

Statistik mit R für Fortgeschrittene

Walter Gruber

2019-04-15

Contents

	2
Vorbemerkung	3
Motivation	4
Teil I: Variabilitäten und Korrelationen	4
Variabilität	4
Korrelationen	13
Pearson Produkt Moment Korrelation	13
Kausalität	16
Partial- Semipartialkorrelation	16
Korrelationstechniken	19
Lösungen	21
Aufgabe_1	21
Aufgabe_2	23
Teil II: Multiple lineare Regression	24
Multiple Regression	25
Definition	25
Modellvergleich	27
Aufgabe MLR 1	30
Wahl relevanter Prädiktoren	30
Sequentielle Modellbildung	31
Modellvergleich durch AIC	32
Kreuzvalidierung	33
Voraussetzungen MLR	33
Lösungen	35
Aufgabe SLR 1 Lsg	35
Aufgabe MLR 1 Lsg	35
Teil III: Dummy-Codierung	35
Kategoriale Prädiktoren	35
Beispiel mit mehrstufigen kategorialen Prädiktor	36
Dummy Kodierung	36
Modellierung mit kategorialen Variablen	37

Teil IV: Mediator-Analyse	38
Mediation	38
Konzeptuelle Modell	38
Effektgrößen der Mediation	40
Fallbeispiel	40
Teil V: Moderator-Analyse	43
Moderation	43
Konzeptuelles Modell	44
Formale Beschreibung des Modells	46
Zentrierung der Variablen	49
Teil VI: Analysis of Covariance	51
Motivation	51
Kovarianzanalyse	51
Voraussetzungen	52
Berechnung einer ANCOVA	53
Nützliche Graphen	64
Homogenität der Steigung	66
Bericht der Ergebnisse	66



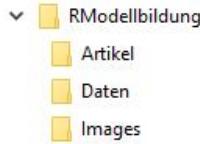


Figure 1: Abbildung 1: Dateistruktur für R-Projekt

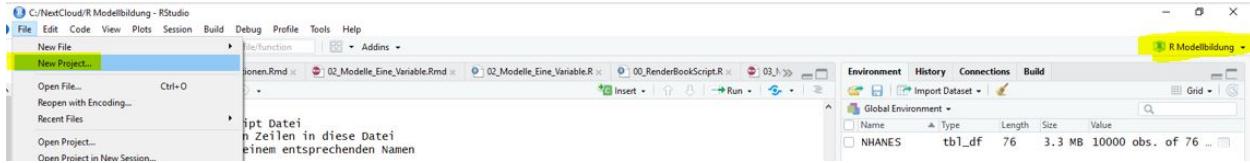


Figure 2: Abbildung 2: R-Projekt definieren

Vorbemerkung

Dieses Skriptum wurde mit dem Paket *bookdown* erstellt. Der verwendete R-Code wird als Teil des Skriptums angeführt und kann auch direkt von diesem Dokument in ein R-Skript übernommen und ausgeführt werden. Erläuterungen zum Code beschränken sich zum Teil auf wesentliche Code-Fragmente. Für detaillierte Angaben zu diversen Funktionen ist die R-Hilfe zu verwenden.

Der nachfolgende Code ist spezifisch für die Erstellung dieses Dokumentes, sowie der Bearbeitung der Beispiele im Kurs von Bedeutung. Es wird in diesem Code-Teil sichergestellt, dass die verwendeten Pakte vorhanden und geladen sind. Daher sollte dieser Code am Anfang jeder neuen R-Datei übernommen werden. Die Vorgehensweise ist:

1. Starten von R-Studio
2. Öffnen einer neuen R-Script Datei
3. Kopiere die nachfolgenden Zeilen in diese Datei
4. Speichere die Datei mit einem entsprechenden Namen
5. Führe diesen Code aus
6. Füge deinen Code nach diesen Zeilen ein

```
# Initialisierung
rm(list = ls())
if (!require("pacman")) install.packages("pacman")
pacman::p_load(corrplot, DAAG, dataMaid, devtools, doBy, DT,
                ggformula, ggplot2, gridExtra, htmlwidgets,
                imager, knitr, labelled, leaps, magick, MASS,
                NHANES, mosaic, mosaicCore, mosaicData, pander,
                pastecs, ppcor, reshape2,
                rockchalk, rpart, rpart.plot)
```

Des Weiteren ist es von Vorteil, zu Beginn einer Auswertung/Datenanalyse mit R eine entsprechende Verzeichnisstruktur im Window-Dateimanager festzulegen und für diese Struktur ein R-Projektfile anzulegen. Die Verzeichnisstruktur richtet sich im Allgemeinen nach der jeweiligen Analyse, folgende Vorgaben haben sich aber bereits schon mehrmals bewährt:

Die Root kann dabei entweder auf der lokalen Festplatte (C:/..) oder einem Server, bzw. Cloud (../NextCloud/R Modellbildung/Images) liegen.

Das Anlegen eines R-Projektes wird im RStudio durchgeführt.

Nachdem bereits eine Verzeichnisstruktur definiert wurde, kann man das Projekt in das bereits definierte Verzeichnis legen (Folge den Schritten die von RStudio vorgegeben werden). Den Vorteil des projektbasierten

Arbeitsens werden wir im Verlauf des Kurses noch näher kennen lernen.

Inhalte, Beispiele und Daten stammen teilweise aus dem Internet, u.a. (Coursera 2018), (DataCamp 2018) und den Büchern (Field 2017), (Bühner 2009) und (Bühner 2017).

Motivation

Modelle werden meist dazu verwendet, um komplexe Sachverhalte zu beschreiben und Erklärungen für deren Wirkungsweise, Ursachen und Zusammenhänge zu finden. In diesem Seminar wollen wir uns mit einer der unzähligen Möglichkeiten der Modellbildung beschäftigen - der einfachen statistischen Modellbildung.

Das Wesentliche und damit gleichzeitig auch das Schwierigste bei der Modellbildung ist die Identifizierung und Zuordnung der einzelnen Bausteine und nicht - wie oft angenommen - die einem statistischen Modell zugrundeliegende Mathematik. Obwohl die mathematischen Grundlagen für die Anwendung bestimmter Modellvorstellungen enorm wichtig und Kenntnisse darüber auch für die Abschätzung der Güte und Gültigkeit eines Modells erforderlich sind, spielen mathematische und formale Details in der Anwendung sehr oft nur eine nebенästhetische Rolle.

In diesem Seminar sollten die Ziele, Einschränkungen, Vor- und Nachteile statistischer Modelle anhand von theoretischen, aber auch praktischen Überlegungen näher gebracht werden. Die Klärung der zentralen Fragen jeder Art von Modellbildung stehen dabei im Vordergrund, d.h.:

- wie kann ich mit *möglichst einfachen Mitteln* die "Wirklichkeit" möglichst gut zu beschreiben?
- wie kann ich beurteilen, ob mein Modell *gut* ist (zumindest im Vergleich mit anderen Modellen)?
- wie kann ich die Wichtigkeit meiner Modellbausteine für sich beurteilen?
- welche Erkenntnisse darf ich aus meinem Modell auf die Wirklichkeit übertragen?

Teil I: Variabilitäten und Korrelationen

Variabilität

Bevor wir uns mit einzelnen Techniken und Verfahren der linearen Modellbildung auseinandersetzen, soll in einem kurzen Exkurs eines der grundlegendsten Prinzipien der statistischen Modellbildung wiederholt und diskutiert werden - die *Varianz* von beobachteten Werten.

Eigentlich ist es die Variabilität von Merkmalen, die statistische Methoden für die Erklärung von Effekten überhaupt erst auf den Plan ruft. Würden Merkmale wie z.B. Leistung einer Person, Persönlichkeitsmerkmale, Wetter, Produktionsgenauigkeit etc. nicht schwanken/variieren, würden wir heute nicht in diesem Raum sitzen und uns mit statistischen Ideen beschäftigen.

Der Begriff Variabilität ist für uns so alltäglich, dass wir ganz selbstverständlich damit umgehen. Doch was steckt wirklich dahinter? Wie können wir Sie nutzen um komplexere Eigenschaften einer Sache oder eines unerklärlichen Phänomens auf die Spur zu kommen?

Betrachten wir zunächst einmal ein sehr einfaches Beispiel. Im nachfolgenden Graphen sind (sehr vereinfacht) mehrere Möglichkeiten dargestellt, wie eine Person mitsamt Hund sich entlang einer Straße bewegt.

Die Daten der Messungen sind in folgender Tabelle gegeben:

Mensch	Hund	MenschD	HundD	KP	ZMensch	ZHund	ZKP
4	3	0.83	1	0.83	0.71	0.71	0.5
5	4	1.83	2	3.66	1.56	1.42	2.22
2	1	-1.17	-1	1.17	-1	-0.71	0.71
3	2	-0.17	0	0	-0.15	0	0

Mensch	Hund	MenschD	HundD	KP	ZMensch	ZHund	ZKP
2	0	-1.17	-2	2.34	-1	-1.42	1.42
3	2	-0.17	0	0	-0.15	0	0

In obiger Tabelle zeigt die Spalte *MenschD* die Differenz jeder Beobachtung zu Mittelwert aller Beobachtungen (für *HundD* gilt dasselbe, eben für die Variable Hund). Die Spalte *KP* zeigt das Kreuzprodukt, also $KP = MenschD \cdot HundD$. Die Spalten *ZMensch* und *ZHund*, sowie die *ZKP* entsprechen den z-Transformierten beobachteten Werten ($z_i^M = (M_i - \bar{M})/sd(M)$ und $z_i^H = (H_i - \bar{H})/sd(H)$).

Statistische Kennwerte für obige Daten (Mittelwert, Varianz und Standardabweichung) sind in folgender Tabelle dargestellt:

	Mensch	Hund	MenschD	HundD	KP	ZMensch	ZHund	ZKP
Mean	3.167	2	-0.003333	0	1.333	-0.005	0	0.8083
Var	1.367	2	1.367	2	2.052	0.9965	1.008	0.7557
SD	1.169	1.414	1.169	1.414	1.433	0.9983	1.004	0.8693

Wenn also ein Hund jeder Bewegung der Person folgt und dabei auch stets denselben Abstand hält, sind deren beobachteten Pfade zwar örtlich gesehen unterschiedlich, aber die Varianz des einen erklärt voll und ganz die Varianz des anderen Pfades. Mit anderen Worten, die beiden Pfade zeigen eine perfekte Kovariation. Formal wird diese *Kovariation* als *durchschnittliche Summe der Kreuzprodukte* ermittelt, also ($M = \text{Mensch}$, $H = \text{Hund}$):

$$cov(M, H) = \frac{\sum_{i=1}^6 (M_i - \bar{M})(H_i - \bar{H})}{N-1}$$

Setzt man die Beispieldaten in diese Berechnungsvorschrift ein, erhält man für die Summe der Kreuzprodukte 8. Die Kovarianz der beobachteten Werte ist (als durchschnittliche Kreuzproduktsumme) somit $cov(M, H) = 1.6$. Die Korrelation berechnet sich dann ganz einfach zu:

$$r(M, H) = \frac{cov(M, H)}{s_M \cdot s_H}$$

Im vorliegenden Beispiel ergibt sich eine Korrelation $r(M, H) = 1$ (also eine positive und perfekte Korrelation).

Der Korrelationskoeffizient hat gegenüber der Kovarianz den Vorteil, dass er durch die Normierung über die Standardabweichungen:

1. einen (einheitenlosen) Wertebereich zwischen $r \in [-1, 1]$ aufweist und damit vergleichbar mit anderen Korrelationswerten wird.
2. Als praktische **Effektgröße** interpretiert werden kann.
3. Das Quadrat des Korrelationskoeffizienten (r^2) Auskunft über die aufgeklärte Varianz gibt. Dieses Maß spielt eine wesentliche Bedeutung sowohl bei der Korrelationsanalyse, als auch bei der multiplen Regression und anderen Verfahren. Häufig findet man die Bezeichnung **Determinationskoeffizient**.

Letztere Eigenschaft sei nochmals anhand des verwendeten Beispiels verdeutlicht:

Im Fall einer perfekten Kovarianz (also 100% Übereinstimmung der Bewegungen von Mensch und Hund), braucht man nur mehr die Bewegung einer Variablen zu wissen (z.B. die des Menschen), um die Bewegungen des Hundes zu bestimmen (erklären). Somit erklärt die Variabilität des Menschen zu 100 % die Variabilität der Bewegungen des Hundes.

Die Beziehung zwischen zwei (intervallskalierten) Variablen lässt sich am besten mit einem Streudiagramm darstellen:

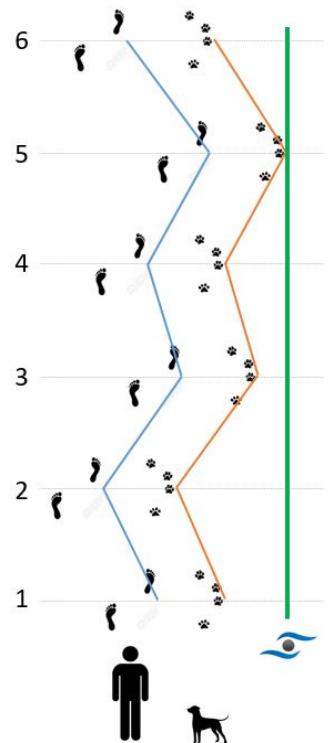


Figure 3: **Abbildung 1:** Gassi gehen mit Blindenhund. Die blaue Linie beschreibt den Weg des Hundehalters, die Orange den des Hundes. Die Grüne Linie ist die Referenzlinie, von welcher aus der Abstand zur jeweiligen Position (Hund und Mensch) zu sechs Beobachtungszeitpunkten gemessen wurde.

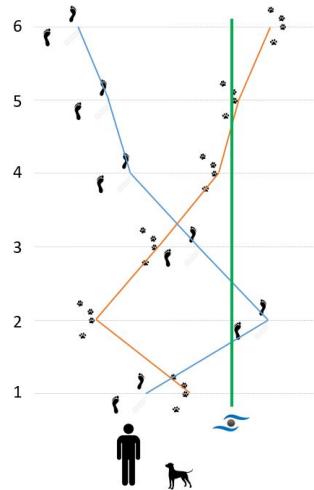
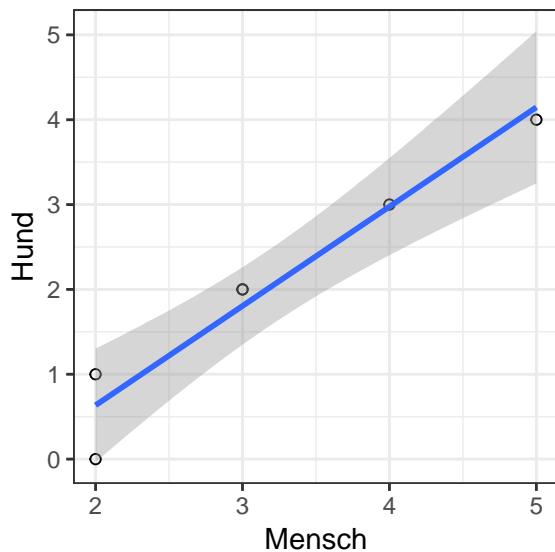


Figure 4: **Abbildung 2:** in diesem Beispiel scheint es sich um eine Hund zu handeln, der bestmöglich das Gegenteil vom Menschen macht. Bestmöglich dahingehen, dass er nicht nur in die genau entgegengesetzte Richtung ausweicht, sondern dabei auch auf den genauen Abstand der Abweichung achtet.



Die folgende Abbildung zeigt ein weiteres Mensch-Hund Beispiel:

In diesem Fall ist die Korrelation auch perfekt, nur eben in die entgegengesetzte Richtung, was zur Folge hat, dass diese Korrelation den Wert $r(M, H) = -1$ zeigt.

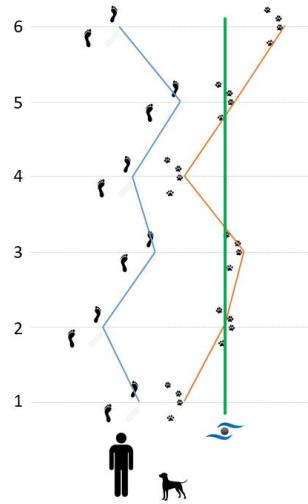
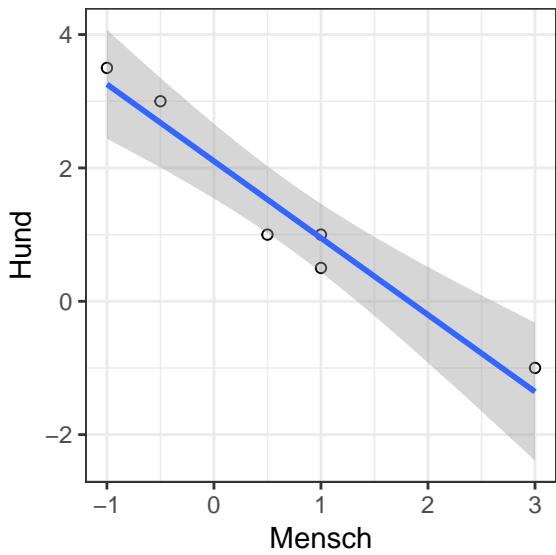
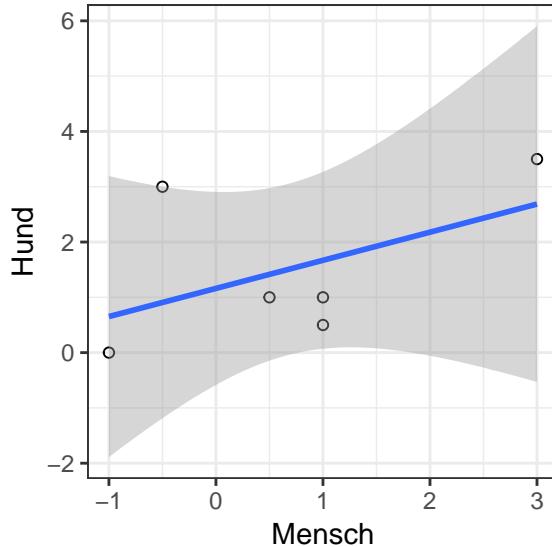


Figure 5: **Abbildung 3:** Hund und Mensch bewegen sich zum Teil unabhängig, zum Teil aber auch synchron. Dies entspricht dann einer Kovarianz, bzw. Korrelation die irgendwo zwischen $r \in [-1, 1]$ angesiedelt ist (im Beispiel ist $r(M, H) = 0.5$).



Interessant und der Praxis am ehesten entsprechend, sind jedoch Fälle, in denen zwei Variablen nur teilweise Gemeinsamkeiten aufweisen. Im folgenden Beispiel wäre



Die Korrelation liegt in diesem Beispiel bei $r(M, H) = 0.5$. Daraus lässt sich auch nochmals eine sehr wichtige Erkenntnis bezüglich der geteilten Varianz der beiden Variablen festhalten:

Bei einer Korrelation von $r(x, y) = 0.5$ entspricht der Determinationskoeffizient $r^2(x, y) = 0.25$. In Prozent ausgedrückt, werden als 25% der Variabilität einer Variablen (z.B. Hund) durch die Variable Mensch erklärt. In welchen Abschnitten der Daten diese gemeinsame Variabilität auftritt, lässt sich durch den r nicht bestimmen.

Diese Feststellung führt uns aber zu einer weiteren Betrachtung von Variabilitäten:

Würde man davon ausgehen, dass sich der Mensch und Hund bei jedem Messzeitpunkt (1 bis 6) jeweils auf der gleichen Höhe befunden haben, dann wird man eine hohe **negative Korrelation** erhalten. Nimmt man jedoch an, dass der Mensch zum Messzeitpunkt (MZP) 1, der Hund aber bereits auf MZP 2 war, dann verschiebt sich die Spur des Hundes einfach um einen MZP nach oben! Korreliert man nun diese beiden Beobachtungen, würde sich eine nahezu perfekte **positive Korrelation** ergeben!

Durch schrittweises Verschieben der Werte einer Variablen um eine Einheit (τ_i) mit anschließender Berechnung der Korrelationskoeffizienten ($r_{\tau_i}(x, y)$), erhält man in Abhängigkeit von der Anzahl der Verschiebungen ($i \in [0, N - 1]$) maximal N neue Korrelationskoeffizienten. Man bezeichnet diese Art der Korrelationsberechnung als **Kreuzkorrelation**.

Für das Beispiel ergibt sich eine normale Korrelation von $r(M, H) = -1$. Die Korrelationen berechnet nach dem Versatzprinzip ergeben folgendes Bild:

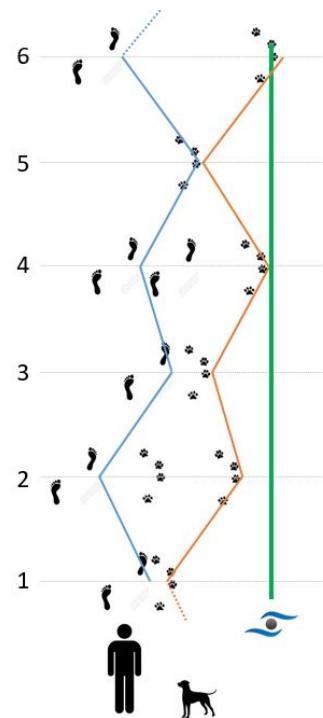
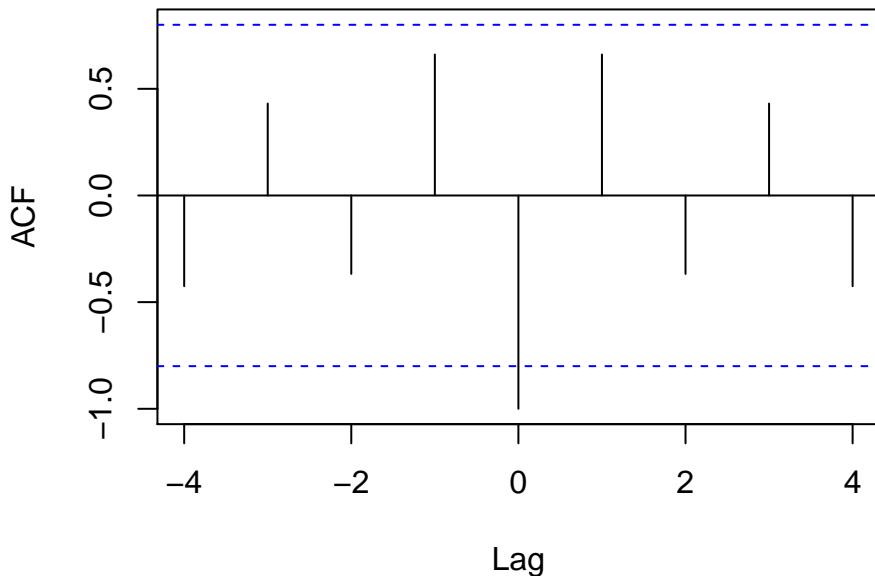


Figure 6: **Abbildung 4:** Hund und Mensch scheinen sich wieder synchron, aber in gegenseitiger Richtung zu bewegen. Man würde also eine negative und hohe Korrelation erwarten. Interessant ist jedoch die Beobachtung, dass ein Versatz der Beobachtungen um eine Einheit zu einem hohen positiven Zusammenhang führen würde!

DF_Gassi_Kreuz\$Mensch & DF_Gassi_Kreuz\$Hund



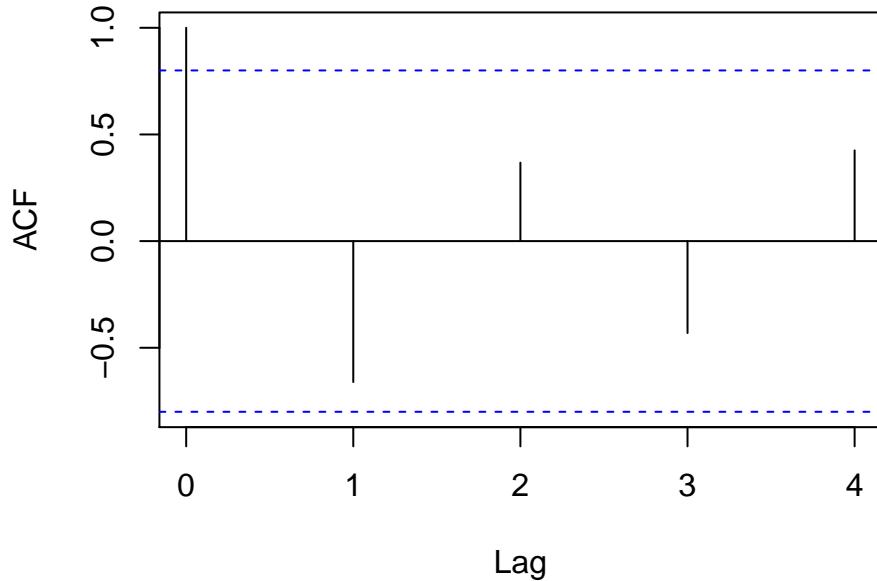
Die Verschiebung τ wurde in diesem Beispiel mit $i = 4$ angegeben, d.h. es wurden die Werte der Variablen Hund um jeweils vier Schritte nach links und vier Schritte nach rechts verschoben. Bei jeder Verschiebung wurde die Korrelation berechnet (im Graphen ist die Verschiebung mit *Lag* auf der x-Achse angegeben). Auf der y-Achse wird der entsprechende Korrelationskoeffizient angezeigt.

Tau	CrossCorr
-4	-0.4253
-3	0.431
-2	-0.3678
-1	0.6609
0	-1
1	0.6609
2	-0.3678
3	0.431
4	-0.4253

Die Werte der Tabelle zeigen nochmals den krassen Wechsel der Korrelation zwischen den Werte $\tau = 0$ (also keiner Verschiebung) und $\tau = 1$. Werden die Werte um nur einen Beobachtungspunkt verschoben, ändert sich die Korrelation von einer perfekt negativen, zu einer sehr hohen positiven Korrelation!

