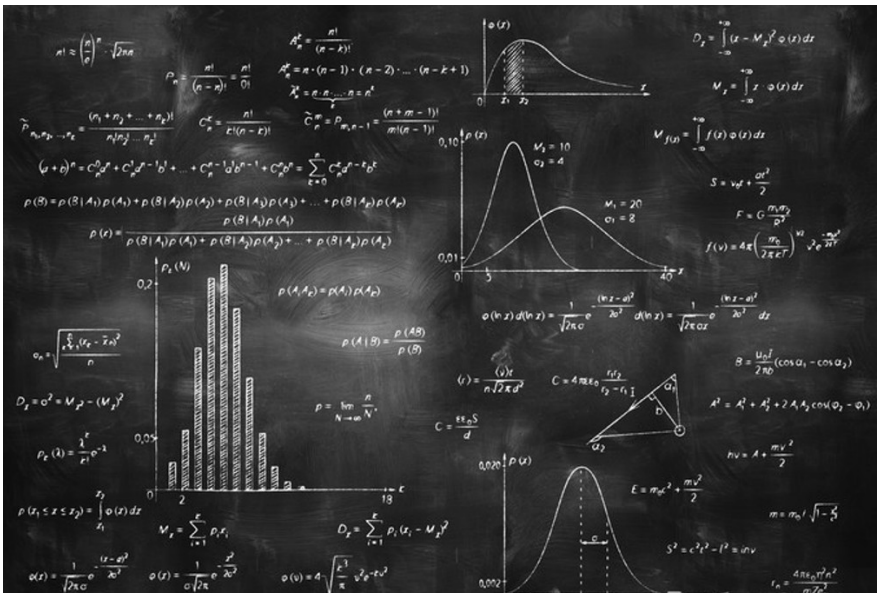
# Lineare Regression

Walter Gruber

2025-03-17

## Contents



# Einfache Lineare Regression

Dieses Skript basiert (großteils) auf der Literatur von Andy Field und Rand Wilcox (&. W. Field A. P. 2017), David Erceg-Hurn et.al. (Hurn 2008), Mair (Mair 2020) and Wilcox (Wilcox 2012). Teile des Inhaltes wurden direkt aus der genannten Literatur übernommen.

## Vorwort

In einer Welt, die zunehmend von Daten geprägt ist, sind statistische Methoden unverzichtbare Werkzeuge, um Muster aufzudecken und fundierte Entscheidungen zu treffen. Die einfache lineare Regression ist eine der grundlegendsten, aber zugleich wirkungsvollsten Techniken in der Statistik. Sie ermöglicht es, den Zusammenhang zwischen zwei quantitativen Variablen zu modellieren und vorherzusagen, wie sich Änderungen in einer unabhängigen Variable auf eine abhängige Variable auswirken.

In diesem Kapitel werden wir die Grundlagen der einfachen linearen Regression erkunden:

- Was bedeutet es, einen linearen Zusammenhang zwischen zwei Variablen zu postulieren?
- Wie wird ein lineares Regressionsmodell aufgestellt und interpretiert?
- Durch anschauliche Beispiele und Schritt-für-Schritt-Anleitungen werden wir die Schlüsselkonzepte und mathematischen Grundlagen dieser Methode erläutern.

## Ziele

Ziel ist es, Ihnen ein solides Verständnis für die einfache lineare Regression zu vermitteln, das als Basis für komplexere statistische Analysen dient.

- Regressionsmodelle verstehen und anwenden.
- Eigenschaften von Quadratsummen verstehen:
  - Totale Quadratsumme.
  - Fehler- oder Residualquadratsumme.
  - Modell Quadratsumme.
- Prüf- und Kenngrößen verstehen und interpretieren:
  - F-Werte.
  - Teststatistiken wie t-Werte.

  - Konfidenzintervalle.
  - b-Gewichte und standardisierte Gewichte ($\beta$).
  - $R^2$ (Determinationskoeffizient, Bestimmtheitsmaß).
- Voraussetzungen und deren Überprüfung bei LM.
- Regression verstehen und anwenden können.
- Interpretation von Ergebnissen und APA-konforme Berichterstattung.
- Problembereiche der einfachen Regression verstehen und erkennen können.

