Universität Zürich[UZH]

Statistik Aufbau

Benjamin Fretwurst

Version 0.84

Zuletzt aktualisiert: 2024-03-05

Kontakt:
Dr. Benjamin Fretwurst
Institut für Kommunikationswissenschaft und Medienforschung – IKMZ
Andreasstrasse 15
8050 Zürich
b.fretwurst@ikmz.ch

Inhalt

Tabellenverzeichnis

Abbildungsverzeichnis

Einleitung und Syllabus

Syllabus

Vorwort

Sicher freuen Sie sich schon auf «Statistik: Aufbau», und ich glaube, Sie haben allen Grund dazu. Manche freuen sich weniger – was ja auch normal und ok ist. Wieder andere, denken lieber daran, wie das Leben so sein wird, wenn Sie «Statistik: Aufbau» hinter sich haben. Ihnen allen soll dieser Begleittext zur Seite stehen, damit Sie aus dem Modul das für sich Beste rausholen. Diejenigen, die in der Statistik ein mächtiges Tool entdecken, will ich ein tiefergehendes Verständnis ermöglichen. Denen, die die Statistik einfach gut absolvieren wollen, soll das Wichtigste vermittelt werden und die mit Graus auf das Modul schauen, soll das Grauen genommen und etwas Greifbares und Handhabbares angeboten werden, das sich – mit zumutbaren Investitionen – lösen lässt. Hier in der Einleitung schreibe ich Ihnen, was ich über den Sinn und die Mächtigkeit von Statistik denke sowie die möglichen Ursachen für das Unbehagen denke.
Liebe Grüsse
Benjamin Fretwurst

Was bringt uns Statistik

Unser Alltag ist von Beobachtungen geprägt, aus denen wir etwas über uns und die Welt lernen. Wir stellen Vermutungen an und haben das Gefühl, dass wir wissen, wie es läuft. Das heisst, wir machen viele Beobachtungen und ziehen unsere Schlüsse daraus. Wir entwickeln also aus empirischen Beobachtungen Theorien. Diese Beobachtungen sind nur nicht sehr systematisch und die Schlüsse, die wir aus ihnen ziehen sind mal mehr von einer Erinnerung und mal mehr von einer anderen Erinnerung geprägt. Wenn wir an dieses Erfahrungswissen etwas wissenschaftlicher herangehen wollen, um systematisch Erkenntnisse zu erlangen, auf die wir uns besser verlassen können, dann machen wir emprische Forschung.

Empirische Forschung ist wiederum dann genau und gültig, wenn sie sehr viele (möglichst unverzerrte) Beobachtungen anstellt. Aber wie können wir nur aus diesen ganzen Daten Informationen extrahieren, wie daraus Schlüsse ziehen? Sie ahnen es: Das macht Statistik. Statistik ist also ein Zweig der Mathematik, mit dessen Hilfe grosse Mengen an Daten auf Kennwerte reduziert werden können, aus denen wir leicht unsere Schlüsse für unser Verhalten im Alltag ziehen können – sei es beruflicher Alltag oder Privates. Die statistische Datenanalyse erlaubt es uns, sehr komplexe Beziehungen in den gemachten Beobachtungen zu finden und zu interpretieren. Die Methoden der Datenerhebung, wie Sie sie in der Einführung kennengelernt haben, ermöglichen (je nach Budget) ein paar Tausend Beobachtungen innerhalb einer Studie, die auf Knopfdruck in Sekundenbruchteilen statistisch analysiert werden können. Daneben gibt es aus dem Alltag der Menschen, der digital erfasst ist Terrabite grosse Datenmengen, die mit denselben statistischen Grundlagen ausgewertet werden können. Mit Hilfe von maschinellem Lernen können aus diesen Datenmengen Prognosen erstellt werden. Dieses «Maschinelle Lernen» (oder «Machine Learning» ML) basiert zu grossen Teilen auf den statistischen Methoden, die Sie in diesem Semester kennenlernen. Sie werden sehen, wie man aus statistischen Modellen generell lernen kann und wie man mit statistischen Methoden Prognosen anstellt, wie sie auch von ML-Algorithmen bereitgestellt werden. Die Art dieser Beziehungen wird aus der Alltagswahrnehmung abgeleitet und durch Formulierung wissenschaftlicher Hypothesen konkretisiert.

Wenn wir zum Beispiel davon sprechen, dass die Leute einfach nur das wichtig finden, was Ihnen die Medien vorgeben, dann wird damit ein Zusammenhang formuliert. Etwas konkreter würde ein KW-ler sagen: Die Menschen lernen aus der Thematisierung in den Medien, was wichtige Themen sind. Und weil das eine Theorie ist, bekommt sie auch noch einen Namen: «Agenda-Setting» (AS).

Gegen den AS könnte man einwenden: «Das gilt nicht immer. Die Leute kriegen schon mit, wenn die Preise steigen – dazu brauchen sie nicht die Medien.» Der AS gilt also nicht für alle Themen, sondern nur für solche, die die Leute nicht am eigenen Leib erfahren können. Es wird also in ‹obtrusive› und ‹nonobtrusive Issues› unterschieden. Jetzt haben wir einen Zusammenhang formuliert, der zusätzlich Randbedingungen enthält. Abgesehen von der Theorie könnte man die Forschungsfrage stellen, ob AS in gleichem Masse für Gebildetere und weniger Gebildete gilt. In der Alltagsbeobachtung wird es jetzt schon kompliziert, da wir diese Randbedingungen schwerlich alle gleichzeitig gegeneinander halten können. Selbst wenn wir den Bildungsstand mitbeobachten können, ist das nicht mit der vollen Differenziertheit möglich. Die wissenschaftliche Datenerhebung dient der Aufzeichnung vieler unabhängiger Beobachtungen. Multivariate Statistik ermöglicht es uns, diese Beobachtungen so zueinander in Beziehung zu setzen, dass wir am Ende einfache Kennwerte bekommen, die für Zusammenhänge stehen.

Was beschreibt die Funktion von Statistik am besten?
Versuchen Sie es mit Ihren eigenen Worten.

Überblick Analysemethoden

Der folgende Überblick zeigt die statistischen Verfahren, mit deren Hilfe kausale Zusammenhänge, Unterschiede und Datengruppierungen analysiert werden können. Diese verschiedenen Analysemethoden ermöglichen es, Daten aus unterschiedlichen Blickwinkeln zu analysieren. Man kann also mit denselben Variablen eine Zusammenhangsanalyse machen oder sie auf Unterschiede hin analysieren oder schauen, ob es Interdependenzen gibt, sie als Gruppen bilden. Die zugrundeliegenden Beziehungen in den Daten sind natürlich immer dieselben. Das liegt daran, dass Unterschiede durch Zusammenhänge entstehen und Zusammenhänge aufgrund von Unterschieden. Beides finden seine Ursache darin, dass Variablen und Fälle Gruppen bilden; und gleichzeitig entstehen die Gruppen durch die Zusammenhänge und Unterschiede.

Die Kennwerte, die aufgrund von Unterschiedsanalysen entstehen sind nicht sehr hoch verdichtet. Daher sind sie leichter zu lernen und für den Einstieg in die Statistik gut geeignet. Sie haben bereits Unterschiedsanalysen kennengelernt, die Masse (gesprochen Maße :-) der zentralen Tendenz auswerten, also zum Beispiel den t-Test für Mittelwertunterschiede zwischen zwei Gruppen. Wir können dabei Variablen aus verschiedenen Teilstichproben (Gruppe der Wähler:innen und Nichtwähler:innen) untersuchen, also «unabhängige Stichproben». Oder wir untersuchen «verbundene Stichproben», wenn zum Beispiel die Mittelwerte von zwei Variablen verglichen werden sollen, die jeweils für die ganze Stichprobe erhoben wurden (zB vor und nach einem experimentellem Eingriff aka Treatment). Oder wir untersuchen die Varianzen von Variablen mit Hilfe von χ^2 oder einem F-Test.

Wenn Sie genau auf die Grafik schauen, finden Sie den X^2-Test einmal bei den Unterschieden und einmal bei den «bivariaten» Zusammenhangsanalysen. Das liegt an der oben angesprochenen Verbundenheit der Konzepte: Unterschiede entstehen, wenn Dinge miteinander zusmamenhängen. Bei den Zusammenhangsanyalysen unterscheiden wir die «bivariten» von den «multivariten Modellen». Die bivariten bringen nur zwei Variablen in Beziehung zueinander, was sie einfacher macht, aber im Grunde zu einfach, um die komplexeren Zusammenhänge in unserer Welt zu erklären. Menschen sind einfach nicht bivariat und unsere Welt ist nicht monokausal. Die multivariaten Modelle sind Erweiterungen der bivariaten Analysemethoden. Bei den «Generalisierten Linearen Modellen» (GLM) geht es also weiter. Analysestrategien der GLM werden nach den Skalenniveaus der Variablen unterschieden, die erklärt werden sollen (also die abhängigen Variablen aka AV) und nach den Skalenniveaus der erklärenden (unabhängigen Variablen aka UV).

Die Analysemethoden sind dann einfacher, wenn das Skalenniveau hoch ist. Darum machen wir den Einstieg auch mit der Regression, bei der die AV und die UVs metrisch sind. Wenn die UVs nominal sind (bzw. nominale vorkommen), wird oft auch von Varianzanalysen (Analysis of Variance aka ANOVA) gesprochen. Wenn die AV nominal ist (dichotom oder polytom) werden logistische Regressionen gerechnet. Wenn Sie nach dem Bachelorstudium mit dem Master weitermachen, lernen Sie die multivariaten Analysemethoden

auf dem «Next Level» kennen – also zumindest einige davon. Wenn Sie dann auch noch in die Wissenschaft weitergehen, befassen Sie sich sicher spezialisierter mit bestimmten Verfahren der statistischen Datenanalyse, die für Ihre Forschung die am besten geeignete ist.

In diesem Semester werden wir uns auch mit Verfahren befassen, die Gruppierungen (aka Interdependenzen) untersuchen. Dazu gehört an erster Stelle die Faktorenanalyse, mit deren Hilfe Faktoren extrahiert werden sollen, die – so die Vermutung – die gemeinsame Ursache für gemessene Variablen sind. Die Idee ist also, dass manifest gemessene Variablen aufgrund von latenten Variablen miteinander zusammenhängen beziehungsweise korrelieren. Das ist schon an sich interessant genug. Darüber hinausgehend, können wir mit Hilfe einer Faktorenanalyse Indizes bauen, die mehrere Variablen auf einmal abbilden. Während die Faktorenanalyse Eigenschaften von Fällen auf zugrundeliegende Gemeinsamkeiten hin untersucht, werden mit Clusteranalysen Fallgruppen gebildet. Zum Beispiel könnten wir untersuchen, ob die Begeisterung und Abneigung gegenüber Mathemaitik, Statistik, Computer-Programmierung, R usw. einen gemeinsamen Kern haben, wie schlechter Matheunterricht oder Identitiätsbildung. Und dann könnten wir mit Clusteranalysen Gruppen identifizieren, je nachdem, wie gross die Begeisterung für Mathe is, für Computer und für Programmiersprachen wie R. Da gibt es sicher die einen und die anderen. Solche, die tollen Matheunterricht hatten und trotzdem mit R auf Kriegsfuss stehen usw. Also, Sie sehen, wir können viel damit anstellen. Das lohnt sich, auch wenn der Weg teils beschwerlich ist.

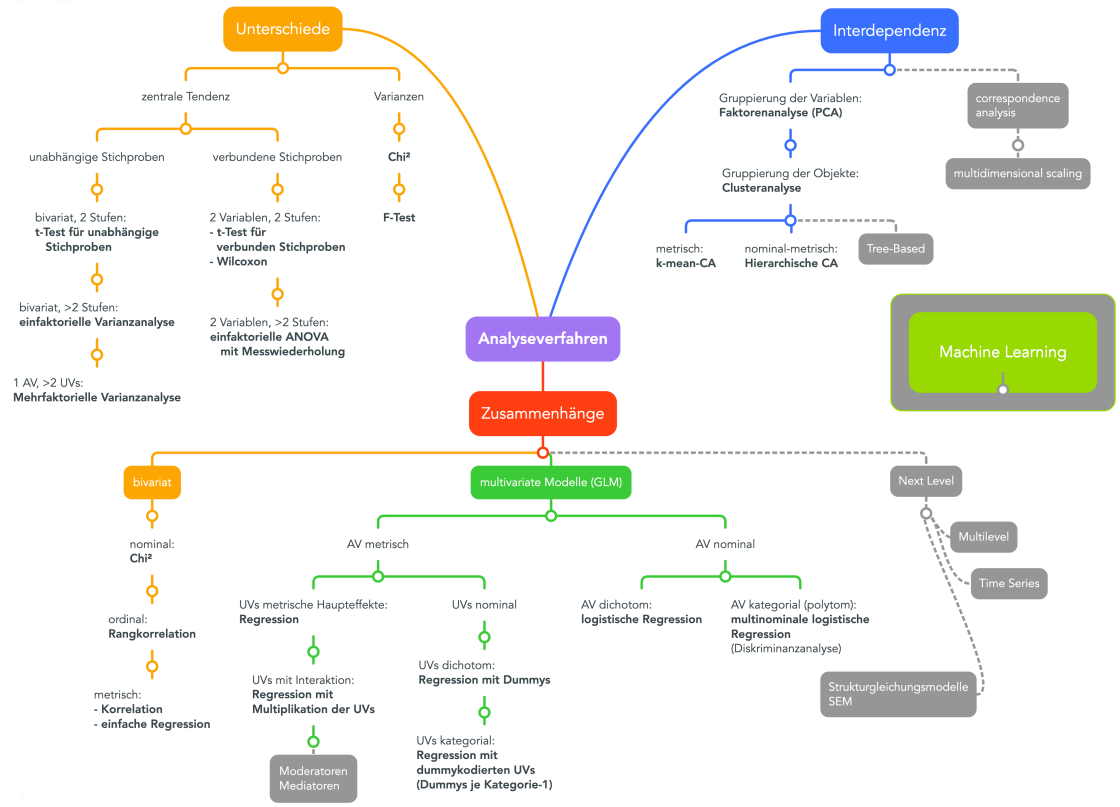


Abbildung 1: Systematik gesamt

Zitation dieser Seite

Zitation: Fretwurst, B. (2022). *Statistik und Datenanalyse: Aufbau. Begleittext zum Modul am IKMZ im HS22.* https://www.ikmz.uzh.ch/static/methoden/Statistik-Aufbau/. Abrufdatum: [aktuelles Datum].

1 Uni- und Bivariate Statistik

1.1 Univariate Statistik

Alle Statistik baut auf Massen zentraler Tendenz und Streumasse auf. Für das, was wir in diesem Semester tun, brauchen wir eigentlich nur den Mittelwert (Durchschnitt) und die Standardabweichung. Der Mittelwert entspricht der Frage: Wo stehen wir im Schnitt? Die Formel gibt an, dass eine Summe ($\sum$) gebildet wird für alle x_i (von i = 1 bis n) und diese durch die Anzahl der Fälle n geteilt wird beziehungsweise mit 1/n multipliziert wird:

$$\overline{x} = \frac{1}{n} \sum_{i}^{n} (x_i) \tag{1.1}$$

Die Standardabweichung zeigt uns, wie homogen die Fälle in einer Variable um den Mittelwert streuen. Das ist ein wichtiger Kennwert, weil derselbe Mittelwert zustande kommt, wenn alle Fälle genau dem Mittelwert entsprechen (die Standardabweichung wäre 0) oder wenn die Hälfte der Werte extrem weit links von der Mitte und die andere Hälfte genauso weit, aber rechts vom Mittelwert liegen (siehe **?@fig-homogenitaet**). Die Standardabweichung (s) ist die Wurzel aus der Varianz (s^2 oder «V»). Wenn Sie genau hinschauen, sehen Sie, dass die Varianz im Grunde der Mittelwert der quadrierten Abweichungen vom Mittelwert ist (vergleichen Sie die Formel mit der Mittelwertformel, wobei Sie gedanklich «$(x_i - \overline{x})^2$» durch «$\overline{x}$» ersetzen).

$$s^2 = V = \frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})^2 \tag{1.2}$$

$$s = \sqrt{\frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})^2} \tag{1.3}$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i}^{n} (x_i - \overline{x})^2 \tag{1.4}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i}^{n} (x_i - \overline{x})^2} \tag{1.5}$$

$$\tag{1.6}$$

IYI: Was das $\frac{1}{n-1}$ bei der Varianz mit Freiheitsgraden zu tun hat

Wenn wir die Varianz als durchschnittliche quadrierte Abweichung vom Mittelwert in der Grundgesamtheit bestimmen wollten und μ kennen würden, ergibt sich folgende Formel für die Varianz:

$$\sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2$$

Das Problem ist, dass wir μ nicht kennen, wenn wir eine Stichprobe gezogen haben. Um den unbekannten Parameter (der GG) μ schätzen zu können, ziehen wir den Mittelwert $\overline{x}$ als Stichprobenkennwert zur Schätzung von μ heran. Dann teilen wir aber nicht mehr durch n, sondern durch n-1. Das will ich hier erklären. Die Schätzformel für die Varianz der GG ist (wie oben):

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i}^{n} (x_i - \overline{x})^2 \tag{1.7}$$

Um deutlich zu machen, dass wir die Varianz der GG nur schätzen, setzen wir dem $\hat{\sigma}^2$ ein Dach auf. Die Formel ist schön kurz und knapp, aber eigentlich sieht man hier nicht gut, was durch das $\frac{1}{n-1}$ korrigiert wird. Ein bisschen klarer wird es, wenn wir deutlich machen, dass es hier einen Korrekturfaktor für die Stichprobenvarianz gibt, der gekürzt wurde. Eigentlich könnte man ja schreiben:

$$\hat{\sigma}^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i}^{n} (x_i - \bar{x})^2 \tag{1.8}$$

$$\hat{\sigma}^2 = \frac{n}{n-1} \cdot s^2 \tag{1.9}$$

Wenn wir den Korrekturfaktor $\frac{n}{n-1}$ separieren, dann ist der Rest die einfach Varianz der Stichprobe s^2. Den Korrekturfaktor bringe ich jetzt mal auf die andere Seite, weil ich damit besser weitermachen kann. Das tue ich, indem ich durch $\frac{n}{n-1}$ teile. Dann stelle ich links noch ein bisschen um.

$$\hat{\sigma}^2 = \frac{n}{n-1} \cdot s^2 \qquad | : \frac{n}{n-1} \tag{1.10}$$

$$\hat{\sigma}^2 \frac{n-1}{n} = s^2 \tag{1.11}$$

$$\hat{\sigma}^2 \frac{n}{n} - \frac{\hat{\sigma}^2}{n} = s^2 \qquad | + \frac{\hat{\sigma}^2}{n} \tag{1.12}$$

$$\hat{\sigma}^2 = s^2 + \frac{\hat{\sigma}^2}{n} \qquad | \sqrt{} \tag{1.13}$$

$$\hat{\sigma} = s + \frac{\hat{\sigma}}{\sqrt{n}} \tag{1.14}$$

$$\hat{\sigma} = s + s_{\bar{x}} \tag{1.15}$$

Nach ein paar Mal umstellen und dann noch die Waruzel aus allem ziehen, kommt raus, dass das Sigma der Grundgesamtheit durch die Standardabweichung in der Stichprobe s plus dem Standardfehler $s_{\bar{x}}$ berechnet wird. Durch die Ableitung wird gezeigt, dass hinter dem Korrekturfaktor, der an die Stichprobenvarianz angelegt wird (also $\frac{1}{n-1}$) eigentlich steht, dass wir die echte Standardabweichung σ in der GG um die Standardabweichung der Mittelwerte (aka Standardfehler) unterschätzen. Der Mittelwert wird je Stichprobe berechnet und wackelt daher, er hat eben seien Freiheiten. Da er ein Kennwert ist, der pro gedachter Stichprobe berechnet wird, hat, nimmmt der Mittelwert der Schätzung der Populationsvarianz σ^2 einen Freiheitsgrad weg, weshalb durch n-1 geteilt wird.
Es gibt Statistiker und Statistikbücher, die sagen, dass auch die Standardabweichung einer Stichprobe oder einer Grundgesamtheit GG ein sinnvoller Wert ist (zB, wenn man eine Lehrevaluation in einem Semester macht und nur etwas über eine Veranstaltung wissen möchte, dann ist das ja eigentlich eine Vollerhebung der GG der Teilnehmenden.) Diese Statistiker schreiben $s^2 = \frac{1}{n} \sum (x_i - \overline{x})^2$. Andere finden, dass die Standardabweichung eigentlich immer eine Schätzung für eine Populationsstreuung ist und schreiben daher immer $s^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$. Wer hat Recht? Is mir egal. Im Zweifel schreib ich lieber $\hat{\sigma}^2$, wenn ich deutlich machen will, dass ein Parameter geschätzt werden soll und s^2 nur, wenn es um die Stichprobe geht.
Im Tutorial schrittweise und graphisch erläutert:
https://www.youtube.com/embed/jXH1ly3Q87U

:::

1.1.1 Z-Transformation

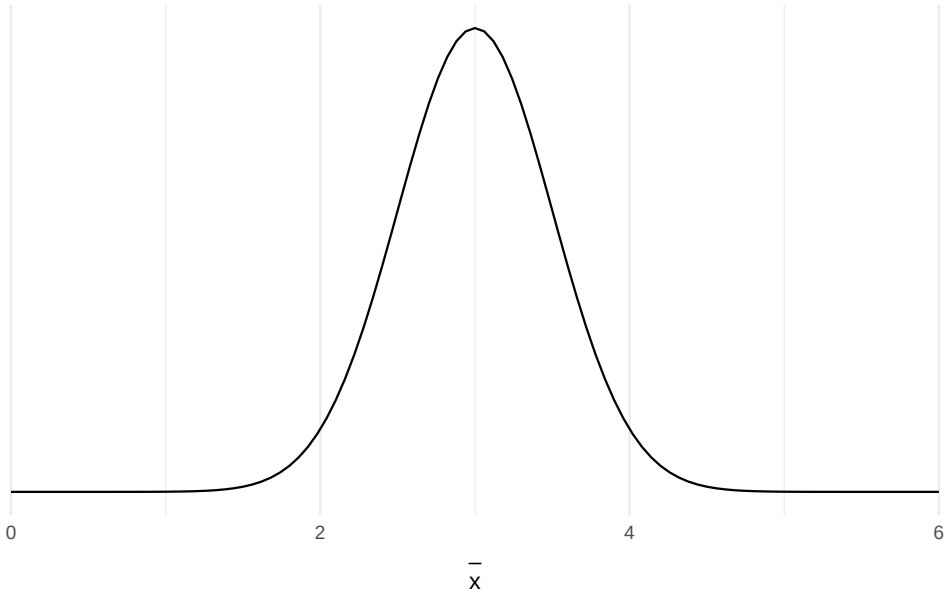
Die z-Transformation (auch «Standardisierung») einer Variable bedeutet, dass man sie so «verschiebt», dass sie um den Mittelwert 0 streut und zwar mit einer Standardabweichung von 1. Dazu wird von jedem Wert x_i der Mittelwert abgezogen und diese Differenz durch die Standardabweichung geteilt:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{1.16}$$

1.1.2 Normalverteilung und Standardnormalverteilung

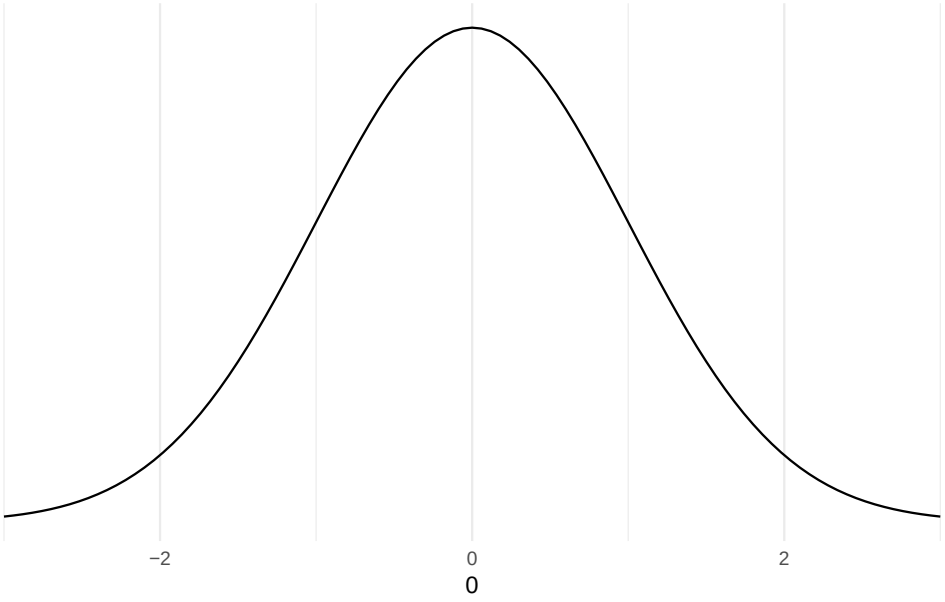
Mit dem Mittelwert und der Abweichung vom Mittelwert können wir Verteilungen beschreiben, die sich nach bekannten Gesetzen verteilen: den Wahrscheinlichkeitsgesetzen. Die Normalverteilung ist dabei die wichtigste Verteilung. Bei dieser gedachten Idealverteilung steht der Mittelwert genau in der Mitte und die Streuung ist symmetrisch um den Mittelwert herum verteilt (wie in Abbildung 1.1, mit dem Mittelwert 2,5 und der Standardabweichung 0,5).

Abbildung 1.1: Beispiel einer Normalverteilung ($\overline{x}$ = 3, s = 0,5)



Wegen dieser Eigenschaften, geben hier also der Mittelwert und die Standardabweichung die volle Information über die Art der Verteilung ab $N(\overline{x}, s_x)$. Die Standardnormalverteilung (nach der z-Transformation) hat einen Mittelwert von 0 und eine Standardabweichung von 1.

Abbildung 1.2: Die Standardnormalverteilung ($\overline{x} = 0$, s $= 1$)



1.1.3 Konfidenzintervalle für Mittelwerte

Konfidenzintervalle geben einen Wertebereich an, in dem die Parameter (GG) der Stichprobenkennwerte mit einer angebbaren Wahrscheinlichkeit liegen.

$$\text{KI:} \overline{X} \pm z_1 \cdot SE \tag{1.17}$$

$$\text{KI:} \overline{X} \pm z_1 \cdot \frac{s_x}{\sqrt{n}} \tag{1.18}$$

$$\text{KI}_{l.05} = \overline{x} - 1.96 \cdot \frac{s_x}{\sqrt{n}} \tag{1.19}$$

$$\text{KI}_{r.05} = \overline{x} + 1.96 \cdot \frac{s_x}{\sqrt{n}} \tag{1.20}$$

Mit dieser kleinen Onlineapp können Sie sich mal einen Eindruck davon verschaffen, wie Konfidenzintervalle reagieren. Schalten Sie sich mal «Show Confidence Intervals» und «Show True Parameter» an und stellen die «Number of Experiments» auf 20. Dann suchen Sie mal, wie viele Konfidenzintervalle den wahren Wert nicht enthalten:

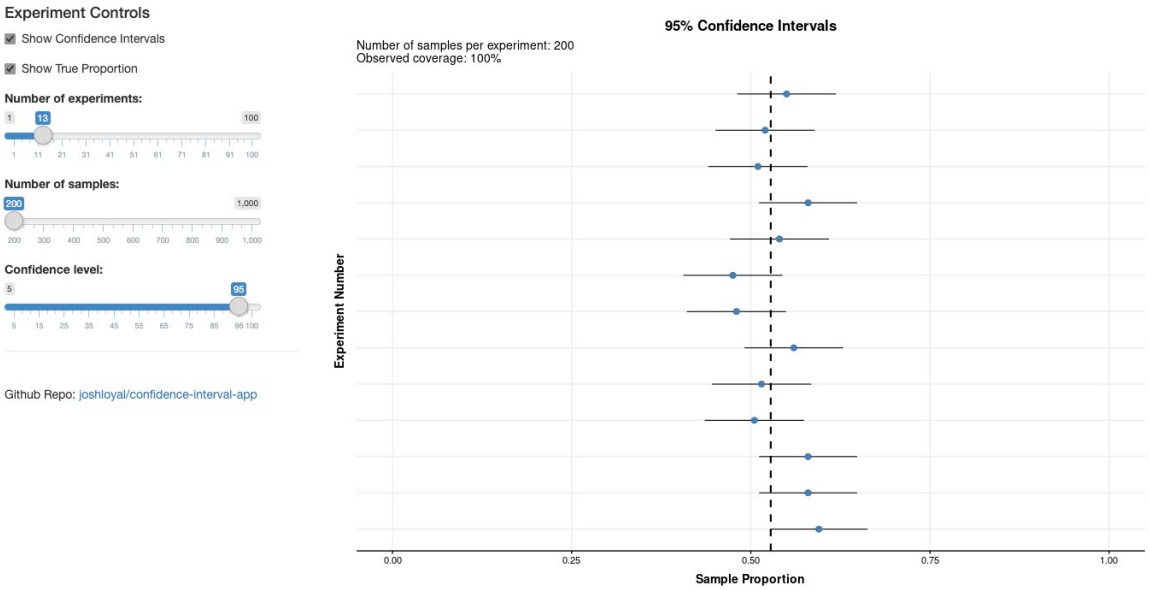


Abbildung 1.3: KI-App

Q&A: Ist es nicht so, dass das Konfidenzintervall bei grösserer und genauerer Sicherheit (zb eben 99% anstatt 95%) dann kleiner wird? Ich dachte grosser KI bedeutet kleine Sicherheit und kleines bzw schmales KI heisst grosse Sicherheit.

Eine sehr gute Frage, echt! Sie haben richtig erfasst, dass Sicherheit bei Konfidenzintervallen etwas mit der Wahrscheinlichkeit der Korrektheit einer Aussage (Sicherheitsniveaus) zu tun hat und gleichzeitig mit Genauigkeit der Aussage, also der Breite des Konfidenzintervalls. Diese beiden Konzepte sind aber gegenläufig! Wenn man in einer Schätzung mehr Sicherheit möchte, geht das nur auf Kosten der Genauigkeit. Stellen Sie sich vor, sie verabreden sich und sagen, dass sie es wahrscheinlich zwischen sechs und sieben schaffen. Ihrer Verabredung ist das aber nicht genau genug und möchte, dass Sie sagen, wann sie garantiert da sein können. Sie würden dann eher nicht das Intervall verkleinern und sagen, dass sie es ganz sicher zwischen 18:13 und 18:21 schaffen, sondern das Intervall eher verbreitern und vielleicht sagen, dass sie es sicher bis 20 Uhr schaffen werden. Wenn dann verlangt wird, dass sie genauer werden bei höherer Sicherheit würden Sie vielleicht vorschlagen, dass man sich ein andermal trifft. Was für ein Stress!
Also, wird eine höhere Sicherheit verlangt (99%-iges statt 95%-iges Signifikanzniveau), wird das Konfidenzintervall breiter.
Sie können das auch gedanklich durchspielen mit der Prognose der Höchsttemperatur für morgen und was passiert, wenn zum einen gefordert wird, dass die Prognose sehr genau ist (enges Intervall) und was passiert, wenn gefordert wird, dass die Prognose mit einer hohen Garantie auf Richtigkeit versehen wird (Signifikanz- bzw. Sicherheitsniveau hoch).

1.2 Bivariate Statistik

1.2.1 Kreuztabellen

Wenn wir Daten analysieren, dann können wir sie visualisieren. Eine sehr starke Form der Visualisierung sind Tabellen. Der Vorteil von Tabellen ist, dass sie sehr dicht an den Daten sind, wir also z.B. sehen können

dass, 57 Prozent der Haushalte noch über mindestens ein Radiogerät verfügen. In der Deutschschweiz (DS) sind es sogar 62 Prozent, während es in der Romandie (FS) 44 Prozent sind und im Tessin und den weiteren Teilen der italienischen Schweiz (IS) 63 Prozent. Man kann aus Kreuztabellen für die beiden berücksichtigten Variablen den ursprünglichen Datensatz rekonstruieren, weil man weiss wie viele Leute befragt wurden, wie viel Prozent aus der DS sind, aus der FS und aus der IS und wie viele Leute jeweils mindestens ein Radio haben. Irgendwann kann ich Ihnen mal zeigen, wie man aus einer Kreuztabelle eine Korrelation berechnen kann – man braucht nicht mehr. Also gut, die Tabellen enthalten viele Informationen und können, mit ein bisschen Anleitung, von fast jedem gelesen werden (anders als unserer schönen Regressions- oder Strukturgleichungsmodelle). Das Problem ist aber, dass man sehr schnell einschläft, wenn einem so eine Tabelle vorgelesen wird. Wenn man sie selbst liesst, ist es einfach sehr viel, worauf man schauen muss, was man vergleichen und dann vielleicht sogar noch texten muss. Eine Korrelation fast in einem direkt lesbaren Wert das Zusammen, was Sie vielleicht aus 5*5, also 25 Tabellenzellen mühsam herauslesen und dann immer noch nur so vom Gefühl her texten können. Tabellen sind also relativ voraussetzungsfrei interpretierbar, aber es ist sehr ermüdend und langwierig.

Beispiel für eine Kreuztabelle mit Link zu mehr (interaktiven) Kreuztballen:

Glaubwürdigkeit der Berichterstattung - Online-Nachrichtenseite

Antworten	Gesamt	Region			Alter			Bildung		
		DE	FR	IT	bis 34	35-54	ab 55	tief	mittel	hoch
(1) Gar nicht glaubwürdig	3%	3%	4%	0%	4%	4%	2%	5%	3%	3%
(2)	12%	11%	17%	7%	13%	13%	9%	7%	13%	13%
(3)	29%	29%	31%	27%	29%	29%	30%	29%	33%	26%
(4)	32%	33%	24%	43%	30%	31%	37%	36%	30%	32%
(5) Sehr glaubwürdig	23%	23%	24%	23%	23%	24%	22%	23%	21%	26%
Gesamt	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Mittelwert	3.6	3.6	3.5	3.8	3.5	3.6	3.7	3.6	3.5	3.6
Fehlermarge (95%)	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.1
Fehlende Angaben	1291	739	431	121	265	614	412	193	519	425
Valide Angaben	2462	1890	472	99	715	1299	448	234	1052	1089

Frage: (GLAUB_InfoOn): Wie glaubwürdig ist für Sie die Berichterstattung von [Online]? Skala: 1 «gar nicht glaubwürdig» bis 5 «sehr glaubwürdig»

Zwischenablage CSV Excel

1.2.2 Covarianz und Korrelation

Vor ein paar Zeilen hatte ich behauptet, man brauche als Kenngrössen in der Statistik nicht mehr als Mittelwert und Standardabweichung. Das sollte der Beruhigung dienen und ist ein bisschen gelogen. Ein Merkmal kann man mit diesen Kenngrössen in der Regel gut beschreiben, aber um das Zusammenspiel zweier Variablen beschreiben zu können, braucht es noch eine Kenngrösse: die Kovariation. Wieder etwas vereinfacht und doch nicht ganz falsch: fast alle multivariate Statistik baut auf Mittelwerten, Varianzen und Kovarianzen auf.

Kovarianz (cov oder auch C) sieht genauso aus, wie die Definition der Varianz, nur, dass nicht die Mittelwertabweichungen einer Variablen quadriert werden, sondern die Mittelwertabweichungen zweier Variablen multipliziert werden. Berechnet man die Kovarianz einer Variable mit sich selbst, kommt wieder die Varianz raus. Da Varianz und Mittelwert nicht standardisiert sind, ist auch die Kovarianz nicht standardisiert. Rechnet man mit standardisierten Variablen, kommt auch ein standardisierter Wert für die Kovarianz raus.

Da der für Vergleiche von unvergleichlicher Bedeutung ist, hat auch dieser Kennwert einen eigenen Namen bekommen: Korrelation (r).

$$cov = C = \frac{1}{n}\sum_{i}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{1.21}$$

$$r = \frac{\sum_{i}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n \cdot s_x \cdot s_y} \tag{1.22}$$

In der folgenden App werden Zusammenhänge illustriert. In dem Beispiel wurden Studierende gebeten, sich selbst in Bezug auf verschiedene Eigenschaften auf einer Skala von 0 bis 10 einzuordnen und was sie sich in Bezug auf dieselbe Eigenschaft von einer:m potentiellen Partner:in wünschen:

Wie man unschwer erkennen kann, hängen diese beiden Variablen miteinander zusammen – nicht perfekt (sonst wären alle Punkte auf der Geraden), aber es ist schon ein deutlicher Zusammenhang. Interessanter Weise ist der Zusammenhang bei Männern schwächer als bei Frauen. Schauen Sie doch mal, was Sie dort finden, vergleichen Sie auch mal die jeweiligen Durchschnittswerte der Selbsteinordnung und dem was sich so von Partnern gewünscht wird. :-)

1.2.3 Bivariate Regression

Die bivariate Regression ist im Grunde eine gehaltvollere Korrelation. Bei der Regression bestimme ich, im Unterschied zur Korrelation, was die abhängige Variable sein soll (AV) und was die unabhängige Variable sein soll (UV). Was die Regression bringt, sieht man in folgenden Formeln. Wenn wir in der ersten Zeile der Formeln @ref(eq:Mittelwert-Modell) davon ausgehen, dass wir keine UV haben, die die AV erklärt, dann haben wir als beste Information nur den Mittelwert. In dem Fall wäre unser Minimodell also: Die Werte in der Variable Y_i sind am besten durch ihren Mittelwert $\overline{Y}$ «erklärt». Übrig bleiben die Abstände zwischen den gefundenen Werten und diesem Minimodell, also dem Mittelwert von Y. In der zweiten Zeile @ref(eq:Regressions-gleichung) sind wir schon klüger und nehmen an, dass eine Variable X dafür verantwortlich ist, dass die Y_i so ausfallen, wie sie ausfallen. In der zweiten Zeile heisst es im Grunde: Die Werte in der AV Y sind abhängig von der UV X, wobei jeder Wert X mit einem b multipliziert werden muss und dann ein Rest e_i bleibt. Damit Y nicht 0 sein muss, wenn X 0 ist (b*0=0), kommt zu dem b_2 noch ein b_1. Fertig ist die bivariate Regressionsgleichung.

$$Y_i = \overline{Y} + e_i \tag{1.23}$$

$$Y_i = b_1 + b_2 X_i + e_i \tag{1.24}$$

$$\hat{Y}_i = b_1 + b_2 X_i \tag{1.25}$$

$$Y_i = \hat{Y}_i + e_i \tag{1.26}$$

Jetzt könnten wir noch die vorhergesagten $\hat{Y}$ (gekennzeichnet durch ein Dach) durch das Modell abbilden, wobei dann einfach kein e_i bleibt (dritte Zeile @ref(eq:Regressions-Modell)). In der untersten Formel @ref(eq:Varianz-Modell-Residuum) haben wir die in der Stichprobe gemessenen Y_i, die gleich gesetzt sind mit den geschätzten Y_i und dem Rest e_i.

Das b_2 gibt jetzt an, um wie viel Y grösser (oder kleiner) ist, wenn X um eine Einheit seiner Skala grösser wird. Die b's sind also skalenabhängig. Weil das nicht immer leicht zu interpretieren ist und uns oft die Skalierung nicht weiter interessiert, wurden die standardisierten Regressionkoeffizienten erfunden. Bei diesen steht im Grunde wieder eine z-Transformation im Hintergrund, die die Skala «herausrechnet», indem durch die Standardabweichung geteilt wird.

Die standardisierten Regressionskoffizienten geben einen Zusammenhang in Standardabweichungen an: Wenn x um eine Standardabweichung grösser ist, um wie viele Standardabweichungen ist dann y grösser (kann negativ sein)?

Sie sind definiert als: $BETA = b \cdot \frac{s_x}{s_y}$

Die BETAs[1] sind den partiellen Korrelationen sehr ähnlich: $+1$ ist ein perfekter positiver Zusammenhang, 0 kein Zusammenhang und -1 ein perfekter negativer Zusammenhang. Interpretieren würde ich ab 0.1, wenn sie signifikant sind.

1.3 Inferenzstatistik

Was in den Daten ist, die wir analysieren, das ist in den Daten – Punkt. Wir können Daten mit all diesen Tools untersuchen. Dafür brauchen wir natürlich vollständige Daten über das, was wir untersuchen wollen. Die haben wir aber oft nicht, weil es schlicht zu teuer ist und alle Menschen unfassbar nerven würde, wenn alle Sozialforscher:innen und Psycholog:innen und Wirtschaftswissenschaftler:innen usw. dauernd alle Menschen befragen würden. Das machen wir nicht, weil eine faszinierende Eigenschaft der Statistik ist, dass wir aus Teildaten Schlüsse auf die «Gesamt-Daten» ziehen können. Dieses Schliessen nennt man Inferenz (besser über das Englische zu merken: to infer). Dafür ist es notwendig, dass wir die Daten ohne bewussten Bias aus den Gesamtdaten entnommen haben.

Einen Bias ausschliessen können wir, wenn wir den Zufall walten lassen. Zufall ist nichts anderes als das Nicht-bewusst-oder-unbewusst-Auswählen. Wenn wir zufällig gezogen haben, dann haben wir eine gute Chance auf ein unverzerrtes Abbild der Grundgesamtheit, die uns interessiert. Dann berechnen wir die ganzen Kennwerte der deskriptiven Statistik für die Stichprobe und können anhand der Zufallsgesetze bzw. Wahrscheinlichkeitstheorie auf die Verteilung in der Grundgesamtheit schliessen. Dabei bleibt eine Unsicherheit, weil der Zufall zufällig ungünstig ausfallen kann (wir haben auch schon Pferde kotzen sehen).

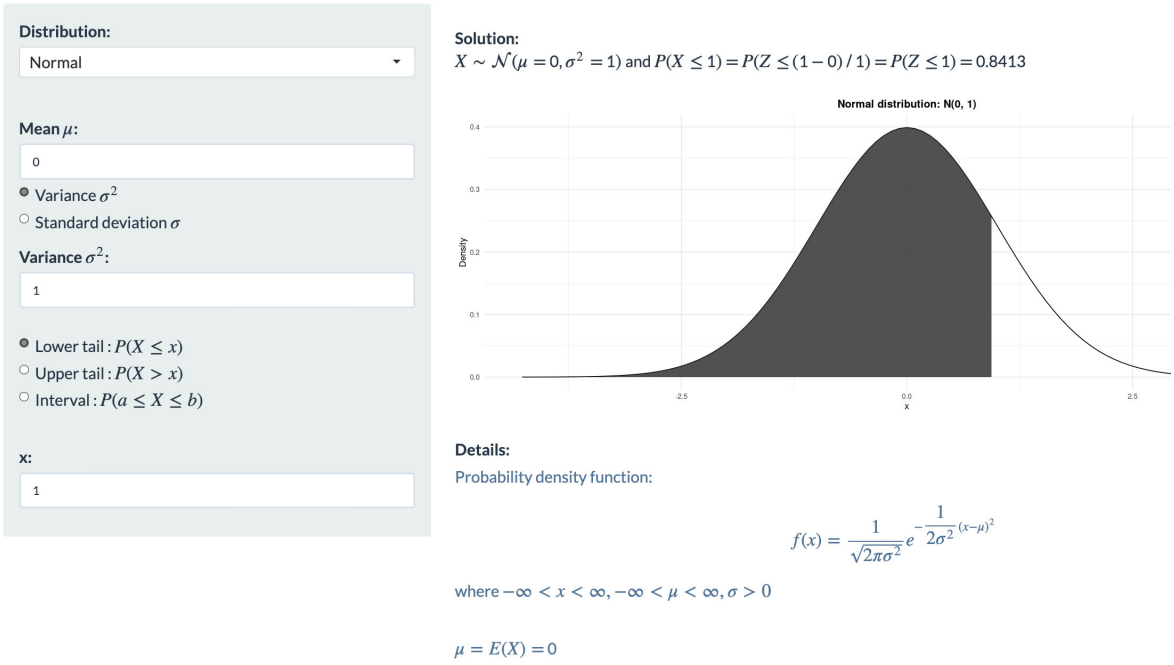
Wir können allerdings sagen, wie wahrscheinlich bzw. unwahrscheinlich es ist, dass wir extrem viel Pech hatten mit unserer Stichprobenziehung. Dabei sind wir sehr vorsichtig und schätzen und testen konservativ. Die Frage ist also immer: Kann ich, wenn ich ganz vorsichtig und konservativ bin, ausschliessen, dass ich sehr viel Pech bei der Ziehung meiner Zufallsstichprobe hatte? Das Pech bei der Zufallsziehung wird als Nullhpyothese bezeichnet (H0). Die sagt nämlich, wir haben eigentlich keinen Unterschied oder Zusammenhang und trotzdem in unserer Stichprobe den Unterschied (abweichend von 0) gefunden, den wir gefunden haben. Wenn wir – bei aller Vorsicht – sagen, dass ein Effekt (Unterschied oder Zusammenhang) so gross ist, dass die Nullhypothese so unwahrscheinlich wird (5% Irrtumswahrscheinlichkeit), dass wir sie ablehnen können, dann haben wir etwas gefunden, etwas Signifikantes!

Ob das signifikante Ergebnis unserer Theorie entspricht oder nicht, das müssen wir dann noch schauen, aber erstmal können wir festhalten, dass wir überhaupt etwas gefunden haben und nicht sagen müssen: Wir haben zwar einen positiven Zusammenhang gefunden, aber statistisch müssen wir die Nullhypothese beibehalten, dass der Zusammenhang positiv sein könnte, aber auch 0 oder negativ. Wenn wir nichts ausschliessen können, wissen wir nichts. Die H0 bedeutet also, dass wir statistisch nichts feststellen können und daher auch nicht zu Wissen gelangen. Wenn Sie irgendwo lesen, dass jemandem in seinen Analysen eine H0 zugute kommt oder die Annahme einer Nullhypothese als Erkenntnis gefeiert wird, seien Sie sehr skeptisch! Meistens liegt ein Denkfehler zugrunde: Dass ich 0 nicht ausschliessen kann, heisst nicht, dass ein Zusammenhang oder Unterschied nicht existiert. Wenn wir z.B. in einer Stichprobe einen Unterschied zwischen zwei Gruppen gefunden haben, dann ist die Wahrscheinlichkeit dafür, dass der Unterschied in der Grundgesamtheit 0 ist genauso gross, wie die Wahrscheinlichkeit, dass der Unterschied in der Grundgesamtheit doppelt so gross ist, wie der den wir gefunden haben. Das liegt daran, dass die Glockenkurve der Fehlerverteilung um einen gefundenen Wert symmetrisch um den gefunden Unterschied verteilt ist.

---

[1] Es ist recht ungünstig, dass die standardisierten Regressionskoeffizienten (fast) genauso heissen wie die Parameter β der b's. Ich versuche das verbal deutlich zu machen, indem ich «beeta» mit langem ee sage, wenn ich die standardisierten Regressionkoeffizienten BETA meine und «betta» sage (mit sehr kurzem e), wenn ich die griechischen β meine. Wenn Sie irgendwo ausserhalb eines Statistiklehrbuchs BETA lesen oder (sogar immer häufiger) β, dann verlassen Sie sich darauf, dass immer die standardisierten Regressionskoeffizienten gemeint sind und nie die Parameter der Regressionskoeffizienten in der Grundgesamtheit, weil wir die nie kennen werden und deshalb in empirisch wissenschaftlichen Publikationen so gut wie nie die Rede von unbekannten Parametern sein wird, sondern immer von den Kennwerten der Stichprobe, also den standardisierten Regressionskoeffizienten BETA. Didaktisch ist das blöd, aber es ist so und wenn ich jetzt was ganz Neues erfinde, dann bringt Ihnen das auch garnix.)

Mit dieser App lässt sich etwas mit der Normalverteilung experimentieren:



1.3.1 Hypothesentesten

Erkenntnistheoretisch prüfen wir unsere Hypothesen, um unser Wissen immer weiter abzusichern, bzw. Veränderungen festzustellen. Eine Theorie muss sich in der empirischen Sozialforschung immer wieder an neuen Daten bewähren. Wiedersprechen meine Daten immer wieder (replizierbar) der Theorie, wird sie modifiziert oder aufgegeben, wenn wir was Besseres haben.

Bei der statistischen Analyse gibt es neben unseren wissenschaftlichen Hypothesen auch rein statistische Fragen. Wenn wir einen Mittelwert in unserer Stichprobe finden, ist der sicher nicht identisch mit dem entsprechenden Parameter μ in der Grundgesamtheit (Population). Es stellt sich also immer die Frage nach möglichen Unschärfen. Über die statistischen Hypthesen können wir Aussagen treffen, wenn wir den Gesetzen der Wahrscheinlichkeit (Stochastik) eine Chance geben, also wenn wir Zufallsstichproben ziehen. Diese Zufallsstichproben ziehen wir aber nicht aus der Grundgesamtheit, also allen Objekten oder Subjekten, für die unsere Theorie Gültigkeit beansprucht. Wir ziehen Stichproben in einem abgesteckten Zeitrahmen und unter den Bedingungen der Machbarkeit. Wer kein Telefon hat, wird nicht angerufen. Um dieser Unterscheidung nachdruck zu verleihen, unterscheide ich Grundgesamtheit und Auswahlgesamtheit. Für letztere gilt: Jedes Element der Auswahlgesamtheit hat eine von 0 verschiedene Chance in die Stichprobe zu gelangen.

1. Könnte in der Auswahlgesamtheit der wahre Wert auch 0 sein, oder ein anderes Vorzeichen haben?
2. Die Nullhypothese ist eine statistische Hypothese gegen Falschentscheidungen aufgrund von Zufallsziehungen.
3. Nullhypothesen werden anhand von bekannten Verteilungen getestet.

Link zu einer guten App zum Probieren:

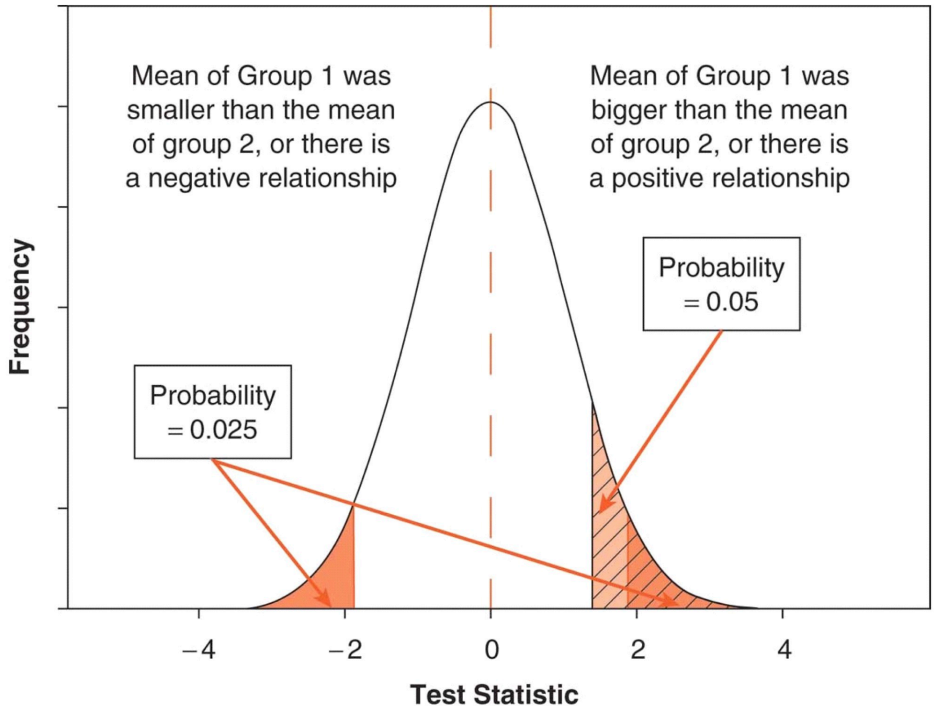
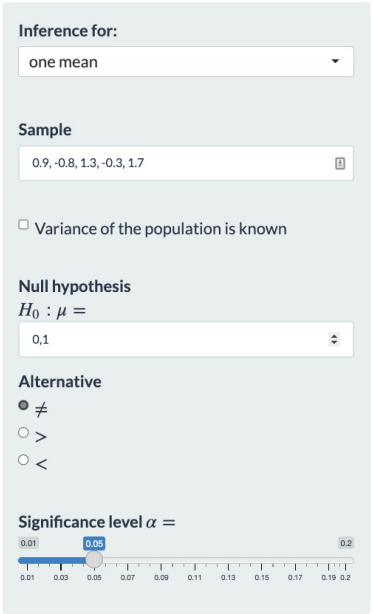


Abbildung 1.4: Hypothesentesten

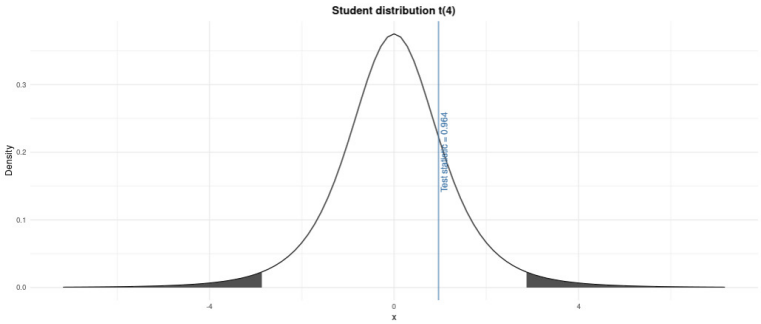


Data
0.9, -0.8, 1.3, -0.3, 1.7
$n = 5$
$\bar{x} = 0.56$
$s = 1.067$

Confidence interval *(two-sided)*
95% CI for $\mu = \bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 0.56 \pm (2.776 * 1.067 / 2.236) = [-0.765; 1.885]$

Hypothesis test
1. $H_0 : \mu = 0.1$ and $H_1 : \mu \neq 0.1$
2. Test statistic : $t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = (0.56 - 0.1) / 0.477 = 0.964$
3. Critical value : $\pm t_{\alpha/2,n-1} = \pm t(0.025, 4) = \pm 2.776$
4. Conclusion : Do not reject H_0

Interpretation
At the 5% significance level, we do not reject the null hypothesis that the true mean is 0.1 (p-value = 0.39).



