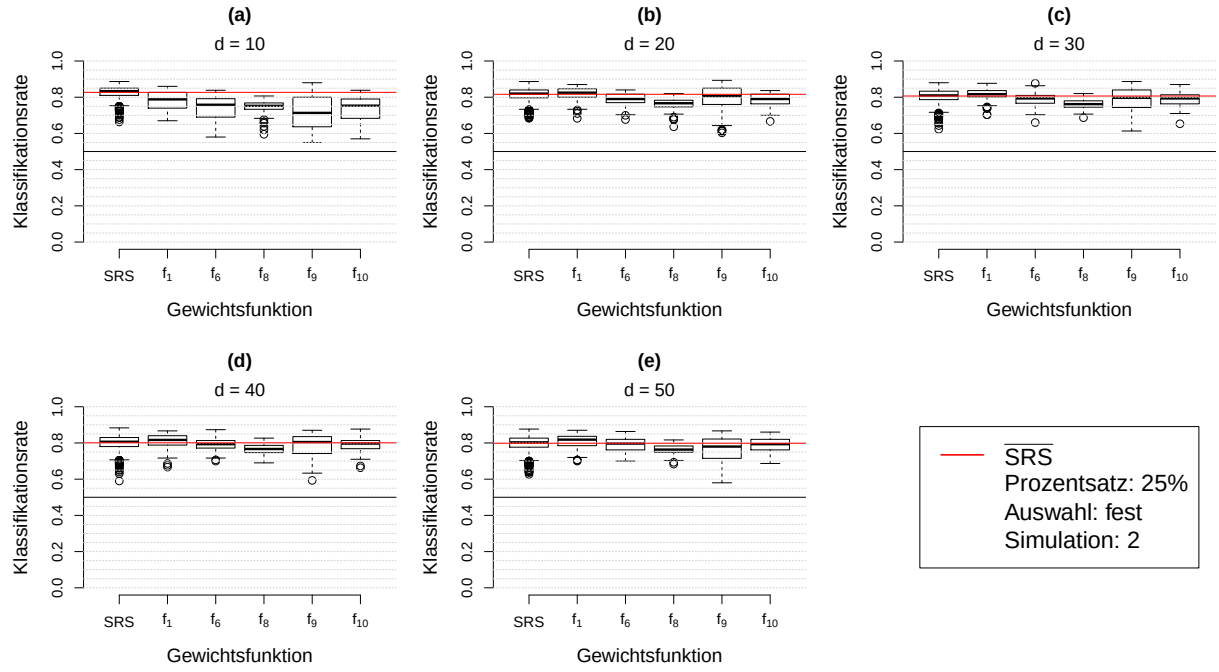

Abbildung A16: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 25% der Daten aus Simulation 2 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

