Statistische Modellbildung

Walter Gruber

2019-03-07

Contents

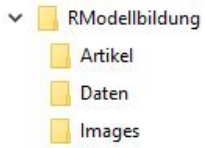


Figure 1: Dateistruktur für R-Projekt

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:

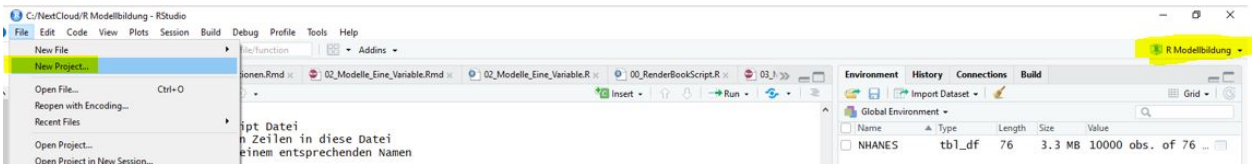


Figure 2: R-Projekt definieren

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 Arbeitens 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 dem Buch (Field 2017).

Einleitung

Donald Knuth, einer der Pioniere und bedeutensten Persönlichkeiten in der Entwicklung von Programmiersprachen, vertrat in seinem Buch Compter *Programming as an art* (Knuth 2008) die Auffassung, dass rein wissenschaftliches Arbeiten durch einen Computer erlernt werden kann (*Science is knowledge which we understand so well that we can teach it to a computer*). Alle andere Formen der Datenanalyse bezeichnete er als Kunst.

Neben statistischen Methoden, maschinellem Lernen, diversen Softwarepaketen und verschiedensten Werkzeugen stehen heutzutage auch entsprechend leistungsfähige Computer zur Verfügung, um Unmengen von Daten aufzuzeichnen und zu verarbeiten. Die *Kunst*, Wirklichkeit in bestmöglicher Form durch ein Modell abzubilden, besteht demnach darin, die:

- richtigen Fragen zu stellen.
- Daten zu sammeln, mit welchen diese Fragen beantwortet werden können.
- entsprechenden Methoden und Werkzeuge anzuwenden.
- Ergebnisse richtig zu interpretieren.
- Ergebnisse zu vermitteln.

In den meisten Fällen ist es nicht möglich, die Wirklichkeit absolut getreu durch Daten abzubilden. Seit Urzeiten verwendet daher der Mensch Modelle, um zumindest mit einer Annäherung diese möglichst genau darzustellen und daraus Rückschlüsse für sonst nicht (oder nur schwer) zu erklärende Phänome zu ziehen. In einem Artilel beschreibt G. Box (Box 1979) sehr treffend die Eigenschaften von Modellen:

> Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $P \cdot V = R \cdot T$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules. For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?". — G. Box

Allgemeine Definition des Begriffs *Modell*

Allgemein versteht man unter Modellbau die Herstellung eines konkreten, dreidimensionalen, physischen Objektes. Der Prozess der Modellbildung abstrahiert dabei mit dem Erstellen eines Modells von der Realität, weil diese meist zu komplex ist, um sie vollständig abzubilden.

Diese Vollständigkeit wird aber auch gar nicht beabsichtigt, vielmehr sollen lediglich die wesentlichen Einflussfaktoren identifiziert und dargestellt werden, die für den realen Prozess und im Modellkontext bedeutsam sind.

Der Begriff Modell entstand im Italien der Renaissance (italienisch *modello*), hervorgegangen aus lateinischesn *modulus*, einem Maßstab in der Architektur. Es wurde bis ins 18. Jahrhundert vorwiegend in der bildenden Kunst als Fachbegriff verwendet.

In der Umgangssprache wird heutzutage das Wort *Modell* für unterschiedliche Sachverhalte verwendet. Im naturwissenschaftlichen Erkenntnisprozess haben Modelle verschiedene Bedeutungen und Funktionen. Merkmale und Charakteristika, an denen ein Modell eindeutig als Modell definiert werden kann, existieren nicht (Upmeier 2010). Die Bedeutung ist zudem vom Kontext abhängig. Beim *Modellsein* werden folgende Aspekte unterschieden (Mahr 2008):

- Modell *für* etwas sein.
- Modell *von* etwas sein.

Dies muss auch bei der Beurteilung eines Modells, also der Frage, ob ein Modell ein gutes Modell ist, berücksichtigen werden. Hier ist die Frage des Zwecks des Modells entscheidend. Mahr benennt drei allgemeine Kriterien zur Beurteilung von Modellen:

1. Das Modell muss die Funktion erfüllen.
2. Durch seine Anwendung etwas von dem, wovon es ein Modell ist, zudem wofür es ein Modell ist, transportieren.
3. Das Modell muss ein Garant von Konsistenz sein.

D.h., dass das Modell garantieren muss, dass es keine Widersprüche enthält, so dass seine Anwendung nicht notwendig zu Widersprüchen führen muss. Das Modell muss über eine ausreichende pragmatische Eignung verfügen und demnach das, wofür oder wovon es ein Modell ist, ausreichend und angemessen repräsentieren.

Die Hauptfunktion in den Naturwissenschaften ist die Untersuchung und Interpretation von naturwissenschaftlichen Phänomenen. Ziel ist die Reduzierung von Komplexität und somit die Erzeugung eines fokussierten Bildes des zu untersuchenden Objekts. Man kann sagen, dass Modelle den Blick auf das Wesentliche eines Phänomens oder Gegenstandes lenken sollen. Das Modell ist somit ein Repräsentant des Originals. Modell und Original können sich aber im Material, der Dimensionierung und der Abstraktion unterscheiden. Sie können gegenständlich, bildhaft, schematisch oder formelhaft sein.

Teil I: Begrifflichkeiten

Statistisches Modell

Ein statistisches Modell, manchmal auch statistischer Raum genannt, ist ein Begriff aus der mathematischen Statistik, dem Teilbereich der Statistik, der sich der Methoden der Stochastik und Wahrscheinlichkeitstheorie bedient. Unter *statistischer Modellbildung* versteht man dabei den Prozess, ein passendes Modell für die Daten einer Beobachtungsreihe zu finden.

- Als *Prozess* wird die strukturierte und gesteuerte Reihe von Arbeitsschritten bezeichnet, welche ein bestimmtes Ergebnis hervorbringen.
- *Daten* sind Werte und Beobachtungen, die im Lauf einer (statistischen) Erhebung gesammelt werden.

- Eine *Erhebung* ist die Sammlung von Daten einer bestimmten Grundgesamtheit zum Zweck der Untersuchung eines speziellen Aspektes. Die Daten werden oft nur von einer Stichprobe der Grundgesamtheit erhoben. Erhebungen sind in der Forschung weit verbreitet.
- Eine *Stichprobe* ist die Teilmenge einer Grundgesamtheit bei einer statistischen Untersuchung, zusammengestellt, um ausgewählte Eigenschaften der Gesamtpopulation zu untersuchen.

Modellbildung

Die Modellbildung abstrahiert mit dem Erstellen eines Modells von der Realität, weil diese meist zu komplex ist, um sie vollständig abzubilden. Man unterscheidet dabei:

- **Strukturelle Modellbildung**: Bei struktureller Modellbildung ist die innere Struktur des Systems bekannt, es wird jedoch bewusst abstrahiert, modifiziert und reduziert. Man spricht hier von einem **Whitebox-Modell**.
- **Pragmatische Modellbildung**: Bei pragmatischer Modellbildung ist die innere Struktur des Systems unbekannt, es lässt sich nur das Verhalten bzw. die Interaktion des Systems beobachten und modellieren. Die Hintergründe lassen sich meist nicht oder nur zum Teil verstehen - hier spricht man von einem **Blackbox-Modell**.
- **Mischformen**: Bei Mischformen sind Teile des Systems bekannt, andere wiederum nicht. Nicht alle Wechselwirkungen und Interaktionen zwischen Teilkomponenten lassen sich nachvollziehen - hier spricht man vom **Greybox-Modell**. Diese Mischform ist die häufigste, weil es aufgrund von Kosten-Nutzen-Überlegungen meist ausreichend ist, das System auf diese Weise abzubilden.

Kennzeichen eines Modelles

Ein Modell ist im Wesentlichen durch drei Merkmale gekennzeichnet:

- **Abbildung**: eines natürlichen oder eines künstlichen Originals, wobei dieses Original selbst auch wiederum ein Modell sein kann.
- **Verkürzung**: es erfasst im Allgemeinen nicht alle Eigenschaften (Attribute) des Originals, sondern nur diejenigen, die dem Modellschaffer bzw. Modellnutzer relevant erscheinen. Diese werden häufig in Form von aggregierenden Maßzahlen (Parameter) beschrieben.
- **Pragmatismus**: sind ihren Originalen nicht eindeutig zugeordnet. Sie erfüllen ihre Ersetzungsfunktion:
  - für bestimmte Subjekte (für wen?)
  - innerhalb bestimmter Zeitintervalle (wann?)
  - unter Einschränkung auf bestimmte gedankliche oder tatsächliche Operationen (wozu?).

Im übertragenen Sinn ist damit ein statistisches Modell:

- Eine vereinfachte mathematisch-formalisierte Methode, sich der Realität anzunähern.
- Die Beschreiben des Zustandes eines Systems vor und nach Änderungen äußerer Relationen, nicht jedoch während einer Änderung.

Das Ziel ist es herauszufinden, ob man in der Natur auftretende Phänomene auf allgemein gültige Gesetzmäßigkeiten zurückführen kann. In der Regel werden Beobachtungen durchgeführt und diese als Daten aufgezeichnet. In diesen Daten gilt es Muster zu finden, die Rückschlüsse auf die Mechanismen zulassen, welche dem Phänomen zugrunde liegen. Auf diese Weise wird ein Modell von der Funktionsweise eines Phänomens erstellt.

Statistische Modelle finden Verwendung für:

- Annäherung (Approximation)
- Erklärung
- Vorhersage

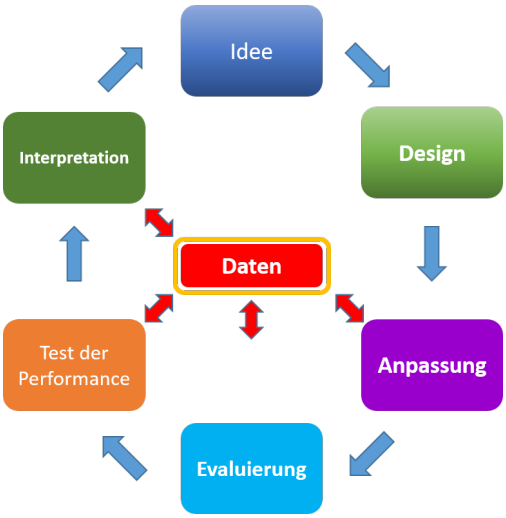


Figure 3: **Abbildung 1**: Zyklus der Modellbildung

Statistische Modelle sind in fast allen Anwendungsbereichen der Wissenschaft, aber auch des praktischen Lebens zu finden, wie z.B.:

- Bei Wettervorhersagen.
- In der Finanzmarktanalyse.
- In der Industrie und Gewerbe.
- Bei Wahlprognosen, Meinungsumfragen.
- In der Informationstechnologie.
- etc.