2 GLM – Regression

2.1 Das lineare Modell (GLM)

2.1.1 Die Idee vom Modell

Für das Ergebnis der Datenerhebung wird ein Modell entworfen, das Zusammenhänge einfach darstellt. Da das Modell nie zu 100% das Ergebnis treffen wird, bleibt ein Rest, den wir Modellfehler oder einfach Fehler nennen.

> Grundmodell
>
> Ergebnis = (Modell) + Fehler
> Zum Beispiel ist der Mittelwert das einfachste univariate «Modell» einer Variablen, wobei $\overline{x}$ das Modell ist und die Abweichungen von x_i sind die Fehler: $x_i = \overline{x} + Fehler_i$.

Das lineare Modell ist die Basis von fast allem. Auch was Sie schon kennen, wird unter dem Konzept «lineares Modell» zusammengefasst:

- Varianzanalyse
- (Korrelation ist auch linear, aber eigentlich kein Modell)
- Regression

Das lineare Modell ist auch nicht auf lineare Zusammenhänge beschränkt. Es kann sehr gut mit kurvilinearen Zusammenhängen umgehen. Also, wenn zum Beispiel bei einer Gesamtnachrichtenlage mit sehr hohem Nachrichtenwert der Umfang des Medienkonsums steigt. Irgendwann erfährt diese Wirkung einen Deckeneffekt, weil niemand auf Dauer 24h am Tag Medien konsumieren kann. Vielleicht steigt der Nachrichtenkonsum mit dem Nachrichtenwert sogar Anfangs exponentiell (wie Coronazahlen) und hat dann bald einen Umkehrpunkt und strebt gegen ein mögliches Maximum. Selbst solche komplexeren Zusammenhänge können in einem linearen Modell dargestellt werden.

Die höhere Statistik wie Strukturgleichungsmodelle, Zeitreihenanalysen (Forcastings oder Laten-Growth-Curve-Modelle) bauen alle auf dem linearen Modell auf. Und auch Computational Science nutzt Modelle und zwar überwiegend als Basis die linearen Modelle.

2.2 Regression Einführung

Die Regression ist das einfachste und gleichzeitig mächtigste Werkzeug multivariater Datenanalyse. Aus den Kovarianzen mehrerer Variablen wird eine Funktion mit wenigen Kennwerten berechnet. Diese Kennwerte geben Auskunft darüber, wie etwas, das wir erklären wollen mit Dingen zusammenhängt, von denen wir glauben (hypothetisch annehmen), dass Sie Erklärungen liefern können. Man muss also vorher sagen, was man erklären will und womit man es erklären will. Die zu erklärende Grösse nennt man in der Sozialwissenschaft (und anderen Disziplinen): abhängige Variable (AV oder DV) und die Erklärungsgrössen nennt man: unabhängige Variablen (UV oder IV).

Die Regression baut auf Kovarianzen auf (bzw. Korrelationen, die wir uns besser vorstellen können). Die Regressionsgerade wird bei einer bivariaten Regression durch eine Konstante (in der Abbildung Abbildung 2.1 ist sie 1) und einem Anstieg je Variable gekennzeichnet (in der Abbildung ist es 0,5 für die eine UV = x).

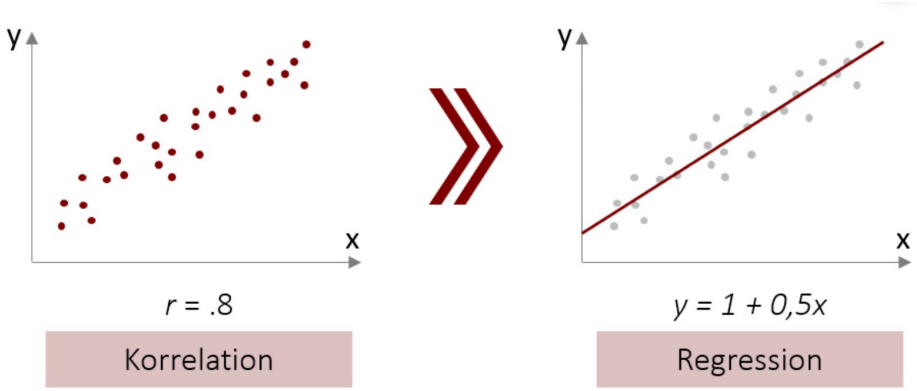


Abbildung 2.1

2.2.1 Notation der (multivariaten) Regression

Auch wenn es sinnvoll war, bei der bivariaten Regression die Konstante «a» zu nennen, soll ab jetzt die Bezeichnung der Regressionskoeffizienten etwas geändert werden:

Wir ändern die Notation etwas:

$$\begin{aligned} Y &= a + bX + e \\ &= b_1 + b_2 X_2 + e \end{aligned}$$

Table 4

Regression results using mpg as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	29.60**	[27.09, 32.11]						
disp	-0.04**	[-0.05, -0.03]	-0.85	[-1.05, -0.65]	.72	[.51, .81]	-.85**	
								R^2 = .718** 95% CI[.51,.81]

Note. * indicates $p < .05$; ** indicates $p < .01$. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights; *beta* indicates the standardized regression weights; sr^2 represents the semi-partial correlation squared; *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

Warum die Schreibweise ändern?

- In Tabellen (auch in R) steht die Konstante (a) in der Spalte der b's (estimates).
- Und weil wir im Multivariaten mehrere X und zugehörige b's haben, nummerieren wir sie durch, wobei wir in der ersten Zeile mit b_1 für die Konstante anfangen.
- Uuuuuund: In der Matrixschreibweise würde man B als Vector für die Regressionskoeffizienten nehmen, wobei die Konstante in der ersten Zeile steht.

Das bedeutet, dass bei einer bivariaten Regression zwei b's die Lage der Regressionsgeraden bestimmen:

Das ist zum einen die Konstante b_1 und zum anderen der Anstieg b_2 für die Gerade. In der Abbildung 2.2auf dieser Seite sehen Sie links zwei Regressionsgeraden mit unterschiedlichen b_2 (rot positiv und grün negativ). Auf der rechten Seite sehen Sie drei Regressionsgeraden mit unterschiedlichen b_1, wobei das b der roten Gerade am grössten ist (knapp 70), grün am kleinsten (bischen über 20) und blau in der Mitte liegt (knapp 40).

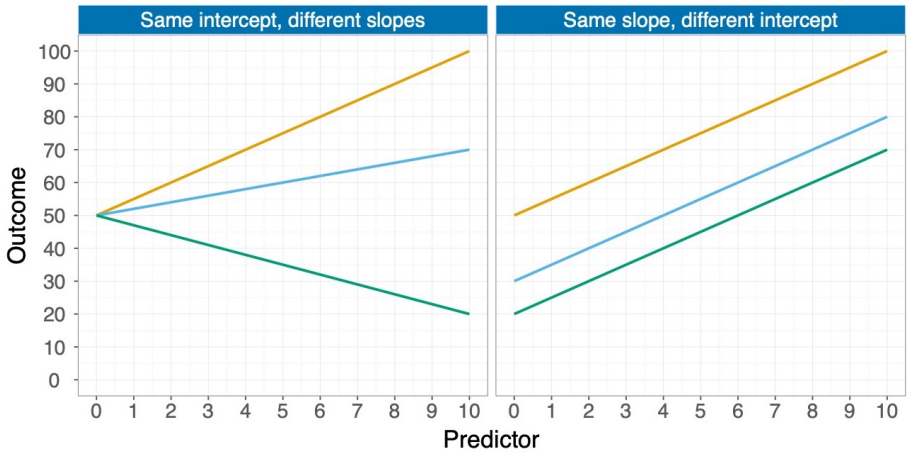


Abbildung 2.2: b's bei bivariaten Regressionen

2.3 Das Modell und die Regressionsgleichung als Schätzung

Die formelle Schreibweise eines Regressionsmodells enthält griechische Buchstaben um zu signalisieren, dass es sich hier um unbekannte Grössen, die Parameter in der Grundgesamtheit, handelt. So lange wir über die Qualität und die Eigenschaften von Regressionsrechnungen sprechen, wird uns der Unterschied zwischen β's und b's interessieren.

Als Gleichung heisst das, dass die Abhängige Variable Y_i durch eine gewichtete Summe (siehe Formel Gleichung 2.1) von einer oder mehreren unabhängigen Variablen erklärt wird. Diese UVs werden in der Regel mit X gekennzeichnet und weil es mehrere davon geben kann, werden sie durchnummeriert. Also mit dem Subscript i für das Durchzählen werden sie griechisch für die Parameter als $\beta_2 X_{i2}$ bezeichnet oder eben als $\beta_3 X_{i3}$ usw. Dann gibt es noch den Rest U_i. Das ist also das theoretische statistische Modell, dessen **Parameter** wir mit **Kennwerten** schätzen wollen.

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i \tag{2.1}$$

Wenn man mal genau schaut was nach der Stichprobenziehung eigentlich noch variabel ist, dann wird klar, dass die Y_i in der Datenerhebung gemessen wurden und damit Werte enthalten, die wir nicht mehr ändern. Das Gleiche gilt für die X_i-Werte der Variablen X_2 und X_3. Also sind diese Grössen eigentlich keine «Variablen» mehr, sondern längst durch echte Werte «fixiert». Zu schätzen sind nur die b's, also b_1, b_2 und b_3 (wie in Gleichung 2.2). Wenn wir die Regressionskoeffizienten, die b's in unserer Stichprobe, berechnet haben, müssen wir uns noch fragen, wie gut, also unverzerrt und genau sie die unbekannten Parameter (βS) messen, also – etwas technischer ausgedrückt – ob die b die β erwartungstreu und effizient schätzen. Dafür gibt es einige Voraussetzungen, die wir uns später [in Kapitel noch nicht da] noch anschauen werden. Am Ende der Formel steht das e_i für die Fehler, also den unerklärten Rest der Varianz, der zwischen den durch

das Modell geschätzten Werten (gekennzeichnet mit einem Dach als $\hat{Y}_i$) und den gemessenen Werten liegt. Während die s's die Schätzer für die βs sind, ist das e_i kein Schätzer für U_i. Das liegt daran, dass das e_i nur eine Fehlerstreuung in der Stichprobe ist und U_i viel mehr angibt, dass unberücksichtigte Einflussgrössen und ein stochastischer Rest nicht vom Modell abgebildet sind.

$$Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i \tag{2.2}$$

Bei einer Regression mit zwei UVs wird praktisch eine Ebene in die Punktwolke gelegt (siehe Abbildung 2.3). Wir schätzen aber eine multivariate Regression, damit wir bivariat interpretieren können, also je UV sagen, wie stark der Effekt auf die AV ist. Insofern interpretieren wir je Variable nur ein b, was dem Ansteig (Zusammenhang) einer UV mit der AV entspricht. Das können wir machen, weil die Statistik beziehungsweise unser Statistikprogramm die «Kontrolle» der übrigen Variablen übernimmt und wir schön die kontrollierte bivariate Beziehung interpretieren können. Die b's beschreiben dabei die Gerade, die die Ebene an der Stelle bildet, die für die andere Variable der Durchschnitt ist. Bei drei UVs spannen die b's zusammen eigentlich einen Raum auf, was sich aber niemand mehr visuell vorstellen kann. Die Statistik kann das aber und erledigt das so für uns, dass wir uns immer nur die Beziehungen anhand der jeweiligen b's der einzelnen UV's anschauen können.

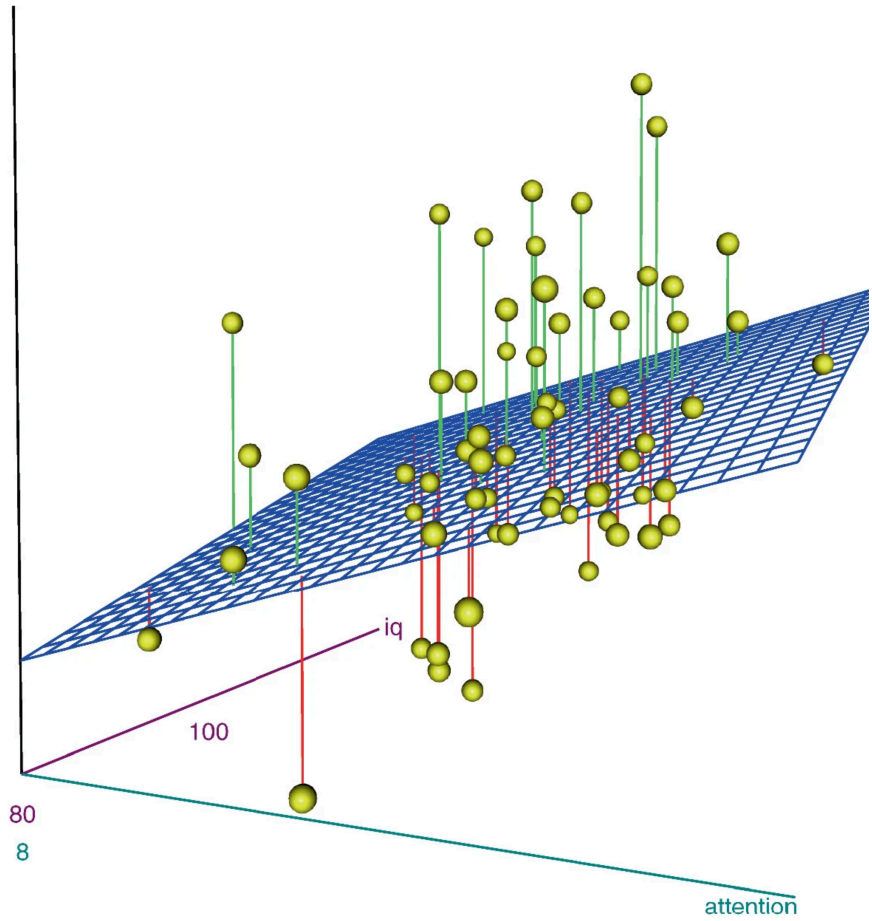


Abbildung 2.3: Regressionsebene bei zwei UVs

Hier wird eine Regression recht gut als Punktwolke visualisiert: http://shiny.calpoly.sh/3d_regression/.

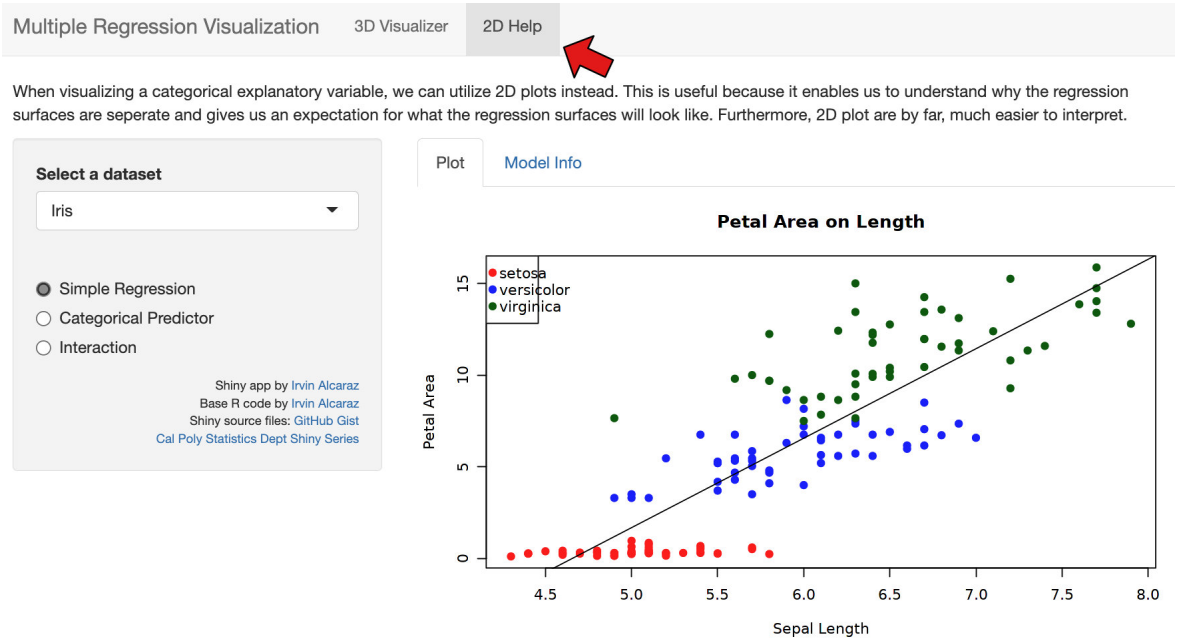


Abbildung 2.4: Regression

Zwischenaufgabe: Schreiben Sie die Formel für die einfache bivariate Regression auf?
Lösung anschauen

$$Y_i = b_1 + b_2 X_{i2} + e_i$$

Q&A: Wann genau bezeichnet man Kennwerte wie den Mittelwert, die Varianz oder die Standardabweichung als unbekannte Parameter? Also wann benutzt man die Symbole μ, σ^2 oder σ und was ändert das im Vergleich zu den Bezeichnung wie s und s^2?

μ, σ^2 oder σ sind Parameter der Grundgesamtheit. Das sind «Kennwerte» der Grundgesamtheit, die wir in der Realität allerdings eigentlich nie kennen. Deshalb wollen wir aus den Kennwerten einer Stichprobe (z.B. Mittelwert: $\overline{x}$, Varianz: s^2 etc.) auf die korrespondierenden Parameter der Grundgesamtheit (Mittelwert: μ, Varianz σ^2 etc.) schliessen.

- Der Mittelwert einer Stichprobe wird mit $\overline{x}$, die Varianz mit s^2 bezeichnet.
- Der Mittelwert in der Grundgesamtheit wird mit μ, die Varianz mit σ^2 bezeichnet

2.4 Die Regressionskoeffizienten b

2.4.1 Im bivariaten Modell

Das b ist im bivariaten Modell: $b = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}$. Wenn Sie genau hinsehen, erkennen Sie über dem Bruch den oberen Teil der Kovarianz $\sum(X-\overline{X})(Y-\overline{Y})$ und im unteren Teil der Varianz von X (ohne dass jeweils durch n geteilt wird). Nachem man das ein bisschen umgestellt hat, erhält man $r_{YX} \cdot \frac{s_y}{s_X}$ und wenn man durch $\frac{s_y}{s_X}$ geteilt hat, steht da, dass $r_{YX} = b \cdot \frac{s_X}{s_Y}$ ist. Im Grunde ist die Korrelation also ein standardisiertes b.

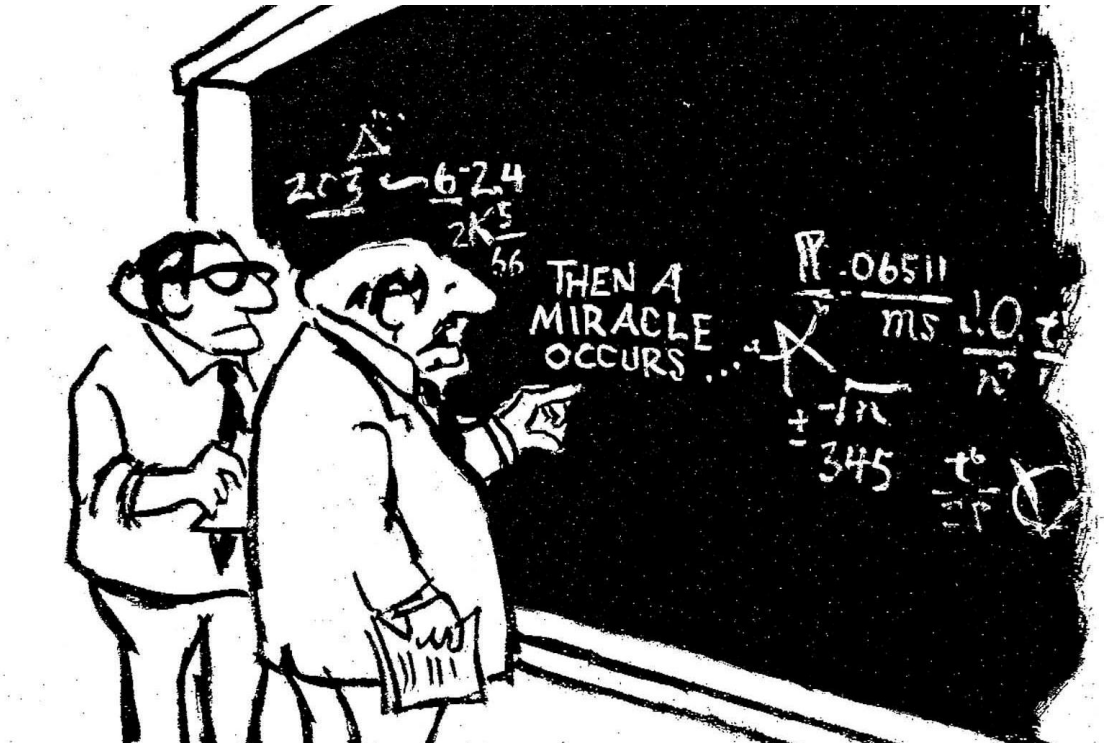
Regressionskoeffizient b (aka Steigungskoeffizient) und r

$$b = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} \tag{2.3}$$

$$= \text{... Then A Miracle Occurs ...} \tag{2.4}$$

$$= r_{YX} \cdot \frac{s_y}{s_X} \tag{2.5}$$

$$r_{YX} = b \cdot \frac{s_X}{s_Y} \tag{2.6}$$



2.4.2 Bei zwei UVs und zwei b's

Wenn wir mit Hilfe der Ordenary-Least-Squares-Methode (OLS) eine Formel für die b's bestimmt haben, kommt folgende Formel Gleichung 2.7 für das b_2 der Variable X_2[1] heraus:

[1] In den tiefergestellten Subscripten steht bei den r's immer nur 2. Das heisst r_{Y2} kennzeichnet die Korrelation zwischen Y und X_2.

$$b_2 = \frac{r_{y2} - r_{23} r_{y3}}{(1 - R^2_{2.3})} \frac{s_y}{s_2} \tag{2.7}$$

Die Formel hat es in sich. Aber schauen Sie sich die Formel mal ganz in Ruhe und stückchenweise an. Als eines der ersten Elemente taucht r_{Y2} auf, was so viel heisst, wie die einfache Korrelation zwischen Y und der ersten X-Variable (also x_2), die ja das b_2 hat und darum kurz und knapp nur noch mit dem Subscript 2 bedacht wird. Also hängt das b mit der Korrelation zwischen der zugehörigen X-Variable und Y zusammen. Da b skalenabhängig ist und r nicht, steht hinten noch dieses $\frac{S_Y}{S_2}$, also die Standardabweichung von Y geteilt durch die Standardabweichung von X_2. Dieser Bruch sorgt nur dafür, dass b in der Skala von Y angegeben ist (darum auch multipliziert mit s_Y) – den Teil können Sie schon mal vergessen.

Interessanter ist der zweite Teil der Gleichung über dem Bruchstrich: Wir ziehen da das Produkt aus r_{23} und r_{Y3} ab. Das heisst, wir gehen von der bivariaten Korrelation aus, rechnen jetzt aber noch die Korrelation raus, die die beiden unabhängigen Variablen X_2 und X_3 untereinander haben. Wir ziehen allerdings nicht einfach r_{23} ab, sondern multiplizieren das auch noch mit r_{Y3}. Das bedeutet, wir haben einen Zusammenhang r_{y2} und rechnen aus dem den Anteil gemeinsamer Varianz, also der Zusammenhänge der Varialbe x_2 heraus (wir subtrahieren sie), die diese mit X_3, wobei wir nur so viel rausrechnen, wie die dritte Variable X_3 wiederum mit Y gemeinsam hat. Wären die beiden Variablen X_2 und X_3 unkorrelliert, dann wäre auch das Produkt $r_{23} r_{Y3} = 0$, weil $0 \cdot r_{Y3} = 0$. Wenn X_2 und X_3 korrellieren, aber X_3 und Y nicht, dann würden wir auch nichts von r_{Y2} abziehen. Im Storchenbeispiel würden wir also sagen, wir sehen den Zusammenhang zwischen Geburtenrate und Anzahl Störche. Wir müssen aber von dieser Korrelation abziehen, dass die Drittvariable X_3 «Bevölkerungsdichte» (Stadt vs. Land) stark mit der zu erklärenden Geburtenrate Y korrelliert und diese mit der Anzahl der Störche (X_2), die in einer Region leben. Wie wir wissen, ist diesem Fall $r_{23} r_{Y3} \approx r_{Y2}$ und darum der kausale Zusammenhang zwischen Storchenpopulation und Geburtenrate nicht gegeben.

2.4.3 Der standardisierte Regressionskoeffizient

Der standardisierte Regressionskoeffizient BETA (nicht β) entspricht im Fall der bivariaten Regression der Korrelation r_{YX}.

Der standardisierte Regressionskoeffizient BETA aka b*

$BETA = b \cdot \frac{s_X}{s_Y} = r_{YX}$
Die standardisierten Regressionskoffizienten geben einen Zusammenhang in Standardabweichungen an: Wenn x um eine Standardabweichung grösser ist, um wie viele Standardabweichungen ist dann y grösser (kann negativ sein)?
Die BETAs sind den Korrelationen sehr ähnlich: +1 ist ein perfekter positiver Zusammenhang, 0 kein Zusammenhang und -1 ein perfekter negativer Zusammenhang. Interpretieren würde ich ab 0.1, wenn sie signifikant sind.

Q&A: Was ist der Unterschied zwischen “b”, “β”“,”std. b”, “b*” und “BETA”? In Statistik Einführung wurde der standardisierte Regressionskoeffizient mit β (ungut bis falsch) bezeichnet.

Die anhand von Stichproben berechneten Regressionskoeffizienten heissen immer und überall die B’s oder b’s (das hat noch ein bisschen was mit Matrizen zu tun, mit den Sie sich aber nicht befassen müssen - die Gross-/Kleinschreibung ist also egal). Die b schätzen aber Werte in der GG und wenn wir über Voraussetzungen sprechen und Erwartungswert oder Erwartungstreue, dann müssen wir über die Parameter reden, die mit den b’s geschätzt werden sollen. So wie zum Beispiel die Standardabweichung s in der Stichprobe den Parameter schätzt, schätzen die b’s das was in der ökonometrischen Literatur schon immer mit bezeichnet wird. Die ’s sind also unbekannte und unstandardisierte Parameter der Grundgesamtheit, die mit Hilfe der b’s geschätzt werden sollen.

Viel später haben sich ein paar Idioten gedacht, dass es doch super praktisch wäre, wenn man die standardisierten Regressionskoeffizienten auch BETA nennen würde. (Das kommt meines Wissens vor allem von SPSS, die das in ihren Outputs auch wirklich so als BETA geschrieben haben und nicht als griechisches). Die standardisierten Regressionskoffizienten haben die Eigenschaften wie Korrelationskoeffizienten und gehen also von -1 bis $+1$ und 0 wäre kein Zusammenhang. Das ist also ein anderer Wert als das b und hat mithin auch nichts mit dem Parameter zu tun. (Es gibt sogar Pakete in R, bei denen die Autoren die unstandardisierten Regressionskoeffizienten b als bezeichnen). Es ist also immer eine Challenge auch in Tabellenoutputs von R oder in Veröffentlichungen herauszufinden, ob da b's oder BETAs stehen (die unbekannten zu schätzenden können nie in einer Spalte mit Werten stehen, da sie ungekannt sind). Üblich sind noch die Bezeichnungen «std. b», «b*», «BETA» und eben leider auch viel zu oft das «β». Wenn ein β in einer Tabelle mit Wersten steht, dann kann es eigentlich nur der standardisierte Regressionskoeffizient sein, weil wir die tatsächlichen Parameterwerte der GG nie kennen.
In der Prüfung kann es sein, dass ich Ihnen eine Regressionstabelle gebe und Sie für zwei Spalten (ohne Spaltenüberschrift) angeben müssen, in welcher Spalte die b's stehen und in welcher die standardisierten b's. Leichter ist das, wenn in einer Spalte Werte mit kleiner -1 oder grösser $+1$ vorkommen, weil das nur die b's können. Wenn in beiden Spalten nur Werte zwischen -1 und $+1$ stehen, ist es schwieriger. Dann wird der Hinweis darin liegen, dass ein sehr kleiner Wert vorkommt (wie wenn zB das Alter in Jahren als UV integriert ist) und dennoch einen sehr kleinen p-Wert hat. Es kann nämlich nicht sein, dass ein Effekt (std. b) sehr klein ist und dennoch signifikant, obwohl andere deutlich grössere Werte alle nicht signifikant sind. In der Spalte mit den sehr kleinen b und sehr kleinem p-Wert (Signifikanz) stehen also die b's.

2.4.4 Signifikanz der b's und BETAs

Die b's und BETAs können daraufhin geprüft werden, ob sie signifikant von 0 verschieden sind. Dafür wird ein t-Test gemacht, wie er auch bei Mittelwertvergleichen oder der Korrelation verwendet wird. Der t-Test spuckt dann einen p-Wert aus und wenn der kleiner ist als .05, dann sagen wir, dass ein b (oder auch sein BETA) signifikant von 0 verschieden sind. Das bedeutet, dass die Wahrscheinlichkeit kleiner als 5 Prozent (0.05) ist, dass das b rein zufällig so gross ist, wie es eben ist. Die Frage lautet also: Bei so einem gegebenem b, können wir da mit hoher Wahrscheinlichkeit ausschliessen, dass das währe β in Wirklichkeit 0 ist oder sogar das entgegengesetzte Vorzeichen hat (bei positivem b also kleiner ist als 0)?

Fehlende Signifikanz sagt nicht, dass das wahre $\beta = 0$ ist!

Wenn wir die Nullhypothese nicht zurückweisen können, weil der t-Test nicht signifikant ist, heisst das nicht, dass das zugehörige wahre β gleich 0 ist. Es heisst nur, es könnte 0 sein. Es ist sogar so, dass die Wahrscheinlichkeit dafür, dass das wahre $\beta = 0$ ist, genauso gross ist, wie die Wahrscheinlichkeit dass β doppelt so gross ist, wie das b, das wir gefunden haben.

Q&A: Könnten Sie noch einmal einen kurzen Input zu den Signifikanz-Kennwerten der Regression machen?

Mach ich! UND SPOILER:
Spoiler zu t- und p-Wert: Wenn Sie einen Effekt (z.B. einen Mittelwertunterschied zwischen zwei Gruppen oder eine Korrelation oder ein b einer Regression) daraufhin untersuchen wie weit der Wert von 0 verschieden ist (Nullhypothese), dann kann diese Differenz zu 0 durch eine Standardisierung als Wert angegeben werden (z-Wert, t-Wert, Chi^2 oder F-Wert). Aus den bekannten Verteilungen dieser Werte kann abgelesen werden, wie wahrscheinlich es ist (p-Wert), zB einen t-Wert zu bekommen, der so gross ist (oder grösser als der t-Wert den Sie ihn in ihrer Stichprobe gefunden haben), obwohl der Wert in der Grundgesamtheit eigentlich 0 ist. Kurz: Ein b kann in einen t-Wert umgerechnet werden,

zu dem ein p-Wert einer Wahrscheinlichkeitsverteilung gehört, der angibt, wie Wahrscheinlich es ist, das konkrete t (und somit das b) zu bekommen, wenn der Effekt in Wirklichkeit 0 ist. Ist p klein (zB kleiner als .05 bei 95%-Signifikanzniveau), also die Richtigkeit der Nullhypothese unwahrscheinlich, dann sagen wir b ist signifikant von 0 verschieden. Wenn p grösser ist als .05, würden wir sagen, dass wir Null nicht ausschliessen können und selbst Parameter mit entgegengesetztem Vorzeichen nicht zu unwahrscheinlich wären. Wir behalten also die Nullhypothese bei und müssen zugeben, dass wir keine statistisch klare Aussage treffen können – traurig, aber fertig t-Test.

2.5 Das Bestimttheitsmass R^2

Das Bestimmtheitsmass gibt an, wie gut die Werte der AV durch die Werte der UV vorhergesagt werden können. Im

Wie viel von der Varianz der AV durch ein Modell aufgeklärt werden kann, stellt man fest, indem zunächst die Summe der quadrierten Abweichungen (**S**um of **S**quares) für alle Y_i Werte gezählt werden. Also die totale Varianz der AV, die geschrieben wird als SS_T (Sum of Squares Total). Jetzt ist die Frage, wie viel von dieser Sum of Squares Total durch die Sum of Squares des Modells (SS_M) erklärt werden kann. Darum setzen wir diese beiden Summen der Quadrate (wenn man jeweils durch n teilen würde, wären das die Varianzen) ins Verhältnis zueinander und bekommen einen Prozentwert. Also rechnen wir $\frac{SS_M}{SS_T}$ und bekommen einen Wert zwischen 0 und 1 bzw. 0% und 100% (% heisst ja «von Hundert» bzw. «geteilt durch 100»). Das ist der aufgeklärte Varianzanteil und den nennen wir R^2.

- SS_T: Summe der quadrierten Abweichungen für die AV (Y).
- SS_M: Summe der quadrierten Abweichungen des Modells (der Punkte auf der Geraden, bzw. die geschätzten $\hat{Y}_i$-Werte).

Also: $R^2 = \frac{SS_M}{SS_T}$

Bei dieser Gleichung 2.8 können wir durch n teilen, also über und unter dem Bruch $1/n$ ergänzen und hätten:

$$R^2 = \frac{SS_M/n}{SS_T/n} \tag{2.8}$$

Was in Worten ausgedrückt bedeutet:

$$R^2 = \frac{\text{aufgeklärte Varianz}}{\text{Gesamtvarianz}} \tag{2.9}$$

In der Abbildung 2.5 ist im ersten Quadrat die Abweichung der gemessenen Werte vom Mittelwert von Y dargestellt. Das ist die Summe der quadrierten (Sum of Squares SS_T) insgesamt (total) der AV, also Y. Würde man, wie oben, durch n teilen, wäre das einfach die univariate Varianz von Y. Wo X dabei liegt, ist völlig unbedeutend, da die Schätzung von Y an jeder Stelle von X gleich ist, also die Gerade parallel zu der x-Achse verläuft. Im zweiten Quadrat sind die Abstände zwischen der Regressionsgeraden und den tatsächlichen Werten dargestellt. Diese Abstände geben die Fehler wieder, die wir auch Residuen nennen, weshalb ihre Quadratsumme mit SS_R gekennzeichnet wird. Im letzten Quadrat ist die Summe der Abstände zwischen dem 0-Modell (orange Linie parallel zur X-Achse) und den geschätzten Werten, also denen, die auf der Modell- beziehungsweise Regressionsgeraden liegen. Deren Quadratsumme wird als SS_M gekennzeichnet.

Mit dem Bestimmtheitsmass können wir angeben, wie gut ein Modell insgesamt ist. Wir werden später noch diskutieren, wie sinnvoll das ist.

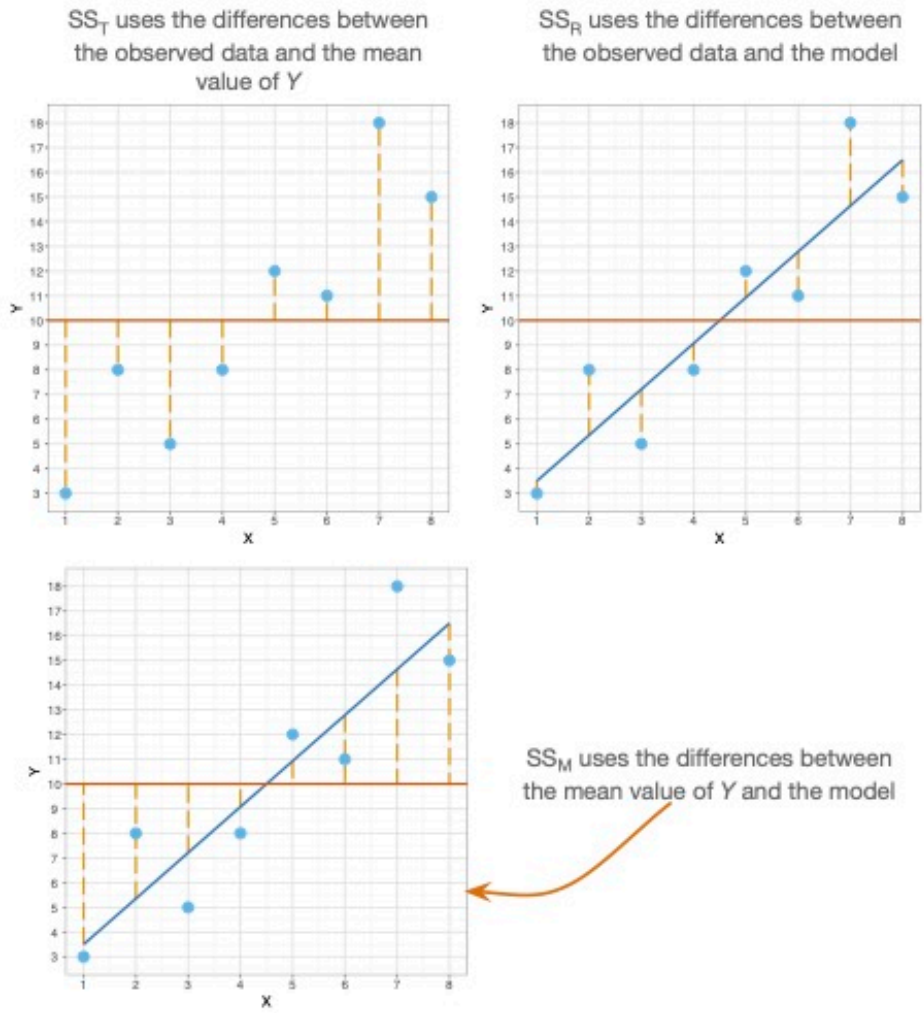


Abbildung 2.5: R-Quadrat

Spoiler

R^2 ist nicht immer sehr sinnvoll, weil es eigentlich mehr eine Stichprobeneigenschaft ist und wenig über die Welt sagt und recht einfach hochgeschraubt werden kann, indem man triviale und langweilige Variablen in ein Modell einbaut.

2.5.1 Das korrigierte R^2

Wenn man in ein Regressionsmodell mehr und mehr UVs aufnimmt, dann kann sich R^2 nur vergrössern, weil die jeweils bestehende Aufklärung der AV nicht verkleinert, wenn man noch fragt, was eine weitere Variable für die Aufklärung der AV leisten kann. Im Gegenteil: Es wird selbst dann ein bisschen von der AV erklärt, wenn es garkeinen Zusammenhang zwischen einer UV und der AV gibt. Es gibt also zufällige «Varianzaufklärung» (die in 95% der Fälle nicht signifikant ist und uns daher eigentlich egal sein könnte). Wenn man aber etliche UVs ins Modell aufnimmt, die alle keinen Zusammenhang mit der AV haben, kann es sein, dass R^2 irreführend gross wird. Darum korrigiert man bei kleinen Stichproben das R^2 ein bisschen um die Anzahl der UVs (die Anzahl der UVs wird mit «k» gekennzeichnet.).

$R^2_{adj.}$

$R^2_{adj.} = R^2 \cdot \frac{n-k-1}{n-1}$

Wenn unser Stichprobenumfang n klein ist und k ähnlich gross, dann ist $n-k-1$ deutlich kleiner als $n-1$ und damit der Korrekturfaktor $\frac{n-k-1}{n-1}$ klein, was zu einer starken Korrektur von R^2 führt.

Besipiel

Nehmen wir als Beispiel eine Stichprobe von nur 30 Fällen. Unser R^2 sei mal 0.2, bei k $= 10$ UVs im Modell. Dann ist das $R^2_{adj.} = 0.13$, also von 0.2 deutlich nach unten korrigiert. Hätten wir stattdessen 3000 Fälle, wäre $R^2_{adj.} = 0.199$. Bei (normal) grossen Stichproben tut diese Korrektur also selbst dann nichts, selbst wenn wir einige UVs ins Modell aufgenommen haben.

2.5.2 Der F-Test zum R^2

F-Test (R^2)

Gibt an, ob durch das Modell insgesamt überzufällig gut Varianz von der AV aufgeklärt wurde. Also, ob die Nullhypothese zurückgewiesen werden kann, dass die AV nicht durch sämtliche UVs im Modell erklärt werden kann.

3 GLM – BLUE

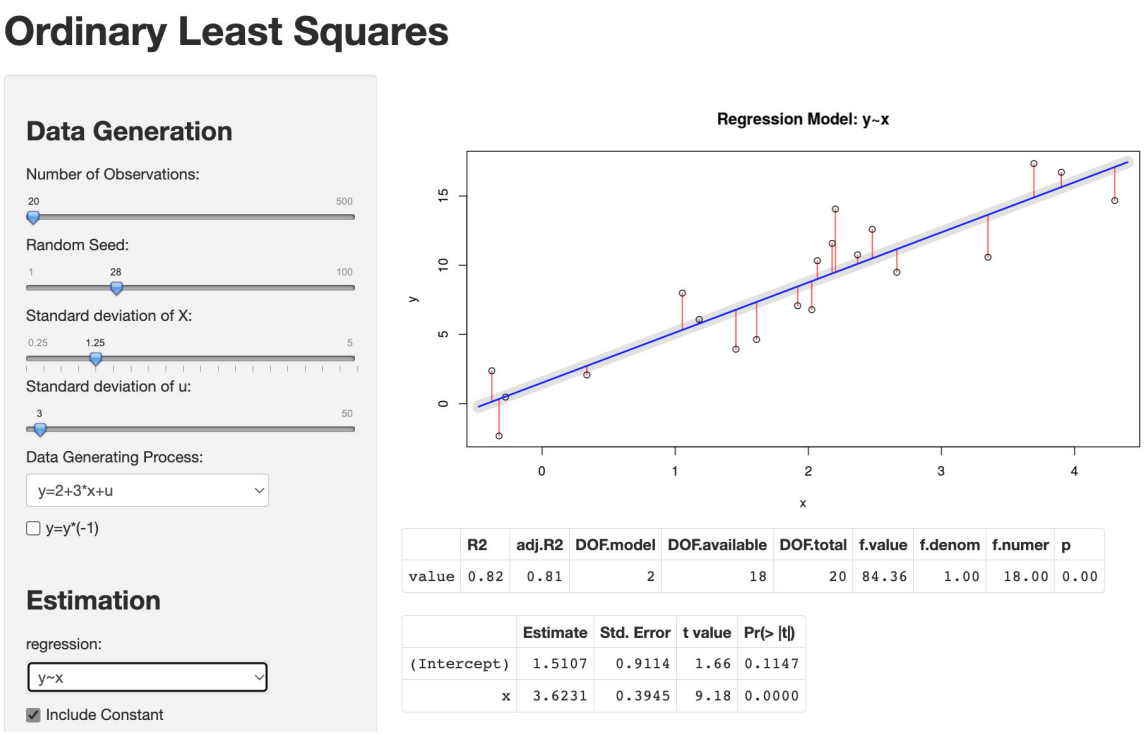
```
## Loading required package: viridisLite
```

Der Vorlesungsmitschnitt

3.1 OLS

Eine der einfacheren und grundlegenden Methoden um die b's zu bestimmen ist die Methode der kleinsten Quadrate bzw. OLS, was das Akronym für **O**rdinary **L**east **S**quares ist. Mit dieser Methode legt die Mathematik eine Gerade in eine Punktwolke, weil sie es nicht visuell und intuitiv machen kann. Das Prinzip ist recht einfach: Man versucht b's zu finden, für die die Fehler möglichst klein sind. Das ist im Grunde die Optimierungsaufgabe der OLS-Methode. Genau das machen wir auch, wenn wir eine Gerade in eine Punktwolke legen, wir bauen sie so ein, dass sie «optimal reinpasst» also die Abstände zu den einzelnen Punkten minimal sind.

Sehr gut hier zum anschauen und spielen:



	R2	adj.R2	DOF.model	DOF.available	DOF.total	f.value	f.denom	f.numer	p
value	0.82	0.81	2	18	20	84.36	1.00	18.00	0.00

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	1.5107	0.9114	1.66	0.1147
x	3.6231	0.3945	9.18	0.0000

Abbildung 3.1: OLS-App

Als Beispiel hatte ich in der Vorlesung gebracht, dass man auch mal überlegen könnte, welcher Wert eine Verteilung einer Variablen optimal repräsentieren würde. Wenn wir dieses Optimierungsproblem an OLS übergeben würden, dann würden wir sagen: Suche einen Wert a aus allen möglichen a-Werten, der für eine Variable x die kleinsten quadrierten Abstände hat. Damit es OLS versteht würden wir schreiben: OLS bitte minimiere folgende Gleichung: $\sum_i (x_i - a)^2$

Jetzt wissen wir, dass die quadrierten Abweichungen gross sein müssen, wenn a links vom Optimum liegt und immer kleiner wird, wenn wir uns dem optimalen a-Wert annähern. Dann wird die Summe der quadratischen Abstände wieder grösser. Also haben wir eine Funktion, die einer quadratischen Funktion folgt (dass die so aussieht, müssen wir garnicht wissen, aber es hilft vielleicht der Vorstellung). Wenn wir wissen wollen, wo diese Funktion ihr Minimum hat, dann können wir die Funktion ableiten und dann nach der Nullstelle der abgeleiteten Funktion suchen. An der Stelle liegt dann der a-Wert, der die Streuung einer jeden Variablen optimal abbildet, weil wir diese Ableitung völlig abstrakt und ohne konkrete Werte gemacht haben und sie daher immer gilt. Also:

$$\frac{df}{da} = \sum_i (x_i - a)^{2'} = 0 \tag{3.1}$$

$$0 = \sum_i [x_i^2 - 2x_i a + a^2]' \tag{3.2}$$

In der ersten Zeile das df/da bedeutet, dass abgeleitet (differenziert) werden soll und zwar die Funktion f nach a. In der zweiten Zeile sehen wir dann schon die Ableitung nach Ableitungsregeln (wer extrem Bock hat, kann sich die ja nochmal angucken) und gleich auch schon mit 0 gleichgesetzt.

In der nächsten Zeile **??**eq-Umstellen1) wird ein bischen aufgelöst und umgestellt (müssen Sie nicht können).

$$0 = -2\sum_i x_i + 2na \qquad | : 2n \; | + \sum_i x_i \tag{3.3}$$

$$\frac{\sum_i x_i}{n} = a \tag{3.4}$$

$$a = \overline{x} \tag{3.5}$$

Am Ende kommt als Lösung für den nach OLS besten Repräsentanten einer Variablen heraus: $\frac{\sum_i x_i}{n} = a$ (**??**)eq-Umstellen2). Der linke Teil ist genau die Definition von $\overline{x}$, also dem Mittelwert. Damit haben wir mit einer Ableitungen der OLS herausgefunden, dass der Mittelwert die kleinste Summe der quadrierten Abstände jedes Wertes zu einem Wert a hat, also der gesuchte beste Repräsentant für eine Variable der Wert $a = \overline{x}$ ist (**??**)eq-Mittelwert-Optimum). Dasselbe könnten wir für die Formel $Y_i = b_1 + b_2 X_i + e_i$ machen. Wenn wir (mit ein paar Annahmen) das für jedes b_1 bis b_3 machen würden, dann hätten wir die b's mit OLS bestimmt. Da das ungleich komplizierter ist als für den Mittelwert, schlage ich vor, wir lassen das an dieser Stelle.

IYI (Klausur): Ableitung der OLS-Funktion

Wir suchen mit Hilfe der OLS-Funktion die b's für das Modell:

$$Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i \qquad | \; -(b_1 + b_2 X_{i2} + b_3 X_{i3}) \tag{3.6}$$

$$e_i = Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3} \tag{3.7}$$

Da es gleich um die e_i gehen wird, haben wir schon mal die Regressionsgleichung nach e_i umgestellt. Ausgangspunkt für die Ableitung der OLS-Funktion ist die Idee, den vom Modell nicht erklärten Rest, also die Residuen (e_i) zu minimieren. Die Residuen sind wie folgt definiert und können nach Formel (3.7) auch als Umstellung der Regressionsgleichung geschrieben werden:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3} \tag{3.8}$$

Die Residuen werden minimmiert, wenn die Summe der quadrierten Fehler aka Residuen (auch Error e) minimiert werden, also die Summe $\sum_{i=1}^{n} e_i^2$ möglichst klein ist. Für die Summe der quadrierten Fehler (Sum of Squared Errors: SSE) können wir schreiben:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_1 - b_2 X_2 - b_3 X_{i3}\right)^2 \tag{3.9}$$

Wenn wir die Summe der quadrierten Fehler (SSE) minimieren wollen, leiten wir die SSE nach den gesuchten b ab (das ∂ steht für differenzieren, also ableiten; das ∂b unter dem Bruchstrich bedeutet, dass nach b abgeleitet wird und nicht etwa, dass irgendwie durch b geteilt wird):

$$\frac{\partial SSE}{\partial b} = \frac{\partial \left(\sum_{i=1}^{n} e_i^2\right)}{\partial b} \tag{3.10}$$

$$= \frac{\partial \left(e_1^2\right)}{\partial b} + \frac{\partial \left(e_2^2\right)}{\partial b} + \cdots + \frac{\partial \left(e_i^2\right)}{\partial b} \tag{3.11}$$

$$= \sum_{i=1}^{n} \frac{\partial \left(e_i^2\right)}{\partial b} \tag{3.12}$$

Nach den Ableitungsregeln kann man die Ableitung einer Summe zerlegen in die Ableitung der einzelnen Summanden. Das steht in (3.12). Nach den ableitungsregeln kann man daraus Folgendes machen (schauen Sie nur darauf, wonach jeweils abgeleitet wird. Alle anderen Teile fallen weg. In (3.8) wird zB nach b_1 abgeleitet, und die Ableitung einer Konstanten (b_1) ist 1 und mit dem Minuszeichen davor, bleibt eben -1 übrig. In (3.12) wird nach b_2 abgeleitet. Darum bleibt aus der Formel $b_2 X_{i3}$ übrig, was nach Ableitungsregeln X_{i2} entspricht und wieder mit einem Minuszeichen aus der Formel versehen ist.):

$$\frac{\partial \left(e_i^2\right)}{\partial b} = \frac{\partial \left(e_i^2\right)}{\partial e_i} \frac{\partial e_i}{\partial b} = 2 e_i \frac{\partial e_i}{\partial b} \tag{3.13}$$

und nach Gleichung (3.8),

$$\frac{\partial e_i}{\partial b_1} = \frac{\partial \left(Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3}\right)}{\partial b_1} = -1, \tag{3.14}$$

$$\frac{\partial e_i}{\partial b_2} = \frac{\partial \left(Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3}\right)}{\partial b_2} = -X_{i2}, \tag{3.15}$$

$$\frac{\partial e_i}{\partial b_3} = \frac{\partial \left(Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3}\right)}{\partial b_3} = -X_{i3}. \tag{3.16}$$

Jetzt müssen alle Formeln von Gleichung 3.14 bis Gleichung 3.16 zusammengefügt und die einzelnen Ableitungen gleich 0 gesetzt werden, um die Gesamtfunktion zu minimieren:

$$\frac{\partial SSE}{\partial b_1} = 2 \sum_{i=1}^{n} e_i \frac{\partial e_i}{\partial b_1} = 2 \sum_{i=1}^{n} \left(Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3}\right)(-1) \tag{3.17}$$

$$= -2 \sum_{i} Y_i + 2 \sum_{i} b_1 + 2 b_2 \sum_{i} X_{i2} + 2 b_3 \sum_{i} X_{i3} \tag{3.18}$$

dafür können wir schreiben:

$$-\sum_i Y_i + nb_1 + b_2 \sum_i X_{i2} + b_3 \sum_i X_{13} = 0 \tag{3.19}$$

Jetzt wollen wir die SSE noch für bzw. nach b_2 ableiten:

$$\frac{\partial SSE}{\partial b_2} = 2\sum_{i=1}^{n} e_i \frac{\partial e_i}{\partial b_2} = 2\sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3})(-X_{i2})$$

oder nach der Zerlegung der Summe in die einzelnen Summanden, die jeweils mit $-X_{i2}$ multipliziert wird und sich darum immer das Vorzeichen umkehrt.

$$-\sum_i Y_i X_{i2} + b_1 \sum_i X_{i2} + b_2 \sum_i X_{i2}^2 + b_3 \sum_i X_{i3} X_{i2} = 0 \tag{3.20}$$

Nun fehlt nur noch die Ableitung der SSE nach b_3:

$$\frac{\partial SSE}{\partial b_3} = 2\sum_{i=1}^{n} e_i \frac{\partial e_i}{\partial b_3} = 2\sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3})(-X_{i3})$$

und wie bei b_2:

$$-\sum_i Y_i X_{i3} + b_1 \sum_i X_{i3} + b_2 \sum_i X_{i2} X_{i3} + b_3 \sum_i X_{i3}^2 = 0. \tag{3.21}$$

Jetzt teilen wir jeweils die Gleichung 3.19 bis Gleichung 3.21 durch die Fallzahl, also n, woraus sich ergibt (etwas konventionaller geschrieben und erstmal übersichtlicher):

$$b_1 + a_1 b_2 + a_2 b_3 = c_1, \tag{3.22}$$
$$a_1 b_1 + a_3 b_2 + a_4 b_3 = c_2, \tag{3.23}$$
$$a_2 b_1 + a_4 b_2 + a_3 b_3 = c_3, \tag{3.24}$$

wobei sich hinter den a's und c's folgende Elemente verbergen, die am Ende eigentlich immer recht einfach ($\bar{X}_2$ und so) ausfallen:

$$\begin{gathered} a_1 = \frac{1}{n}\sum X_{i2} = \bar{X}_2, \quad a_2 = \frac{1}{n}\sum X_{i3} = \bar{X}_3, \quad a_3 = \frac{1}{n}\sum X_{i2}^2, \\ a_4 = \frac{1}{n}\sum X_{i2} X_{i3}, \quad a_5 = \frac{1}{n}\sum X_{i3}^2, \\ c_1 = \frac{1}{n}\sum Y_i = \bar{Y}, \quad c_2 = \frac{1}{n}\sum Y_i X_{i2}, \quad c_3 = \frac{1}{n}\sum Y_i X_{i3}. \end{gathered} \tag{3.25}$$

Durch Einsetzten erhalten wir also:

$$\bar{Y} = b_1 + b_2 \bar{X}_2 + b_3 \bar{X}_3 \quad \text{umgestellt} \quad b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 \bar{X}_3 \tag{3.26}$$

und Gleichung 3.27 sowie Gleichung 3.21 sind

$$\bar{X}_2 b_1 + \left(\frac{1}{n}\sum X_{i2}^2\right) b_2 + \left(\frac{1}{n}\sum X_{i2} X_{i3}\right) b_3 = \frac{1}{n}\sum Y_i X_{i2} \tag{3.27}$$
$$\bar{X}_3 b_1 + \left(\frac{1}{n}\sum X_{i2} X_{i3}\right) b_2 + \left(\frac{1}{n}\sum X_{i3}^2\right) b_3 = \frac{1}{n}\sum Y_i X_{i3}. \tag{3.28}$$

Wenn man jetzt das b_1 aus Gleichung 3.26 einsetzt, ergibt sich

$$b_2\left(\frac{1}{n}\sum X_{i2}^2 - \bar{X}_2^2\right) + b_3\left(\frac{1}{n}\sum X_{i2}X_{i3} - \bar{X}_2\bar{X}_3\right) = \left(\frac{1}{n}\sum Y_i X_{i2} - \bar{Y}\bar{X}_2\right) \tag{3.29}$$

$$b_2\left(\frac{1}{n}\sum X_{i2}X_{i3} - \bar{X}_2\bar{X}_3\right) + b_3\left(\frac{1}{n}\sum X_{i3}^2 - \bar{X}_3^2\right) = \left(\frac{1}{n}\sum Y_i X_{i3} - \bar{Y}\bar{X}_3\right) \tag{3.30}$$

Die Varianzen der Variable X ist $\left[V_X = (1/n)\sum X_i^2 - \bar{X}^2\right]$ und die Kovarianz von X und Y ist $\left[C_{XY} = (1/n)\sum X_1 Y_i - \bar{X}\bar{Y}\right]$, also kann man für die (3.30) etwas übersichtlicher schreiben:

$$b_2 V_{X_2} + b_3 C_{X_2X_3} = C_{YX_2}, \quad b_2 C_{X_2X_3} + b_3 V_{X_3} = C_{YX_3}$$

Das ist damit auch das Ergebnis der ganzen Ableitung: Die b's lassen sich aus den Varianzen und Kovarianzen der Variablen bestimmen!
Um eine noch übersichtlichere Schreibweise zu bekommen, lassen wir jetzt noch die Subscripte der ganzen X weg. Also schreiben wir ddie Varianzt von X_2 nicht mehr als V_{X_2}, sondern einfach als V_2 und die Kovarianz zwischen X_2 und X_3 statt $C_{X_2X_3}$ als C_{23}. Dann vereinfacht sich das Ganze für b_2 zu:

$$b_2 = \left(V_3 C_{Y2} - C_{23} C_{Y3}\right) / \left(V_2 V_3 - C_{23}^2\right). \tag{3.31}$$

und für b_3:

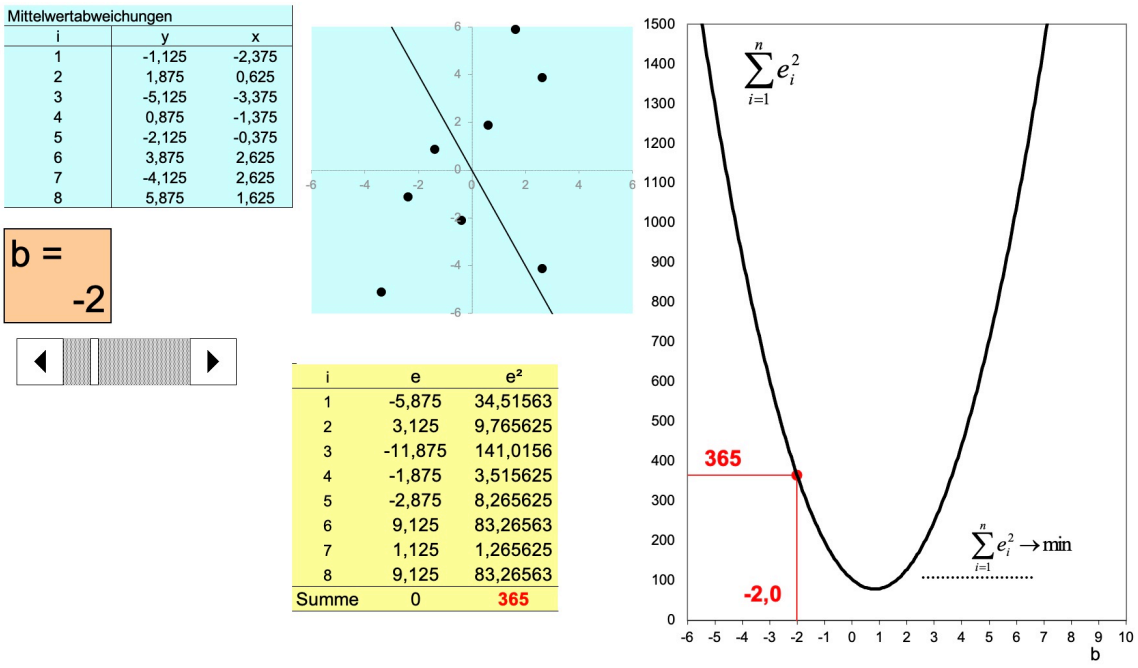
$$b_3 = \left(V_2 C_{Y3} - C_{23} C_{Y2}\right) / \left(V_2 V_3 - C_{23}^2\right). \tag{3.32}$$

Und weil die Korrelelation $r_{Y2} = C_{Y2}/S_2 S_Y$ ist und die Varianz $V = S^2$, kann man für die **?@eq-2.15** kann man, statt der Covarianzen und Varianzen, Korrelationen schreiben:

$$b_2 = \frac{\left(V_3 C_{Y2} - C_{23} C_{Y3}\right)}{\left(V_2 V_3 - C_{23}^2\right)} = \frac{r_{Y2} - r_{23} r_{Y3}}{\left(1 - r_{23}^2\right)} \frac{S_Y}{S_2}. \tag{3.33}$$

(Wer Lust hat, zeigt, dass das die (3.33) stimmt.)