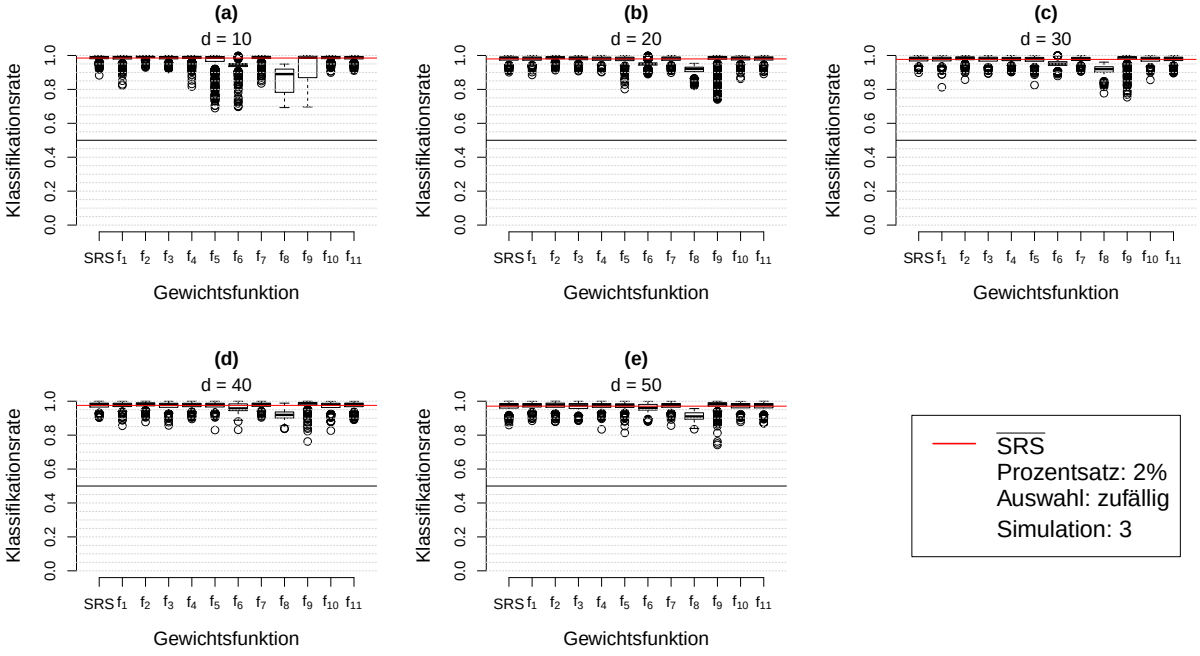

Abbildung A17: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 2% der Daten aus Simulation 3 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

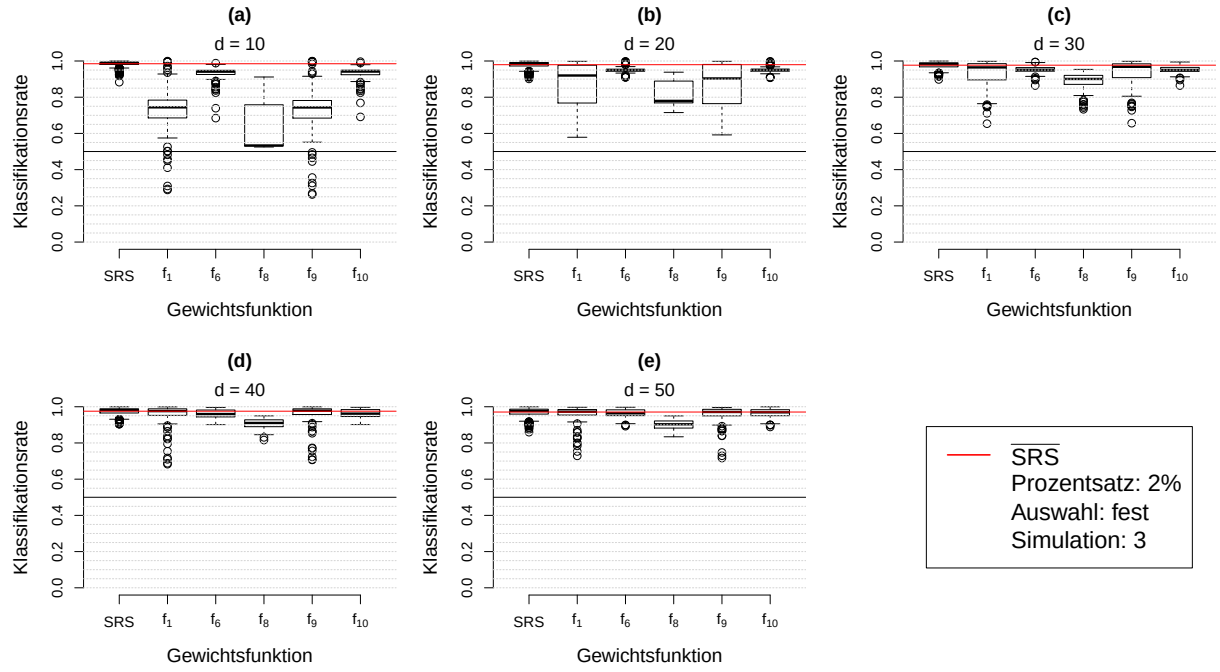

Abbildung A18: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 2% der Daten aus Simulation 3 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