Eine weitere wichtige Eigenschaft die mit Hilfe dieser Vorgehensweise geprüft werden kann, ist die der sogenannten **Autokorrelation**. Diese funktioniert im Prinzip wie die eben beschriebene Kreuzkorrelation, mit dem Unterschied, dass eine Variable mit verschobenen “Eigenversionen” korreliert wird. Folgendes Beispiel zeigt das Ergebnis für die Variable Mensch unseres Beispiels:

Series DF_Gassi_Kreuz\$Mensch



Die Verschiebung τ wurde in diesem Beispiel mit $i = 4$ angegeben, d.h. es wurden die Werte der Variablen Mensch um schrittweise in eine Richtung verschoben. Bei jeder Verschiebung wurde die Korrelation berechnet (im Graphen ist die Verschiebung mit *Lag* auf der x-Achse angegeben). Auf der y-Achse wird der entsprechende Korrelationskoeffizient angezeigt.

Tau	CrossCorr
0	1
1	-0.6609
2	0.3678
3	-0.431
4	0.4253

Die perfekte positive Korrelation bei einer Verschiebung um den Wert $\tau = 0$ ist bei der Autokorrelation trivial, da es sich ja um einen direkten Vergleich der Variablen mit sich selbst handelt. Bei $Lag = 1$ wird jedoch ersichtlich, dass sich die Korrelation ändert (auf $r = -0.66$), springt dann wieder auf $r = +0.37$ usw.

Es ist zu beachten, dass dieser Datensatz nur zu Demonstrationszwecken erzeugt wurde. Eine inhaltliche Interpretation wäre im gegebenen Fall nicht angebracht.

Nichts desto trotz sollte durch diese Beispiel gezeigt werden, dass sowohl die Kreuzkorrelation als auch die Autokorrelation vor allem in der Zeitreihenanalyse (und damit auch bei Längsschnittstudien) wichtige Erkenntnisse über die betrachteten Variablen liefern können. Vor allem kann eine vorliegende **Autokorrelation** bei der MLR zu beträchtlichen Einschränkungen der Gültigkeit eines Modells beitragen. Bei den MLR-Methoden werden wir noch über Möglichkeiten sprechen, bei MLR auf statistische Signifikanz einer Autokorrelation zu prüfen (Stichwort: Durbin-Watson).

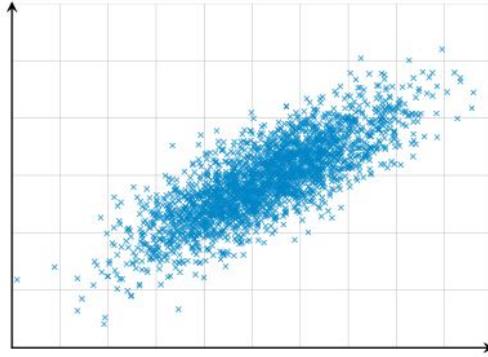


Figure 7: Abbildung 5: Endliche Varianz

Korrelationen

Korrelationen sind ein Maß für den statistischen Zusammenhang zwischen zwei Datensätzen. Unabhängige Variablen sind daher stets unkorreliert. Korrelation impliziert daher auch stochastische Abhängigkeit - ohne jedoch auf kausale Zusammenhänge schließen zu können. Bei der Berechnung einer Korrelation wird die lineare Abhängigkeit zwischen zwei Variablen quantifiziert.

Korrelationen werden i.A. der *deskriptiven Statistik* zugeordnet. Durch eine Reihe von Verfahren, wie z.B. partielle Korrelation, multiple Korrelation oder Faktorenanalyse, kann die einfache Korrelation zweier Variablen auf Beziehungen zwischen zwei Variablen unter Berücksichtigung des Einflusses weiterer Variablen werden.

Korrelationen sind ein unverzichtbares Werkzeug für viele Forschungsgebiete und stehen häufig am Beginn jeder weiteren Datenanalyse, wie z.B.:

- multiple Regression
- Faktorenanalyse
- Clusteranalyse
- Mediator- und Moderator-Analyse

Pearson Produkt Moment Korrelation

Die häufigst verwendete Form der Korrelationsberechnung ist die Pearson-Produkt-Moment Korrelation. Bei dieser Methode wird die Beziehung zwischen zwei metrische Variablen (bzw. eine metrische und eine dichotome Variable) als Kennzahl mit dem Wertebereich $r \in [-1, 1]$ berechnet.

Die Berechnung einer Korrelation ist für sich gesehen an keine Voraussetzungen gebunden. Hingegen fordern eine sinnvolle Interpretationen der berechneten Kennwerte und vor allem die statistischen Tests von Korrelationskoeffizienten folgende inhaltliche und formale Überlegungen¹:

- **Skalenniveau:** der Korrelationskoeffizient liefert zuverlässige Ergebnisse wenn die Variablen mindestens intervallskaliert sind oder für dichotome Daten².
- **Endliche Varianz (und Kovarianz):** bei Erhöhung des Stichprobenumfangs darf sich die Variabilität nicht immer weiter erhöhen, sondern sollte sich stabilisieren. Bei Variablen, die bivariat normalverteilt sind, ist diese Voraussetzung automatisch gegeben. Der Korrelationskoeffizient ist damit auch gleichzeitig der *Maximum-Likelihood Schätzer* des Korrelationskoeffizienten in der Grundgesamtheit (asymptotisch erwartungstreu und effizient).

¹die Abbildungen wurden der Website Matheguru entnommen

²dieser Spezialfall ist unter biserialer, bzw. punktbiserialer Korrelation bekannt.

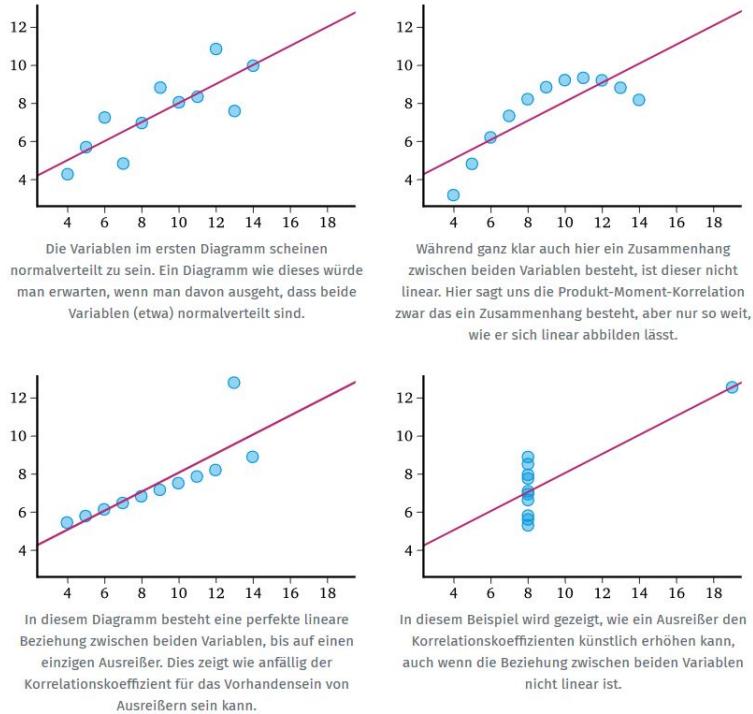


Figure 8: **Abbildung 6:** Linearität und Korrelation

- **Linearität:** die Korrelation ist ein Maß für **lineare Abhängigkeit**. Abweichungen der Daten von dieser Linearitätsannahme führen zu einer mehr oder weniger starken Verzerrung des Korrelationskoeffizienten, wie in den nachfolgenden Beispielen gezeigt wird:

Vor allem zur Prüfung der Signifikanz einer Korrelation sollte man weitere Voraussetzungen überprüfen:

- **Normalverteilung:** Korrelation berechnen sich aus dem Kreuzprodukt von z-standardisierten Werten zweier Variablen. Für diese Berechnung wird der Mittelwert als zentraler Kennwert verwendet, welcher nur dann ein "sinnvoller" Kennwert für die Daten ist, wenn diese zumindest symmetrisch und im besten Fall normalverteilt sind.
- **Homoskedastizität:** bedeutet unterschiedliche Streuung innerhalb einer Datenmessung. Die exogene und die endogene Variable³ sind nicht mehr identisch verteilt, d.h. sie ändern ihre Variabilität mit zu/abnehmenden Werten einer Variablen. Das hat zur Folge, dass die KQ⁴-Schätzer nicht mehr effizient sind und der Standardfehler der Koeffizienten verzerrt und nicht konsistent wird.
- **KEINE Ausreißer:** der Korrelationskoeffizient ist nicht robust gegenüber Ausreißern. Dies bedeutet, dass Ausreißer den Korrelationskoeffizienten sowohl künstlich erhöhen als auch künstlich senken können.
- **KEINE Kluster:** es kann vorkommen, dass zwei oder mehr Gruppen eine Korrelation zeigen, die eigentlich getrennt untersucht werden müssten. Dieses Problem wird oft auch mittels **partieller Korrelation** umgangen, bei der mögliche Drittvariablen statistisch konstant gehalten werden.

³eine *exogene* Variable ist eine erklärende Variable, die mit der Störgröße unkorreliert ist (sogenannte Exogenität). Eine *endogene* erklärende Variable in einem multiplen Regressionsmodell ist eine erklärende Variable, die entweder aufgrund einer ausgelassenen Variablen, eines Messfehlers oder wegen Simultanität mit der Störgröße korreliert ist (sogenannte Endogenität).

⁴KQ steht für kleinste Quadrate (auch MLS - Minimum Least Square, oder OLS - Ordinary Least Square) und ist eine einfache Schätzung über minimierte quadratische Abstände der Residuen (Fehler) zu einem Modell (Mittelwert, Gerade, etc.)

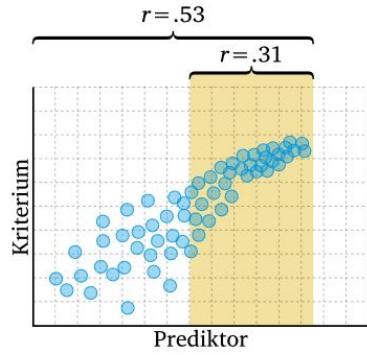


Figure 9: **Abbildung 7:** Variabilität(einschränkung) und Korrelation

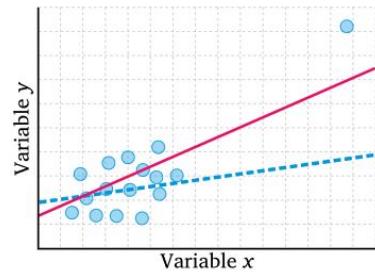


Figure 10: **Abbildung 8:** Einfluss von Ausreißer bei linearer Modellbildung

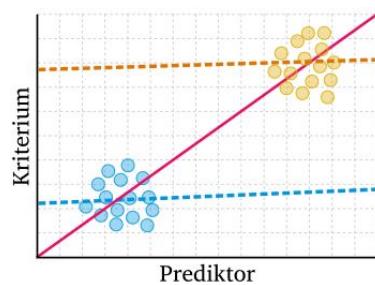


Figure 11: **Abbildung 9:** Kluster und deren Auswirkung bei linearer Modellierung

Beispiel Pearson Korrelation

Im folgenden, fiktiven Beispiel werden die Zusammenhänge von Klausurperformanz (*EP*), Intelligenz (*IQ*), Vorbereitungszeit (*VZ*) und Prüfungsangst (*PA*) korreliert. Der Code zum Laden der Daten sowie die Daten selbst sind in nachfolgender Ausgabe/Tabelle dargestellt:

```
load("Daten/CorrBsp1.Rda")
```

EP	IQ	VZ	PA
74	109	16	117
67	96	18	122
72	106	13	108
66	89	12	97
63	93	14	98
67	102	15	106

Aufgabe 1

Übernehmen Sie den obigen Code zum Lade der Daten in eine R-Script-Datei. Führen Sie nun folgende Aufgaben aus:

1. Ermitteln Sie mit einer geeigneten Funktion die Korrelationen und prüfen Sie diese auch auf statistische Signifikanz.
2. Zeichnen Sie einen Korrelationsplot mit dem Paket *corrplot*.
3. Berechnen Sie die Teststärke der Korrelation $r(IQ, EP)$ (**Hinweis**: verwenden Sie die Funktion *pwr.r.test* des Pakets *pwr*).
4. Verwenden diese Funktion (*pwr.r.test*) um für eine Korrelation $r(x, y) = 0.21$ den optimalen Stichprobenumfang zu berechnen.
5. Prüfen Sie mit Hilfe der Funktion *mvn* aus dem Paket *MVN* die Voraussetzung der bivariaten Normalverteilung der Variablenpaare (EP,IQ), (EP, VZ) und (EP,PA).
6. Berechnen Sie die durchschnittliche Korrelation von $r_1(EP, IQ)$, $r_1(EP, VZ)$ und $r_1(EP, PA)$. Beachten Sie, dass zur Berechnung von durchschnittlichen Korrelationswerten eine Fisher-Z-Transformation notwendig ist (**Hinweis**: verwenden Sie die *fisherz()* und *fisherz2r()* des Pakets *psych*).
7. Prüfen Sie, ob der Unterschied der Korrelationskoeffizienten $r(EP, IQ) = 0.47$ und $r(EP, VZ) = 0.36$ statistisch signifikant ist. Verwenden Sie die Funktion *paired.r()* aus dem Paket *psych*.

Lösung Aufgabe 1

Kausalität

Eine relevante (statistisch signifikante) Korrelation liefert keinen Beleg für die Kausalität. Vor allem in der Medizin und Psychologie suchen Forscher nach Kriterien für Kausalität. Es existieren mehrere Ansätze zur Erklärung der Ursächlichkeit einer Korrelation (siehe z.B. die 9 Bradford-Hill-Kriterien).

Partial- Semipartialkorrelation

Die *partielle Korrelation* ist die bivariate Korrelation zweier Variablen, welche mittels linearer Regression vom Einfluss einer Drittvariablen bereinigt wurden.

Eine *Semipartialkorrelation* ist ein Zusammenhang zwischen einer residualisierten und einer nicht-residualisierten Variable.

Beispiel Partial- Semipartialkorrelation

Folgendes Beispiel verdeutlicht die Wirkungsweise einer Partial- und Semipartialkorrelation. Kopier den folgenden Code in ein R-Script und führe diesen dann aus. Diskutiere die Ergebnisse.

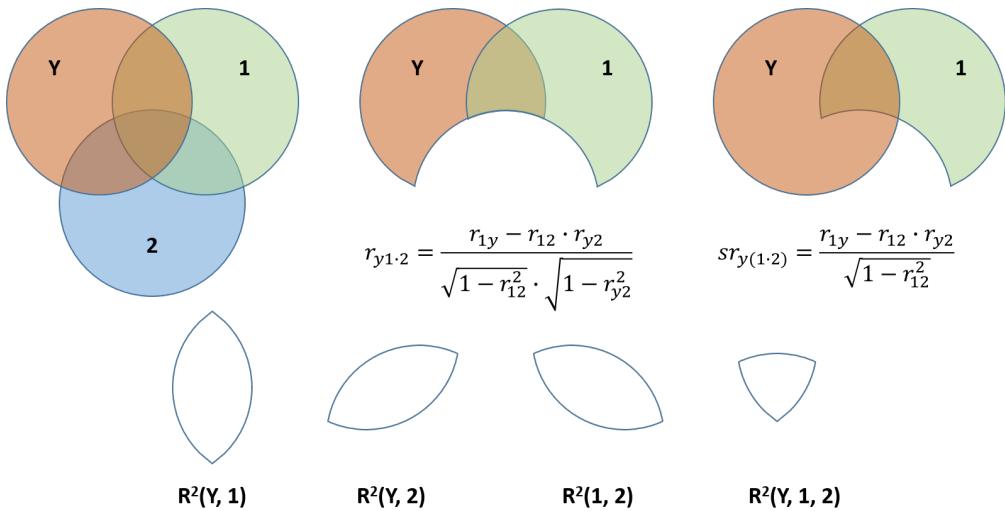


Figure 12: Abbildung 10: Partial und Semipartialkorrelation in einem Venn-Diagramm dargestellt

```

examData  <- read.delim("Daten/Exam Anxiety.dat", header = TRUE)
examData2 <- examData[, c("Exam", "Anxiety", "Revise")]

# Normale Korrelation
pander::pander(round(cor(examData2), 2))

# Partielle Korrelation
# library(ppcor)
pander::pander(round(ppcor::pcor(examData2)$estimate, 2))

# Partialkorrelation mit Linearen Modell
Mod1      <- lm(Exam ~ Revise, data = examData2)
Res_Exam_Rev <- residuals(Mod1)
Mod2      <- lm(Anxiety ~ Revise, data = examData2)
Res_Anx_Rev <- residuals(Mod2)
pr_Exam_Anx_Rev <- round(cor(Res_Exam_Rev, Res_Anx_Rev), 2)

# Semipartialle Korrelation
pander::pander(round(ppcor::spcor(examData2)$estimate, 2))
# Semipartialkorrelation mit Linearen Modell
sr_Exam_Anx_Rev <- round(cor(examData2$Exam, Res_Anx_Rev), 2)
  
```

In diesem Code wurde zur Veranschaulichung der Wirkungsweise einer Partial/Semipartialkorrelation eine lineare Regression verwendet. Was dabei genau passiert sei durch nachfolgende Abbildung nochmals veranschaulicht:

1. Examperformance wird durch Revisiontime vorhergesagt. Die Residuen sind jener Anteil an Variabilität der Examperformance, der nicht durch Revisiontime vorhergesagt werden können⁵. Diese über die durch Revisiontime erklärbare Variabilität von Examperformance kann zurückgeführt werden auf:
 - andere erklärende Merkmale, bzw.
 - Messfehler
2. Anxiety wird durch Revisiontime vorhergesagt. Auch hier gilt wieder, dass die Residuen der Variabilität von Anxiety, bereinigt von Revisiontime entsprechen.
3. Die Korrelation der Residuen entspricht nun genau der Partialkorrelation $r_{Y1.2}$

Bei der Semipartialkorrelation bereinigt man nun nicht beide Variablen, sonder eben nur einen Teil (z.B. wird

⁵ anderenfalls würden ja alle beobachteten Werte auf der Gerade liegen!

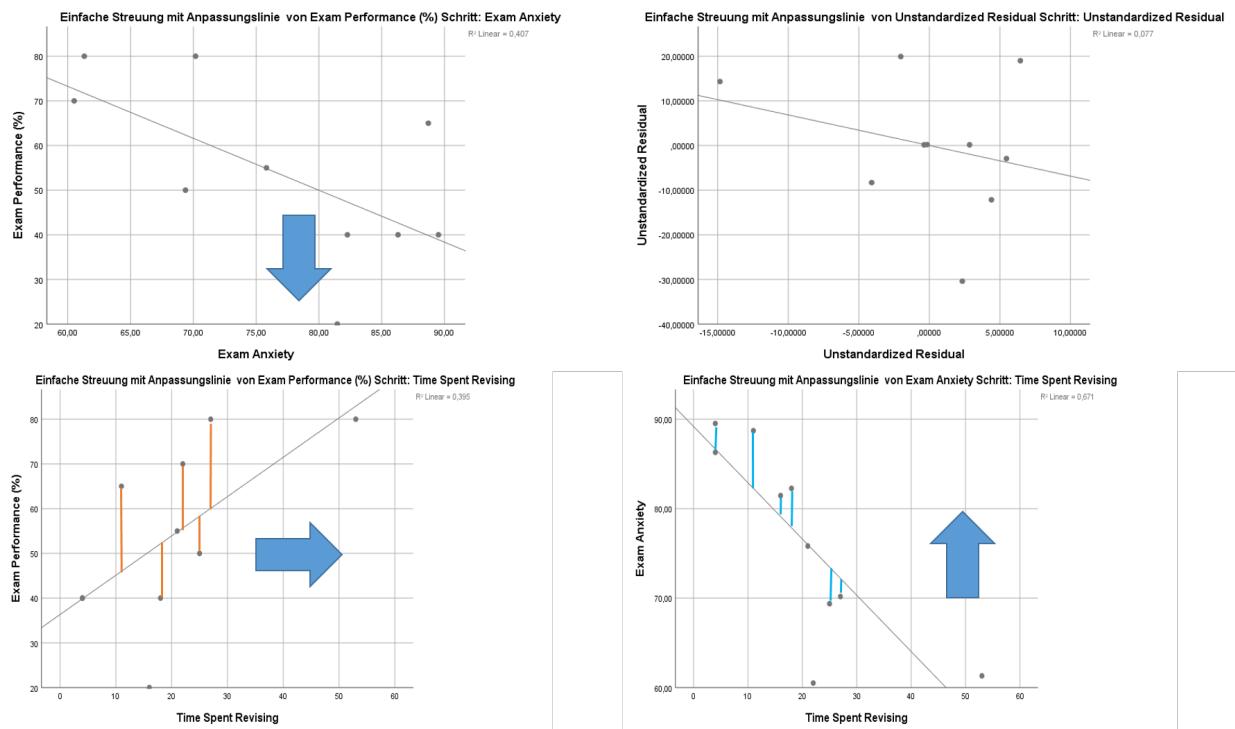


Figure 13: **Abbildung 11:** Partial und Semipartialkorrelation als lineares Regressionsmodell. Für die Beziehung Examperformance und Exam Anxiety soll der Effekt von Revisiotime berücksichtigt werden. Die roten Linien entsprechen den Residuen der Regression Revisiontime mit Exam Performance. Die blauen den Residuen der Regression Revisiontime mit Exam Anxiety. Der linke obere Graph stellt die Beziehung von Anxiety und Examperfomance bereinigt von Recisiontime dar. Details siehe nachfolgendemn Text.

Nominalskaliert					
dichotom					
Intervallskaliert	Ordinalskaliert	künstlich	natürlich	polytom	
Intervallskaliert	<input type="checkbox"/> Pearson Produkt-Moment-Korrelation <input type="checkbox"/> Kendall's Tau <input type="checkbox"/> polychorische Korrelation	<input type="checkbox"/> Spearman's Rho <input type="checkbox"/> biserial Korrelation <input type="checkbox"/> polychorische Korrelation	<input type="checkbox"/> punktbiserial Korrelation <input type="checkbox"/> biserial Korrelation	<input type="checkbox"/> punktbiserial Korrelation	<input type="checkbox"/> η -Koeffizient
Ordinalskaliert		<input type="checkbox"/> Spearman's Rho <input type="checkbox"/> Kendall's Tau <input type="checkbox"/> polychorische Korrelation	<input type="checkbox"/> biserial Rangkorrelation <input type="checkbox"/> polychorische Korrelation	<input type="checkbox"/> biserial Rangkorrelation	<input type="checkbox"/> Cramér's V
Nominalskaliert (künstlich dichotom)			<input type="checkbox"/> Punkttetrachorische Korrelation (φ -Koeffizient) <input type="checkbox"/> Tetrachorische Korrelation	<input type="checkbox"/> Punkttetrachorische Korrelation (φ -Koeffizient) <input type="checkbox"/> v -Koeffizient	<input type="checkbox"/> Cramér's V
Nominalskaliert (natürlich dichotom)				<input type="checkbox"/> Punkttetrachorische Korrelation (φ -Koeffizient) <input type="checkbox"/> Yule's Y	<input type="checkbox"/> Cramér's V
Nominalskaliert (polytom)					<input type="checkbox"/> Cramér's V

Figure 14: Abbildung 12: verschiedene Korrelationskoeffizienten

nur die Anxiety von Revisiontime bereinigt).

Kopiere den nachfolgenden Code in ein R-Script und führe diesen aus. Diskutiere die Ergebnisse!

Korrelationstechniken

Neben dem Pearson-Produkt-Moment-Korrelationskoeffizienten r existieren noch etliche weitere Korrelationskoeffizienten und Zusammenhangsmaße. Die meisten hiervon sind Sonderfälle der Pearson-Produkt-Moment-Korrelation. Nachfolgende Tabelle zeigt, wann welcher Koeffizient berechnet werden soll. Die Verwendung unterschiedlicher Korrelationsberechnungen ist i.A. abhängig vom Skalenniveau der beteiligten Variablen.

Spearman und Kendall

Für die Berechnung des Pearson-Korrelationskoeffizienten (r) ist das Vorliegen von kontinuierlichen Variablen erforderlich. Bei **ordinalskalierten Daten** wird eine der folgenden Rangkorrelation berechnet:

- **Spearman r_s :** Spearman-Rangkorrelation setzt voraus, dass Ränge gleichabständig sind⁶ und keine Ausreißer vorliegen.
- **Kendall τ :** Ränge müssen nicht gleichabständig sein und Ausreißer beeinflussen diesen Korrelationskoeffizienten weit weniger als z.B. den r_s . Beim Kendall-Maßen unterscheidet man noch drei

⁶diese Voraussetzung ist eher selten erfüllt. Sie ist gleichzusetzen mit der Annahme, dass in einem Skirennen der erste, zweite, dritte, etc. Platz genaus die gleichen Zeitabstände aufweisen. Ist diese nicht gegeben, sollte Kendalls τ verwendet werden.

unterschiedliche Maße⁷: * Kendalls τ_a : Rangbindungen werden nicht berücksichtigt. * Kendalls τ_b : Rangbindungen werden berücksichtigt. * Kendalls τ_c : für nicht quadratische Kontingenztafeln.

Zur Veranschaulichung der verschiedenen Rangbasierten Korrelationsmaße sind folgende Aufgaben zu bearbeiten:

1. Berechnen Sie zuerst nochmal die Pearson-Korrelation $r(EP, IQ)$ des bereits geladenen Datensatzes und rechnen Sie dann eine Spearman Korrelation. Verwenden Sie nun die Funktion `cor()` des Basispaketes. Vergleichen Sie die Ergebnisse!
2. Vererwenden Sie die Funktion `rank()` um den Variablen *EP* und *IQ* Ränge zuzuordnen. Speichern Sie die Ergebnisse in *EP_Ranks* und *IQ_Ranks* und berechnen Sie anschließend eine Pearson-Korrelation. Vergleichen Sie die Ergebnisse mit dem vorherigen Pearson-r.