# Das einfache lineare Modell

Formal wird ein einfaches lineares Regressionsmodell (***ELR***) definiert durch:

$$y_i = b_0 + b_1 \cdot x_i + \varepsilon_i$$

mit:

- $y_i$: Kriterium, abhängige Variable - also der beobachtete Wert des Kriteriums der $i$-ten Person.
- $b_0$: konstanter Term, intercept (Wert für $y_i$ wenn $x_i = 0$).
- $b_1$: Steigung, regression coefficient, gradient, slope
- $x_i$: Prädiktor, unabhängige Variable: also der beobachtete Wert des Prädiktors der $i$-ten Person.
- $\varepsilon_i$: Fehler, error term
- $i$: Index für betrachteten Fall

Eine sehr bedeutende Rolle in dieser Gleichung nehmen die sogenannten **Regressionskoeffizienten** $b_0$ und $b_1$ ein. Man nennt diese auch die **Parameter** (= bestimmenden Elemente) der Gleichung. Sind diese Werte bekannt, kann man für jedes beliebige $x_i \in \mathbb{R}$ einen entsprechenden $y_i$-Wert auf der Gerade bestimmen. Damit ist also durch diese Parameter die Lage und Steigung der Gerade eindeutig bestimmt!

Will man formal jene Werte beschreiben, die durch das Modell für einen bestimmten Prädiktorwert $x_i$ vorhergesagt werden, schreibt man:

$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

Aus diesen beiden Darstellungsformen lässt sich eine wichtige Eigenschaft ableiten. Der beobachtete Wert des Kriteriums $y_i$ wird in den allermeisten Fällen nicht mit dem vom linearen Modell vorhergesagten Wert $\hat{y}_i$ übereinstimmen. Daher gilt:

$$\varepsilon_i = y_i - \hat{y}_i$$

Man bezeichnet das $\varepsilon_i$ auch als **Residuum**, oder **Fehler**, oder **Fehlerterm**. Das $\varepsilon_i$ ist also der Abstand von einem beobachteten Wert zu einem vom Modell vorhergesagten Wert an der Stelle $x_i$.

## Fehlerterm

Über den Fehler $\varepsilon_i$ kann man auch die Modellannahme für die optimale Lage der Geraden innerhalb der beobachteten Wertepaare $(x_i, y_i)$ bestimmen. Eine simple und relativ leicht zu berechnende Annahme für den Fehlerterm lautet:

> Die Summe der quadratischen Abweichungen sollte ein Minimum sein!

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \text{Minimum}$$

Setzt man für $\hat{y}_i$ obige Gleichung ein, erhält man:

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2 \rightarrow \text{Minimum}$$

Durch partielle Ableitung und Null-setzen von $\frac{\partial f(b_0,b_1)}{\partial b_0} = 0$ und $\frac{\partial f(b_0,b_1)}{\partial b_1} = 0$ lassen sich die Koeffizienten bestimmen.

> Diese Methode/Annahme ist unter den Namen **O**rdinary **L**east **S**quare, bzw. **MKQ** für **M**ethode der **K**leinsten **Q**uadrate bekannt.

Für das einfache lineare Modell ergibt sich für $b_0$:

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

und für $b_1$:

$$b_1 = \frac{n \cdot \sum x_i \cdot y_i - \sum x_i \cdot \sum y_i}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

Um die Residuen und die Anpassung einer Geraden an die Daten besser zu verstehen, eignet sich folgende **Visualisierung einer Regression** recht gut.

# Modellvorhersage & Interpretation

Das $b_0$ wird auch konstanter Term, oder Interzept genannt. Es ist jener $y$-Wert, der durch das Modell für die Stelle $x = 0$ vorhergesagt wird. Damit ist es auch jener Punkt auf der $y$-Achse, durch welche die Gerade geht.

Inhaltlich entspricht dieser Wert dem vorhergesagtem Wert des Kriteriums $\hat{y}$ an dem der Prädiktor $x$ den Wert Null hat. Für unser Beispiel also jener Lernerfolg, der durch 0 Stunden Lernen zu erwarten wäre.

Die Steigung $b_1$ gibt die erwartete Veränderung des Kriteriums $\hat{y}$ an, die einer Erhöhung des Prädiktors $x$ um einen Einheit entspricht.

## Beispiel

Der Lernerfolg $\hat{y}$ soll durch die Anzahl der Stunden für die Vorbereitung zur Klausur $x$ vorhergesagt werden. Sei $b_0 = 10$ und $b_1 = 3$. Die Regressionsgleichung lautet daher:

$$\hat{y} = 10 + 3 \cdot x$$

Hat jemand $x = 0$ Stunden gelernt, wird anhand des Modells ein Lernerfolg von 10 vorhergesagt. Bei einem Lernaufwand von $x = 1$ Stunde, würde das Modell einen Lernerfolg von $\hat{y} = 13$, also eine um 3 Einheiten besseren Lernerfolg vorhersagen.