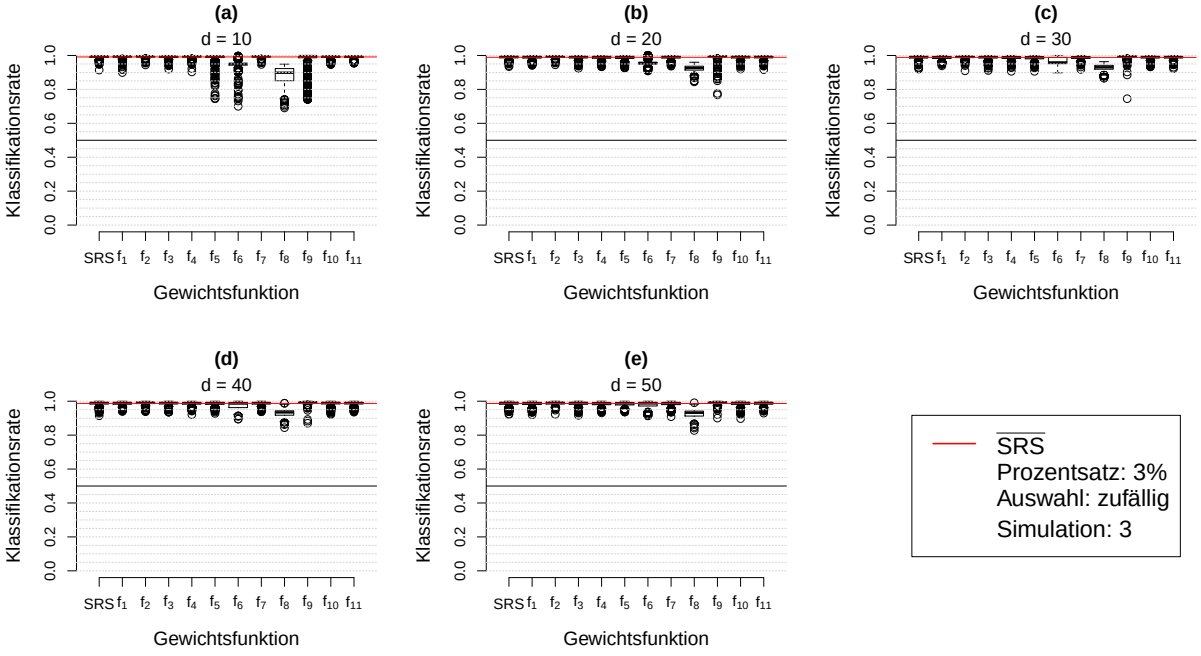

Abbildung A19: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 3% der Daten aus Simulation 3 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