Lösung Aufgabe 2

Biserial Korrelation

Biserial Korrelationen kommen zur Anwendung, wenn ein Merkmal **Intervall-** oder **Ordinalskaliert** und das zweite Merkmal **dichotom Nominalskaliert** ist. Für das Nominalskalierte Merkmal unterscheidet man noch zwischen:

- **Echt dichotome Variable:** natürlich vorkommende Gruppenteilung wie z.B. wahr/falsch, männlich/weiblich, etc. Der Zusammenhang einer solchen mit einer intervallskalierten Variablen wird durch die **punktbiserial Korrelation** beschrieben.
- **Künstlich dichotome Variable:** wird eine kontinuierliche Variable in zwei Gruppen aufgeteilt, wie z.B. zwei Altersgruppen (jung, alt), oder hohe Leistungsfähigkeit vs. niedrige Leistungsfähigkeit, etc., dann spricht man von einer künstlich dichotomen Variablen. Zusammenhänge dieser mit einer intervallskalierten Variablen werden durch die **biserial Korrelation** beschrieben.

Betrachten wir folgendes Beispiel: eine Gruppe von Personen hat eine Dpressionsbehandlung erhalten, einer weiteren nicht. Beiden Gruppen wurden über den BDI gemessen um die Symptombelastung zu messen. Die Daten liegen

Phi-Koeffizient

Korrelationen zwischen echt-dichotomen Variablen (männlich/weiblich, etc.) können mit dem Phi-Koeffizienten berechnet werden. Um den Phi-Koeffizienten zu berechnen, werden Häufigkeiten in Form einer Vier-Felder-Tafel benötigt.

Folgendes einfaches Beispiel zeigt die Berechnung des Phi-Koeffizienten sowie dessen Äquivalenz mit einer Pearson-Korrelation:

```
Geschlecht <- c(1, 1, 0, 0, 1, 0, 1, 1, 1)
Bestanden <- c(1, 1, 1, 0, 0, 0, 1, 1, 1)

VFT      <- table(Geschlecht, Bestanden)
pander::pander(VFT)

cor(Geschlecht, Bestanden)

phi(VFT)
CST <- chisq.test(VFT, correct = FALSE)
pander::pander(CST)
qchisq(p=.95,df=1) # kritischer Chi-Square-Wert bei einem Freiheitsgrad und Alpha = 5%
```

⁷Details zu den unterschiedlichen Kendalls- τ sind der Literatur zu entnehmen. Weitere Betrachtungen beziehen sich auf das Kendalls- τ_b

	Nicht bestanden	Bestanden	
Männlich	2 (a)	1 (b)	3 (a+b)
Weiblich	1 (c)	5 (d)	6 (c+d)
	3 (a+c)	6 (b+d)	9 (a+b+c+d)

$$\phi = \frac{(a \cdot d) - (b \cdot c)}{\sqrt{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}}$$

$$\chi^2 = \frac{n \cdot [(a \cdot d) - (b \cdot c)]^2}{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}$$

Figure 15: Abbildung 13: Beispiel und Berechnung des Phi-Koeffizienten

Die Anzahl der Freiheitsgrade beträgt in diesem Fall immer eins, da wir es mit zwei dichotomen Merkmalen zu tun haben.

Lösungen

Aufgabe_1

```
# library(Hmisc) fÃ¼r Hmisc::rcorr
# library(corrplot) fÃ¼r corrplot
# library(pwr)

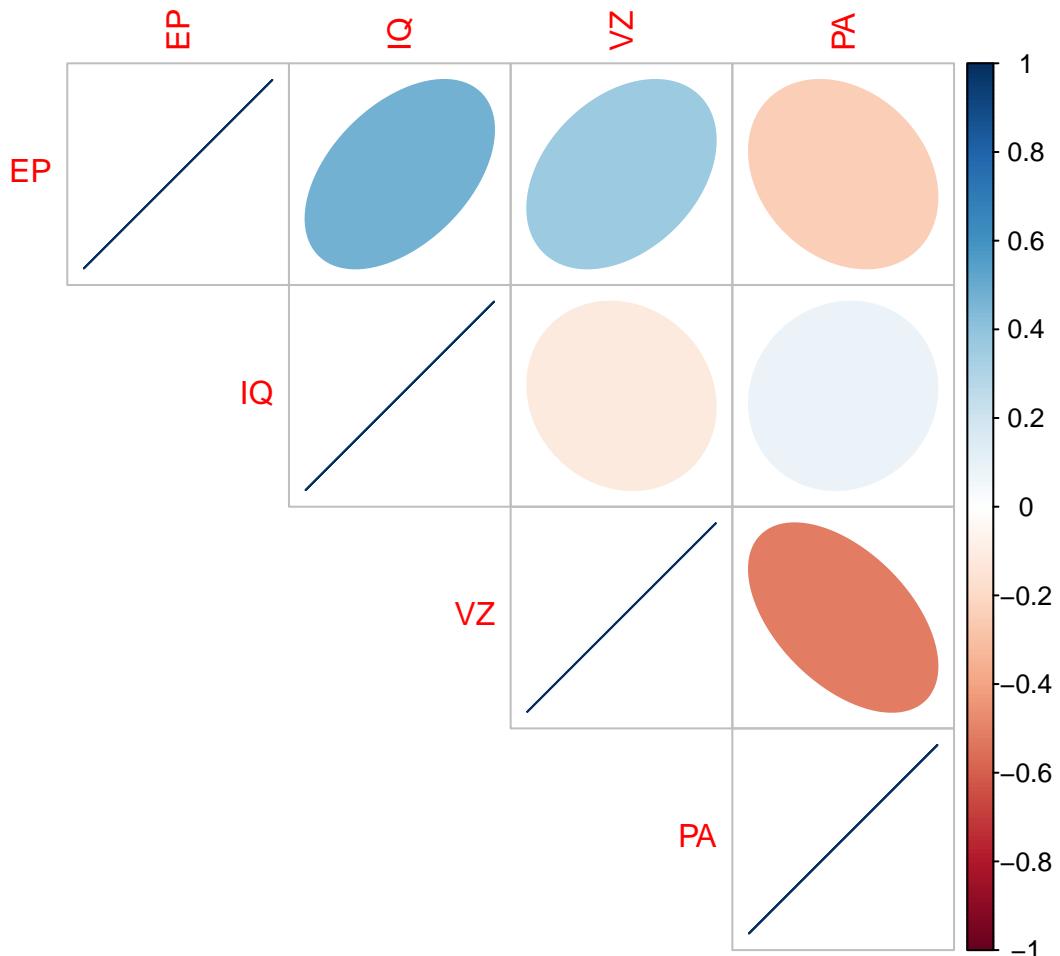
# 1. Ermitteln Sie mit einer geeigneten Funktion die Korrelationen und prÃ¼fen Sie
# diese auch auf statistische Signifikanz.
CorRes <- Hmisc::rcorr(as.matrix(DF_Korr), type="pearson") # type can be pearson or spearman
pander::pander(round(CorRes$r, 2))
```

	EP	IQ	VZ	PA
EP	1	0.47	0.36	-0.25
IQ	0.47	1	-0.12	0.08
VZ	0.36	-0.12	1	-0.52
PA	-0.25	0.08	-0.52	1

```
pander::pander(round(CorRes$P , 2))
```

	EP	IQ	VZ	PA
EP	NA	0	0	0
IQ	0	NA	0.06	0.2
VZ	0	0.06	NA	0
PA	0	0.2	0	NA

```
# 2. Zeichnen Sie einen Korrelationsplot mit dem Paket *corrplot*.
corrplot::corrplot(cor(DF_Korr), type="upper", method = "ellipse")
```



```
# 3. Berechnen Sie die Teststärke der Korrelation $r(IQ, EP)$
#   (Hinweis: verwenden Sie die Funktion *pwr::pwr.r.test* des Pakets *pwr*).
N      <- dim(DF_Korr)[1]
PwrA <- pwr::pwr.r.test(n = N, r = 0.47, sig.level = 0.05, alternative = 'two.sided')
pander::pander(data.frame(Kennwerte = unlist(PwrA)))
```

	Kennwerte
n	248
r	0.47
sig.level	0.05
power	0.999999999220975
alternative	two.sided
method	approximate correlation power calculation (arctangh transformation)

```
# 4. Verwenden diese Funktion (*pwr::pwr.r.test*) um für eine Korrelation  $r(x,y) = 0.21$ 
# den optimalen Stichprobenumfang zu berechnen.
OptN <- pwr::pwr.r.test(r = 0.21, sig.level = 0.05, power = 0.95, alternative = 'greater')
pander::pander(data.frame(Kennwerte = unlist(OptN)))
```

	Kennwerte
n	240.15014639158
r	0.21
sig.level	0.05
power	0.95
alternative	greater
method	approximate correlation power calculation (arctanh transformation)

```
# 5. Prüfen Sie mit Hilfe der Funktion *mvn* aus dem Paket *MVN*
# die Voraussetzung der bivariaten Normalverteilung der
# Variablenpaare (EP, IQ), (EP, VZ) und (EP, PA).
# library(MVN)
# mvn(DF_Korr[, c("EP", "IQ")], multivariatePlot = "persp")
# mvn(DF_Korr[, c("EP", "VZ")], multivariatePlot = "persp")
# mvn(DF_Korr[, c("EP", "PA")], multivariatePlot = "persp")

# 6. Berechnen Sie die durchschnittliche Korrelation von  $r_{-1}(EP, IQ)$ ,  $r_{-1}(EP, VZ)$  und  $r_{-1}(EP, PA)$ .
round(fisherz2r(mean(c(fisherz(.47), fisherz(.36), fisherz(-.25)))), 2)
```

```
## [1] 0.21

# 7. Prüfen Sie, ob der Unterschied der Korrelationskoeffizienten  $r(EP, IQ) = 0.47$  und  $r(EP, VZ) = 0.36$ 
# statistisch signifikant ist. Verwenden Sie die Funktion *psych::paired.r()* aus dem Paket *psych*
# library(psych)
psych::paired.r(xy      = .47,
                xz      = .36,
                n       = N,
                twotailed = TRUE)
```

```
## Call: psych::paired.r(xy = 0.47, xz = 0.36, n = N, twotailed = TRUE)
## [1] "test of difference between two independent correlations"
## z = 1.5 With probability = 0.14
```

zurück zu Aufgabe

Aufgabe_2

```
# library(Hmisc)

# 1. Berechnen Sie zuerst nochmal die Pearson-Korrelation  $r(EQ, IQ)$  des bereits geladenen Datensatzes
# und rechnen Sie dann eine Spearman Korrelation. Verwenden Sie nun die Funktion cor() des Basis-Pakets
# Vergleichen Sie die Ergebnisse!

EP <- DF_Korr$EP
IQ <- DF_Korr$IQ

CorPearson <- round(cor(EP, IQ, method = "pearson"), 2)
```

```

CorSpearman <- round(cor(EP, IQ, method = "spearman"), 2)
CorKendall  <- round(cor(EP, IQ, method = "kendall"), 2)

pander::pander(data.frame(Pearson  = CorPearson,
                           Spearman = CorSpearman,
                           Kendall   = CorKendall))

```

Pearson	Spearman	Kendall
0.47	0.47	0.34

```

# Alternativ kann auch cor.test() verwendet werden, dabei werden die Tests auf
# Signifikanz gleich mitgerechnet.
# cor.test(x = EP,
#           y = IQ,
#           alternative = 'two.sided',
#           method = 'pearson')
# cor.test(x = EP,
#           y = IQ,
#           alternative = 'two.sided',
#           method = 'spearman')
# cor.test(x = EP,
#           y = IQ,
#           alternative = 'two.sided',
#           method = 'kendall')
# Bemerkung: cor.test() mit Kendall bringt Warnung bezüglich der Rangbindungen.
# Alternativ kann man daher die Funktion Kendall() des Paketes Kendall verwenden:
# library(Kendall)
# Kendall(x = EP,
#          y = IQ)
# 2. Verwenden Sie die Funktion rank() um den Variablen EP und IQ Ränge zuzuordnen.
# Speichern Sie die Ergebnisse in EP_Ranks und IQ_Ranks und berechnen Sie anschließend
# eine Pearson-Korrelation. Vergleichen Sie die Ergebnisse mit dem vorherigen Pearson-r.
EP_Ranks <- rank(EP, na.last = TRUE,
                  ties.method = c("average"))
IQ_Ranks <- rank(IQ, na.last = TRUE,
                  ties.method = c("average"))
round(cor(EP_Ranks, IQ_Ranks, method = "pearson"), 2)

## [1] 0.47
zurück zu Aufgabe

```

Teil II: Multiple lineare Regression

```

# Initialisierung
rm(list = ls())
graphics.off()
if (!require("pacman")) install.packages("pacman")
pacman::p_load(car, DAAG, ggplot2, pander, ppcor, mosaicData, reshape2, rockchalk)
# 11.04.2019 rockchalk entfernt da Problem beim Laden

```

Multiple Regression

Man könnte nun die bereits erwähnte Variable Erfahrung (*exper*) ins Modell aufnehmen. Der bereits aus der Korrelation ersichtliche (negative) Zusammenhang mit der Ausbildung *educ* lässt den Schluss auf eine Kovariabilität der beiden Variablen zu. Man nennt derartige Variablen auch **Kovariate**. Im linearen Modell wird diese jedoch wie eine weitere Variable (ein weiterer Prädiktor) zur Vorhersage des Kriteriums verwendet.

Definition

Die formale Definition eines multiplen linearen Modells ist:

$$y_i = b_0 + b_1 \cdot x_{1i} + \cdots + b_k \cdot x_{ki} + \varepsilon_i \quad (1)$$

Die wesentlichen Parameter dieses Modells sind:

1. Intercept b_0 : jener Wert den y_i einnimmt, wenn $x_{ji} = 0$ ist (mit $j \in [1, k]$).
2. Steigung b_i : die Zunahme von y_i , wenn x_{ji} sich um eine Einheit erhöht, bei gleichzeitigem Konstanthalten der restlichen Prädiktorwerte x_{mi} (mit $m \in [1, k]$ und $m \neq j$)!

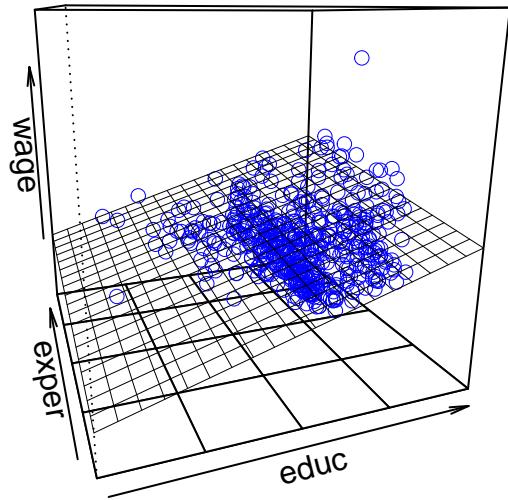
Des Weiteren berücksichtigt auch dieses Modell wieder einen Fehler (ε_i). Betrachtet man das multiple Modell isoliert (also ohne Fehlerterm), ist folgende Schreibweise üblich:

$$\hat{y}_i = b_0 + b_1 \cdot x_{1i} + \cdots + b_k \cdot x_{ki} \quad (2)$$

Betrachten wir an unseren Beispieldaten folgendes Modell mit zwei Prädiktoren:

$$\hat{wage}_i = b_0 + b_1 \cdot educ_i + b_2 \cdot exper_i$$

```
# library(mosaicData) fÃ¼r CPS85
model_1      <- lm(wage ~ educ, data = CPS85)
model_2      <- lm(wage ~ educ + exper, data = CPS85)
Det_model_2 <- pander::pander(summary(model_2))
# 11.04.2019 entfernt, da es Probleme beim Laden gibt
rockchalk::plotPlane(model = model_2, plotx1 = "educ", plotx2 = "exper")
```



Dabei entspricht der Koeffizient b_2 der Zunahme des Gehaltes \hat{y}_i wenn sich die Erfahrung x_{2i} um eine Einheit erhöht und die Ausbildung x_{1i} konstant gehalten wird. In nachfolgender Tabelle sind die Werte der Vorhersagen des Modells für den vorliegenden Datensatz auszugsweise dargestellt:

```

MinExp      <- min(CPS85$exper)
MaxExp      <- max(CPS85$exper)
RowSeq      <- seq(from = 1, to = MaxExp, by = 1)
educVon     <- 10
educBis     <- 18
AnzCols    <- educBis - educVon + 1
Predicted   <- matrix(NA, nrow = MaxExp, ncol = AnzCols)
for (i in seq(from = 1, to = MaxExp, by = 1)) {
  new_input   <- data.frame(educ = educVon:educBis, exper = i)
  Predicted[i,] <- predict(model_2, newdata = new_input)
}
Predicted      <- data.frame(seq(from = 1, to = MaxExp, by = 1), Predicted)
colnames(Predicted) <- c("Exp", "Edu10", "Edu11", "Edu12", "Edu13",
                         "Edu14", "Edu15", "Edu16", "Edu17", "Edu18")
TabRows2Disp   <- c(1:3, 53:55)
Predicted2Disp  <- Predicted[TabRows2Disp,]
row.names(Predicted2Disp) <- NULL
pander::pander(Predicted2Disp, style = "rmarkdown")

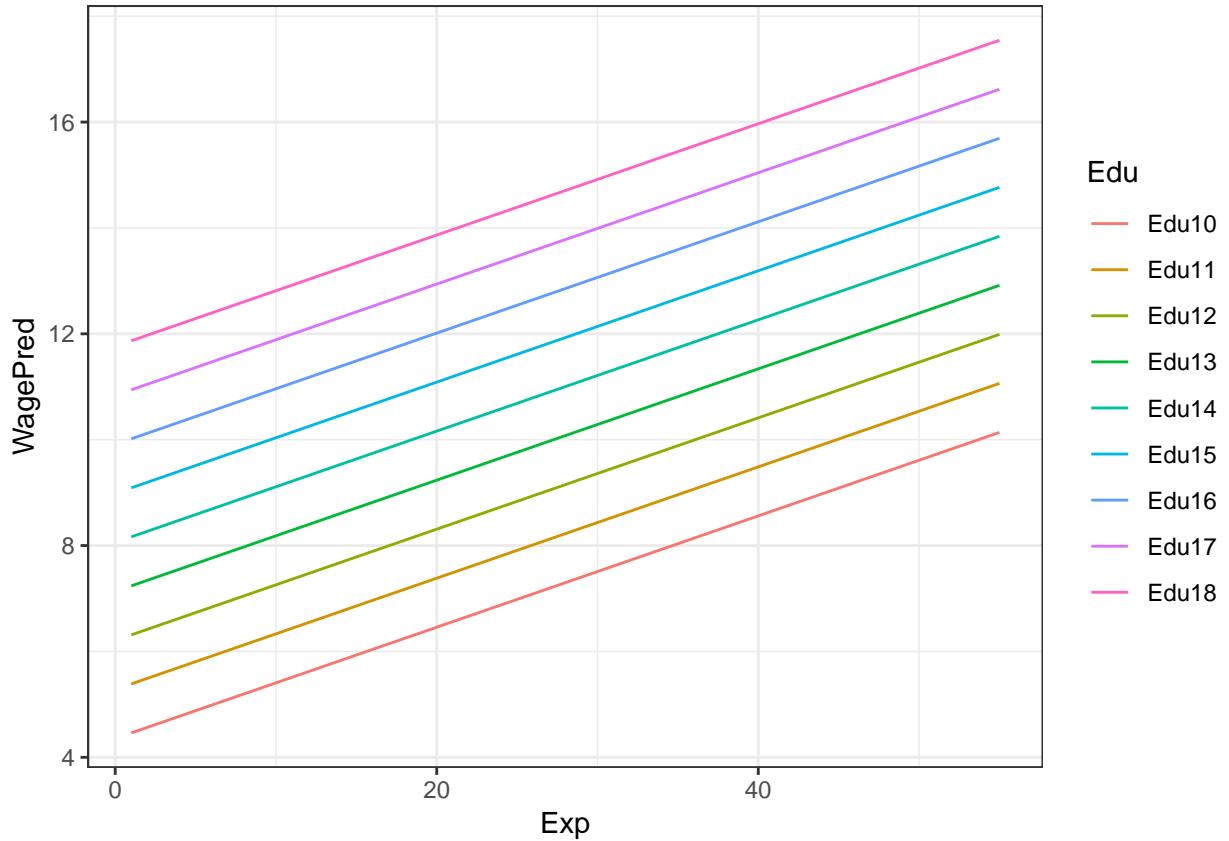
```

Exp	Edu10	Edu11	Edu12	Edu13	Edu14	Edu15	Edu16	Edu17	Edu18
1	4.46	5.386	6.312	7.238	8.164	9.09	10.02	10.94	11.87
2	4.565	5.491	6.417	7.343	8.269	9.195	10.12	11.05	11.97

Exp	Edu10	Edu11	Edu12	Edu13	Edu14	Edu15	Edu16	Edu17	Edu18
3	4.671	5.597	6.522	7.448	8.374	9.3	10.23	11.15	12.08
53	9.927	10.85	11.78	12.71	13.63	14.56	15.48	16.41	17.33
54	10.03	10.96	11.88	12.81	13.74	14.66	15.59	16.51	17.44
55	10.14	11.06	11.99	12.92	13.84	14.77	15.69	16.62	17.55

```
# library(reshape2) #for melt
# library(ggplot) #for Plots
CPS852Disp      <- reshape2::melt(Predicted,
                                    id.vars = "Exp",
                                    measure.vars = c("Edu10", "Edu11", "Edu12",
                                                    "Edu13", "Edu14", "Edu15",
                                                    "Edu16", "Edu17", "Edu18"))

CPS852Disp$Exp   <- rep(1:55, 9)
colnames(CPS852Disp) <- c("Exp", "Edu", "WagePred")
p                  <- ggplot(CPS852Disp, aes(x = Exp, y = WagePred, color = Edu)) +
                        geom_line() +
                        theme_bw()
print(p, comment = FALSE)
```



Modellvergleich

Ein Modell sollte die Wirklichkeit mit möglichst großer Genauigkeit abbilden. Bei der Erstellung des Modells wurden aufgrund einer Stichprobe aus der Grundgesamtheit die Modellparameter (z.B. die b 's) bestimmt.

Um nun festzustellen, inwieweit das Modell brauchbare Vorhersagen liefert, sollte man das Modell evaluieren. In den vorangegangen Beispielen wurden zwei Modelle (*model_1* und *model_2*) erstellt.

Der Vergleich der Modelle ist über den Fehler des jeweiligen Modells möglich. Je kleiner der Fehler, desto besser bildet das Modell die beobachteten Werte ab. Im Idealfall (Fehler = 0), würden alle beobachteten Werte gleich den vorhergesagten Werten sein und damit auf der Linie liegen.

```
M      <- data.frame(wage = CPS85$wage, educ = CPS85$educ, exper = CPS85$exper)
MV_Data <- data.frame(educ = M$educ, exper = M$exper)
MSE_Model1 <- round(mean(resid(model_1)^2), 2)
#MSE_Model1 <- mean((M$wage - predict(model_1, newdata = MV_Data))^2)
StdResid <- rstandard(model_1)
#StdResid <- (resid(model_1) - mean(resid(model_1))) / sd(resid(model_1))
MSE_Model2 <- round(mean((M$wage - predict(model_2, newdata = MV_Data))^2), 2)
```

Der Modellvergleich der obigen Beispiele ergibt für das Modell 1 einen $MSE_1 = 22.52$ und für Modell 2 einen $MSE_2 = 21.04$.

Bei diesen Ergebnis lässt sich zunächst nur feststellen, dass der MSE_2 kleiner als der MSE_1 ist. Ob diese Verringerung des MSE von statistischer und/oder praktischer Signifikanz ist, wird im folgenden noch genauer betrachtet.

Mit einer einfachen ANOVA lässt sich nun auch die statistische Signifikanz der Änderungen im Fehler bei den verwendeten Modellen berechnen. Betrachten wir zunächst die statistische Änderung die Modell 1 im Vergleich zum Mittelwertsmodell erzielt:

```
# ANOVA Tests auf signifikante Änderungen model_1 vs Mittelwertsmodell
# Berechnung der Quadratsummen fÄr die Regression (educ)
preds_1      <- predict(model_1, newdata = CPS85)
AnzPred       <- 2 # b_0 und b_1
SS_Regression_1 <- sum((preds_1 - mean(preds_1))^2)
Zdf_Regression_1 <- AnzPred - 1
MSS_Regression_1 <- round(SS_Regression_1 / Zdf_Regression_1, 2)
# Berechnung der Quadratsummen des Fehlers (Residuals)
Residuals_1    <- CPS85$wage - preds_1
SS_Residuals_1 <- sum(Residuals_1^2)
Ndf_Residuals_1 <- nrow(CPS85) - AnzPred
MSS_Residuals_1 <- round(SS_Residuals_1 / Ndf_Residuals_1, 2)
# Berechnung der Teststatistik
F_Wert         <- round(MSS_Regression_1 / MSS_Residuals_1, 2)
# Berechnung der totalen Quadratsumme
SS_Total_1     <- sum((CPS85$wage - mean(CPS85$wage))^2)
CPS85_Total   <- nrow(CPS85) - 1
# Vergleich mit den Ergebnissen der ANOVA
pander::pander(anova(model_1))
```

Table 12: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
educ	1	2053	2053	90.85	5.474e-20
Residuals	532	12023	22.6	NA	NA

Das Ergebnis zeigt uns, dass Modell 1 im Vergleich zum Mittelwertsmodell zu einer statistisch signifikanten Fehlerreduktion führt. Bei der händischen Berechnung der Prüfgrößen erhalten wir für die mittlere Quadratsumme der Regression (also der Varianz der Werte die durch das Modell vorhergesagt werden) einen Wert

von $\$MSS_{\text{Regression}} = \$ 2053.29$, welcher ident mit dem Wert der ANOVA-Tabelle ist.

Die restlichen Kennwerte stimmen auch mit dem Ergebnis der ANOVA überein ($MSS_{\text{Residual}} = 22.6$, $F(1,532) = 90.85$).