Ich habe Ihnen eine Excel-Datei gebaut, mit der Sie sich das Prinzip von OLS interaktiv anschauen können:



Mittelwertabweichungen		
i	y	x
1	-1,125	-2,375
2	1,875	0,625
3	-5,125	-3,375
4	0,875	-1,375
5	-2,125	-0,375
6	3,875	2,625
7	-4,125	2,625
8	5,875	1,625

i	e	e²
1	-5,875	34,51563
2	3,125	9,765625
3	-11,875	141,0156
4	-1,875	3,515625
5	-2,875	8,265625
6	9,125	83,26563
7	1,125	1,265625
8	9,125	83,26563
Summe	0	365

Abbildung 3.2: OLS-xlsx

Welche Funktion und Eigenschaften hat OLS
Versuchen Sie es in Ihren Worten.

3.2 Vorraussetzung für BLUE

Damit unsere b's aus der OLS die besten linearen unverzerrten Schätzer (BLUE:**B**est **L**inear **U**nbiased **E**stimator) für die βs sind, müssen ein paar Voraussetzungen erfüllt sein. Diese Voraussetzungen gucken wir uns in diesem Kapitel an. Zusammengefasst sind es:

V1. Die UVs und die AV dürfen keine Konstanten sein.
V2. Das Skalenniveau der UVs muss metrisch oder dichotom (0/1) sein.
V3. Die Werte der X müssen fix sein.
V4. Das Modell muss voll spezifiziert sein. D.h.: Keine Korrelation mit externen Variablen.
V5. Es darf keine **perfekte** oder **heftige** Multikollinearität geben.
V6. Die Residuen müssen bei jedem Wert jeder UV gleich streuen (Homoskedastizität).
V7. Die Residuen müssen grob normalverteilt sein.
V8. Die Residuen dürfen nicht autokorreliert sein.

Was verbirgt sich hinter demm Akronym BLUE (ausgeschrieben)?
Lösung anschauen
Best **L**inear **U**nbiased **E**stimator

Q&A: Was ist mit $E(b_2)$ gemeint? Erwartungstreue und Erwartungswert erklärt:

Ein Koeffizient soll seinen zugehörigen Parameter unverzerrt schätzen, also zB b_2 das β_2 oder auch $\overline{x}$ das μ. Da wir Zufallsstichproben ziehen, sind die Koeffizienten zu ihren Parametern sehr selten genau gleich. Vielmehr streuen die Kennwerte um den wahren Parameter, wenn man viele Stichproben aus einer Grundgesamtheit zieht. Diese Streuungen der Kennwerte um den wahren Wert des Parameters kennen Sie als Normalverteilung.
In der Abbildung 3.3 ist ein fiktives Beispiel für einen Parameter als rote Linie dargestellt (es könnte ein $\overline{x}$ oder ein b_2). Der soll geschätzt werden. Wenn wir eine Stichprobe ziehen, kommt ein Wert der blauen Linie heraus. Wenn wir das elendig oft machen, kommt eine Verteilung raus, die in der Regel die Form einer Normalverteilung hat, wie sie in der Abbildung dargestellt ist. Wenn diese Kennwertverteilung um den wahren Parameter symmetrisch verteilt ist, also ihr Maximum gleich dem Parameter ist, sprechen wir von einem «erwartungstreuen» Kennwert. In der Abbilung ist der Kennwert nicht erwartungstreu. Sie fragen sich vielleicht, wann soetwas vorkommt. Das passiert, wenn zum Beispiel eine verzerrte Stichprobe gezogen wird. Nehmen wir an, es die Wahrscheinlichkeit geschätzt werden, dass Leute abstimmen gehen. Die liegt im Beispiel der Abbildung 3.3 bei 50% der Abstimmungsberechtigten. Wenn wir allerdings eine Befragung gemacht hätten, wo wir ganz repräsentativ die gesamte Wohnbevölkerung befragt hätten, wären viele Befragte in der Stichprobe, die nicht abstimmungsberechtigt wären. Wir hätten also praktisch unser Modell unterspezifiziert, weil wir nicht berücksichtigt haben, dass es in der Wohnbevölkerung zwei Gruppen gibt, wobei eine abstimmungberechtigt ist und die andere nicht.
Der **Erwartungswert** ist der Mittelwert der Kennwertverteilung. Die Schreibweise $E(b_2)$ kommt nun daher, dass wir fragen, ob ein Kennwert **erwartungstreu** ist, also ob der Erwartungswert gleich dem zu schätzenden Parameter ist, also bei b_2 das β_2. Ob ein Kennwert erwartungstreu ist, können wir für jede Art von Kennwert fragen, also zum Beispiel Korrelationskoeffizienten, Covarianz, Varianz, Standardabweichung, Regressionskoeffizienten standardisiert oder nicht.

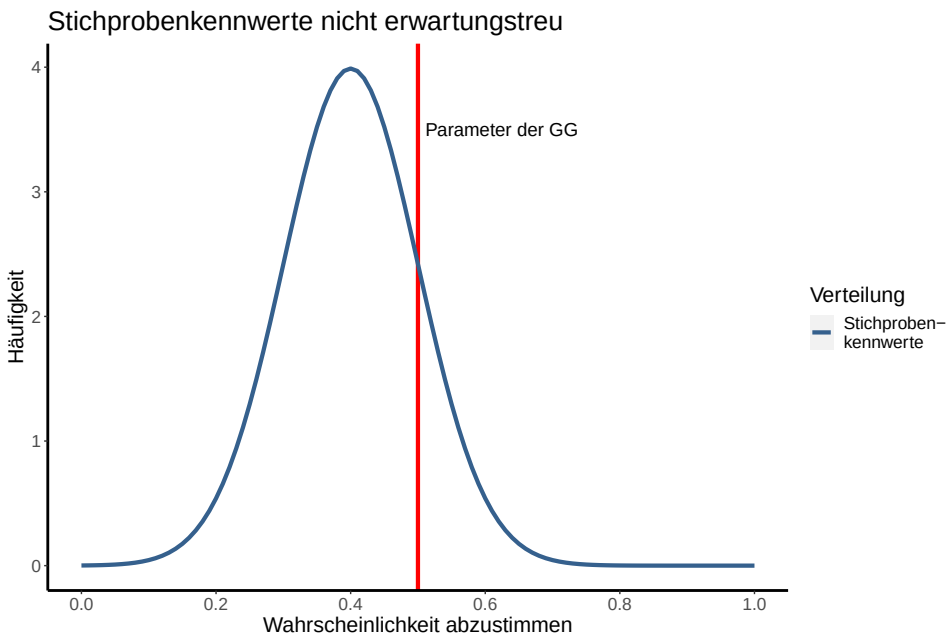


Abbildung 3.3: Nicht erwartungstreuer Kennwert eines Parameters

3.3 Variablenskalierung (V1.-V2.)

Die beiden ersten Voraussetzungen (V1. und V2.) betreffen die Skalierung der Variablen.

3.3.0.1 Variablen dürfen keine Konstanten sein (V1.)

Die UVs und die AV dürfen keine Konstante sein. Das ist insofern recht trivial, als dass eine Konstante mit nichts kovariieren kann, weil Konstanten nicht variieren. Je grösser «π, desto ...» macht einfach keinen Sinn. Da Konstanten nicht variieren (keine Varianz haben), können sie nicht kovariieren und können daher in keinen Erklärungsmodellen als Variablen einbezogen werden. An dieser Stelle klingt das sehr trivial. Und doch kommt es immer wieder vor, dass in Hypothesen Variablen einfliessen, die in der gewählten Stichprobe konstant sind. Zum Beispiel ist in der Hypothese «Wenn über Sport berichtet wird, zählen Superlative besonders.» Das Konstrukt «über Sport berichtet» ist eine Konstante, wenn nur der Sportteil untersucht werden soll. Hypothesen sind keine Annahmen über Zusammenhänge mehr, wenn eines der Konstrukte, die in Hypothesen zusammengebracht werden, in den Daten eine Konstante ist. Oftmals kommen solche Hypothesen mit Konstanten zustande, wenn der Fokus auf eine Ausprägung einer Variablen gelegt wird und die Abweichung von dieser Ausprägung nicht erhoben wird. Annahmen über den Wandel von Kriegsberichterstattung kann als zeitlicher Prozess nicht untersucht werden, wenn nur das Heute untersucht wird. Oft genug kommen Konstanten in Hypothesen vor, wenn das Forschungsinteresse aus dem Interesse der Forschenden eigentlich deskriptiv ist, also nur die Verteilung von einzelnen Variablen gefragt ist, und dann posthoc Hypothesen formuliert werden sollen, weil das von den Dozierenden oder Reviewern verlangt bzw. erwartet wird. ;-)

3.3.0.2 Variablen sollen metrisch sein (V2.)

Die AV und die UVs sollen metrisch sein. Das klingt nach einer recht harten Voraussetzung. Allerdings gibt es die schöne Eigenschaft von Dummyvariablen (0/1), dass sie sich verhalten wie metrische Variablen, weil ihr Mittelwert und ihre Streuung sinnvoll interpretierbar sind. Dummyvariablen können also gut als UVs

eingesetzt werden. Nun ist diese spezielle Form der dichotomen Variable (zwei Ausprägungen) nur die eine Form der nominalen Variablen. Dichotome Variablen können immer als Dummyvariable dargestellt werden. Man muss ja nur eine Ausprägung in 0 umkodieren und die andere in 1. Bei den kategorialen Variablen gibt es mehr Ausprägungen. Zum Beispiel Gender mit 1 = weiblich, 2 = männlich, 3 = divers[1]. Das Gute wiederum ist, dass kategoriale Variablen vollständig mit Dummyvariablen abgebildet werden können. Das geht dann so: Man baut eine Variable «Weiblich», die die Ausprägungen 1 = «trifft zu» und 0 = «trifft nicht zu» hat. Dann gibt es eine zweite Variable für «männlich» mit 0 und 1 und auch eine Dummy für «Divers». Diesem Vorgehen sind eigentlich keine Grenzen gesetzt. Man könnte also auch noch erweitern oder differenzieren in «transgender», «genderqueer», «genderfluid», «bigender», «pangender», «trigender», «agender», «demigender», «abinär» und zur Sicherheit in Deutschland auch «Taucher»[2].

In den linearen Modellen können Sie also auch kategoriale Variablen einbauen[3]. Auch die AV kann eine Dummyvariable sein. Das führt allerdings zu ein paar Problemen mit dem einfachen linearen Modell. Deshalb werden bei einer AV mit nur den Ausprägungen 0 und 1 logistische Regressionen gerechnet. Damit befassen wir uns später. Es geht auch, dass die AV kategorial ist. Das ist dann so ähnlich wie mit den Dummys als UV, weil dann mehrere Regressionen mit mehreren Dummys für die AV gerechnet werden. Das wird multinominale Regression genannt (auch bekannt als Diskriminanzanalyse).

Dann bleiben im Grunde nur die ordinalen Variablen übrig, die mehr Informationen über Ordnung der Ausprägungen (Rangordnung) enthalten, aber die Zahlenwerte (numerisches Relativ) mit ihren identischen Abständen (1 zu 2 wie 2 zu 3 und 3 zu 4 usw.) nicht abbilden, dass die Abstände der gemessenen Ausrägungen (empirisches Relativ) nicht annähernd gleich sind (1 = «arm», zwei = «reich», 3 gleich «superreich»). Dafür gibt es drei Lösungen, um ordinale Variablen auch in lineare Modelle einbeziehen zu können.

1. Ordinale Variablen werden als metrisch oder quasimetrisch behandelt und wie metrische in ein Modell aufgenommen. Das geschieht praktisch häufig, wenn z.B. Schulnoten einfach in ein lineares Modell aufgenommen werden. Wir wissen, dass die Abstände zwischen der Schweizer Bestnote 6.0 und 5.5 nicht genauso gross sind, wie zwischen 5.5 und 5.0 oder gar 4.0 und 3.5. Dennoch sind die Schätzer der linearen Modelle relativ robust gegen diese Verletzung. Gerade wenn es eigentlich nur darum geht, zu prüfen, ob Schulnoten einen signifikanten Effekt auf eine AV haben, dann kann man diese ordinalen Variablen getrost als «quasimetrisch» verwenden. In diesen Fällen sollte man nur etwas vorsichtiger sein, wenn eine Signifikanzschwelle nur knapp gerissen wurde oder b als Effekt nur knapp die Schwelle der Interpretierbarkeit übersprungen hat, dann sollte man bescheiden sein und klar machen, dass aufgrund der Datenlage und dem Skalennivau der Variablen die Zahlen nicht überinterpretiert werden sollten.

2. Es gibt auch die Möglichkeit, ordinale Variablen als kategoriale Variablen zu behandeln (womit ihr Datenniveau aber eigentlich herabgestuft wird). Dann würden wir die Ausprägungen der ordinalen UVs wiederum in Dummyvariablen umkodieren und nur die Dummys interpretieren. Im besten Fall werden in solche Interpretationen die zugrundeliegende Rangfolge der Dummys berücksichtigt, also die erste Gruppe mit der zweiten, die zweite mit der Dritten und dann die erste mit der Dritten, aber mit Rücksicht auf die Bedeutung der Rangfolge.

3. Wenn eine oder mehrere UVs klar ordinal sind, also die Abstände zwischen den Zahlenwerte deutlich auseinandergehen oder vielleicht sogar variieren (Laufwettkampf mit mal sehr knappen Unterschieden und mal sehr grossen von Platz eins zu Platz zwei, wenn Kipchoge mitläuft), dann sollten die ordinalen nicht einfach als metrische betrachtet werden. Wenn solche ordinalen Variablen zentral sind, dann kann auch nicht einfach auf Dummys ausgewichen werden. Dafür gibt es aber inzwischen Analysemethoden der ordinalen Regression, die in diesen Fällen eingesetzt werden können. Mit dem Verständnis der normalen linearen Modelle ist es nicht mehr schwer, sich so gut selbständig in die ordinale Regression einzuarbeiten, dass sie gewinnbringend eingesetzt werden kann.

[1] In einem offiziellen Anmeldeformular, das in Deutschland für Impfungen aufgeschaltet war, stand als dritte Option «Taucher», was der Autor für eine nicht sehr gelungene Übersetzung des Wortes «divers» hält.

[2] Noch besser ist es, wenn die Geschlechterfrage in Fragebögen halboffen gestaltet ist und die offenen Antworten in Dummys kodiert werden.

[3] Wenn nur kategoriale Variablen in der oder den UVs stecken, haben wir das, was mal Varianzanalyse genannt wurde.

3.4 Modellspezifikation und Multikollinearität (V3.-V5.)

IYI (Klausur): Standardfehler steigen bei Multikollinearität

Die Fehlervarianz s^2 ist definiert als Abstand zwischen den Fehlern e_i und dem Durchschnitt der Fehler, wobei wir davon ausgehen, dass wir die Fehler um 0 streuen, also die Regressionsgerade erwartungstreu schätzt und keinen Bias hat:

$$s^2 = \frac{1}{n-3}\sum (e_i - \bar{e})^2 = \frac{1}{n-3}\sum e_i^2. \tag{3.34}$$

Es wird also vorausgesetzt, dass $\bar{e} = 0$. Interessant ist noch der Zähler vor der Summe mit $n-3$ der zeigt, dass drei Kennwerte in die Berechnung für das Regressionsmodell eingegangen sind, die alle für sich auch eine Fehlerstreuung haben, weshalb wir nicht durch n teilen, sondern bei n-3 Freiheitsgraden eben durch n-3.
Die Fehlervarianz (Wir machen das mit der Varianz, weil lauter Wurzeln die Übersichtlichkeit nicht gerade steigern würden.) der Regressionskoeffizienten, also der b's ist durch folgende Formel definiert:

$$s_{b_2}^2 = \frac{s^2}{n}\frac{1/V_2}{1-r_{23}^2} \tag{3.35}$$

Also ist die Fehlervarianz von b_2 wie beim Standardfehler $s_{\bar{x}}$ gleich der der Varianz der Fehler s^2 geteilt durch n. Die Streuung der Fehler können wir durch gute Modellbildung verringern. Die Fallzahl n können wir durch die Vergrösserung der Stichprobe erhöhen (das kostet einfach Ressourcen aka Geld). Wir können also die Sicherheit bzw. Unsicherheit unserer Messungen durch die Vergrösserung der Stichproben verkleinern.
In der (3.35) steht aber mehr. Es volgt noch $\frac{1}{V_2}$. Das bedeutet, dass die Streuung der Regressionskoeffizienten auch etwas mit V_2 der Varianz von X_2 zu tun hat. Je mehr Varianz die Variable X_2 hat, desto kleiner ist die Streuung von b_2. Anders ausgedrückt: Je mehr Varianz beziehungsweise Unterschiede in der UV vorhanden ist, mit der Unterschiede in der AV aufgeklärt werden können, desto besser können wir den Zusammenhang beziehungsweise die grösse von b_2 schätzen. Wenn alle unsere Befragten fast dieselbe Antwort gegeben hätten, woran soll man dann festmachen, wann sich Y aufgrund von X_2 wie ändert? Wenn die UV breit streut, also von allem was drin ist, dann kann man auch schauen, was wie mit der AV, also Y einhergeht. Darum ist die Varianz von V_2 umgekehrt proportional zur Fehlerstreuung $s_{b_2}^2$.
Am Ende steht noch ein Faktor: $\frac{1}{1-r_{23}^2}$. In diesem Teil steckt die Musik für das aktuelle Kapitel: Hier wird die Multikollinearität abgebildet. Wenn nähmlich die Quadrierte Korrelation zwischen X_2 und Y gross ist, dann wird $1-r_{23}^2$ klein und damit geht die Fehlervarianz $s_{b_2}^2$ hoch. Wenn wir uns mal nur diesen «Multikollinearitätsfaktor» anschauen, dann wird deutlich, dass, wenn die Varianz der Variablen X_2 von der Variablen X_3 sagen wir zu 50% aufgeklärt wird, dann steht unter dem Bruch 0,5 und damit der Faktor 2 für die Varianz der Fehler von b_2. Darum wird dieser Faktor auch «Varianz-Inflations-Faktor» genann, oder kurz: VIF. Das Prinzip ist ähnlich wie bei dem Faktor $\frac{1}{V_2}$, für den wir oben bemerkt haben, dass wenig Erklärungsvarianz zu wackeligen b's führt, also solchen mit hoher Fehlerstreuung. In der multivariaten Regression rechnen wir aber immer die gemeinsame Varianz der UVs gegenseitig heraus. Wir nehmen also der Variablen X_2 viel Erklärungskraft, wenn wir mehrere weiter UVs in das Modell mit aufnehmen, die X_2 zu grossen Prozentanteilen Varianz klauen. Es kann also V_2 von X_2 in der Erhebung mehr oder weniger Varianz haben, oder es wird ihm durch andere UVs im Modell noch Varianz genommen. Der VIF ist in der (3.35) mit kleinem r geschrieben. Das liegt daran, dass es sich hier um das quadrierte bivariate r handelt, dass recht vertraut aussieht. Wenn wir mehr als zwei UVs haben, also neben X_2 und X_3 noch ein X_4, dann würden wir unter dem Bruch das multiple R^2 hinschreiben und im Subscipt an erster Stelle die Kennung der Variablen zu der das $s_{b_2}^2$ geschätzt werden soll, also X_2 und dann alle anderen UVs hinter einem Punkt. Wir

würden also schreiben: $R^2_2.34$, wenn wir noch die Varialben X_3 und X_4 im Modell berücksichtigen. Damit das nicht zu lang wird, deuten wir das nur an und schreiben bei noch mehr Variablen einfach $R^2_2.34....$
Das Gleiche gilt natürlich für die Fehlervarianz vn b_3, also $s^2_{b_3}$ spiegelbildlich:

$$s^2_{b_3} = \frac{s^2}{n} \frac{1/V_3}{1 - r^2_{23}} \tag{3.36}$$

Die Standardfehler von b_2 und b_3 sind einfach die Wurzeln, also $\sqrt{s_{b_2}}$ beziehungsweise $\sqrt{s_{b_3}}$.

3.5 V3. Fixe X

Dass die UVs fix sein sollen, bedeutet im Grunde nur, dass sich die UVs nicht ständig ändern sollen, sondern in unserer GG (beziehungsweise Auswahlgesamtheit) stabil sind. Wenn sich zum Beispiel die Berichterstattung insgesamt häufig stark ändert, dann wäre es nicht gut, wenn wir mit der Stichprobe einer Inhaltsanalyse arbeiten, die in einer sehr speziellen Zeit erhoben wurde (z.B. ein Kriegsanfang). Diese Stichprobe in einer «Spezialzeit» würde zu verzerrt geschätzten b's in der Normalzeit führen [vgl. @Wolling2015]. Da wir nicht davon ausgehen können und wollen, dass unsere Theorien in der Sozialwissenschaft immer und ewig gelten, verlangen wir nur mittelfristig gültige Theorien («middle range theory» [@Merton2012]) und dass unsere Variablen mittelfristig relativ stabil bzw. fix sind. Das bedeutet insbesondere, dass wir bei der Stichprobenziehung aufpassen müssen, dass wir nicht eine sehr spezielle Stichprobe in einer ganz besonderen Phase erheben, die Effekte hat, die sonst sehr untypisch sind. Das ist das, was mit fixe X gemeint ist.

IYI: Fixe X in Formeln abgeleitet

Im Folgenden soll abgeleitet werden, warum die fixierten X für die Schätzung von erwartungstreuen (also unverzerrten) b's wichtig ist. gleich in der ersten Formel führen wir eine neue vereinfachte Nomenklatur ein. Es soll ab hier V für die Varianz stehen und C für die Covarianz, während die Subscripte kennzeichnen, von welcher Variablen die Varianz gemeint ist und welche zwei Variablen die Covarianz aufweisen. Also ist zB $V_3C_{2y} das Produkt aus der Varianz von X_3 und der Covarianz von Y und X_2. unter dem Bruch steht ein «D» in dem wir die konstante Differenz von $V_2V_3-C_^2{23} erstmal verstecken, weil sie uns nicht sonderlich interessiert.

$$b_2 = \frac{V_3 C_{2Y} - C_{23} C_{3Y}}{D} \tag{3.37}$$

Also gut: Da wir wissen wollen, unter welchen Bedingungen unsere b's nicht erwartungstreu wären, nehmen wir die Gleichung für die Differenz der tatsächlichen Werte Y_i und dem tatsächlichen Mittelwert von Y, also $\overline{Y}$:

$$= \frac{(1/n) \sum_{i=1}^{n} \left\{ \left[V_3 \left(X_{i2} - \bar{X}_2 \right) - C_{23} \left(X_{i3} - \bar{X}_3 \right) \right] \left(Y_i - \bar{Y} \right) \right\}}{D} \tag{3.38}$$

Das ist viel, aber wenn man sich traut hinzuschauen, sieht man schnell, dass oben links vom -Zeichen einfach die Regressionsgleichen für 2 UVs steht, die wir schätzen wollen und rechts vom -Zeichen steht in Klammern dieselbe Formel, aber immer für die Mittelwerte der UVs ($\overline{X_2}$ und $\overline{X_3}$ sowie $\overline{U}$). In der zweiten Zeile werden die Mittelwerte ihren UVs zugeordnet.

$$Y_i - \bar{Y} = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i - \left(\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{U} \right) \tag{3.39}$$

$$= \beta_2 \left(X_{i2} - \bar{X}_2 \right) + \beta_3 \left(X_{i3} - \bar{X}_3 \right) + \left(U_i - \bar{U} \right) \tag{3.40}$$

Der Mittelwert des unbekannten Rests $\overline{U}$ ist der Mittelwert der Fehlerterme in der Stichprobe. Wenn wir das überall einsetzen, wird es noch unübersichtlicher, aber, wenn Sie das Muster erkenne, wird deutlich, dass statt der Variablen jetzt überall die Differenzen der Variablen und ihren Mittelwerten stehen:

$$b_2 = \frac{1}{ND}\left\{\sum_{i=1}^{n}\left[V_3\left(X_{i2}-\bar{X}_2\right)-C_{23}\left(X_{i3}-\bar{X}_3\right)\right]\right. \tag{3.41}$$

$$\left.\times\left[\beta_2\left(X_{i2}-\bar{X}_2\right)+\beta_3\left(X_{i3}-\bar{X}_3\right)+\left(U_i-\bar{U}\right)\right]\right\} \tag{3.42}$$

$$= \frac{1}{D}\left\{\frac{\beta_2}{N}V_3\sum\left(X_{i2}-\bar{X}_2\right)^2-\frac{\beta_2}{N}C_{23}\sum\left(X_{i3}-\bar{X}_3\right)\left(X_{i2}-\bar{X}_2\right)\right. \tag{3.43}$$

$$+\frac{\beta_3}{N}V_3\sum\left(X_{i2}-\bar{X}_2\right)\left(X_{i3}-\bar{X}_3\right)-\frac{\beta_3}{N}C_{23}\sum\left(X_{i3}-\bar{X}_3\right)^2 \tag{3.44}$$

$$\left.+\frac{1}{N}\sum\left[V_3\left(X_{i2}-\bar{X}_2\right)-C_{23}\left(X_{i3}-\bar{X}_3\right)\right]\left(U_i-\bar{U}\right)\right\} \tag{3.45}$$

Die erste elendige Summe kann geschrieben werden als: $\beta_2 V_3(1/N)\Sigma\left(X_{i2}-\bar{X}_2\right)^2$ was sich zu $\beta_2 V_3 V_2$ deuuutlich vereinfachen lässt.
Jetzt knöpfen wir uns zwei wesentliche Teile vor: Den Erwartungswert der Covarianz aus X_2 und dem unbekannten Rest U $(E(C_{2U}))$, sowie den Erwartungswert $E(C_{2U})$:

$$E\left(C_{2U}\right) = \frac{1}{N}\sum\left(X_{i2}-\bar{X}_2\right)E\left(U_i-\bar{U}\right) \tag{3.46}$$

$$= \frac{1}{N}\sum\left(X_{i2}-\bar{X}_2\right)\left(\bar{U}_i-\mu\right) \tag{3.47}$$

$$= C_{2\bar{U}_i} \tag{3.48}$$

Das bedeutet, dass nach der Ableitung rauskommt, dass der Erwartungswert der Covaranz zwischen X_2 und dem unbekannten Rest die Covarianz zwischen X_2 und dem Mittelwert vom unbekannten Rest U ist (das Gleiche kann man für $E(C_{3U})$ ableiten, aber das sparen wir uns [wer will, kann ja]). Wenn jetzt jedes U_i gleich um $\bar{U}$ streut, dann ist $E(U_i-\bar{U})=0$. Es dürfen also die UVs nicht mit dem unbekannten Rest korrelieren. Das tun sie nicht, wenn im Rest nur Rauschen ist und keine Erklärungsvariablen, die wir nicht im Modell haben.
Jetzt schauen wir uns die nächsten Elemente an. Das ist zum Einen die Konstante b_1:

$$b_1 = \bar{Y}-b_2\bar{X}_2-b_3\bar{X}_3 = \beta_1+\beta_2\bar{X}_2+\beta_3\bar{X}_3+\bar{U}=b_2\bar{X}_2-b_3\bar{X}_3 \tag{3.49}$$

$$= \beta_1+\left(\beta_2-b_2\right)\bar{X}_2+\left(\beta_3-b_3\right)\bar{X}_3+\bar{U} \tag{3.50}$$

Wenn wir davon wieder den Erwartungswert suchen, ergibt sich:

$$E(b_1) = E(\beta_1)+E(\beta_2-b_2)\bar{X}_2+E(\beta_3-b_3)\bar{X}_3+E(\bar{U})$$

Es ist dann b_1 ein erwartungstreuer Schätzer von β_1, wenn die U_i unabhängig sind von X_2 und X_3 und b_2 sowie b_3 unverzerrt sind, also gilt: $E(\beta_2-b_2)=0$ und $E(\beta_3-b_3)=0$. Und es müssen die Erwartungswerte der Unbekannten Null sein, also $E(\bar{U})=0$.

Wann sind also b_2 und b_3 erwartungstreu und unverzerrt? Wir nehem also erstmal wieder unsere mit βs gespickte Formel:

$$b_2 = \frac{\beta_2 V_3 V_2 - \beta_2 C_{23}^2 + \beta_3 V_3 C_{23} - \beta_3 C_{23} V_3}{D} + \frac{V_3 C_{2U} - C_{23} C_{3U}}{D} \tag{3.51}$$

$$= \frac{\beta_2 \left(V_2 V_3 - C_{23}^2\right)}{D} + \frac{V_3 C_{2U} - C_{23} C_{3U}}{D} \tag{3.52}$$

$$= \beta_2 + \frac{V_3 C_{2U} - C_{23} C_{3U}}{D}. \tag{3.53}$$

Wichtig ist wieder die letzte Zeile der Ableitung. Hier zeigt sich, dass b_2 gleich β_2 ist, wenn der zweite Summand rechts gleich Null ist. Das ist er wieder, wenn es keine Covarianz zwischen X_2 und U gibt.
Für b_3 gilt dasselbe:

$$b_3 = \beta_3 + \frac{V_2 C_{3U} - C_{23} C_{2U}}{D}.$$

Wenn wir jetzt wieder den Erwartungswert suchen, um zu sehen, wovon er abhängig ist, schreiben wir:

$$E\left(b_2\right) = E\left(\beta_2\right) + E\left[\frac{V_3 C_{2U} - C_{23} C_{3U}}{D}\right] \tag{3.54}$$

Da die wahren Regressionkoeffizienten β Konstanten sind, können wir sie einfach so hinschreiben, ohne das E() drumrum. Aus Den Ausdrücken für die Erwartungswerte können wir auch V_2 und V_3 rausholen, weil das auch Konstanten sind:

$$E\left(b_2\right) = \beta_2 + \frac{V_3 E\left(C_{2U}\right)}{D} - \frac{C_{23} E\left(C_{3U}\right)}{D}, \tag{3.55}$$

Und wieder gilt, dass es darauf ankommt, dass es keine Covarianz zwischen den Variablen im Modell gibt und dem unbekannten Rest U, damit die b's die βs erwartungstreu schätzen.
Wie oben im einfachen Teil schon gesagt, können wir nicht statistisch prüfen, ob unsere Annahmen stimmen. Wir müssen also kritisch und erfinderisch nach Forschungsmethoden suchen, um aus dem Unbekannten die Einflussgrössen zu holen, die eventuell noch mit unseren Konzepten in unseren theoretischen Modellen korrelieren!

3.5.0.1 V4. Voll spezifizierte Modelle

Unsere B's sind nur dann unverzerrt, wenn das Modell voll spezifiziert ist in Bezug auf Einflüsse, die mit unseren B's in Wirklichkeit zusammenhängen. Wenn wir vergessen in unsere Überlegungen und Messungen einzubeziehen, dass die Storchenpopulation einer Gegend nur darum mit der Geburtenrate zu tun hat, weil in ländlichen Regionen die Geburtenrate höher ist und mehr Störche leben als in der Stadt; wenn wir also diesen Dritteinfluss vergessen, dann scheint es einen Zusammenhang zwischen Geburtenrate und Storchenpopulation zu geben. Wir würden falsche Schlüsse ziehen, weil der Zusammenhang verzerrt geschätzt würde. Journalistinnen vom Berliner Kurier könnten glauben, dass der Storch die Kinder bringt. Wir müssen also theoretisch erarbeiten, welche Einflüsse von Bedeutung sein könnten für unsere AV oder den Zusammenhang zwischen den UVs und der AV beeinflussen könnten. Das ist Theoriearbeit. Dieser Zusammenhang muss sich auch mathematisch in der Statistik abbilden, was er auch tut.

Wenn wir mal annehmen, dass die wahren Zusammenhänge gut durch die Formel **??**eq-3.81) dargestellt wären, aber die Theorie zu dem Thema auf dem Stand ist, dass die einfacheren Zusammenhänge aus der Formel **??**eq-3.82) gelten, also eine wichtige Einflussgrösse (X_4) nicht berücksichtigt wurde. Wenn dem so wäre, dann würde das Unbekannte (U_i) in Formel **??**eq-3.82) nicht nur den einfachen stochastischen Rest umfassen, sondern zusätzlich $\beta_4 X_{i4}$. Dann wäre der Erwartungswert (also der Wert, um den unsere Stichprobenparameter b streuen) nicht mehr das erhoffte β_2 sondern $\beta_2 + \beta_4 b_{42}$, wie in Formel **??**eq-3.83).

Das würde zu einem Fehler führen, der bei $\frac{r_{42}-r_{32}r_{43}}{1-r_{32}^2}\sqrt{\frac{V_4}{V_2}}$ liegt. Wenn wir also ewig Stichproben ziehen würden und jedes Mal ein b_2 bestimmen würden, dann würden diese b_2s nicht um β_2 streuen. Das Mass, um das wir uns verschätzen würden, wäre so gross wie in (**??**)eq-Spez4) notiert. Auch unsere Signifikanztests wären falsch und die Konfidenzintervalle würden an der falschen Stelle liegen. Unsere ganze Analyse wäre falsch.

$$\text{wahr:} Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + U_i \tag{3.56}$$

$$\text{geschätzt: } Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i^\star \quad \text{wobei} \quad U_i^\star = \beta_4 X_{i4} + U_i \tag{3.57}$$

$$\text{also: } E(b_2) = \beta_2 + \beta_4 b_{42} \tag{3.58}$$

$$\text{mit: } b_{42} = \frac{r_{42} - r_{32}r_{43}}{1 - r_{32}^2}\sqrt{\frac{V_4}{V_2}} \tag{3.59}$$

Wie geht man nun mit dieser Tyrannei um, dass man alle Einflüsse kennen sollte, die schlicht unbekannt sind. Nur Chuck Norris weiss, wann ein Modell voll spezifiziert ist. Wir können nie wissen, wann wir am Ende der Wissenschaft angekommen sind, weil wir alles vollständig und für immer gültig spezifiziert haben. Es geht bei dieser Überlegung der Spezifikation mehr darum, dass wir die Spezifikation der bestehenden Modelle verbessern. Das kann heissen, dass wir falsche Alltagsvorstellungen korrigieren, indem wir den Kindern irgendwann sagen, dass das bivariate Regressionsmodell mit den Störchen und den Kindern, nicht voll spezifiziert ist und Sex, Verhütung und viele mehr einen gewissen Einfluss hat auf die Geburtenrate. Wir klären aber nicht nur in der Alltagswelt auf, sondern verbessern auch unsere Modelle stetig, indem wir uns fragen, welche Einflussgrössen bei der Erklärung eines Phänomens noch eine Rolle spielen könnten.

Die statistisch, mathematische Anforderung an die Modellspezifikation bedeutet also, dass wir unsere Theorie gut und gründlich entwickeln müssen. Bei einer schlechten Theorie und entsprechend zu wenig erfasster oder einbezogener Modells sind unsere Ergebnisse verzerrt und damit falsch oder mindestens nicht state of the art. Darum muss man immer erst schauen, was der Forschungsstand ist. Der kann repliziert und damit kontrolliert werden, und wenn wir das Modell weiter spezifizieren und neue Ergebnisse erlangen, dann haben wir die Theorie erweitert und einen wissenschaftlichen Mehrwert geschaffen. Es werden auch noch Generationen nach uns und Ihnen kommen, die unsere Theorien überarbeiten und dabei feststellen, dass wir unserer Modelle unterspezifiziert hatten. Das ist dann der wissenschaftliche und zivilisatorische Fortschritt. Wissenschaft wird also nicht irgendwann fertig sein und wichtig bleiben.

IYI (Klausur): Ableitung Modellspezifikation

Das Basismodell für die wahre zu schätzende Realität sei wieder: $Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + U_i$. Das heisst, es kommt ein Dritteinfluss hinzu, der als $\beta_4 X_{i4}$ Teil des ganzen Kausalzusammenhangs ist. Im Folgenden untersuchen wir, was passiert, wenn eben dieses $\beta_4 X_{i4}$ nicht im Modell ist, dieses also unterspezifiziert ist. Also:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + U_i,$$

Aber statt dieses vollständige Modell zu schätzen, nehmen wir mal ein unterspezifziertes Modell, wobei im U_i^* das $\beta_4 X_{i4}$ steckt, also:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i^* \quad \text{wobei} \quad U_i^* = \beta_4 X_{i4} + U_i$$

Auch das schauen wir uns genauer für b_2 an und tauschen alles entsprechend aus, wenn wir das auch noch für b_3 machen wollten:

$$b_2 = \frac{V_3 C_{2Y} - C_{23} C_{3Y}}{V_2 V_3 - C_{23}^2} = \beta_2 + \frac{V_3 C_{2U^*} - C_{23} C_{3U^*}}{V_2 V_3 - C_{23}^2}, \tag{3.60}$$

Wenn man jetzt $U_i^* = \beta_4 X_{i4} + U_i$ an den Stellen einsetzt, wo U^* stand, wird es wieder voll, vereinfacht sich aber auch gleich wieder:

$$C_{2U^*} = \frac{1}{n}\sum\left(X_{i2} - \bar{X}_2\right)\left(U_i^* - \bar{U}^*\right) \tag{3.61}$$
$$= \frac{1}{n}\sum\left(X_{i2} - \bar{X}_2\right)\left(\beta_4 X_{i4} + U_i - \beta_4\bar{X}_4 - \bar{U}\right) \tag{3.62}$$
$$= \frac{1}{n}\beta_4\sum\left(X_{i2} - \bar{X}_2\right)\left(X_{i4} - \bar{X}_4\right) + \frac{1}{n}\sum\left(X_{i2} - \bar{X}_2\right)\left(U_i - \bar{U}\right) \tag{3.63}$$
$$= \beta_4 C_{24} + C_{2U} \tag{3.64}$$

Nehmen wir jetzt wieder die Erwartungswerte von b_2 und b_3, mit der Grundannahme, dass die X fix sind und $E\left(U_i\right) = 0$, dann geht es gut weiter mit:

$$E\left(b_2\right) = \beta_2 + \beta_4\left(\frac{V_3 C_{24} - C_{23}C_{34}}{V_2 V_3 - C_{23}^2}\right) + E\left[\frac{V_3 C_{2U} - C_{23}C_{3U}}{V_2 V_3 - C_{23}^2}\right] = \beta_2 + \beta_4 b_{42} \tag{3.65}$$

Und genau das bringt uns zu der Formel, die beschreibt, um wie viel wir das b_2 verzerrt schätzen, wenn wir die wichtige Einflussgrösse X_4 nicht mit im Modell haben (für b_3 gilt wieder das Gleiche mit ein bisschen ausgetauschten Subscripten.):

$$b_{42} = \frac{\left(r_{42} - r_{32}r_{43}\right)}{1 - r_{32}^2}\sqrt{\frac{V_4}{V_2}}$$

3.5.0.2 Keine perfekte oder heftige Multikollinearität (V5.)

Wenn perfekte Multikollinearität vorliegt, dann kann eine Variable perfekt aus den übrigen Variablen vorhergesagt werden (technischer: eine UV ist eine Linearkombination der übrigen UVs). Ein lineares Modell gibt dann keine Antwort auf die ihm gestellte Frage, wenn zwei UVs identisch sind, also untrennbar verwoben. Das liegt daran, dass die Frage an das lineare Modell ist: «Wie starkt ist der Effekt jeder einzelnen UV, wenn die Effekte der übrigen UV herausgerechnet werden?». Wenn eine Variable eine Linearkombination der übrigen Variablen ist, dann bleibt von ihr exakt nichts übrig, wenn die Linearkombination der übrigen Variablen aus ihr herausgerechnet werden. Ist ihre Varianz dadurch 0, ist sie im Grunde eine Konstante, und wie in V1. diskutiert, kann mit Konstanten keine Kovarianz und damit auch kein lineares Modell gerechnet werden. Jedes Statistikprogramm würde also an dieser Stelle aussteigen und ihnen sagen, dass das Modell so nicht gerechnet werden kann, weil perfekte Multikollinearität vorliegt. Das muss also nicht extra getestet werden.

Perfekte Multikollinearität entsteht meistens, wenn eine Variable aus dem Rohdatensatz umkodiert wurde und die Originalvariable und die einfach umkodierte mit im Modell sind. Die schuldige Variable findet man recht schnell. Etwas weniger direkt ersichtlich ist so eine perfekte Multikollinearität durch Datenaufbereitung, wenn ein Index und alle Variablen, aus denen der Index berechnet wurden, mit in das Modell aufgenommen wurden. Wenn Sie also z.B. die Durchschnittsnote im Abi in das Modell packen und alle Noten der einzelnen Fächer auch, die zusammen exakt die Durchschnittsnote ergeben. Suchen Sie in solchen Fällen nach den Indizes. Wenn Sie in dem Beispiel die Durchschnittsnote rausnehmen oder ein paar Fächer, die ihnen für die Erklärung der AV nicht so wichtig erscheinen, dann wird das Problem der perfekten Multikollinearität schnell gelöst sein.

Etwas Multikollinearität ist allerdings nicht nur erlaubt, sondern der Grund dafür, dass wir multivariate Modelle rechnen. Wären die UVs untereinander alle unkorreliert, dann wären alle B's dieselben, wenn nur bivariate Regressionen gerechnet werden würden. In der Formel (**??**)eq-Bs1) für b_2 sieht man das auch sehr gut: Wenn $r_{23} = 0$, also keine Multikollinearität beim Modell mit zwei UVs (X_2 und X_3), dann kommt für b_2 dasselbe raus, wie ohne X_3 (in (**??**)eq-Bs1) wird $r_{23} = 0$ gesetzt und in (**??**)eq-Bs3) sieht man, dass X_3 oder r_3 keine Rolle spielen).

$$b_2 = \frac{r_{Y2} - r_{23} r_{Y3}}{(1 - r_{23}^2)} \frac{S_y}{S_2} \tag{3.66}$$

$$b_2 = \frac{r_{Y2} - 0 \cdot r_{Y3}}{(1 - 0^2)} \frac{S_y}{S_2} \tag{3.67}$$

$$b_2 = r_{Y2} \frac{S_y}{S_2} \tag{3.68}$$

Wenn es etwas Multikollinearität gibt, wird das Produkt aus $r_{23} r_{Y3}$ aus dem bivariaten b_2 subtrahiert (herausgerechnet). Zusätzlich wird mit einer Korrektur unter dem Bruchstrich von $1 - r_{23}^2$ angepasst. In Worten bedeutet das so viel wie: Wenn wir untersuchen wollen, ob der Storch (UV) die Kinder bringt (AV), aber wissen, dass das auch noch mit Urbanität (X_3) zusammenhängt, dann müssen wir berücksichtigen (herausrechnen) wie stark Urbanität (X_3) und Storchenpopulation (X_2) zusammenhängen (r_23), wenn bzw. in dem Masse, wie auch die Geburtenrate (Y) mit der Urbanität zusammenhängt (r_{Y2}). Das steht über dem Bruch der Formel (**??**)eq-Bs1). Da wir nicht mehr mit den vollen 100% der Varianz von X_2 rechnen können, wird unter dem Bruchstrich der Formel (**??**)eq-Bs1) auch noch herausgerechnet, um wie viel X_2 durch X_3 beklaut wird ($1 - r_{23}^2$). Über diesen Teil der Formel lohnt es sich, etwas länger nachzudenken.

Toleranz und VIF

Wenn Multikollinearität bedeutet, dass eine Variable durch eine andere stark bestimmt wird, haben wir für die Bestimmtheit einer Variablen durch andere ein Mass: Das Bestimmtheitsmass R^2. In der Formel (**??**)eq-Bs1) steht unter dem Bruch ein r_{23}^2, das man besser auch schreiben könnte als $R_{2.3}^2$, einfach um deutlicher zu machen, dass es um eine multiple Korrelation geht und darum, dass die Regression auf X_2 gemeint ist, von allen übrigen Variablen. Wenn es mehr als nur die X_3 gibt, würden wir in der Formel für b_2 schreiben $R_{2.34567...}^2$ und bei b_3 $R_{3.24567...}^2$. Nun ist Multikollinearität nichts Gutes, sondern ein Problem. Darum steht in Formel (**??**)eq-Bs1) auch $1 - r_{23}^2$. Hier ist also angegeben, wie viel von den 100% Varianz von b_2 übrig bleiben, wenn man herausgerechnet hat, wie stark die übrigen UVs die Variable X_2 bestimmen ($R_{2.34567...}^2$). Man könnte auch sagen, dass damit für die Multikollinearität angegeben ist, wie stark ihre Toleranz gegenüber den übrigen Variablen ist. Wenn also zum Beispiel die übrigen Variablen 40% der Variable X_2 erklären, dann wäre die Toleranz $1 - 0.4$, also 60%. Diesen Toleranzwert (TOL) sollte man sich bei jeder Regression mit rausgeben lassen, um zu prüfen, wie stark die einzelnen Variablen von Multikollinearität betroffen sind. In Publikationen sieht man diese Werte oft nicht, weil sie von den Forschenden geprüft und für nicht problematisch befunden wurden (wenn diese Forschenden gründlich arbeiten).