Zyklus der Modellbildung

Das Bilden von statistischen Modellen ist ein iterativer Vorgang, welcher durchaus mehrere zyklische Entwicklungsschritte beinhalten kann.

Grundlagen der statistischen Modellbildung

Grundlage ist eine entsprechende Fragestellung, die auf theoretischen Grundlagen basiert und möglichst präzise formuliert werden soll. Bei den Daten, die zur Beantwortung der Fragestellung geeignet sind, sollte man speziell achten auf:

- woher Sie kommen.
- was und wie Sie ein erhobenes/gemessenes Objekt abbilden.

Charakteristiken von guten Fragen sind:

- **Interessant für andere**: wie z.B. für Mitarbeiter, wissenschaftliche Gemeinschaft, Geldgeber, Allgemeinheit?
- **Noch nicht beantwortet**: wurde die Fragestellung bereits bearbeitet/beantwortet? Das erfordert eine intensive Auseinandersetzung mit dem Thema (Literaturrecherche, Kongressbeiträge, etc.).
- **Sinnvoll**: kann durch die Beantwortung eine Erklärung gefunden werden, wie etwas funktioniert?
- **Beantwortbarkeit**: kann die Fragestellung überhaupt beantwortet werden?
- **Spezifisch**: wie präzise ist die Fragestellung? Gesundes Essen führt zu besserer Gesundheit ist weniger präzise wie z.B. 5 Mal am Tag Früchte und Gemüse führt zu einer geringeren Wahrscheinlichkeit an Atemwegserkrankungen zu erkranken.

Bei den Qualitätskriterien für Daten ist folgendes zu beachten:

- **Korrektheit** → müssen mit der Realität übereinstimmen.
- **Konsistenz** → dürfen in sich und zu anderen Datensätzen keine Widersprüche aufweisen.
- **Zuverlässigkeit** → Entstehung der Daten muss nachvollziehbar sein.
- **Vollständigkeit** → muss alle notwendigen Attribute enthalten.
- **Genauigkeit** → müssen in der jeweils geforderten Exaktheit vorliegen (Beispiel: Nachkommastellen).
- **Aktualität** → müssen jeweils dem aktuellen Zustand der abgebildeten Realität entsprechen.
- **Relevanz** → Der Informationsgehalt muss den jeweiligen Informationsbedarf erfüllen.
- **Einheitlichkeit** → Die Informationen müssen einheitlich strukturiert sein.
- **Eindeutigkeit** → muss eindeutig interpretierbar sein.
- **Verständlichkeit** → müssen in ihrer Begrifflichkeit und Struktur mit den Vorstellungen der Fachbereiche übereinstimmen.
- **Redundanzfreiheit**: Innerhalb der Datensätze sollen/dürfen keine Dubletten vorkommen.

Aus der Testtheorie sind in auch die Gütekriterien (*Testgütekriterien*) der empirischen Forschung für die statistische Modellbildung anzuwenden. Diese sind:

- **Reliabilität**: Indikator für die Replizierbarkeit der Ergebnisse. Fragen müssen z.B. so eindeutig formuliert sein, dass sie nicht höchst unterschiedlich verstanden werden können.
- **Validität**: wenn die gewählten Indikatoren, Fragen und Antwortmöglichkeiten wirklich und präzise das messen, was gemessen werden soll.
- **Objektivität**: wenn die Wahl der Messenden, InterviewerInnen, PrüferInnen keinen Einfluss auf die Ergebnisse hat.

Bei der Analyse und Interpretation der Daten unterscheidet man:

- **Deskriptiv**: beschreibend (innerhalb Daten)
- **Explorativ**: Hypothesen generierend (innerhalb Daten)
- **Inferentiell**: Statement über etwas, was nicht beobachtet wird, also aus den erhobenen Daten auf die Ursachen zu schliessen, die die Daten erzeugt haben könnten (außerhalb der Daten).
- **Prädiktiv**: Vorhersage für Werte die noch nicht beobachtet wurden (außerhalb der Daten)
- **Korrelativ**: Beschreibung der Zusammenhäng zwischen Beobachtungen (innerhalb und außerhalb der Daten).
- **Mechanistisch**: nicht nur die Frage ob ein Zusammenhang besteht, sondern wie und warum der Zusammenhang besteht ist von Interesse.

Deskriptive Statistik

Die deskriptive Statistik hat zum Ziel, empirische Daten durch Tabellen, Kennzahlen (auch: Maßzahlen oder Parameter) und Grafiken übersichtlich darzustellen und zu ordnen.

Die Methoden der deskriptiven Statistik umfassen:

- Tabellen
- Diagramme
- Parameter (Maßzahlen, Kennzahlen, Kennwerte)
  - Lagemaßen: Maße der zentralen Tendenz, wie z.B. Mittelwerte, Median, Modus.
  - Streuungsmaße: für die Variabilität (Dispersion), wie z.B. Range, Varianz, Standardabweichung, Standardfehler.
  - Zusammenhangsmaße: wie z.B. die Korrelation.

Die deskriptive Statistik ist ein zentrales Element jeder formalen Analyse von Daten, hat jedoch in ihrer Eigenschaft folgende Einschränkungen:

- liefert keine Aussagen zu einer über die untersuchten Fälle hinausgehenden Grundgesamtheit .
- Es ist keine Überprüfung von Hypothesen möglich.
- Verwendet keine stochastischen Modelle (Grundlage der induktiven Statistik).
- getroffenen Aussagen können nicht durch Fehlerwahrscheinlichkeiten abgesichert werden.

Explorative Datenanalyse (EDA)

Die explorative Datenanalyse/Statistik hat zum Ziel, bisher unbekannte Strukturen und Zusammenhänge in den Daten zu finden und hierdurch neue Hypothesen zu generieren. Diese auf Stichprobendaten beruhenden Hypothesen können dann im Rahmen der schließenden Statistik mittels wahrscheinlichkeitstheoretischer Methoden auf ihre Allgemeingültigkeit untersucht werden.

Die Methoden der explorativen Statistik sind meist identisch mit denen der deskriptiven Statistik, unterscheiden sich aber bezüglich der *Ziele der Analyse*. Bei der EDA sollten vor allem die nachfolgenden Fragen geklärt werden:

- Haben wir die richtigen Daten zur Beantwortung der Fragestellung?
- Brauchen wir mehr Daten?
- Müssen wir die Fragestellung verfeinern?
- Was kann aus den vorliegenden Daten abgeleitet werden?

Die Ziele der EDA sind:

- bisher unbekannte Strukturen und Zusammenhänge in den Daten zu finden.
- Annahmen (Hypothesen) über die Ursache und den Grund der beobachteten Daten zu bilden.
- Annahmen einzuschätzen, worauf statistische Inferenz basieren kann.
- Die Auswahl von passenden statistischen Werkzeugen und Techniken zu unterstützen.
- Eine Basis für die weitere Datensammlung durch Umfragen oder Design von Experimenten bereitzustellen.

Speziell in Bezug auf die erforderliche Stichprobengröße sollte vor Begin der Datenerhebung die mindest notwendigen Fallzahlen (optimaler Stichprobenumfang) bestimmt werden, der für den Nachweise praktisch bedeutsamer Effekte notwendig ist. Sowohl in R als auch in diversen Anwednungen gibt es die Möglichkeit, diese optimale Stichprobengröße a priori zu bestimmen.

Voraussetzung dafür sind jedoch Kenntnisse über den zu erwartenden Effekt. Dieser kann entweder durch einer Meta-Analyse oder durch entsprechende Erfahrungswerte bestimmt werden. Zusätzlich ist dann noch die Irrtumswahrscheinlichkeit, die gewünschte Teststärke und die verwendete statistischen Methode festzulegen. Je größer eine Stichprobe ist, desto genauer werden auch die daraus abgeleiteten Kennwerte und Teststatistiken Auskunft über die in der Population gültigen Werte liefern. Aus praktischen/finanziellen/zeitlichen Gründen wird man jedoch i.A. danach trachten, eine möglichst kleine Stichprobe zu erheben. Generell lassen sich diesbezüglich folgende Aussagen treffen:

- bei großen Stichproben werden auch kleine Effekte im statistischen Sinn signifikant (damit ist nicht gesagt, dass auch eine praktische Signifikanz des Effektes vorliegt).
- bei kleinen Stichproben wird ein kleiner Effekt oft statistisch nicht signifikant. Wenn doch, ist das oft nur ein Zufallsergebnis und kann i.A. nicht reproduziert werden.
- eine sorgfältige Stichprobenplanung geht mit einer intensiven inhaltlichen Auseinandersetzung einher. Der Vorteil dieser Planung liegt also nicht nur in der Wahrscheinlichkeit, einen vorhandenen Effekt auch stastisch absichern zu können, sondern vor allem auch darin, dass eventuell schon beantwortete Fragen rechtzeitig entdeckt werden!

Inferenzstatistik

Bei der inferentiellen Statistik wird eine Zufallsstichprobe aus einer Grundgesamtheit entnommen, um Rückschlüsse auf die Grundgesamtheit zu ziehen und diese zu beschreiben. Weitere Begriffe sind:

- analytische Statistik
- inferenzielle Statistik
- induktive Statistik
- schließende Statistik

Die inferentielle Statistik ist in den Situationen nützlich, in denen es nur schwer oder gar nicht möglich ist, eine vollständige Grundgesamtheit zu untersuchen. Wesentliche Merkmale der Inferenzstatistik sind:

- trifft Wahrscheinlichkeitsaussagen über Populationswerte.
- Daten werden in Form von (Zufalls-) Stichproben aus der Grundgesamtheit entnommen.
- Signifikanztests und Intervalleschätzungen bieten die Entscheidungsgrundlage hinsichtlich der postulierten Hypothesen.

Prädiktive Statistik

Prädikative Statistik (predictive analytics) verwendet (historische) Daten, um zukünftige Ereignisse vorherzusagen. Im Allgemeinen werden vorliegende (historische) Daten verwendet, um ein mathematisches (prädiktives Modell) Modell zu erstellen. Dieses Modell soll bestmöglich wichtige Trends in den Daten erfassen.

Dieses wird dann auf aktuelle Daten angewendet, um zukünftige Ereignisse vorherzusagen, oder um Aktionen vorzuschlagen, mit denen optimale Ergebnisse erreicht werden können. Vor allem in den letzten Jahren hat diese Form der Statistik in den Bereichen von Big Data und Machine Learning sehr an Bedeutung gewonnen.