Wird das Modell 1 erweitert (auf Modell 2), stellt sich die Frage, ob diese Erweiterung im statistischen Sinn zu einer signifikanten Verbesserung führt. Bei diesem Vergleich wird nun die Änderung (Change Statistic) zwischen Modell 1 und Modell 2 auf Signifikanz geprüft.

```
# ANOVA Tests auf signifikante Änderungen model_1 vs model_2 (Änderung signifikant?)
pander::pander(anova(model_1, model_2))
```

Table 13: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
532	12023	NA	NA	NA	NA
531	11233	1	790.6	37.37	1.893e-09

Zum Verständnis dieser Statistik greifen wir kurz zurück auf die verschiedenen Möglichkeiten der Berechnung von Korrelationskoeffizienten zurück. Diese sind:

1. Pearson Korrelationskoeffizient (r_{xy}): entspricht der Kovarianz der z -transformierten Variablen.
2. Partielle Korrelationskoeffizient ($r_{xy.z}$): ist die bivariate Korrelation zweier Variablen, welche mittels linearer Regression vom Einfluss einer Drittvariablen bereinigt wurden.
3. Semipartialkorrelation ($sr_{k \cdot x_j}$): zwischen Kriterium und dem j -ten Prädiktor ergibt sich als Korrelation von y mit dem Residuum x_j^* der linearen Regression des j -ten Prädiktors auf den anderen Prädiktor. Mit anderen Worten, die Semipartialkorrelation gibt den alleinigen Beitrag eines Prädiktors x_j (bereinigt um die gemeinsamen Anteile mit den restlichen Prädiktoren) am Kriterium an. Das Quadrat dieses Koeffizienten wird unter anderm auch als Nützlichkeit des Prädiktors U_k bezeichnet und findet sich z.B. in SPSS als R^2_{change} wieder. Formal: $sr_{k \cdot 12 \dots (k-1)}^2 = R^2_{y,12 \dots k} - R^2_{y,12 \dots k-1}$

```
# Korrelationen, Partial- und Semipartialkorrelationen
Korr_Data      <- data.frame(wage = M$wage, educ = M$educ, exper = M$exper)
PearsonKorr    <- cor(Korr_Data)
ModVgl_Korr    <- pander::pander(PearsonKorr)
R2Change_mod_1 <- PearsonKorr[2]^2
# Partial Korrelation zwischen "wage" und "educ" gegeben "exper"
PartKorr_1     <- ppcor::ppcor.test(Korr_Data$wage, Korr_Data$educ, Korr_Data$exper)
ModVgl_PartKorr_1 <- pander::pander(PartKorr_1)
# Partial Korrelation zwischen "wage" und "exper" gegeben "educ"
PartKorr_2     <- ppcor::ppcor.test(Korr_Data$wage, Korr_Data$exper, Korr_Data$educ)
ModVgl_PartKorr_2 <- pander::pander(PartKorr_2)
# Semi-Partial (part) Korrelation zwischen "wage" und "educ" gegeben "exper"
SemiPartKorr_1  <- ppcor::spcor.test(Korr_Data$wage, Korr_Data$educ, Korr_Data$exper)
ModVgl_SemParKorr_1 <- pander::pander(SemiPartKorr_1)
# Semi-Partial (part) Korrelation zwischen "wage" und "exper" gegeben "edu"
SemiPartKorr_2  <- ppcor::spcor.test(Korr_Data$wage, Korr_Data$exper, Korr_Data$educ)
ModVgl_SemParKorr_2 <- pander::pander(SemiPartKorr_2)
R2Change_mod_2   <- round(SemiPartKorr_2$estimate^2, 3)
pander::pander(summary(model_2))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.904	1.219	-4.024	6.564e-05
educ	0.926	0.0814	11.37	5.563e-27

	Estimate	Std. Error	t value	Pr(> t)
exper	0.1051	0.0172	6.113	1.893e-09

Table 15: Fitting linear model: wage ~ educ + exper

Observations	Residual Std. Error	R ²	Adjusted R ²
534	4.599	0.202	0.199

Im vorliegenden Beispiel sind daher die beiden Nützlichkeitsmaße $U_{educ} = 0.146$ und $U_{exper} = 0.056$ von Interesse. Ersteres bedeutet, dass die Varianzaufklärung aufgrund der Verwendung der Variablen *educ* 14.6% ist. Wird im Modell dann noch der Prädiktor *exper* aufgenommen, werden zusätzliche 5.6% an Varianz des Kriteriums *wage* erklärt. Insgesamt werden somit $R^2 = 0.202$ oder 20.2% der Varianz des Kriteriums erklärt. Der Test ($t(531) = 11.37, p < .001$) bestätigt für den Prädiktor *educ*, sowie ($t(531) = 6.11, p < .001$) für den Prädiktor *exper* die statistische Signifikanz.

Aufgabe MLR 1

Öffne ein neues R-Script und kopiere die bereits bekannte Kopfzeile in diese Datei. Speichere anschließend das Skript unter dem Namen *SLR_Aufgabe2.R*. Bearbeite nun folgende Aufgabenstellungen:

- Lade die Datei “*Album Sales 2.dat*”
- erstelle ein lineares Modell zur Vorhersage der Verkaufszahlen (*sales*) durch die Variable *adverts*.
- erstelle ein weiteres lineares Modell zur Vorhersage der Verkaufszahlen (*sales*) durch die Variable *adverts*, *airplay* und *attract*.
- Zeige die Ergebnisse des ersten Modells an.
- Zeige die Ergebnisse des zweiten Modells an.
- Vergleiche die beiden Modelle mit einer ANOVA und interpretiere die Ergebnisse.
- Berechne zur Überprüfung der Multikolinearität den Kennwert *Tol* und *VIF* (verwende die Funktion *vif()*). Hinweis: die Toleranz ist der Kehrwert von *VIF*)

Lösung Aufgabe MLR 1

Wahl relevanter Prädiktoren

Eine wichtige Frage bei der Modellerstellung betrifft die Wahl der besten Prädiktoren. Prinzipiell muss bereits im Vorfeld der statistischen Analyse bestimmt werden, welche Merkmale für die Modellierung der abhängigen Variablen am geeignetsten sind. Ausreichende theoretische und praktischen Kenntnisse sind daher unbedingt erforderlich. Die Erfassung von potentiellen Prädiktoren ist stets mit zeitlichen und/oder finanziellen Aufwand verbunden. Prädiktoren sind dann gut geeignet, wenn Sie folgende Eigenschaften erfüllen:

1. jeder Prädiktor erklärt möglichst viel der Variabilität des Kriteriums.
2. die Prädiktoren (z.B. x_1 und x_2) sind im günstigsten Fall voneinander unabhängig ($r(x_1, x_2) \approx 0$)

Diese Eigenschaft kann man durch eine einfache paarweise Korrelation prüfen. Vor allem wenn die zweite Eigenschaft nicht gegeben ist, also wenn einen hohen Korrelationen zwischen zwei Prädiktoren vorliegt, wird es bei der Modellierung zu maßgeblichen Problemen (Multikollinearität) kommen (siehe: Voraussetzungen der multiplen Regression).

Neben der Frage nach der Güte einzelner Prädiktoren ist es auch wichtig sich Gedanken über die Anzahl der zu verwendenden Prädiktoren zu machen. Einerseits führt trivialerweise eine höhere Anzahl von Prädiktoren auch zu einer besseren Aufklärung der Varianz im Kriterium. Ausgenommen von Prädiktoren die in keiner Beziehung zum Kriterium stehen, wird jeder zusätzliche Prädiktor mehr oder weniger der verbleibenden Varianz erklären. In den meisten Fällen ist es aber aus zeitlichen/finanziellen oder sonstigen Gründen nicht sinnvoll, eine möglichst große Menge an Prädiktorvariablen zu erheben.

Werden zu viele erklärende Variablen zur Spezifizierung eines Modells verwendet, wird die tatsächliche (geringere) Anpassungsgüte verschleiert. Das Modell wird zwar besser auf die Daten der Stichprobe angepasst, allerdings besteht aufgrund fehlender Generalität keine Übertragbarkeit auf die Grundgesamtheit. Grundsätzlich sollte wie bereits erwähnt die Wahl der Prädiktoren auf theoretisch und praktisch fundierten Grundlagen erfolgen. Welche der zur Verfügung stehenden Prädiktoren im Endeffekt für das Modell verwendet werden, kann anhand der Modellvergleiche auch im statistischen Sinn evaluiert werden.

Bei der bisher besprochenen Vorgehensweise der Modellerstellung obliegt es dem Analysten, die zu verwendenden Prädiktoren zu bestimmen. Eine weitere Möglichkeit bietet die sogenannte sequentielle Vorgehensweise, bei der die Ein- und Ausschlusskriterien für Prädiktoren durch statistische Kriterien getroffen werden.

Sequentielle Modellbildung

In manchen Fällen sind nicht ausreichende theoretische Grundlagen und Erfahrungswerte bezüglich der Wirksamkeit und Wichtigkeit von Prädiktoren vorhanden. In solchen Fällen kann ein exploratives Vorgehen bei der Modellerstellung sehr hilfreich sein. Die nachfolgend beschriebene sequentielle Modellierung entspricht einem solchen Ansatz.

Bei der sequentiellen Modellbildung wird ein Modell schrittweise mit unabhängigen Variablen erweitert. In der Regel wird jene Variable, die das R^2 am meisten vergrößert und damit die Vorhersage am meisten verbessert hinzugefügt.

Abhängig von der Anzahl der verfügbaren Prädiktoren wird die Bildung neuer Modelle entweder abgebrochen, wenn weitere Variablen keinen weiteren statistischen signifikanten Beitrag zur Varianzaufklärung mehr leisten, oder wenn keine weiteren Variablen zur Verfügung stehen.

Aufgrund der statistischen (maschinellen) Entscheidung über die Verwendung von Prädiktoren, wird diese Vorgehensweise vielfach kritisiert. Nehmen wir in einem sehr einfachen Beispiel einmal an, es stehen 2 Prädiktoren (x_1, x_2) zur Vorhersage der abhängigen Variablen zur Verfügung. Der Prädiktor x_1 klärt geringfügig weniger Varianz des Kriteriums auf als Prädiktor x_2 , ersterer ist aber inhaltlich sinnvoller, leichter zu interpretieren und vor allem weit kostengünstiger zu erfassen. Bei der sequentiellen Methode könnte aber aufgrund des Abbruchkriteriums (Signifikanz des Beitrags) genau dieser Prädiktor vom Modell ausgeschlossen werden.

Bei der sequentiellen Methode unterscheidet man noch unterschiedliche Vorgehensweisen hinsichtlich des Hinzufügens/Entfernens von Variablen:

1. Schrittweise (STEPWISE): Diese Methode ist ähnlich wie "Vorwärts"-Selektion, es wird aber zusätzlich bei jedem Schritt getestet, ob die am wenigsten "nützliche" Variable entfernt werden soll.
2. Vorwärts-Selektion (FORWARD): Die Variablen werden sequenziell in das Modell aufgenommen. Diejenige unabhängige Variable, welche am stärksten mit der abhängigen Variable korreliert wird zuerst zum Modell hinzugefügt. Dann wird jene der verbleibenden Variablen hinzugefügt, die die höchste partielle Korrelation mit der abhängigen Variablen aufweist. Dieser Schritt wird wiederholt, bis sich die Modellgüte (R-Quadrat) nicht weiter signifikant erhöht oder alle Variablen ins Modellaufgenommen worden sind.
3. Rückwärts-Elimination (BACKWARD): Zunächst sind alle Variablen im Regressionsmodell enthalten und werden anschließend sequenziell entfernt. Schrittweise wird immer diejenige unabhängige Variable entfernt, welche die kleinste partielle Korrelation mit der abhängigen Variable aufweist, bis entweder keine Variablen mehr im Modell sind oder keine die verwendeten Ausschlusskriterien erfüllen. Im Unterschied zur STEPWISE-Methode wird nicht mehr geprüft, ob die am wenigsten nützliche Variable entfernt werden soll - diese bleibt somit im Modell!

Diese Methoden unterscheiden sich von der sogenannten Einschlussmethode (ENTER), bei der alle Variablen gleichzeitig in das Modell eingefügt werden. Diese Methode wird angewendet, wenn das Modell auf theoretischen Überlegungen basiert. Das heißt, sie eignet sich um Theorien zu testen, während die übrigen Methoden eher im Rahmen explorativer Studien eingesetzt werden.

Modellvergleich durch AIC

Nach einer (explorativen) Analyse der Daten und der Wahl einer passenden Modellklasse, geht es darum das bestmögliche Modell zu den vorliegenden Daten zu finden (siehe FUB). Daher stellt sich die Frage, was "bestmögliches" Modell bedeutet und wie ein solches bestimmt werden kann. In diesem Zusammenhang wird der Gedanke aufgegriffen, dass mit keinem Regressionsmodell die Realität eins zu eins abgebildet werden kann. Nimmt man zu viele erklärende Variablen auf, läuft man in Gefahr das Modell zu "overfitten" (überanpassen). Ein überangepasstes Modell erklärt die zum Schätzen verwendete abhängige Variable meist sehr gut, schneidet jedoch in der Vorhersage von Daten außerhalb der verwendeten Stichprobe häufig schlecht ab. Auf der anderen Seite kann ein Modell auch "underfitted" sein, d.h. die aufgenommenen unabhängigen Variablen können die abhängige Variable nur sehr unzureichend erklären.

Das Thema der Modellselektion ist ein allgegenwärtiges in der Statistik/ Regressionsanalyse. Dennoch gibt es keine absoluten, objektiven Kriterien anhand derer entschieden werden kann, ob das eine oder das andere Modell gewählt werden sollte. Vielmehr existieren viele verschiedene Verfahren, die versuchen zwischen möglichst viel Erklärungsgehalt des Modells und möglichst wenig Komplexität (siehe dazu Ockhams Rasiermesser) abzuwegen.

In einem Artikel von (Yamashita 2007) wurden folgende Methoden:

- a. Partial F
- b. Partial Correlation
- c. Semi-Partial Correlation
- d. Akaike Information Criteria (AIC)

für den Vergleich von Regressionsmodellen untersucht. Die Autoren schließen aus den Ergebnissen ihrer Untersuchung, dass alle Methoden zu den gleichen Ergebnissen, d.h. zur gleichen Modellentscheidung gelangen. Da aber der AIC einerseits leicht zu interpretieren und andererseits auch auf nichtlineare Modelle und Modelle die auf nicht normalverteilten Daten beruhen zu erweitern ist, wird die Anwendung dieses Kriteriums empfohlen.

Das AIC dient also dazu, verschiedene Modellkandidaten zu vergleichen. Dies geschieht anhand des Wertes der log-Likelihood, der umso größer ist, je besser das Modell die abhängige Variable erklärt. Um nicht komplexere Modelle als durchweg besser einzustufen wird neben der log-Likelihood noch die Anzahl der geschätzten Parameter als Strafterm mitaufgenommen.

$$AIC_k = 2 \cdot |k| - 2 \cdot \hat{L}_k \quad (3)$$

In der Formel steht k für die Anzahl der im Modell enthaltenen Parameter und \hat{L}_k für den Wert der log-Likelihoodfunktion.

Das Modell mit dem kleinsten AIC wird bevorzugt.

Das AIC darf nicht als absolutes Gütemaß verstanden werden. Auch das Modell, welches vom Akaike Kriterium als bestes ausgewiesen wird, kann eine sehr schlechte Anpassung an die Daten aufweisen. Die Anpassung ist lediglich besser als in den Alternativmodellen.

Die praktische Bedeutung soll anhand eines einfachen Beispiels und der Verwendung des Kriteriums bei unseren Beispieldaten erläutert werden.

Nehmen wir an, dass drei Modellvergleiche (mod_1, mod_2, mod_3) folgende AIC-Werte ergeben haben:

$AIC_1 = 100, AIC_2 = 102, AIC_3 = 110$. Berechnet man $e^{(AIC_{min} - AIC_i)/2}$, kann das Ergebnis folgendermaßen interpretiert werden:

- Beim mod_2 ist es um das $e^{(100-102)/2} = 0.368$ -fache wahrscheinlicher den Informationsverlust zu verringern als bei Modell 1 (mod_1).
- Beim mod_3 ist es um das $e^{(100-110)/2} = 0.007$ -fache wahrscheinlicher den Informationsverlust zu verringern als bei Modell 1 (mod_1).

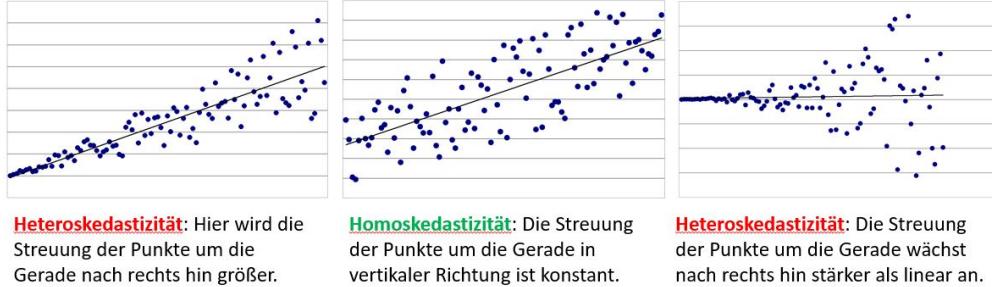


Figure 16: Abbildung 14: Homoskedastizität vs. Heteroskedastizität

Bei diesem Beispiel würde man also *mod_3* für weitere Betrachtungen ausschließen. Nachdem aber die Modelle *mod_1* und *mod_2* sehr nahe beisammen liegen, ist es mit den vorliegenden Daten nicht möglich, eine klare Entscheidung für eines der beiden Modelle zu treffen.

Man könnte durchaus noch zusätzliche Daten erheben um dadurch eventuell eine klarere Trennung der beiden Modelle (*mod_1*, *mod_2*) zu erkennen. Ist das nicht möglich, könnte man beide Modelle mit der relativen likelihood gewichten und auf eine statistische Signifikanz testen, oder davon ausgehen, dass mit den vorliegenden Daten eine Modellwahl eben nicht eindeutig zu treffen ist.

Kreuzvalidierung

Betrachten wir im Folgenden ein Modell (*mod_1*) mit den Prädiktoren *sector* (Berufsgruppe), *exper* (Erfahrung), sowie das um den Prädiktor *age* (Alter) erweiterte Modell (*mod_2*).

Die Vorhergehensweise bei der Kreuzvalidierung ist relativ simpel:

1. Erstelle ein/mehrere Modell(e) und berechne die jeweiligen Modellparameter b_i^j (mit $j = j\text{-tes Modell}$ und $i = i\text{'ter Parameter}$) mit einer Teilmenge der zur Verfügung stehenden Daten (z.B. *Training_Data* $\subset DF$).
2. Verwende die restlichen Daten um mit den entsprechenden Modellen Vorhersagen zu berechnen.
3. Berechne die Differenz der beobachteten Daten und der vorhergesagten Daten. Diese Differenz entspricht dem Fehler des Modells ($\rightarrow \epsilon_i$).
4. Berechne den mittleren quadratischen Fehler der Differenzen.

Voraussetzungen MLR

Folgende Voraussetzungen müssen/sollten bei der linearen Modellierung mit mehreren Prädiktoren erfüllt sein, damit die Ergebnisse auch sinnvoll interpretiert werden können (Bemerkung: im folgenden sei die abhängige Variable y und die Prädiktoren mit den Zahlen $1, 2, \dots, k$ bezeichnet):

1. **Lineare Beziehung** zwischen den Variablen (keine Ausreißer): eine einfache Prüfung erfolgt visuell mit Streudiagrammen, wobei alle Beziehungen, also $r_{y \cdot 1}, r_{y \cdot 2}, \dots, r_{y \cdot k}, \dots, r_{1 \cdot 2}, r_{1 \cdot k}, \dots, r_{(k-1) \cdot k}$ zu betrachten sind!
2. **Varianzgleichheit der Residuen** (Homoskedastizität): auch diese Voraussetzung kann visuell geprüft werden. Dabei wird ein Streudiagramm der Residuen erstellt, in welchem auf der x-Achse die standardisierten vorhergesagten Werte und auf der y-Achse die standardisierten Residuen aufgetragen werden. Heteroskedastizität liegt vor, wenn die Punktwolke nicht gleichverteilt um die Gerade liegen!

Bekannte Verfahren, um die Nullhypothese „Homoskedastizität“ zu überprüfen sind der:

- * Levene-Test
- * Goldfeld-Quandt-Test
- * White-Test
- * Glejser-Test

* RESET-Test

* Breusch-Pagan-Test

3. **Normalverteilung der Residuen:** mittels Histogramm der Fehler zu prüfen - sollte halbwegs normalverteilt sein mit einem Erwartungswert des Fehlers $E(\varepsilon) = 0$.
4. **Unabhängigkeit der Residuen** (keine Autokorrelation): verletzt wird diese Voraussetzung, wenn aufeinanderfolgende Werte abhängig sind (z.B. auf einen hohen Wert folgt ein hoher Wert, etc.). Vor allem bei Längsschnittdaten ein Thema, bei welchen die Prüfung durch die Durbin-Watson-Methode empfohlen wird. Es gilt: $d = \frac{\sum_i (e_i - e_{i-1})^2}{\sum_i (e_i)^2}$ mit $d \approx 2$, Werte zwischen $1.5 < d < 2.5$ sind noch akzeptabel.
5. **Vollständig spezifizierte Modelle:** werden maßgebliche Prädiktoren nicht im Modell berücksichtigt, wird es auch kaum gelingen, die Varianz des Kriteriums zufriedenstellend zu erklären. Andererseits bewirken Modelle mit vielen Prädiktoren, dass die β -Gewichte entsprechend klein werden. Bei derartigen Gegebenheiten ist die Stichprobe entsprechend groß zu wählen.
6. **Keine Multikollinearität:** Multikollinearität bedeutet, dass Prädiktoren existieren, die hoch miteinander korrelieren (z.B. $r_{1,2} > 0.8$). Damit wird es für das Modell schwer, den jeweiligen Beitrag den Prädiktoren zuzuordnen. Besteht rein das Interesse an maximaler Varianzaufklärung des Kriteriums, ist eine hohe Multikollinearität zu vernachlässigen - die β -Gewichte der einzelnen Prädiktoren darf man dann allerdings nicht interpretieren. Spielen jedoch gerade diese eine wichtige Rolle, kann man entweder hoch korrelierte Prädiktoren zusammenfassen (eventuell Faktorenanalyse/Clusteranalyse vorher durchführen), oder entsprechende Prädiktoren ausschließen. Allerdings sollte man vor dem Ausschluss von Prädiktoren diese auf eventuelle Suppressionseffekte prüfen.
 - *Negative und reziproke Suppression:* man spricht von Suppressionseffekten, wenn ein Prädiktor aus einem anderen Prädiktor irrelevante Varianz unterdrückt (suppression) und dadurch die Beziehung zwischen diesem Prädiktor und dem Kriterium erhöht. Solche Effekte können durchaus beträchtlich sein und u.U. auch einen Prädiktor, der nichts mit dem Kriterium an sich zu tun hat ($r_{y,k} \approx 0$), als wichtigen Bestandteil des Modells werden lassen. Die Aufnahme des Suppressors in das Regressionsmodell hat somit den Effekt, den anderen Prädiktor von diesen Fehlereinflüssen zu bereinigen. Erkennbar sind Suppressionseffekte einerseits durch Vorzeichenwechsel bei Korrelationen (Nullter Ordnung, also der Produkt-Moment-Korrelation) vs. β -Gewichten (negative Suppression, bzw. NET-Suppression). D.h., dass für nicht-negative Validitäten⁸ ist der Prädiktor 2 ein negativer Suppressor, falls seine partielle Steigung negativ ist, d. h., falls $B_2 < 0$. Eine *reziproke Suppression* liegt vor, wenn für nicht-negative Validitäten die Korrelation der Prädiktoren negativ ist, d. h., falls $r_{1,2} < 0$. Weitere Details zu Suppressionseffekten siehe Literatur und Diskriminanzanalyse.
7. **Hohe Reliabilität der Prädiktoren und des Kriteriums:** Variablen sind hochreliabel, wenn sie weitgehend frei von Zufallsfehlern sind, also bei Messwiederholung ähnliche Ergebnisse liefern.
8. **Keine Varianzeinschränkung:** eine Einschränkung führt i.A. zu eingeschränkten (niedrigeren) Korrelationen. Z.B.: aus 500 Personen werden 100 augrund eines Aufnahmeverfahrens zu einem Studium zugelassen. Will man die Validität des Aufnahmeverfahrens anhand der Beziehung Studienerfolg und Leistung beim Aufnahmetest prüfen, wird es aufgrund der eingeschränkten Variabilität durch die Aufnahmekriterium zu einer Unterschätzung kommen.
9. **Unabhängigkeit der Beobachtungseinheiten:** eine Verletzung dieser Voraussetzung, kann zu einer maßgeblichen Reduktion der Teststärke des Modells führen. Z.B. soll die Teamorientierung in einem Unternehmen untersucht werden. Diese wird sicher zwischen den einzelnen Personen variieren, aber darüber hinaus kann diese auch abhängig von der Abteilung sein, in welcher Personen arbeiten. Die Variabilität kann dadurch bei bestimmten Abteilungen stark eingeschränkt sein, was einer Reduktion des Stichprobenumfangs und damit einer Teststärkenreduktion gleichzusetzen ist. In solchen Fällen könnte man eine Multilevel-Analyse (gemischtes hierarchisches Modell) einsetzen!

⁸Die Korrelationen des Kriteriums mit den Prädiktorvariablen bezeichnen wir als Validitäten, d. h. die Validität der j -ten Prädiktorvariablen ist gleich ihrer Korrelation mit dem Kriterium.

Zusammenfassend lässt sich festhalten, dass eine Verletzung einer/mehrerer dieser Voraussetzungen meistens dazu führt, dass die Genauigkeit der Vorhersage gemindert wird. Relativ einfach zu prüfen sind die ersten drei Voraussetzungen (graphisch, Kennwerte wie Korrelation, etc.). Bei der Überprüfung der restlichen Voraussetzung muss man i.A. auf entsprechende statische Verfahren zurückgreifen, die hier aber nicht näher besprochen werden. Einen Überblick über die Möglichkeiten zur Überprüfung der Voraussetzungen finden Sie z.B. unter (UZH 2018), oder MR2 - (Hemmerich 2018).

Lösungen

Aufgabe SLR 1 Lsg

[zurück zur Aufgabenstellung](#)

Aufgabe MLR 1 Lsg