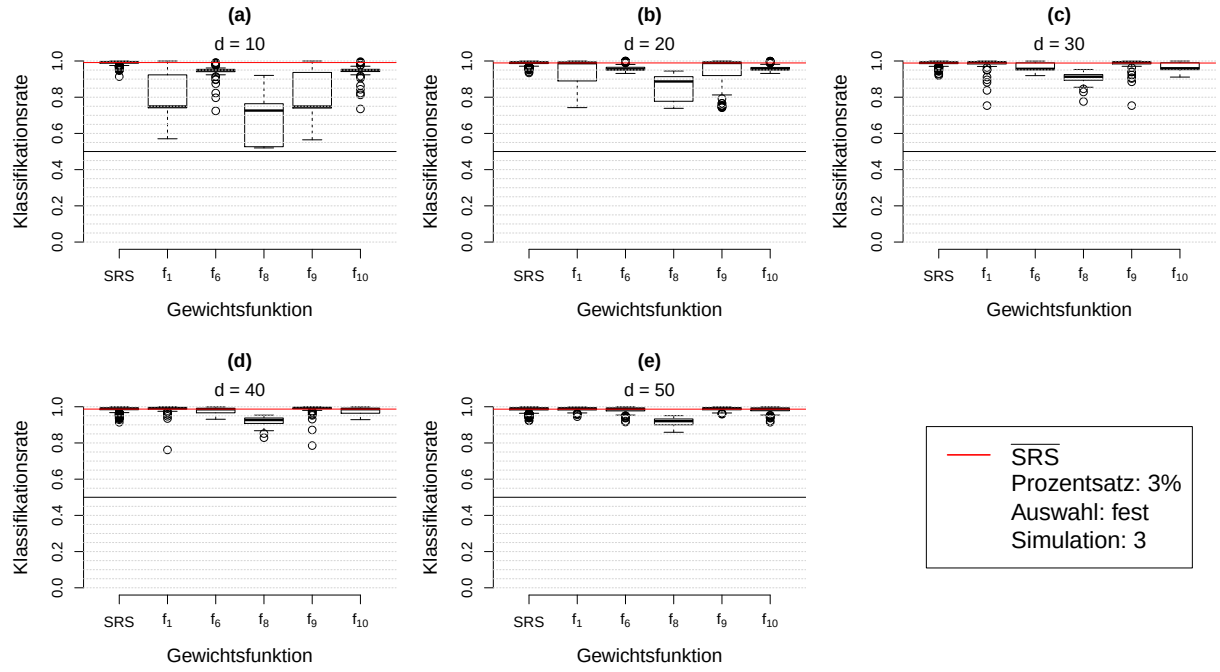

Abbildung A20: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 3% der Daten aus Simulation 3 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

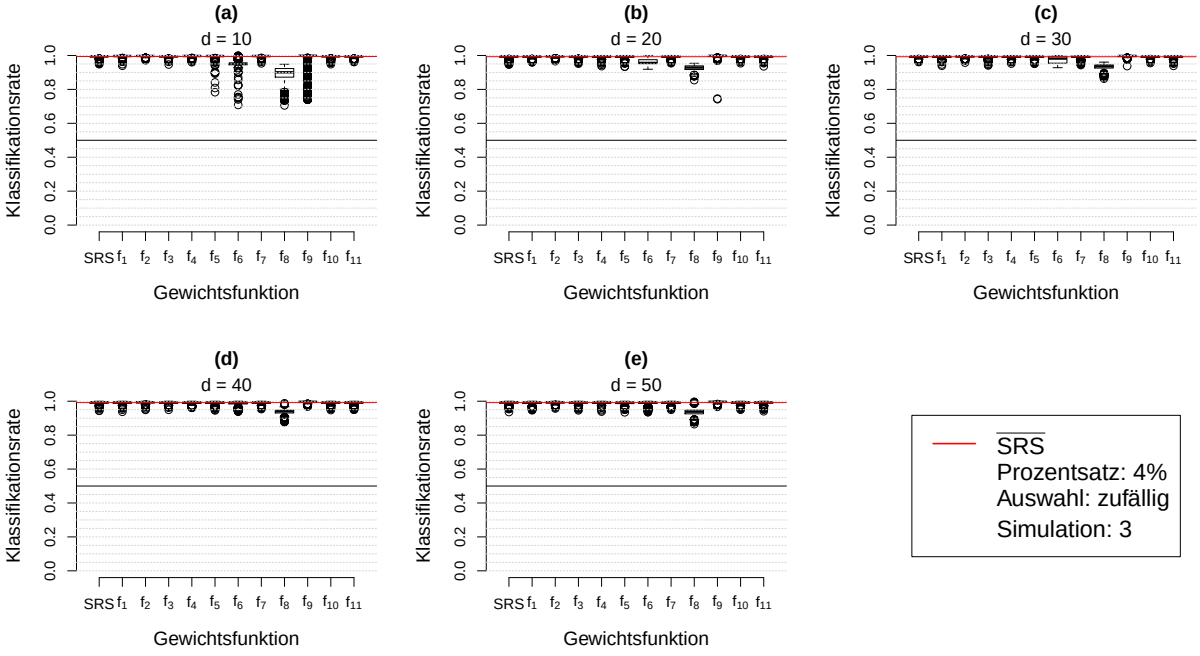

Abbildung A21: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 4% der Daten aus Simulation 3 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