Mechanististische Statistik

Zusammenhängen zwischen mehreren Variablen mit gleichzeitiger Interpretation der Kausalitäten werden häufig mit Hilfe eines Strukturgleichungsmodells (SEM[1]) analysiert. Dabei handelt es sich um ein statistisches Modell, das das Schätzen und Testen korrelativer Zusammenhänge zwischen abhängigen Variablen und unabhängigen Variablen sowie den verborgenen Strukturen dazwischen erlaubt.

Eine Besonderheit von Strukturgleichungsmodellen ist das Überprüfen latenter (nicht direkt beobachtbarer) Variablen. Spezialfälle von Strukturgleichungsmodellen sind:

- Pfadanalyse
- Faktorenanalyse
- Regressionsanalyse

Ein Strukturgleichungsmodell stellt wiederum einen Spezialfall eines sogenannten Kausalmodells dar.

Teil II: Modelle mit einer Variablen

Datensatz

Betrachten wir zunächst einen einfachen Datensatz aus dem Projekt MOSAIC Data Sets (Paket mosaicData, **CPS85**)[2].

Diese Datei beinhaltet $N =$ `534` Beobachtungen und $k =$ `11` Variablen, deren Namen in folgender Tabelle nochmals separat angeführt sind:

[1] *S*tructural *E*quation *M*odeling

[2] Project MOSAIC, is a community of educators working to develop a new way to introduce mathematics, statistics, computation and modeling to students in colleges and universities.

LNr	Variablenname
1	wage
2	educ
3	race
4	sex
5	hispanic
6	south
7	married
8	exper
9	union
10	age
11	sector

Angenommen Sie müssten auf Basis der vorliegenden Daten für eine Person das durchschnittliche Einkommen (*wage*) schätzen, ohne dabei andere Variablen zu berücksichtigen. Welchen Wert würden Sie wählen?

Deskriptive Statistik

Bevor man diese Frage beantwortet, sollte man sich die deskriptive Statistik der entsprechenden Variablen genauer ansehen. In der vorliegenden Fragestellung handelt es sich um eine intervallskalierte Variable, daher ist die Betrachtung der Kennwerte für zentrale Tendenzen (Mittelwert, Median, Modus, Minimum, Maximum und Range), der Dispersion (Varianz, Standardabweichung, Quartile, Standardfehler, Konfidenzintervalle, Schiefe und Kurtosis), sowie die Darstellung der Verteilung in einem Histogramm sehr hilfreich.

Unter bestimmten Voraussetzungen, eignet sich der Mittelwert als bester Schätzer (bzw. als einfachste Modellvorstellung). Bevor man sich jedoch der Auswertung von Daten widmet, ist es sehr empfehlenswert die zugrundeliegende Datenstruktur zu analysieren und auch zu dokumentieren. Im nachfolgenden Kapitel wird ein sehr nützliches Paket für genau diese Analyse kurz vorgestellt.

Codebooks in R

In R hat man die Möglichkeit, mit Hilfe des Pakets *codebook* eine genaue Beschreibung der Daten (inklusive einer deskriptiven Statistik für jede Variable) zu erstellen. Für den vorliegenden Datensatz wurde auszugsweise eines erstellt, welches in Kapitel Codebook CPS85 zu finden ist.

Tabellen

Im Codebook werden neben den deskriptiven Kennwerten (für kategorielle Variablen) Häufigkeitstabellen angegeben. Wir wollen uns daher einen kurzen Überblick über Häufigkeitstabellen in R verschaffen. Kopiere den nachfolgenden Code in den Editor und führe in aus. Diskutiere die Ergebnisse.

```
Income <- CPS85$wage
library(pastecs)
  kable(stat.desc(Income))
  # DT::datatable(data.frame(stat.desc(Income)))
library(psych)
  kable(describe(Income))
  # DT::datatable(data.frame(describe(Income)))

# HÃ¤ufigkeitstabellen
SR  <- table(CPS85$sex, CPS85$race)
kable(SR)
SRM <- table(CPS85$sex, CPS85$race, CPS85$married)
kable(SRM)
# HÃ¤ufigkeitstabellen mit Randsummen
```

```r
x0 <- addmargins(table(CPS85$sex, CPS85$race))
kable(x0)
# HÃ¤ufigkeitstabellen in Prozent
x1 <- addmargins(round(100*prop.table(table(CPS85$sex, CPS85$race)), 2))
kable(x1)
```

Mittelwerts-Modell

Wie bereits erwähnt, wäre unter bestimmten Voraussetzungen (Verteilungseigenschaften) der Mittelwert ein guter Schätzer, da dieser die folgende Eigenschaft besitzt:

- Die Summe der quadrierten Abweichungen der Beobachtungswerte x_i von einem beliebigen Punkt m wird minimal, wenn dieser Punkt $m = \bar{x}$, also der arithmetische Mittelwert ist!

Formal berechnet sich das arithmetische Mittel:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

Die Aussagekraft des arithmetischen Mittels beschränkt sich jedoch ganz wesentlich, wenn nicht weitere Kennwerte der Daten bekannt sind. Vor allem ist es von Interesse, die Streung (Variabilität) der Werte um den Mittelwert zu kennen. Diese wird durch das Streuungsmaß, welches als durchschnittliche Abweichung der Messwerte um den Mittelwert gesehen werden kann, beschrieben:

$$s = \sqrt{\sum_{i=1}^{N} (x_i \frac{x_i - \bar{x})^2}{N-1}} \tag{2}$$

Modell *idealer* Daten

Betrachten wir zunächst einen Datensatz, in dem das Gehalt (*wage*) symmetrisch und in Form einer Glockenkurve gegeben ist. Kopiere den nachfolgenden Code in dein R-Script und führe diesen dann aus.

```r
options(digits = 2)
DF      <- CPS85
set.seed(825)
DF$wage <- rnorm(n = nrow(DF), mean = 8, sd = 2.3)
Income  <- DF$wage

# Kennwerte berechnen
MW <- mean(Income)
SD <- sd(Income)
MD <- median(Income)
RA <- range(Income)
# Daten anzeigen
# p1 <- hist(Income)
# p2 <- boxplot(Income)

# starke Abweichungen entfernen
TrimmedIncome <- Income[!Income %in% boxplot.stats(Income)$out]
# Kennwerte berechnen
MW_T <- mean(TrimmedIncome)
SD_T <- sd(TrimmedIncome)
```

```r
MD_T <- median(TrimmedIncome)
RA_T <- range(TrimmedIncome)
# Daten anzeigen
# p3   <- hist(TrimmedIncome)
# p4   <- boxplot(TrimmedIncome)
```

Die Ergebnisse der statistischen Kennwerte $\bar{x} = (8.0952214)$, $med = (8.2289139)$, $sd = (2.3218623)$ und vor allem (das hier nicht angezeigte) angezeigte Histogramm lassen vermuten, dass der Mittelwert als *Modell* durchaus geeignet ist. Vor allem wenn man noch durch *Beschneidung (trim)* der Daten die kleinsten und größten Werte entfernt, nehmen der Mittelwert $\bar{x}_{trim} = 8.1089335$ und Median $med_{trim} = 8.2306598$ den gleichen Wert ein.

Modell *schiefer* Daten

Um die Eigenschaften des Mittelwertes bei vorliegen von starken Abweichungen in den Daten noch besser zu verdeutlichen, verwenden wir einerseits die Originaldaten (welche für sich schon schiefverteilt sind) und setzen zusätzlich bei 50 zufällig gewählten Personen das Einkommen drastisch hinauf. Kopiere den folgenden Code ins RStudio und führe diesen dann aus.

```r
IncomeNew     <- CPS85$wage
ID            <- sample(1:534, 50)
IncomeNew[ID] <- IncomeNew[ID] + 18
```

Aufgabenstellung 1

Berechne für die neuen Daten folgende Kennwerte und zeichne sowohl ein Histogramm, als auch einen Boxplot.

- Mittelwert
- Standardabweichung
- Median
- Range
- Getrimmten Mittelwert, wobei jeweils 10% der Daten vom unteren und oberen Wertebereich unberücksichtigt bleiben sollen.

Diskutiere die Ergebnisse. Die Lösung zu diesen Aufgaben findest du in Lösung Aufgabe 1.

Graphische Darstellung

Zur Veranschaulichung von Verteilungseigenschaften einer Variablen eignen sich vor allem Histogramme, Boxplots und Q-Q-Plots. In Kombination mit den entsprechenden Tabellen, können bereits durch die einfache deskriptive Statistik wertvolle Aussagen über die statistischen Eigenschaften der Daten gewonnen werden. Kopiere den folgenden Code ins RStudio und führe diesen dann aus. Diskutiere die Ergebnisse.

```r
# Histogramme und Density-Plots
p1 <- ggplot(CPS85, aes(x = wage)) + geom_histogram()
p2 <- ggplot(CPS85, aes(x = age)) + geom_histogram(binwidth = 4)
p3 <- ggplot(CPS85, aes(x = exper)) + geom_density()
# Boxplots
x1 <- mosaic::mean_(wage ~ sex, data = CPS85)
x2 <- mosaic::sd(wage ~ sex, data = CPS85)
x3 <- mosaic::quantile(wage ~ sex, data = CPS85)
x4 <- mosaic::favstats(wage ~ sex, data = CPS85)
x5 <- gf_boxplot(wage ~ sex, data = CPS85)
x6 <- gf_point(wage ~ sex, data = CPS85)
# Q-Q Plots
```

```r
qqnorm(CPS85$wage, pch = 1, frame = FALSE)
qqline(CPS85$wage, col = "steelblue", lwd = 2)
```