```
album2      <- read.delim("Daten/Album Sales 2.dat", header = TRUE)
# Erstes Modell
  albumSales.2 <- lm(sales ~ adverts, data = album2)
# zweites Modell
  albumSales.3 <- lm(sales ~ adverts + airplay + attract, data = album2)
# Ausgabe Ergebnisse
  pander::pander(summary(albumSales.2))
  pander::pander(summary(albumSales.3))
# Modellvergleich
# library(car) fÃ¼r VIF
  anova(albumSales.2, albumSales.3)
# library(car) fÃ¼r VIF
  Tol <- 1/car::vif(albumSales.3)
  VIF <- car::vif(albumSales.3)
```

[zurück zur Aufgabenstellung](#)

Teil III: Dummy-Codierung

```
# Field R - S302 ff

rm(list = ls())
graphics.off()
if (!require("pacman")) install.packages("pacman")
pacman::p_load(pander)
options(digits=2)
```

Kategoriale Prädiktoren

Bei linearen Modellen ist häufig neben intervallskalierten Prädiktorvariablen auch die Verwendung von kategorialen Variablen von Interesse. So lange der verwendete Prädiktor nur zwei Ausprägungen hat (z.B. männlich/weiblich, Ja/Nein, etc.), stellt dies auch kein Problem dar.

Beispiel mit mehrstufigen kategorialen Prädiktor

Während eines dreitägigen Musikfestivals wurde bei einer Anzahl freiwilliger TeilnehmerInnen der "Hygienezustand" gemessen (Variablen *day1*, *day2*, *day3*). Der Wertebereich der Messung liegt zwischen 0 und 4, mit 0 = smell like s.t., bis 4 = smell like freshly baked bread. Darüber hinaus wurden die TeilnehmerInnen über ihre jeweilige Zuordnung zu einer bestimmten, persönlich bevorzugten Musikrichtung (*music*) befragt. Bei dem Festival gaben die TeilnehmerInnen insgesamt vier verschiedenen Musikrichtungen an: *Metaller*, *Crusty*, *Indie*, *NMA* (= No Music Affiliation). Nach Erfassung der Daten wurde die Differenz der Hygienewerte zwischen dem letzten und dem ersten Tag des Festivals berechnet und in der Variablen *change* gespeichert:

	ticknumb	music	day1	day2	day3	change
1	2111	Metaller	2.65	1.35	1.61	-1.04
2	2229	Crusty	0.97	1.41	0.29	-0.68
10	2504	No Musical Affiliation	1.11	0.44	0.55	-0.56
12	2510	Crusty	0.82	0.2	0.47	-0.35
14	2515	No Musical Affiliation	1.76	1.64	1.58	-0.18
21	2549	Crusty	2.17	0.7	0.76	-1.41

Offenbar liegt bei der Variablen *music* ein Faktor mit mehr als 2 Stufen (es sind 4) vor. Da die Verwendung von kategorialen Variablen in einem linearen Modell eine Stufenanzahl von 2 voraussetzt, kann durch geschicktes Kodieren der Variablen diese Voraussetzung auch für mehrstufige Variablen erreicht werden.

Dummy Kodierung

Man nennt diesen Vorgang auch **Dummy Kodierung**. Die Vorgehensweise ist dabei:

1. Die Anzahl der neuen (Dummy) Variablen ist die Anzahl der Stufen des Prädiktors - 1 ($N_{\text{DummyVars}} = N_{\text{Stufen}} - 1$)
2. Man legt so viele neue Variablen (Dummy-Variablen) an, wie man (im ersten Schritt) als Anzahl der Gruppen berechnet hat.
3. Wahl einer Bezugsgruppe (Baseline-Bedingung). üblicherweise die Kontrollgruppe, falls keine vorhanden wählt man am besten die Gruppe, in der die meisten Personen/Fälle vorliegen.
4. Allen Dummy-Variablen für die gewählte Baselinegruppe den Zahlenwert 0 zuweisen.
5. Der ersten Dummy-Variablen für die erste Gruppe die man gegen die Baselinegruppe vergleichen will den Wert 1 zuweisen, den restlichen Gruppen den Wert 0.
6. Wiederholung des Schrittes 5, bis alle Dummy-Variablen entsprechend codiert wurden.
7. Alle Dummy-Variablen ins Modell aufnehmen!

	DVar1	DVar2	DVar2
Crusty	1	0	0
Indie Kid	0	1	0
Metaller	0	0	1
No Affiliation	0	0	0

Bei der linearen Modellierung in R werden kategoriale Daten im Modell automatisch Dummy-Kodiert. Will man jedoch eine spezielle Anordnung der Gruppen, sollte man wissen, wie eine händische Kodierung einfach durchgeführt werden kann. Im folgenden Code werden diese Möglichkeiten dargestellt:

```
# Automatisch ohne Bezeichnung der Dummyvariablen
contrasts(DF$music) <- contr.treatment(4, base = 4)
# Manuel mit Bezeichnung der Dummyvariablen
crusty_v_NMA <- c(1,0,0,0)
```

```

indie_v_NMA <- c(0,1,0,0)
metal_v_NMA <- c(0,0,1,0)
contrasts(DF$music) <- cbind(crusty_v_NMA, indie_v_NMA, metal_v_NMA)
pander(attr(DF$music, "contrasts"))

```

	crusty_v_NMA	indie_v_NMA	metal_v_NMA
Crusty	1	0	0
Indie Kid	0	1	0
Metaller	0	0	1
No Musical Affiliation	0	0	0

Modellierung mit kategorialen Variablen

Sind die Dummy-Variablen angelegt, kann damit auch das Modell erstellt werden. Im nachfolgenden Beispiel wird die Variable *change* durch die Dummy-Kodierten Prädiktoren modelliert. Die erste Tabelle zeigt die durchschnittlichen *change*-Werte pro Musikzugehörigkeitsgruppe.

```
pander(round(tapply(DF$change, DF$music, mean, na.rm = TRUE), 3))
```

Crusty	Indie Kid	Metaller	No Musical Affiliation
-0.966	-0.964	-0.526	-0.554

```

mod_dummy_1 <- lm(change ~ music, data = DF)
AllRes      <- summary(mod_dummy_1)
pander(summary.lm(mod_dummy_1))

```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.5543	0.09036	-6.134	1.154e-08
musiccrusty_v_NMA	-0.4115	0.167	-2.464	0.01518
musicindie_v_NMA	-0.41	0.2049	-2.001	0.04771
musicmetal_v_NMA	0.02838	0.1603	0.177	0.8598

Table 21: Fitting linear model: change ~ music

Observations	Residual Std. Error	R ²	Adjusted R ²
123	0.6882	0.07617	0.05288

Wesentliche Kennzahlen des Ergebnisses:

- $R^2 = 0.076$: 7.6% der Variabilität in der Änderung der Hygenewerte zwischen ersten und dritten Tag (*change*) werden durch die Zugehörigkeit zu einer Musikgruppe erklärt.
- $F(3, 119) = 3.27; p = .053$ gibt an, dass die 7.6% Varianzaufklärung statistisch signifikant ist. Das Modell ist also signifikant besser als kein Modell zu verwenden.
- *musiccrusty_vs_NMA*: Differenz zwischen der *NMA* und *crusty* Gruppe. Betrachtet man die Differenz der Mittelwerte (siehe obige Tabelle) zwischen *crusty* - *NMA* = $-0.966 - (-0.554) = -0.412$, stellt man fest, dass diese Differenz dem Estimate, also dem *b*-Koeffizienten entspricht. Offenbar ist die Änderung der Hygenewerte bei *crusty* höher als bei der *NMA* → *crusties* sind größere Schweindln wie die *NMA* Leute. **Die *b*-Werte geben also die relative Änderung zur Baselinegruppe an!**

- $t = -2.46, p = .015$: tested ob die Differenz signifikant unterschiedlich zu einer Null-Differenz (kein Unterschied) in den Hygienebedingungen ist. Im vorliegenden Fall handelt es sich um eine signifikante Abnahme der Hygienewerte, wenn man von *NMA* auf *crusty* wechselt.

Die restlichen Koeffizienten sind in gleicher Weise zu interpretieren.

Teil IV: Mediator-Analyse

```
# Field SPSS 5 - S497
rm(list = ls())
graphics.off()
if (!require("pacman")) install.packages("pacman")
pacman::p_load(pander, SciViews)

# install.packages("sjPlot")
# install.packages("yaml", dependencies = TRUE)
# install.packages("sjPlot", dependencies = TRUE)
# install.packages("sjmisc", dependencies = TRUE)
# install.packages("stringi", dependencies = TRUE)
# install.packages("httpuv", dependencies = TRUE)
# require(yaml)
# require(sjPlot)
# require(sjmisc)
#
options(digits=3)
# Verzeichnisse InitialisierenCPS85
```

Mediation

Sowohl bei Moderation als auch bei Mediation geht es um die Zusammenhänge zwischen drei Variablen X , Y und M . Interessant dabei ist der Effekt eines Prädiktors oder Faktors X auf einen abhängige Variable Y . Das kann in einem Regressionsmodell mit X als unabhängige und Y als abhängige Variable untersucht werden. Zusätzlich gibt es eine dritte Variable M . Sie ist entweder der Moderator oder der Mediator.

Bei der Mediation steht der Mediator (M) sowohl in Beziehung zu X als auch zu Y . Der **direkte Effekt** zwischen X und Y wird durch den **indirekten Effekt** über M erklärt, also durch $X \rightarrow M \rightarrow Y$.

Konzeptuelle Modell

Eine Mediatorvariable ist eine Variable, welche die Beziehung zweier anderer Variablen vermittelt/erklärt.

Um ein Mediatormodell zu prüfen, sind eine Reihe von Regressionsanalysen erforderlich:

1. Eine Regression mit X_2 als Prädiktor und Y als Kriterium. Der daraus resultierende Regressionskoeffizient b_1 entspricht dem **c** im konzeptuellen Modell.
2. Eine Regression mit X_2 als Prädiktor und X_1 als Kriterium. Der daraus resultierende Regressionskoeffizient b_2 entspricht dem **a** im konzeptuellen Modell.
3. Eine Regression mit X_1 und X_2 (= Mediator) als Prädiktoren und Y als Kriterium. Der daraus resultierende Regressionskoeffizient des Prädiktors b_3 entspricht dem **c'** und der des Mediators b_4 entspricht dem **b** im konzeptuellen Modell.

Folgende Ergebnisse dieser Modelle würden für den Effekt des Mediators sprechen:

Einfache linear Beziehung



Totaler Effekt = Direkter Effekt + Indirekter Effekt

$$c = c' + a \cdot b$$

Linear Beziehung mit Mediator

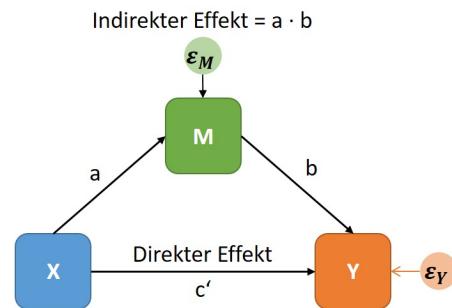


Figure 17: Abbildung 15: Konzeptuelles Modell Mediation

1. Die Regressionskoeffizienten b_1, b_2 und b_3 (also die Pfade **a**, **b**, **c** im Modell) zeigen ein signifikantes Ergebnis.
2. Der Regressionskoeffizient b_3 muss kleiner sein als b_1 ($c' < c$ im Modell)

Obwohl die Regressionsanalyse die grundlegende Idee der Mediationsanalyse gut zeigt, hat sie den Nachteil, dass der Effekt (Bedeutsamkeit) der Reduktion nicht wirklich klar ersichtlich ist.

Häufig findet man noch folgende Kriterien (Baron and Kenny's Method) für die Entscheidung ob eine Mediation vorliegt:

Eine Mediation liegt vor, wenn die Beziehung zwischen Prädiktor und Kriterium ohne Berücksichtigung des Mediators signifikant ($p < .05$) und mit Berücksichtigung des Mediators nicht mehr signifikant ist.

Diese Entscheidungsgrundlage entspricht dem NHST-Testen (all or nothing) und kann zu *maßgeblichen Fehlentscheidungen* führen, denn:

- Ein b -Wert kann sich unter Umständen nur um ein wenig ändern, der dazugehörige p -Wert kann sich dabei jedoch ohne weiteres von signifikant auf nicht signifikant ändern!
- Bei einer großen Änderung des b -Wertes kann es aber auch durchaus vorkommen, dass beide signifikant bleiben!

Als alternative Möglichkeiten zur Entscheidungsfindung haben sich folgende Verfahren bewährt:

- **Sobel Test:** testet den indirekten Effekt (kombinierter Pfad **a** und **b**). Liefert dieser Test ein signifikantes Ergebnis, liegt eine signifikante Mediation vor. Allerdings zeigt dieser Test folgende Problembereiche: + Die Annahme, dass $a \cdot b$ eine normalverteilte Stichprobenverteilung besitzt, ist vor allem bei kleinen Stichproben zweifelhaft. + Der Test besitzt eine schlechte Power womit große Konfidenzintervall einhergehen. Die Präzision des Tests ist damit in Frage zu stellen!
- **Bootstrap Test:** berechnet Konfidenzintervalle für den indirekten Effekt. Liegt der Null-Effekt im CI, kann man davon ausgehen, dass keine Mediation vorliegt, anderenfalls hat man einen Mediator-Effekt gefunden. Diese Methode gibt über den Sobel-Test hinaus auch noch Auskunft über die Güte (Breite des CIs) des gefundenen Mediator-Effektes. Wenn möglich, sollte diese Methode zur Absicherung des

Effektes gewählt werden.

Effektgrößen der Mediation

Die einfachste Effektgröße ist der Regressionskoeffizient für den indirekten Effekt und das dazugehörige CI. Der indirekte Effekt ergibt sich aus den kombinierten Effekten der Pfade **a** und **b**, also:

Unstandardisierter indirekter Effekt: $UIE = a \cdot b$

Um einerseits den Effekt mit anderen Mediationsmodellen vergleichen zu können, und andererseits einen Kennwert zu berichten, der vor allem in einer Meta-Analyse verwendet werden kann, standardisiert man den indirekten Effekt (**index of mediation**):

Standardisierter indirekter Effekt: $SIE = a \cdot b \cdot \frac{s_{X_i}}{s_Y}$

Fallbeispiel

In einer Lehr-Lernstudie mit Einzeltutoren wurden von $N = 10$ StudentInnen Daten zur *Motivation*, *Lernleistung* und *Unterrichtsgüte* erfasst.

Motivation	Lernleistung	Unterrichtsgüte
98	4	5
98	5	6.5
103	6	7
101	7	5.5
100	8	9
106	10	6.5
111	11	8
125	12	10.5
120	14	8
115	15	10

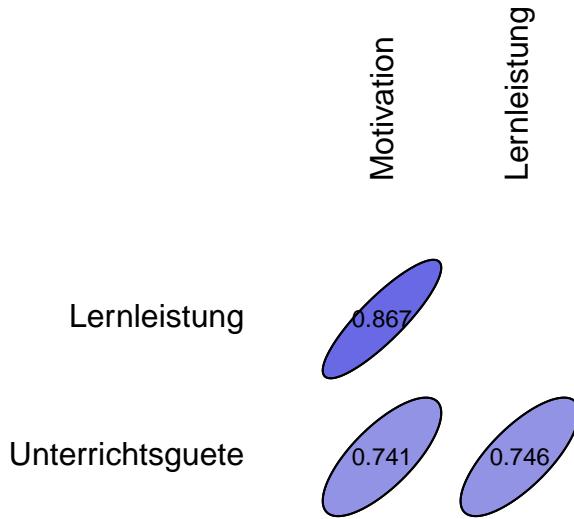
Regressionsanalyse

Wenn wir von der Hypothese ausgehen, dass die *Motivation* zu lernen durch die *Güte des Unterrichts* sowie der *Lernleistung* beeinflusst wird, ist es zunächst von Interesse festzustellen, ob dieses Merkmale überhaupt in Beziehung zueinander stehen. In folgenden rechnen wir daher zuerst die Korrelationen zwischen den Merkmalen. Kopier den Code in ein R-Script und führe diesen aus. Diskutiere die Ergebnisse!

```
# library(SciViews) # fÃ¼r Funktion correlation
KorTab <- SciViews::correlation(DF)
pander(KorTab, digits = 3)
```

	Motivation	Lernleistung	Unterrichtsgüte
Motivation	1	0.867	0.741
Lernleistung	0.867	1	0.746
Unterrichtsgüte	0.741	0.746	1

```
plot(KorTab, type = "lower", digits = 3)
```



Wir verwenden nun die Merkmale *Lernleistung* und *Unterrichtsgüte* als Regressoren und *Motivation* als Regressand, also:

$$Motivation = b_0 + b_1 \cdot Lernleistung + b_2 \cdot Unterrichtsguete$$

Kopiere den nachfolgenden Code in dein R-Script und führe diesen aus. Diskutiere die Ergebnisse und vergleiche diese mit nachfolgendem Pfadmodell.

```

Mod1      <- lm(Motivation ~ Lernleistung + Unterrichtsguete, data = DF)
Mod1_Std <- lm(scale(Motivation) ~ scale(Lernleistung) +
                 scale(Unterrichtsguete),
                 data = DF)

pander(summary(Mod1), digits = 2)
pander(summary(Mod1_Std), digits = 2)
    
```

Die zweite Tabelle zeigt die Ergebnisse in standardisierten Werten an (β_{LL}, β_{UG}). In einem Pfadmodell lassen sich die bisher gewonnenen Ergebnisse folgendermaßen zusammenfassen:

Dem Ergebnis ist zu entnehmen, dass beide Prädiktoren das Ausmaß der Motivation vorhersagen, wobei die Lernleistung ($\beta_{LL} = 0.71$ und somit $R^2_{LL} = 0.504$) sich als der bessere Prädiktor als die Unterrichtsgüte ($\beta_{UG} = 0.044$) herausstellt.

Man könnte allerdings auch davon ausgehen, dass die *Motivation* nicht ein Effekt der *Lernleistung* ist, sondern dass die Kausalwirkung umgekehrt verläuft, also:

- je höher die *Unterrichtsgüte*, desto höher die *Motivation*
- je höher die *Motivation* der Studenten, desto höher die *Lernleistung*
- die Modellvorstellung wäre demnach *Unterrichtsgüte* → *Motivation* → *Lernleistung*

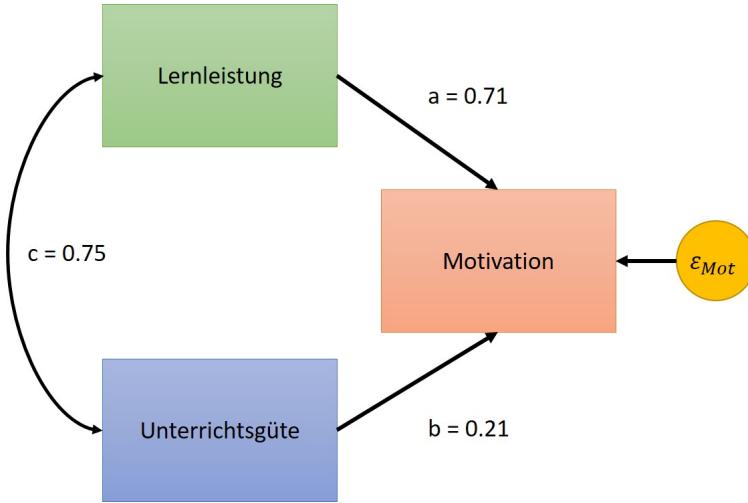


Figure 18: Abbildung 16: Pfadmodell der standardisierten Regressionskoeffizienten

In diesem Fall würde die *Unterrichtsgüte* indirekt über die Mediatorvariable *Motivation* auf die *Lernleistung* wirken. Darüber hinaus ist auch anzunehmen, dass die *Unterrichtsgüte* auch eine direkte Wirkung auf die *Lernleistung* ausübt. Diese Modellvorstellung lässt sich im folgenden Pfadmodell abbilden:

Diese Modellvorstellung kann man in folgenden Regressionsberechnungen zerlegen:

In der ersten Regression wird die abhängige Variable (Output, Kriterium) *Lernleistung* durch die unabhängige Variable (Prädiktor) *Unterrichtsgüte* vorhergesagt.

Im vorliegenden Beispiel bezeichnen wir den Steigungskoeffizienten mit b_1 . Dieser entspricht im konzeptuellen Modell dem Pfad **c**. Formal entspricht das Modell:

$$Lernleistung = b_0 + b_1 \cdot Unterrichtsguete$$

Kopiere den nachfolgenden Code in dein R-Script und führe diesen aus. Diskutiere die Ergebnisse und vergleiche diese mit obigen Pfadmodell.

```

Med_Mod_C      <- lm(Lernleistung ~ Unterrichtsguete, data = DF)
Med_Mod_C_Std <- lm(scale(Lernleistung) ~ scale(Unterrichtsguete), data = DF)

pander(summary(Med_Mod_C), digits = 2)
pander(summary(Med_Mod_C_Std), digits = 2)

```

In der zweiten Regression betrachten wir den Pfad *Unterrichtsgüte* zu *Motivation* (im konzeptuellen Modell der Pfad **a**), alsoe:

$$Motivation = b_0 + b_1 \cdot Unterrichtsguete + \varepsilon_{Mot}$$

Kopiere den nachfolgenden Code in dein R-Script und führe diesen aus. Diskutiere die Ergebnisse und vergleiche diese mit obigen Pfadmodell.

```

Med_Mod_A      <- lm(Motivation ~ Unterrichtsguete, data = DF)
Med_Mod_A_Std <- lm(scale(Motivation) ~ scale(Unterrichtsguete), data = DF)

pander(summary(Med_Mod_A), digits = 2)
pander(summary(Med_Mod_A_Std), digits = 2)

```

In der dritten Regression wird nun die Beziehung *Unterrichtsgüte* und *Lernleistung*, wie auch *Motivation* und *Lernleistung* (im konzeptuellen Modell der Pfad **b** und **c'**) überprüft. Das formale Modell lautet also:

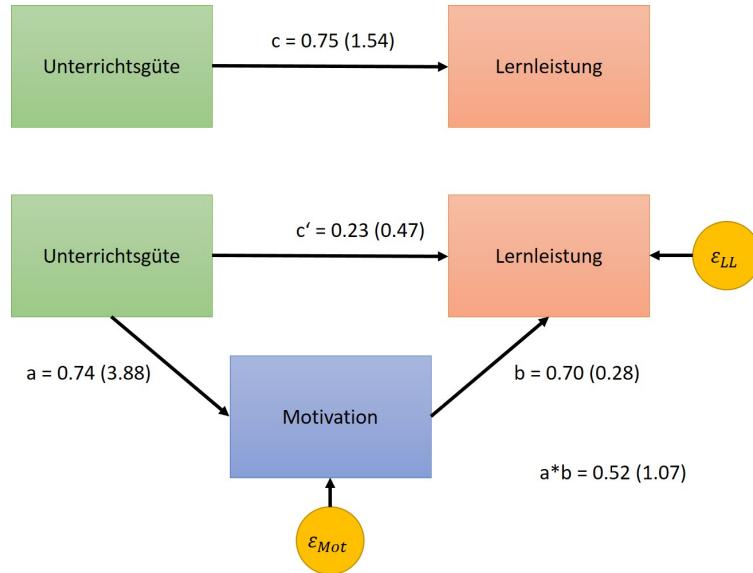


Figure 19: Abbildung 17: Pfadmodell der standardisierten Regressionskoeffizienten in einem Mediator-Modell

$$Lernleistung = b_0 + b_1 \cdot Unterrichtsguete + b_2 \cdot Motivation + \varepsilon_{LL}$$

Kopiere den nachfolgenden Code in dein R-Script und führe diesen aus. Diskutiere die Ergebnisse und vergleiche diese mit obigen Pfadmodell.