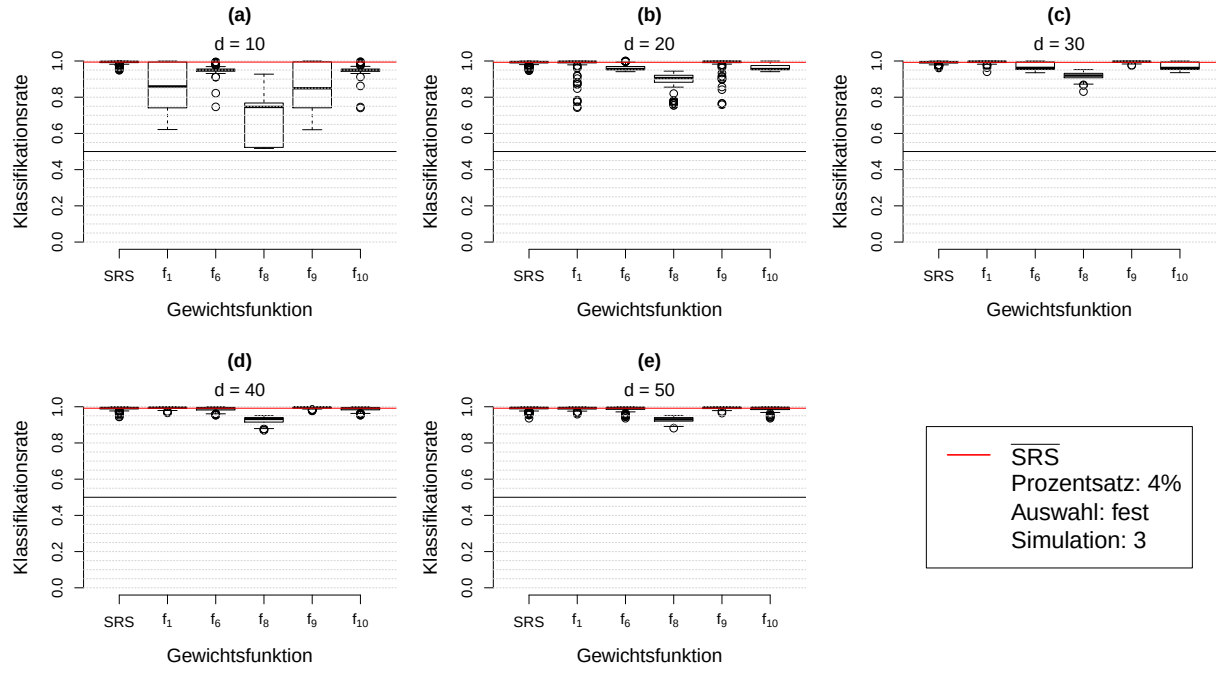

Abbildung A22: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 4% der Daten aus Simulation 3 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