Multikollinearität hat vor allem auch eine Bedeutung für die Fehlervarianz der B's, also wie unsicher oder wackelig die b's sind. Darum steckt in der Formel für die $s_{b_2}^2$ auch das $1 - R_{23}^2$ unter dem Bruchstrich des Faktors drin, der hinten steht. Dieser hintere Faktor ist demnach der Faktor, um den die Fehlervarianz der B's steigt, wenn die Toleranz ($1 - R_{2.3}^2$) klein ist, weil die jeweilige UV stark durch die übrigen Variablen bestimmt wird ($R_{2.3}^2$). Mit diesem Faktor wird auch gearbeitet, indem in Regressionsanalysen in Outputs häufig der **V**arianz-**I**nflations-**F**aktor (VIF) mit angezeigt wird. Wenn also zum Beispiel die Varianz der Variablen X_2 zu 90% durch die übrigen Variablen im Modell aufgeklärt wird, dann ist die Wert TOL nur noch $1 - .9 = .1$. Der Variablen X_2 würden also nur noch 10% seiner Ursprungsvarianz bleiben, um die AV erklären zu können. Das ist nicht viel, worauf eine stabile Regressionsgerade angepasst werden könnte. Darum wackelt das b_2 viel mehr, als wenn die anderen Variablen nicht berücksichtigt worden wären. Die Unsicherheit wurde um den Faktor $\frac{1}{1 - R_{2.34567...}^2}$ inflationiert, also um das Zehnfache! Da muss man sich dann schon fragen, was da eigentlich übrig bleibt.

$$s_{b_2}^2 = \frac{s^2}{n} \cdot \frac{1}{V_2} \cdot \frac{1}{1 - R_{2.3}^2} \tag{3.69}$$

$$s_{b_3}^2 = \frac{s^2}{n} \cdot \frac{1}{V_3} \cdot \frac{1}{1 - R_{3.2}^2} \tag{3.70}$$

3.6 Homoskedastizität (V6.)

IYI (Klausur): Effizienz als Varianz von b_2 und b_3

Gehen wir davon aus, dass b_2 unverzerrt ist, entspricht ist die Fehlervarianz durch die Streuung der b_2 um den wahren Wert β_2 definiert:

$$b_2 - \beta_2 = \frac{V_3 C_{2U} - C_{23} C_{3U}}{D} \tag{3.71}$$

$$\begin{aligned}
\operatorname{var}(b_2) &= E\left[b_2 - E(b_2)\right]^2 && (3.72)\\
&= E\left(b_2 - \beta_2\right)^2 && (3.73)\\
&= E\left(\frac{V_3 C_{2U} - C_{23} C_{3U}}{D}\right)^2 && (3.74)\\
&= \frac{1}{D^2} E\left(V_3^2 C_{2U}^2 - 2 V_3 C_{23} C_{2U} C_{3U} + C_{23}^2 C_{3U}^2\right) && (3.75)\\
&= \frac{1}{D^2}\left[V_3^2 E\left(C_{2U}^2\right) - 2 V_3 C_{23} E\left(C_{2U} C_{3U}\right) + C_{23}^2 E\left(C_{3U}^2\right)\right] && (3.76)
\end{aligned}$$

{#eq-3.5}
Um das Problem in Teile zu zerlegen, die uns Auskunft über die Streuung der b's geben, ziehen wir nacheinander die Teile der Formel heraus, die mit Erwartungen $E()$ versehen sind, also $E(C_{2U}^2)$ (für $E(C_{3U}^2)$ spiegelbildlich) und $E(C_{2U}C_{3U})$. Es geht also wieder darum, wie die Covarianz der UVs mit dem unbekannten Rest aussieht und wie es um das Produkt dieser Covarianzen bestellt ist.
Auflösen von $E(C_{2U}^2)$
Für die Covarianz setzen wir erstmal die bekannte Formel ein. Da der Erwartungswert ja gedanklich die Wiederholung von Stichproben ist, ergeben sich lauter Covaranzen mit jeweils unterschiedlichen $U_i - \bar{U}$ in einer laaaangen Summe.

$$\begin{aligned}
E\left(C_{2U}^2\right) &= E\left\{\frac{1}{n^2}\left[\sum\left(X_{i2} - \bar{X}_2\right)\left(U_i - \bar{U}\right)\right]^2\right\} && (3.77)\\
&= \frac{1}{n^2} E\big[\left(X_{12} - \bar{X}_2\right)\left(U_1 - \bar{U}\right) + \left(X_{22} - \bar{X}_2\right)\left(U_2 - \bar{U}\right) && (3.78)\\
&\quad + \cdots + \left(X_{n2} - \bar{X}_2\right)\left(U_T - \bar{U}\right)\big]^2 && (3.79)
\end{aligned}$$

Um die ganze lange Summe herum steht ein Quadrat. Das lösen wir jetzt auf, indem wir es in die Klammer reinziehen und nach binomischer Formel umstellen:

$$\begin{aligned}
&= \frac{1}{n^2} E\left[\sum_{i=1}^{n}\left(X_{i2} - \bar{X}_2\right)^2\left(U_i - \bar{U}\right)^2 && (3.80)\right.\\
&\left. + 2\sum_{i=1}^{n-1}\sum_{s=i+1}^{n}\left(X_{i2} - \bar{X}_2\right)\left(X_{s2} - \bar{X}_2\right)\left(U_i - \bar{U}\right)\left(U_s - \bar{U}\right)\right] && (3.81)
\end{aligned}$$

Das Ziel der Übung ist, dass wir die Erwartungswerte definieren als Erwartungswert der Streuung von X_2 und einem Teil, wo der Erwartungswert nur für U's drinsteht. Dann können wir nämlich etwas über Bedingungen für die U's sagen. Also:

$$= \frac{1}{n^2}[\sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2 E(U_i-\bar{U})^2 \tag{3.82}$$

$$. + 2\sum_{i=1}^{n}\sum_{s=t+1}^{n}(X_{i2}-\bar{X}_2)(X_{s2}-\bar{X}_2)E(U_i-\bar{U})(U_s-\bar{U})] \tag{3.83}$$

Aus der Bedingung der fixierten X können wir für $E(U_i-\bar{U})(U_s-\bar{U})$ das σ_{is} einsetzen. Damit verschwinden also die U's aus der Gleichung, was ja unser Ziel war:
Das gilt also, wenn $\sigma_i^2 = E\left(U_i-\bar{U}\right)^2$ und $\sigma_{is} = E\left(U_i-\bar{U}\right)\left(U_s-\bar{U}\right)$. Wenn wir diese Umstellung für $E\left(C_{2U}C_{3U}\right)$ und $E\left(C_{3U}^2\right)$ ebenfalls durchführen kommen wir zu:

$$E\left(C_{3U}^2\right) = \frac{1}{n^2}\left[\sum\left(X_{i3}-\bar{X}_3\right)^2\sigma_i^2 + 2\sum\sum\left(X_{i3}-\bar{X}_3\right)\left(X_{s3}-\bar{X}_3\right)\sigma_{is}\right], \tag{3.84}$$

$$E\left(C_{2U}C_{3U}\right) = \frac{1}{n^2}\left[\sum\left(X_{i2}-\bar{X}_2\right)\left(X_{i3}-\bar{X}_3\right)\sigma_i^2\right. \tag{3.85}$$

$$\left. +2\sum\sum\left(X_{i2}-\bar{X}_2\right)\left(X_{s3}-\bar{X}_3\right)\sigma_{is}\right]. \tag{3.86}$$

$$\tag{3.87}$$

OK, das sind schon sehr komplizierte Formeln. Mit weiteren Annahmen über die Fehlerverteilung können wir das aber weiter vereinfachen. Damit erhalten wir Auskünfte über weitere Eigenschaften der OLS-Schätzer im Vergleich zu anderen Schätzern. Wenn die Fehlerterme homoskedastisch sind, also überall gleich (und überall denselben Mittelwert haben), dann ergibt sich folgende Vorraussetzung, die wir «Homoskedastizität» nennen:

$$E\left(U_i-\bar{U}\right)^2 = \sigma_i^2 = \sigma^2 \quad \text{für alle } i.$$

Heisst also, dass die Streuung der b's für alle Fälle gleich sein soll und nicht in Abhängigkeit der X_i mal schmaler und mal breiter um den wahren Wert β streuen.
Wenn nun alle Fälle unabhängig voneinander, also in einer ordentlichen Zufallsstichprobe gezogen wurden, dann sind sie unabhängig voneinander und korrelieren nicht miteinander. Das bedeutet, die Covarianz im unbekannten Rest ist 0, für unterschiedliche Fälle:

$$E\left(U_i-\bar{U}\right)\left(U_s-\bar{U}\right) = \sigma_{is} = 0 \quad \text{für} \quad i \neq s.$$

Wenn wir diese beiden Annahmen haben, dass die Fehler überall gleich sind (Homoskedastizität) und die Fehler unkorreliert sind, dann vereinfachen sich die Terme der Erwartungswerte von oben zu:

$$E\left(C_{2U}^2\right) = \frac{1}{n^2}\sum\left(X_i-X_2\right)^2\sigma^2 = \frac{\sigma^2 V_2}{n}, \tag{3.88}$$

$$E\left(C_{3U}^2\right) = \frac{\sigma^2 V_3}{n}, \tag{3.89}$$

Wenn wir die jetzt wieder in die **?@eq-3.5** einsetzen, dann wird es wieder komplizierter, aber bei Weitem nicht so kompliziert wie oben:

$$\mathrm{var}\left(b_2\right) = \frac{\sigma^2}{nD^2}\left[V_3^2V_2 - 2V_3C_{23}^2 + C_{23}^2V_3\right] = \frac{\sigma^2V_3}{nD^2}\left[V_3V_2 - C_{23}^2\right] \tag{3.90}$$

$$= \frac{\sigma^2V_3}{nD} = \frac{1}{n}\sigma^2\left[\frac{V_3}{V_2V_3-C_{23}^2}\right] = \frac{\sigma^2}{n}\left[\frac{1/V_2}{1-r_{23}^2}\right]. \tag{3.91}$$

Für b_3 wieder nach demselben Prinzip:

$$\operatorname{var}(b_3) = \frac{1}{n}\sigma^2 \left[\frac{V_2}{V_2V_3 - C_{23}^2}\right] = \frac{\sigma^2}{n}\left[\frac{1/V_3}{1-r_{23}^2}\right].$$

Nur zum Spass: Was Sie jetzt probieren könnten: Zeigen Sie, dass im bivariaten Fall (also nur X_2 als UV) Folgendes gilt, indem Sie alles rauskürzen, was die Beziehung der UVs untereinander wiedergibt:
$\operatorname{var}(b_2) = \sigma^2/\Sigma\left(X_i - \bar{X}\right)^2 = \sigma^2/n\operatorname{var}(X)$.

Homoskedastizität bedeutet, dass die Streuung der Fehler um die Regressionsgerade überall ungefähr gleich (homo) gross sein sollte. Heteroskedastizität bedeutet, dass die Fehlerstreuung um unsere Regressionsgerade mit der grösse unserer UVs unterschiedlich ist, also z.B. grösser wird, weil Kodierer:innen wenn sie sehr lange nacheinander (weil vielleicht in letzter Minute) kodieren, mit der Zeit immer mehr Fehler machen. Oder weil Kodierer:innen regelmässig ein bisschen kodieren und dabei immer besser werden und immer weniger Fehlerstreuung entsteht. Wenn diese Streuung um die Regressionsgerade mit einer Variablen korreliert wie in Abb. @ref(fig:Heteroskedastizitaet), dann sind die Standardfehler der b's nicht gut und gültig geschätzt. Mithin sind die t-Werte nicht korrekt, damit die p-Werte und Konfidenzintervalle falsch und schliesslich unsere Entscheidung über die Gültigkeit oder auch die Entscheidbarkeit der Hypothese (H0 oder H1) falsch.

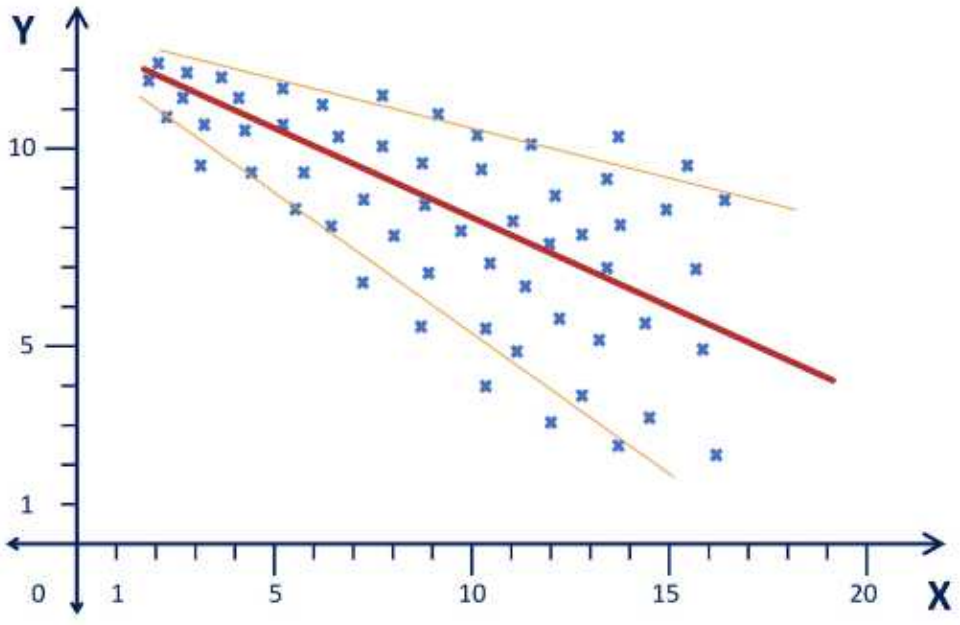


Abbildung 3.4: Heteroskedastizität

Neben diesem breiter oder schmaler werden der Streuung um die Regressionsgerade entsteht Heteros-

kedastizität oftmals, wenn wir eine Gerade in einen kurvlinearen Zusammenhang einpassen. In der Abb. @ref(fig:Hetero-Nicht-Linearitaet) ist gut zu erkennen, dass in (a) die Verteilung der standardisierten Fehler recht gleichmässig ist. In (b) geht eben die Schultüte (bzw. Tüte Marroni) auseinander und stellt damit Heteroskedastizität dar. In (c) kommt die Heteroskedastizität durch eine erzwungene Gerade bei gegebener kurvlinearer Beziehung zwischen den Variablen (das sieht in (b) recht kubisch aus). In (d) wäre es beides zusammen, also ein (vermutlich quadratischer) Zusammenhang, bei dem mit steigendem X auch noch die Streuung steigt.

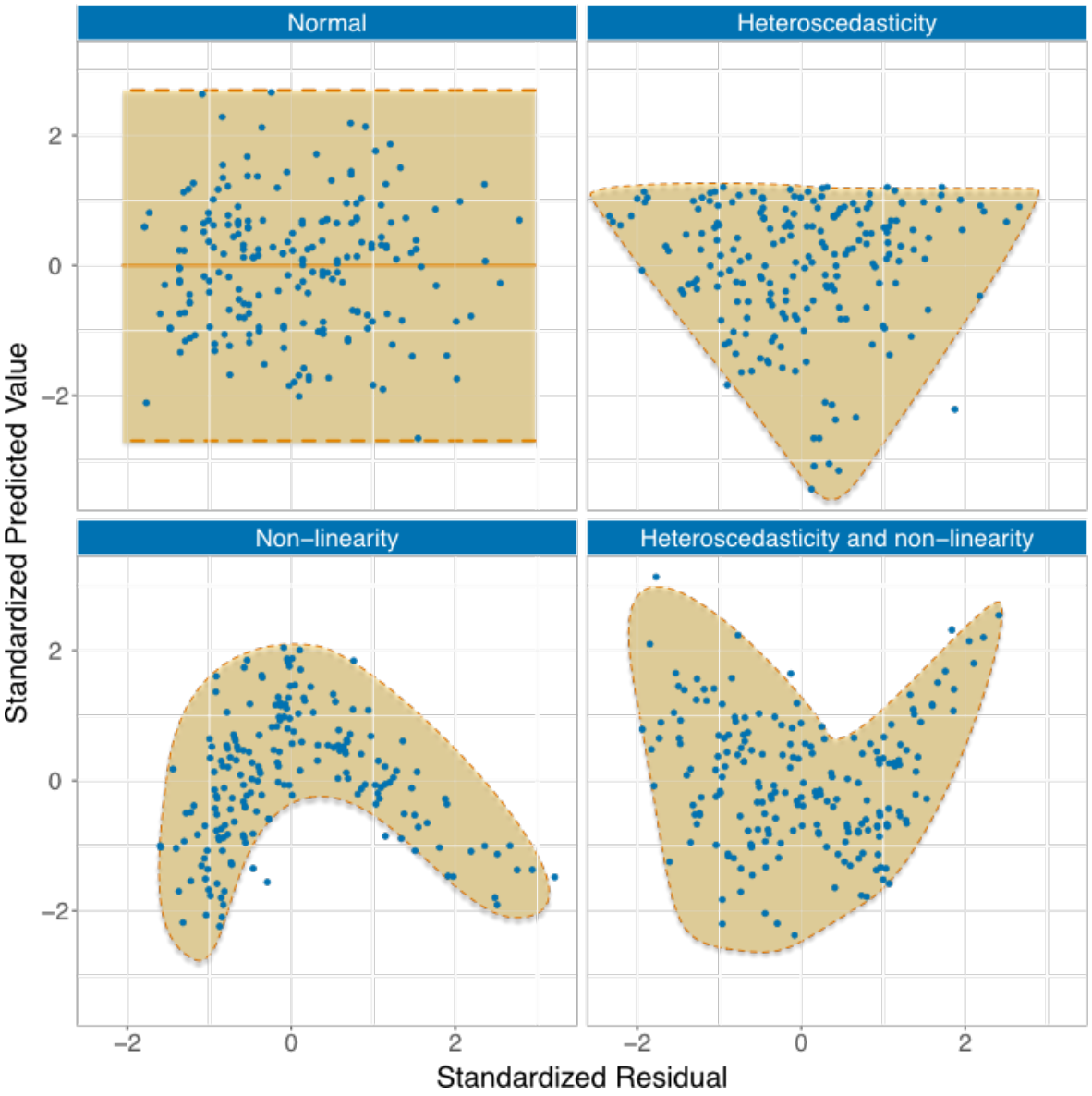


Abbildung 3.5: Nicht-Linearität der Beziehungen

Lösen kann man die Probleme mit der Heteroskedastizität, indem man GLS rechnet, also (**G**eneralized **L**east **Squares**) und dabei zunächst das korrekte b bestimmst, dann die Streuung berechnet und im 2-Stage-Least-Squares mit den gewichteten Residuen rechnen würde. Das zu vermitteln geht über diesen Kurs hinaus. Einfacher ist es mit den kurvilinearen Beziehungen. Die können wir linearisieren. Wir schauen uns

also die Verteilung der Residuen an und wenn wir da so eine kurvlineare Beziehung sehen, dann modellieren wir die so, dass sie linear geschätzt werden kann. Das ist gut in Abb. @ref(fig:Kurvlineare) abgebildet. Dabei ist nicht entscheidend, dass Sie jetzt schon den Aufbau der Formel verstehen, sondern, dass es komplexere Formeln gibt als die einfache additiv lineare, und durch diese Formeln doch wieder das lineare Modell angewendet werden kann, weil die Formeln für eine «Linearisierung» (Transformation) sorgen.

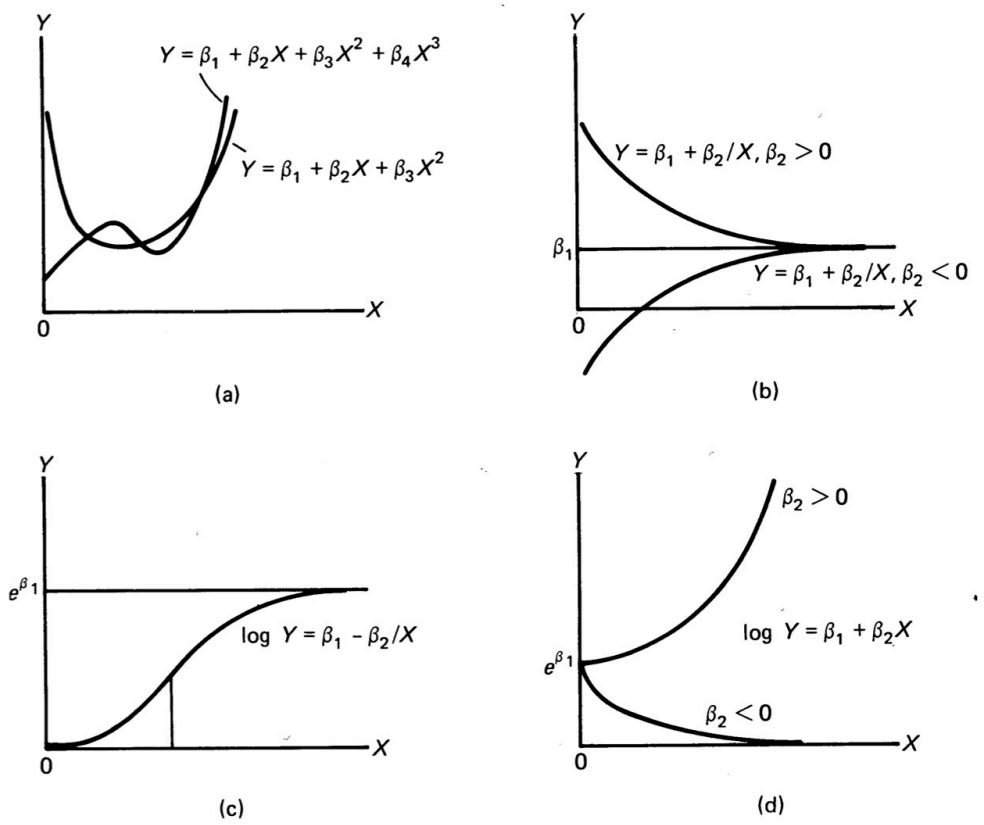


Abbildung 3.6: Linearisierung kurvlinearer Beziehungen

IYI (Klausur): Herleitung, wann OLS "Best" ist

The proof that b_2 is best is only sketched here. A complete proof is shown in Appendix 5.1. The proposition to be demonstrated is that, among all linear and unbiased estimators of β_2 and β_3, the least squares estimators b_2 and b_3 have the minimum variance when assumptions A.1-A.4 hold. We first define an arbitrary linear estimator $b_2^{\#}$. Linear refers to the fact that the estimator is a linear function of the $Y_i, b_2^{\#} = \Sigma C_{i2}^{\#}\left(Y_i - \bar{Y}\right)$, where $C_{i2}^{\#}$ is any set of weights. (The weights are

$$C_{i2} = \frac{1}{n}\frac{V_3\left(X_{i2} - \bar{X}_2\right) - C_{23}\left(X_{i3} - \bar{X}_3\right)}{D}$$

for the least squares estimator of β_2.) With complete generality, we can write $C_{i2}^{\#}$ as the least squares weight plus an arbitrary number $g_{i2}, C_{i2}^{\#} = C_{i2} + g_{i2}$. The restriction of unbiasedness implies that $E\left[\Sigma C_{i2}^{\#}\left(Y_i - \bar{Y}\right)\right] = \beta_2$. However, for OLS we showed that $E\left[\Sigma C_{i2}\left(Y_i - \bar{Y}\right)\right] = \beta_2$. This implies that $E\left[\Sigma g_{i2}\left(Y_i - \bar{Y}\right)\right] = 0$ since $E\left[\Sigma C_{i2}^{\#}\left(Y_i - \bar{Y}\right)\right] = E\left[\Sigma C_{i2}\left(Y_i - \bar{Y}\right) + \Sigma g_{i2}\left(Y_i - \bar{Y}\right)\right]$.

Using this restriction and assumption A.4, the variance of $b_2^{\#}$ is

$$\operatorname{var}\left(b_2^{\#}\right) = \operatorname{var}\left(b_2\right) + \sigma^2 \sum_{i=1}^{n} g_{i2}^2,$$

where $\operatorname{var}\left(b_2\right)$ is the variance of the least squares estimator. Since $g_{i2}^2 \geqslant 0$, $\operatorname{var}\left(b_2^{\#}\right)$ cannot be less than the variance of the least squares estimator b_2. Further, it can equal $\operatorname{var}\left(b_2\right)$ only if each perturbation $\left(g_{i2}\right)$ from the least squares weight is identically zero. (Similar developments can be done for b_1 and b_3.)
An important aspect of this development, however, is that we have accepted and used the assumptions about the error distribution and fixed X 's. That is, this proof holds when $E\left(U_i - \bar{U}\right) = 0, E\left(U_i - \bar{U}\right)^2 = \sigma^2$, and $E\left(U_i - \bar{U}\right)\left(U_s - \bar{U}\right) = 0$ for all i and $s \neq i$.

3.7 Verteilung der Residuen (V7. und V8.)

Ein Modell und die zugrundeliegenden Beziehungen ist oft dann gut, wenn die Verteilung der nicht erklärten Varianzanteile sich wie eine einfache Zufallsverteilung verhält bzw. wie Schrott.

3.7.1 Normalverteilung der Fehler (V7.)

Die Residuen (also der nicht erklärte Rest bzw. Modellfehler oder einfach Fehler) bezieht sich immer auf die nicht erklärte Streuung in der AV. Wenn wir also unser Modell haben und mit unseren Daten berechnen, dann bekommen wir vorhergesagte Werte und den Rest. Wenn wir den Rest anschauen, dann sollte der nicht zu stark von einer Normalverteilung abweichen.

In der Abb. @ref(fig:Hetero1) sieht man recht gut, dass links eine relativ gleichmässige Verteilung vorliegt, also kein Zusammenhang zwischen Fehlern und geschätzten Werten zu erkennen ist (Wäre perfekt 0, wenn die rote Linie exakt auf der gestrichelten Null-linie liegen würde.). Im zweiten Fall namens «Case 2» sieht man deutlich, dass es hier eine kuvlineare Abweichung gibt. Hier würde es sich sicher lohnen, ein quadratisches Modell anzupassen.

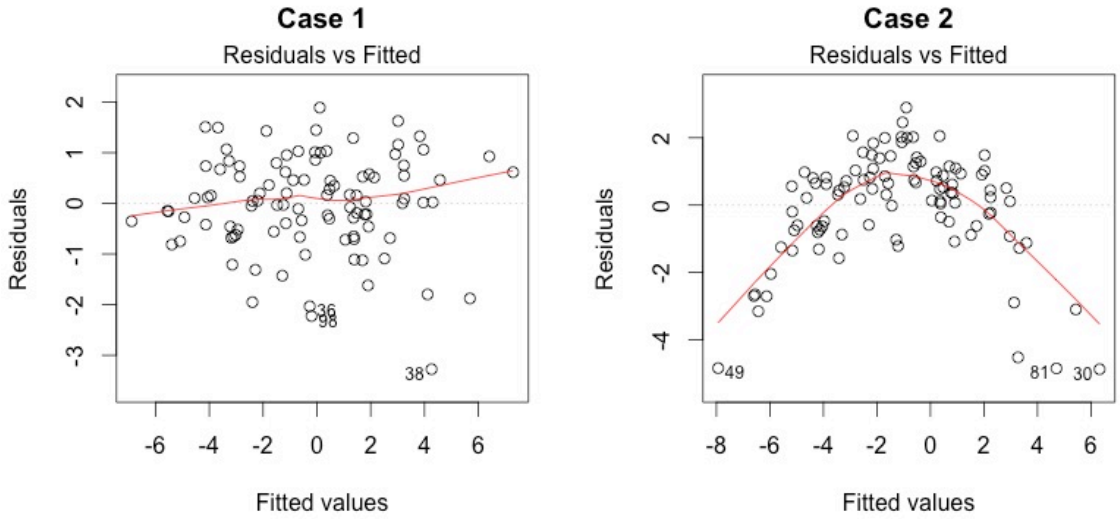


Abbildung 3.7: Residuen gegenüber Modell

In der Grafik @ref(fig:Hetero2) sind Normal Q-Q-Plots abgebildet. Bei dieser visuellen Darstellung werden die standardisierten Residuen gegen die theoretischen Quantile abgetragen, wobei «theoretisch» hier die zu erwartende Verteilung nach Wahrscheinlichkeitstheorie also nach Normalverteilung. Wenn die Punkte alle auf der Gerade liegen, dann ist der Normalverteilung der Residuen nicht stark widersprochen. Wenn sie, wie im zweiten Fall (typisch Case 2!) abweichen, dann ist die Annahme der Normalverteilung verletzt. Dann würden wir nach einem R-Paket suchen, das mit diesem Problem umgehen kann.

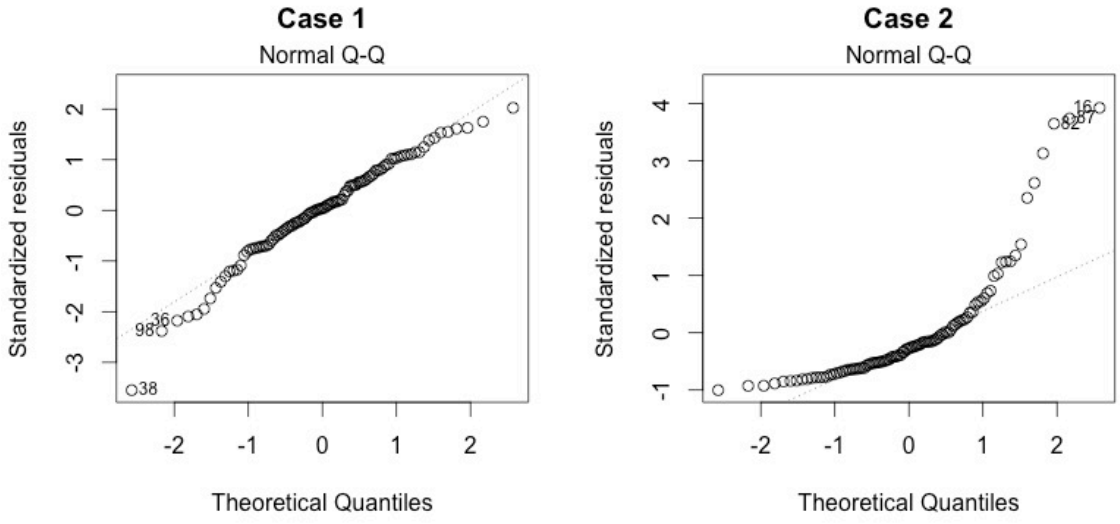


Abbildung 3.8: Normal Q-Q

IYI (Klausur): Ableitung für $U_i \sim N(0, \sigma^2)$.

If $U_i \sim N(0, \sigma^2)$, and independent of U_s, then the b 's, which are linear functions of U_i, are normally distributed with a variance given in Eq. (3.5a), that is, $b_2 \sim N(\beta_2, \sigma^2/nV_2(1-r_{23}^2))$. This implies that

$$Z_2 = (b_2 - \beta_2)/\sigma_{b_2} \quad \text{and} \quad Z_3 = (b_3 - \beta_3)/\sigma_{b_3}$$

are distributed according to the standard normal distribution $N(0,1)$. (This distribution is discussed in Appendix I.) Since the Z's are $N(0,1)$, standard tables of cumulative normal distributions would yield the probability of Z being greater than any given value, or in turn would give the probability of the estimated value being more than any given distance from the true value of β, i.e., $b-\beta$. However, (3.16) depends upon σ, which is unknown. s^2 is an unbiased estimator of σ^2 and is substituted for σ in our expression for σ_b. However, any inferences employing s^2 will not be as precise as they would be if we knew σ^2 and did not have to rely upon the random variable s^2. In order to allow for this additional imprecision, we do not use probabilities from the normal distribution. Instead, we shall rely upon the t-distribution. The definition of the t-distribution is as follows: if Z is a standard normal variable, i.e., Z is $N(0,1)$, and if W^2 is an independently distributed chisquared with $n-3$ degrees of freedom, then $Z/\sqrt{W^2/(n-3)}$ is distributed according to the t-distribution with $n-3$ degrees of freedom. We have demonstrated that $(b-\beta)/\sigma_b$ is $N(0,1)$. It can further be shown that (Hoel, 1962, pp. 262-268)

$$W^2 = \sum e_i^2/\sigma^2 \quad \text{is} \quad \chi^2_{n-3}.$$

For b_2 we get

$$t_{b_2} = \frac{Z}{\sqrt{W^2/(n-3)}} = \frac{(b_2-\beta_2)/\sigma_{b_2}}{\sqrt{\sum e_i^2/\sigma^2(n-3)}} \tag{3.92}$$

$$= \frac{(b_2-\beta_2)}{(\sigma/n)\left[1/V_2(1-r_{23}^2)\right]} \frac{1}{\sqrt{\sum e_i^2/\sigma\sqrt{n-3}}} = \frac{b_2-\beta_2}{(s/n)\left((1/V_2)/(1-r_{23}^2)\right)} \tag{3.93}$$

$$= \frac{b_2-\beta_2}{s_{b_2}}, \tag{3.94}$$

where $s^2 = \sum e_i^2/(n-3)$. This variable t_{b_2} is distributed as Student's t with $n-3$ degrees of freedom.

$$t_b = |b|/s_b > t_{\text{crit }(\alpha/2, n-3)},$$

$t_{\text{crit }(\alpha/2,n-3)}$ is the critical value for $n-3$ degrees of freedon ificance level of α. The significance level α is the size of a type I probability of rejecting the null hypothesis when it is in fact ole terms, the further b is from zero (i.e., the higher t_b), the less lik β is really zero. The general form of the hypothesis test for H_0 :

$$t_b = |b - \beta^*|/s_b.$$

$> t_{\text{crit }(\alpha/2,n-3)}$, the null hypothesis is rejected. In other words

3.7.2 Unabhängigkeit der Fehler (V8.)

Die Unabhängigkeit der Fehler ist eigentlich nur dann ein echtes Problem, wenn die Fehler in eine Reihenfolge gebracht werden können. Das wiederum passiert eher nur bei Zeitreihen, also wenn die Werte einer Erhebung zeitlich angeordnet sind. Dafür gibt es dann allerdings die recht komplexen Zeitreihenanalysen, die eher Statistik IV im Master darstellen. Wir können uns in der R-Übung mal den Durbin-Watson-Test anschauen (zum Spass die Formel (**??**)eq-DWT), wo man schon sieht, dass nicht der Index i für Fälle, sondern t durchläuft für **t**ime), der prüft, ob die Fehler autokorreliert sind, also hoch mit der um eine

Zeiteinheit versetzten Version ihrer selbst korrelieren. Was Sie mitnehmen sollten ist, dass sie bei Erhebungen über die Zeit (Longitudinalstudien), noch prüfen müssen, ob bzw. inwieweit die Fehler miteinander korrelieren.

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{3.95}$$

4 Übung: GLM I

Die Inhalte zur R-Übung finden Sie derzeit nur online.

Wer Probleme mit dem import der Daten von «SosciSurvey» hat, kann die Daten auch hier herunterladen. Speichern Sie die im Unterordner «data» in dem Ordner, in dem auch Ihre «Uebung_1_ab.qmd» liegt, also die Datei, die Sie für die Übung angelegt haben. Wer die Lösung der Übung als Quarto-Datei (Uebung_1_ab.qmd) insgesamt runterladen möchte, kann das hier tun (speichern Sie sich die Datei in einem für die Übung geeigneten Ordner auf Ihrem Recher ab). Und Sie können hier eine die Installation.R herunterladen , die Ihnen im besten Fall noch bei der Installation hilft. Inzwischen ist das aber auch in der Uebung_1_ab.qmd von oben integriert, so dass Sie nach dem Ausführen des entsprechenden Chunks zum Download der Installation.R, diese Datei in Ihrem Projektunterordner «files» finden sollten.

5 GLM – kategoriale UV

Auch als Gruppenvergleiche oder Varianzanalyse (ANOVA) bzw. Multivariate Varianzanalyse (MANOVA/MANCOVA) bekannt.

5.1 Gruppenvergleiche (ANCOVA)

Nominale UVs sind dichotome Variablen (zwei Ausprägungen) oder kategoriale (mehr als zwei Ausprägungen). Wenn wir mit solchen nominalen Variablen Unterschiede erklären wollen (die auf Zusammenhänge bzw. Kausalität zurückgehen), bilden wir mit diesen Variablen Gruppen der Fälle, die wir dann vergleichen. Der Vergleich besteht in der Regel in der Prüfung der Signifikanz von Unterschieden. Das können t-Tests für Mittelwertunterschiede sein. Varianzanalysen für zwei oder mehr Gruppen. Oder auch Regressionen, wo die nominalen Variablen als UV bzw. UVs eingehen.

Der Datensatz, der für dieses Beispiel herangezogen wird, ist von Andy Field. Er hatte einen Artikel darüber gelesen, dass Ingineure einen Stoff entwickeln, der wie unsichtbar macht, indem irgendwie der Hintergrund auf den Umhang projeziert wird oder so. Jedenfalls hat sich Andy Field dann überlegt, was die Leute mit so einem Unsichtbarkeitsumhang (Cloak) wohl für Schabernack (Mischief) anstellen würden, wenn sie von ihrer Umgegung nicht mehr beobachtet werden. Dafür hat Andy Field ein Experiment in der Zukunft imaginiert, bei dem 12 Personen kein Umhang gegeben wird und 12 ein Tarnmantel, der unsichtbar macht. Dann wird gemessen, wie viel Unsinn die Leute jeweils anstellen. Die durchschnittliche Anzahl von Schabernackstaten (Mischief) wird dafür zwischen der Experimental- (hat einen Cloak an) und der Kontrollgruppe (kein Cloak) verglichen.

5.2 Visualisierung und Deskriptives

Gruppenvergleiche können schon gut mit Boxplots gemacht werden. Dabei wird der Mittelwert in einer Box als Linie dargestellt und die das untere sowie das obere Quartil (25% bzw. 95% der Verteilung) als Ränder der Box.

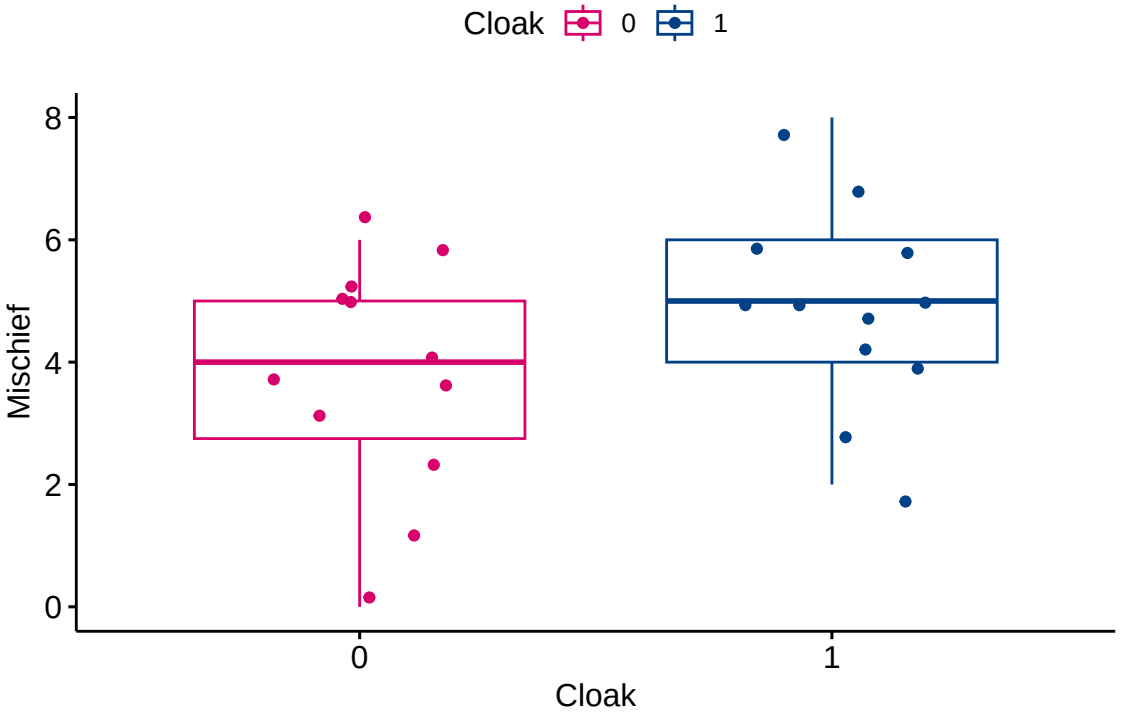


Abbildung 5.1: Mittelwerte (Boxplot) für Gruppenvergleich

Es können aber auch Histogramme erstellt werden, mit Mittelwerten für zwei Gruppen, wobei die Balken für die einzelnen Werte bzw. Wertegruppen überlagert sind.

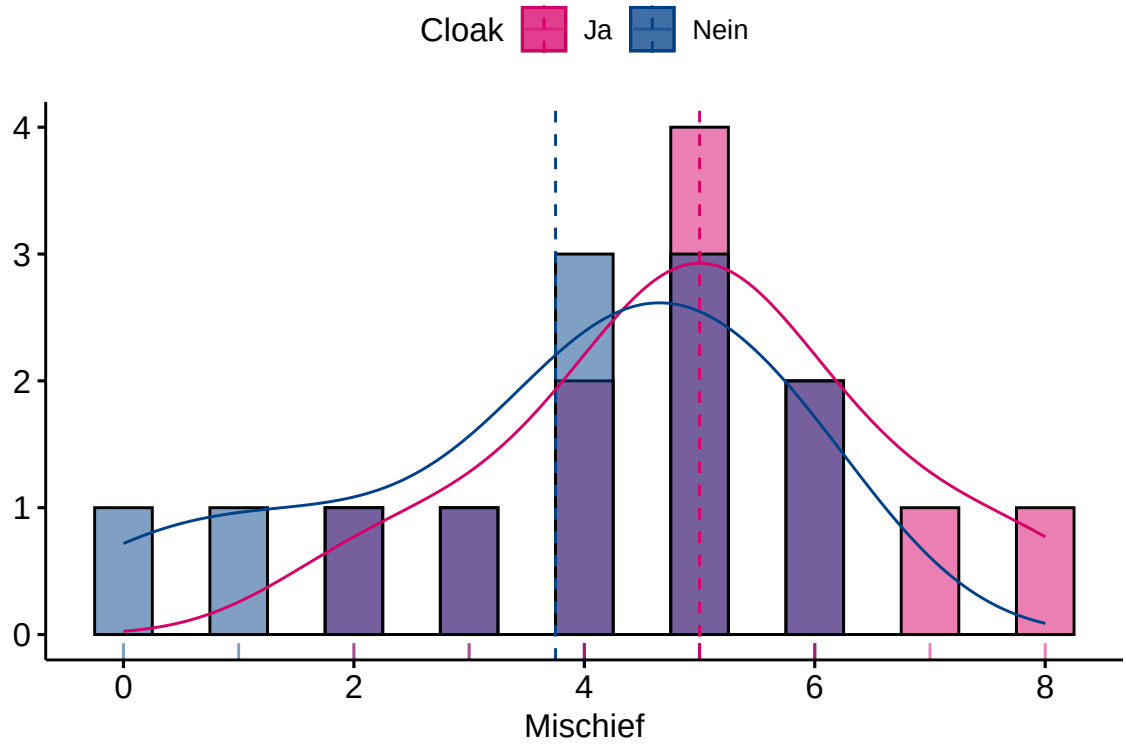


Abbildung 5.2: Mittelwertvergleich der Unsichtbarkeit

Spätestens an dieser Stelle sollte man sich die Ausgangsvariablen mal angucken, um zu sehen, wie die verteilt ist und wo ihr Mittelwert liegt und wie sie um ihren Mittelwert streut und all das. Dafür ist es immer sinnvoll sich die Variablen als Häufigkeitsauszählung anzusehen.

```
## Cloak of invisibility (Cloak) <numeric>
## # total N=24 valid N=24 mean=0.50 sd=0.51
##
## Value |    Label |  N | Raw % | Valid % | Cum. %
## ------------------------------------------------
##     0 | No Cloak | 12 |    50 |      50 |     50
##     1 |    Cloak | 12 |    50 |      50 |    100
##  <NA> |     <NA> |  0 |     0 |    <NA> |   <NA>
##
## Mischievous Acts (Mischief) <numeric>
## # total N=24 valid N=24 mean=4.38 sd=1.86
##
## Value | N | Raw % | Valid % | Cum. %
## ----------------------------------
##     0 | 1 |  4.17 |    4.17 |   4.17
##     1 | 1 |  4.17 |    4.17 |   8.33
##     2 | 2 |  8.33 |    8.33 |  16.67
##     3 | 2 |  8.33 |    8.33 |  25.00
##     4 | 5 | 20.83 |   20.83 |  45.83
##     5 | 7 | 29.17 |   29.17 |  75.00
##     6 | 4 | 16.67 |   16.67 |  91.67
##     7 | 1 |  4.17 |    4.17 |  95.83
```

```
##     8 | 1 |  4.17 |    4.17 | 100.00
##  <NA> | 0 |  0.00 |    <NA> |   <NA>
```

Und man sollte sich die Mittelwerte der AV (hier Mischief) und die Gruppierungsvariable (auch UV und hier Cloak) ausgeben lassen.

```
## # A tibble: 2 x 2
##   Cloak        Mittelwerte
##   <dbl+lbl>          <dbl>
## 1 0 [No Cloak]        3.75
## 2 1 [Cloak]           5
```

5.3 Mittelwertvergleich für zwei Gruppen

5.3.1 mit dem t-Test

Mit dem t-Test kann geprüft werden, ob sich die Mittelwerte der beiden Gruppen unterscheiden. Es wird ein t-Test für unabhängige Stichproben gemacht. Dabei wird die Differenz der beiden Mittelwerte berechnet und gegen die H0 getestet, dass sie 0 sein könnte, also in der GG kein Unterschied zwischen der Gruppe Cloak = 1 und der Gruppe Cloak = 0.

```
##
##  Welch Two Sample t-test
##
## data:  Mischief by Cloak
## t = -1.7135, df = 21.541, p-value = 0.101
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.764798  0.264798
## sample estimates:
## mean in group 0 mean in group 1
##            3.75            5.00
```

Der Output sagt uns, dass die Mittelwerte von Mischief aufgeteilt nach Cloak angeschaut werden. Der t-Wert unter Annahme der Nullhypothese H0 ist -1.7135 und der zugehörige p-Wert ist 0.101. Im Text steht noch, dass die Alternativhypothese lautet: Der wahre Mittelwertunterschied zwischen der 0-Gruppe und der 1-Gruppe ist nicht gleich 0. Darunter steht das 95-prozentige Konfidenzintervall der Mittelwertdifferenz. In der untersten Zeile werden die beiden Mittelwerte der beiden Gruppen nochmals ausgegeben.

5.3.2 Mit Korrelation

Wenn die Gruppenvariable eine Dummyvariable ist (also dichotom und nur aus 0 und 1 bestehend), dann kann auch eine Korrelation gerechnet werden, wobei der t-Wert dann derselbe ist, wie beim t-Test von Mittelwertvergleichen für unabhängige Stichproben.

```
##
##  Pearson's product-moment correlation
##
## data:  Invisibility$Mischief and Invisibility$Cloak
## t = 1.7135, df = 22, p-value = 0.1007
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06994687  0.65575942
## sample estimates:
```

```
##       cor
## 0.3431318
```

Vergleichen Sie mal den t-Wert und den p-Wert der Korrelation mit dem des t-Test für Mittelwertunterschiede. Die sind (bis auf Rundungsunterschiede) identisch.

Es gibt die einfache Varianzanalyse. Dabei wird geprüft, ob die Gruppierungsvariable signifikant Varianz der AV aufklärt. Der p-Wert ist derselbe, wie oben beim t-Test und der Korrelation, weil es dieselben Daten und Variablen sind.

5.3.3 Gruppenvergleich mit Varianzanalyse

```
# Mache eine Varianzanalyse (Analysis of Varian (aov bzw. ANOVA)) mit einer
# UV (one.way)
one.way <- aov(Mischief ~ Cloak, data = Invisibility)

# Gib die Zusammenfassung der aov raus
summary(one.way)
##             Df Sum Sq Mean Sq F value Pr(>F)
## Cloak        1   9.38   9.375   2.936  0.101
## Residuals   22  70.25   3.193

# Berechne mal das R^2 durch die Quadratsumme (Sum Sq), die die Gruppierung
# (hier nach Cloak) aufklärt, durch die Gesamtquadratsumme (Sum Sq der Cloak +
#der der Residuals). Dann runde auf 4 Nachkommastellen.

R2 <- round(9.38/(9.38 + 70.25),4)

# Binde das R^2 in die Ausgabe ein, einfach für später
paste0("R2 = Sum_Sq_Cloak / (Sum_Sq_Cloak + Risiduals_SumSq): ", R2," (12%)")
## [1] "R2 = Sum_Sq_Cloak / (Sum_Sq_Cloak + Risiduals_SumSq): 0.1178 (12%)"
```

Die Varainzanalyse prüft, ob die Mittelwerte in einer AV für jede der UV-Gruppen identisch ist. Was in der Tabelle steht, sind lauter Hilfswerte für den einen relevantenn Wert: dem p-Wert (hier "Pr(>F)). Der p-Wert ist wieder derselbe wie oben bei der Korrelation und dem Mittelwertvergleich.

5.4 Mittelwertvergleich mit Regression

Am besten kann mit einer Regression ein Mittelwertvergleich durchgeführt werden. Das R^2 entspricht dem Quadrat der Korrelation. Der F-Wert zum R^2 ist gleich dem F-Wert aus der Varianzanalyse. Der b-Wert (hier von «Cloak of invisibility») in der Regression entspricht dem Mittelwertunterschied zwischen den beiden Gruppen. Der «Intercept» entspricht dem Mittelwert der 0-Gruppe (keine Cloak). Mit der Regression kann also alles abgedeckt werden, was mit den anderen Auwertungsmethoden auch erledigt wird. Die Regression kann aber mehr!

Tabelle 5.1: Regression mit einer Dummmy als UV

UVs	B	std. B	se	t	p
Intercept	3.75	NA	0.516	7.270	<0.001
Cloak	1.25	0.343	0.730	1.713	0.101

$$R^2 = 0.12; \text{ adj. } R^2 = 0.08; \text{ F} = 2.94; \text{ p} = 0.10$$

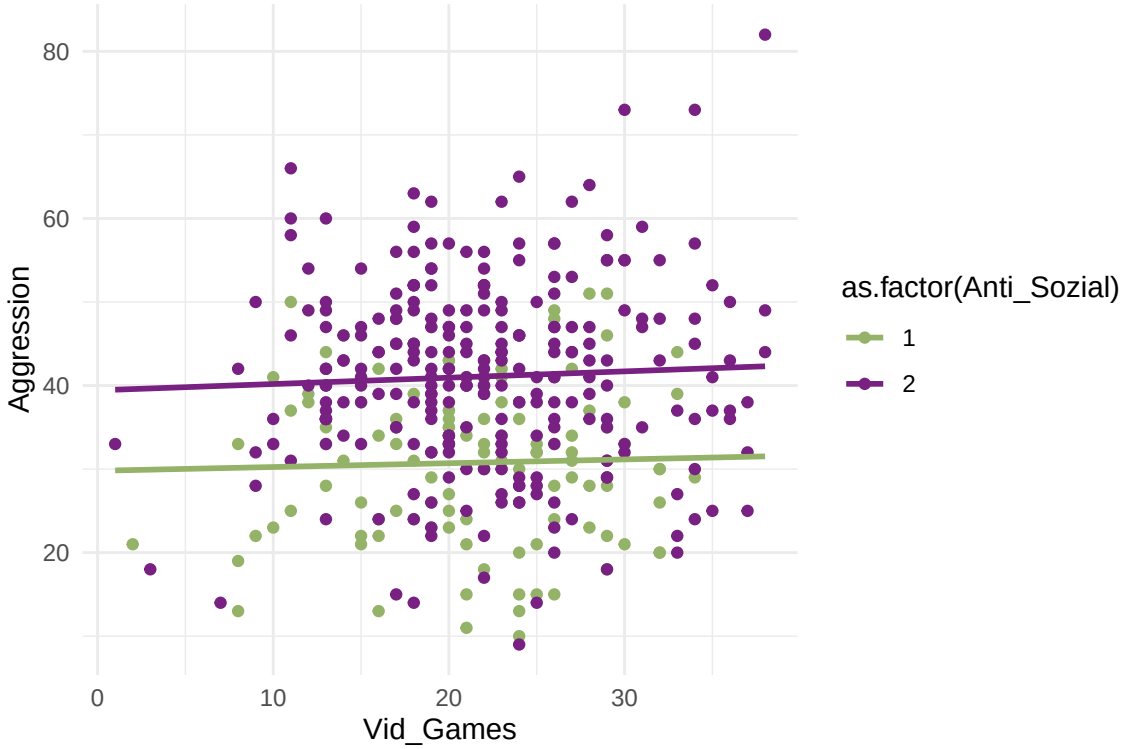
5.5 Interaktionseffekte

Werden wir mal erwachsen und schauen uns ein anderes Beispiel an, das auf eine Medienwirkgungsfrage zurück geht. Gehen wir also jetzt der Frage nach, ob gewalthaltige Videospiele antisozial machen. Dazu hat das britische Ofcom (Office of Communication) 2008 eine Studie herausgegeben [@Ofcom2008]. Für die Studie wurden 442 Jugendliche befragt. Im folgenden Chunkg wird der dazugehörige Datensatz heruntergeladen, umgewandelt und im Datenobjekt «Video_Games» gespeichert. Den analysieren wir im Folgenden. Die Variablen sind «Aggression» als Messung aggressiver Verhaltensweisen, «CaUnTs» als callous unemotional traits (affektiv-soziale defizite) und «Vid_Games» in Stunden Nutzung von Videospielen.

```
## # A tibble: 442 x 4
##       ID Aggression Vid_Games CaUnTs
##    <dbl>      <dbl>     <dbl>  <dbl>
## 1    69         13        16      0
## 2    55         38        12      0
## 3     7         30        32      0
## 4    96         23        10      1
## 5   130         25        11      1
## 6   124         46        29      1
## # i 436 more rows
```

Schauen wir uns das mal genauer an:

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Anti_Sozial = sjmisc::rec(CaUnTs, rec = "0:10 = 1 [gering]; 11:30 = 2
##   [mittel]; 31:max = NA [hoch]")`.
## Caused by warning in `FUN()`:
## ! NAs introduced by coercion
## `geom_smooth()` using formula = 'y ~ x'
```



Wie man sieht gibt es die zwei Gruppen. Wenn die Tage pro Woche mit Videospielen steigt, dann steigt die Aggression kaum an. Das ist für beide Gruppen so, ber für die 2-er-Gruppe (mittleres Antisoziales Verhalten) liegen die Werte im Mittel höher. Das haben wir jetzt gesehen, aber geschätzt und getestet haben wir es noch nicht. Das geht aber gut mit der Regression.

5.6 Eine Dummy als UV

Wenn wir eine Dummyvariable als UV haben, dann haben wir es eigentlich mit einem Unterschiedstest zu tun, also einem Mittelwertvergleich. Vergleichen werden dabei die Mittelwerte der UV für zwei Gruppen. Die Gruppen wiederum werden durch in der Dummyvariable festgelegt: Die eine Gruppe (G0) hat die 0 und die andere Gruppe (G1) die 1. Es wird also die Differenz in den Y-Werten (Y_Diff) durch die Dummyvariable erklärt.

$$\overline{Y}_{Diff} = \overline{Y}_{G1} - \overline{Y}_{G0} \tag{5.1}$$

$$Y_i = b_1 + b_2 X_{i2} \tag{5.2}$$

$$Y_i = b_1 \quad \text{wenn } X_{i2} = 0 \tag{5.3}$$

$$Y_i = b_1 + b_2 \quad \text{wenn } X_{i2} = 1 \tag{5.4}$$

$$\text{Also ist: } \overline{Y}_{Diff} = b_2 \tag{5.5}$$

$$t = \frac{\overline{Y}_{G1} - \overline{Y}_{G0}}{se_{\overline{Y}_{Diff}}} = \frac{b_2}{se_b} \tag{5.6}$$

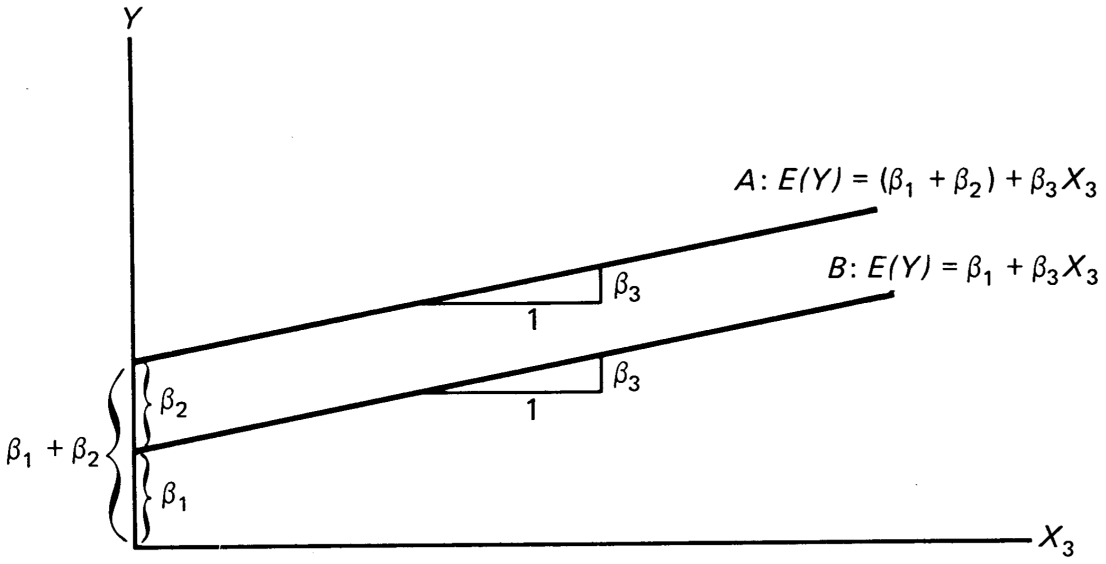


FIGURE 4.3 Bivariate relationship with intercept dummy variable.

Hier ist nur «Anti_Soz_mittel» als UV im Modell.

Agression
Predictors
b
std. b
CI
standardized CI
p
(Intercept)
30.76
-0.00
28.71 – 32.81
-0.09 – 0.09
<0.001
Callous UnemotionalTraits
10.33
0.40
7.91 – 12.75
0.30 – 0.49
<0.001
Observations
381
R2 / R2 adjusted
0.157 / 0.155
Regression mit einer Dummy als UV

Interpretation der Regression: Der (Intercept) hat im b eine 30.76 und zeigt daher in diesem Modell an, wie gross der Mittelwert für die Referenzgruppe ist (0 für Anti_Soz_mittel «nicht mittel»). Das b für die «Callous Unemotional Traits» liegt bei 10.33. Das bedeutet, dass der Mittelwert der Gruppe Anti_Soz_mittel = 1 um 10.33 grösser ist als der Mittelwert der 0-Gruppe, also 41.09. Dieser Unterschied entspricht einem Zusammenhang von .4 als Korrelation, was an dem standardisierten b abgelesen werden kann, weil die standardisierten Regressionskoeffizienten (oft auch als BETA bezeichnet) sehr dicht an den

Korrelationskoeffizienten sind. Das Konfidenzintervall für den Mittelwertunterschied liegt zwischen 7.91 und 12.75. Da 0 nicht mit im Intervall liegt, sehen wir schon, dass der Mittelwertunterschied signifikant ist. Wir sehen aber nicht nur, dass der Mittelwertunterschied signifikant von 0 verschieden ist, sondern auch, dass er signifikant von z.B. 5 verschieden ist. Wenn jetzt zum Beispiel andere Forscherinnen das Phänomen vorher schon untersucht gehabt hätten und die Mittelwertunterschied zwischen 1.93 und 4.25 gefunden hätten, dann könnten wir mit der Analyse hier sagen, dass sich die beiden Konfidenzintervalle nicht überschneiden, also unser Ergebnis signifikant von dem der anderen Forscher ist. Das geht schon in die Richtung Metaanalyse. Wenn wir nochmal in die Tabelle schauen, dann sehen wir hinten auch, dass die p-Werte unter .05 liegen, was eine Signifikanz auf dem 95%-igem Signifikanzniveau anzeigt. Das wussten wir über die CI aber auch schon vorher und da wussten wir sogar mehr!

5.6.1 Dummykodierung

Q&A: Wie viele Dummyvariablen brauchen Sie, um die volle Information einer kategorialen Variablen mit vier Ausprägungen abzubilden?

Sie brauchen 3 Dummys im Modell. Das kommt daher, dass Sie 4 Ausprägungen einer kategorialen in 4 Dummys umkodieren würden. Wenn Sie zum Beispiel die Sprachregionen der Schweiz abgefragt haben, bestünde die kategoriale Variable aus 1. Deutsch, 2. Französisch, 3. Italienisch, 4. Rätoromanisch. Da Sie nach den Sprachregionen gefragt haben, in denen die Befragten ihren Wohnsitz haben, schliessen sich die Antwortmöglichkeiten aus (sind disjunkt). Nur deshalb können Sie überhaupt in einer Variable erfasst werden. Würden Sie danach fragen, welche Sprachen die Leute verstehen, würden Sie 4 Variablen anlegen, bei der jede:r Befragte auch zwei, drei oder alle vier Sprachen angeben könnte. Jede Sprache würde durch eine Dummyvariable gekennzeichnet sein, also eine Dummy für DE, eine für FR, eine für IT und eine für RR. Jeweils hätten die eine 1, wenn die jeweilige Sprache verstandne wird und eine 0, wenn nicht. So eine Dummykodierung können Sie aber auch für die Sprachregion machen, also die kategoriale Variable in vier Dummys für die Sprachregion umkodieren, in der die Leute leben. Das können Sie durch Umkodierung machen, indem man je Sprachregion sagt:
Für die Deutschschweiz DS:

- Wenn in der (kategorialen) Sprachregion eine 1 (für DS), dann in der Dummy DS eine 1, sonst immer eine 0.
- Wenn in der Sprachregion eine 2 (für FS), dann in der nächsten Dummy FS eine 1, sonst immer 0.
- Wenn in der Sprachregion eine 3 (für IS), dann in der nächsten Dummy IS eine 1, sonst immer 0.
- Wenn in der Sprachregion eine 4 (für RRS), dann in der nächsten Dummy RRS eine 1, sonst immer 0.

Damit hätten Sie die Kategoriale in 4 Dummys umkodiert und könnten die in eine Regression integrieren. R (und kein anderes Regressionsprogramm) würde das dann berechnen, weil es eine 100% Multikollinearität zwischen den Variablen gäbe: Wenn Sie die Ausprägungen von drei der Dummys kennen, können Sie exakt sagen, welche Ausprägung die Vierte hat. Also müssen Sie in der Regression eine Dummy weglassen. Das sollte immer am besten die grösste Gruppe sein, die damit zur Referenzgruppe wird. Im Beispiel würde man also die DS rauslassen.
Antwortsatz in der Klausur: Eine kategoriale Variable mit 4 Ausprägungen wird in Form von 3 Dummyvariablen in das Modell integriert (weil sie in 4 Dummys kodiert würde und eine der Dummys weggelassen wird, die damit die Referenzkategorie darstellt).

Tabelle 5.2: Dummmykodierung einer kategorialen Variable mit 4 Ausprägungen

Region	Dummy_DS	Dummy_FS	Dummy_IT	Dummy_RRS
DS	1	0	0	0
FS	0	1	0	0
IS	0	0	1	0
RRS	0	0	0	1

IYI: Effektkodierung

Bei der Dummykodierung muss man immer eine Referenzkategorie rauslassen, mit der dann die b's der Dummys verglichen werden (Mittelwertunterschied zwischen den Gruppen die in den Dummys eine 1 haben und der Referenzgruppe, wenn es keine Interaktionen gibt). Nun möchte man vielleicht nicht immer eine Gruppe raus haben und gegen die Gruppe vergleichen, sondern Aussagen darüber treffen, ob die einzelnen Gruppen signifikant über oder unter dem Gesamtdurchschnitt liegen. Das geht, indem man eine sogenannte «Effektkodierung» vornimmt.
Bei der Effektkodierung bekommt immer eine Gruppe bei allen Zugehörigen eine -1 und die anderen Gruppen eine 1. Dann werden alle Dummys in das Modell mitaufgenommen. Die b's dieser Effektkodierten Dummys geben immmer den Abstand zum Gesamtmittelwert wieder. Sind die b's signifikant, ist der Unterschied zum Gesamtdurchschnitt signifikant.

Tabelle 5.3: Effektkodierung einer kategorialen Variable mit 4 Ausprägungen

Region	Dummy_DS	Dummy_FS	Dummy_IT	Dummy_RRS
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	-1

5.7 Dummy und Covariate

Jetzt wird das Modell um eine Covariate ergänzt. Mit `olsrr::ols_vif_tol(Modell3)` werden die Toleranz und der VIF berechnet.