```
Med_Mod_CS_B      <- lm(Lernleistung ~ Unterrichtsguete + Motivation, data = DF)
Med_Mod_CS_B_Std <- lm(scale(Lernleistung) ~ scale(Unterrichtsguete) +
                         scale(Motivation),
                         data = DF)

pander(summary(Med_Mod_CS_B), digits = 2)
pander(summary(Med_Mod_CS_B_Std), digits = 2)
```

Teil V: Moderator-Analyse

```
# Field SPSS 5 - S481

rm(list = ls())
graphics.off()
if (!require("pacman")) install.packages("pacman")
pacman::p_load(DT, ggplot2, interactions, pander)
options(digits=3)
# Verzeichnise InitialisierenCPS85
```

Moderation

Fördert das Spielen von gewalttätigen Videos unsoziales, bzw. aggressives Verhalten?

In einer Studie wurde der Zusammenhang zwischen dem Spielen von aggressiven Videos wie z.B. Manhunt,

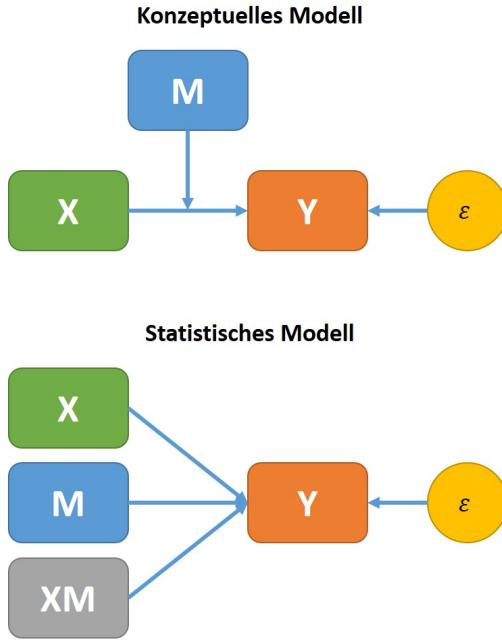


Figure 20: Abbildung 18: Konzeptuelles Modell Moderation

Grand Theft Auto und MadWorld mit Aggression untersucht. Dabei wurden $N = 442$ jugendliche bezüglich ihres aggressiven Verhaltens (Aggression), den gefühls- und emotionslosen Charaktereigenschaften (Callous Unemotional Traits - CUT, CaUnTs) und der Dauer des Videospielen in Stunden pro Woche (VidGames) aufgezeichnet. Die Daten finden Sie unter *Video Games.csv*.

ID	Aggression	Vid_Games	CaUnTs	CaUnTs_Grp
1	27	20	7	Low
2	30	34	14	Low
3	37	20	8	Low
4	29	29	13	Low
5	22	22	15	Low
6	29	34	7	Low

Das Ziel der Untersuchung ist es, die Beziehung zwischen Spieldauer (= Prädiktor, *Vid_Games*) und Aggression (*Aggression*) genauer zu untersuchen.

Konzeptuelles Modell

Eine Moderatorvariable ist eine Variable, welche die Beziehung zweier anderer Variablen beeinflusst.

Im Prinzip stellt sich also die Frage, ob eine Variable (X_1) einen Interaktionseffekt auf die Beziehung der beiden anderen Variablen (Y, X_2) bewirkt. Ein derartiger Interaktionseffekt ist im statistischen Sinne gleichzusetzen mit einem Moderationseffekt (konzeptueller Begriff).

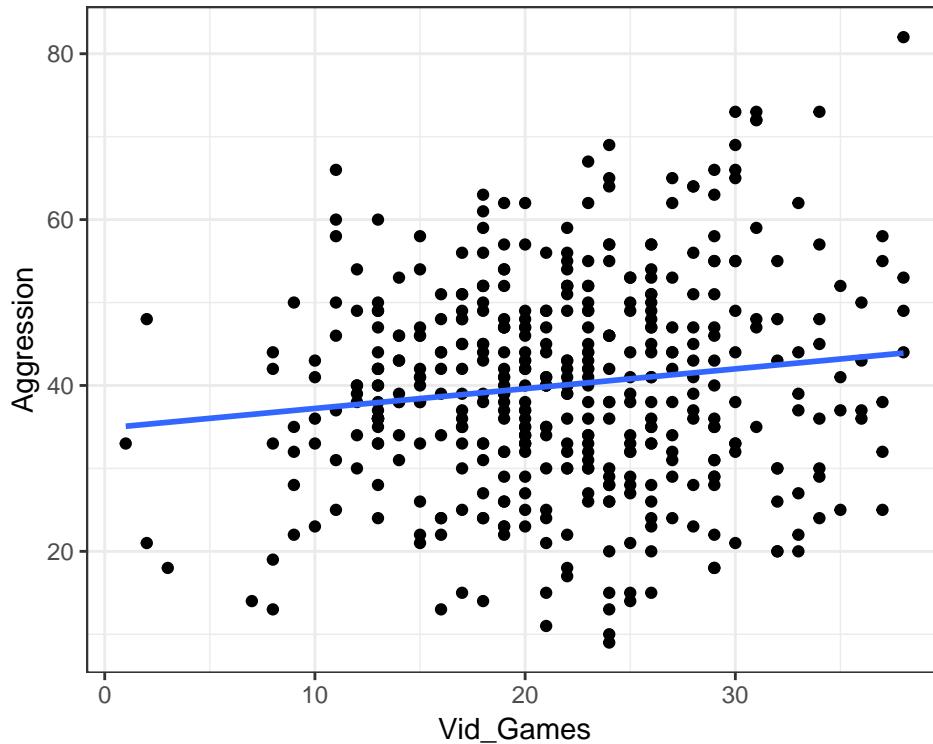
Betrachtet man die Beziehung zwischen der abhängigen Variablen (*Aggression*) und der erklärenden Variablen (*Vid_Games*):

```
p <- ggplot(DF, aes(x = Vid_Games, y = Aggression)) +
  geom_point() +
```

```

geom_smooth(method = lm, se = FALSE) +
theme_bw()
print(p, comment = FALSE)

```



```

SumMod1 <- summary(lm(Aggression ~ Vid_Games, data = DF))
pander(SumMod1)

```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.84	1.96	17.77	1.166e-53
Vid_Games	0.2385	0.08552	2.789	0.005515

Table 26: Fitting linear model: Aggression ~ Vid_Games

Observations	Residual Std. Error	R ²	Adjusted R ²
442	12.5	0.01737	0.01514

scheint die Spieldauer und Aggression nur in einem geringen Ausmaß miteinander in Beziehung zu stehen. Die aufgeklärte Varianz beträgt gerade einmal 1.737%.

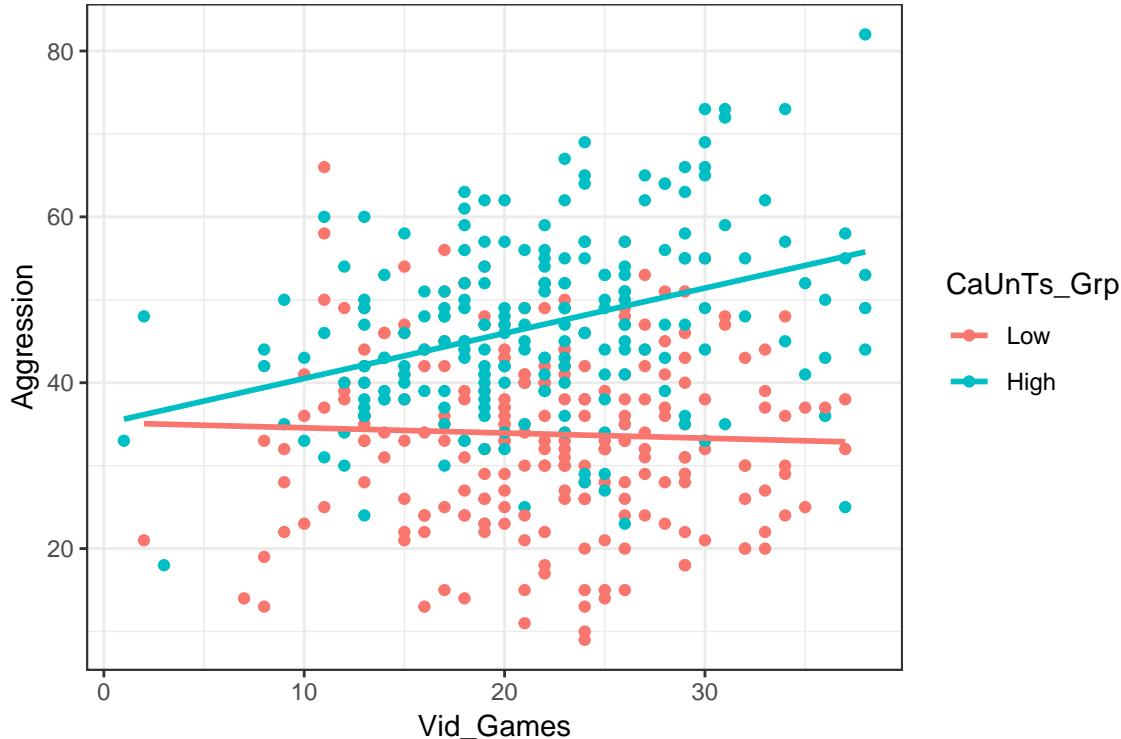
Ein anderes Bild ergibt sich jedoch, wenn man die Daten unter Berücksichtigung der Variablen *CaUnts* betrachtet. Teilt man die Variable *CaUnts* z.B. durch einen Mediansplit in zwei Gruppen (*CaUnts_Grp* = Low und High), zeigt sich bereits ein sehr unterschiedliches Bild:

```

p1 <- ggplot(DF, aes(x = Vid_Games, y = Aggression, color = CaUnts_Grp)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  theme_bw()

```

```
print(p1, comment = FALSE)
```



Durch die Trennung können folgende Eigenschaften in den Daten beobachtet werden:

- Bei Personen die keine CUT aufweisen, besteht keine Beziehung der Spieldauer und der Aggression.
- Personen die einen CUT aufweisen, zeigen eine positive Beziehung, d.h. je mehr Zeit sie spielen, desto höher wird ihr Aggressionslevel.
- Die CUT beeinflusst (moderiert) daher die Beziehung der Spieldauer und Aggression.

Den Effekt der Variablen *CaUnTs* (Moderatorvariablen) kann man sich gut durch folgende Darstellung vorstellen:

```
Mod_1 <- lm(Aggression ~ Vid_Games * CaUnTs, data=DF)
# plotPlane(model = Mod_1, plotx1 = "Vid_Games", plotx2 = "CaUnTs")
```

Dass die Moderatorvariable einen Einfluss auf die Beziehung zwischen *Vid_Games* und *Aggression* hat, zeigt sich vor allem dadurch dass bei niedrigen *CaUnTs*-Werten eine negative und bei hohen *CaUnTs*-Werten eine positive Beziehung zwischen Spieldauer und Aggression besteht. Damit wird auch der Kernpunkt eines Moderationseffektes angesprochen. Ein Moderationseffekt liegt vor, wenn sich die Beziehung zweier Variablen vom Wertebereich des Moderator abhängig ist.

Formale Beschreibung des Modells

Das lineare Modell einer Moderationsanalyse erweitert die bereits bekannte multiple Regression um den Interaktionsterm:

$$\widehat{Aggression}_i = (b_0 + b_1 \cdot Spieldauer_i + b_2 \cdot Callous_i + b_3 \cdot Interaktion_i)$$

In den meisten Statistikprogrammen (R, SPSS, SAS, etc.) gibt es Pakete/Makros, mit denen die Moderationsanalyse (u.v.m) speziell aufbereitet wird. Im vorliegenden Fall sollte anhand des einfachen Modells die grundlegende Idee vorgestellt werden.

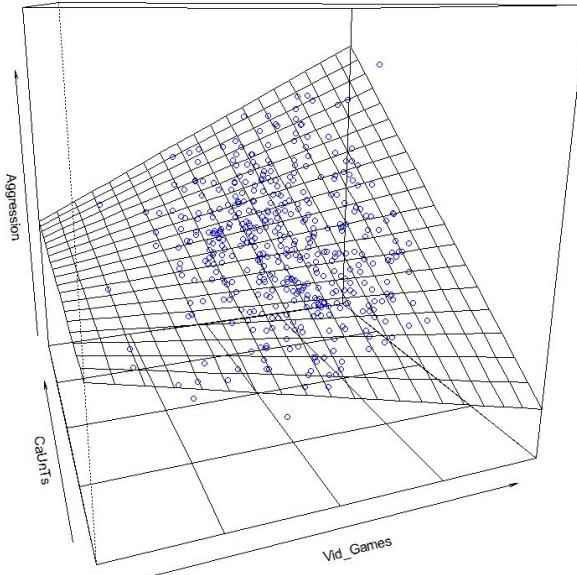


Figure 21: Abbildung 19: Moderator-Effekt

Bei der obigen Formel sind bis auf die Interaktion alle Daten bereits im geladenen Datenmaterial verfügbar. Die Interaktion kann nun sehr leicht aus diesen Daten berechnet und als weitere Variable im Datenframe abgespeichert werden. Dazu braucht man nur die beiden Variablen *Vid_Games* und *CaUnTs* multiplizieren und speichern.

```
DF$IA <- DF$Vid_Games*DF$CaUnTs
```

Im nachfolgenden Ergebnis ist vor allem der Interaktionseffekt von Interesse. Ist die Interaktion signifikant, kann man davon ausgehen, dass ein bedeutsamer Moderationseffekt vorliegt.

```
pander(summary(Mod_1))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.12	3.427	9.664	3.729e-20
Vid_Games	-0.3336	0.1508	-2.212	0.0275
CaUnTs	0.1689	0.161	1.049	0.2947
Vid_Games:CaUnTs	0.02706	0.006981	3.877	0.0001221

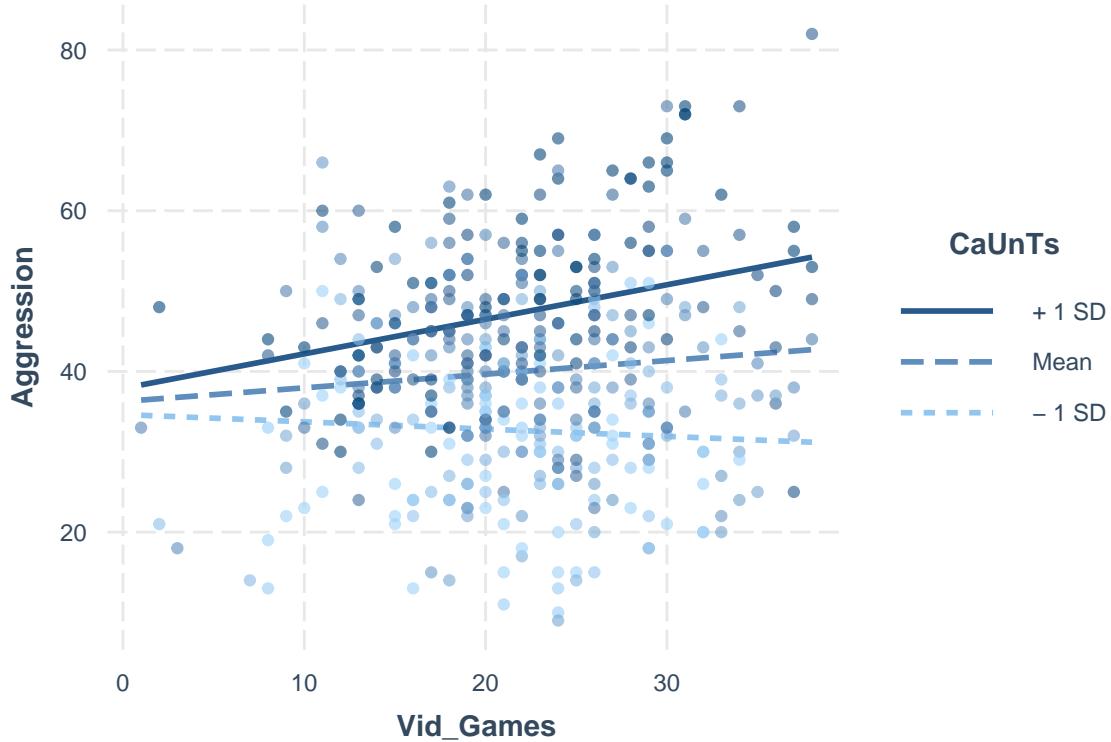
Table 28: Fitting linear model: Aggression ~ Vid_Games * CaUnTs

Observations	Residual Std. Error	R ²	Adjusted R ²
442	9.976	0.3773	0.373

Bei vorliegender signifikanter Interaktion kann durch eine **Simple Slope Analyse** (SSA) die Beziehung zwischen Prädiktor und Kriterium für verschiedene Werte des Moderators untersucht werden.

Häufig werden dafür drei Werte des Moderators (niedrig = $\bar{x}_{Mod} - sd_{Mod}$, mittel = \bar{x}_{Mod} , hoch = $\bar{x}_{Mod} + sd_{Mod}$) verwendet. In der folgenden Abbildung wurden die Datenpunkte noch zusätzlich nach Ihrer Ausprägung (Größe) bei *CaUnTs* gewichtet.

```
# library(interactions)
# # https://cran.r-project.org/web/packages/jtools/vignettes/interactions.html
interact_plot(Mod_1,
              pred = "Vid_Games",      47
              modx = "CaUnTs",
              plot.points = TRUE)
```



Für eine feiner Aufteilung, bzw. Darstellung des Wirkungsbereiches vom Moderator kann über die Johnson und Neymann Methode sehr viele Werte des Moderators berechnen werden. Die entsprechenden b 's werden ermittelt und daraus Signifikanz-Zonen berechnet. Damit werden die Werte des Moderators ermittelt, ab welchen der Prädiktor ein signifikantes Ergebnis liefert.

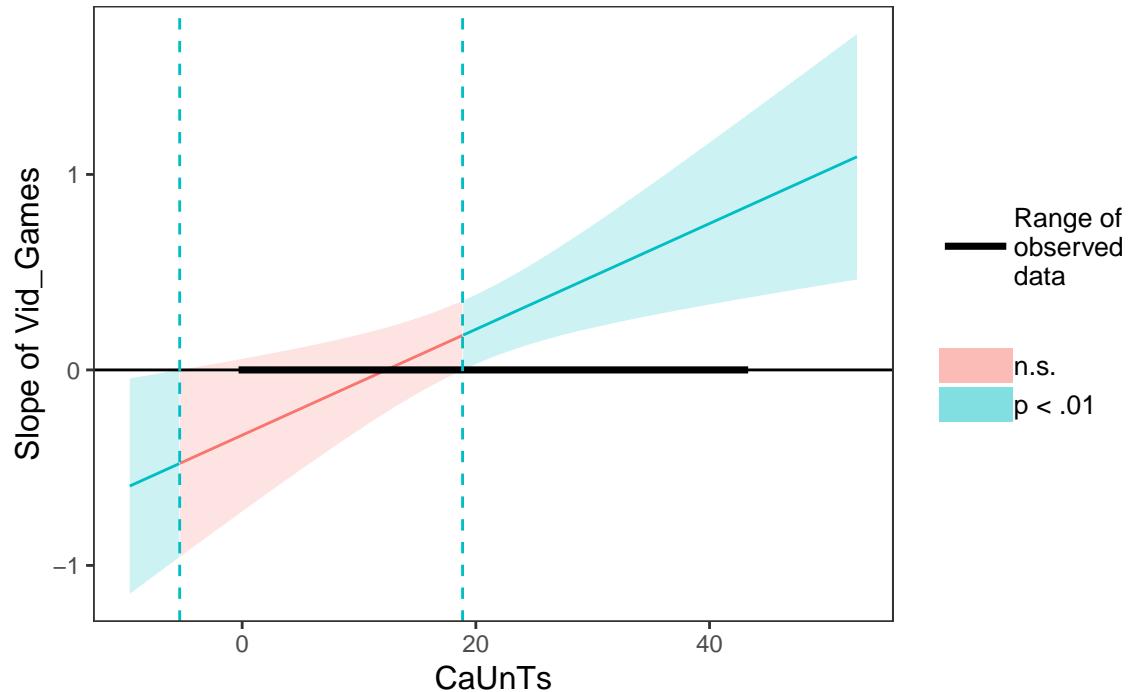
```
Mod_SS <- lm(Aggression ~ Vid_Games + CaUnTs + (Vid_Games * CaUnTs), data=DF)
SS_Res <- sim_slopes(Mod_SS, pred = Vid_Games, modx = CaUnTs, johnson_neyman = TRUE)
johnson_neyman(Mod_SS, pred = Vid_Games, modx = CaUnTs, alpha = 0.01)
```

JOHNSON-NEYMAN INTERVAL

When CaUnTs is OUTSIDE the interval [-5.35, 18.86], the slope of Vid_Games is $p < .01$.

Note: The range of observed values of CaUnTs is [0.00, 43.00]

Johnson–Neyman plot



Zentrierung der Variablen

Durch die Berücksichtigung der Interaktion, bekommen die Koeffizienten eine spezielle Bedeutung.

Ein Prädiktor X_i repräsentiert die Regression des Kriteriums, wenn die anderen Prädiktoren $X_{j \neq i}$ gleich Null sind!

Im Beispiel:

- b_1 entspricht dem Koeffizienten einer Regression wenn CaUnTs gleich Null ist, also keine Gefühls und emotionslosen Charaktereigenschaften aufweist.
- b_2 entspricht dem Koeffizienten einer Regression wenn VidGames gleich Null ist, also 0 Stunden Video gespielt hat.

Im gegebenen Beispiel sind für beide Prädiktoren Null-Werte durchaus möglich. Würde man aber anstelle der Charaktereigenschaften die Herz-Rate der Jugendlichen messen, wäre ein Wert von Null nicht sinnvoll.

Aus diesem Grund werden in einer Moderationsanalyse die Prädiktoren transformiert. Man zentriert die Daten auf die Abweichungen des *grand averages*.

Eigenschaften der Zentrierung

Das Zentrieren der Prädiktoren ist ähnlich der bekannten z-Transformation, wobei bei der Zentrierung die Division durch die Standardabweichung nicht relevant ist.

Formal entsprechen die zentrierten Daten also:

$$X_{1i}^C = X_{1i} - \bar{X}_{11}$$

Die Zentrierung hat keinen Einfluss auf den Koeffizienten, welcher die meisten Variablen (highest-order predictor) verknüpft. Im gegebenen Beispiel entspricht das dem Koeffizienten b_3 , also der Interaktion. Für die Koeffizienten b_1 und b_2 bewirkt die Zentrierung:

- b_2 zeigt den Effekt zwischen Aggression und CUT, wenn jemand die durchschnittliche Spielzeit mit Videospielen verbringt.
- b_1 zeigt den Effekt zwischen Aggression und Spieldauer, wenn jemand den durchschnittlichen CUT-Wert haben würde.

Gründe für Zentrierung

Aus Gründen der Interpretierbarkeit, z.B. y = verbale Fähigkeiten eines Kindes, x_1 = Vokabular von der Mutter, x_2 = Alter des Kindes:

- b_0 ist nicht sinnvoll interpretierbar, wenn $x_1 = 0$ und $x_2 = 0$
- b_0 ist aber dann sinnvoll interpretierbar, wenn die Variablen zentriert wurden, denn dann sind die Werte 0 für beide Variablen eben der Mittelwert der zentrierten Variablen!
- b_1 entspricht der Steigung von x_1 unter der Annahme eines durchschnittlichen Wertes von x_2 . Liegt keine Moderation vor, wir b_1 konsistent über alle Werte der x_2 -Verteilung sein.
- Liegt ein Moderationseffekt vor, ist b_1 nicht konsistent über die Verteilung der x_2 -Werte.

Aus statistischen Gründen:

- korrelieren x_1 und x_2 dann kann es sein, dass auch die Interaktion $x_1 \cdot x_2$ hoch korreliert! Wenn zwei Prädiktoren in einem GLM hoch miteinander korrelieren, dann sind sie im wesentlichen redundant. Es liegt eine hohe Multikollinearität vor! Es wird auch schwierig, die b -Werte den jeweiligen Prädiktoren zuzuordnen.
- Bei Zentrierung der Daten kann dieses Problem entschärft werden.

Interpretation der b 's

Die Zentrierung ermöglicht die Interpretation der Koeffizienten b_1 und b_2 (lower-order-predictors) auch dann, wenn Null-Werte bei Prädiktoren nicht sinnvoll zu interpretieren sind!

Die Interpretation von lower-order-predictors ist nur sinnvoll, wenn die higher-order-predictors nicht signifikant sind!

Bei zentrierten Prädiktoren können die b 's von einzelnen Prädiktoren folgendermaßen interpretiert werden:

1. sie zeigen den Effekt des jeweiligen Prädiktors beim Mittelwert der Stichprobe.
2. sie zeigen den durchschnittlichen Effekt des Prädiktors über die Spannweite der Werte des anderen Prädiktors.

Bemerkung zu 2: stellen Sie sich vor, dass Sie für jeden möglichen Wert der Spieldauer (VidGames = von 0 bis max(Spieldauer)) jeweils ein lineares Modell rechnen, also ein Modell zur Vorhersage für Aggression durch CUT für die Personen die 0 Std. gespielt haben. Dann dasselbe Modell für jene die 1 Std. gespielt haben, etc.

Dadurch ergeben sich z.B. N unterschiedliche b 's, wobei jedes davon den Zusammenhang zwischen CUT und Aggression für unterschiedlich lange Spieldauer zeigt. Würde man den Mittelwert dieser b 's berechnen, dann wäre dieser Wert gleich dem b -Wert für CUT (zentriert) wenn dieser im Moderationsmodell gemeinsam mit der zentrierten Spieldauer (VidGames) und der Interaktion berechnet wird!

Interaktion

Besteht eine signifikante Interaktion zwischen den beiden Prädiktoren, liegt ein Moderationseffekt vor!

Ist die Interaktion $CaUnTs \times Vid_Games$ ein signifikanter Prädiktor für Aggression, dann wissen wir zwar dass ein Moderationseffekt vorliegt, aber nicht in welcher Weise! Es könnten folgende Zusammenhänge bestehen:

- Es könnte sein, dass die Spieldauer immer einen negativen Einfluss auf den Aggressions hat, aber diese Beziehung bei höher werdenden CUT noch wesentlich verstärkt wird.
- Es könnte sein, dass bei Personen mit niedrigem CUT die Spielzeit die Aggression verringert, aber bei Personen mit hohen CUT die Aggression verstärkt.

Die Simple Slope Analysis (SSA) bietet bei der Interpretation der Interaktionseffekte eine wesentliche Hilfestellung.

Teil VI: Analysis of Covariance

Motivation

Die Kovarianzanalyse (ANCOVA) ist ein Spezialfall der ANOVA. Beide werden verwendet, um die Auswirkungen kategorischer Variablen (Faktoren) auf eine intervall- oder ratio-level abhängige Variable zu testen.

Die ANCOVA gibt uns jedoch die zusätzliche Möglichkeit, gleichzeitig die Wirkung anderer kontinuierlicher Variablen auf die abhängige Variable zu beurteilen oder zu kontrollieren. Kontinuierliche Variablen, die als Unabhängige in einem ANOVA-Design enthalten sind und die mit einer abhängigen Variablen kovariabel sind, nennt man *Kovariablen*.

Bei der ANCOVA geht es im Wesentlichen darum, die Fehlervarianz bei randomisierten Gruppenexperimenten weiter zu verringern.

ANCOVA wird jedoch häufiger eingesetzt, wenn eine Randomisierung nicht möglich ist. In diesen Fällen müssen wir uns oft mit so genannten "nicht-äquivalenten" (nicht zufällig zugeordneten) Gruppen zufrieden geben. Per Definition können sich solche Gruppen in erheblicher Weise unterscheiden, auch bei Merkmalen, die die Ergebnisvariable beeinflussen können. Solange sie nicht berücksichtigt werden, kann das Vorhandensein dieser Hintergrund- oder Fremdvariablen die Fehlervarianz erhöhen, unser "Signal-Rausch-Abstand" verringern und es letztendlich schwieriger machen, einen echten Unterschied zwischen den Gruppen zu erkennen.

Kovarianzanalyse

Im folgenden betrachten wir ein Beispiel, welches die Auswirkungen von Viagra auf den Libido untersucht.

Es ist naheliegend anzunehmen, dass auch andere Faktoren (wie andere Medikamente, Müdigkeit, etc.) den Libido beeinflussen. Wenn diese Variablen (die Kovariablen) gemessen werden, ist es möglich, den Einfluss, den sie auf die abhängige Variable haben, durch Einbeziehung in das Regressionsmodell zu steuern/kontrollieren.