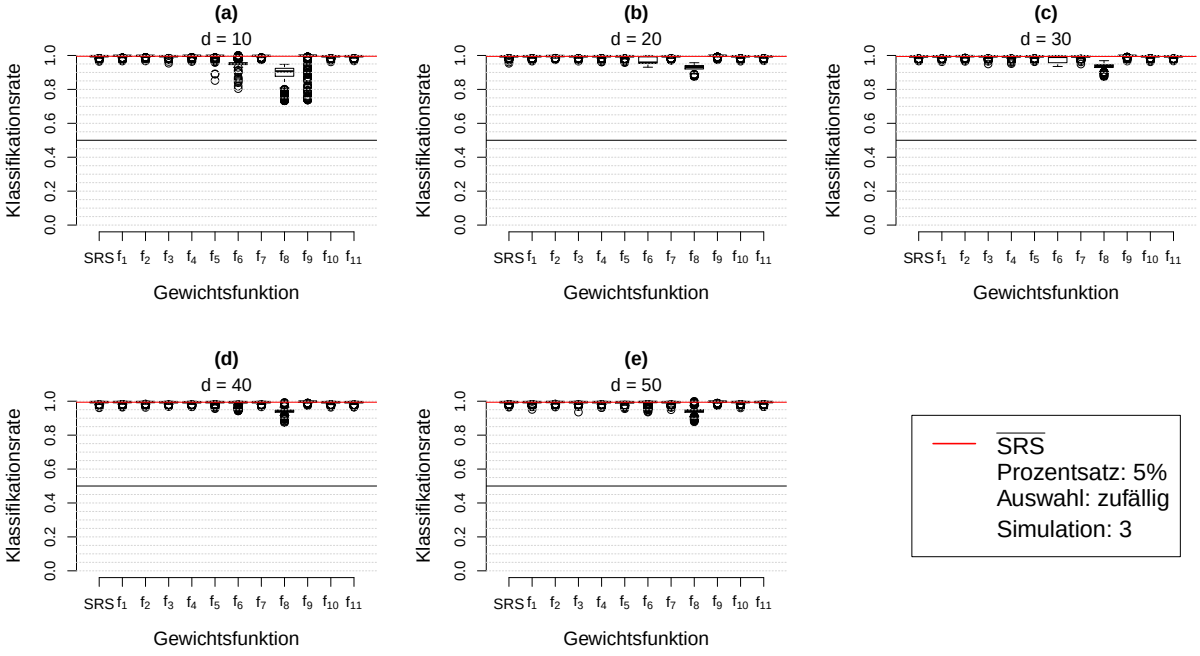

Abbildung A23: Boxplots der Klassifikationsraten durch wiederholte Anpassung logischer Regressionsmodelle mit dem logicFS-Ansatz auf 5% der Daten aus Simulation 3 bei zufälliger Auswahl der Beobachtungen proportional zu den Leverage Scores.