```
##       Variables Tolerance     VIF
## 1     Vid_Games 0.9998605 1.00014
## 2 Anti_Soz_hoch 0.9998605 1.00014
```

Predictors	b	std. b	Agression CI	standardized CI	p
(Intercept)	33.13	-0.00	29.54 – 36.73	-0.09 – 0.09	**<0.001**
Video Games(Hours per week)	0.23	0.13	0.07 – 0.39	0.04 – 0.21	**0.004**

Callous Unemotional Traits	13.65	0.37	10.51 – 16.79	0.29 – 0.46	**<0.001**
Observations	442				
R^2 / R^2 adjusted	0.157 / 0.153				

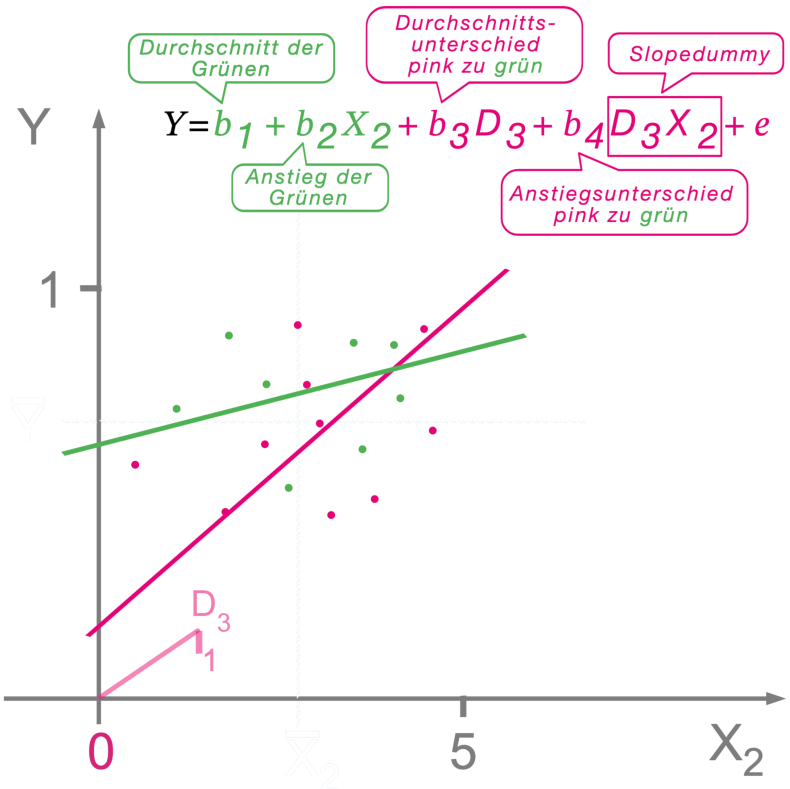
Die Toleranzwerte sind sehr hoch und daher völlig ok. Der Varianzinflationsfaktor ist fast genau 1. Es gibt also eigentlich keine Inflation der Fehlerstreuung der b's (und allem was darauf aufbaut, wie die standardisierten Regressionskoeffizienten, Konfidenzintervalle, t-Wert zum t-Test und also auch die p-Werte). Also ist hier alles gut.

6 GLM – Interaktionen

6.1 Interaktion mit Slope-Dummy

Eine Slope-Dummy ist ein Produkt aus einer metrischen Variable und einer Dummyvariabeblen. Durch diese Kombination wird es möglich, dass nicht nur der Mittelwert von zwei Gruppen unterschiedlich sein kann, sondern auch die Anstiege der Regressionsgeraden unterschiedlich sein können. Das bedeutet im Grunde, dass man prüfen kann, ob der Zusammenhang einer metrischen Variable für Gruppen unterschiedlich ist. Anders gesagt kann die Fragestellung beantwortet werden, ob zwei Gruppen sich in der Stärke des Zusammenhangs unterscheiden. Also zum Beispiel, ob ein Nachrichtenfaktor bei einer Rezipient:innengruppe anders wirkt (ein anderes Gewicht hat) als bei einer anderen Gruppe.

In der Grafik ist gut zu erkennen, dass die grüne Gruppe und die pinke Gruppe unterschiedliche Anstiege haben und damit die Lage der Werte besser abbilden kann, als würde man nur erlauben, dass die Mittelwerte unterschiedlich sind.



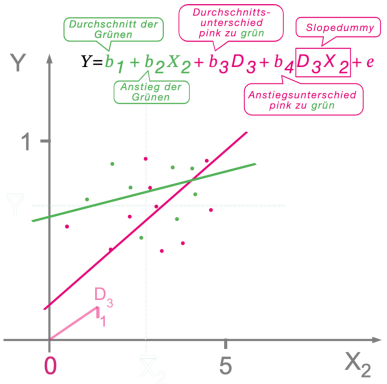
Das b_1 steht für den Schnittpunkt mit der Y-Achse ($X_2 =$) der 0-Gruppe (grün) und b_3 für den Schnittpunkt der pinken mit der Y-Achse. Diese beiden Werte sind nicht ohne Weiteres interpretierbar und sorgen nur dafür, dass die Regressionsgeraden sich frei an die Werte der Gruppen anpassen könnnen. Interpretierbar wird das Ganze, wenn man die metrische Variable X_2 zentriert, also in ihren Mittelwert verschiebt. Dann ist das b_1 der Mittelwert der grünen 0-Gruppe und das b_3 der Mittelwertunterschied zwischen der pinken 1-Gruppe und der 0-Gruppe.

In Worten bedeteutet das also:

Regression mit Slope-Dummy

$$Y = b_1 + b_2X_2 + b_3D_3 + b_4D_3X_2 + e$$

- im Modell gibt es eine metrische Variable, eine Dummy und eine Slope-Dummy
- das b_1 gibt den Schnittpunkt mit der Y-Achse (X=0) der 0-Gruppe wieder
- das b_2 gibt den Anstieg der 0-Gruppe wieder
- das b_3 gibt den Unterschied des Schnittpunkts mit der Y-Achse der 1-Gruppe wieder
- das b_4 gibt den Unterschied im Anstieg der 1-Gruppe wieder

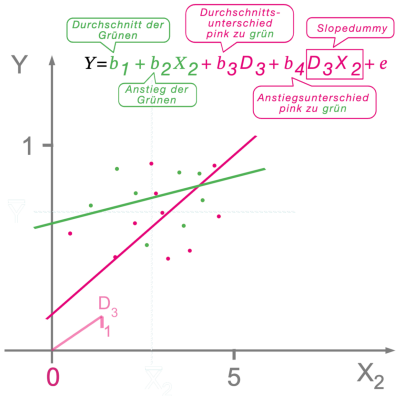


Setzt man einzelne Werte gedanklich auf 0, wird die Formel jeweils klarer. Beachten Sie, dass auch das *1 weggelassen wird, wenn D = 1 ist.

Regression mit Slope-Dummy

$$Y = b_1 + b_2X_2 + b_3D_3 + b_4D_3X_2 + e$$

$$\begin{aligned}
Y &= b_1 + e && |X_2 = 0, D_3 = 0 \\
Y &= b_1 + b_3 + e && |X_2 = 0, D_3 = 1 \\
Y &= b_1 + b_2X_2 + e && |D_3 = 0 \\
Y &= b_1 + b_2X_2 + b_3 + b_4X_2 && |D_3 = 1
\end{aligned}$$



6.2 Beispiel zu Videospielen und Aggression

Mit dem folgenden Beispiel läuft auch die Übung 2. Sie werden also selbst mit denselben Daten arbeiten. Dafür müssen sie zunächst geladen werden.

Wie Sie am Kopf der Datendatei sehen, besteht sie aus drei inhaltlichen Variablen und einer ID.

1. Agression: wurde auf einer breit angelegten Skala gemessen.
2. Vid_Games: wurde als Stunden pro Woche abgefragt.
3. Antisoziales Varhalten wurde ebenfalls mit einer Skala gemessen und kann daher hohe Werte annehmen und ist metrisch.

```
download.file(
  "http://www.discoveringstatistics.com/docs/ds_data_files/SPSS%20Data%20Files/Video%20Games.sav",
  "data/Video_Games.sav", quiet = TRUE)

DATEN <- haven::read_spss("data/Video_Games.sav")
```

```
head(DATEN)
## # A tibble: 6 x 4
##      ID Aggression Vid_Games CaUnTs
##   <dbl>      <dbl>     <dbl>  <dbl>
## 1    69         13        16      0
## 2    55         38        12      0
## 3     7         30        32      0
## 4    96         23        10      1
## 5   130         25        11      1
## 6   124         46        29      1
```

6.3 Zusammenhang Videospiele und Aggression

Die Variable «CaUnTs» wird zunächst in einer kategoriale Variable umkodiert, wobei die 1 für geringes antisoziales Verhalten steht, die 2 für mittleres und die 3 für hohes. Ganz aggressionsfrei ist kaum jemand, aber etwa ein Viertel der Befragten zeigte eher tiefe Werte in der neu gebildeten Variable «Antisozial». Die meisten liegen in der Mitte (62%) und zum Glück nur wenige bei hohen Werten für antisoziales Verhalten (14%).

```
DATEN <- DATEN  |>
  mutate(Anti_Sozial = case_match(CaUnTs,
    c(0:10) ~ 1,
    c(11:30) ~ 2,
    c(31:200) ~ 3,
    .default = NA
  )) |>
  sjlabelled::var_labels(Anti_Sozial = "Antisoziales Verhalten")

DATEN |> sjmisc::frq(Anti_Sozial)
## Antisoziales Verhalten (Anti_Sozial) <numeric>
## # total N=442 valid N=442 mean=1.89 sd=0.61
##
## Value |   N | Raw % | Valid % | Cum. %
## --------------------------------------
##     1 | 108 | 24.43 |   24.43 |  24.43
##     2 | 273 | 61.76 |   61.76 |  86.20
##     3 |  61 | 13.80 |   13.80 | 100.00
##  <NA> |   0 |  0.00 |    <NA> |   <NA>
```

```
# Speichere in dem Datensatz Video_Games_AS_gering man nur die Fälle mit mittlerer
# oder geringem Antisozialem Verhalten.

Video_Games_AS_gering <- DATEN  |>
  filter(CaUnTs < 31)

# Mache einen Scatterplot (geom_point) für Vid_Games und Agression, unterteilt nach
# Anti_Sozial und lege da mit geom_smooth jeweils eine Regressionsgerade rein.

Video_Games_AS_gering  |>
  ggplot2::ggplot(aes(x = Vid_Games, y = Aggression, color = as.factor(Anti_Sozial))) +
  geom_point()+
```

```
  scale_color_manual(values=c(c(Farben[3], Farben[4]))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE) +
  theme_minimal()
## `geom_smooth()` using formula = 'y ~ x'
```

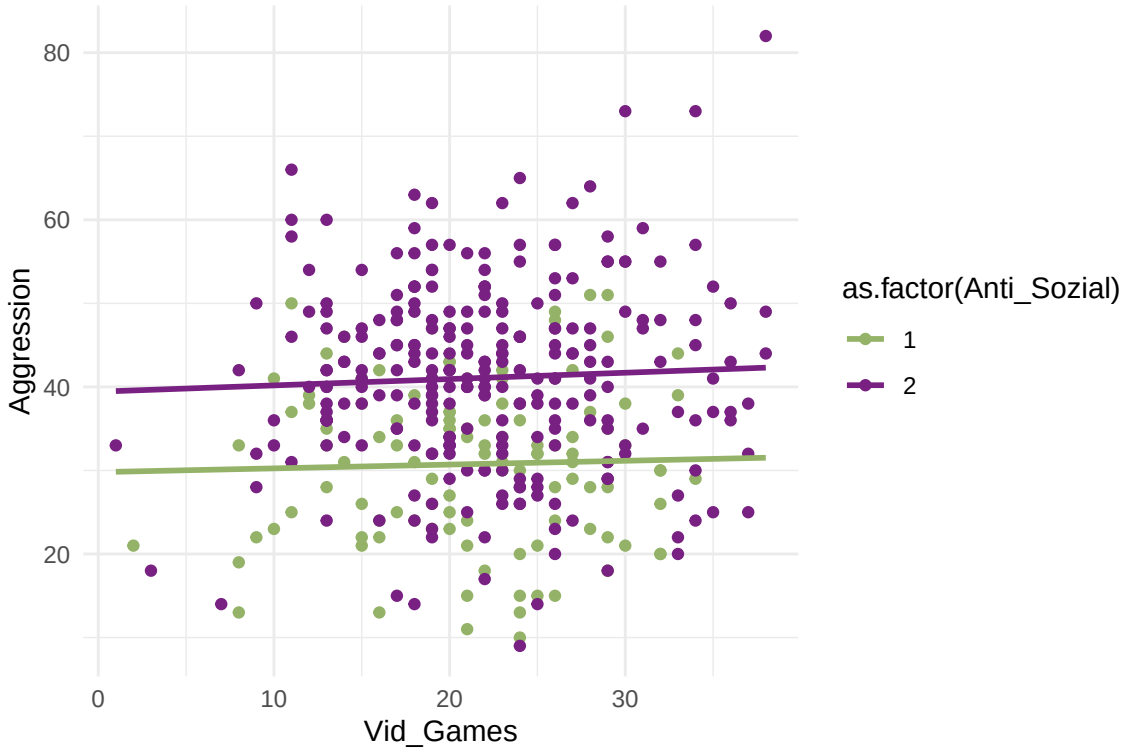


Abbildung 6.1: Antisoziales Verhalten und Aggression

In der Grafik sind die Punktverteilungen gut zu sehen und auch, dass die Gruppe mit hohem Werten in «Antisozial» einen deutlich stärkeren Zusammenhang zwischen der Nutzung von Videospielen und Aggression zeigen.

[Mehr wird hier erstm nicht gespoilert, weil die Interpreltationen der Werte Teil der Übung sind]

6.3.1 Grafik Videogames

```
DATEN <- DATEN  |>
  mutate(Anti_Soz_hoch = case_match(Anti_Sozial,
   3 ~ 1,
   .default = 0
  ),
  Anti_Soz_mittel = case_match(Anti_Sozial,
  2 ~ 1,
  .default = 0))

DATEN  |> sjmisc::frq(Anti_Soz_hoch)
## Anti_Soz_hoch <numeric>
```

```
## # total N=442 valid N=442 mean=0.14 sd=0.35
##
## Value |   N | Raw % | Valid % | Cum. %
## --------------------------------------
##     0 | 381 | 86.20 |   86.20 |  86.20
##     1 |  61 | 13.80 |   13.80 | 100.00
##  <NA> |   0 |  0.00 |    <NA> |   <NA>

DATEN  |>
  ggplot2::ggplot(aes(x = Vid_Games, y = Aggression,
                      color = as.factor(Anti_Soz_hoch))) +
  geom_point()+
  scale_color_manual(values=c(c(Farben[3], Farben[4]))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE) +
  theme_minimal()
## `geom_smooth()` using formula = 'y ~ x'
```

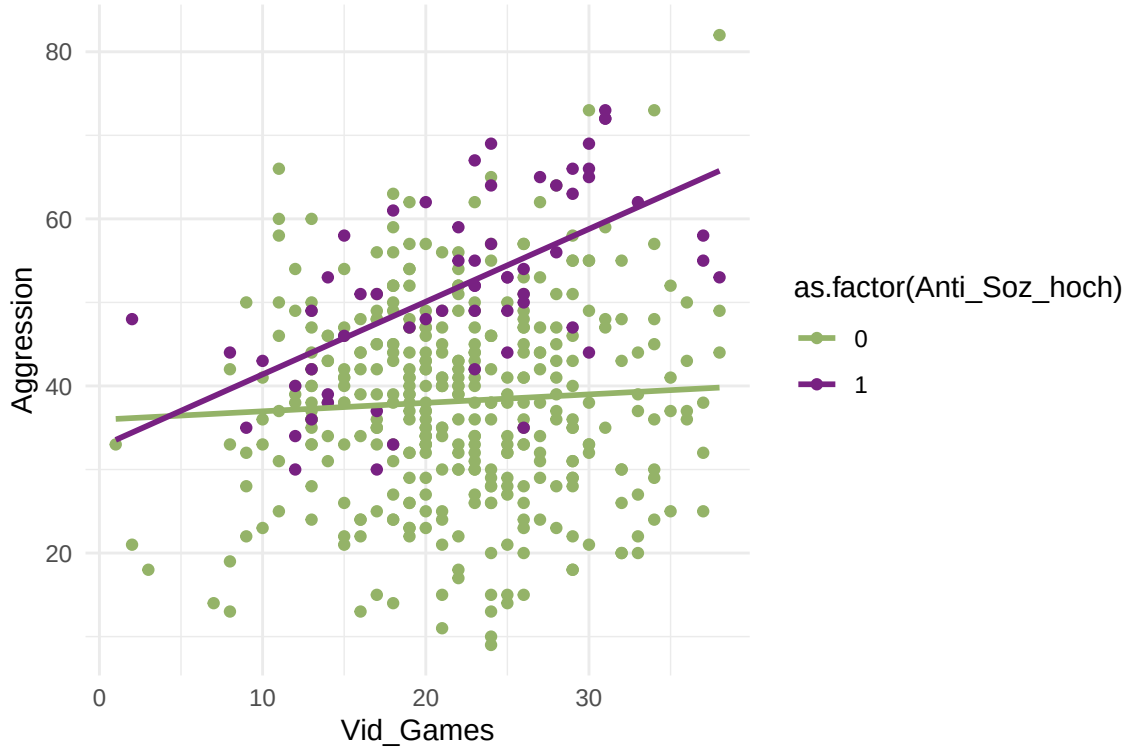


Abbildung 6.2: Zusammenhang Videospiele zu Aggression für Menschen mit hohem vs. geringerem antisozialen Verhalten

6.3.2 Multikollinearität bei Slope-Dummys und Lösungsansätze

Q&A: Welches besondere Problem gibt es bei Slop-Dummys mit Multikollinearität und wie löst man es?

Die Slope Dummy ist das Produkt aus der ursprünglichen Dummy und der metrischen. Da die Metrische und die Dummy aber auch noch im Modell sein müssen, korreliert die Slope-Dummy, wie Sie richtig sagen, mit der Metrischen, aber eben auch mit der Dummy, da die Slope-Dummy bei allen Fällen eine 0 hat, wo die Dummy eine 0 hat und immer dann, wenn die Slope-Dummy positiv ist, die Dummy auch positiv ist (eine 1) hat. Das ist das besondere (eher technische) Problem der Multikollinearität der Slope-Dummy. Das können wir lösen, indem wir die Metrische vorher zentrieren. Dann ist die Slope-Dummy zwar immernoch 0, wenn die Dummy 0 ist, aber sie streut um 0, wenn die Dummy 1 ist. Damit gibt es für den Fall diese technische Multikollinearität nicht mehr. Es bleibt noch, dass die Slope-Dummy mit der Metrischen identisch ist, wenn die Dummy 1 ist. Das ist dann problematisch, wenn die 1-en viel sind, weil die Metrische dann in vielen Fällen mit der Slope-Dummy übereinstimmt. Das können wir lösen, indem wir die Dummy umdrehen (und gut im Kopf behalten, dass wir sie umgedreht haben), also die selteneren 0-en zur 1 machen und die 1-en zur 0. Wir würden also nicht
Beispiel: Wenn wir die Wahrscheinlichkeit zu wählen damit vorhersagen wollten, dass jemand gebürtige:r Schweier:in ist und wie viele Minuten Nachrichten er:sie in der Woche konsumiert, wäre die Annahme sicher, dass gebürtige Schweizer:innen eher wählen gehen und der Umfang des Nachrichtenkonsums auch positiv mit der Wahrscheinlichkeit zu tun hat, dass jemand wählen gehen würde Variable WAHL. Wenn wir dann noch annehmen, dass der Umfang des Nachrichtenkonsums einen stärkeren Zusammenhang für Schweizerinnen hat als auf die Nicht-Schweizer:innen, dann bauen wir noch die Interaktion ein CH*Nachrichten_Dauer. Die Interaktionsvariable CH*Nachrichten_Dauer korelliert dann mit CH und Nachrichten_dauer. Wenn wir vorher Nachrichten_dauer zentrieren (also minus ihrem Mittelwert), dann korelliert CH*Nachrichten_Dauer nicht mehr mit CH, aber noch recht stark mit Nachrichten_dauer. Also kodieren wir um und machen aus der Dummy CH die Dummy Nicht_CH. Dann korelliert die Nicht_CH kaum noch mit der Nicht_CH*Nachrichten_Dauer. Wenn die Slope-Dummy dann signifikant negativ ist, würden wir sagen, dass der Zusammenhang zwischen dem Umfang der Nachrichtennutzung bei Nicht-Schweizer:innen geringer ist als bei gebürtigen Schweizer:innen.
Antwortsatz in der Klausur: Da die Slope-Dummy stark mit der Dummy korelliert, wenn die Metrische immer im positiven (oder negativen) Bereich liegt, gibt es hohe Multikollinearität, die dadurch verringert werden kann, dass die Metrische vorher zentriert wird. Überwiegen in der Dummy die 1-en deutlich, ist die Multikollinearität zwischen der Metrischen und der Slope-Dummy eventuell noch ein Problem. Dann sollte die Dummy umkodiert werden.

6.4 Regression (unzentriert)

```
Modell4 <- lm(Aggression ~ Vid_Games + Vid_Games * Anti_Soz_hoch, data = DATEN)

olsrr::ols_vif_tol(Modell4)
##                 Variables Tolerance      VIF
## 1               Vid_Games 0.8318770 1.202101
## 2           Anti_Soz_hoch 0.1049793 9.525692
## 3 Vid_Games:Anti_Soz_hoch 0.1024746 9.758512

sjPlot::tab_model(Modell4,
                  show.ci = FALSE,
                  show.std = TRUE, # zeige die standardisierten Koeffizienten
```

```
                  show.est = TRUE, # zeige die unstandardisierten estimates
                  show.r2 = TRUE # zeige R^2
                  )
```

	Agression			
Predictors	Estimates	std. Beta	p	std. p
(Intercept)	35.95	-0.00	**<0.001**	0.968
Video Games(Hours per week)	0.10	0.11	0.238	**0.008**
Anti Soz hoch	-3.28	0.37	0.501	**<0.001**
Vid_Games:Anti_Soz_hoch	0.77	0.15	**<0.001**	**<0.001**
Observations	442			
R^2 / R^2 adjusted	0.183 / 0.177			

6.5 Regression nach Zentrierung

```
DATEN |>
  summarize(Aggressions_Mittel = mean(Aggression, na.rm = TRUE), .by = Anti_Soz_hoch)
## # A tibble: 2 x 2
##   Anti_Soz_hoch Aggressions_Mittel
##           <dbl>              <dbl>
## 1             0               38.2
## 2             1               51.9

51.9 - 38.2
## [1] 13.7

# zentriere Vid_Games (Mittelwert = 0):
DATEN_z <- DATEN %>%
  mutate(Vid_Games = Vid_Games - mean(Vid_Games, na.rm = TRUE))

Modell4 <- lm(Aggression ~ Vid_Games +
                           Anti_Soz_hoch +
                           Vid_Games * Anti_Soz_hoch,
                           data = DATEN_z)

olsrr::ols_vif_tol(Modell4)
##                  Variables Tolerance      VIF
## 1               Vid_Games 0.8318770 1.202101
## 2           Anti_Soz_hoch 0.9993801 1.000620
## 3 Vid_Games:Anti_Soz_hoch 0.8314800 1.202675

sjPlot::tab_model(Modell4,
                  show.ci = FALSE,
                  show.std = TRUE, # zeige die standardisierten Koeffizienten
                  show.est = TRUE, # zeige die unstandardisierten estimates
                  show.r2 = TRUE, # zeige R^2
                  show.fstat = TRUE
                  )
```

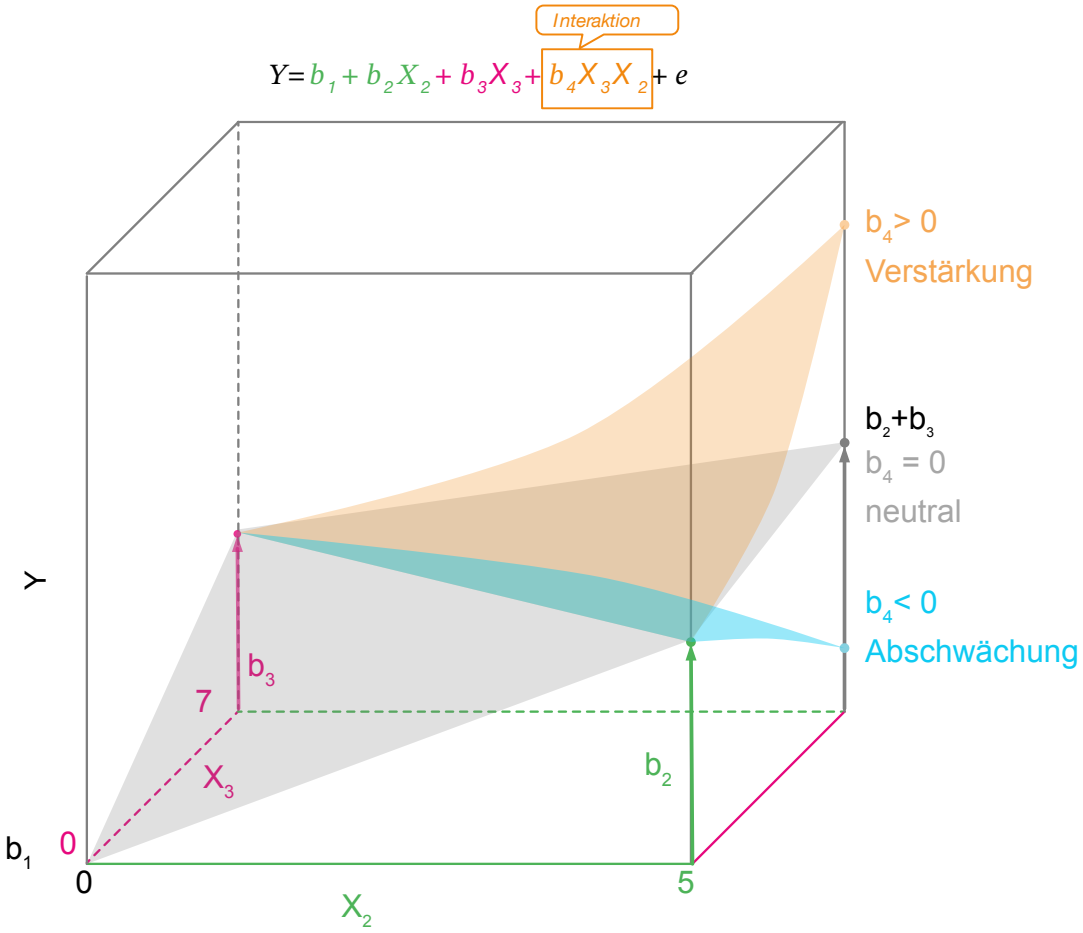
		Agression		
Predictors	Estimates	std. Beta	p	std. p
(Intercept)	38.17	-0.00	**<0.001**	0.968
Video Games(Hours per week)	0.10	0.11	0.238	**0.008**
Anti Soz hoch	13.52	0.37	**<0.001**	**<0.001**
Vid_Games:Anti_Soz_hoch	0.77	0.15	**<0.001**	**<0.001**
Observations	442			
R^2 / R^2 adjusted	0.183 / 0.177			

6.6 Kategoriale UV

```
Modell5 <- lm(Aggression ~ Vid_Games + Anti_Soz_hoch + Anti_Soz_mittel +
                Vid_Games * Anti_Soz_hoch + Vid_Games * Anti_Soz_mittel,
              data = DATEN_z)

olsrr::ols_vif_tol(Modell5)
##                    Variables Tolerance      VIF
## 1                  Vid_Games 0.2266636 4.411824
## 2              Anti_Soz_hoch 0.7390622 1.353066
## 3            Anti_Soz_mittel 0.7373241 1.356256
## 4   Vid_Games:Anti_Soz_hoch 0.5739909 1.742188
## 5 Vid_Games:Anti_Soz_mittel 0.2730809 3.661919

sjPlot::tab_model(Modell5,
                  show.std = TRUE, # zeige die standardisierten Koeffizienten
                  show.est = TRUE, # zeige die unstandardisierten estimates
                  show.ci = FALSE,
                  show.r2 = TRUE, # zeige R^2
                  show.fstat = TRUE,
                  string.est = "b",
                  string.std = "std. b"
                  )
```

		Agression		
Predictors	b	std. b	p	std. p
(Intercept)	30.79	-0.00	**<0.001**	0.958
Video Games(Hours per week)	0.05	0.10	0.766	**0.015**
Anti Soz hoch	20.90	0.57	**<0.001**	**<0.001**
Anti Soz mittel	10.29	0.40	**<0.001**	**<0.001**
Vid_Games:Anti_Soz_hoch	0.82	0.16	**<0.001**	**<0.001**
Vid_Games:Anti_Soz_mittel	0.03	0.01	0.865	0.865
Observations	442			
R^2 / R^2 adjusted	0.299 / 0.291			

7 Interaktion zweier metrischer Variablen



Interaktion zweier metrischer Variablen (in Worten)

```
# verändere alle numerischen Variablen, indem sie z-transformiert werden (scale)
DATEN_z <- DATEN  |>
  mutate(across(everything(), ~.x - mean(.x, na.rm = TRUE)))

Modell5 <- lm(Aggression ~ Vid_Games * CaUnTs,
              data = DATEN_z)

olsrr::ols_vif_tol(Modell5)
##          Variables Tolerance      VIF
## 1        Vid_Games 0.9930553 1.006993
## 2           CaUnTs 0.9974262 1.002580
## 3 Vid_Games:CaUnTs 0.9950094 1.005016

sjPlot::tab_model(Modell5,
                  show.std = TRUE, # zeige die standardisierten Koeffizienten
                  show.est = TRUE, # zeige die unstandardisierten estimates
                  show.ci = FALSE,
                  show.r2 = TRUE, # zeige R^2
                  show.fstat = TRUE,
                  string.est = "b",
                  string.std = "std. b"
                  )
```

		Agression	
Predictors	b	std. b	p
(Intercept)	-0.09	-0.01	0.854
Video Games(Hours per week)	0.17	0.09	**0.014**
Callous Unemotional Traits	0.76	0.58	**<0.001**
Vid_Games:CaUnTs	0.03	0.14	**<0.001**
Observations	442		
R^2 / R^2 adjusted	0.377 / 0.373		

onen zwischen metrischen Variablen zeigen an, inwiefern der Anstieg der einen UV mit dem grösser werden der anderen UV so:

Nutzung von Videogames hat einen signifikanten, aber sehr geringen Einfluss auf Aggression.
soziale Persönlichkeitsmerkmale korrelieren stark mit Aggression
öher die antisozialen Persönlichkeitsmerkmale, desto stärker wird der Zusammenhang zwischen der Nutzung von Video-Games Aggression
Je mehr Videospiele jemand spielt, desto grösser wird der Zusammenhang zwischen Antisozialen Merkmalen und Aggression.

8 Übung: GLM II

Hier schonmal die Uebung_02.qmd zum Download. Im Vorlesungspodcast habe ich die Datei etwas verändert. Das Resultat können Sie hier herunterladen

8.1 Vorlesungspodcast (online only)

In dem Podcast spreche ich die Übung_2 anhand der oben verlinkten Datei durch und verändere sie, bis sie so aussieht wie der zweite Downloadlink. Es geht auch viel um die R-Befehle und Tipps, wie man mit R gut arbeiten kann und wie Sie besser mit den Herausforderungen in der Arbeit mit R umgehen können. Der Input dauert etwas über 90 Minuten. Ich rate Ihnen, ihn sich eher in Etappen anzusehen und vielleicht erstmal den ersten Teil anzuschauen und dann vielleicht nochmal in die Aufgaben der Übung 2 und versuchen Sie, die noch übrigen Aufgaben zu lösen. Oder Sie schauen es sich nach und nach ganz an und machen danach Änderungen an der Übungsdatei und versuchen es nochmal, wenn Sie im ersten Durchgang Probleme hatten. Dabei lernen Sie mehr als nur mit der Übung allein. :-)

https://www.youtube.com/embed/QE11BYuTXaI

9 Faktorenanalysen

Auch als (explorative) Faktorenanalyse bekannt.

9.1 Multikollinearität und Dimensionsreduktion

Wenn zwei oder mehr Variablen stark miteinander zusammenhängen, ergibt sich eine hohe Multikollinearität. Bei zwei Variablen ergib sich also keine Punktwolke in einem zweidimensionalen Koordinatensystem, sondern eher eine Reihe Punkte die sehr nahe um eine (Regressions)Gerade liegen. Im Grunde kann man sie statt auf zweidimensionalen Ebene auf einer eindimensionalen Gerade darstellen. Wenn wir zB 5 UV's hätten, die alle sehr hoch miteinander korrelieren (die r's $> .7$), dann können wir diese fünf Dimensionen ohne grosse Verluste auf einer Dimension darstellen. Wir könnten einen Index bauen, der diese fünf UV's abbildet, ohne das wir viele Informationen verlieren würden (wir hoffen, dass wir nur Messrauschen «verlieren» und kaum substantielle Varianz).

9.2 Indices

Indices sind Variablen, die viele andere zusammenfassen. Das ist bei Aktienindices so und bei Mittelwert- oder Summenindices in der Statistik auch. Die Faktoren als Ergebnis einer explorativen Faktorenanalyse sind auch Indices, die die zugrundeliegenden (latenten) Variablen optimal abbilden, und zwar in dem Sinne, dass sie die Varianz aller Variablen optimal erfassen (also nicht alle gleichberechtigt, sondern die bevorzugt, die untereinander stark korrelieren).

9.3 Faktorenanalyse in Worten

In einem ersten Schritt wird eine Regressionsgerade so in die «Punktwolke» aller Dimensionen gelegt, dass diese Gerade alle Variablen möglichst gut abbildet, also ihre Varianz maximal abbildet. In der Regel bleibt dabei einige Varianz übrig, da die Punkte der Wolke nicht alle beziehungsweise mehrheiltich nicht auf der Geraden liegen. Diese ganze Varianz kann nun eine zweite Gerade haben, die so in die übrigen Punkte gelegt wird, dass sie nicht (steht senkrecht auf der ersten Geraden) oder kaum mit der ersten Gerade korreliert (Winkel gleich oder nahe 90°).

9.4 Explorative Faktorenanalyse (Principal Component Analysis – PCA)

Merkmale von Fällen wie Personen, Gruppen oder Inhalten usw. werden gemessen und als Variablen gespeichert. Die möglichen Merkmale einer zu messenden Population (GG) spannen einen Merkmalsraum mit vielen Merkmalsdimensionen auf. Das kann man sich tatsächlich auch räumlich vorstellen. Wenn zum Beispiel allen Befragten mehrere Fragen zu ihrer Medienzufriedenheit auf einer 5er-Skala gestellt werden, dann kann jeder bei jedem zu bewertenden Medium. Zum Beispiel könnte Axel bei der SRG1 eine 2 vergeben, weil er sie nicht so gut findet, Bernd eine 4 für eine bessere Bewertung usw. Bei der Bewertung von 3+ gibt Axel eine 4 und Bernd eine 3. Die Bewertungen für die SRG1 könnten wir auf eine X-Achse legen und die Bewertungen für 3+ auf eine Y-Achse. Damit hätten wir einen Merkmalsraum von zwei Dimensionen (da klingt es mit dem Raum noch komisch, wird aber einfach immer so genannt). Hätten wir noch ein drittes Medium, wie zum Beispiel RTL+, dann könnte man das auf eine Z-Achse packen, die dann die

dritte Dimension wäre und damit auch im alltagssprachlichen ein schöner Merkmalsraum mit drei Merkmalsdimensionen. Allerdings gibt es da noch TV-Ostschweiz, tele bärn, 4+, Pro7, ARD, SRG2 usw. Das heisst, wir haben in der Normalität einen viel grösseren Merkmalsraum mit etlichen Merkmalsdimensionen. Nun kann es interessant sein, ob hinter den unterschiedlichen Bewertungen der Medien unterschiedliche zugrundeliegende Vorlieben stecken. Es könnte doch sein, dass viele Leute aus einem bildungsbürgerlichen Anspruch heraus eher Arte, 3SAT, öffentlich rechtlichen wie die Sender der SRG oder ARD und ZDF besser bewertet als die privaten kommerziellen bzw. regionalen Sender. Andere finden vielleicht generell die öffentlich rechtlichen Sender blöd, weil sie ihnen Staatsnähe unterstellen oder sie Gebühren zahlen müssen. Es könnte also zugrundeliegende bzw. latente Merkmale geben, die zu den Einzelbewertungen auf den Dimensionen führen. Wenn dem so ist, dann müssten die Variablen, die zu einer latenten Dimension gehören, stark miteinander korrelieren, also z.B. alle Bewertungen zu den öffentlich-rechtlichen Sendern. Das bedeutet, wir könnten diese Bewertungen auch auf diese latente Dimension reduzieren. Zack fertig: Dimensionsreduktion!

Faktorenanalysen dienen genau dieser Dimensionsreduktion. Sie werden eingesetzt, um latente Konstrukte zu identifizieren, die die Ausprägungen der gemessenen (manifesten) Variablen bestimmen bzw. determinieren. Wir setzen die Faktorenanalyse aber auch ein, um z.B. das Problem der Multikolliniearität bei Regressionsanalysen in den Griff zu kriegen. Das Ziel der Faktorenanalyse ist es daher, eien Vielzahl an Variablen auf wenige zugrundeliegende Faktoren zu reduzieren, die man dann gut und klar unterscheiden kann, die also nicht oder nur wenig miteinander korrelieren.

9.5 Ablauf einer Faktorenanalyse

Die Schritte der Faktorenanalyse sind: 1. Voranalyse über Korrelationstabellen (Ausschluss von Variablen, die mit keiner anderen korrelieren) 2. Extraktion der Faktoren 3. Bestimmung der Anzahl Faktoren (Scree Plots der Eigenwerte) 4. Rotation der Faktoren (bessere Verteilung der Varianzaufklärung) 5. Eignung für die Variablen über Kommunalitäten (ggf. Ausschluss gering abgebildeter Variablen mit Kommunalitäten $<$ .4) 6. Interpretation und Benennung der Faktoren 7. Speichern der Faktoren (eine Form der Indizes) für weitere Analysen, wie Regressionen

Um eine explorative Faktorenanalyse durchzuführen, müssen:

1. die Variablen ausgewählt werden, für die Faktoren extrahiert werden sollen.
2. Vor der eigentlichen Faktorenanalyse sollte man sich die Korrelationsmatrix anschauen, sowie
3. KMO und Bartlett-Test durchlaufen lassen, um festzustellen, ob die Variablen für eine FA geeignet sind.
4. Im Anschluss wählt man die Extraktionsmethode und Rotationsmethode aus, die verwendet werden sollen.
5. Dann werden die Faktoren extrahiert und man erstellt einen Screeplot, um anhand des Ellenbogenkriteriums zu bestimmen, wie viele Faktoren eine gute Faktorlösung ergeben. Ein gängiges Kriterium ist, dass die Eigenwerte der Faktoren grösser als 1 sein sollen.
6. Die Rotation der Faktoren ist nicht zwingend, aber so üblich und empfohlen, dass sie im Grunde dazugehört. Die Faktoren können orthogonal (unkorreliert) oder oblique (leichte Korrelationen zugelassen) rotiert werden.
7. Dann kann man anhand der Faktorladungen jeder Variable die Faktoren interpretieren. Dabei charakterisieren hohe Faktorladungen die Faktoren.
8. Haben einzelne Variablen kleine Kommunalitäten ($<$.4) und eine hohe Uniqueness ($>$.6), entfernt man diese Variablen aus den Faktorenanalysemodell und führt die Schritte 2. bis 6. nochmals ohne die entsprechenden Variablen aus.
9. Am Ende schaut man sich noch die Gesamtvarianzaufklärung der Faktorlösung an, die Auskunft darüber gibt, wie gut die Faktoren insgesamt die Variablen abbilden.
10. Will man die Faktorwerte als Indizes weiterverarbeiten, speichert man sie am Ende in seinen Daten ab.

9.6 The R Anxiety

Als Beispiel wird hierfür die R-Angst-Skala von Andy Field [@Field2022] verwendet. Die Fragen und zugehörigen Variablen sind:

1. **raq_01**: *Statistics make me cry*
2. **raq_02**: *My friends will think I'm stupid for not being able to cope with R*
3. **raq_03**: *Standard deviations excite me*
4. **raq_04**: *I dream that Pearson is attacking me with correlation coefficients*
5. **raq_05**: *I don't understand statistics*
6. **raq_06**: *I have little experience of computers*
7. **raq_07**: *All computers hate me*
8. **raq_08**: *I have never been good at mathematics*
9. **raq_09**: *My friends are better at statistics than me*
10. **raq_10**: *Computers are useful only for playing games*
11. **raq_11**: *I did badly at mathematics at school*
12. **raq_12**: *People try to tell you that R makes statistics easier to understand but it doesn't*
13. **raq_13**: *I worry that I will cause irreparable damage because of my incompetence with computers*
14. **raq_14**: *Computers have minds of their own and deliberately go wrong whenever I use them*
15. **raq_15**: *Computers are out to get me*
16. **raq_16**: *I weep openly at the mention of central tendency*
17. **raq_17**: *I slip into a coma whenever I see an equation*
18. **raq_18**: *R always crashes when I try to use it*
19. **raq_19**: *Everybody looks at me when I use R*
20. **raq_20**: *I can't sleep for thoughts of eigenvectors*
21. **raq_21**: *I wake up under my duvet thinking that I am trapped under a normal distribution*
22. **raq_22**: *My friends are better at R than I am*
23. **raq_23**: *If I am good at statistics people will think I am a nerd*

9.7 Korrelationsmatrix

Die Korrelationsmatrix ist die Basis für Faktorenanalysen (im Grunde braucht man nur die Korrelationsmatrix (+ Fallzahl) und die ursprünglichen Daten nicht). Mit dem folgenden Befehlen kann man sich die Korrelationsmatrix rausgeben lassen.

```
# Lade den Datensatz "raq.csv" aus dem Ordner discovr_csv, den man hier
# herunterladen kann: https://www.discovr.rocks/csv/discovr_csv.zip
raq.tib <- readr::read_csv("data/discovr_csv/raq.csv")
## Rows: 2571 Columns: 24
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): id
## dbl (23): raq_01, raq_02, raq_03, raq_04, raq_05, raq_06, raq_07, raq_08, raq_09, raq_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

# Lösche die Variable "id", die ganz vorne im Datensatz steht
raq_items_tib <- raq.tib |>
  select(-id)

# Berechne die Korrelationen für alle Variablen mit allen Variablen (items)
raq_cor <- raq_items_tib |>
```

```
  cor()

# Gebe einen Korrelationsplot mit dem Paket "psych" raus.
#psych::corPlot(raq_cor, upper = FALSE)

# Sortiere die Variablen danach, wie stark sie miteinander korrelieren
order.FPC <- corrplot::corrMatOrder(raq_cor, order = 'FPC')
order.hc <- corrplot::corrMatOrder(raq_cor, order = 'hclust')

# Speichere die Ordnung als Matrix
raq_cor.FPC <- raq_cor[order.hc, order.hc]

# Schönere Korrelationsplots gibt es mit dem Paket "corrplot" und dem Befehl "corrplot"
corrplot::corrplot(raq_cor.FPC, tl.col='black', tl.cex=.75)
```

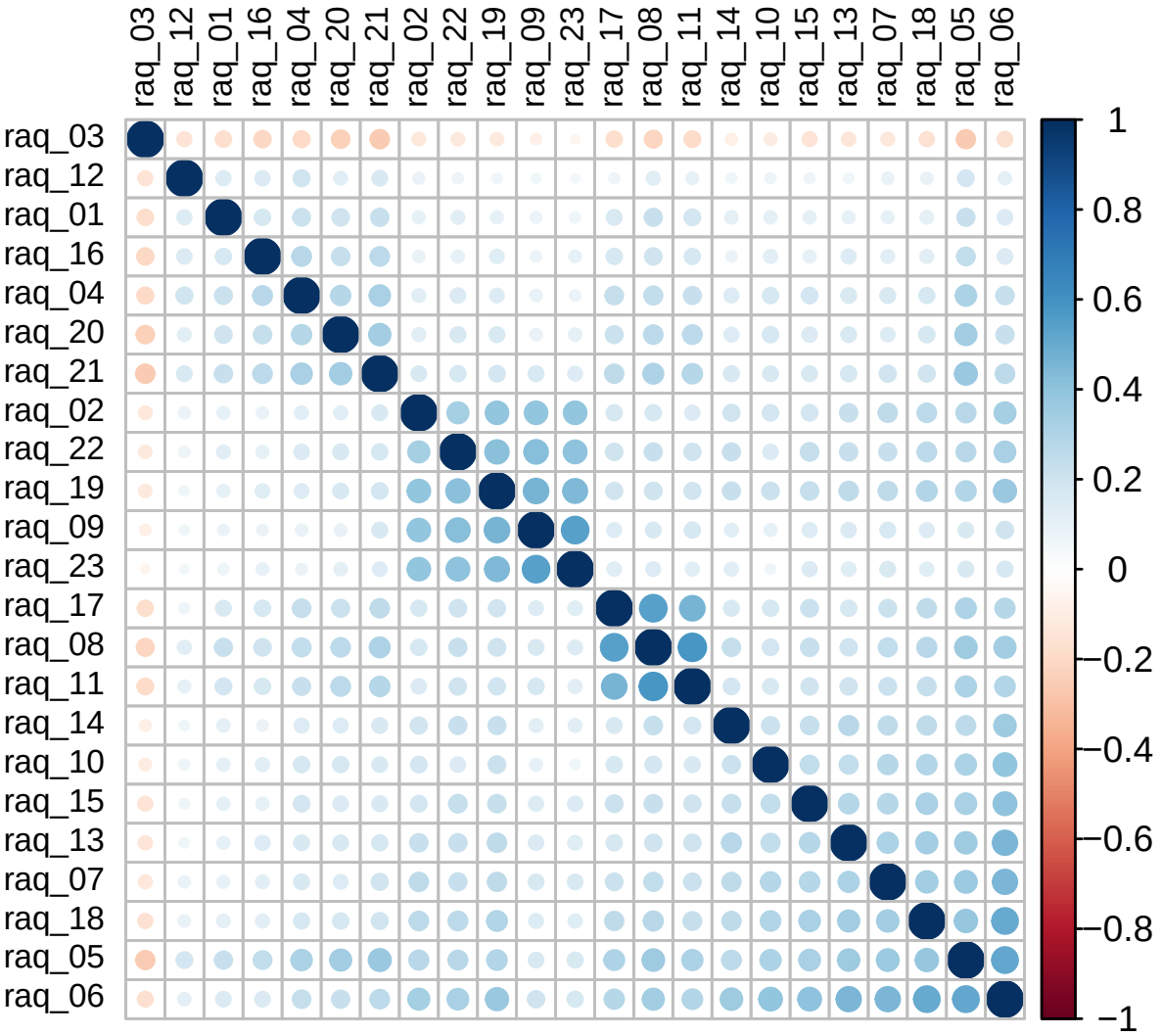


Abbildung 9.1: Korrelationsplot

Man sieht hier schon, dass die Korrelationen nicht wahnsinnig gross sind, aber sich wie Haufen bilden. Die Variablengruppen, die untereinander hoch korrelieren, gehen vermutlich auf ein gemeinsames latentes Konstrukt zurück. Diese latenten Konstrukte werden im Folgenden auch als Faktoren bezeichnet.

9.8 Anzahl Faktoren bestimmen

Wenn wir eine Faktorenlösung suchen, müssen wir erstmal die Anzahl sinnvoller Faktoren bestimmen. Das geht mit dem «psych»-Paket und der Analyse «fa.parallel». Dort werden die Eigenwerte (eigen values) der Faktoren angezeigt. Die Eigenwerte sind der Anteil der Varianzaufklärung eines Faktors relativ zur Anzahl der Variablen in der Faktorenanalyse. Wenn also ein Eigenvalue bei 1 ist, erklärt ein Faktor so viel wie eine einzelne Variable.