Bemerkung Ausreißer

Eine der häufigsten Ursachen für Verzerrungen in den Verteilungseigenschaften einer Variablen sind Ausreißer. Die Behandlung von Ausreißern ist ein eigenes und heftig diskutiertes Thema in der Statistik. Eine (wenngleich nicht unbedenkliche) Methode ist das bereits verwendete *Trimmen* der Daten. Nachfolgendes Beispiel zeigt eine weitere Möglichkeit[3], Ausreißer aus einer Analyse zu entfernen. Es sei an dieser Stelle nochmals explizit darauf hingewiesen, dass ein beliebiges Weglassen von *störenden* Werten durchaus bedenklich ist und eigentlich nur im Sinne einer explorativen Analyse von Daten (was wäre, wenn die Daten keine Ausreißer hätten?) gerechtfertigt werden kann! Kopiere den folgenden Code ins RStudio und führe diesen dann aus. Diskutiere die Ergebnisse.

```r
# panderOptions("table.split.table", 120)
# pander(head(CPS85), style = "rmarkdown")
MW_Wage    <- mean(CPS85$wage, na.rm = TRUE)
Med_Wage   <- median(CPS85$wage, na.rm = TRUE)
SD_Wage    <- sd(CPS85$wage, na.rm = TRUE)
CPS85_1    <- CPS85[!CPS85$wage %in% boxplot.stats(CPS85$wage)$out,]
MW_Wage_1  <- mean(CPS85_1$wage, na.rm = TRUE)
Med_Wage_1 <- median(CPS85_1$wage, na.rm = TRUE)
SD_Wage_1  <- sd(CPS85_1$wage, na.rm = TRUE)
# Modell0_Graph_1 ----
ggplot(CPS85, aes(x = wage)) +
  geom_histogram(aes(y    = ..density..),
                 binwidth = .5,
                 colour   = "black", fill = "white") +
  geom_density(alpha = .2, fill = "#FF6666") +
  geom_vline(aes(xintercept = mean(wage, na.rm = T)),
             color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept=median(wage, na.rm=T)),
             color = "blue", linetype = "dotted", size = 1) +
  theme_bw()
# Modell0_Graph_2
ggplot(CPS85_1, aes(x = wage)) +
  geom_histogram(aes(y    = ..density..),
                 binwidth = .5,
                 colour   = "black",
                 fill     = "white") +
  geom_density(alpha = .2,
               fill  = "#FF6666") +
  geom_vline(aes(xintercept = mean(wage, na.rm = T)),
             color    = "red",
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = median(wage, na.rm = T)),
             color    = "blue",
             linetype = "dotted",
             size     = 1) +
  theme_bw()
```

[3] wir haben bereits bei der Mittelwertsfunktion *mean()* das Argument *trim* kennengelernt.

```r
pander(shapiro.test(CPS85_1$wage), style = "rmarkdown")
```

Güteschätzung des Mittelwertsmodells

Ein wichtiger Bestandteil einer Modellbildung ist die Abschätzung der Güte des jeweilig erstellten Modells. Für das Mittelwertsmodell eignet sich der Standardfehler (siehe Eq. (5)) als Kennwert zu Abschätzung der Genauigkeit des Modells.

$$SE = \frac{s}{\sqrt{N}} \tag{3}$$

$$s = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \tag{4}$$

$$\tag{5}$$

Der Standardfehler (englisch: standard error, meist SE abgekürzt) ist die Standardabweichung der Stichprobenkennwertverteilung einer Stichprobenfunktion. In der Regel bezieht sich der Standardfehler dabei auf den Mittelwert und wird meistens dann als *standard error of the mean* (SEM abgekürzt) bezeichnet.

Erläuterung zum SE

Wenn wir viele zufällige Stichproben aus derselben Grundgesamtheit ziehen und jeweils den Mittelwert berechnen, würden diese Mittelwerte in der Regel unterschiedlich sein.

Die Mittelwerte haben ihre eigene Verteilung (die wiederum ihren eigenen Mittelwert und ihre eigene Standardabweichung hat). Der Standardfehler des Mittelwerts (also der SEM und damit die Schätzung des Mittelwerts der Grundgesamtheit aus dem Mittelwert der Stichprobe) ist die Standardabweichung der Mittelwerte für alle möglichen Stichproben (mit jeder möglichen Stichprobengröße) die aus der Grundgesamtheit gezogen werden können.

Offenbar spielt bei der Berechnung dieser Kennwerte die Stichprobengröße N eine Rolle. Welche Werte nimmt s und respektive SE ein, wenn $N \to \infty$ geht?

Wir halten fest, dass die Stichprobenstreuung s abhängig ist von:

- der Streuung σ in der Grundgesamtheit
- der Stichprobengröße N

Die Streuung in der Grundgesamtheit ist (auch wenn meist unbekannt) ein fixer Wert. Wird N sehr groß nähert sich die Standardabweichung diesem Wert. Im Extremfall, also wenn $N = N_{Pop}$, streuen die Werte genau mit σ!

Beim Standardfehler hingegen nähert sich mit zunehmenden N der Wert von SE der Null! Im Extremfall, also wenn $N = N_{Pop}$, gibt es nur mehr einen Mittelwert (und der ist gleich μ), welcher auch nicht mehr streut $\Rightarrow$ die Streuung $SE = 0$.

Konfidenzintervall um den Mittelwert

Aus Stichproben errechnen wir einen oder mehrere verschiedene Werte, die Schätzwerte für die Grundgesamtheit darstellen sollen. Man spricht hier von *Punktschätzer*, da eben jeweils genau ein Wert (Anteils-, Mittelwert oder andere Größe, z. B. Regressionskoeffizient) geschätzt wird.

Wünschenswerte Eigenschaften von Schätzern sind:

- *Erwartungstreue*: Der Erwartungswert (Mittelwert) der Kennwerteverteilung soll dem wahren Parameter in der Grundgesamtheit entsprechen.

- *Effizienz*: Die Streuung des Schätzers soll möglichst klein sein (d. h., die Schätzwerte sollen möglichst häufig möglichst nahe am wahren Wert liegen)
- *Konsistenz*: Mit zunehmendem Stichprobenumfang sollen Abweichungen vom wahren Wert geringer werden.

Der Punktschätzer ist der beste Schätzer für den (unbekannte) Parameter der Grundgesamtheit. Dennoch ist es recht unwahrscheinlich, dass der Punktschätzer genau dem Parameter entspricht[4]. Daher sollte man die Punktschätzung durch eine Intervallschätzung ergänzen, die eine größere Wahrscheinlichkeit aufweist – um den Preis einer größeren Bandbreite.

Die Intervallschätzung zielt nun darauf ab, einen Bereich anzugeben, der mit einer gewissen (von der Forscherin gewählten) Wahrscheinlichkeit den wahren Wert enthält (überdeckt). Dieser Bereich heißt *Konfidenzintervall*. Die Wahrscheinlichkeit, mit der das Intervall den wahren Wert enthält, sollte in der Regel möglichst hoch sein. Der trade-off: Je größer die gewählte Wahrscheinlichkeit, desto breiter das resultierende Intervall.

Das Konfidenzintervall berechnet sich aus:

$$KI = \bar{x} \pm SE \cdot t_{1-\frac{\alpha}{2};n-1} \tag{6}$$

Die Eigenschaften des Konfidenzintervalls lassen sich sehr schön in einer Simulation von Geoff Cumming veranschaulichen.

Codebook CPS85

Das nachfolgende Codebook zeigt die Datenstruktur des Datensatzes *CPS85*.

```
================================================================================
```

wage

Storage mode: double

```
   Min.:   1.000
1st Qu.:   5.250
 Median:   7.780
   Mean:   9.024
3rd Qu.:  11.250
   Max.:  44.500
```

```
================================================================================
```

educ

Storage mode: integer

```
   Min.:   2.000
1st Qu.:  12.000
 Median:  12.000
   Mean:  13.019
3rd Qu.:  15.000
   Max.:  18.000
```

```
================================================================================
```

race

[4] siehe Prof. Dr. Wolfgang Ludwig-Mayerhofer, Uni Siegen, Punkt- und Intervallschätzungen, oder Springer

---

Storage mode: integer Factor with 2 levels

Values and labels N Percent

```
          1 'NW'   67   12.5
          2 'W'   467   87.5
```

==========

sex

Storage mode: integer Factor with 2 levels

Values and labels N Percent

```
           1 'F'  245   45.9
           2 'M'  289   54.1
```

==========

hispanic

Storage mode: integer Factor with 2 levels

Values and labels N Percent

```
        1 'Hisp'   27    5.1
        2 'NH'    507   94.9
```

==========

south

Storage mode: integer Factor with 2 levels

Values and labels N Percent

```
          1 'NS'  378   70.8
          2 'S'   156   29.2
```

==========

married

Storage mode: integer Factor with 2 levels

Values and labels N Percent

```
     1 'Married'  350   65.5
     2 'Single'   184   34.5
```

==========

exper

Storage mode: integer

```
   Min.:   0.000
1st Qu.:   8.000
 Median:  15.000
   Mean:  17.822
3rd Qu.:  26.000
   Max.:  55.000
```

==========

union

Storage mode: integer Factor with 2 levels

Values and labels N Percent

```
1 'Not'    438   82.0
2 'Union'   96   18.0
```

==========

age

Storage mode: integer

```
   Min.:  18.000
1st Qu.:  28.000
 Median:  35.000
   Mean:  36.833
3rd Qu.:  44.000
   Max.:  64.000
```

==========

sector

Storage mode: integer Factor with 8 levels

Values and labels N Percent

```
1 'clerical'   97   18.2
2 'const'      20    3.7
3 'manag'      55   10.3
4 'manuf'      68   12.7
5 'other'      68   12.7
6 'prof'      105   19.7
7 'sales'      38    7.1
8 'service'    83   15.5
```

Lösungen

Aufgabe_1

```
IncomeNew     <- CPS85$wage
set.seed(21430)
ID            <- sample(1:534, 50)
IncomeNew[ID] <- IncomeNew[ID] + 180
```

```r
# Kennwerte berechnen
MW_A1         <- mean(IncomeNew)
SD_A1         <- sd(IncomeNew)
MD_A1         <- median(IncomeNew)
RA_A1         <- range(IncomeNew)
MW_A1_Trimmed <- mean(IncomeNew, trim = .1)

# Daten anzeigen
# p5 <- hist(IncomeNew)
# p6 <- boxplot(IncomeNew)
```

zurück zur Aufgabenstellung

Teil III: Modelle mit mehr Variablen

Korrelationen

Korrelation ist ein Maß für den statistischen Zusammenhang zwischen zwei Datensätzen. Unabhängige Variablen sind daher stets unkorreliert. Korrelation impliziert daher auch stochastische Abhängigkeit. 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.

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).

Linearität

Ein Korrelationskoeffizient zeigt die Stärke eines *linearen Zusammenhangs* zwischen zwei Variablen. Aber der Wert von r charakterisiert nicht die genaue Art des Zusammenhangs oder das Aussehen des Punktdiagramms beider Variablen[5].

Korrelationskoeffizienten

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.

Weiter Infos zu den einzelnen Korrelationskoeffizienten sind der Literatur zu entnehmen. Eine übersichtliche Darstellung findet man auch auf der Website von MatheGuru.