Ziel ist es, den Effekt der Kovariaten auf die Zielvariable zu *entfernen*. Durch diese Maßnahme sollte sich die Wirkungsweise der unabhängigen Variablen (Viagra) auf den Libido *besser* zeigen. Es gibt im Wesentlichen zwei Gründe⁹ für die Aufnahme von Kovariablen in die ANOVA:

- **Reduzierung der Fehlervarianz:** in der ANOVA und beim t-Tests wird die Wirkung eines Experiments anhand der erklärbaren Variabilität in den Daten, mit der nicht erklärbaren Variabilität verglichen. Wenn ein Teil dieser *unerklärten Varianz (SSR)* einer anderen Variablen (*Kovariablen*) zugeordnet werden kann, reduziert sich die Fehlervarianz. Damit kann die Wirkung der unabhängigen Variable (*SSM*) genauer beurteilt werden.
- **Eliminierung von Confounds:** in jedem Experiment kann/wird es Variablen geben, welche nicht gemessen/erhoben wurden, die aber die Ergebnisse der Zielvariablen durchaus beeinflussen können. Sind diese bekannt, kann mit Hilfe der ANCOVA dieser Einfluss beseitigt werden.

Im vorliegenden Beispiel gehen wir davon aus, dass der Libido der Sexualpartner den eigenen Libido beeinflusst¹⁰. Das entsprechende Regressionsmodell erweitert sich demnach zu:

$$libido_i = b_0 + b_3 \cdot covariate_i + b_2 \cdot high_i + b_1 \cdot low_i + \varepsilon_i$$

⁹Es gibt noch andere Gründe für die Aufnahme von Kovariablen in die ANOVA, welche in der Berechnung der ANCOVA detailliert beschrieben werden. Siehe (Stevens 2002), (Wildt 2009).

¹⁰wie oft sie versuchten, sexuellen Kontakt aufzunehmen.

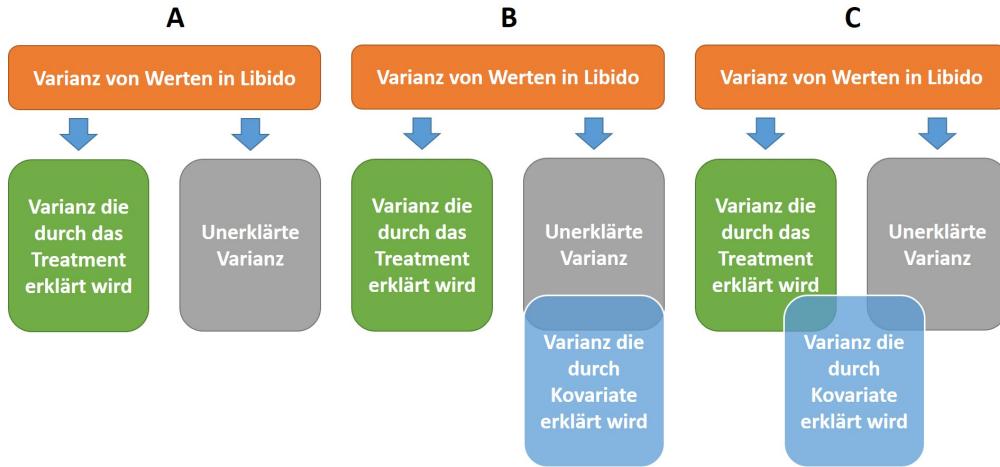


Figure 22: Abbildung 20: Wirkungsweise von Kovariate

Voraussetzungen

Die ANCOVA hat die gleichen Annahmen wie ANOVA, zu welchen es jedoch noch zwei wichtige zusätzliche Überlegungen gibt:

1. **Unabhängigkeit von der Kovariablen- und dem Treatmenteffekt.** Abbildung 4-A zeigt die bereits aus der ANOVA bekannte Zerlegung der Varianzen in eine Fehlervarianz und einer Treatmentvarianz. Abbildung 4-B stellt eine ideale Voraussetzung für die Verwendung einer Kovariaten dar. Hierbei wird durch die Kovariate ein Teil der Fehlervarianz erklärt, ohne den Effekt des Treatment zu beeinflussen. Abbildung 4-C hingegen zeigt das Problem bei einer fälschlicherweise verwendeten Kovariaten. Die Kovariate verringert zwar nach wie vor die Fehlervarianz, aber gleichzeitig wird auch der Treatmenteffekt beeinflusst. Statistisch gesehen können wir nur festhalten, dass die Kovariate und das Treatment Varianz gemeinsam erkären. Eine Trennung dieser gemeinsamen Varianz in Anteile Viagra und Kovariate ist nicht möglich! Eine einfache Möglichkeit die Kovariate auf ihre Eigenschaft zu prüfen, ist ein einfacher Mittelwertsvergleich (t-Test, ANOVA) der nach Viagragruppen aufgeteilten Kovariaten. Wenn die Gruppen sich nicht unterscheiden, kann von einer Unabhängigkeit ausgegangen werden und sofern die anderen Voraussetzungen erfüllt sind, die Kovariate verwendet werden. Auch durch eine Randomisierung der Gruppenzuordnung kann man unerwünschte Effekte (in Bezug auf die Wirkung der Kovariaten) zwischen den Gruppen evtl. vermeiden.

Zum besseren Verständnis der mit den ANOVA-Verfahren verbundenen Varianzaufteilung betrachten wir nochmals im Detail die Eigenschaften der verschiedenen Varianzanteile.

Die Gesamtvarianz (im vorigen Graphen die Varianz von Werten im Libido) wird folgendermaßen ermittelt:

Die Treatmentvarianz (im vorigen Graphen die Varianz die durch das Treatment erklärt wird) entspricht der Variabilität der Mittelwerte der jeweiligen Gruppen (in unserem Fall der Dosierungsstufen):

Die Fehlervarianz wird aus den durchschnittlichen Abweichungen der beobachteten Werte zu den jeweiligen Gruppenmittelwerten bestimmt (geschätzt). Anhand dieser Darstellung wird auch klar, warum die Varianzgleichheit über die Gruppen hinweg gleich sein sollte. Wäre das nämlich nicht gegeben, würde die Additivität ($QS_1 + \dots + QS_k$) nicht gegeben sein.

2. **Homogenität der Regressionssteigungen.** Bei einer ANCOVA wird die Gesamtbeziehung zwischen dem Ergebnis (abhängige Variable) und der Kovariablen analysiert. D.h., es wird eine *Regressionslinie an den gesamten Datensatz angepasst* und man *ignoriert*, zu welcher Gruppe eine Person gehört. Bei der Anpassung dieses Gesamtmodells gehen wir daher davon aus, dass diese Gesamtbeziehung für alle Teilnehmergruppen gilt. Diese Annahme ist für die ANCOVA sehr wichtig. Der beste Weg diese Annahme zu kontrollieren, ist eine Darstellung der Kovariablen (*Partner's Libido*) auf der einen und dem

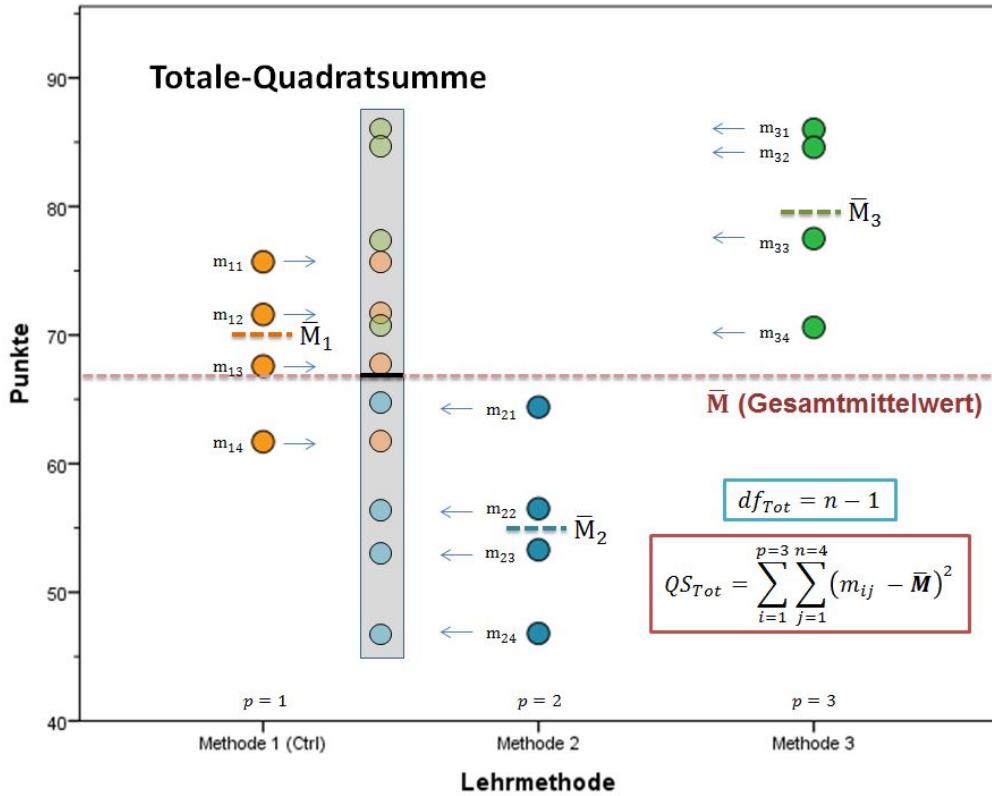


Figure 23: Abbildung 21: Totale Quadratsumme

Ergebnis (*Libido*) auf der anderen Achse, getrennt nach den Gruppen (*Dosierung*). Die Regressionslinien sollten dann mehr oder weniger gleich aussehen (d.h. die Werte von b in jeder Gruppe sollten gleich sein). Im nachfolgender Darstellung wäre diese Voraussetzung nicht erfüllt!

Berechnung einer ANCOVA

Bei der Berechnung einer ANCOVA sollten folgende Schritte durchgeführt werden:

1. Grafischen Darstellung der Daten und der Berechnung einiger deskriptiver Statistiken. Dabei sollten auch die Verteilungsannahmen überprüfen und den Levene-Test durchgeführt werden (Homogenitätstest).
2. Überprüfen der Kovariablen und alle unabhängigen Variablen auf Unabhängigkeit, d.h. eine ANOVA mit der Kovariablen als Ergebnis und alle unabhängigen Variablen als Prädiktoren durchführen. Damit wird sichergestellt, dass sich die Kovariablen auf den Ebenen dieser Variablen nicht signifikant unterscheidet. Wenn man ein signifikantes Ergebnis erhält, dann ist die Analyse bei diesem Schritt beendet¹¹.
3. Durchführen der ANCOVA.
4. Berechnung der *Kontraste* oder *post hoc-Tests* (falls signifikante Ergebnisse vorliegen).
5. Überprüfen der Homogenität der Regressionssteigungen. Dies kann graphisch (siehe oben) durchgeführt werden, oder man kann auch die ANCOVA erneut ausführen und die Interaktion zwischen der unabhängigen Variable und der Kovariablen ins Modell aufnehmen. Wenn diese Interaktion signifikant ist, kann man nicht von einer Homogenität der Regressionsflanken ausgehen!

Deskriptive, graphisch und Homogenität

Um die Verteilung von Daten darzustellen, kann man z.B. Boxplots für *Libido* als auch für *Libido des Partners* erzeugen. Darüber hinaus ist es hilfreich, die Beziehung zwischen der Ergebnisvariablen und der Kovariablen

¹¹möglicherweise kann man eine robuste Version des Tests ausführen, Details später.

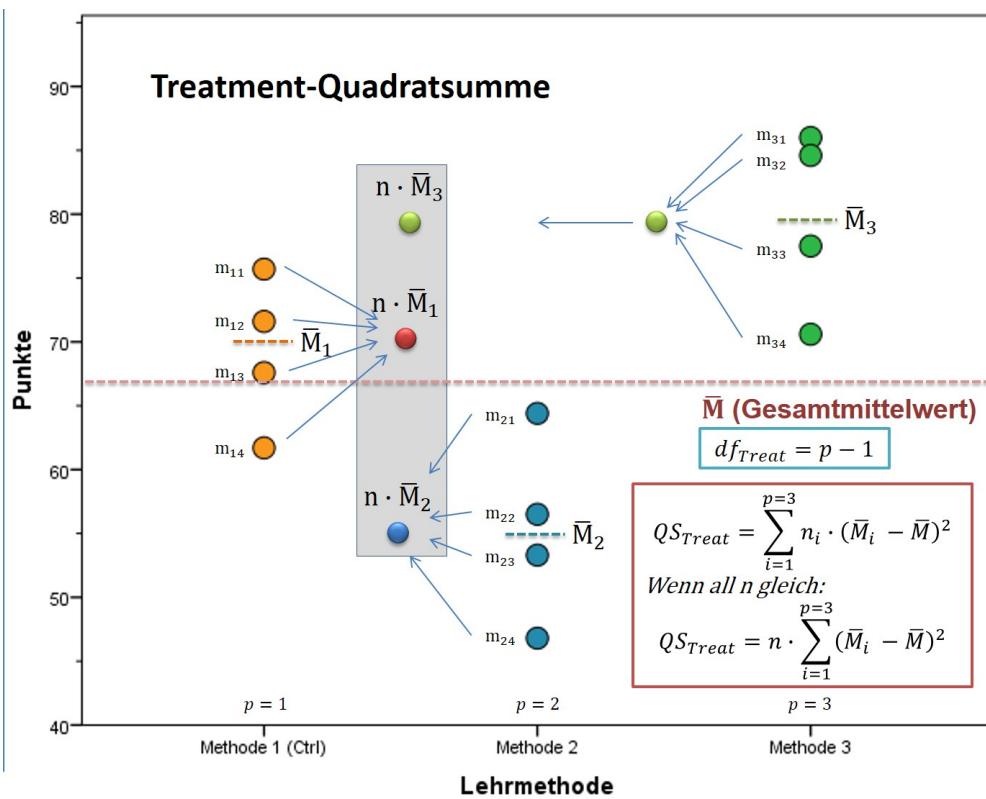


Figure 24: Abbildung 22: Treatmentquadratsumme

innerhalb jeder Gruppe zu betrachten (dies sagt uns etwas über die Homogenität der Steigungen aus).

```
options(digits = 3)

viagraData      <- read.delim("C:/NextCloud/DATEN/CSV_Text/ViagraCovariate.dat", header = TRUE)
viagraData$dose <- factor(viagraData$dose, levels = c(1:3), labels = c("Placebo", "Low Dose", "High Dose"))

restructuredData <- reshape2::melt(viagraData,
                                    id = c("dose"),
                                    measured = c("libido", "partnerLibido"))
names(restructuredData) <- c("dose", "libido_type", "libido")

boxplot <- ggplot(restructuredData, aes(dose, libido)) +
  geom_boxplot() +
  facet_wrap(~libido_type) +
  labs(x = "Dose", y = "Libido") +
  theme_bw()
boxplot
```

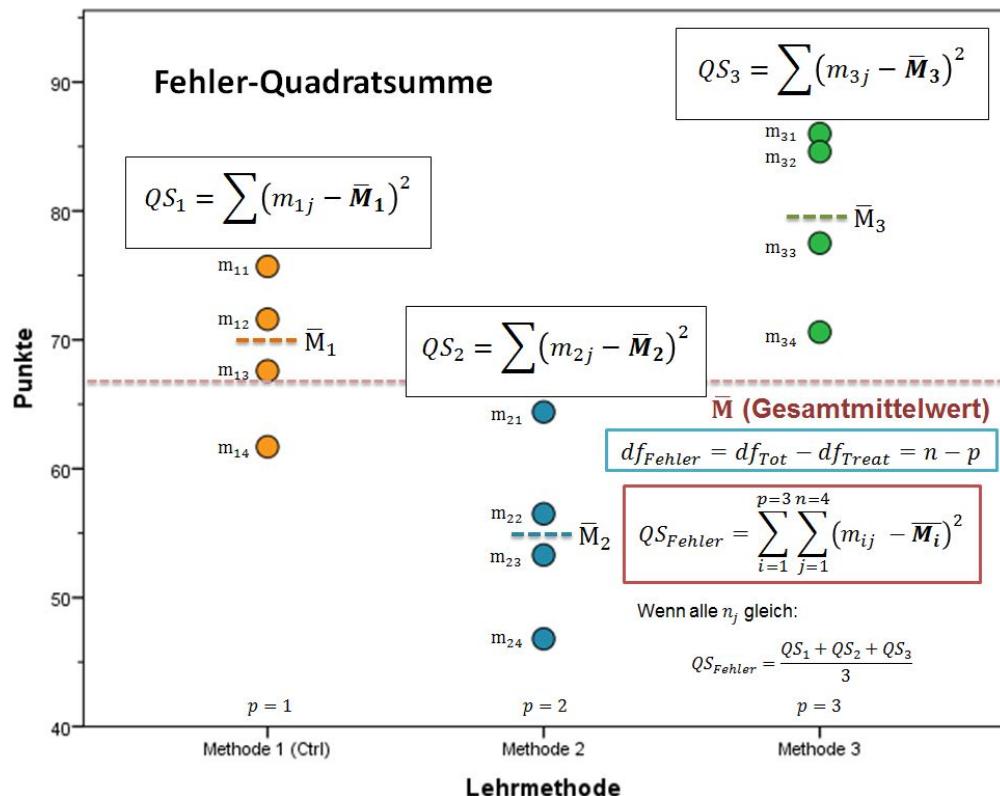


Figure 25: Abbildung 23: Fehlerquadratsumme

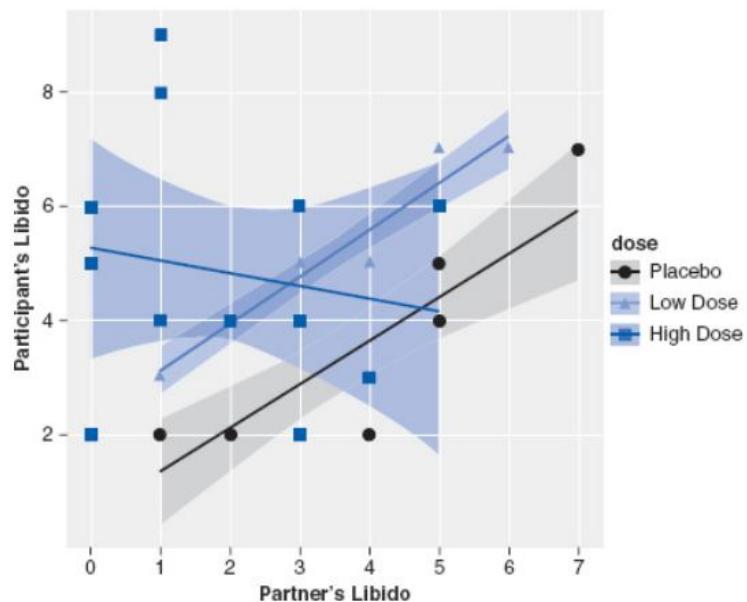
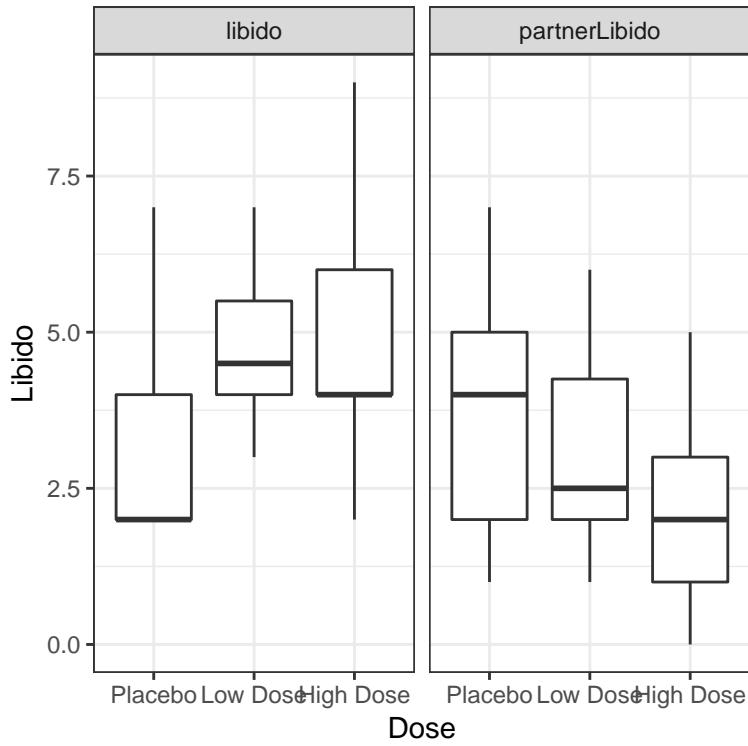


Figure 26: Abbildung 24: Verletzung der Homogenitätsbedingung



Die Boxplots zeigen den Libido bei den Teilnehmern und ihren Partnern über die drei Dosen von Viagra. Die Libido scheint für die Teilnehmer mit zunehmender Dosis von Viagra zu steigen, aber das Gegenteil gilt für ihre Partner.

Neben der graphischen Darstellung sind auch die deskriptiven Werte aufschlussreich, da diese Kennwerte wie die Streung (sd) und Mittelwerte (\bar{x}), Konfidenzintervalle (CI), etc. ausgegeben werden.

```
# library(pastecs) fÃ¼r stat.desc()
Res1 <- by(viagraData$libido, viagraData$dose, stat.desc, basic = F, simplify = TRUE)
pander(Res1$Placebo)
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
2	3.222	0.5958	1.374	3.194	1.787	0.5547

```
pander(Res1$`Low Dose`)
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
4.5	4.875	0.5154	1.219	2.125	1.458	0.299

```
pander(Res1$`High Dose`)
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
4	4.846	0.5867	1.278	4.474	2.115	0.4365

```
# Res1 <- data.frame(unlist(by(viagraData$libido, viagraData$dose, stat.desc, basic = F)))
# colnames(Res1) <- c("Statistic Libido")
# pander(Res1)
```

```
Res2 <- by(viagraData$partnerLibido, viagraData$dose, stat.desc, basic = F)
pander(Res2$Placebo)
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
4	3.444	0.6894	1.59	4.278	2.068	0.6005

```
pander(Res2$`Low Dose`)
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
2.5	3.125	0.6105	1.444	2.982	1.727	0.5526

```
pander(Res2$`High Dose`)
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
2	2	0.4529	0.9868	2.667	1.633	0.8165

```
# Res2 <- data.frame(unlist(by(viagraData$partnerLibido, viagraData$dose, stat.desc, basic = F)))
# colnames(Res2) <- c("Statistic Partners Libido")
# pander(Res2)
```

```
pander(stat.desc(viagraData$libido, basic = F))
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
4	4.367	0.3571	0.7304	3.826	1.956	0.448

```
pander(stat.desc(viagraData$partnerLibido, basic = F))
```

median	mean	SE.mean	CI.mean.0.95	var	std.dev	coef.var
2.5	2.733	0.3388	0.6929	3.444	1.856	0.6789

Der Test auf Varianzhomogenität wird mit dem Levene's-Test durchgeführt. Dabei zeigt sich der Test mit dem Median als zentraler Kennwert robuster als die Schätzung durch den Mittelwert (*Bemerkung:* man kann auch das Verhältnis der größten zur kleinsten Varianz¹² (aus deskriptiver Statistik) bilden und in einer entsprechenden Tabelle auf Signifikanz prüfen).

```
# library(car) fÃ¼r Levenes Test
pander(leveneTest(viagraData$libido, viagraData$dose, center = median)) # fÃ¼r robustere SchÃ¤tzung!
```

¹²Hartely's F_{max} variance ratio.

Table 37: Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	0.3256	0.7249
	27	NA	NA

```
pander(leveneTest(viagraData$libido, viagraData$dose, center = mean))
```

Table 38: Levene's Test for Homogeneity of Variance (center = mean)

	Df	F value	Pr(>F)
group	2	0.7112	0.5
	27	NA	NA

Unabhängigkeit

Die Unabhängigkeit kann man relativ einfach durch eine ANOVA mit *partnerLibido* als Ergebnis und *Dosis* als Prädiktor durchführen.

```
checkIndependenceModel <- aov(partnerLibido ~ dose, data = viagraData)
pander(summary(checkIndependenceModel))
```

Table 39: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dose	2	12.77	6.385	1.979	0.1577
Residuals	27	87.1	3.226	NA	NA

```
pander(summary.lm(checkIndependenceModel))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.444	0.5987	5.753	4.062e-06
doseLow Dose	-0.3194	0.8727	-0.366	0.7172
doseHigh Dose	-1.444	0.7788	-1.855	0.0746

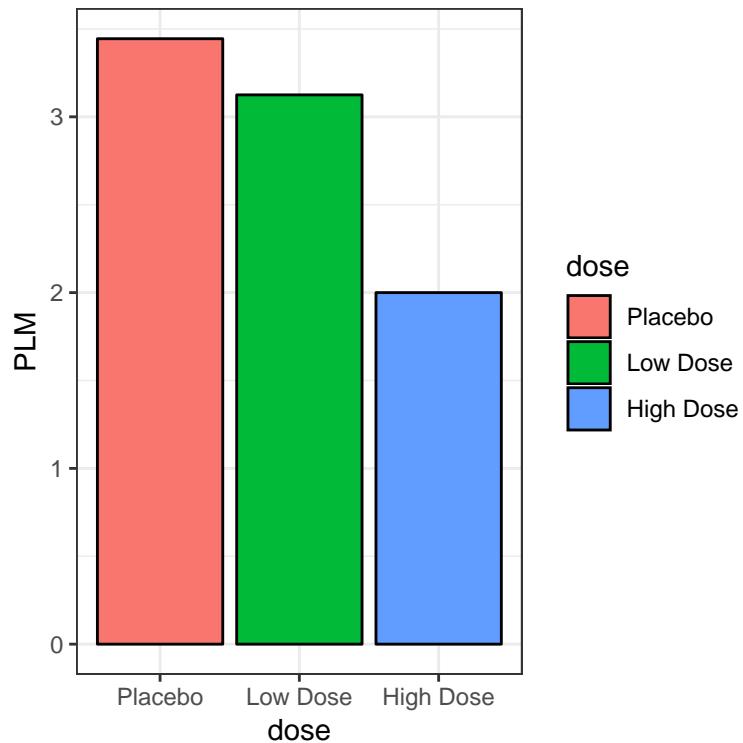
Table 41: Fitting linear model: partnerLibido ~ dose

Observations	Residual Std. Error	R ²	Adjusted R ²
30	1.796	0.1279	0.06326

```
PD1 <- data.frame(doBy::summaryBy(partnerLibido ~ dose, data = viagraData, FUN = mean))

colnames(PD1) <- c("dose", "PLM")
ggplot(data = PD1, aes(x = dose, y = PLM, fill = dose)) +
  geom_bar(colour="black", stat="identity") +
```

```
theme_bw()
```



```
guides(fill=FALSE)
```

```
## $fill  
## [1] FALSE  
##  
## attr(),"class")  
## [1] "guides"
```

Bei den Koeffizienten (Estimate) des Modells entspricht der Intercept den Mittelwert der ersten Dosierungsstufe (= Placebo) und die weiteren den jeweiligen Abstand zum Mittelwert der Placebodosierung!