```
## Parallel analysis suggests that the number of factors =  4  and the number of components =  NA
```

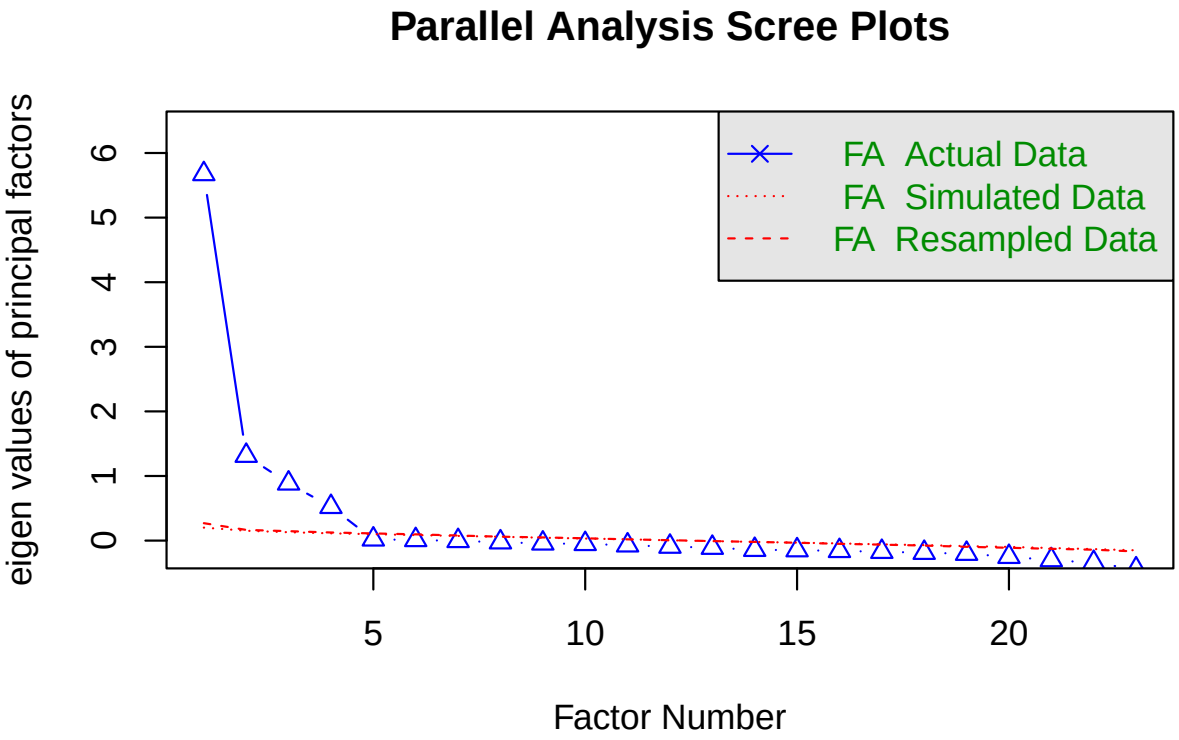


Abbildung 9.2: Analyse zur Bestimmung der Faktoren (über roter Linie)

Im Plot kann man sehen, dass der erste Faktor einen Eigenwert von knapp 6 hat, also so viel Varianz aufnimmt, wie sechs Variablen im Ursprung. Der zweite Faktor ist noch über 1. Das bedeutet, er erklärt etwas mehr als eine Ursprungsvariable. Der dritte und der vierte Faktor erklären etwas weniger als eine Variable. Da das alte Kaiser-Kriterium (Eigenwerte müssen über 1 sein) etwas sehr holzschnittartig ist, haben sich findige Statistiker ausgedacht, dass man die FA simulieren könnte, unter der Annahme, dass die Faktoren nichts erklären. Diese Simulation durch mehrfaches ziehen von Stichproben aus den Daten (FA Resampled Data) ergibt, dass 4 Faktoren mehr besser sind als die informationslose Simulation. Also ist die Faktorlösung 4.

9.9 Faktorladungen und Uniqueness

Mit dieser Analyse können wir jetzt die Faktorenanalyse rechnen. Als Anzahl «n» der Faktoren (nfactors) geben wir die 4 aus der Analyse von oben ein (siehe Abbildung @ref(fig-Parallelanalyse)auf Seite ??).

Die Uniqueness ist der Varianzanteil, den eine Variable ganz alleine hat, also nicht mit den anderen teilt. Die Uniqueness ist das Gegenteil von Kommunalität (Communality), also der gemeinsamen Varianz mit der Faktorenlösung. Rechnerisch ergibt sich je Variable die Uniqueness aus 1 - Kommunalität. Hohe Uniqueness bedeutet, dass eine Variable nicht gut in die Faktorenanalyse passt, weil sie eben nicht gut durch die Faktoren abgebildet wird, sondern einzigartig (unique) ist. Für die Variable selbst und ggf. für ihre Integration in ein Modell ist eine hohe Uniqueness gut, da sie auch bedeutet, dass es keine Probleme mit Multikollinearität gibt. Die Variable kann also getrost aus der Faktorenanalyse entfernt und als eigenständige Variable in eine Modell aufgenommen werden.

Die Complexity gibt an, wie viele Faktoren gebraucht werden, um die Variable abzubilden. Wenn sie 1 ist, dann wird eine Variable von einem Faktor abgebildet. Ist sie zum Beispiel 1.97 braucht es zwei Faktoren, um die Variable darzustellen. Geringe Komplexität ist in dem Fall gut, da sie zu einer klaren Faktorenlösung führt.

```
## Loading required namespace: GPArotation
## # Rotated loadings from Factor Analysis (oblimin-rotation)
##
## Variable | MR1   | MR2   |  MR4   | MR3   | Complexity | Uniqueness
## ----------------------------------------------------------------
## raq_06   | 0.84 |       |        |       |    1.00    |    0.27
## raq_18   | 0.63 |       |        |       |    1.03    |    0.57
## raq_13   | 0.57 |       |        |       |    1.02    |    0.66
## raq_07   | 0.56 |       |        |       |    1.02    |    0.65
## raq_10   | 0.49 |       |        |       |    1.08    |    0.74
## raq_15   | 0.48 |       |        |       |    1.04    |    0.71
## raq_05   | 0.45 |       | 0.39   |       |    1.97    |    0.46
## raq_14   | 0.42 |       |        |       |    1.06    |    0.78
## raq_09   |       | 0.81 |        |       |    1.02    |    0.38
## raq_23   |       | 0.79 |        |       |    1.02    |    0.41
## raq_19   | 0.26 | 0.56 |        |       |    1.41    |    0.50
## raq_22   |       | 0.52 |        |       |    1.29    |    0.59
## raq_02   | 0.25 | 0.48 |        |       |    1.54    |    0.62
## raq_21   |       |       | 0.59   |       |    1.04    |    0.60
## raq_04   |       |       | 0.56   |       |    1.01    |    0.67
## raq_20   |       |       | 0.54   |       |    1.02    |    0.68
## raq_16   |       |       | 0.51   |       |    1.02    |    0.77
## raq_03   |       |       | -0.43 |       |    1.01    |    0.80
## raq_01   |       |       | 0.39   |       |    1.06    |    0.83
## raq_12   |       |       | 0.37   |       |    1.07    |    0.89
## raq_08   |       |       |        | 0.88 |    1.00    |    0.25
## raq_11   |       |       |        | 0.72 |    1.00    |    0.45
## raq_17   |       |       |        | 0.68 |    1.00    |    0.51
##
## The 4 latent factors (oblimin rotation) accounted for 40.12% of the total variance of the original
```

9.10 Interpretation der Faktorenanalyse

Mit dieser Faktorenlösung können wir jetzt die Faktoren interpretieren.

Der erste Faktor lädt hoch auf folgenden Items. Wir können diesen Faktor als / «Probleme mit Computern»** labeln:

- **raq_05**: *I don't understand statistics*
- **raq_06**: *I have little experience of computers*
- **raq_07**: *All computers hate me*
- **raq_10**: *Computers are useful only for playing games*
- **raq_13**: *I worry that I will cause irreparable damage because of my incompetence with computers*
- **raq_14**: *Computers have minds of their own and deliberately go wrong whenever I use them*
- **raq_15**: *Computers are out to get me*
- **raq_18**: *R always crashes when I try to use it*

Beachte: Das Item **«raq_05»** lädt auch hoch auf dem zweiten Faktor MR2.

Wenn man die Fragen anschaut, die hoch auf dem zweiten Faktor MR2 laden, deuten darauf hin, dass es die Befragten Angst haben, von ihren Peers komisch angesehen zu werden. Nennen wir diesen Faktor **«Angst vor sozialer Bewertung»**:

- **raq_02**: *My friends will think I'm stupid for not being able to cope with R*
- **raq_09**: *My friends are better at statistics than me*
- **raq_19**: *Everybody looks at me when I use R*
- **raq_22**: *My friends are better at R than I am*
- **raq_23**: *If I am good at statistics people will think I am a nerd*

Beim Faktor MR3 wird deutlich, dass es hier eine Angst vor Statistik gibt. Nennen wir den Faktor **«Angst vor Stastik»**:

- **raq_01**: *Statistics make me cry*
- **raq_03**: *Standard deviations excite me*
- **raq_04**: *I dream that Pearson is attacking me with correlation coefficients*
- **raq_05**: *I don't understand statistics*
- **raq_12**: *People try to tell you that R makes statistics easier to understand but it doesn't*
- **raq_16**: *I weep openly at the mention of central tendency*
- **raq_20**: *I can't sleep for thoughts of eigenvectors*
- **raq_21**: *I wake up under my duvet thinking that I am trapped under a normal distribution*

Bei den übrigen Fragen, die auf dem Faktor MR4 laden, geht es eher um Mathematik. Wir könnten also sagen der Faktor MR4 ist **«Angst vor Mathe»**.

- **raq_08**: *I have never been good at mathematics*
- **raq_11**: *I did badly at mathematics at school*
- **raq_17**: *I slip into a coma whenever I see an equation*

Wir können noch schauen, ob die Variablen mit Doppelladungen plausibel sind. Also schauen wir zum Beispiel auf das Item **raq_05** «*I don't understand statistics*». Das scheint mit einer geringen Selbstwirksamkeit in Bezug auf Computer und Statistik zusammenzuhängen. Es spiegelt die Selbsteinschätzung wieder, dass man Statistik und Computer «nicht kann».

9.11 Faktorendiagramm

Faktorenanalysen kann man mit solchen Diagrammen darstellen. Hier sieht man auch, wie stark die einzelnen Faktoren miteinander korrelieren, wenn man die Faktoren nicht gezwungen hat, orthogonal zu sein, also unkorrliert.

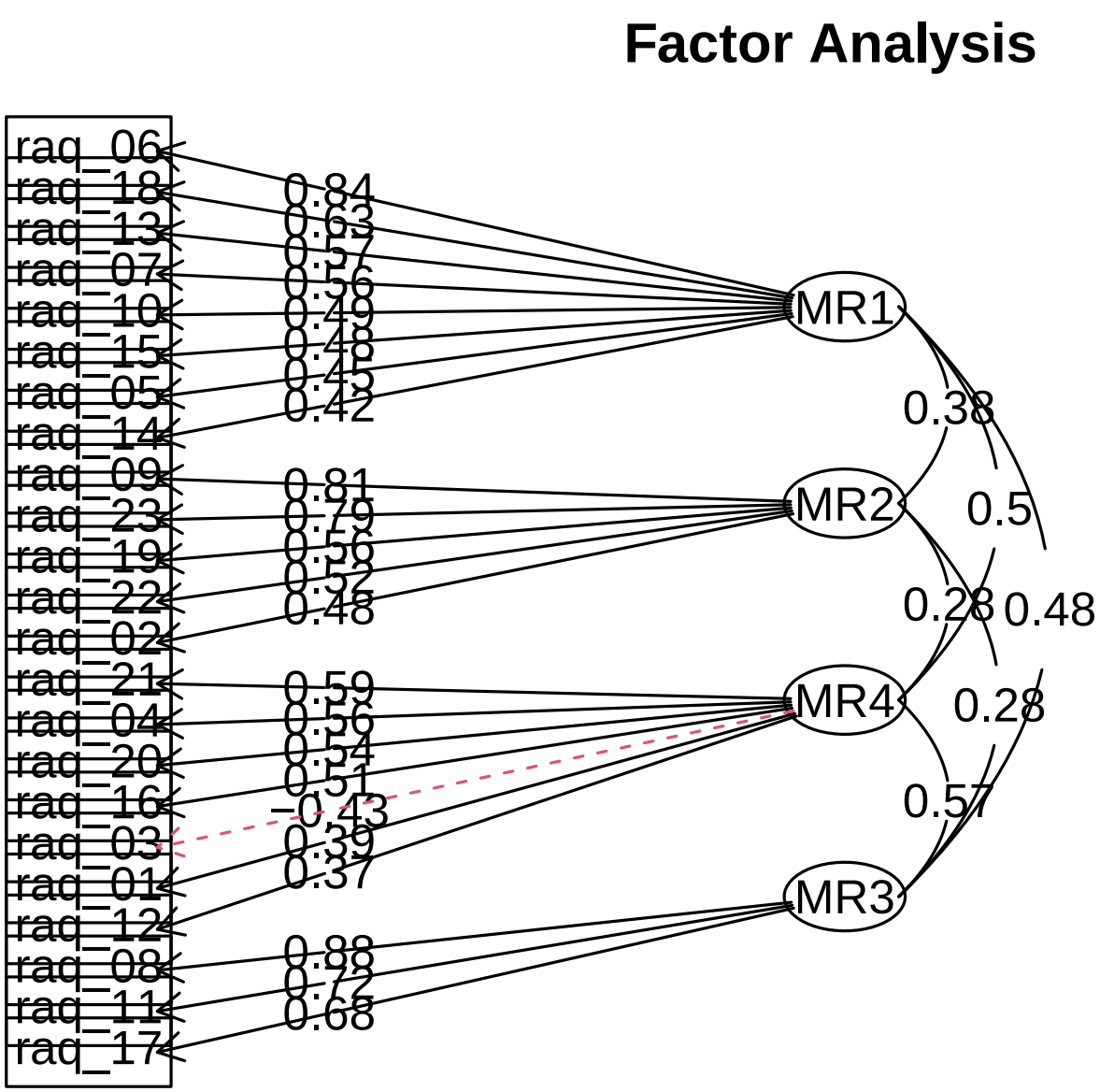


Abbildung 9.3: Faktorendiagramm

```
## Saving 6 x 4 in image
```

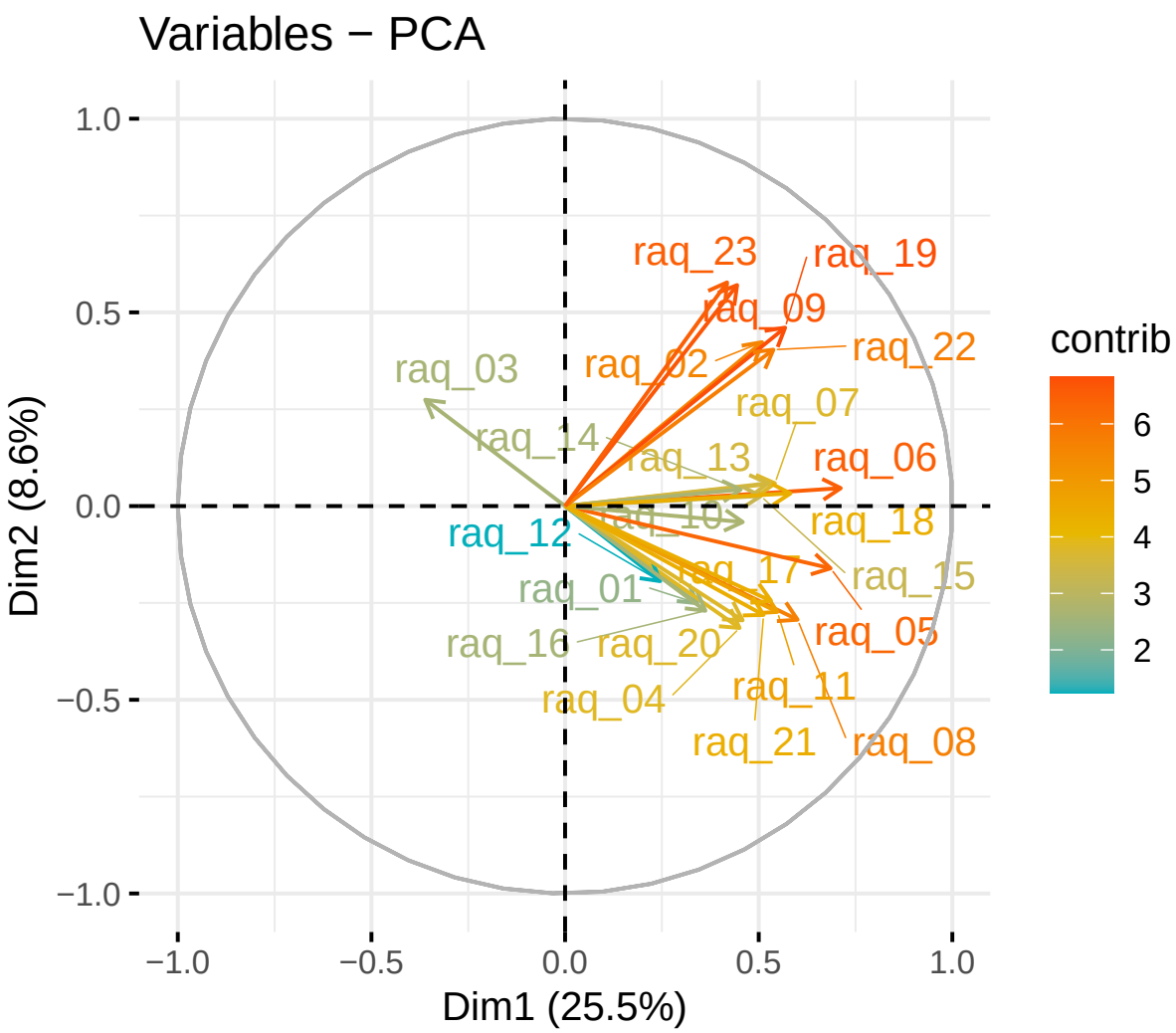


Abbildung 9.4: Variablen PCA

10 Übung: Dimensionsreduktion

11 Machine Learning und logistische Regression

11.1 Learn from Desaster (Titanic)

In diesem Kapitel beschäftigen wir uns mit Daten des Untergangs der Titanic.

11.1.1 Daten einlesen

Hier können Sie den Datensatz und die Beschreibung der Daten finden. Suchen und speichern Sie den Datensatz «train.csv».

Downlaod der Daten: https://www.kaggle.com/competitions/titanic/data

```
## Loading required package: viridisLite
```

Die PassengerId ist einefach eine Identifikationsnummer. Es gibt dann eine Variable, die «Survived» heisst, die ein Minimum von 0 hat und ein Maximum von 1. Das deutet sehr auf eine Dummy hin. Da der Durchschnitt («Mean») = 0.38 ist, wissen wir jetzt schon, dass 38 Prozent der Passagiere überlebt haben (der Mittelwert einer Dummy ist immer der Prozentsatz der 1er-Gruppe). Dann kommt noch der Name als Zeichenvariable, das Alter, das von 0.42 bis 80 geht. Von 177 Personen fehlen die Altersangaben. Bei den übrigen Variablen muss man nochmal nachschauen auf «kaggle». Dort steht:

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

11.1.2 Daten in Trainings- und Testdaten aufteilen

Wenn von Machine Learning (ML) die Rede ist, dann wird (wenn es um supervised learning geht) zunächst ein Modell an Trainingsdaten trainiert bzw. angelernt und später an Testdaten getestet. Darum nehmen wir den vorliegenden Datensatz mal auseinander und teilen die Fälle (Passagiere) zufällig den Trainingsdaten zu und später zu verwendenden Testdaten. In der Regel wird der Trainingsdatensatz grösser gewählt: Ich habe ihn auf 75% des Ursprungsdatensatzes festgelegt. Der Rest (#anti_join) wird für später als Testdatensatz aufbewahrt.

```
# Setze eine Zufallszahl, damit die Ergebnisse replizierbar sind, also nicht jedes Mal
# eine neue Zufallszahl gesetzt wird und die Ergebnisse (bisschen) abweichen
set.seed(12345)

# Ziehe eine Zufallsstichprobe aus dem Filmdatensatz und bezeichne ihn als "train", also
# Trainingsdatensatz, mit dem die "Maschine" trainiert wird
train <- DATEN_titanic |>
  sample_frac(.75)

# Bilde aus dem Rest der nicht für "train" gezogenen Fälle einen Test-Datensatz, indem nach
# 'id' die Fälle aus "Filme" das Gegenteil (anti) von zusammengetan (join) werden.
test  <- anti_join(DATEN_titanic, train, by = 'PassengerId')
```

Sehen kann man nach diesem r-Chunk übrigens nichts, weil nur Datensätze im Hintergrund aufgeteilt wurden. Also suchen wir mal nach guten Datenvisualisierungen.

11.1.3 Datenvisualisierung

Eine einfache Darstellungsmöglichkeit ist ein sogenannter «Scatterplot» der die Lage von Fällen in ein Koordinatensystem einteilt, das durch zwei Variablen gebildet wird. Im Beispiel ist es das Alter der Passagiere «Age» und der Fahrpreis «Fare». Als Zweites haben wir ein Balkendiagramm für die Passagierklassen. Im letzten sehr aufwendigen Plotgrafik werden die Zweierbeziehungen aller Variablen dargestellt, also wie sie miteinander korrelieren (obere Nebendiagonalen), wie ihre Verteilung ist (auf der Diagonale mit Namen) und wie ihre gemeinsame Streuung ist, also ein Scatterplot in der unteren Nebendiagonalen. Mehr zu diesen SPLOM finden Sie hier: https://cran.r-project.org/web/packages/psych/vignettes/intro.pdf.

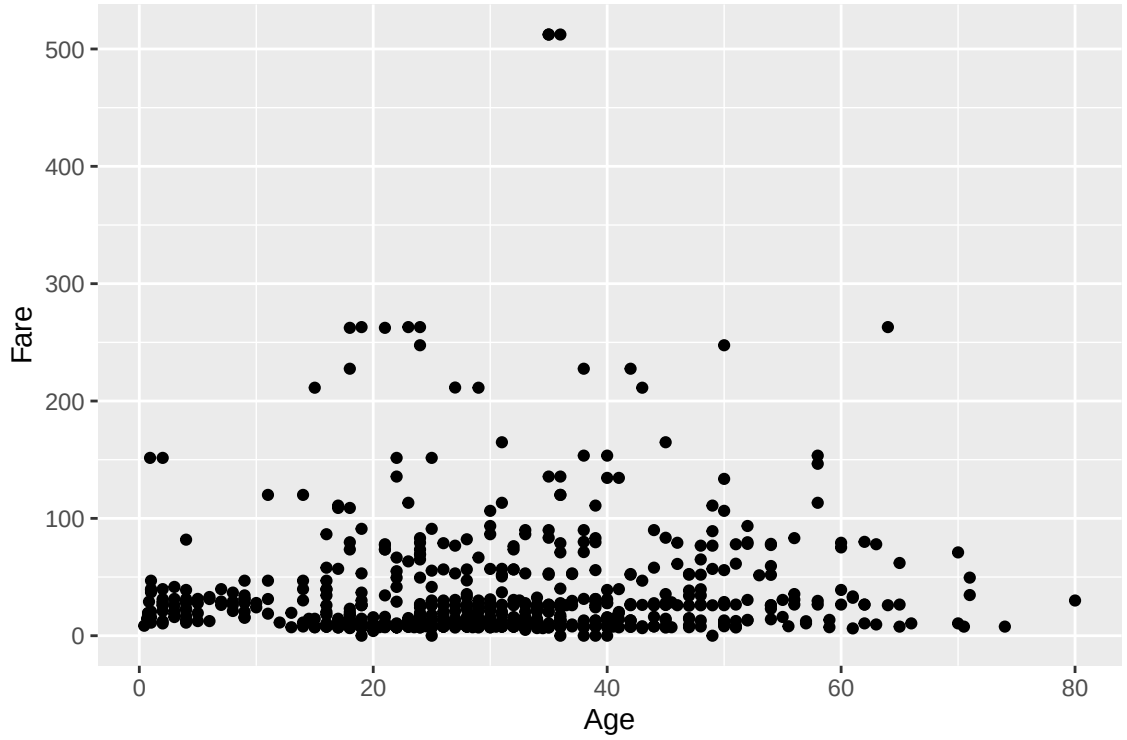


Abbildung 11.1

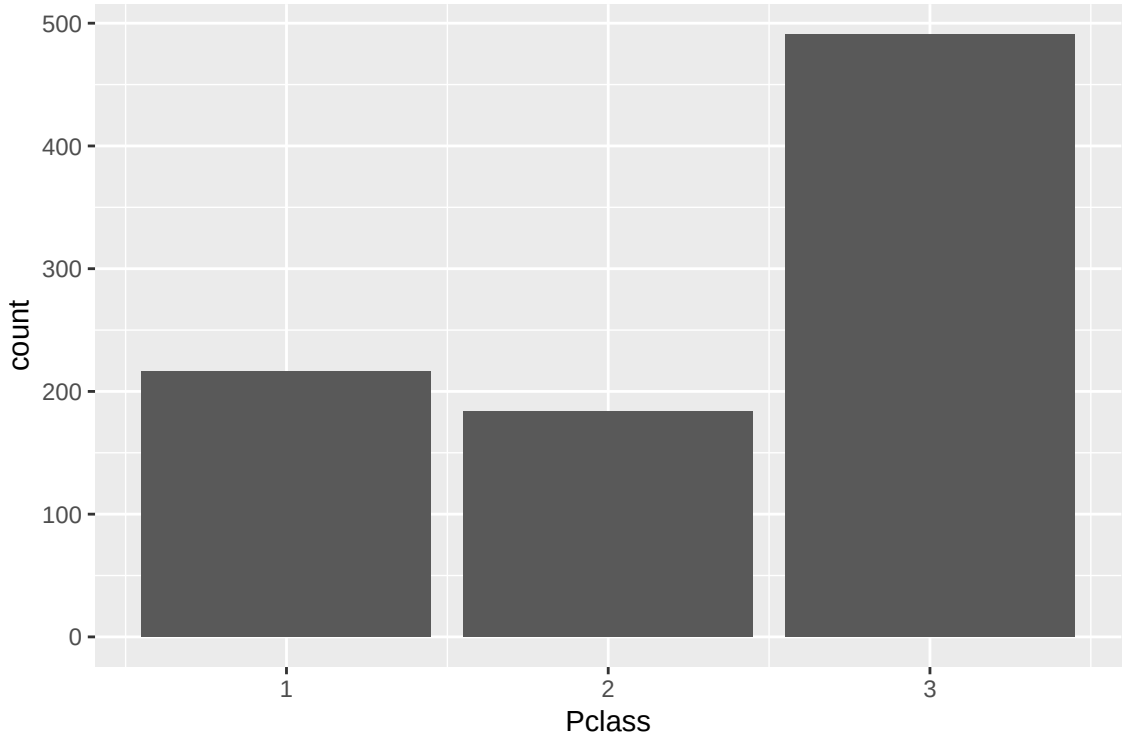


Abbildung 11.2

11.2 Modellbildung für den Fahrpreis

Jetzt kommt das erste Modell. Wenn Sie mit diesem Syntax experimentieren, dann kopieren Sie sich mal die Zeile für das Modell, löschen aus von «Age_z» … bis «Kinder» heraus und schätzen Sie mal das mit den summarys. Schauen Sie sich die gut an und achten Sie darauf, was passiert, wenn Sie die Summanden für I(Age_z^2) usw. wieder in das Modell tun. Am Ende können Sie versuchen das Modell durch weitere Variablen ergänzen und verbessern oder andere Teile wieder herausnehmen. [1].

```
fit_titanic <- lm(log(Fare + 1) ~
                                    Sex + Age_z +
                                    I(Age_z^2) +
                                    Survived + Pclass_f +
                                    Kinder,
                                     data = train)

# summary(fit_titanic)
```

In der Summary des fits_titanic sehen wir zunächst ganz oben die Formel. Da der Preis schnell hoch ging, wird für den Preis mit «log» der natürliche Logarithmus gebildet (Es wird Fare + 1 gerechnet, weil der log für 0 nicht definiert ist und für einige Passagiere angegeben ist, dass ihr Fahrpreis 0 war.) Das Geschlecht ist eine Dummyvariable mit 1 für männlich. Das Alter ist eine zentrierte Variante der Variable Age. Die Zentrierung machen wir, weil die resultierende zentrierte Variable nicht mehr sehr stark mit ihrer quadrierten Version. Den Befehl für die Zentrierung finden Sie in der «Datenaufbereitung.Rmd». Dann folgt der etwas komische Ausdruck für die quadrierte Version der Altersvariable «I(Age_z^2)». Die Quadrierung

[1] Manche Variablen wurden erst noch erstellt (zB «Kinder» oder «Age_z»). Die entsprechende Datenaufbereitung.Rmd können Sie hier schnell

machen wir, weil die Beziehung zwischen Alter und Fahrpreis vermutlich nicht linear ist, sondern kurvilinear quadratisch (Alter hat oft einen quadratischen Einfluss, weil in der Regel Ältere und Jüngere etwas weniger Geld haben als die mittlere Altersgruppe). Wenn nur das Alter als quadratische Funktion in der Gleichung wäre, dann müsste die quadratische Funktion einer um 0 zentrierten Variable immer wie eine summetrische Schüssel aussehen, die um 0 liegt. In der Regel ist die Schüssel aber gekippt. Dafür wird noch gebraucht, dass die Altersvariable «Age_z» auch noch in ihrer nichttransformierten Form Teil der Gleichung ist.

Wir lesen nun in der Tabelle die b, die in der Spalte «Estimate» stehen. Der Intercept hat ein b_1 von 2.66 und keiner weiss, wie man das interpretieren soll. Ist aber auch nicht wichtig. Dann kommt schon die Variable «Sex» für das Geschlecht mit 1 für «männlich». Wir können hier sehen, dass männliche Mitreisende einen tieferen Fahrpreis gezahlt haben und zwar das auch signifikant, da der Pr($>$|t|) bzw. einfach der p-Wert kleiner ist als .05 (Die «wissenschaftliche» Schreibweise ist etwas mühsam zu lesen. Der p-Wert für Sexmale ist 0.00000763). Der Einfluss des linearen Alters ist klein, negativ und nicht signifikant. In der Stichprobe findet es sich also, dass die Regressionsgerade für das Alter leicht nach unten schräg ist. Die «Schüssel», die vom nachfolgenden quadrierten Alter gebildet wird, neigt also leicht nach rechts. Da das quadrierte Alter auch negativ ist, sieht es aus als wäre die Beziehung zwischen Alter und Fahrpreis ein umgekehrter Bogen einer quadratischen Funktion, die sehr flach ist und rechts ein bischen stärker nach unten gebogen als links. Allerdings ist auch der Einfluss des quadrierten Alters nicht signifikant. Probieren Sie es mal ohne das quadrierte Alter, was dann passiert.

Dann folgt die Variable «Survived», die anscheinend angibt, dass Passagiere, die überlebt haben, etwas weniger zahlen mussten. Das ist intuitiv und theoretisch natürlich Quatsch, da hier die Kausalität zeitlich auf den Kopf gestellt wird. Die Varialbe müsste aus einem seriösen Modell wieder raus. Gut also, dass Sie das hier gelesen haben.

Dann kommen noch zwei Varialben für die Passagierklasse «Pclass». Hier hat R automatisch die Variable Pclass, die drei Ausprägungengen hatte in drei Dummys aufgeteilt, wobei die Ausprägung die im Level des Faktors an erster Stelle steht, als Referenzkategorie genommen wird, was im Titanicmodell die Klasse 3 ist. Wäre die auch im Modell, hätten wir perfekte Multikollinearität, weil wir immer schon wüssten, in welcher Klasse jemand eingecheckt sein musste, wenn es nicht die «first class» oder «second class» war. Wir sehen nun, dass die Passagiere der zweiten Klasse mit einem b von 0.49 einen signifikant höheren Preis bezahlt haben als die Personen der 3. Klasse. Der Unterschied ist vergleichsweise stark, was man daran erkennen kann, dass der t-Wert vergleichsweise hoch ist bei 6.859. Allerdings ist der t-Wert für den Unterschied zwischen 3. Klasse und 1. Klasse noch viel grösser. Für die «first class» musste also noch deutlich mehr gezahlt werden. Interessanterweise haben auch Kinder etwas mehr bezahlt als Erwachsene, obwohl die Zugehörigkeit der Klasse schon rausgerechnet ist. Offenbar waren Kinder auf dem Schiff eher besser untergebracht als der Schnitt oder wie würden Sie sich das erklären?

Der Summary-Befehl gibt keinen sehr guten Output raus und lässt sich auch nicht gut anpassen. Für einen Abdruck in einem Forschungsbericht eignet sich das noch weniger. Daher versuchen wir hier eine schönere Ausgaben hinzubekommen

11.2.1 Regressionsoutput

Mit der Funktion von `sjPlot::tab_model` können wir die standardisierten Regressionkoeffizienten rausgeben lassen (und sie auch als std. b bezeichnen). Zudem werden uns ein Konfidenzintervall «CI» für die b rausgegeben und eine «standardized CI» für die standardisierten Regressionskoeffizienten. Speziell ist, dass einfach p rausgegeben werden und standardisierte p. R gibt dazu als Hinweis, dass die p-Werte hier unteschiedlich sein können, weil in der Formel logarithmen und quadrierte Beziehungen drin sind. Im Zweifel interpretieren Sie besser die einfachen p und nicht die standardisierten «std. p».

	log(Fare+1)					
Predictors	b	std. b	CI	standardized CI	p	std. p
(Intercept)	2.66	0.83	2.51 – 2.81	0.80 – 0.87	**<0.001**	**<0.001**
Sex [male]	-0.32	-0.06	-0.45 – -0.18	-0.10 – -0.02	**<0.001**	**0.003**
Age z	-0.00	-0.02	-0.01 – 0.00	-0.05 – 0.00	0.314	0.063
Age z^2	-0.00	-0.00	-0.00 – 0.00	-0.02 – 0.01	0.685	0.639

Survived	-0.08	-0.00	-0.22 – 0.07	-0.02 – 0.02	0.294	0.914
Pclass f [2]	0.49	0.06	0.35 – 0.63	0.02 – 0.10	**<0.001**	**0.004**
Pclass f [1]	1.78	0.44	1.63 – 1.93	0.39 – 0.48	**<0.001**	**<0.001**
KinderTRUE	0.61	0.05	0.27 – 0.95	-0.04 – 0.15	**<0.001**	0.272
Observations	536					
R^2 / R^2 adjusted	0.578 / 0.572					

Als auch nicht ganz schlechte Alternative kann gtsummary::tbl_regression benutzt werden, was durch weitere gepipte Befehle angepasst und ergänzt werden kann. Hier sind die b zu sehen und deren CI, sowie die p-Werte und gleich noch zwei Varianzinflationswerte. Wenn Sie schauen wollen, ob es ein Problem gibt, helfen die «Adjusted GVIF» schon, weil sie für grössere Stichproben kleiner sind als die VIF und damit die Probleme durch Multikollinearität angemessener abbilden [@Fox1992]. Ein Adjusted GVIF wird unangenehm, wenn es über 2 ist, also der VIF etwa bei 4. Schön an der Tabelle @ref(tab:gtsummary-tabelle) ist auch, dass immer auch die Referenzkategorie mit ausgewiesen wird, auch wenn sie immer die Ausstriche für die Tabellen aufweisen, wie zB bei Sex:female. Unter der Tabelle stehen noch einige Gütemasse für das Modell, wie R^2 (.578) und die zugehörige F-Statistik als p-Wert (schön als <0.001, wobei ich die führende 0 weglassen würde, wenn die Werte praktisch immer etwas mit nach dem Komma sind).

Recht blöd finde ich an gtsummary, dass die *unstandardisierten* Regressionskoeffizienten als «beta» bezeichnet werden und die *standardisierten* Regressionskoeffizienten (die hin und wieder auch als «BETA» bezeichnet werden) garnicht in die Ausgabe gepackt werden können.

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**b**	**95% CI**	**p-value**	**GVIF**	**Adjusted GVIF**
Sex				1.5	1.2
female	—	—			
male	-0.32	-0.45, -0.18	<0.001		
Age_z	0.00	-0.01, 0.00	0.3	3.1	1.8
I(Age_z^2)	0.00	0.00, 0.00	0.7	2.5	1.6
Survived	-0.08	-0.22, 0.07	0.3	1.7	1.3
Pclass_f				1.4	1.1
3	—	—			
2	0.49	0.35, 0.63	<0.001		
1	1.8	1.6, 1.9	<0.001		
Kinder				3.5	1.9
FALSE	—	—			
TRUE	0.61	0.27, 0.95	<0.001		

Bevor wir unsere Therie anhand unserer Ergebnisse auf den Prüfstand stellen können, müssen wir die Voraussetzungen für die Regressionsmodelle überprüfen.

11.2.2 Voraussetzungschecks

Die möglichen Verletzungen der Voraussetzungen sind:

- Modellspezifikation
- Multikollinearität
- Deutliche Abweichung von der Normalverteilung in den Fehlern
- Heteroskedastizität

11.2.2.1 Modellspezifikation

An dieser Stelle müsste die Forschungsliteratur aufgearbeitet werden. Der Stand der Erkenntnis über die Zusammenhänge wird dann in einem Modell formuliert. Dann würde man weiter darüber nachdenken, ob das Modell aus den eigenen Erfahrungen und Erkenntnissen mit dem aktuellen Stand der Forschung voll spezifiziert ist oder aus unserer Sicht wichtige Einflussgrössen bisher nicht berücksichtigt wurden, die wir dann kunstvoll hinzufügen (erstmal erheben und dann in der Analyse ergänzen). Wenn das Grundmodell mit dem Stand der Forschung hinreichend korrespondiert, gehen wir davon aus, dass wir den Stand der Theorie bestätigen. Dann fügen wir unserer neuen argumentativ hergeleiteten Einflussgrössen als Variablen in das Modell ein. Wenn das dazu führt, dass die AV besser erklärt werden kann und/oder andere (bisher als wichtig erachtete) UVs an Erklärungskraft verlieren, dann können wir sagen, dass wir einen wissenschaftlichen Mehrwert geschaffen haben, da wir ein bestehendes Modell verbessert beziehungsweise korrigiert haben. Das heisst, wir haben ein unterspezifiziertes Modell (unzureichende Theorie) besser spezifiziert und unsere verbesserte oder neue Theorie hat sich sogar an Daten bewährt. Der Anspruch, einen wissenschaftlichen Mehrwert zu schaffen, wird nicht an Bachelorstudierende gestellt und in der Regel auch nicht an Masterstudierende. Das ist ein Anspruch, der in der Regel erst an Dissertationen gerichtet ist.

11.2.2.2 Multikollinearität

Die Toleranzen bei Age_z und I(Age_z^2) sind nur knapp über .3 und knapp .4, also schon recht klein, auch wenn die VIF noch nicht zu sehr quietschen. Da aber beide Altersvariablen keine signifikanten Werte gezeigt haben, würde ich eine der beiden eliminieren und zwar das quadrierte Alter «I(Age_z^2)». Dann sollte sich auch der sehr kleine Toleranzwert bei «KinderTRUE» erledigt haben bzw. kleiner sein - vielleicht auch nicht. Schauen Sie sich das mal an! Wenn Age_z auch ohne die quadrierte Altersvariable nicht signifikant ist, würde ich nur «KinderTRUE» im Modell lassen. Könnte man eine Faktorenanalyse für die zwei aufgrund der Altersvariablen gebildeten Variablen machen? Neeeee! Was soll da für ein latenter Faktor rauskommen? Wieder das Alter? Das macht keinen Sinn. Die Variable «Kinder» ist ja direkt aus dem Alter bestimmt worden. Da brauchen wir nicht nach latenten Faktoren zu suchen.

```
## Es gibt ein Paket "olsrr" für die Prüfung der OLS-Voraussetzungen,
## wo man sich den VIF und die Toleranz rauslassen kann:s

olsrr::ols_vif_tol(fit_titanic) |>
  kableExtra::kable()|>
  kableExtra::kable_styling()
```

Variables	Tolerance	VIF
Sexmale	0.6817534	1.466806
Age_z	0.3252169	3.074871
I(Age_z^2)	0.3995618	2.502742
Survived	0.5967148	1.675842
Pclass_f2	0.8292064	1.205972
Pclass_f1	0.6550188	1.526674
KinderTRUE	0.2878347	3.474217

11.2.2.3 Residualplot

```
# Mache mal ein Histogramm der Residuen. Die sollten annähernd normalverteilt sein.
olsrr::ols_plot_resid_hist(fit_titanic)
```

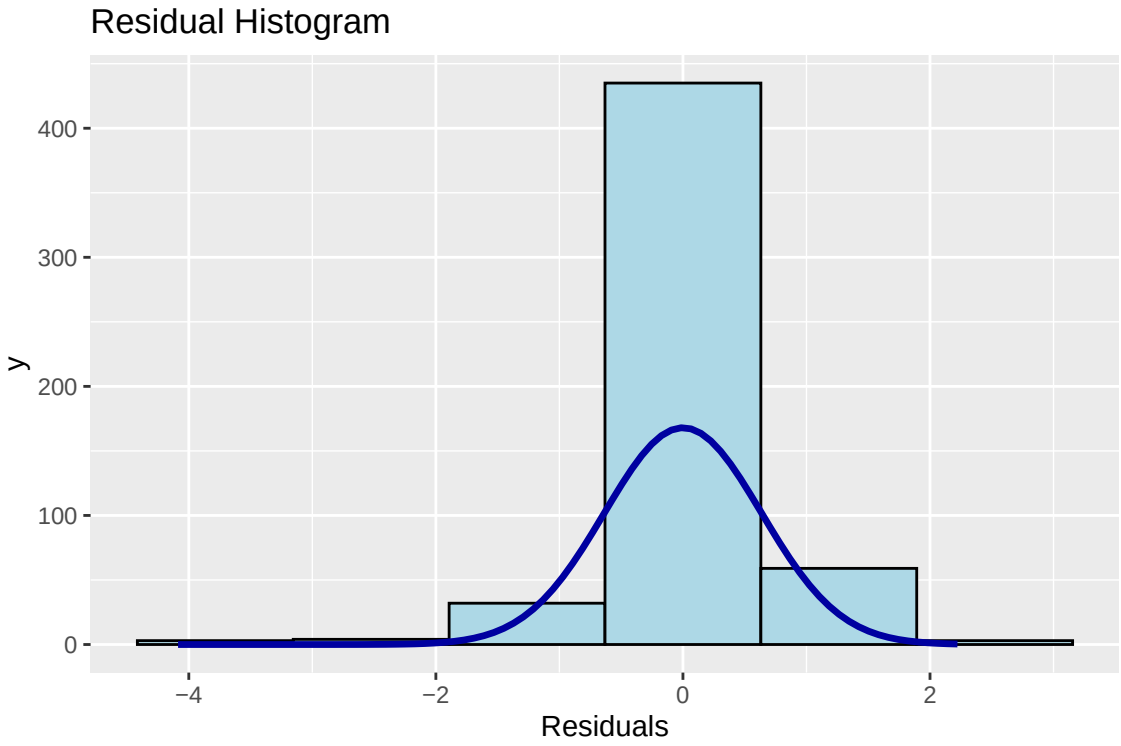
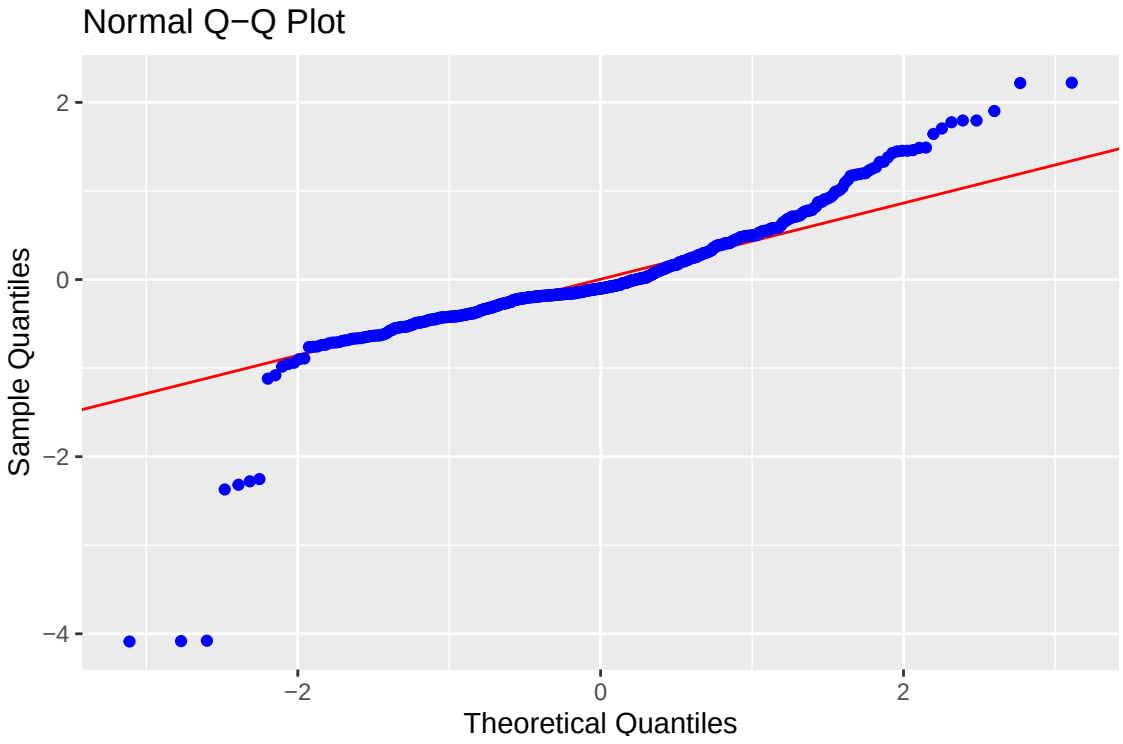


Abbildung 11.3: Histogramm der Residuen

Im Residualplot ohne den Logarithmus für die AV «Fare» war schon ganz schön schief. Hier sieht es etwas besser aus mit der Verteilung. Das ist ja schon fast normalverteilt, auch wenn die mittlere Kategorie wie ein Stinkefinger in der Landschaft steht. Schauen wir mal den N-Q-Q-Plot an, wie der aussieht ...

11.2.2.4 N-Q-Q

```
# Führe einen Normal-Q-Q-Plot aus
olsrr::ols_plot_resid_qq(fit_titanic)
```



OK, es gibt rechts eine paar Werte über der Referenzlinie, die für die Normalverteilung steht und links ein paar unter der Referenzlinie. Das heisst, die Normalverteilung ist nicht perfekt getroffen. Wir sollten also zunächst die Resultate unseres Modells nicht überinterpretieren. Es gibt noch viele Anpassungen, mit denen man diese kleineren Verletzungen der Voraussetzungen für die Regression auflösen kann. Das geht aber für die Flughöhe dieser Veranstaltung zu hoch bzw. zu tief, wie Sie wollen.

11.2.2.5 Heteroskedastizität

```
# Plotte die geschätzten Werte auf der Regressionsgeraden (Y-Hut)
# auf der X-Achse und die Residuen auf der Y-Achse
olsrr::ols_plot_resid_fit(fit_titanic)
```

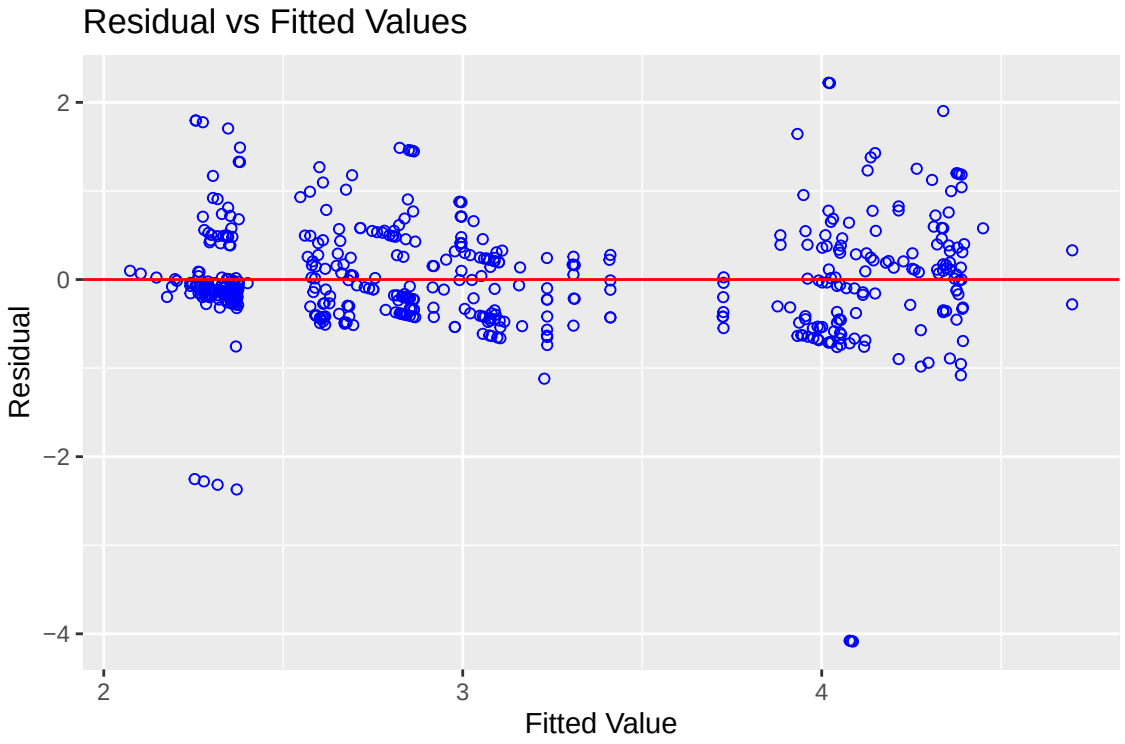


Abbildung 11.4: Plot für Fit und Residuen

Es gibt auch hier eine gewisse Heteroskedastizität, aber eigentlich sind die Werte schon relativ gleichmässig um 0 verteilt. Wir können das ja mal testen.

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                  Data
##  -----------------------------------------
##  Response : log(Fare + 1)
##  Variables: fitted values of log(Fare + 1)
##
##          Test Summary
##  -------------------------------
##  DF            =    1
##  Chi2          =    51.25669
##  Prob > Chi2   =    8.104396e-13
```

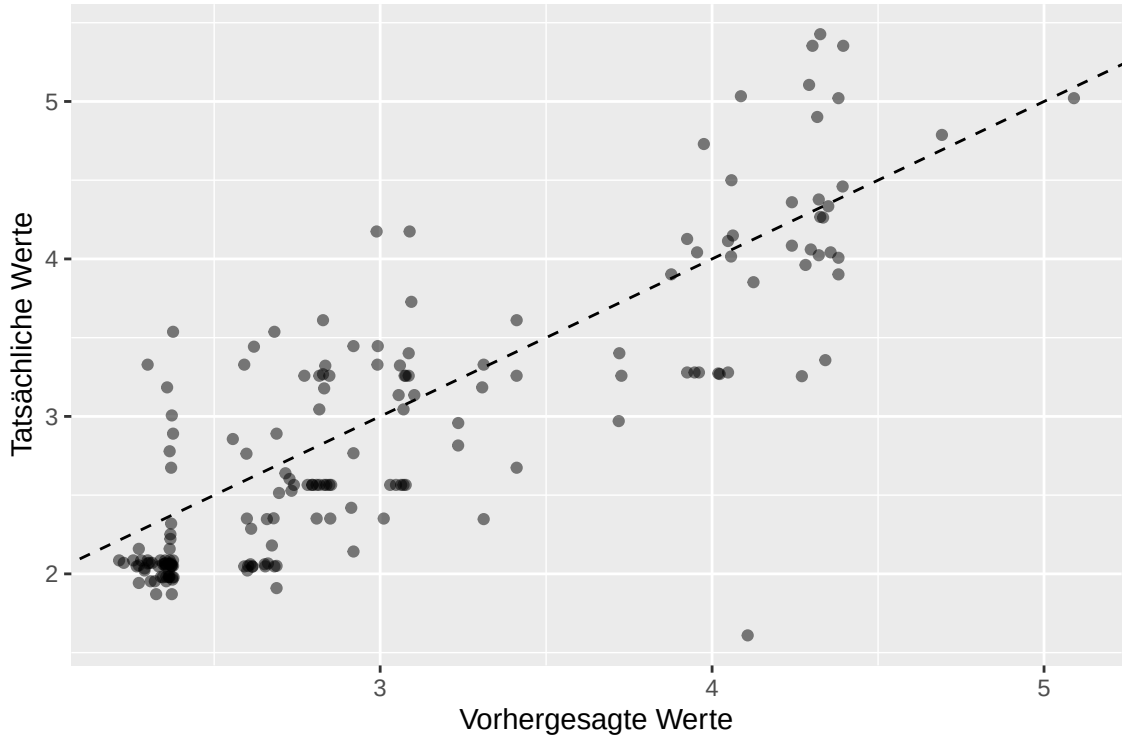
Oha, Breusch und Pagan finden, dass unsere Varianz der Residuen heftig heterogen ist. Der p-Wert des Chi2 ist sehr weit von 0 entfernt. Das ist allerdings schnell so, wenn die Stichprobe recht gross ist. Dann wird irgendwann jedes Chi2 signifikant. Also auch hier: Wir gehen vorsichtig mit den Ergebnissen um, interpretieren nicht exakt die Nachkommastellen von bs und sagen auch bei einem p-Wert von .03, dass die Signifikanz hier nicht ganz klar ist (vor dem Hintergrund, dass einige Voraussetzungen verletzt sind).