[5] Abbildungen aus Matheguru

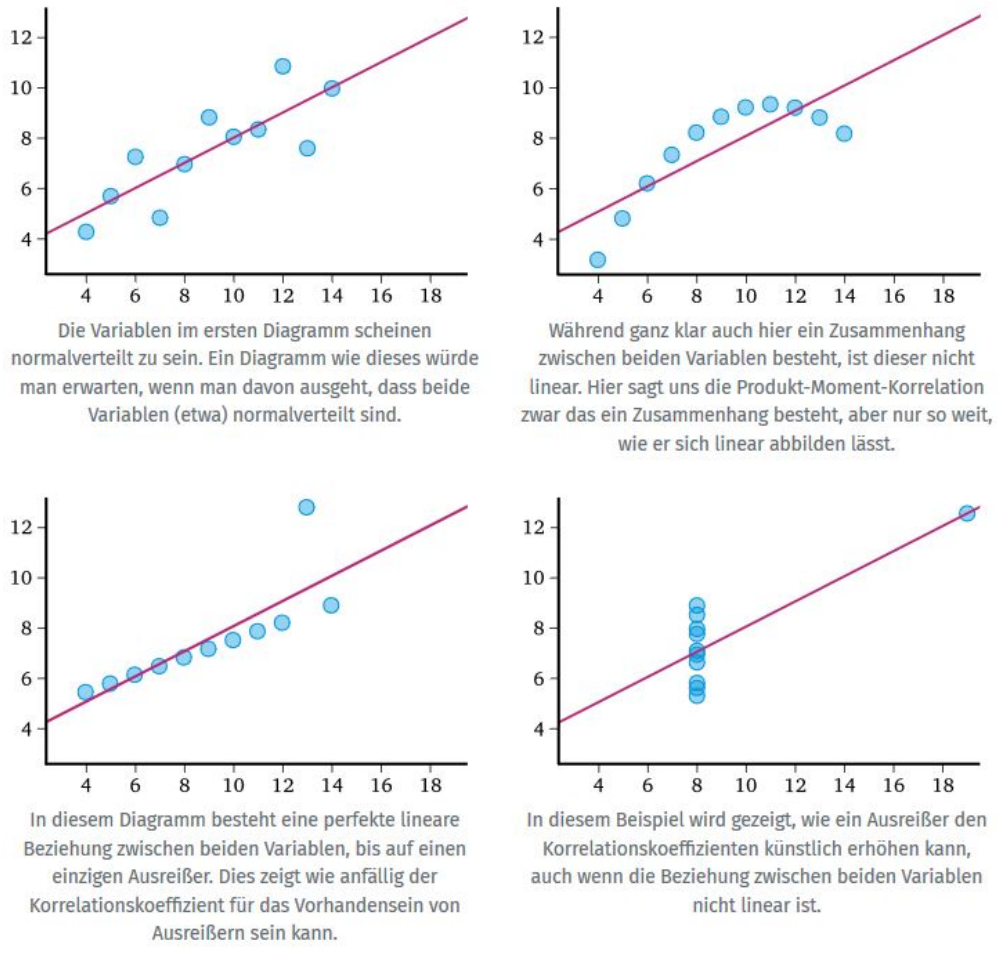


Figure 4: **Abbildung 1**: Korrelation und Linearität

			Nomialskaliert		
			dichotom		
	Intervallskaliert	Ordinalskaliert	künstlich	natürlich	polytom
Intervallskaliert	Pearson Produkt-Moment-Korrelation	Spearman's Rho Kendall's Tau polychorische Korrelation	punktbiseriale Korrelation biseriale Korrelation	punktbiseriale Korrelation	η-Koeffizient
Ordinalskaliert		Spearman's Rho Kendall's Tau polychorische Korrelation	biseriale Rangkorrelation polychorische Korrelation	biseriale Rangkorrelation	Cramér's V
Nomialskaliert (künstlich dichotom)			Punkttetrachorische Korrelation (φ-Koeffizient) Tetrachorische Korrelation	Punkttetrachorische Korrelation (φ-Koeffizient) ν-Koeffizient	Cramér's V
Nomialskaliert (natürlich dichotom)				Punkttetrachorische Korrelation (φ-Koeffizient) Yule's Y	Cramér's V
Nomialskaliert (polytom)					Cramér's V

Figure 5: **Abbildung 2**: verschiedene Korrelationskoeffizienten

Herleitung

Bereits bei der deskriptiven Statistik haben wir mit dem Maß der Varianz (s^2) einen Kennwert definiert, der die Schwankungen bezüglich des entsprechenden Mittelwertes beschreibt. Per Definition ist die Varianz die durchschnittliche Summe der quadrierten Abweichungen zum Mittelwert, also:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \tag{7}$$

Betrachtet man zwei (normalverteilte) intervallskalierte Variablen x und y, dann lässt sich diese Idee auch als ein Kennwert der gemeinsamen Variablität der beiden Variablen definieren:

$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{N-1} \tag{8}$$

Dieser Kennwert nennt sich *Kovarianz* (*cov*). Da dieser Kennwert an die entsprechenden Einheiten der Variablen gebunden ist, normiert man i.A. dieses Maß durch das Produkt der jeweiligen Standardabweichung s_x und s_y. Dieses normierte Maß bezeichnet man als *Korrelationskoeffizient* (r):

$$r(x,y) = \frac{cov(x,y)}{s_x \cdot s_y} \tag{9}$$

Beispiel

Anhand des bereits verwendeten Datensatzes (*CPS85*) wollen wir die Beziehung der Variablen Gehalt (*wage*), Ausbildung (*educ*) und Berufserfahrung (*exper*) berechnen und graphisch darstellen. Kopiere den folgenden Code ins RStudio und führe diesen dann aus. Diskutiere die Ergebnisse.

```
M      <- data.frame(wage = CPS85$wage, educ = CPS85$educ, exper = CPS85$exper)
Korr_1 <- cor(M)
pander(Korr_1, style = "rmarkdown")
# DT::datatable(round(Korr_1,2))
corrplot(cor(M), method = "ellipse")
```

Einfache Regression

Will man bei der Korrelationsanalyse den Zusammenhang von Variablen beschreiben, versucht man in der Regressionsanalyse eine Variable mittels einer linearen Funktion durch eine (oder mehrere) andere Variablen zur *erklären*. Nichts desto trotz sind Korrelation und Regression sehr eng miteinander verknüpft.

Der Begriff Regression tauchte erstmalig 1877 in einer von Sir Francis Galton abgefassten wissenschaftlichen Studie auf. In einer späteren Studie über die Körpergröße von Vätern und deren Söhnen wendete er den Gedanken der Regressionsanalyse erneut an.

Er fand heraus, dass Söhne sehr großer (kleiner) Väter zwar groß (klein), aber etwas kleiner (größer) sind als diese. Die Körpergröße entwickelt sich somit immer wieder in Richtung des Durchschnitts zurück. Als Engländer bezeichnete Galton diesen Prozess als *Regression*[6].

Zwischen der Körpergröße der Söhne und der Väter besteht somit ein Zusammenhang, dessen Stärke mit Hilfe der Korrelation ausgedrückt werden könnte. Im Unterschied zur Korrelationsanalyse unterstellt man bei der Regressionsanalyse jedoch sehr oft auch die kausale Richtung des Zusammenhangs:

> Die Körpergröße der Söhne ist abhängig von der Körpergröße des Vaters und nicht umgekehrt.

[6] was mit *Rückschritt*, *Rückkehr* oder *rückläufige Entwicklung* übersetzt werden kann (siehe Deskriptive Statistik und moderne Datenanalyse).

Entsprechend bezeichnete Galton:

- die Größe der Söhne als *abhängige Variable* (*dependent variable*, **DV**) und
- die Größe der Väter als *unabhängige Variable* (*independent variable*, **IV**).

Häufig werden die Variable die vorhergesagt werden soll bei der Regression *Kriterium* (y_i)und die Variable(n) die für die Vorhersage eingesetzt wird/werden *Prädiktor(n)* (x_{1i})[7] genannt. Anhand des Prädiktors wird demzufolge das Kriterium vorhergesagt.

Der Schluss, dass die Regression die Kausalität von Zusammenhängen *beweist*, ist damit allerdings nicht (immer) erlaubt. Die Kausalität (Wirkungsrichtung) muss zuvor theoretisch abgeleitet werden, bevor sie empirisch (mit Hilfe der Regression) bewiesen werden kann. So ist die Richtung der Kausalität bei Fragen wie:

- ist es das Alter des Bräutigams, welches das Alter der Braut bestimmt, oder umgekehrt?
- beeinflusst sich das Alter der verheirateten Paare gar gegenseitig?

nicht zu bestimmen. Manchmal ist die Kausalität jedoch sehr offensichtlich:

- der Blutdruck hat keinen Einfluss auf das Alter, sondern das Alter hat einen Einfluss auf den Blutdruck.
- die Körpergröße hat einen Einfluss auf das Körpergewicht, aber umgekehrt lässt sich dieser Zusammenhang wohl theoretisch kaum herleiten.
- mit Zunahme des CO_2 Gehaltes in der Atmosphäre steigt die durchschnittliche Temperatur, eine umgekehrte Wirkungsrichtung ist aber auszuschließen (da hätten wir in südlichen Ländern ein kleines Problem!).

Die Regression ermöglicht jedenfalls unter bestimmten Umständen[8] gute, bzw. bestmögliche Vorhersage für eine Variable. Folgernd aus dem eben gesagten, sollte nochmals klargestellt werden, dass im Gegensatz zur Korrelation festgelegt werden muss, welche Variable durch eine andere Variable vorhergesagt werden soll.

Definitionen

Die formale Definition eines einfachen linearen Modells ist:

$$y_i = b_0 + b_1 \cdot x_{1i} + \varepsilon_i \tag{10}$$

Die wesentlichen Parameter dieses einfachen Modells sind:

1. Konstanter Term (intercept) b_0: jener Wert den y_i einnimmt, wenn $x_{1i} = 0$ ist.
2. Steigung (slope) b_1: die Zunahme von y_i, wenn x_{1i} sich um eine Einheit erhöht.

Des Weiteren berücksichtigt dieses Modell auch einen Fehler (ε_i). Damit kommt auch ein ganz zentraler Teil bei der Modellbildung zum Ausdruck. Die meisten Modelle definieren sich also aus:

$$\text{wahrer Wert} = \text{Modell} + \text{Fehler} \tag{11}$$

Daraus lässt sich auch folgende Erkenntnis bezüglich des Modells direkt ableiten:

1. Je kleiner die Summe der Fehler sind, desto besser ist das Modell.
2. Je genauer das Modell, desto kleiner wird auch der Fehler sein.

Mit dieser Erkenntnis wird auch klar, dass i.A. ein ganz einfaches Modell (mit einem einzigen Prädiktor) nur zu einer bedingten Reduktion des Fehlers geeignet ist. Wir werden uns im weiteren Verlauf mit erweiterten Modellen beschäftigen, wollen aber zunächst die Eigenschaften des einfachen linearen Modells näher betrachten. Im folgenden Link findet man eine gute Veranschaulichung des einfachen linearen Modells.

[7] wobei die 1 für den ersten (einzigen) Prädiktor und i als Index für die i-te Beobachtung steht.