Berechnung ANCOVA

Nach Überprüfung der Voraussetzungen können wir die ANCOVA berechnen.

```
pander(car::Anova(aov(libido ~ partnerLibido + dose, data = viagraData), type = "III"))
```

Table 42: Anova Table (Type III tests)

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	12.94	1	4.257	0.0492
partnerLibido	15.08	1	4.959	0.03483
dose	25.19	2	4.142	0.02745
Residuals	79.05	26	NA	NA

Betrachtet man die Signifikanz-Werte, so ist klar, dass die Kovariable die abhängige Variable signifikant vorhersagt, da $F(1, 26) = 4.96, p = .035$ ist. Es ist also davon auszugehen, dass der Libido der Person durch

die Libido des Partners beeinflusst wird.

Interessant ist, dass nach Berücksichtigung der Wirkung des Libido's vom Partners die Wirkung von Viagra signifikant ist ($F(2, 26) = 4.14, p = .028$).

Wenn wir das nochmals mit den Ergebnissen einer ANOVA (also ohne Berücksichtigung der Kovariaten vergleichen), stellen wir fest, dass durch die Kovariate sich ein nicht signifikantes in ein signifikantes Ergebnis geändert hat.

```
pander(car:::Anova(aov(libido ~ dose, data = viagraData), type = "III"))
```

Table 43: Anova Table (Type III tests)

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	93.44	1	26.81	1.891e-05
dose	16.84	2	2.416	0.1083
Residuals	94.12	27	NA	NA

```
pander(summary.lm(aov(libido ~ partnerLibido + dose, data = viagraData)))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.789	0.8671	2.063	0.0492
partnerLibido	0.416	0.1868	2.227	0.03483
doseLow Dose	1.786	0.8494	2.102	0.04535
doseHigh Dose	2.225	0.8028	2.771	0.01018

Table 45: Fitting linear model: libido ~ partnerLibido + dose

Observations	Residual Std. Error	R ²	Adjusted R ²
30	1.744	0.2876	0.2055

Interpretation ANCOVA

Es scheint ziemlich klar zu sein, dass die signifikante ANOVA einen Unterschied zwischen der Placebogruppe und den beiden experimentellen Gruppen widerspiegelt.

Table 46: Anova Table (Type III tests)

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	76.07	1	25.02	3.342e-05
partnerLibido	15.08	1	4.959	0.03483
dose	25.19	2	4.142	0.02745
Residuals	79.05	26	NA	NA

Dieser Effekt kann damit begründet werden, da niedrig- und hochdosierten Gruppen sehr ähnliche Mittel haben ($\bar{x}_{Low} = 4.88$, $\bar{x}_{High} = 4.85$, während der Mittelwert der Placebogruppe bei $\bar{x}_{Placebo} = 3.22$ viel niedriger ist.

	Libido	Libido_Partner	Libido_Adj
	Libido	Libido_Partner	Libido_Adj
Placebo	3.222	3.444	2.926
Low Dose	4.875	3.125	4.712
High Dose	4.846	2	5.151

Eigentlich können wir aber diese Gruppenmittel nicht interpretieren, da sie nicht um den Effekt der Kovarianz bereinigt wurden. Diese ursprünglichen Mittel sagen uns nichts über die Gruppenunterschiede, die sich in der signifikanten ANCOVA widerspiegeln! Daher müssen für diesen Vergleich die um den Effekt der Kovariaten bereinigten Mittelwerte verwendet werden. Diese sind in obiger Tabelle in Spalte *Libido_Adj* angegeben!

Geplante Kontraste

Für die Berechnung von Kontrasten können entweder vordefinierte Kontrastcodes, oder eigene Kontrastekodierungen angegeben werden¹³. In R lässt sich z.B. ein Kontrast durch folgende Eingabe definieren:

```
# für orthogonale Kontraste nach Helmert
contrasts(viagraData$dose) <- contr.helmert(3)
# für Vergleich von Placebo vs. low- und highdose (-2,1,1), sowie low vs. high
contrasts(viagraData$dose) <- -cbind(c(-2,1,1), c(0,-1,1))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.126	0.625	5.002	3.342e-05
partnerLibido	0.416	0.1868	2.227	0.03483
dose1	0.6684	0.24	2.785	0.009852
dose2	0.2196	0.4056	0.5414	0.5928

Die Ausgabe des zweiten - oben angegebenen - Kontrastes lässt sich folgendermaßen interpretieren:

- die erste Variable (*Dosis1*) vergleicht die Placebogruppe mit der Niedrig- und Hochdosisgruppe. Als solches vergleicht es den angepassten Mittelwert der Placebogruppe ($\bar{x}_{Placebo} = 2.93$) mit dem Durchschnitt der angepassten Mittelwerte für die niedrig- und hochdosierten Gruppen ($(4.71 + 5.15)/2 = 4.93$).
- der b-Wert für die erste Variable sollte daher die Differenz zwischen diesen Werten sein: $4.93 - 2.93 = 2$.
- dieser Wert wird durch die Anzahl der Gruppen innerhalb des Kontrastes (d.h. 3) dividiert und somit $2/3 = .67$ (wie in der Ausgabe) beträgt. Die zugehörige t-Statistik ist signifikant, was darauf hindeutet, dass sich die Placebogruppe signifikant vom kombinierten Mittelwert der Viagra-Gruppen unterschied.
- die zweite Variable (*Dosis2*) vergleicht die niedrig- und hochdosierten Gruppen, so dass der b-Wert die Differenz zwischen den eingestellten Mitteln dieser Gruppen sein sollte: $5.15 - 4.71 = 0.44$. Dieser Wert wird durch die Anzahl der Gruppen innerhalb des Kontrastes (d.h. 2) dividiert wird und somit $0.44/2 = 0.22$ (wie in der Ausgabe) beträgt.
- die zugehörige t-Statistik ist nicht signifikant ($p = .590$), was darauf hindeutet, dass die hochdosierte Gruppe keine signifikant höhere Libido produzierte als die niedrigdosierte Gruppe.

¹³für eine Liste vordefinierter Kontraste siehe Literatur. Kontraste können sowohl in SPSS wie auch in R durch entsprechende Kontrastcodierungen definiert werden. Bei R ist darauf zu achten, dass bei orthogonalen Kontrasten die Type III sum of squares verwendet wird, da sonst die Quatratsummen für derartige Kontraste nicht stimmen!

- der Wert für die *Kovariable* beträgt ($b = 0.416$). Wenn also der Libido eines Partners um eine Einheit zunimmt, sollte der Libido der Person um knapp eine halbe Einheit zunehmen (obwohl es nichts gibt, was auf einen kausalen Zusammenhang zwischen den beiden hinweist).
- das Vorzeichen dieses Koeffizienten zeigt in welche Richtung die Beziehung zwischen der Kovariablen und dem Ergebnis geht. Da der Koeffizient in diesem Beispiel positiv ist, bedeutet dies also, dass die Libido des Partners in einem positiven Verhältnis zur Libido des Teilnehmers steht:
- mit dem einen steigt auch der andere.
- ein negativer Koeffizient würde das Gegenteil bedeuten: wenn einer steigt, sinkt der andere.

Interpretation Kovariate

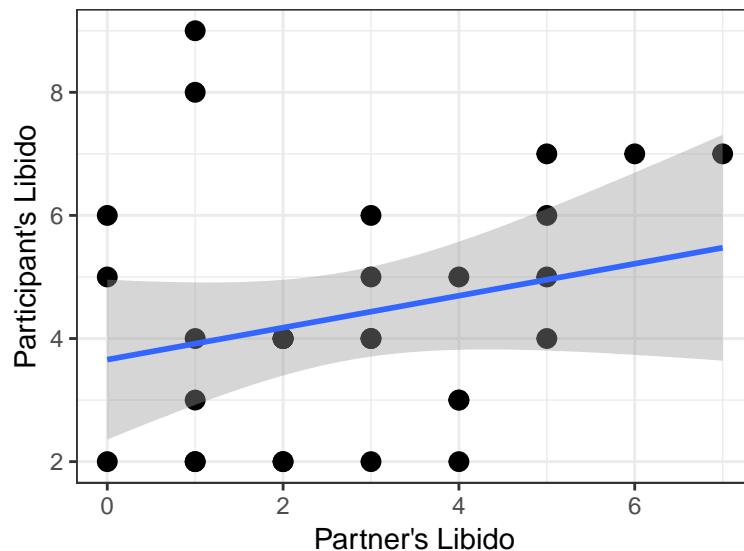
Für die Interpretation der Kovariaten verwendet man am besten die Parameterschätzungen (b) in folgender Weise:

- wenn der b -Wert für die Kovariable positiv ist, haben die Kovariable und die Ergebnisvariable eine positive Beziehung, also mit zunehmenden Werten der Kovariable steigt auch das Ergebnis!
- wenn der b -Wert negativ ist, bedeutet das das Gegenteil.

Für diese Daten war der b -Wert positiv, was darauf hindeutet, dass mit zunehmender Libido des Partners auch die Libido des Teilnehmers steigt. Eine weitere Möglichkeit, das Gleiche zu entdecken, besteht darin, einfach einen Streudiagramm der Kovariablen gegen das Ergebnis zu zeichnen.

Abschließend wird durch den Scatterplot nochmals bestätigt, was wir bereits wissen: die Kovariable bewirkt, dass mit zunehmender Partnerlibido auch die Libido des Teilnehmers zunimmt (wie die Steigung der Regressionslinie zeigt).

```
scatter <- ggplot(viagraData, aes(partnerLibido, libido)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", alpha = 0.4) +
  labs(x = "Partner's Libido", y = "Participant's Libido") +
  theme_bw()
scatter
```



Post hoc Tests

Wie bereits aus der ANOVA bekannt sein sollte, werden bei den Post hoc Tests alle Stufen der unabhängigen Variablen paarweise miteinander verglichen. Im Unterschied zur herkömmlichen ANOVA wenden jedoch bei der ANCOVA die adjustierten Mittelwerte verwendet!

Das Ergebnis zeigt die drei Vergleiche (niedrige Dosis vs. Placebo, hohe Dosis vs. Placebo, hohe Dosis vs. niedrige Dosis).

```
postHocs <- multcomp::glht(viagraModel, linfct = mcp(dose = "Tukey"))
PostHocRes <- summary(postHocs)
```

Verglichen werden die Differenzen zu den adjustierten Gruppenmitteln

- die Schätzung für die niedrige Dosis vs. Placebo beträgt $4.71 - 2.93 = 1.78$
- für die hohe Dosis vs. Placebo beträgt sie $5.15 - 2.93 = 2.22$ und
- für die niedrige vs. hohe $5.15 - 4.71 = 0.44$

Der angegebene Standardfehler bezieht sich auf die Differenz zwischen den adjustierten Mittelwerten.

Der t -Test (Differenz zwischen den Mitteln geteilt durch den Standardfehler) und dem zugehörigen p -Wert deutet auf signifikante Unterschiede zwischen der Hochdosis- und Placebogruppe ($t = 2.77, p < .050$) hin.

Kein Unterschied besteht zwischen der Niedrigdosisgruppe und der Placebogruppe ($t = 2.10, p = .120$) und der Hochdosisgruppe ($t = 0.54, p = .850$).

Die Konfidenzintervalle bestätigen diese Schlussfolgerung (weil sie für den Vergleich der Hochdosis- und Placebogruppen Null nicht enthalten).

```
PostHocRes
```

```
## 
##   Simultaneous Tests for General Linear Hypotheses
##
##   Multiple Comparisons of Means: Tukey Contrasts
##
## 
## Fit: aov(formula = libido ~ partnerLibido + dose, data = viagraData)
##
## Linear Hypotheses:
##                               Estimate Std. Error t value Pr(>|t|)    
## Low Dose - Placebo == 0     1.786     0.849    2.10    0.109    
## High Dose - Placebo == 0    2.225     0.803    2.77    0.027 *  
## High Dose - Low Dose == 0   0.439     0.811    0.54    0.852    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)

confint(postHocs)

## 
##   Simultaneous Confidence Intervals
##
##   Multiple Comparisons of Means: Tukey Contrasts
##
## 
## Fit: aov(formula = libido ~ partnerLibido + dose, data = viagraData)
##
## Quantile = 2.48
## 95% family-wise confidence level
```

```

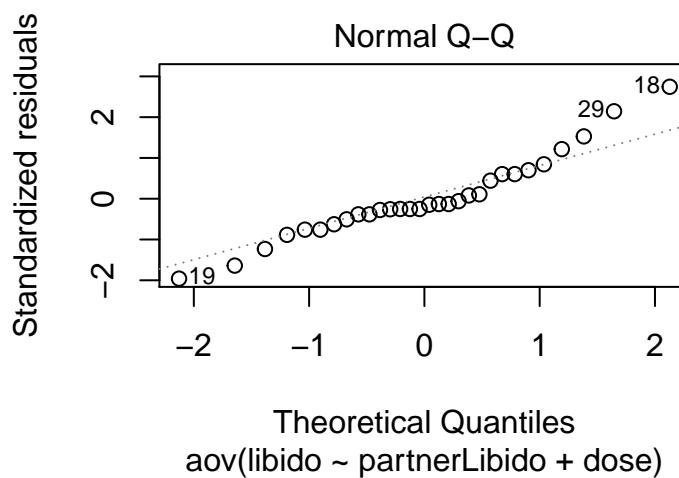
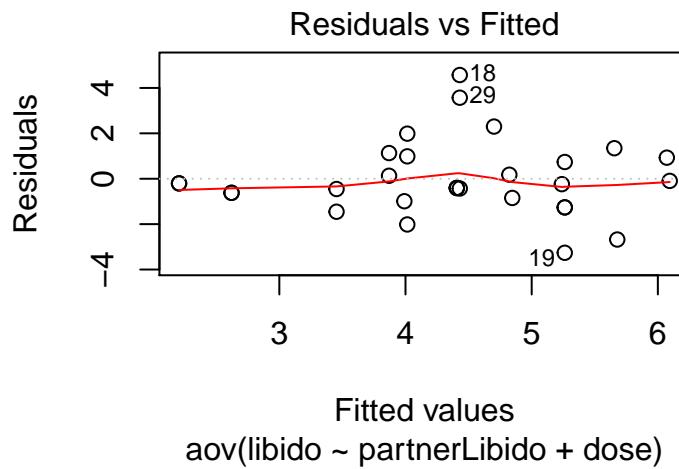
## 
## 
## Linear Hypotheses:
##                               Estimate lwr      upr
## Low Dose - Placebo == 0    1.786   -0.324  3.895
## High Dose - Placebo == 0   2.225    0.231  4.219
## High Dose - Low Dose == 0  0.439   -1.576  2.454

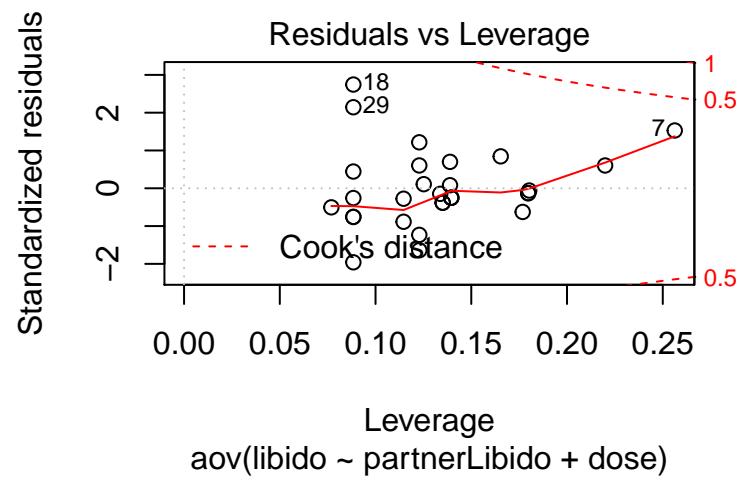
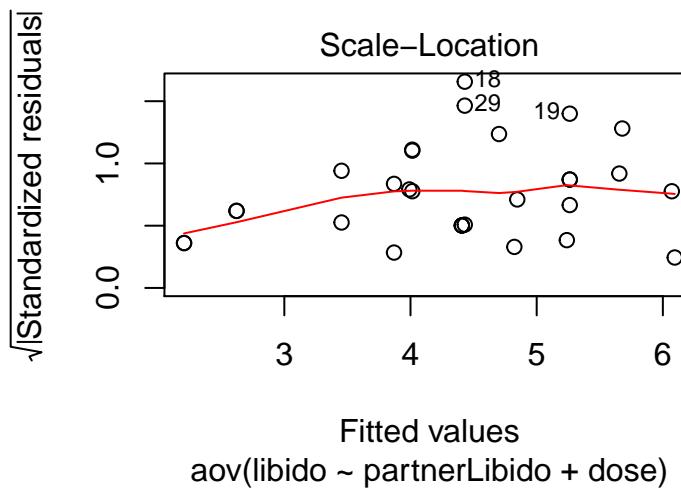
```

Nützliche Graphen

Zur Überprüfung von Voraussetzungen können sich folgende Graphiken als hilfreich erweisen:

```
plot(viagraModel)
```





Von den vier Diagramme sind die ersten beiden die wichtigsten:

- Ersterer kann zur Überprüfung der Varianzhomogenität der Varianz verwendet werden. Es zeigt sich, dass die Verteilung der Scores an einigen Stellen breiter als an anderen ist (Funneling). Die Residuen sind also heteroscedastisch.
- Das zweite Diagramm ist ein Q-Q-Diagramm. Die Punkte im Diagramm sollten nahe der diagonalen Linie liegen. Die vorliegende Verteilung deutet darauf hin, dass keine Normalverteilung vorliegt und daher eher eine robuste ANCOVA¹⁴ angebracht wäre.
- Die dritte Graphik wird auch als *Spread-Location-Plot* bezeichnet. Diese Darstellung zeigt, ob die Residuen gleichmäßig über die Bereiche der Prädiktoren verteilt sind. So können Sie die Annahme der gleichen Varianz (Homoscedastizität) überprüfen. Es ist gut, wenn man eine horizontale Linie mit gleichmäßig (zufällig) gespreizten Punkten sehen - was hier nicht der Fall ist.
- Die vierte Graphik identifiziert einflussreiche Fälle (Ausreißer). Nicht alle Ausreißer beeinflussen die linearen Regressionsanalyse im negativen Sinn, d.h. die Ergebnisse wären nicht viel anders, wenn wir sie entweder einbeziehen oder von der Analyse ausschließen würden. Sie folgen in den meisten Fällen dem Trend und sind nicht wirklich wichtig. Andererseits können einige Fälle sehr einflussreich sein,

¹⁴robuste ANCOVAs werden in dieser LV nicht näher besprochen - siehe Literatur.

auch wenn sie sich in einem angemessenen Bereich der Werte bewegen. Sie können Extremfälle gegen eine Regressionslinie sein und die Ergebnisse verändern, wenn wir sie von der Analyse ausschließen. Im Gegensatz zu den anderen Graphiken sind hier Muster nicht relevant. Man achtet auf die äußereren Werte in der oberen rechten Ecke oder in der unteren rechten Ecke. Diese Punkte sind die Orte, an die für eine Regressionslinie einflussreich sein können. Man beachtet vor allem Fälle, die *außerhalb einer gestrichelten Linie* (Cook's Distance) sind. Werte die außerhalb liegen, sind für die Regressionsergebnisse von Bedeutung. Die Regressionsergebnisse werden geändert, wenn wir diese Fälle ausschließen!

Homogenität der Steigung

Bereits im Scatterplot der nach Gruppen getrennten Regressionen konnten wir feststellen, dass die Annahme der Homogenität der Regressionssteigungen für die *Hochdosisgruppe* unterschiedlich verletzt wird. Um einen statistischen Test dieser Annahme durchzuführen, wird die ANCOVA unter Einbeziehung des Interaktionseffektes zwischen der Kovariaten und dem Prädiktor nochmals durchgeführt.

```
hoRS <- aov(libido ~ partnerLibido + dose + dose:partnerLibido, data = viagraData)
# Ident mit obiger Zeile!
# hoRS <- aov(libido ~ partnerLibido*dose, data = viagraData)
hoRS_Res <- car::Anova(hoRS, type="III")
pander(hoRS_Res)
```

Table 49: Anova Table (Type III tests)

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	53.54	1	21.92	9.323e-05
partnerLibido	17.18	1	7.035	0.01395
dose	36.56	2	7.484	0.00298
partnerLibido:dose	20.43	2	4.181	0.02767
Residuals	58.62	24	NA	NA

Die Auswirkungen der Dosis von Viagra und der Libido des Partners sind immer noch signifikant, aber da die Interaktion (partnerLibido:dose) signifikant ($p = .028$) ist, ist die Annahme der Homogenität der Regressionsgeraden verletzt.

Obwohl dieser Befund nicht überraschend ist (vgl. Graphik oben), gibt er Anlass zur Sorge über die Hauptanalyse. Dieses Beispiel veranschaulicht, warum es wichtig ist, Annahmen zu testen und nicht nur die Ergebnisse einer Analyse blind zu akzeptieren!

Bericht der Ergebnisse

Der Ergebnisbericht einer ANCOVA ist weitgehend identisch mit der einer ANOVA. Hinzu kommt lediglich die Wirkung der Kovariablen. Für die Kovariablen und den experimentellen Effekt berichten wir Details über das F -Verhältnis und die Freiheitsgrade, aus denen es berechnet wurde. In beiden Fällen wurde das F -Verhältnis aus der Division der mittleren Quadrate für den Effekt durch die mittleren Quadrate der Residuen ermittelt. Die Freiheitsgrade zur Beurteilung des F -Wertes sind daher die Freiheitsgrade für die Wirkung des Modells ($df_M = 1$ für die Kovariablen und 2 für die experimentelle Wirkung) und die Freiheitsgrade für die Residuen des Modells ($df_R = 26$ für die Kovariablen und die experimentelle Wirkung). Der Bericht könnte folgendermaßen abgefasst werden:

Die Kovariablen (Libido des Partners) zeigt einen signifikanten Zusammenhang mit dem Libido des Teilnehmers ($F(1, 26) = 4.96, p < .050, r = 0.40$). Kontrolliert man für den Effekt des Libido's des Partners, dann zeigt sich auch ein signifikanter Effekt der Dosis von Viagra auf den Libido ($F(2, 26) = 4.14, p < .050, \eta^2_{part} = .24$).

Die geplanten Kontraste zeigten, dass die Einnahme einer hohen oder niedrigen Dosis von Viagra den Libido im Vergleich zur Einnahme eines Placebos signifikant erhöht ($t(26) = 2,79, p < .010, r = 0.48$). Es gab keinen signifikanten Unterschied zwischen der hohen und niedrigen Dosis von Viagra ($t(26) = 0.54, p = .590, r = 0.11$).

Die Tukey-Post-Hoc-Tests zeigten, dass der über die Kovariate angepasste Mittelwert der Hochdosis-Gruppe signifikant größer war als der des Placebos (Differenz = 2.22, $t = 2.77, p < .050, d = 1, 13$). Es gab jedoch keinen signifikanten Unterschied zwischen der Niedrigdosis- und Placebogruppe (Differenz = 1.79, $t = 2.10, p = .110, d = 1.04$) und zwischen der Niedrigdosis- und der Hochdosisgruppe (Differenz = 0.44, $t = 0.54, p = .850, d = 0.11$).

Trotz der fehlenden Bedeutung zwischen der Niedrigdosis- und der Placebogruppe war die Effektgröße ziemlich groß.

Bühner, M. 2009. *Statistik Für Psychologen Und Sozialwissenschaftler*. Erste Auflage. Martin-Kollar-Str. 10-12, D-81829 München/Germany: Pearson Education Deutschland GmbH.

———. 2017. *Statistik Für Psychologen Und Sozialwissenschaftler*. Zweite Auflage. Lilienthalstraße 2, 85399 Hallbergmoos, Germany: Pearson Deutschland GmbH.

Coursera. 2018. “Coursera Take the World’s Best Courses.” <https://www.coursera.org/>.

DataCamp. 2018. “DataCamp Learn Data Science.” <https://www.datacamp.com/>.

Field, A. 2017. *Discovering Statistics Using R*. 2nd ed. 1 Olivers Yard, 55 City Road, London EC1Y 1SP: SAGE Publications Ltd.

Hemmerich, W. A. 2018. “StatistikGuru Multiple Lineare Regression in Spss, Version 1.96.” <https://statistikguru.de/spss/multiple-lineare-regression/einleitung-2.html>.

Stevens, J. P. 2002. “Applied Multivariate Statistics for the Social Sciences.” *APA PsycNET*.

UZH. 2018. “Multiple Regressionsanalyse.” https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/zusammenhaenge/mreg.html.

Wildt, A. 2009. *Analysis of Covariance*. 12th ed. 1 Olivers Yard, 55 City Road, London EC1Y 1SP: SAGE Publications Ltd.

Yamashita, T. 2007. “A Stepwise Aic Method for Variable Selection in Linear Regression.” *Communications in Statistics Theory and Methods*, No. 36:13:2395–2403. <https://doi.org/10.1080/03610920701215639>.