11.2.3 Algorithmus für den Fahrpreis

Wir haben jetzt also ein Modell und festegstellt, dass unsere Regression mit der log-Transformation für die AV und einer quadrierten UV nicht super durch die Prüfung der Voraussetzungen kommt. Also sind wir vorsichtig, lassen aber trotzdem mal einen Schätzalgorithmus für den Fahrpreis raus. Der Algorithmus ist schon da: Es ist der Fit des Modells. Dieses gefittete Modell können wir jetzt auf den Testdatensatz ansetzen und mal schauen, wie gut das Modell zu den Testdaten passt, anhand derer es nicht gebaut wurde, die aber auch die Outcomes enthalten, also den Fahrpreis.

Wir berechnen als vorhergesagte Werte die «preds» mit `predict(fit_titanic, test)`. Dann Binden wir die an den Test-Datensatz «test», wobei wir dort auch noch schnell den natürlichen Logarithmus für «Fare» bilden, indem wir `log(test$Fare)` einsetzen, mit cbind (Spalten zusammenbinden) zusammenfassen und als tibble (tidydatentabelle) speichern. In der Grafik sieht man, das die Prognosen nicht perfekt sind, aber ok. Wenn Sie es mal ohne den schwieriger zu interpretierenden log machen, dann sehen Sie spätestens hier die Probleme, weil die hohen Fahrpreise schwer kalkuliert werden können.

```
# mache aufgrund der Testdaten mit dem Modell "fit_titanic" eine
# Vorhersage (prediction)
preds <- predict(fit_titanic, test) |>
  as_tibble()

# Binde die tatsächliche logarithmierte Fare-Variable aus test mit den
# Vorhergesagten Werten zusammen
# und mit rename die Spaltenbeschriftung sinnvoller

modelEval <- cbind(log(test$Fare), preds) |>
  rename(actual = "log(test$Fare)", predicted = "value")

plot <- modelEval |>
  na.omit()|>
  ggplot(aes(x = predicted, y = actual)) +
  # Create a diagonal line:
  geom_abline(lty = 2) +
  geom_point(alpha = 0.5) +
  labs(y = "Tatsächliche Werte", x = "Vorhergesagte Werte")

plot
```



Jetzt hätten Sie also einen Algorithmus, den Sie nicht nur auf Test-Daten anlegen könnten, sondern an jede andere Konstellation von Daten, die die UVs enthält. Sie könnten also mit neuen Daten über `predict(fit_titanic, neudaten)` festlegen, was jede Person im Datensatz schätzungsweise für einen Preis für die Titanicüberfahrt gezahlt hätte. Das Modell ist nicht perfekt, aber Sie könnten es mit etwas Nachsteuern fair gestalten und jedem sagen, dass das die beste Anlehnung an die damalige (sicher eher analogen) Preisgestaltung ist.

Das ist zwar nicht völlkommen überflüssig, aber wenn wir an «Titanic» denken, denken wir an Leonardo DiCaprio und an das Überleben und Sterben vor dem Fernseher und natürlich damals auf der Titanic. Damit haben wir nicht gleich angefangen, weil das «Überleben» eine dichotome Variable ist und damit eine Dummy als AV. Das macht das Ganze schon etwas komplizierter, aber klar, schauen wir uns das an.

11.3 Logistische Regression

Wenn man an die Titanic denkt, grübelt man in der Regel nicht lange, wie wohl die Preise auf der Titanic waren. Viel mehr ist «Titanic» mit dem Schiffsunglück verbunden (ok und mehr oder weniger guten Verfilmungen). Wenn wir von dem Unglück etwas lernen wollen («Learning from Desaster»), dann ist es sinnvoll, Prognosemodelle für die Überlebenswahrscheinlichkeit zu machen. Die Überlebenschancen für verschiedene Personengruppen auf der Titanic ergeben sich daraus, wie viele Personen der Gruppen überlebt haben. Ein Erklärungsmodell hat also zur abhängigen Variable (AV), ob eine Person überlebt hat (1) oder nicht (0). Die AV ist also eine Dummyvariable. Wenn die AV eine Dummyvariable ist, dann verweigert es R nicht, eine lineare Regression zu rechnen (das macht es ein bischen «gefährlich», weil viele Kolleg:innen und Reviewer:innen normale lineare Regressionen bei Dummys in der AV als grossen Spezifikations-Fehler betrachten).

```
DATEN_titanic <- readRDS("data/titanic/train.RDS")

model <- lm(Survived ~ Pclass_f + Sex + Age_z + I(Age_z^2) + Kinder,
            data=train)
```

```
summary(model)
##
## Call:
## lm(formula = Survived ~ Pclass_f + Sex + Age_z + I(Age_z^2) +
##     Kinder, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96849 -0.25350 -0.06132  0.22493  0.97743
##
## Coefficients:
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.57899126  0.03753055  15.427   < 2e-16 ***
## Pclass_f2    0.17998008  0.04200296   4.285 0.0000217 ***
## Pclass_f1    0.37609658  0.04287552   8.772   < 2e-16 ***
## Sexmale     -0.50783615  0.03528671 -14.392   < 2e-16 ***
## Age_z       -0.00292748  0.00200019  -1.464     0.144
## I(Age_z^2)  -0.00001618  0.00008790  -0.184     0.854
## KinderTRUE   0.12521089  0.10287435   1.217     0.224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.382 on 529 degrees of freedom
##   (132 observations deleted due to missingness)
## Multiple R-squared:  0.4033, Adjusted R-squared:  0.3965
## F-statistic: 59.59 on 6 and 529 DF,  p-value: < 2.2e-16
```

Die Werte sind jedenfalls nicht intuitiv interpretierbar, weil die AV nur 0 und 1 annehmen kann und keine Zwischenwerte. Zudem streuen die Fehler stark um die Normalverteilungskurve (N-Q-Q-Plot), sind also stark heteroskedastisch.

11.3.1 Grundidee und Herangehensweise

Da eine lineare Regression b-Werte zur Folge hätte, die für reale Werte in den UVs Werte unter 0 und über 1 für die AV vorhersagen würde, wird eine logistische Regression gerechnet. Werte unter 0 und über 1 können nicht existieren, weil die AV eben eine Dummy ist und nur die Werte 0 und 1 kennt.

Die Formel einer logistischen Regression sieht eigentlich so aus:

$$P(y=1) = \frac{1}{1+e^{-z}} \quad \text{der Logit z ist die Regressionsgleichung} \tag{11.1}$$

$$Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i \tag{11.2}$$

$$P(Y_i) = \frac{1}{1+e^{-(b_1+b_2 X_{2i})}} \tag{11.3}$$

$$P(Y_i) = \frac{1}{1+e^{-(b_1+b_2 X_{2i}+b_3 X_{3i}+\cdots+b_n X_{ni})}} \tag{11.4}$$

Die Gerade ist keine Gerade mehr, sondern eine S-Kurve und sieht ein bisschen idealisiert so aus:

In der Abbildung 11.5 ist die UV auf der X-Achse sehr schön metrisch. Hier ist die Verteilung von UV und AV noch durch Histogramme dargestellt. In der Verteilung unten sieht man, dass die meisten Fälle mit einer 0 in der AV (darum ja unten) in der UV um den Wert 5 herum streuen; ein paar wenige Werte gibt es sogar unter 0 und auch ein paar über 10. Die Fälle mit einer 1 in der AV streuen etwas mehr nach rechtsalso eher um die 10 und schöpfen das ganze Spektrum von 0 bis 20 in der UV aus. Man sieht also

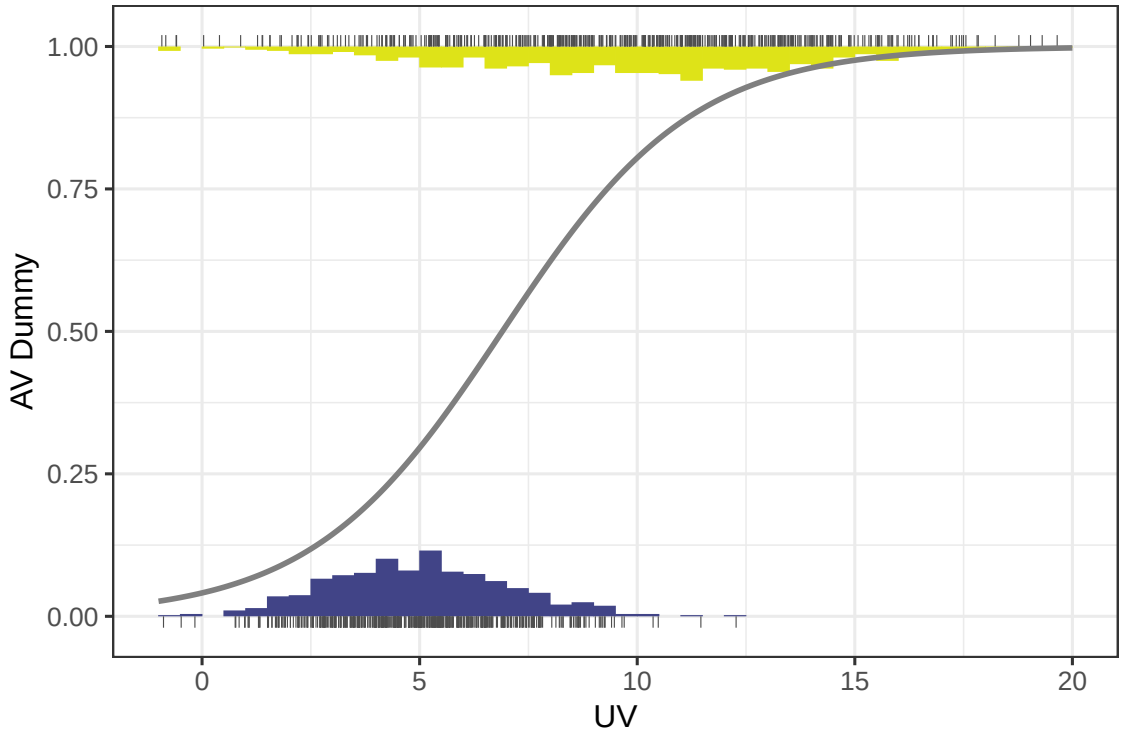


Abbildung 11.5

schon, dass mit höheren Werten in der UV die Wahrscheinlichkeit steigt, eine 1 zu haben. Darum ist die logistische Schätzkurve auch wie ein gedehntes S geformt und nicht umgekehrt wie ein Fragezeichen. Das wäre eben dann der Fall, wenn die 0 eher bei hohen UV-Werten vorkäme und die 1 bei niedrigen Werten in der UV.

Die berechneten b's sind kaum inhaltlich interpretierbar. Im «summary» Output stehen in der Spalte «Estimates» die b's. Was man sehen und sagen kann ist, dass die Mitfahrenden der 2. Klasse, im Vergleich zur 3. Klasse, eine bessere Chance hatten, zu überleben (der Estimate (b und keine OR) ist signifikant positiv). Die Mitreisenden der ersten Klasse hatten eine noch grössere Chance zu überleben (b ist positiv, grösser als bei Pclass_f2, hat auch einen grösseren z-Wert und einen kleineren p-Wert).

```
##
## Call:
## glm(formula = Survived ~ Pclass_f + Sex + Kinder, family = binomial,
##     data = train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2353     0.2209   1.065  0.28684
## Pclass_f2     1.1302     0.2864   3.946 7.94e-05 ***
## Pclass_f1     2.1581     0.2871   7.518 5.57e-14 ***
## Sexmale      -2.7026     0.2438 -11.087  < 2e-16 ***
## KinderTRUE    1.1727     0.3732   3.142  0.00168 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 724.29  on 535  degrees of freedom
## Residual deviance: 483.87  on 531  degrees of freedom
##   (132 observations deleted due to missingness)
## AIC: 493.87
##
## Number of Fisher Scoring iterations: 5
```

Besser als die b's können die exponentiellen b's EXP(B) gelesen werden. Sie geben eine «Odds Ratio» an. Das kann so gelesen werden, wie Multiplikatoren von Wahrscheinlichkeiten. Die «Odds Ratios» in Tabelle @ref(tab:Publikationsoutput1) bwz. die «OR» in Tabelle @ref(tab:Publikationsoutput2) geben diese Werte raus, die man mit (Wett)quoten übersetzen könnte. Sie fangen bei >0 an und können unendlich gross werden. Wenn eine Variable keinen Einfluss auf die Wahrscheinlichkeit des Ausgangs der AV hat, dann ist ihr b $= 1$. Im Beipspiel kann man ablesen, dass im Vergleich zur 3. Passagierklasse (Pclass_f ist die Referenz und darum in Tabelle @ref(tab:Publikationsoutput1) gar nicht zu sehen und in Tabelle @ref(tab:Publikationsoutput2) ausgestrichen) die 2. Passagierklasse eine 3.1-fache Überlebenschance hatte und die 1. Passagierklasse eine 8.65-fache.

Tabelle 11.3: Überlebensanalyse zum Titanicunglück mit sjPlot

		Survived	
Predictors	Odds Ratios	CI	p
(Intercept)	1.27	0.82 – 1.96	0.287
Pclass f [2]	3.10	1.77 – 5.47	**<0.001**
Pclass f [1]	8.65	4.99 – 15.41	**<0.001**
Sex [male]	0.07	0.04 – 0.11	**<0.001**
KinderTRUE	3.23	1.56 – 6.78	**0.002**
Observations	536		
R^2 Tjur	0.404		

Die Modellausgabe kann auch mit `gtsummary` erfolgen (und es gibt einige weitere Pakete). Bei gtsummary werden die Spalten für die Analyse durch Befehle in der Pipe ergänzt. In der folgenden Variante werden die Odds-Ratios (OR) rausgelassen und ihre Konfidenzintervalle (CI) sowie die p-Werte und den (generalisierten) Varianzinflationsfaktor. Der kommt sogar noch mit einer Anpassung, dem «Adjusted GVIF». So lange der unter 2 liegt, wird die Analyse nicht zu sehr von Multikollinearität gestört.

```
model |>
  gtsummary::tbl_regression(exponentiate = TRUE) |>
  gtsummary::add_vif() |>
#  gtsummary::add_global_p() |> # damit könnte man die Signifikanz ganzer
                               # Faktoren testen, statt jede Ausprägung gegen
                               # die Referenz
  gtsummary::add_glance_source_note() |>
  gtsummary::modify_caption(
    "**Überlebensanalyse zum Titanicunglück mit gtsummary**")
```

11.4 Voraussetzungschecks

11.4.0.1 Multikollinearität

```
## Es gibt ein Paket "olsrr" für die Prüfung der OLS-Voraussetzungen,
## wo man sich den VIF und die Toleranz rauslassen kann:s
lm(Survived ~ Pclass_f + Sex + Age_z + I(Age_z^2) + Kinder,  family=binomial,
          data=train) |>
  olsrr::ols_vif_tol()
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
##    Variables Tolerance      VIF
## 1  Pclass_f2 0.8579868 1.165519
## 2  Pclass_f1 0.7502936 1.332811
## 3    Sexmale 0.9486830 1.054093
## 4      Age_z 0.3265338 3.062470
## 5 I(Age_z^2) 0.3995874 2.502582
## 6 KinderTRUE 0.2886407 3.464515
```

11.4.1 Residualplot

```
plot(model)
```

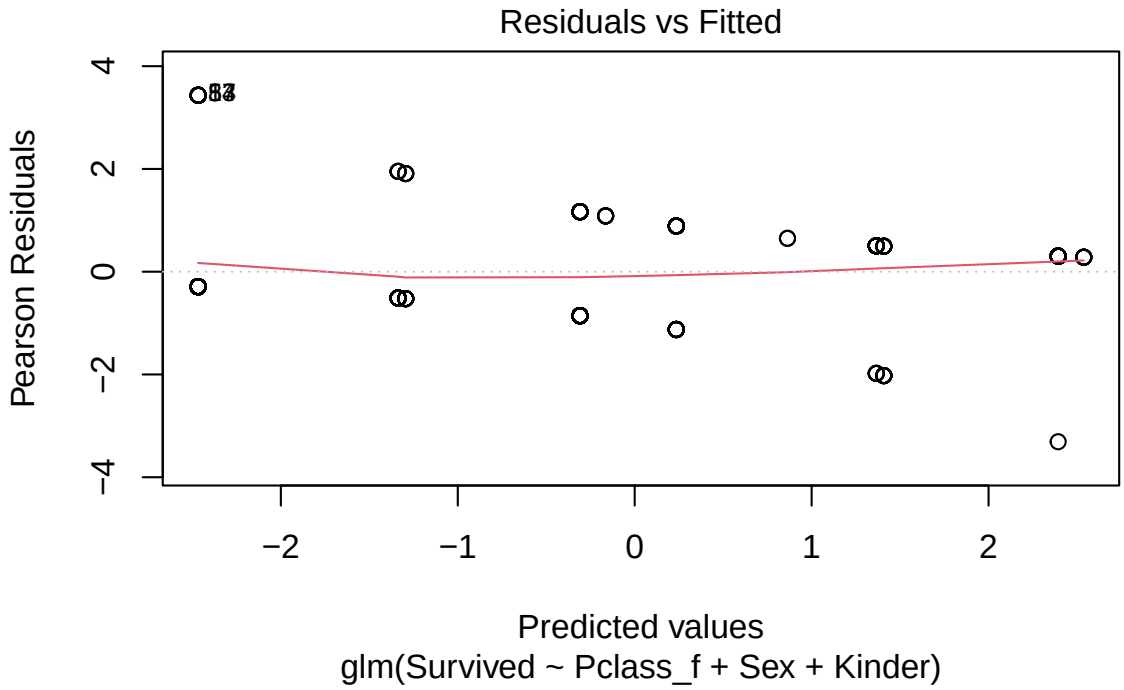


Abbildung 11.6: Histogramm der Residuen

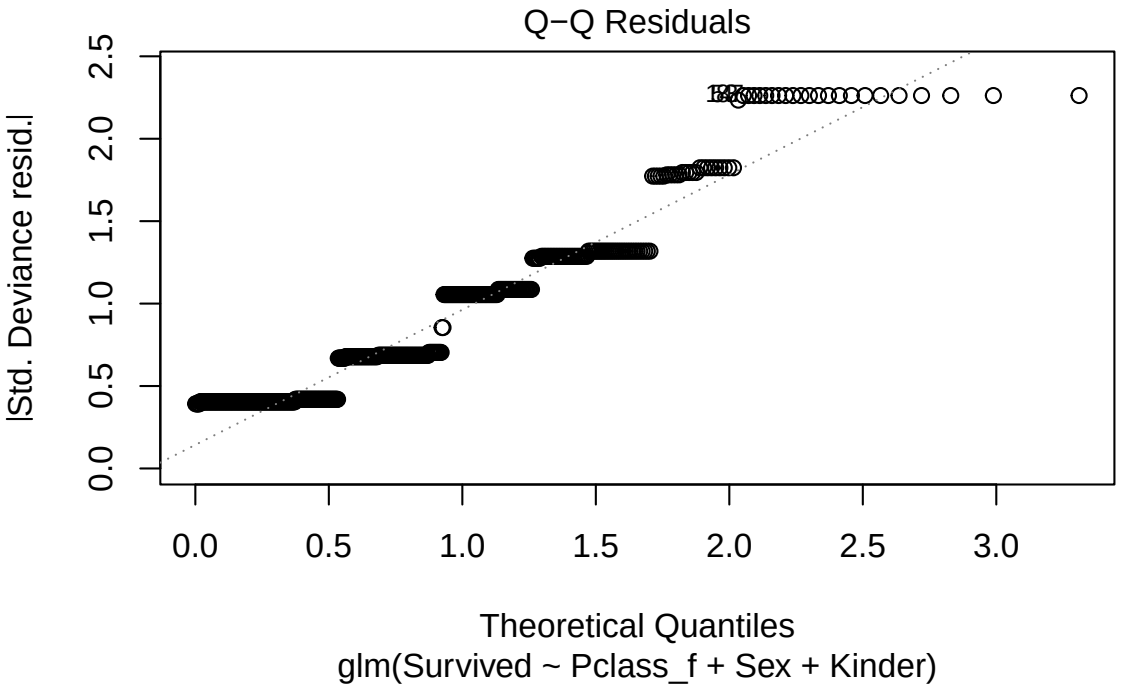


Abbildung 11.7: Histogramm der Residuen

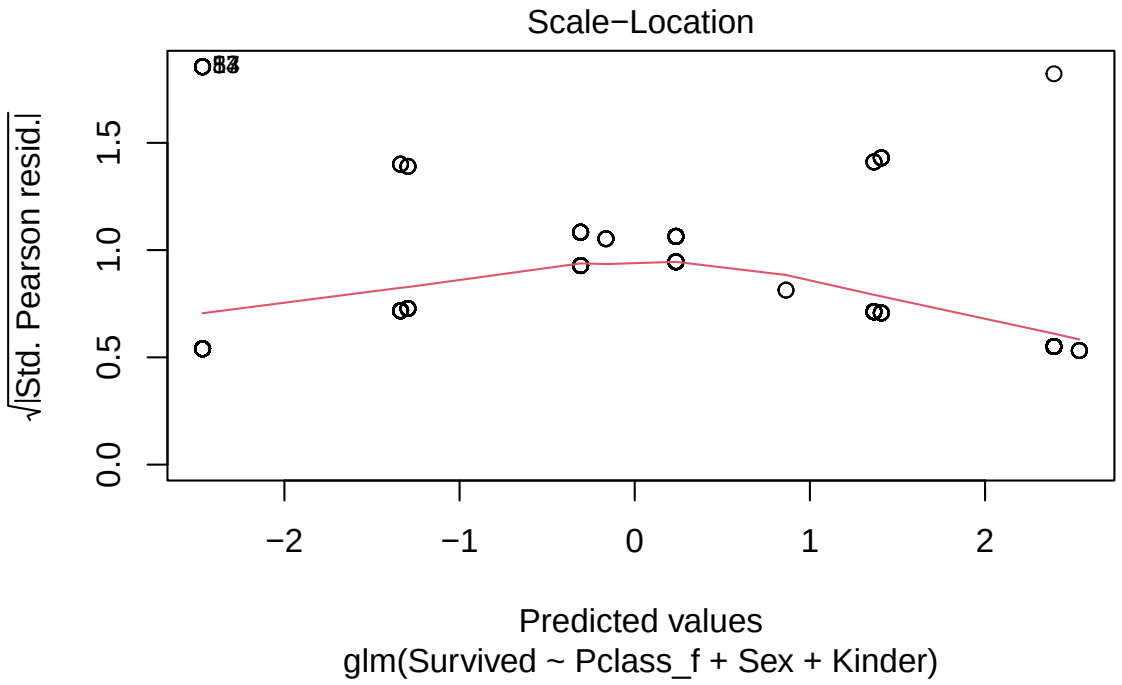


Abbildung 11.8: Histogramm der Residuen

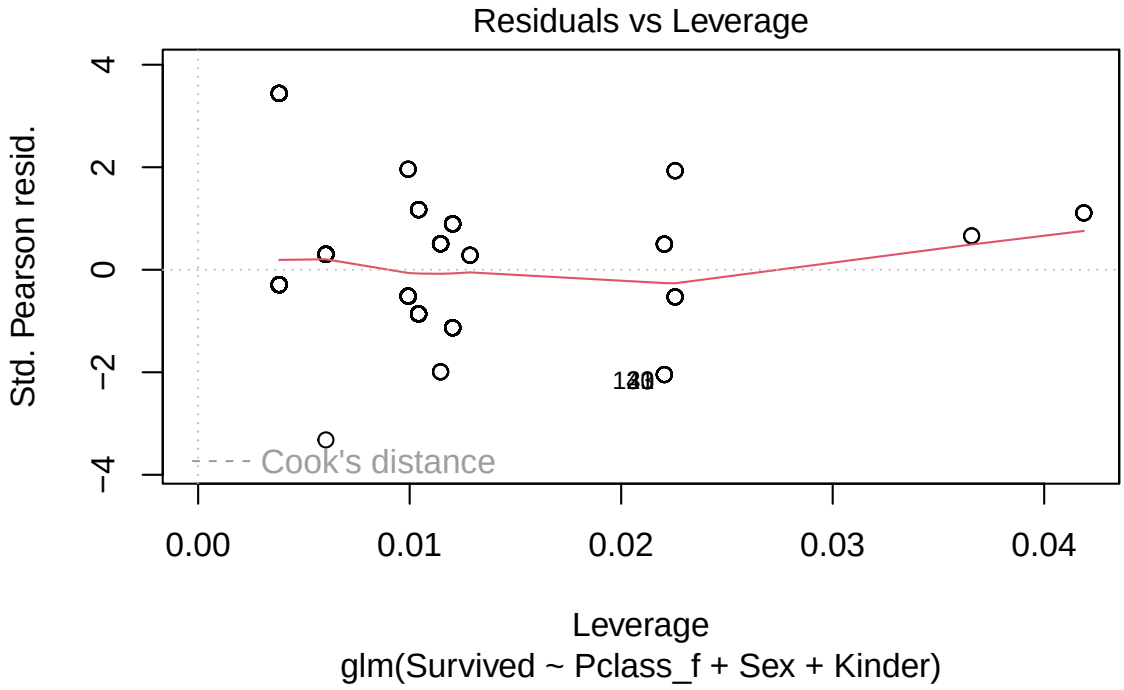


Abbildung 11.9: Histogramm der Residuen

11.4.1.1 Vorhersagetest

```
preds <- predict(model, test)

modelEval <- cbind(test$Survived, preds) |>
  as_tibble() |>
  mutate(preds = ifelse(preds > 0.5,1,0)) |>
  rename(actual = 'V1', predicted = 'preds')

modelEval |> # sjmisc::flat_table()
  count(actual, predicted) |>
  pivot_wider(names_from = predicted, values_from = n)|>
  kableExtra::kable()|>
  kableExtra::kable_styling()|>
  kableExtra::add_header_above(c("", "predicted" = 3))
```

	predicted		
actual	0	1	NA
0	100	6	33
1	31	41	12

```
modelEval |>
  dplyr::filter(!is.na(predicted)) |>
  mutate(test = actual == predicted, na.rm = TRUE) |>
```

```
  summarise(Accuracy = mean(test))
## # A tibble: 1 x 1
##   Accuracy
##      <dbl>
## 1    0.792
```

12 Übung: Machine Learning

Zum Termin bitte durchgehen: Übung 4 und Text Zerback Wirz! lesen! (liegt beides auch in OLAT unter Materialien bzw. Texte)

Übung 4

In der Übung beschäftigen wir uns mit Daten des Untergangs der Titanic. Schauen Sie sich jeweils die Befehle an und finden Sie heraus, was die Befehle machen. Machen Sie sich Notizen, wenn Sie etwas nicht verstehen.

Daten einlesen

Legen Sie für diese Übung einen Ordner an, erstellen Sie dort eine qmd und kopieren Sie die Folgenden Daten in einen dortigen Unterordner «data»:

Suchen und nehmen Sie den Datensatz «train.csv»: Downlaod der Daten: https://www.kaggle.com/competitions/titanic/data

```
# DATEN_titanic <- read_csv("data/titanic/train.csv") # lese beim ersten Mal die Daten ein

DATEN_titanic <- readRDS("data/titanic/train.RDS") # Lese nach dem ersten Mal so die Daten ein.
saveRDS(DATEN_titanic, "data/titanic/train.RDS") # speichere die Daten
```

12.1 Datenaufbereitung

```
DATEN_titanic <- DATEN_titanic |>
    mutate(Kinder = Age < 14) |> # Kinder werden als unter 14 Jährige definiert
    mutate(Age_z = Age - mean(Age, na.rm = TRUE)) |> # Das alter um den Mittelwert verschieben (zenti
    mutate(Pclass_f = factor(Pclass, levels = c(3, 2, 1)),
           Cabin_D = ifelse(is.na(Cabin), 0, 1))
```

… und mal die Daten angucken:

```
DATEN_titanic |> # mache eine Zusammenfassung der Daten
  summary()
##   PassengerId       Survived          Pclass          Name               Sex
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##       Age            SibSp           Parch           Ticket               Fare
##  Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00
##  1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91
##  Median :28.00   Median :0.000   Median :0.0000   Mode  :character   Median : 14.45
##  Mean   :29.70   Mean   :0.523   Mean   :0.3816                      Mean   : 32.20
##  3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00
##  Max.   :80.00   Max.   :8.000   Max.   :6.0000                      Max.   :512.33
##  NA's   :177
```

```
##     Cabin           Embarked    Kinder          Age_z          Pclass_f    Cabin_D
##  Length:891         S   :644   Mode :logical   Min.   :-29.279   3:491    Min.   :0.000
##  Class :character   C   :168   FALSE:643       1st Qu.: -9.574   2:184    1st Qu.:0.000
##  Mode  :character   Q   : 77   TRUE :71        Median : -1.699   1:216    Median :0.000
##                     NA's:  2   NA's :177       Mean   :  0.000            Mean   :0.229
##                                                3rd Qu.:  8.301            3rd Qu.:0.000
##                                                Max.   : 50.301            Max.   :1.000
##                                                NA's   :177
```

Die PassengerId ist einefach eine Identifikationsnummer.

- Es gibt dann eine Variable, die «Survived» heisst, die ein Minimum von 0 hat und ein Maximum von 1. Das deutet sehr auf eine Dummy hin. Da der Durchschnitt («Mean») = 0.38 ist, wissen wir jetzt schon, dass 38 Prozent der Passagiere überlebt haben (der Mittelwert einer Dummy ist immer der Prozentsatz der 1er-Gruppe).
- Dann kommt noch der Name als Zeichenvariable,
- das Alter, das von 0.42 bis 80 geht. Von 177 Personen fehlen die Altersangaben.

Informieren Sie sich über die übrigen Variablen auf (https://www.kaggle.com/competitions/titanic/data)[«Kaggle»].

Daten in Trainings- und Testdaten aufteilen

Was passiert hier?

```
# Setze eine Zufallszahl, damit die Ergebnisse replizierbar sind, also nicht jedes Mal eine neue Z
set.seed(12345)

# Ziehe eine Zufallsstichprobe aus dem Filmdatensatz und bezeichne ihn als "train", also Trainings
train <- DATEN_titanic |>
  sample_frac(.75)

# Bilde aus dem Rest der nicht für "train" gezogenen Fälle einen Test-Datensatz, indem nach 'id' d
test  <- anti_join(DATEN_titanic, train, by = 'PassengerId')
```

Sehen kann man nach diesem r-Chunk übrigens nichts, weil nur Datensätze im Hintergrund aufgeteilt wurden. Also suchen wir mal nach guten Datenvisualisierungen.

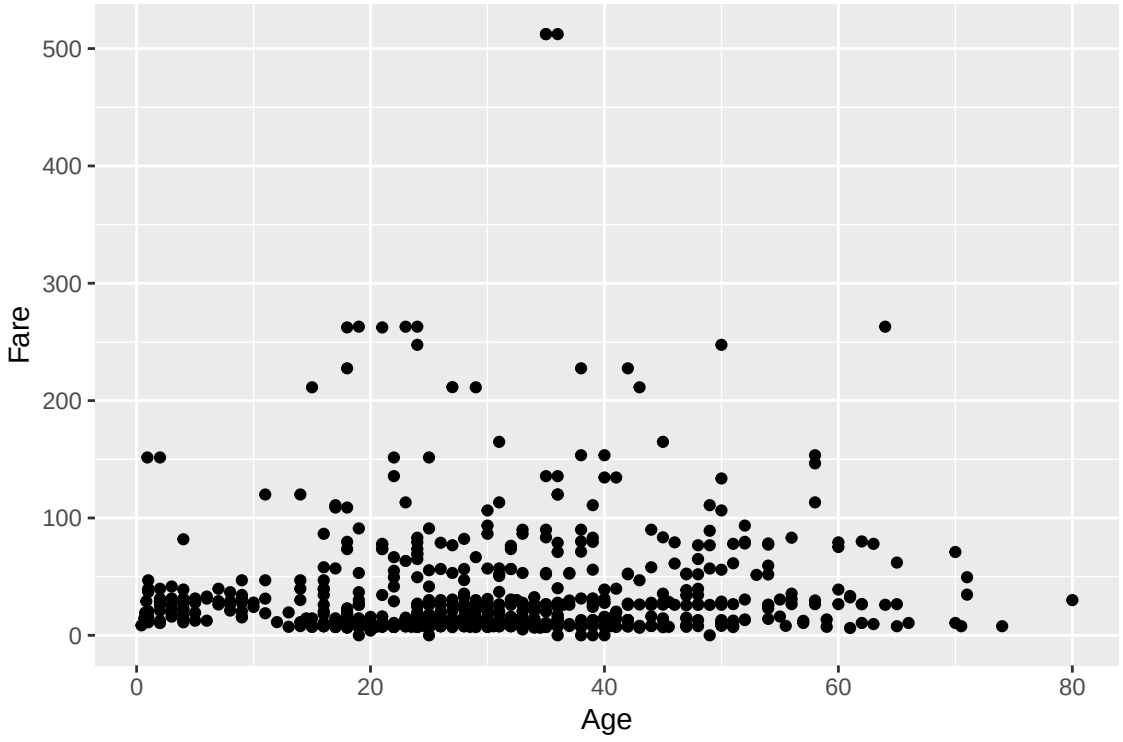
Übung: Experimentieren Sie mit verschiedenen Zahlen in set.seed() und grösseren und kleineren Werten in sample_frac().

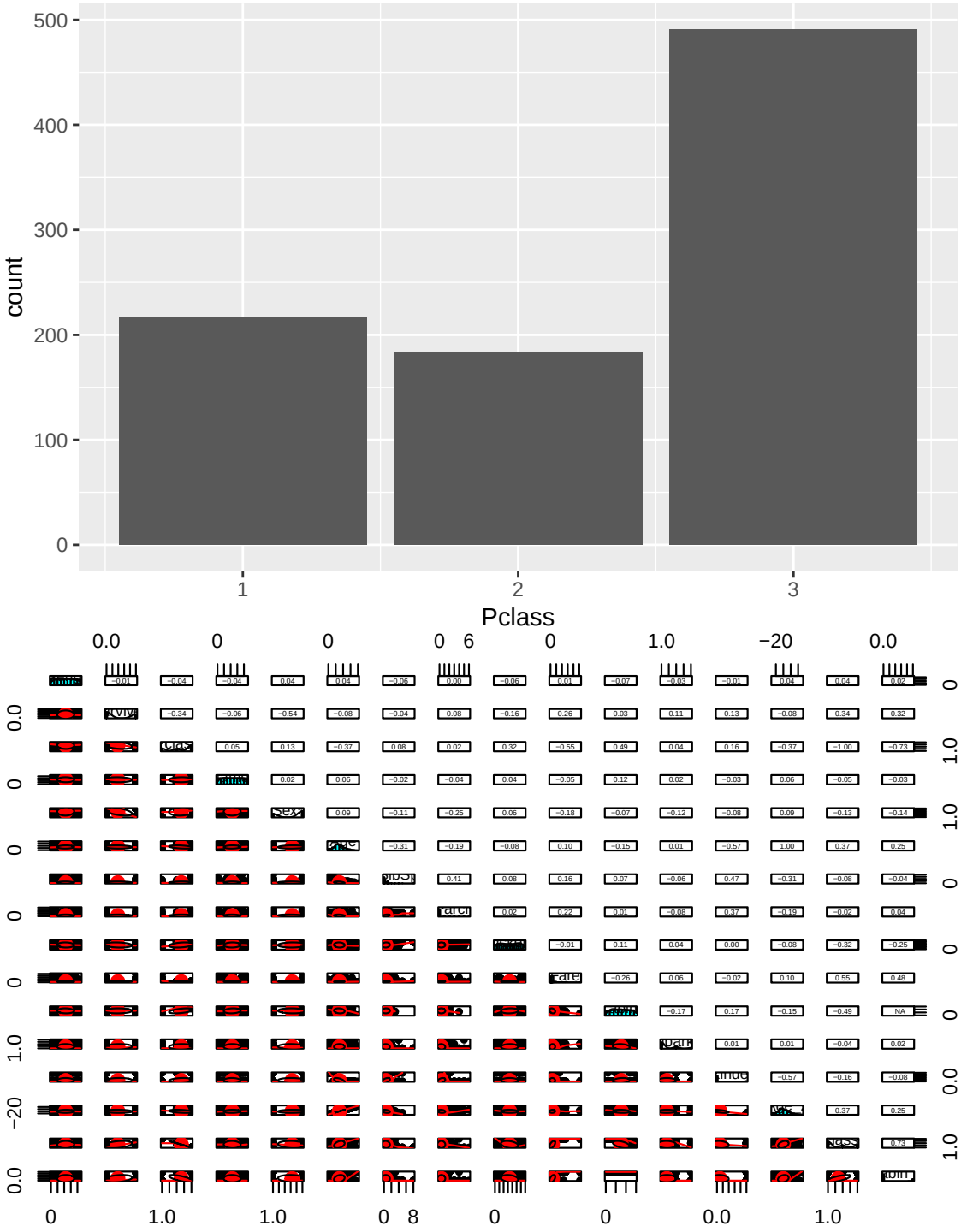
Datenvisualisierung

Eine einfache Darstellungsmöglichkeit ist ein sogenannter «Scatterplot» der die Lage von Fällen in ein Koordinatensystem einteilt, das durch zwei Variablen gebildet wird. Im Beispiel (Abbildung @ref(titanic-Datenvisualisierungen)auf Seite ??) ist es das Alter der Passagiere «Age» und der Fahrpreis «Fare». Als Zweites haben wir ein Balkendiagramm für die Passagierklassen. Im letzten sehr aufwendigen Plotgrafik werden die Zweierbeziehungen aller Variablen dargestellt, also wie sie miteinander korrelieren (obere Nebendiagonalen), wie ihre Verteilung ist (auf der Diagonale mit Namen) und wie ihre gemeinsame Streuung ist, also ein Scatterplot in der unteren Nebendiagonalen. Mehr zu diesen SPLOM finden Sie hier: https://cran.r-project.org/web/packages/psych/vignettes/intro.pdf.

```
# Erstelle einen Scatterplot (Punktewolke)

DATEN_titanic |>
```

```
  ggplot(aes(x = Age, y = Fare)) +
  geom_point()
## Warning: Removed 177 rows containing missing values (`geom_point()`).

# Erstelle ein Balkendiagramm
DATEN_titanic |>
  ggplot(aes(x=Pclass)) +
  geom_bar()

# Alle auf einmal
psych::pairs.panels(DATEN_titanic)
## Warning in cor(x, y, use = "pairwise", method = method): the standard deviation is zero
## Warning in cor(x, y, use = "pairwise", method = method): the standard deviation is zero
```





Modellbildung für den Fahrpreis

Übung: Experimentieren Sie mit dem Syntax! Kopieren Sie sich die Zeile für das Modell, löschen von «Age_z» bis «Kinder» alles heraus und schätzen Sie mal. Schauen Sie sich das Ergebnis gut an und achten Sie darauf, was passiert, wenn Sie die Summanden für I(Age_z^2) usw. wieder in das Modell tun. Am Ende können Sie versuchen das Modell durch weitere Variablen ergänzen und verbessern oder andere Teile wieder herausnehmen. Manche Variablen wurden erst noch erstellt (zB «Kinder» oder «Age_z»). Die

entsprechende Datenaufbereitung.qmd können Sie hier abrufen .

Übung: Interpretieren Sie den Output!

```
fit_titanic <- lm(log(Fare + 1) ~ Sex + Age_z + I(Age_z^2) + Survived + Pclass_f + Kinder, data = tra

broom::tidy(fit_titanic) |>
  mutate(across(where(is.numeric), ~round(.,3)))|>
  gt::gt()
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.657	0.076	35.047	0.000
Sexmale	-0.316	0.070	-4.520	0.000
Age_z	-0.003	0.003	-1.009	0.314
I(Age_z^2)	0.000	0.000	-0.406	0.685
Survived	-0.077	0.073	-1.050	0.294
Pclass_f2	0.492	0.072	6.859	0.000
Pclass_f1	1.778	0.077	23.104	0.000
KinderTRUE	0.607	0.173	3.515	0.000

```
broom::glance(fit_titanic)|>
  round(2)|>
  gt::gt()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.58	0.57	0.64	103.29	0	7	-517.85	1053.7	1092.26	216.7	528	536

Regressionsoutput mit `sjPlot::tab_model`

```
sjPlot::tab_model(fit_titanic,
                  show.std = TRUE, # zeige die standardisierten Koeffizienten
                  show.est = TRUE, # zeige die unstandardisierten estimates
                  show.r2 = TRUE, # zeige R^2
                  show.fstat = FALSE, # zeige die F-Statistik
                  string.est = "b", # beschrifte die Estimator mit b
                  string.std = "std. b" # beschrifte die standardisierten b mit std. b
                  )
## Formula contains log- or sqrt-terms.
##   See help("standardize") for how such terms are standardized.
```

	log(Fare+1)					
Predictors	b	std. b	CI	standardized CI	p	std. p
(Intercept)	2.66	0.83	2.51 – 2.81	0.80 – 0.87	**<0.001**	**<0.001**
Sex [male]	-0.32	-0.06	-0.45 – -0.18	-0.10 – -0.02	**<0.001**	**0.003**
Age z	-0.00	-0.02	-0.01 – 0.00	-0.05 – 0.00	0.314	0.063
Age z^2	-0.00	-0.00	-0.00 – 0.00	-0.02 – 0.01	0.685	0.639
Survived	-0.08	-0.00	-0.22 – 0.07	-0.02 – 0.02	0.294	0.914

Pclass f [2]	0.49	0.06	0.35 – 0.63	0.02 – 0.10	**<0.001**	**0.004**
Pclass f [1]	1.78	0.44	1.63 – 1.93	0.39 – 0.48	**<0.001**	**<0.001**
KinderTRUE	0.61	0.05	0.27 – 0.95	-0.04 – 0.15	**<0.001**	0.272
Observations	536					
R^2 / R^2 adjusted	0.578 / 0.572					

Als auch nicht ganz schlechte Alternative kann gtsummary::tbl_regression benutzt werden, was durch weitere gepipte Befehle angepasst und ergänzt werden kann.

Regressionsoutput mit `gtsummary::tbl_regression`

Übung: Was für ein Problem sehen Sie im folgenden Output? Bei welcher Befehlszeile müssten Sie das # wegnehmen, um das Problem zu lösen?

```
## Alternative zum sjPlot-Ouptut. Sie können entscheiden, was Sie besser finden.
fit_titanic |>
  gtsummary::tbl_regression() |>
  gtsummary::add_vif() |>
#  gtsummary::add_global_p() |> # Gebe die Signifikanzen je Variable raus (Varianzaufklärung)
#  gtsummary::modify_header(estimate = "**b**")   |> # Sorgt dafür, dass die b's nicht Beta heisse
  gtsummary::add_glance_source_note()
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	Beta	95% CI	p-value	GVIF	Adjusted GVIF
Sex				1.5	1.2
female	—	—			
male	-0.32	-0.45, -0.18	<0.001		
Age_z	0.00	-0.01, 0.00	0.3	3.1	1.8
I(Age_z^2)	0.00	0.00, 0.00	0.7	2.5	1.6
Survived	-0.08	-0.22, 0.07	0.3	1.7	1.3
Pclass_f				1.4	1.1
3	—	—			
2	0.49	0.35, 0.63	<0.001		
1	1.8	1.6, 1.9	<0.001		
Kinder				3.5	1.9
FALSE	—	—			
TRUE	0.61	0.27, 0.95	<0.001		

Voraussetzungschecks

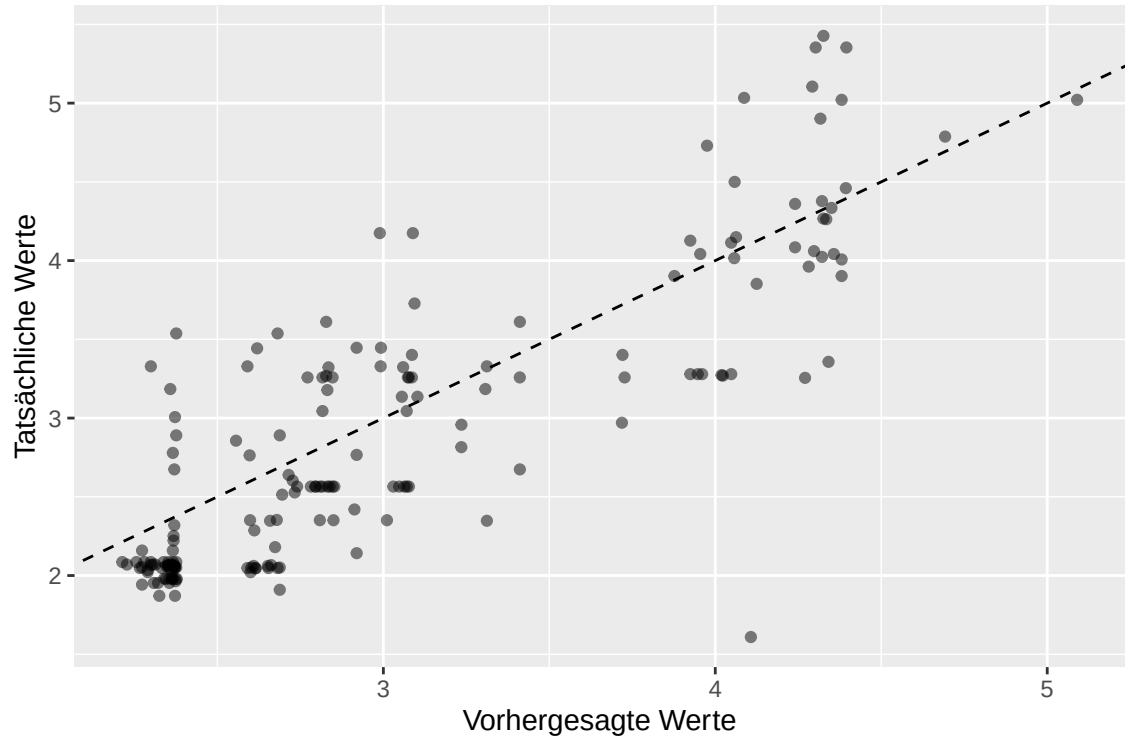
Übung: Suchen Sie die Befehle und führen Sie sie aus:
Die möglichen Verletzungen der Voraussetzungen sind:

- Multikollinearität
- Deutliche Abweichung von der Normalverteilung in den Fehlern
- Heteroskedastizität
- Outlier

Algorithmus für den Fahrpreis

Prognosewerte erstellen:

```
preds <- predict(fit_titanic, test)

modelEval <- cbind(log(test$Fare), preds) |>
  as_tibble()
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if `.name_repair`
## is omitted as of tibble 2.0.0.
## i Using compatibility `.name_repair`.

# Beschrifte die Spalten
colnames(modelEval) <- c('actual', 'predicted')

plot <- ggplot(modelEval, aes(x = predicted, y = actual)) +
  # Create a diagonal line:
  geom_abline(lty = 2) +
  geom_point(alpha = 0.5) +
  labs(y = "Tatsächliche Werte", x = "Vorhergesagte Werte")

plot
## Warning: Removed 45 rows containing missing values (`geom_point()`).
```



Überlebensprognose

Zweite Analyse: Überlebensprognose (Dummy) mit linearer Regression. Was denken Sie?

```
DATEN_titanic <- readRDS("data/titanic/train.RDS")

model1 <- lm(Survived ~ Pclass_f + Sex + Age_z + I(Age_z^2) + Kinder, data=train)

broom::tidy(model1) |>
 mutate(p.value = scales::pvalue(p.value)) |> # damit der p-Wert nicht mit endlosen Nullen dargest
 mutate(across(where(is.numeric), ~round(.x, 3))) |> # über alle numerischen Variablen hinweg, run
 gt::gt()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.579	0.038	15.427	<0.001
Pclass_f2	0.180	0.042	4.285	<0.001
Pclass_f1	0.376	0.043	8.772	<0.001
Sexmale	-0.508	0.035	-14.392	<0.001
Age_z	-0.003	0.002	-1.464	0.144
I(Age_z^2)	0.000	0.000	-0.184	0.854
KinderTRUE	0.125	0.103	1.217	0.224

```
broom::glance(model1)|>
 mutate(p.value = scales::pvalue(p.value)) |> # damit der p-Wert nicht mit endlosen Nullen dargest
 mutate(across(where(is.numeric), ~round(.x, 3))) |> # über alle numerischen Variablen hinweg, run
 gt::gt()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nc
0.403	0.397	0.382	59.587	<0.001	6	-241.157	498.314	532.587	77.177	529	5

Logistische Regression

```
model2 <- glm(Survived ~ Pclass_f + Sex + Kinder,  family=binomial, data=train)

# Berechnung der Odds-Ratios (OR) für Pclass_f2 und Pclass_f1 um sie unten im Text verwenden zu kö
OR_Pclass_f2 <- round(exp(summary(model2)$coefficients["Pclass_f2", 1]), 2)
OR_Pclass_f1 <- round(exp(summary(model2)$coefficients["Pclass_f1", 1]), 2)

broom::tidy(model2)|>
 mutate(p.value = scales::pvalue(p.value)) |> # damit der p-Wert nicht mit endlosen Nullen dargest
 mutate(across(where(is.numeric), ~round(.x, 3))) |> # über alle numerischen Variablen hinweg, run
 gt::gt()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.235	0.221	1.065	0.287
Pclass_f2	1.130	0.286	3.946	<0.001
Pclass_f1	2.158	0.287	7.518	<0.001
Sexmale	-2.703	0.244	-11.087	<0.001

KinderTRUE	1.173	0.373	3.142	0.002

```
broom::glance(model2)|>
 mutate(across(where(is.numeric), ~round(.x, 3))) |> # damit der p-Wert nicht mit endlosen Nullen da
  gt::gt()
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
724.287	535	-241.935	493.871	515.291	483.871	531	536

Kennwerte

Übung: Interpretieren Sie den Output

```
model2 <- glm(Survived ~ Pclass_f + Sex +  Kinder,  family=binomial, data=train)

sjPlot::tab_model(model2,title = "Überlebensanalyse zum Titanicunglück mit sjPlot",
                  show.est = TRUE, # zeige die estimates
                  show.r2 = TRUE, # zeige R^2
                  show.fstat = TRUE # zeige die F-Statistik
                  )
```

Tabelle 12.3: Überlebensanalyse zum Titanicunglück mit sjPlot

		Survived	
Predictors	Odds Ratios	CI	p
(Intercept)	1.27	0.82 – 1.96	0.287
Pclass f [2]	3.10	1.77 – 5.47	**<0.001**
Pclass f [1]	8.65	4.99 – 15.41	**<0.001**
Sex [male]	0.07	0.04 – 0.11	**<0.001**
KinderTRUE	3.23	1.56 – 6.78	**0.002**
Observations	536		
R^2 Tjur	0.404		

Wieder auch mit `gtsummary::`

```
model2 |>
  gtsummary::tbl_regression(exponentiate = TRUE) |>
  gtsummary::add_vif() |>
#  gtsummary::add_global_p() |> # damit könnte man die Signifikanz ganzer Faktoren testen, statt jed
  gtsummary::add_glance_source_note() |>
  gtsummary::modify_caption("**Überlebensanalyse zum Titanicunglück mit gtsummary**")
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Tabelle 12.4: **Überlebensanalyse zum Titanicunglück mit gtsummary**

Characteristic	OR	95% CI	p-value	GVIF	Adjusted GVIF
Pclass_f				1.2	1.0
3	—	—			
2	3.10	1.77, 5.47	<0.001		
1	8.65	4.99, 15.4	<0.001		
Sex				1.1	1.0
female	—	—			
male	0.07	0.04, 0.11	<0.001		
Kinder				1.1	1.0
FALSE	—	—			
TRUE	3.23	1.56, 6.78	0.002		

Voraussetzungschecks

Übung: Führen Sie die Checks für die Voraussetzungen aus!

Vorhersagetest

Was sagt dieser Test?

```
preds <- predict(model2, test)

modelEval <- cbind(test$Survived, preds) |>
  as.tibble() |>
  mutate(preds = ifelse(preds > 0.5,1,0))

# Beschrifte die Spalten
colnames(modelEval) <- c('actual', 'predicted')

modelEval |>
sjmisc::flat_table()

modelEval |>
  filter(!is.na(predicted)) |>
  mutate(test = actual == predicted, na.rm = TRUE) |>
  summarise(Accuracy = mean(test))
```

13 Clusteranalyse

Mit der Clusteranalyse wird versucht, Elemente (Fälle) nach ihren Merkmalen (mehrere Variablen) in Gruppen (Clustern) zusammenzufassen. Die Problemstellung lautet als: Wie können Fälle in einem Datensatz nach mehreren Variablen gruppiert werden?

Die Clusteranalyse gehört zu den explorativen Verfahren. Im Kontext des Machine Learnings (ML) wird sie auch zu den «unsupervised learning»-Verfahren gezählt.

Das Grundsätzliche Vorgehen ist bei allen Arten von Clusteranalysen gleich bzw. ähnlich: Wir suchen Gruppen (Cluster) von Fällen, die sich untereinander so stark wie möglich ähneln (homoge Cluster) und so stark von den anderen Gruppen unterscheiden wie möglich. Es geht also um Segmentierung anhand von Mustern in den Daten und nicht um Sortierung anhand vorgegebener Kategorien. Wir wollen also stark Vereinfachen (wenige Cluster), aber auch wenig Heterogenität in den Clustern, wobei die eben immer kleiner wird, je mehr Cluster man bildet. Das führt zu einem Optimierungsproblem (siehe Abbildung 13.1), das die Clusteranalyse ganz gut lösen kann, da in der Regel die Heterogenität innerhalb der Cluster stark sinkt, wenn man ein Cluster in zwei aufteilt und auch dann, wenn man drei Cluster versucht zu finden und zu optimieren, aber schon etwas weniger als von eins auf zwei. Wenn man statt drei, vier Cluster zulässt und verteilt, dann sinkt die Heterogenität nochmals weiter, aber wieder etwas weniger als im Schritt davor. Das schauen wir weiter unten nochmal technisch an, wenn es um den «Heterogenitätsknick» geht, der als «Ellenbogen» gesehen werden und damit als Kriterium für eine gute Clusterlösung herhalten kann.