[8] intervallskaliertes Kriterium, linearer Zusammenhang zw. Kriterium und Prädiktor(en), Zufallsstichprobe, Normalverteilung der Fehler, Homoskedastizität, Unabhängigkeit der Fehler. Details dazu später.

Betrachtet man das Modell isoliert (also ohne Fehlerterm), ist folgende Schreibweise üblich:

$$\hat{y}_i = b_0 + b_1 \cdot x_{1i} \tag{12}$$

Berechnung der Koeffizienten

Für die Berechnung der Koeffizienten wird das Kriterium der kleinste Quadrate (MLS) angewandet. Einfach ausgedrückt wird eine Gerade durch die beobachteten Daten gesucht, die folgenden Eigenschaften aufweist:

1. die Summe der quadratischen Abstände jeder Beobachtung zum entsprechenden Punkt auf der Geraden ist ein Minimum, also $\sum_{i=1}^{N} \varepsilon_i^2 = min.$
2. es gibt keine andere Gerade die eine kleinere Summe dieser Fehler liefert.

Die Berechnung der Parameter entspricht daher einer Extremwertaufgabe, d.h. die partiellen Ableitungen werden auf Null gesetzt. Daraus lassen sich dann die Parameter b_0, b_1 berechnent. Details dazu siehe Wikipedia.

Modellanwendung

Zur Anwendung eines einfachen linearen Modell betrachten wir wiederum die bereits bekannten Daten aus dem Datensatz *CPS85*. Diese Mal wollen wir das Gehalt (*wage*) durch die Ausbildungsdauer (*educ* in Jahren) vorhersagen. Formal lautet das Modell demnach:

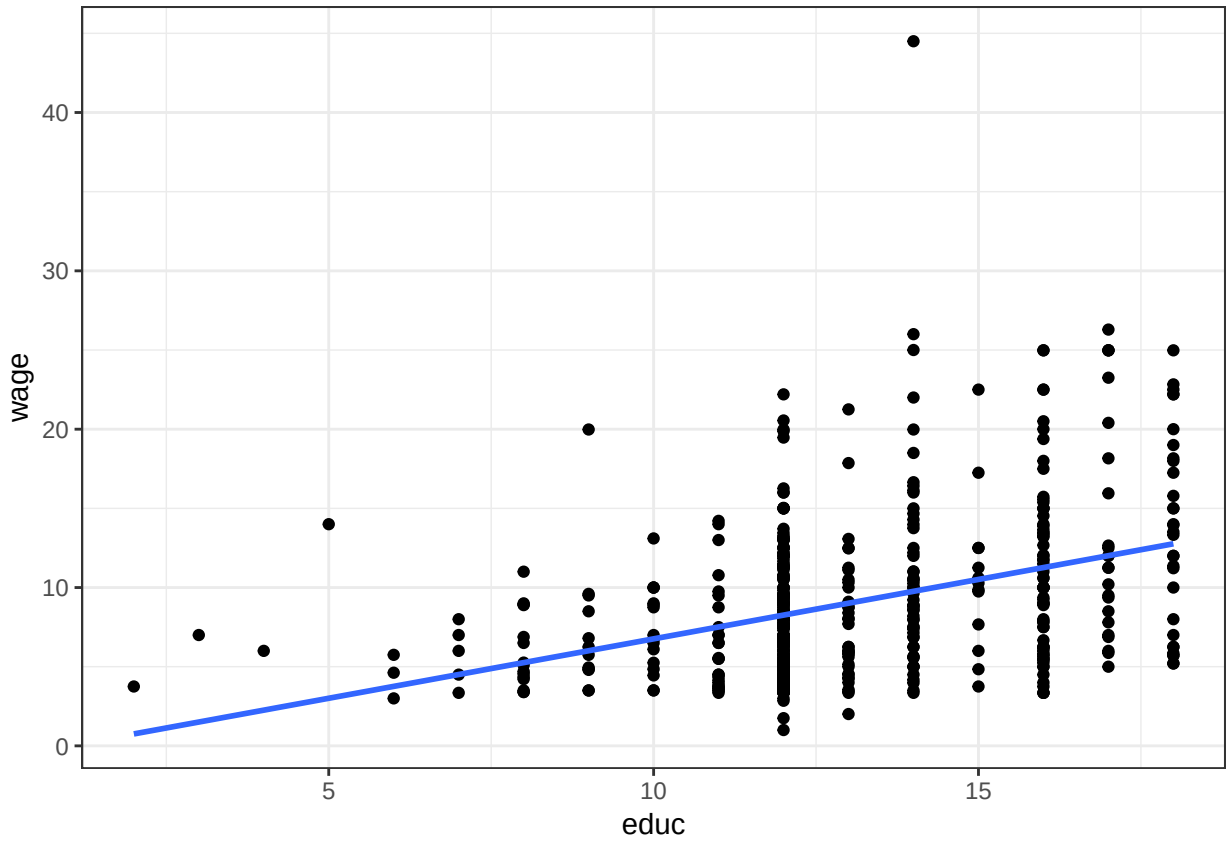
$$\hat{\text{wage}} = b_0 + b_1 \cdot \text{educ} \tag{13}$$

Die Werte der Parameter b_0, b_1 können für dieses Beispiel entsprechend der obigen Erläuterung folgendermaßen interpretiert werden:

1. Für eine Person mit keiner Ausbildung (wage $= x_{1i} = 0$) wird durch das Modell ein Einkommen $y_i = b_0$ vorhergesagt.
2. Erhöht man die Ausbildungsdauer x_{1i} um ein Jahr, steigt der Gehalt y_i um das b_1-fache an.

Kopiere zur Veranschaulichung folgenden Code in dein R-Script und führe diesen aus.

```
DF <- CPS85
ggplot(CPS85, aes(x = educ, y = wage)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE) +
  theme_bw()
```




```
model_1     <- lm(wage ~ educ, data = CPS85)
pander(summary(model_1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.746	1.045	-0.7135	0.4758
educ	0.7505	0.07873	9.532	5.474e-20

Table 2: Fitting linear model: wage ~ educ

Observations	Residual Std. Error	R^2	Adjusted R^2
534	4.754	0.1459	0.1443

Die in der Tabelle angegebenen Werte der Spalte **Estimate** entsprechen dabei den Parametern b_0, b_1 des Modells. Eine weitere wesentliche Kennzahl für die Interpretation des Modells ist der Spalte R^2 zu entnehmen. Dieser Wert wird als *Determinationskoeffizient*[9] bezeichnet. Umgerechnet in % (im vorliegenden Beispiel also 14.43%) besagt der Wert, wie viel der Variablität des Gehaltes durch den Prädiktor Ausbilung erklärt wird. Wir werden im weiteren Verlauf noch öfter auf diesen Kennwert zurückkommen.

Welche Gehälter würden für Ausbildungszeiten zwischen 10 und 14 Jahren vorhergesagt werden? Kopiere folgenden Code in dein R-Script und führe diesen aus. Änder auch den Wertebereich der Prädiktoren und beobachte was dabei passiert!

[9] häufig auch als Varianzaufklärung

Im Vergleich zum Mittelwert-Modell zeigt sich mit steigender Ausbildung ein höheres Einkommen. Der Fehler bei der Vorhersage des Einkommens wird sich daher durch diese Modellvorstellung verringern (mehr zur Abschätzung der Fehlerreduktion später).

Aufgabe SLR 1

Öffne ein neues R-Script und kopiere die bereits bekannte Kopfzeile in diese Datei. Speichere anschließend das Skript unter dem Namen *SLR_Aufgabe1.R*. Bearbeite nun folgende Aufgabenstellungen:

- Lade die Datei "*Album Sales 1.dat*"
- erstelle ein einfaches Streudiagramm mit Sales auf der x- und adverts auf der y-Achse.
- erstelle ein einfaches lineares Modell zur Vorhersage der Verkaufszahlen (*sales*) durch die Variable *adverts*.
- Wie stark korrelliert der Prädiktor mit dem Kriterium?
- Wie viel Varianz wird vom Kriterium durch den Prädiktor aufgeklärt?
- Ist das erstellte Modell signifikant besser, als dan Null-Modell?
- Welchen Wert würde das Modell für Werbeausgaben = 100 vorhersagen?

Lösung Aufgabe SLR 1

Residualanalyse

Ein zentrales Thema der Modellbildung ist die Beurteilung und (statistische) Auswertung der Abweichungen des Modells von den Beobachtungen (Fehler, Residum). Folgende Kennwerte bilden die Möglichkeit, die Güte des Modells abzuschätzen:

1. *Vorhergesagte Werte*: vorhergesagte Werte der Regressionsgleichung (= Werte die auf der Geraden liegen).
   - Nicht standardisiert: der Wert, den das Modell für die abhängige Variable vorhersagt.
   - Standardisiert: z-Transformierte vorhergesagte Werte.
   - Korrigiert: der vorhergesagte Wert für einen Fall, wenn dieser Fall von der Berechnung der Regressionskoeffizienten ausgeschlossen ist.
   - Standardfehler des Mittelwerts: Standardfehler der vorhergesagten Werte. Ein Schätzwert der Standardabweichung des Durchschnittswertes der abhängigen Variablen für die Fälle, die dieselben Werte für die unabhängigen Variablen haben.
2. *Residuen*: tatsächliche Wert der abhängigen Variablen minus des vorhergesagten Werts aus der Regressionsgleichung.
   - Nicht standardisiert: Die Differenz zwischen einem beobachteten Wert und dem durch das Modell vorhergesagten Wert.
   - Standardisiert: Der Quotient aus dem Residuum und einer Schätzung seiner Standardabweichung. Standardisierte Residuen, auch bekannt als Pearson-Residuen, haben einen Mittelwert von 0 und eine Standardabweichung von 1.
   - Studentisiert: Ein Residuum, das durch seine geschätzte Standardabweichung geteilt wird, die je nach der Distanz zwischen den Werten der unabhängigen Variablen des Falles und dem Mittelwert der unabhängigen Variablen von Fall zu Fall variiert.
   - Ausgeschlossen: Das Residuum für einen Fall, wenn dieser Fall nicht in die Berechnung der Regressionskoeffizienten eingegangen ist. Dies ist die Differenz zwischen dem Wert der abhängigen Variablen und dem korrigierten Schätzwert.
   - Studentisiert und ausgeschlossen: Der Quotient aus dem ausgeschlossenen Residuum eines Falles und seinem Standardfehler. Die Differenz zwischen einem studentisierten ausgeschlossenen Residuum und dem zugehörigen studentisierten Residuum gibt an, welchen Unterschied die Entfernung eines Falles für dessen eigene Vorhersage bewirkt.
3. *Distanzen*: Maße zum Auffinden von Fällen mit ungewöhnlichen Wertekombinationen bei den unabhängigen Variablen und von Fällen, die einen großen Einfluss auf das Modell haben könnten.
   - Mahalanobis: Dieses Maß gibt an, wie weit die Werte der unabhängigen Variablen eines Falls vom Mittelwert aller Fälle abweichen. Eine große Mahalanobis-Distanz charakterisiert einen Fall, der

bei einer oder mehreren unabhängigen Variablen Extremwerte besitzt.
   - Cook: Ein Maß dafür, wie stark sich die Residuen aller Fälle ändern würden, wenn ein spezieller Fall von der Berechnung der Regressionskoeffizienten ausgeschlossen würde. Ein großer Wert der Cook-Distanz zeigt an, dass der Ausschluss eines Falles von der Berechnung der Regressionskoeffizienten die Koeffizienten substanziell verändert.
   - Hebelwerte: Werte, die den Einfluss eines Punktes auf die Anpassung der Regression messen. Der zentrierte Wert für die Hebelwirkung bewegt sich zwischen 0 (kein Einfluss auf die Anpassung) und $(N-1)/N$.
4. *Vorhersageintervalle*: obere und untere Grenzen sowohl für Mittelwert als auch für einzelne Vorhersageintervalle.
   - Mittelwert: Unter- und Obergrenze (zwei Variablen) für das Vorhersageintervall für den mittleren vorhergesagten Wert.
   - Individuell: Unter- und Obergrenzen (zwei Variablen) für das Vorhersageintervall der abhängigen Variablen für einen Einzelfall.
   - Konfidenzintervall: Geben Sie einen Wert zwischen 1 und 99,99 ein, um das Konfidenzniveau für die beiden Vorhersageintervalle festzulegen. Wählen Sie *Mittelwert* oder *Individuell* aus, bevor Sie diesen Wert eingeben. Typische Werte für Konfidenzniveaus sind 90, 95 und 99.
5. *Einflussstatistiken*: Änderung in den Regressionskoeffizienten (DfBeta(s)) und vorhergesagten Werten (DfFit), die sich aus dem Ausschluss eines bestimmten Falls ergeben.
   - Differenz in Beta: entspricht der Änderung im Regressionskoeffizienten, die sich aus dem Ausschluss eines bestimmten Falls ergibt. Für jeden Term im Modell, einschließlich der Konstanten, wird ein Wert berechnet.
   - Standardisiertes DfBeta: die Änderung des Regressionskoeffizienten, die sich durch den Ausschluss eines bestimmten Falls ergibt. Es empfiehlt sich, Fälle mit absoluten Werten größer als $2/\sqrt{N}$ zu überprüfen, wenn N die Anzahl der Fälle darstellt. Für jeden Term im Modell, einschließlich der Konstanten, wird ein Wert berechnet.
   - DfFit: Differenz im Anpassungswert ist die Änderung im vorhergesagten Wert, die sich aus dem Ausschluss eines bestimmten Falls ergibt.
   - Standardisiertes DfFit: Änderung des vorhergesagten Werts, die sich durch den Ausschluss eines bestimmten Falls ergibt. Es empfiehlt sich, Fälle mit absoluten Werten $> 2/\sqrt{p/N}$ zu überprüfen, wobei p die Anzahl der unabhängigen Variablen im Modell und N die Anzahl der Fälle darstellt.
   - Kovarianzverhältnis: Verhältnis der Determinante der Kovarianzmatrix bei Ausschluss eines bestimmten Falls von der Berechnung der Regressionskoeffizienten zur Determinante der Kovarianzmatrix bei Einschluss aller Fälle. Wenn der Quotient dicht bei 1 liegt, beeinflusst der ausgeschlossene Fall die Kovarianzmatrix nur unwesentlich.

Nachfolgender Code und Tabelle zeigen die Auswertung der Residualanalyse für das oben erstellte *model_1*:

```
CPS85_Res <- data.frame(Res     = round(resid(model_1), 2),
                        StdRes  = round(rstandard(model_1), 2),
                        StudRes = round(rstudent(model_1), 2),
                        # Cook    = round(cooks.distance(model_1), 2),
                        # DFBeta  = round(dfbeta(model_1), 2),
                        DF5Fit  = round(dffits(model_1), 2),
                        # Lev     = round(hatvalues(model_1), 2),
                        CovRat  = round(covratio(model_1), 2))
pander(head(CPS85_Res))
```

Res	StdRes	StudRes	DF5Fit	CovRat
2.24	0.47	0.47	0.03	1.01
-2.76	-0.58	-0.58	-0.03	1
-4.46	-0.94	-0.94	-0.04	1
2.24	0.47	0.47	0.02	1.01
6.74	1.42	1.42	0.07	1
-2.26	-0.48	-0.48	-0.03	1.01

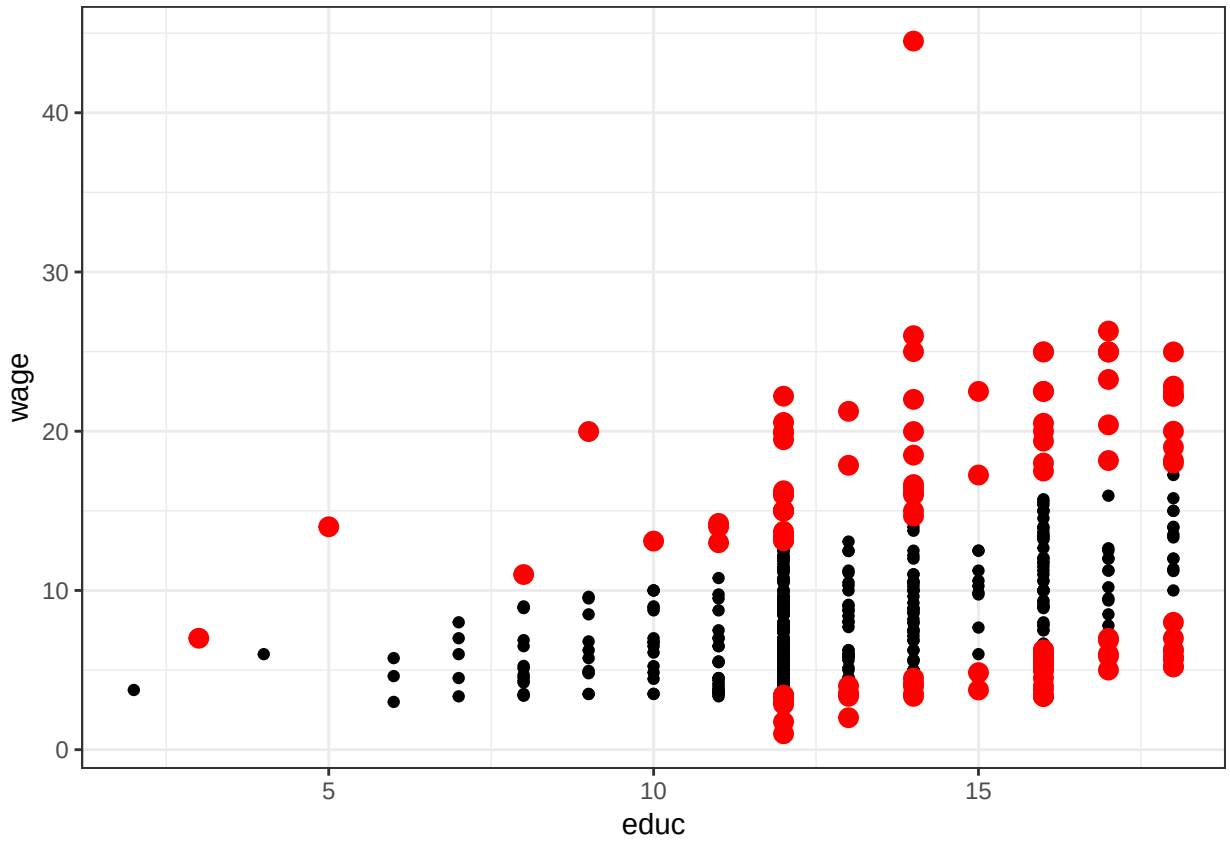
Res	StdRes	StudRes	DF5Fit	CovRat

Mit der Residualanalyse kann man auf relativ einfache Weise jene Werte ermitteln (und auch graphisch darstellen), die z.B. um mehr als eine Standardabweichung abweichen. Diese Werte könnte man nochmals genauer untersuchen und gegebenenfalls vor einer weiterführenden Analyse ausschließen[10]. Keinesfalls sollte sie jedoch dazu verwendet werden, um einen erwünschten Effekt durch schrittweises löschen störender Daten zu erreichen! Kopiere folgenden Code in dein R-Script und führe diesen aus.

```
# Liste standardisierte Residuen > |1|
Ind_Res <- which((CPS85_Res$StdRes > 1 | CPS85_Res$StdRes < -1) == TRUE)
# Anzeige der Werte von wage und educ sowie der Standardisierten Residuen
# fÃ¼r jene FÃ¤lle, deren Residuen Ã¼ber 1 SD abweichen.

# pander(data.frame(Indizes = Ind_Res,
#                   wage = CPS85$wage[Ind_Res],
#                   educ = CPS85$educ[Ind_Res] ,
#                   CPS85_Res$StdRes[Ind_Res]))

p_Res1 <- ggplot(CPS85, aes(x = educ, y = wage)) +
          geom_point() +
          geom_point(data=CPS85[Ind_Res,],colour="red",size=3) +
          theme_bw()
print(p_Res1, comment = FALSE)
```



[10]der Ausschluss von Werten ist nur dann erlaubt, wenn eine entsprechende Begründung (nachvollziehbarer Messfehler, falsche Datenübertragung, etc.) vorliegt!

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:

$$y_i = b_0 + b_1 \cdot x_{1i} + \cdots + b_k \cdot x_{ki} + \varepsilon_i \tag{14}$$

Die wesentlichen Parameter dieses Modells sind:

1. Intercept b_0: jener Wert den y_i einnimmt, wenn $x_{ji} = 0$ ist (mit $j \epsilon [1, k]$).
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 $m[1, k]$ und $m \neq j$)!

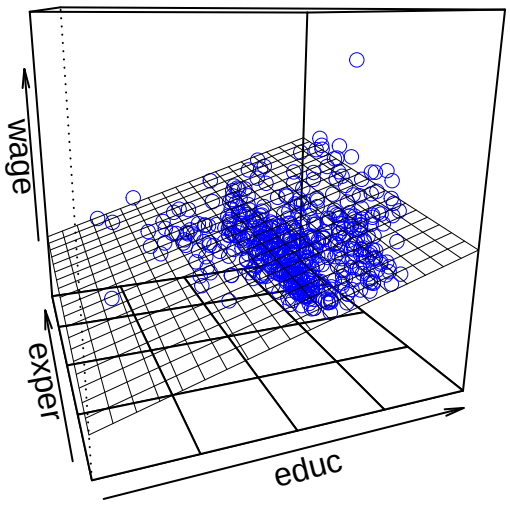
Des Weiteren berücksichtigt auch dieses Modell wieder einen Fehler (ε_i). Betrachtet man das multiple Modell isoliert (also ohne Fehlerterm), ist folgende Schreibweise üblich:

$$\hat{y}_i = b_0 + b_1 \cdot x_{1i} + \cdots + b_k \cdot x_{ki} \tag{15}$$

Betrachten wir an unseren Beispieldaten folgendes Modell mit zwei Prädiktoren:

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