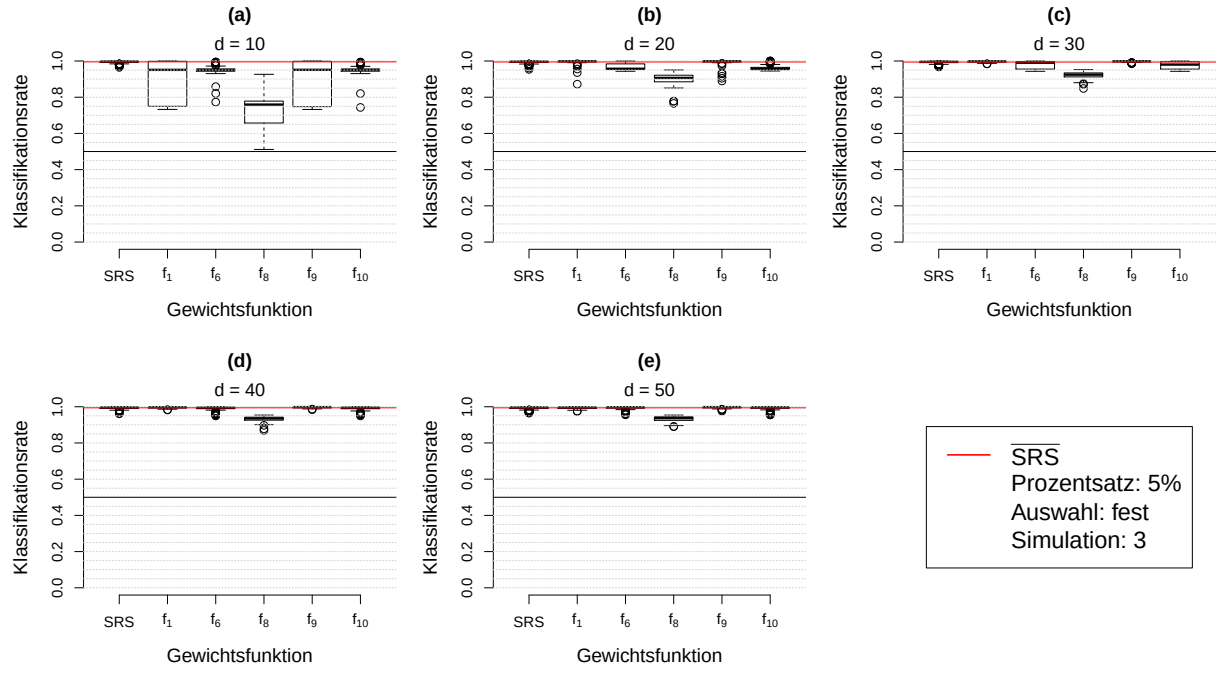

Abbildung A24: Boxplots der Klassifikationsraten durch das Anpassen logischer Regressionsmodelle mit dem logicFS-Ansatz auf 5% der Daten aus Simulation 3 bei fester Auswahl der Beobachtungen anhand der Leverage Scores.

B Tabellen

Tabelle B1: Anzahl der Datensätze der jeweils 100 Datensätze der Simulation 4, in denen sich eine Binärvariable X_i, $i = 1, 2, 3, 5, 6, 7, 9, 10, 11, 12$, in den $2d' = 20$ niedrigsten bzw. höchsten Cross Leverage Scores einer anderen Binärvariable X_j, $j = 1, 2, 3, 5, 6, 7, 9, 10, 11, 12$, in den Daten wiederfinden lässt.

			X_1	X_2	X_3	X_5	X_6	X_7	X_9	X_{10}	X_{11}	X_{12}
$d = 250$	f_9	X_1	100	99	16	14	0	1	2	0	10	1
		X_2	99	100	73	60	0	0	0	0	75	0
		X_3	16	70	100	13	0	1	0	0	24	0
		X_5	13	60	13	100	52	0	0	0	16	0
		X_6	1	0	0	70	99	4	1	4	0	3
		X_7	1	0	1	0	3	100	8	44	11	0
		X_9	2	0	0	0	0	8	100	100	11	0
		X_{10}	0	0	0	0	3	48	100	100	82	0
		X_{11}	22	85	30	32	0	22	21	87	100	18
		X_{12}	1	0	0	0	2	0	1	0	35	100
	f_8	X_1	0	0	1	0	4	7	4	25	0	8
		X_2	0	0	0	0	14	44	32	99	0	69
		X_3	1	0	0	0	5	10	3	38	0	13
		X_5	0	0	0	0	0	15	9	40	0	11
		X_6	9	14	7	0	0	1	1	0	2	0
		X_7	4	43	11	14	0	0	1	0	0	16
		X_9	4	30	3	7	0	3	0	0	0	7
		X_{10}	29	98	43	45	0	0	0	0	0	62
		X_{11}	0	0	0	0	7	0	0	0	0	0
		X_{12}	17	81	28	26	0	28	14	86	0	0
$d = 500$	f_9	X_1	100	100	31	35	0	0	0	0	1	0
		X_2	100	100	82	88	0	0	0	0	40	0
		X_3	25	77	100	34	0	0	0	0	0	0
		X_5	34	80	35	100	25	0	1	0	4	0
		X_6	0	1	0	66	99	2	3	2	0	0
		X_7	0	0	0	0	1	100	25	49	3	0
		X_9	0	0	0	1	0	25	100	100	1	0
		X_{10}	0	0	0	0	1	66	100	100	32	0
		X_{11}	23	78	31	32	0	20	17	64	100	0
		X_{12}	0	0	0	0	0	0	0	0	0	100
	f_8	X_1	0	0	0	0	0	17	10	45	0	2
		X_2	0	0	0	0	3	65	61	100	0	34
		X_3	0	0	0	0	1	15	11	62	0	1
		X_5	0	0	0	0	0	18	18	63	0	1
		X_6	4	4	8	0	0	0	0	0	1	0
		X_7	15	49	13	17	0	0	0	0	0	0
		X_9	9	47	10	16	0	0	0	0	0	0
		X_{10}	58	99	74	70	0	0	0	0	0	21
		X_{11}	0	0	0	0	1	0	0	0	0	0
		X_{12}	23	83	24	31	0	17	14	73	0	0

Eidesstattliche Erklärung des Urhebers

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht habe.

Dortmund, den 13.12.2016 Unterschrift