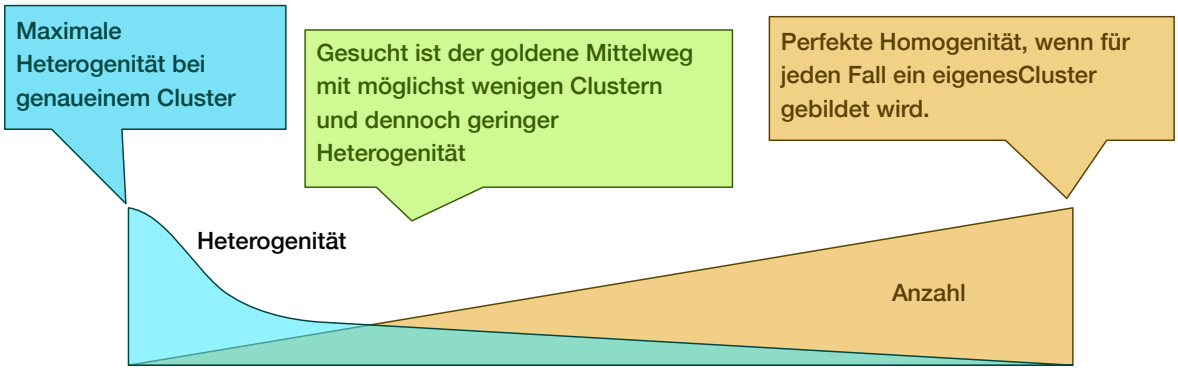


Abbildung 13.1: Optimierungsproblem der Clusteranalyse

13.1 Voraussetzungen von Clusteranalysen generell

1. Nicht zu viele fehlende Werte, da fehlende Werte die Clusterbildung verzerren.

2. Das Skalenniveau spielt grundsätzlich keine Rolle, da Methoden der Clusteranalyse gibt (hierarchische Clusteranalyse), die auch mit kategorialen (nominale mit mehreren Ausprägungen) Variablen und ordinalen Variablen gut umgehen kann. Bei Clusteranalysemethoden mit Distanzmassen (wie k-Means-Clustering) müssen die Distanzen interpretierbar sein und mithin metrisch skaliert, was aber auch von den Dummys erfüllt wird.

3. Die Fallzahl sollte nicht zu klein sein. Es werden für brauchbare Clusteranalysen ordentliche Fallzahlen benötigt. Vor allem gilt das, wenn einzelne Variablenkombinationen (wichtig für die Gruppenbildung) dünn besetzt sind.

4. Die Variablen sollten ähnlich skaliert sein, damit nicht eine Variable mit einem deutlich grösserem Gewicht in die Clusteranalyse eingeht, nur weil sie breiter skaliert ist. Bei sehr unterschiedlichen Skalierungen empfiehlt sich eine vorherige Standardisierung (z-Transformation bzw. scale) der Variablen (bei vorheriger Faktorenanalyse ist das schon gegeben, weil Faktoren immer standardisiert sind ($\overline{x} = 0$, sd $= 1$)).[1]

13.2 Vorgehensweise

1. Als Grundlage von Clusteranalysen dienen Variablen im Datensatz. Man wählt Variablen aus, von denen man annimmt, dass sie charackteristisch sind für die Gruppenbildung, die man anstrebt.

2. Aufbereitung der Variablen:

- Die Variablen für die Clusteranalyse sollten nicht zu stark korrelieren. Sollte das der Fall sein, kann eine explorative Faktorenanalyse durchgeführt werden, um dann die Clusteranalyse mit den Faktoren und separaten Variablen durchzuführen, die eine hohe Uniqueness (kleine Kommunalitäten) in der Faktorenenalyse haben (siehe Kapitel 9).
- Sind die Variablen sehr unterschiedlich skaliert (zB Alter 15-107 und Wahlabsicht 0 und 1), sollte man die Variablen standardisieren (z-transformieren), damit jede Variable in die Clusteranalyse mit demselben Gewicht eingeht (Es sei denn, man möchte einzelen Variablen höher gewichten. In dem Fall empfehle ich, die Variablen nach der Standardisierung mit einem Faktor zu multiplizieren, also zB das standardisierte Alter mit 0.5 und die standardisierte Wahlabsicht mit 2, wenn die höher gewichtet sein soll.)

3. Bevor man zur eigentlichen Clusteranalyse schreitet, muss man ein gutes Mass für die Ähnlichkeit oder die Distanz wählen («Proximitätsmasse»).

4. An dieser Stelle weichen verschiedene Clusteranalysen voneinander ab.

- Bei der **hierarchischen Clusteranalyse** werden an diesem Punkt die Cluster zusammengelegt, also vom Ausgangspunkt, dass jeder Fall ein Cluster ist, bis dahin, dass alle Fälle in einem Cluster zusammenglegt sind. Das passiert schrittweise, indem in jedem Schritt alle Cluster miteinander verglichen werden und jeweils die beiden ähnlichsten zusammengelegt werden.
- Bei der **K-mean-Clusteranalyse** muss an dieser Stelle mit Hilfe eines Screeplots bestimmt und festgelegt werden, wie viele Cluster es geben soll.

5. Da der Schritt 4 für verschiedene Clusteranalysen unterschiedlich ist, ist es auch der Schritt 5.

- Bei der **hierarchischen Clusteranalyse** wird hier die Anzahl Cluster anhand des Dendrogramms (oder alternativ durch einen Screeplot) bestimmt.
- Bei der **K-mean-Clusteranalyse** werden in diesem Schritt die Cluster extrahiert, indem zunächst die Clusterzentren zufällig in den Merkmalsraum (multidimensionales Koordinatensystem aller Variablen in der Analyse) gelegt werden. Dann werden diesen Clusterzentren alle ihnen jeweils nächsten Fälle zugeordnet. Dann werden die Clusterzentren ins Zentrum ihrer Fälle verschoben und dann die Fälle neu zugeordnet. Das passiert so oft, bis die Clusterzentren nicht mehr ihre Position verändern (und alle Fälle in ihren Clustern verbleiben).

6. Die Clusterzuorndung wird in den Daten gespeichert.

7. Jetzt werden die Cluster charakterisiert (interpretiert), indem geschaut wird, wie die Variablen in den verschiedenen Clustern aussehen, welche Mittelwerte sie zum Beispiel haben oder wie hoch die Prozentanteile einzelner Werte (1 bei Dummys) sind.

[1] Wenn man einzelne Variablen besonders wichtig findet in der Clusteranalyse, kann man sie erstmal standardisieren und dann mit einem Faktor multiplizieren, damit sie um den Faktor mit höherem Gewicht in die Analyse eingehen.

13.3 Proximitätsmasse

Diese «Proximitätsmasse» können folgende Distanzmasse sein: euklidische Distanzen bei metrischen Variablen:

> Die Distanz d ist die Wurzel der Summe aller quadrierten Abstände in den Richtungen der Variablendimensionen: $d = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$. Das kommt aus der euklidischen Geometrie und kennen Sie sicher aus der Formel für die Seitenlängen in rechtwinkligen Dreiecken mit $a^2 + b^2 = c^2$, also Ankathete zum Quadrat plus Gegenkathete zum Quadrat ist gleich der Hypothenuse zum Quadrat. Das nur schnell als Antwort auf ihre alte Frage aus Schulzeiten, wozu sie das wohl jemals brauchen würden. Siehe auch Abbildung 13.2.

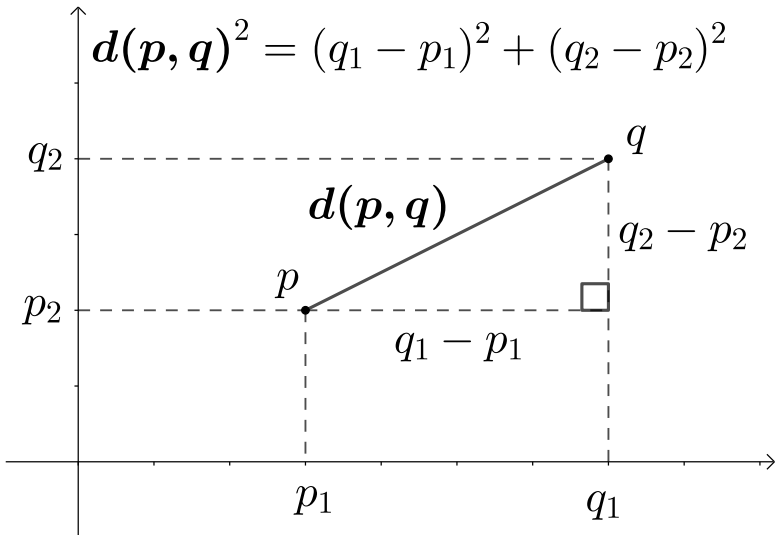


Abbildung 13.2: Euklidische Distanz

Es gibt weitere Distanzmasse auch für metrische Variablen. Bei dichotomen Variablen gibt es noch den M-Koeffizienten, der schlicht die Übereinstimmungen wiedergibt (also in allen Variablen eine 1 oder in allen Variablen eine 0 oder nur in Teilen usw.).

Neben diesen Massen für die Distanz bzw. Nähe im geometrischen Sinne, gibt es noch Ähnlichkeitsmasse. Dazu zählt zum Beispiel der Q-Korrelationskoeffizient, der dasselbe ist, wie Pearsons Korrelationskoeffizient. Die Korrelationen können bei stetigen Variablen verwendet werden. Bei kategoriellen Merkmalen kann χ^2 verwendet werden. Eine kleine Systematik der Clusteranalyse findet sich in Abbildung 13.3.

Dann muss noch der Cluster-Algorithmus gewählt werden. Es gibt die «Hierarchische Clusteranalyse» mit «Single-Linkage» und mit «Complete Linkage». Für metrische Variablen können partionierende Clusteranalysen eingesetzt werden, wie der k-Means-Algorithmus oder der Two-Stage-Algorithmus. Die einfachste Methode ist im Grunde die k-Means-Cluster-Methode, die daher hier als erste etwas genauer angeschaut und in R berechnet werden sollte.

Nehmen Sie also den Code und probieren Sie es aus. So können Sie praktisch üben.

13.4 Die k-Means-Cluster-Methode

In der schrittweisen «Animation» der k-Means-Clusterung sieht man, wenn man genau hinschaut, wie die Cluster am Anfang zufällig verteilt werden, dann alle Fälle, den ihnen am nächsten gelegenen Clusterzentrum zugeordnet werden und die «Clusterzentren» eigentlich erst dann in das Zentrum ihres Clusters gelegt werden. Dann kann es vorkommen, dass einzelne Fälle dichter an einem anderen Cluster liegen und werden

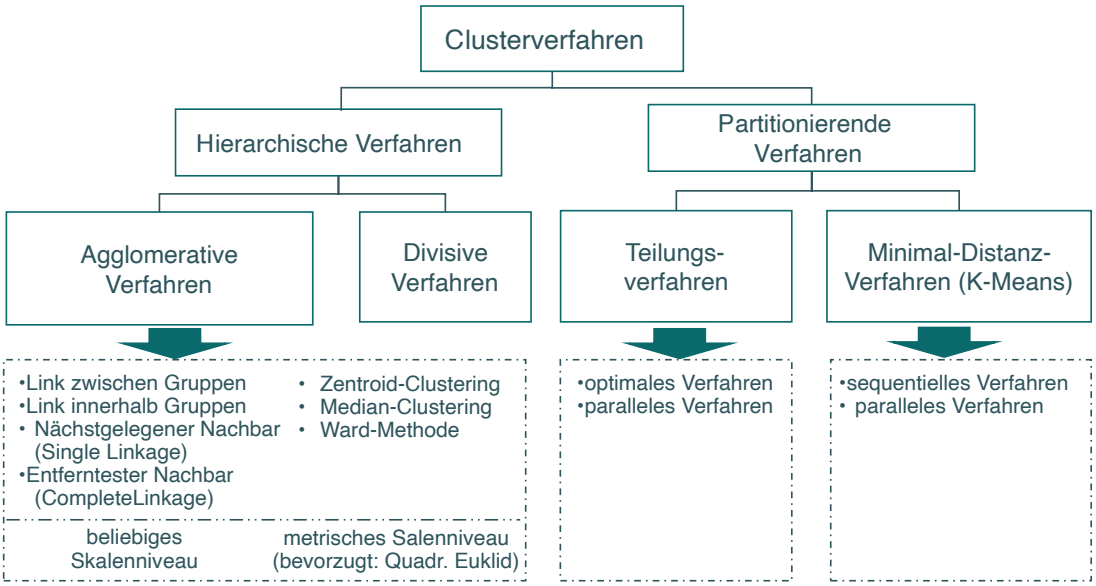


Abbildung 13.3: Cluster-Analyse-Systematik

deshalb eben dem Cluster zugeordnet, in dessen Nähe sie liegen. Danach liegen die Clusterzentren wieder nicht mehr genau im Zentrum ihres eigenen Clusters und werden erneut so verschoben, dass sie genau in dessen Mitte liegen. Das «konvergiert» dann in der Regel irgendwann. Das heisst, bei jedem neuen Schritt würde kein Fall sein Cluster wechseln und darum auch die Clusterzentren unbewegt da bleiben wo sie sind. Das ist dann die Lösung!

Mit diesen Befehlen kann man eine schöne Clusteranalyse laufen lassen.

```
DATEN <- iris |>
  select(-Species)

# Compute k-means with k = 3
set.seed(123)
res.km <- kmeans(scale(DATEN), 3, nstart = 25)

# Dimension reduction using PCA
res.pca <- prcomp(DATEN,  scale = TRUE)

# Coordinates of individuals
ind.coord <- as.data.frame(factoextra::get_pca_ind(res.pca)$coord)
# Add clusters obtained using the K-means algorithm
ind.coord$cluster <- factor(res.km$cluster)
# Add Species groups from the original data sett
ind.coord$Species <- iris$Species
# Data inspection
# head(ind.coord)

# Percentage of variance explained by dimensions
eigenvalue <- round(factoextra::get_eigenvalue(res.pca), 1)
```

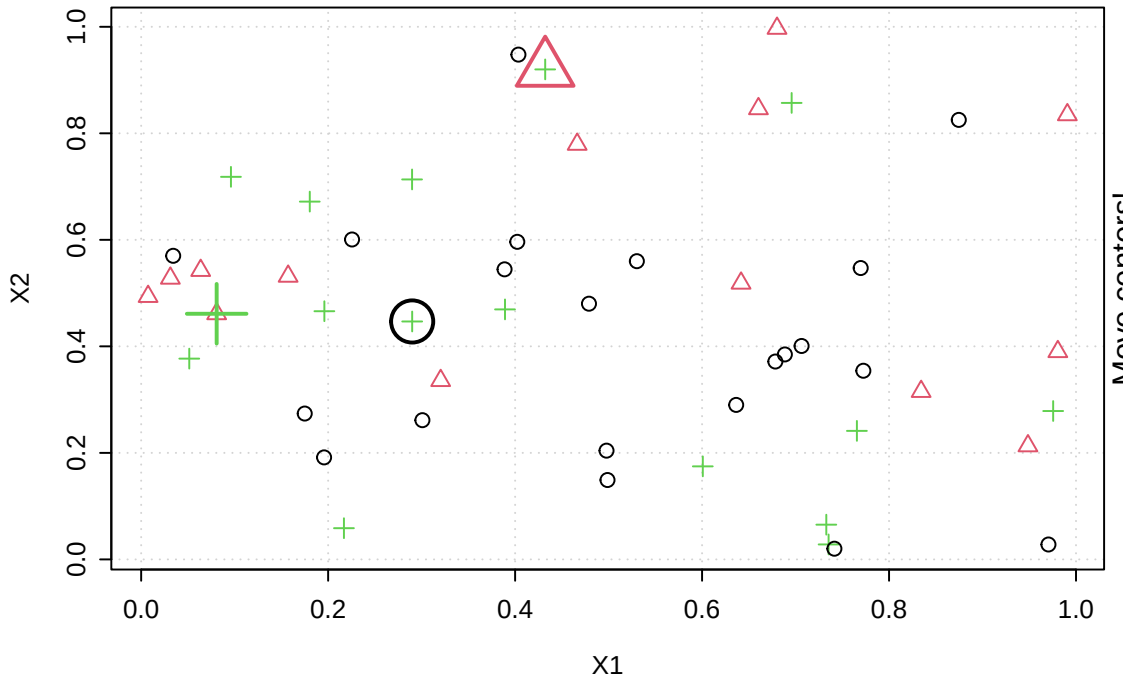


Abbildung 13.4: k-means-Cluster Prinzip

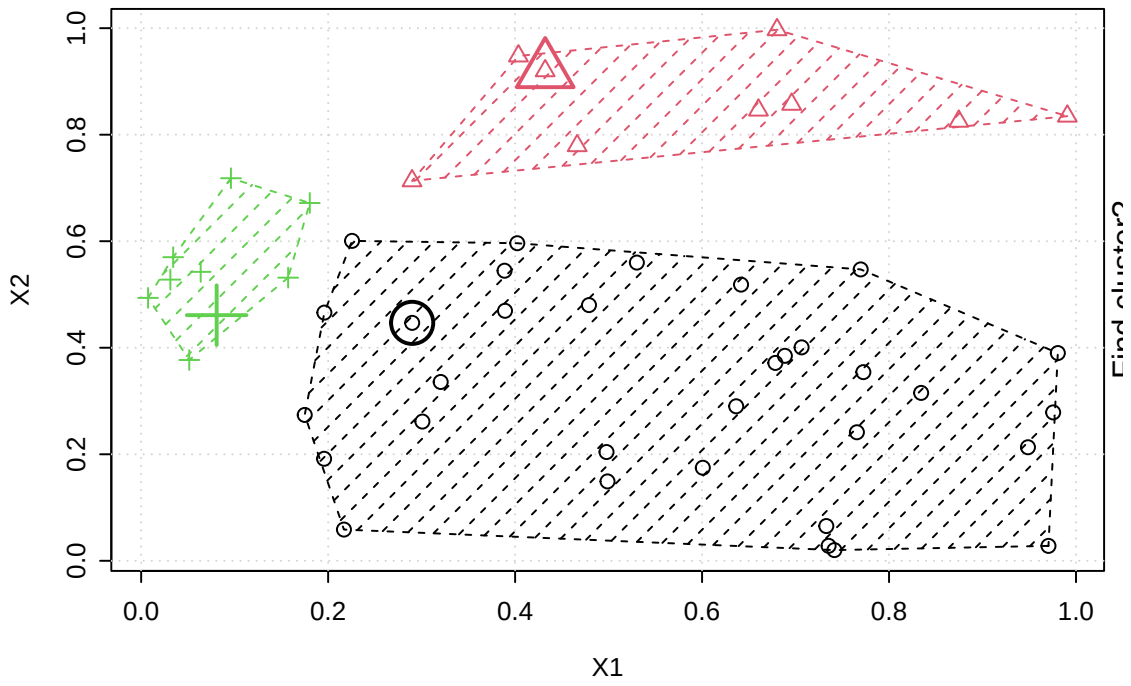


Abbildung 13.5: k-means-Cluster Prinzip

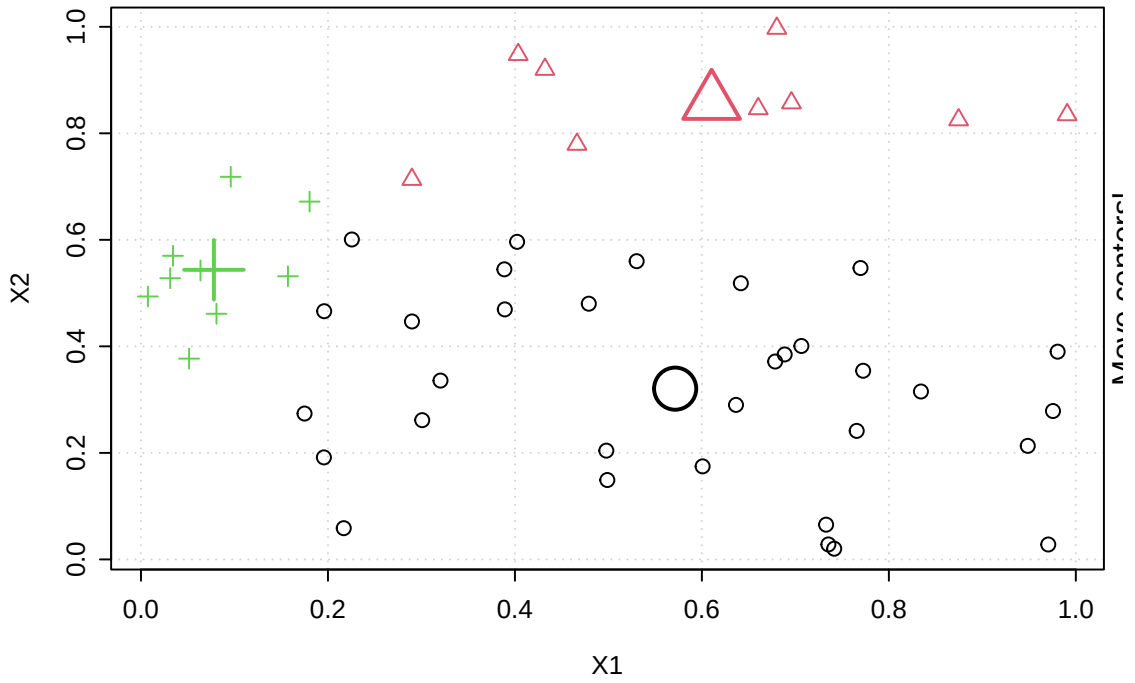


Abbildung 13.6: k-means-Cluster Prinzip

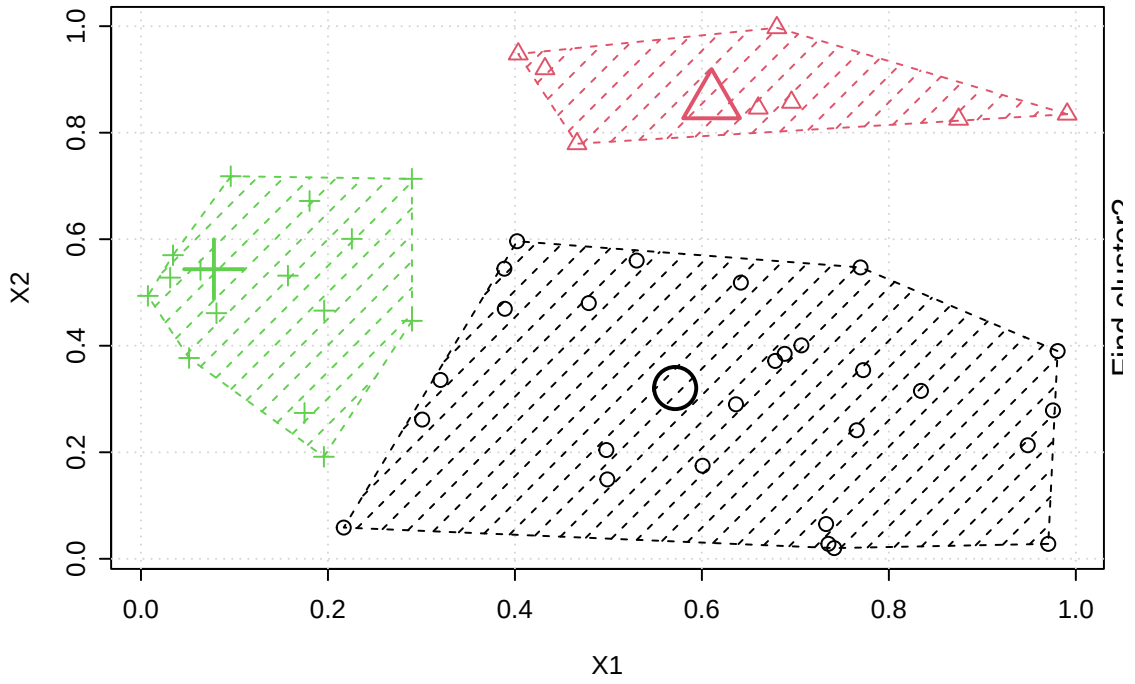


Abbildung 13.7: k-means-Cluster Prinzip

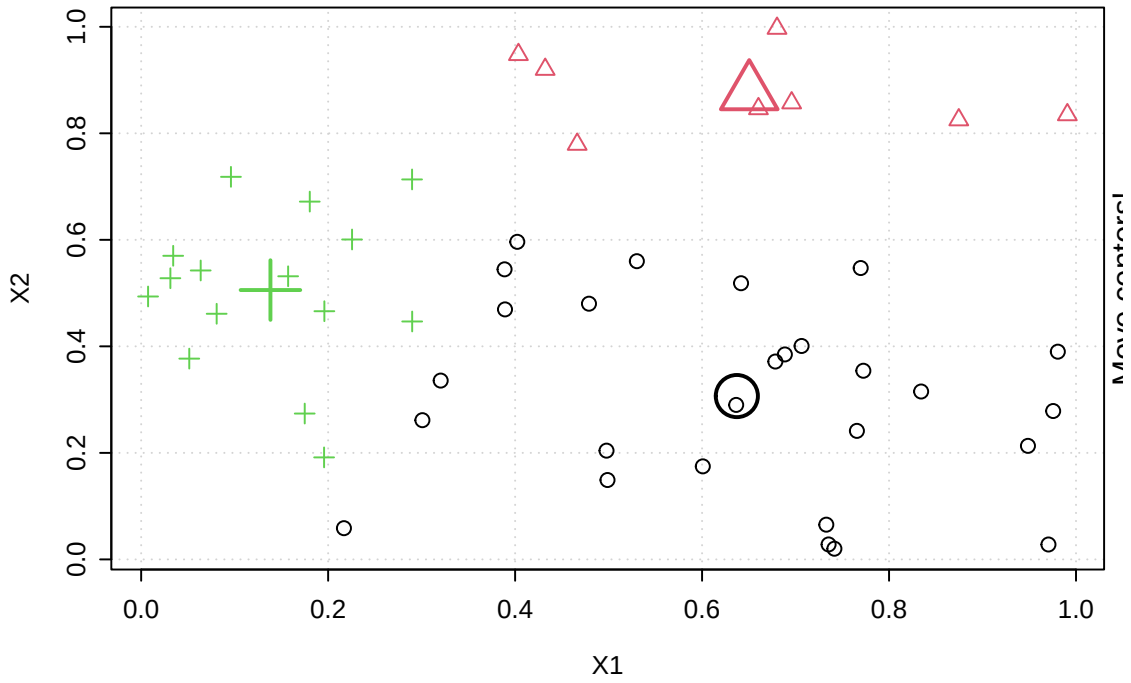


Abbildung 13.8: k-means-Cluster Prinzip

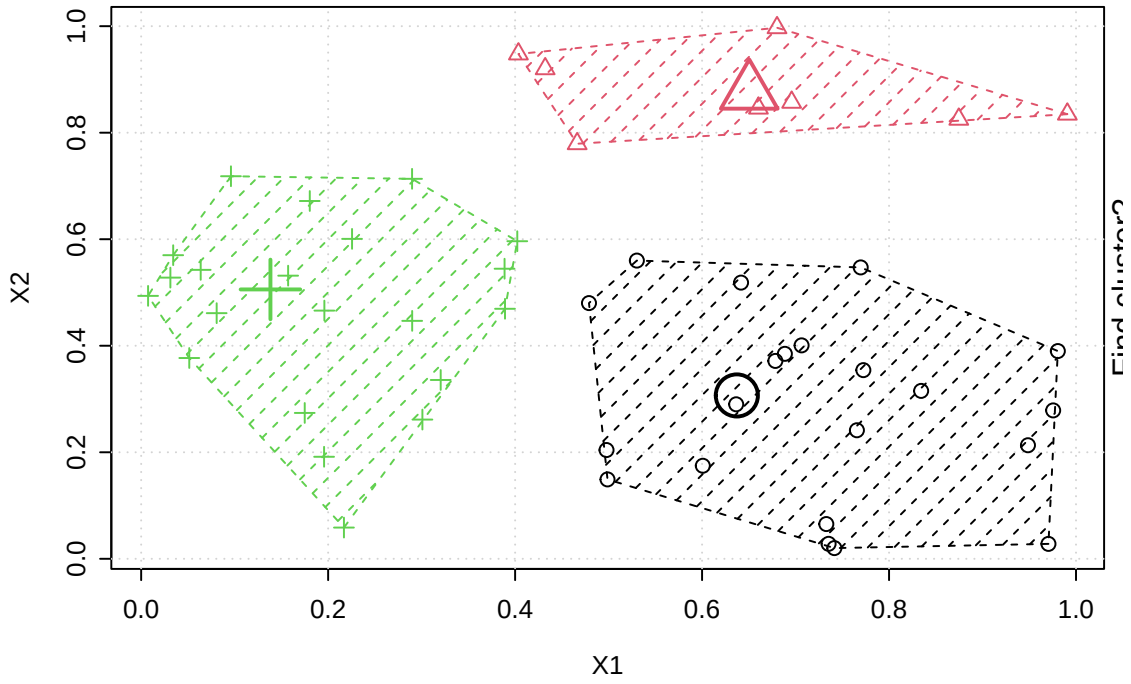
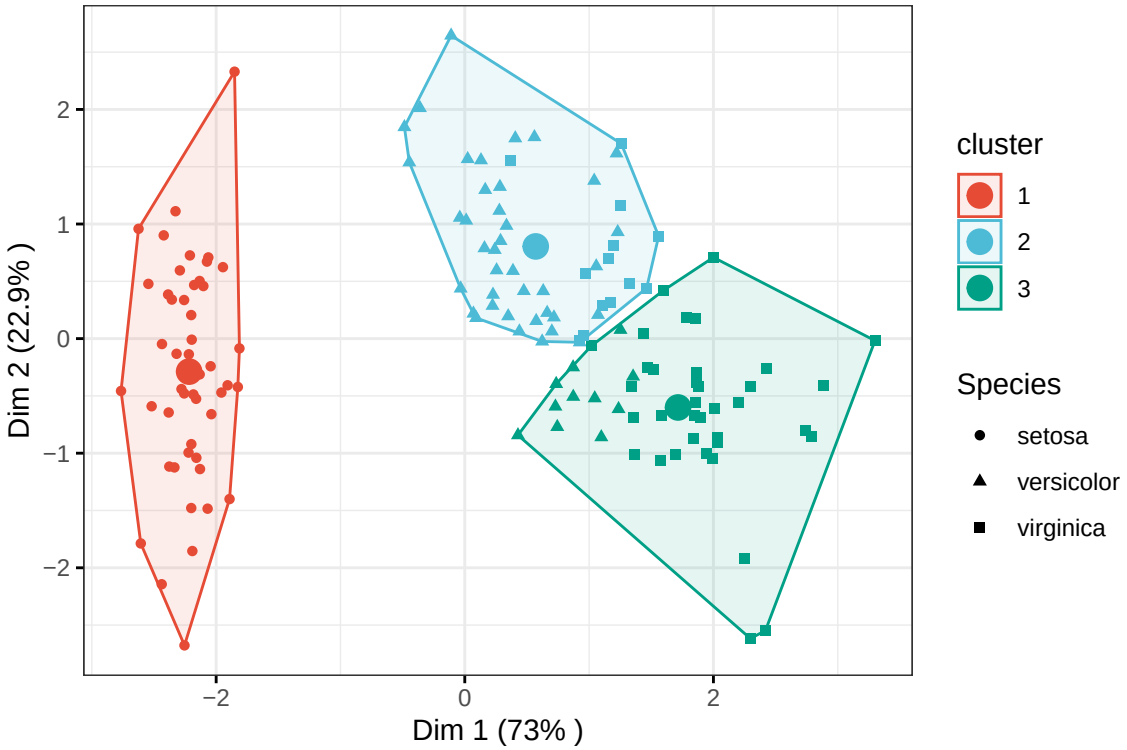


Abbildung 13.9: k-means-Cluster Prinzip

```
variance.percent <- eigenvalue$variance.percent

# head(eigenvalue)

ggpubr::ggscatter(
  ind.coord, x = "Dim.1", y = "Dim.2",
  color = "cluster", palette = "npg", ellipse = TRUE, ellipse.type = "convex",
  shape = "Species", size = 1.5,  legend = "right", ggtheme = theme_bw(),
  xlab = paste0("Dim 1 (", variance.percent[1], "% )" ),
  ylab = paste0("Dim 2 (", variance.percent[2], "% )" )
) +
  ggpubr::stat_mean(aes(color = cluster), size = 4)
```



14 Fokussierte Zusammenfassung

Die Folien zur Sitzung

Vodcast Zusammenfassung

https://www.youtube.com/embed/gz0x-WSROe4?si=wb3Bd6qydbgmw0sS (Unter dem Video auf Youtube.com finden Sie das Inhaltsverzeichnis ebenfalls verlinkt.)

Einstieg
00:08 Hinweise zur Prüfung
12:06 - Intro und Überblick über die erlernten Verfahren
Bivariates
25:28 - Covarianz und Korrelation
29:16 - Regression bivariat
34:01 - b's (bivariat)
36:50 - BETAs bivariat
42:32 - Standardfehler der b's bivariat
46:35 - R^2 Modellgüte bivariat
54:47 - t-Tests der b's bivariat
57:02 - R-Output Regression bivariat
Multivariates lineares Modell (GLM)
01:01:51 - Multivariate Regression
01:07:29 - OLS
01:08:54 - b's multivariat
Voraussetzungen OLS
01:14:23 - Voraussetzungen OLS für BLUE
01:20:11 - Multikollinearität
01:26:19 - Lineartransformation
01:27:45 - Heteroskedastizität
01:30:53 - Residualverteilung
Dummys als UV
01:32:28 - Dummy UVs
01:42:07 - Slope-Dummys
Interaktionen
01:52:25 - Interaktion zweier metrischer
Weitere Analysemethoden
01:54:15 - Faktorenanalysen
02:09:43 - Logistische Regression
02:18:15 - Machine Learning ML
02:20:11 - Clusteranalysen
02:35:49 - Nächste Woche

Die verschiedenen Analysemethoden, die Sie in diesem Semester kennengelernt haben, ermöglichen es, Daten aus unterschiedlichen Blickwinkeln zu analysieren. Man kann also mit denselben Variablen eine Zusammenhangsanalyse machen oder sie auf Unterschiede hin analysieren oder schauen, ob es Interdependenzen gibt, sie als Gruppen bilden. Die zugrundeliegenden Beziehungen in den Daten sind natürlich immer dieselben. Das liegt daran, dass Unterschiede durch Zusammenhänge entstehen und Zusammenhänge aufgrund von Unterschieden. Beides finden seine Ursache darin, dass Variablen und Fälle Gruppen bilden; und

gleichzeitig entstehen die Gruppen durch die Zusammenhänge und Unterschiede.

Die Kennwerte, die aufgrund von Unterschiedsanalysen entstehen sind nicht sehr hoch verdichtet. Daher sind sie leichter zu lernen und für den Einstieg in die Statistik gut geeignet. Sie haben bereits Unterschiedsanalysen kennengelernt, die Masse (gesprochen Maße :-) der zentralen Tendenz auswerten, also zum Beispiel den t-Test für Mittelwertunterschiede zwischen zwei Gruppen. Wir können dabei Variablen aus verschiedenen Teilstichproben (Gruppe der Wähler:innen und Nichtwähler:innen) untersuchen, also «unabhängige Stichproben». Oder wir untersuchen «verbundene Stichproben», wenn zum Beispiel die Mittelwerte von zwei Variablen verglichen werden sollen, die jeweils für die ganze Stichprobe erhoben wurden (zB vor und nach einem experimentellem Eingriff aka Treatment). Oder wir untersuchen die Varianzen von Variablen mit Hilfe von χ^2 oder einem F-Test.

Wenn Sie genau auf die Grafik schauen, finden Sie den χ^2 -Test einmal bei den Unterschieden und einmal bei den «bivariaten» Zusammenhangsanalysen. Das liegt an der oben angesprochenen Verbundenheit der Konzepte: Unterschiede entstehen, wenn Dinge miteinander zusmamenhängen. Bei den Zusammenhangsanyalysen unterscheiden wir die «bivariten» von den «multivariten Modellen». Die bivariten bringen nur zwei Variablen in Beziehung zueinander, was sie einfacher macht, aber im Grunde zu einfach, um die komplexeren Zusammenhänge in unserer Welt zu erklären. Menschen sind einfach nicht bivariat und unsere Welt ist nicht monokausal. Die multivariaten Modelle sind Erweiterungen der bivariaten Analysemethoden. Bei den «Generalisierten Linearen Modellen» (GLM) geht es also weiter. Analysestrategien der GLM werden nach den Skalenniveaus der Variablen unterschieden, die erklärt werden sollen (also die abhängigen Variablen aka AV) und nach den Skalenniveaus der erklärenden (unabhängigen Variablen aka UV).

Die Analysemethoden sind dann einfacher, wenn das Skalenniveau hoch ist. Darum machen wir den Einstieg auch mit der Regression, bei der die AV und die UVs metrisch sind. Wenn die UVs nominal sind (bzw. nominale vorkommen), wird oft auch von Varianzanalysen (Analysis of Variance aka ANOVA) gesprochen. Wenn die AV nominal ist (dichotom oder polytom) werden logistische Regressionen gerechnet. Wenn Sie nach dem Bachelorstudium mit dem Master weitermachen, lernen Sie die multivariaten Analysemethoden auf dem «Next Level» kennen – also zumindest einige davon. Wenn Sie dann auch noch in die Wissenschaft weitergehen, befassen Sie sich sicher spezialisierter mit bestimmten Verfahren der statistischen Datenanalyse, die für Ihre Forschung die am besten geeignete ist.

In diesem Semester werden wir uns auch mit Verfahren befassen, die Gruppierungen (aka Interdependenzen) untersuchen. Dazu gehört an erster Stelle die Faktorenanalyse, mit deren Hilfe Faktoren extrahiert werden sollen, die – so die Vermutung – die gemeinsame Ursache für gemessene Variablen sind. Die Idee ist also, dass manifest gemessene Variablen aufgrund von latenten Variablen miteinander zusammenhängen beziehungsweise korrelieren. Das ist schon an sich interessant genug. Darüber hinausgehend, können wir mit Hilfe einer Faktorenanalyse Indizes bauen, die mehrere Variablen auf einmal abbilden. Während die Faktorenanalyse Eigenschaften von Fällen auf zugrundeliegende Gemeinsamkeiten hin untersucht, werden mit Clusteranalysen Fallgruppen gebildet. Zum Beispiel könnten wir untersuchen, ob die Begeisterung und Abneigung gegenüber Mathemaitik, Statistik, Computer-Programmierung, R usw. einen gemeinsamen Kern haben, wie schlechter Matheunterricht oder Identitiätsbildung. Und dann könnten wir mit Clusteranalysen Gruppen identifizieren, je nachdem, wie gross die Begeisterung für Mathe is, für Computer und für Programmiersprachen wie R. Da gibt es sicher die einen und die anderen. Solche, die tollen Matheunterricht hatten und trotzdem mit R auf Kriegsfuss stehen usw. Also, Sie sehen, wir können viel damit anstellen. Das lohnt sich, auch wenn der Weg teils beschwerlich ist.

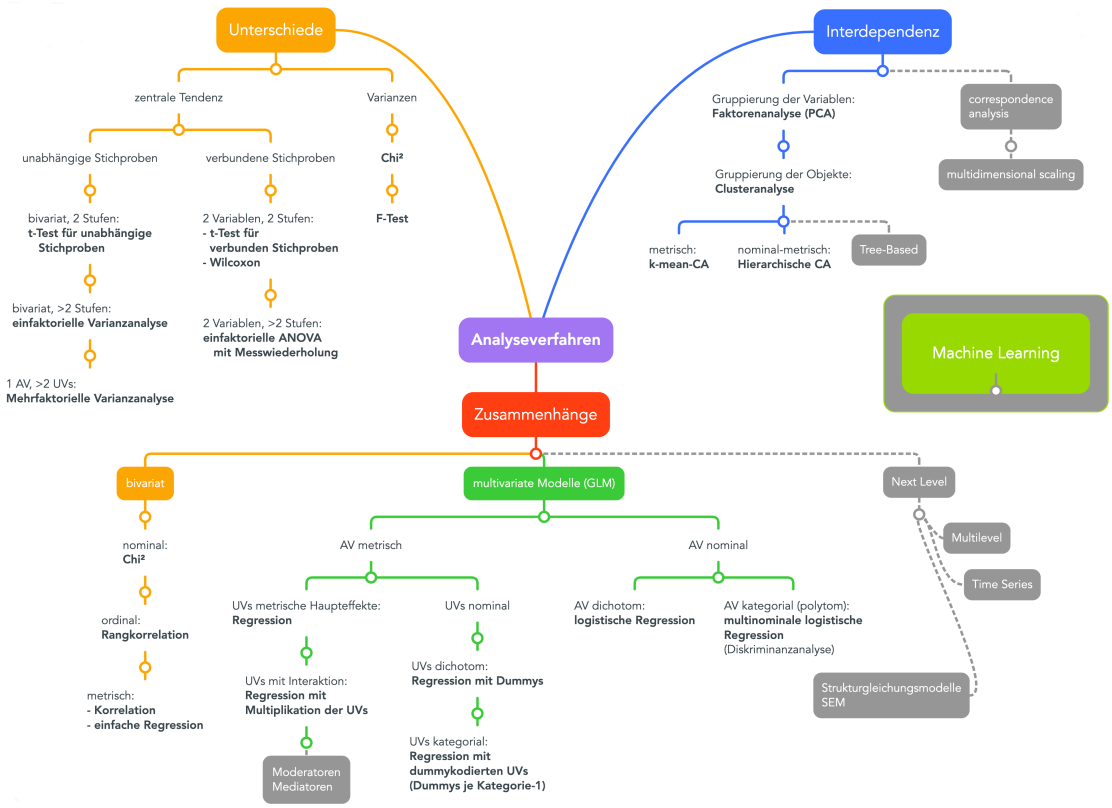


Abbildung 14.1: Systematik gesamt

15 Besprechung der LEF

Formelsammlung

Univariat

Mittelwert $\bar{x}$ und μ

- μ ist der Parameter «Mittelwert» in der Grundgesamtheit GG, den wir nicht kennen
- $\overline{x}$ ist der Kennwert «Mittelwert» in der Stichprobe

$$\mu = \frac{1}{N} \sum_{i}^{N} (x_i) \tag{15.1}$$

$$\overline{x} = \frac{1}{n} \sum_{i}^{n} (x_i) \tag{15.2}$$

Varianz und Standardabweichung

$$s^2 = V = \frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})^2 \tag{15.3}$$

$$s = \sqrt{\frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})^2} \tag{15.4}$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i}^{n} (x_i - \overline{x})^2 \tag{15.5}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i}^{n} (x_i - \overline{x})^2} \tag{15.6}$$

(15.7)

Standardfehler

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} \tag{15.8}$$

z-Transformation

$$z_i = \frac{x_i - \overline{x}}{s} \tag{15.9}$$

Konfidenzintervalle der Mittelwerte

$$\text{KI:}\overline{X} \pm z_1 \cdot se \tag{15.10}$$

$$\text{KI:}\overline{X} \pm z_1 \cdot \frac{s_x}{\sqrt{n}} \tag{15.11}$$

$$\text{KI}_{l.05} = \overline{x} - 1.96 \cdot \frac{s_x}{\sqrt{n}} \tag{15.12}$$

$$\text{KI}_{r.05} = \overline{x} + 1.96 \cdot \frac{s_x}{\sqrt{n}} \tag{15.13}$$

Bivariat

Kovarianz und Korrelation

$$cov = C = \frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{15.14}$$

$$r = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{n \cdot s_x \cdot s_y} \tag{15.15}$$

Bivariate Regression

$$Y_i = \overline{Y} + e_i \tag{15.16}$$

$$Y_i = b_1 + b_2 X_i + e_i \tag{15.17}$$

$$\hat{Y}_i = b_1 + b_2 X_i \tag{15.18}$$

$$Y_i = \hat{Y}_i + e_i \tag{15.19}$$

Multivariate Regression

Das Basismodell

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i$$

$$Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i$$

Die b's

$$b_2 = \frac{r_{y2} - r_{23} r_{y3}}{(1 - R^2_{2.3})} \frac{s_y}{s_2} \tag{15.20}$$

BETAs

$$BETA = b \cdot \frac{s_X}{s_Y} = r_{YX} \tag{15.21}$$

R^2

$$R^2 = \frac{SS_M/n}{SS_T/n} \tag{15.22}$$

$R^2_{adj.}$

$R^2_{adj.} = R^2 \cdot \frac{n-k-1}{n-1}$

Die Residuen

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{i2} - b_3 X_{i3} \tag{15.23}$$

Standardfehler der Regressionskoeffizienten b

$$s^2_{b_2} = \frac{s^2}{n} \frac{1/s^2_2}{1 - R^2_{2.3}} \tag{15.24}$$

mit s^2:

$$s^2 = \frac{1}{n-3} \sum (e_i - \bar{e})^2 = \frac{1}{n-3} \sum e_i^2 \tag{15.25}$$

$$s^2_{b_3} = \frac{s^2}{n} \frac{1/V_3}{1 - r^2_{23}} \tag{15.26}$$

Erwartungstreue b_2

$$E(b_2) = \beta_2 + \frac{V_3 E(C_{2U})}{D} - \frac{C_{23} E(C_{3U})}{D}, \tag{15.27}$$

Toleranz (TOL)

$$TOL_{b_2} = 1 - R^2_{2.3} \tag{15.28}$$

$$TOL_{b_3} = 1 - R^2_{3.2} \tag{15.29}$$

Varianz-Inflationsfaktor (VIF)

$$VIF_{b_2} = \frac{1}{TOL_{b_2}} = \frac{1}{1 - R^2_{2.3}} \tag{15.30}$$

$$VIF_{b_3} = \frac{1}{TOL_{b_3}} = \frac{1}{1 - R^2_{3.2}} \tag{15.31}$$

Glossar

Sitzung	Inhalt	Deutsch	Englisch
4	Statistiken mit R, R-Studio	AV	dependent variable (dv)
3	Voraussetzungen	Allgemeine kleinste Quadrate	Generalized Least Squares (GLS)
1	Wiederholung	Auswahlgesamtheit	sampling population
3	Voraussetzungen	BLUE	BLUE
3	Voraussetzungen	Bias	Bias
10	Clusteranalyse	Cluster	cluster
10	Clusteranalyse	Clusteranalyse (CA)	cluster analysis
10	Clusteranalyse	Clusterzentren	cluster center
1	Wiederholung	Codebuch	codebook, CB
7	EFA, CFA, Lesekompetenz	Cronbachs α	Cronbachs α
8	ML, Logit-Modelle, Multinominal	Datensplit	data splitting
6	Faktorenanalysen	Dimensionsreduktion	dimensional reduction
10	Clusteranalyse	Distanzmasse	distance metrics
5	Kategoriale im linearen Modell	Dummy	dummy
3	Voraussetzungen	Effizienz	efficiency
6	Faktorenanalysen	Eigenwerte	eigenvalues
10	Clusteranalyse	Ellbogenkriterium	elbow criterion
10	Clusteranalyse	Euklidische Distanz	euclidean distance
8	ML, Logit-Modelle, Multinominal	Exponentielle Beta	exponential beta
6	Faktorenanalysen	Faktorenanalyse	factor analysis
6	Faktorenanalysen	Faktorladung	factor loadings
6	Faktorenanalysen	Faktorlösung	Solution
6	Faktorenanalysen	Faktorrotation	factor rotation
8	ML, Logit-Modelle, Multinominal	Feature Engineering	Feature Engineering
3	Voraussetzungen	Fehlervarianz	error variance
8	ML, Logit-Modelle, Multinominal	Generalisiertes Lineares Modell	Generalized Linear Model (GLM)
1	Wiederholung	Grundgesamtheit	population
6	Faktorenanalysen	Hauptkomponentenanalyse	Principle Component Analyisis (PCA)
10	Clusteranalyse	Heterogenität	heterogeneity
3	Voraussetzungen	Heteroskedastizität	heteroscedasticity
10	Clusteranalyse	Hierarchische Clusteranalyse	hierarchical cluster analysis
10	Clusteranalyse	Homogenität	homogeneity
3	Voraussetzungen	Homoskedastizität	homoscedasticity
1	Wiederholung	Hypothese	hypothesis
2	Regression	Hypothesentest	hypothesis testing
1	Wiederholung	Kennwerte	characteristic
6	Faktorenanalysen	Kommunalitäten	communalities
2	Regression	Konfidenzintervall	confidence interval
2	Regression	Korrelation	correlation
2	Regression	Kovarianz	covariance
5	Kategoriale im linearen Modell	Kovariate	covariates
2	Regression	Lineares Modell	linear model (lm)
8	ML, Logit-Modelle, Multinominal	Logistische Regression	logistic regression
8	ML, Logit-Modelle, Multinominal	Machine Learning (ML)	machine learning (ML)
8	ML, Logit-Modelle, Multinominal	Machine Learning, reinforcement	machine learning, reinforcement
8	ML, Logit-Modelle, Multinominal	Machine Learning, supervised	machine learning, supervised
8	ML, Logit-Modelle, Multinominal	Machine Learning, unsupervised	machine learning, unsupervised
4	Statistiken mit R, R-Studio	Markdown	Markdown
1	Wiederholung	Messniveau (Skalenniveau)	level of measurement
1	Wiederholung	Mittelwert	mean (value)
2	Regression	Mittlerer quadratischer Fehler	mean squared error
4	Statistiken mit R, R-Studio	Modell	model
3	Voraussetzungen	Modellspezifikation	model specification

(continued)

Sitzung	Inhalt	Deutsch	Englisch
3	Voraussetzungen	Multikollinearität	multicollinearity
8	ML, Logit-Modelle, Multinominal	Multinominale Regression	multinominal logistic Regression
2	Regression	Multivariate Regression	multivariate regression
2	Regression	Nullhypothese	null hypothesis
10	Clusteranalyse	Nächster Nachbar	single linkage
2	Regression	Parameter	parameter
10	Clusteranalyse	Proximitätsmass	proximity
2	Regression	Quadratsumme	sum of squares
4	Statistiken mit R, R-Studio	Quarto	Quarto
8	ML, Logit-Modelle, Multinominal	Quote	Odds Ratio (OR)
4	Statistiken mit R, R-Studio	R-Chunks	R-Chunks
2	Regression	Regression	regression
2	Regression	Regressionsgerade	regression line
4	Statistiken mit R, R-Studio	Regressionskoeffizient	regression coefficient
1	Wiederholung	Schätzen	estimate
6	Faktorenanalysen	Scree Plot	Scree Plot
1	Wiederholung	Signifikanz	significance
1	Wiederholung	Skala	scale
1	Wiederholung	Skala, Interval-	interval scale
1	Wiederholung	Skala, Nominal-	nominal scale
1	Wiederholung	Skala, Ordinal-	ordinal scale
1	Wiederholung	Skala, Ratio-	ratio scale
1	Wiederholung	Skala, metrische	metric scale
1	Wiederholung	Skalenniveau	level of measurement
7	EFA, CFA, Lesekompetenz	Skalenreliabilität	scale reliability
1	Wiederholung	Standardabweichung	standard deviation
2	Regression	Standardisierung	Standardization
1	Wiederholung	Stichprobe	sample
1	Wiederholung	Stichprobenumfang	sample / sampling size
6	Faktorenanalysen	Strukturgleichungsmodell (SEM)	structural equation model (SEM)
8	ML, Logit-Modelle, Multinominal	Testdaten	test data
1	Wiederholung	Testen	testing
3	Voraussetzungen	Toleranz (TOL)	tolerance
8	ML, Logit-Modelle, Multinominal	Trainingsdaten	train data
6	Faktorenanalysen	Uniqueness	Uniqueness
3	Voraussetzungen	Unterspezifikation	under estimation
8	ML, Logit-Modelle, Multinominal	Unterspezifikation	under fitting
3	Voraussetzungen	Unverzerrtheit	unbiasedness
1	Wiederholung	Variable	variable
1	Wiederholung	Variable, dichotome	dichotomous variable
1	Wiederholung	Variable, polytom	polytomous variable
2	Regression	Varianz	variance
3	Voraussetzungen	Varianzinflationsfaktor (VIF)	variance inflation factor
10	Clusteranalyse	Weitester Nachbar	complete linkage
1	Wiederholung	Zentrale Tendenz	central tendency
1	Wiederholung	Zufallsstichprobe	random sample / sampling
1	Wiederholung	abhängige Variable (AV)	dependent variable
2	Regression	bivariate Regression	bivariate regression
4	Statistiken mit R, R-Studio	dplyr::arrange()	dplyr::arrange()
4	Statistiken mit R, R-Studio	dplyr::filter()	dplyr::filter()
4	Statistiken mit R, R-Studio	dplyr::group_by()	dplyr::group_by()
4	Statistiken mit R, R-Studio	dplyr::mutate()	dplyr::mutate()
4	Statistiken mit R, R-Studio	dplyr::select()	dplyr::select()
4	Statistiken mit R, R-Studio	dplyr::summarise()	dplyr::summarise()
10	Clusteranalyse	explorativ	explorative
6	Faktorenanalysen	explorative Faktorenanalyse	explorative factor analysis (EFA)
10	Clusteranalyse	k-means	k-means
6	Faktorenanalysen	konfirmatorische Faktorenanalyse	confirmatory factor analysis (CFA)
6	Faktorenanalysen	latente Variablen	latent variable
4	Statistiken mit R, R-Studio	lm	lm
10	Clusteranalyse	monothetisch	monothetic

(continued)

Sitzung	Inhalt	Deutsch	Englisch
6	Faktorenanalysen	oblique Rotation	oblique rotation
6	Faktorenanalysen	orthogonale Rotation	orthogonal rotation
10	Clusteranalyse	polythetisch	polythetic
10	Clusteranalyse	Ähnlichkeitsmasse	similarity measure
3	Voraussetzungen	Überspezifikation	over estimation
8	ML, Logit-Modelle, Multinominal	Überspezifikation	over fitting

Literatur