```
model_2     <- lm(wage ~ educ + exper, data = CPS85)
Det_model_2 <- pander(summary(model_2))
plotPlane(model = model_2, plotx1 = "educ", plotx2 = "exper")
```



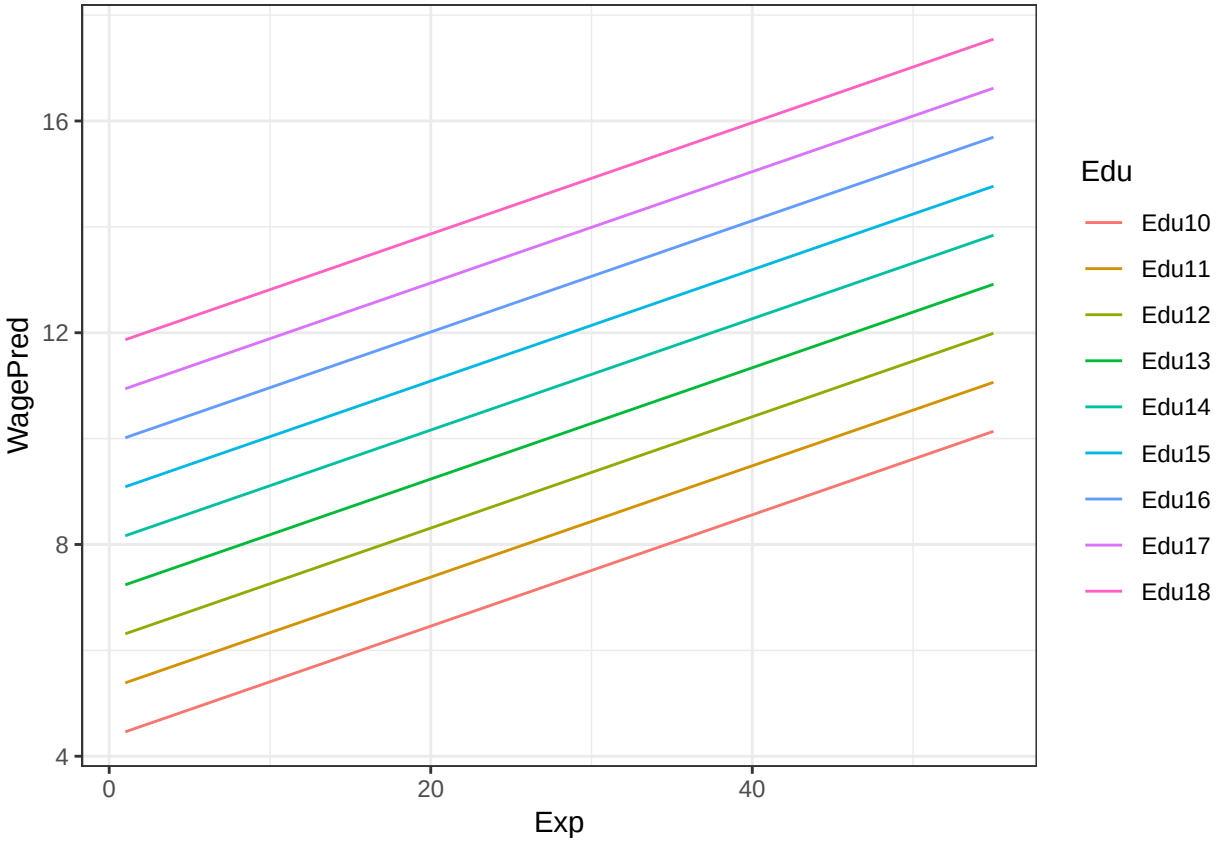
Dabei entspricht der Koeffizient b_2 der Zunahme des Gehaltes $\hat{y}_i$ wenn sich die Erfahrung x_{2i} um eine Einheit erhöht und die Ausbildung x_{1i} konstant gehalten wird. In nachfolgernder Tabelle sind die Werte der Vorhersagen des Modells für den vorliegenden Datensatz auszugsweise dargestellt:

```
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(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

```
CPS852Disp           <- 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 Ã„nderungen model_1 vs Mittelwertsmodell
# Berechnung der Quadratsummen fÃ¼r 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(anova(model_1))
```

Table 5: 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_{Regression} = $ 2053.29, welcher ident mit dem Wert der ANOVA-Tabelle ist.

Die restlichen Kennwerte stimmen auch mit dem Ergebnis der ANOVA überein ($MSS_{Residual} = 22.6$, 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 Ã„nderungen model_1 vs model_2 (Ã„nderung signifikant?)
pander(anova(model_1, model_2))
```

Table 6: 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 Korrealtionskoeffizient ($r_{xy \cdot z}$): 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: $sr^2_{k \cdot 12 \cdots (k-1)} = R^2_{y,12 \cdots k} - R^2_{y,12 \cdots k-1}$

```
# Korrelationen, Paritial- und Semipartialkorrelationen
Korr_Data      <- data.frame(wage = M$wage, educ = M$educ, exper = M$exper)
PearsonKorr    <- cor(Korr_Data)
ModVgl_Korr    <- pander(PearsonKorr)
R2Change_mod_1 <- PearsonKorr[2]^2
# Partial Korrelation zwischen "wage" und "educ" gegeben "exper"
PartKorr_1       <- pcor.test(Korr_Data$wage, Korr_Data$educ, Korr_Data$exper)
ModVgl_ParKorr_1 <- pander(PartKorr_1)
# Partial Korrelation zwischen "wage" und "exper" gegeben "educ"
PartKorr_2       <- pcor.test(Korr_Data$wage, Korr_Data$exper, Korr_Data$educ)
ModVgl_ParKorr_2 <- pander(PartKorr_2)
# Semi-Partial (part) Korrelation zwischen "wage" und "educ" gegeben "exper"
SemiPartKorr_1      <- spcor.test(Korr_Data$wage, Korr_Data$educ, Korr_Data$exper)
ModVgl_SemParKorr_1 <- pander(SemiPartKorr_1)
# Semi-Partial (part) Korrelation zwischen "wage" und "exper" gegeben "edu"
SemiPartKorr_2      <- spcor.test(Korr_Data$wage, Korr_Data$exper, Korr_Data$educ)
ModVgl_SemParKorr_1 <- pander(SemiPartKorr_2)
R2Change_mod_2      <- round(SemiPartKorr_2$estimate^2,3)
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
exper	0.1051	0.0172	6.113	1.893e-09

Table 8: 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 hohe 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 interpretiern 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) abzuwägen.

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.

$$AIC_k = 2 \cdot |k| - 2 \cdot \hat{L}_k \tag{16}$$

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*).

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.

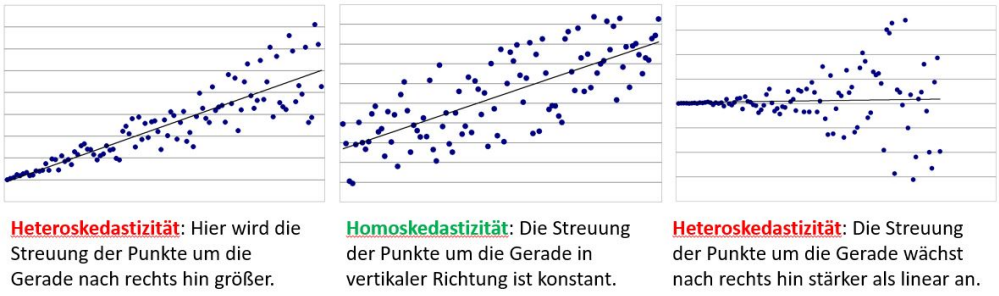


Figure 6: **Abbildung 3**: Homoskedastizität vs. Heteroskedastizität

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 Model (*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 $j = j$-tes Modell und $i = i$'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 $1, 2, \cdots, k$ bezeichnet):

1. **Lineare Beziehung** zwischen den Variablen (keine Ausreißer): eine einfache Prüfung erfolgt visuell mit Streudiagrammen, wobei alle Beziehungen, also $r_{y\cdot 1}, r_{y\cdot 2}, \cdots, r_{y\cdot k}, \cdots, r_{1\cdot 2}, r_{1\cdot k}, \cdots, r_{(k-1)\cdot k}$ zu betrachten sind!
2. **Varianzgleichheit der Residuen** (Homoskädasdizität): auch diese Vorausstung 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 doe standardisierten Residuen aufgetragen werden. Heteroskedastizität liegt vor, wenn die Punktewolke 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. $r_{1 \cdot 2} > 0.8$). 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 ($r_{y \cdot k} \approx 0$), 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 Validitäten[11] 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 $r_{1 \cdot 2} < 0$. 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. **Unabgängigkeit 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!

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

[11] Die 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.

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

```
  album1       <- read.delim("Daten/Album Sales 1.dat", header = TRUE)
  ggplot(album1, aes(x = adverts, y = sales)) +
    geom_point() +
    geom_smooth(method=lm, se=FALSE) +
    theme_bw()
  albumSales.1 <- lm(sales ~ adverts, data = album1)
  pander(summary(albumSales.1))

# #---- Modell1_Pred_1
#
#   new_input <- data.frame(educ = 10:14)
#   pander(predict(model_1, newdata = new_input), style = "rmarkdown")
```

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(summary(albumSales.2))
  pander(summary(albumSales.3))
  # Modellvergleich
  anova(albumSales.2, albumSales.3)
  Tol <- 1/vif(albumSales.3)
  VIF <- vif(albumSales.3)
```

zurück zur Aufgabenstellung

Box, G.E.P. 1979. "Robustness in the Strategy of Scientific Model Building." *Academic Press*.

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.

Knuth, D. 2008. "Ein Modell Des Modellseins – Ein Beitrag Zur Aufklärung Des Modellbegriffs." *Peter Lang Verlag*, 187–220.

Mahr, Bernd. 2008. "Ein Modell Des Modellseins – Ein Beitrag Zur Aufklärung Des Modellbegriffs." *Peter*

Lang Verlag, 187–220.

Upmeier, D., A. und Krüger. 2010. "Modellkompetenz Im Biologieunterricht." *Zeitschrift Für Didaktik Der Naturwissenschaften*.

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

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