In vielen Fällen ist jedoch eine Vorhersage eines $\hat{y}$-Werte bei einem $x = 0$ nicht sinnvoll. Würde man z.B. den Lernerfolg mit Intelligenz vorhersagen und wären z.B. die aus beobachteten Daten ermittelten Koeffizienten $b_0 = -9$ und $b_1 = 0.2$, dann würde der Lernerfolg bei einer Intelligenz von $x = 0$ dem Wert $\hat{y} = -9$ vorhersagen. Das wäre aber offensichtlich eine unbrauchbare Vorhersage, da es einerseits keinen negativen Lernerfolg gibt und auf einer herkömmlichen IQ-Skala der Wert $x = 0$ auch keine Bedeutung hat.

## Korrelation/Regression

Zusammenhang Korrelation und Regression:

$$b_1 = r(x, y) \cdot \frac{s_y}{s_x}$$

Daraus leiten sich folgende Erkenntnisse ab:

- Die Steigung ist positiv (negativ), wenn die Korrelation positiv (negativ) ist.
- Ist die Korrelation null, dann ist auch die Steigung null (Gerade ist parallel zur $x$-Achse)
- $b$ ist abhängig von $s_x$ und $s_y$. Daher führt eine Änderung der Messeinheit in einer der beiden Variablen auch zu einer Änderung der Steigung!
- Aus voriger Folgerung kann man erkennen, dass $b$ kein standardisiertes Maß für den Einfluß von $x$ auf $y$
- $b$ ist direkt proportional zu $r$

## Eigenschaften der Gerade

Folgende Eigenschaften der Regressionsgeraden sind bemerkenswert:

- eine zur $x$-Achse parallele Gerade bedeutet, dass kein Zusammenhang zwischen den beiden Variablen besteht. Für jeden beliebigen Wert des Prädiktors $x$ wird stets der gleiche $y$-Wert (Kriterium) vorhergesagt.
- eine nach rechts oben steigend bedeutet, dass mit Zunahme der Werte des Prädiktors auch die Werte des Kriteriums steigen und damit eine Abhängigkeit der beiden Variablen gegeben ist (positiver Zusammenhang).
- eine nach rechts fallende bedeutet, dass mit Zunahme der Werte des Prädiktors die Werte des Kriteriums fallen und damit ebenfalls eine Abhängigkeit der beiden Variablen gegeben ist (negativer Zusammenhang).
- die Regressionsgerade wurde aus den Daten einer Stichprobe berechnet.
  - Die Parameter der Regressionsgerade $b_0, b_1, s_b$ sind Schätzwerte der in der Population vorhandenen Parameter $\beta_0, \beta_1, \sigma_b$.

## Beta-Gewichte

Werden die in die Regression eingehenden Variablen (Kriterium und Prädiktorvariable(n)) vor Berechnung der Koeffizienten z-transformiert, erhält man standardisierte Koeffizienten. Diese Koeffizienten werden allegemein auch als **Beta-Gewichte** (**B_1**, oder $\beta_1$) bezeichnet. Es gilt:

$$\hat{z}_y = B_0 + B_1 \cdot z_x$$

Auch für das $B_1$ besteht ein direkt proportionaler Zusammenhang der Steigung zur Korrelation:

$$B_1 = r(z_x, z_y) \cdot \frac{s_{z_y}}{s_{z_x}} = r(x, y) \cdot \frac{1}{1} = r$$

Wie halten folgende wichtige Zusammenhänge zwischen $r(x, y)$ und $B_j$ fest:

- Die Korrelation $r(x, y)$ und der standardisierte Steigungskoeffizient $B$ sind identisch.[1]
- Beide Maße ($r(x, y)$ und $B$) sind unabhängig von der Messeinheit der beiden Variablen.
- $B$ und damit auch $r(x, y)$ zeigen den Effekt, den die Änderung des Prädiktors um eine Standardabweichung auf das z-transformierte Kriterium hat.
- $r$ ist somit eine Maßzahl, die bei Regressionsmodellen die Effektstärke widerspiegelt.

[1] WICHTIG: dies gilt prinzipiell für ELR, aber unter der Bedingung, dass alle verwendeten Prädiktoren voneinander unabhängig sind (also nicht korrelieren) auch für die Beta-Gewichte der multiplen Regression!}

- Nach Cohen gelten folgende Richtwerte: $r = 0.1$ entspricht einem kleinen, $r = 0.3$ entspricht einem mittleren und $r = 0.5$ entspricht einem starken Effekt.
- der $\beta$-Koeffizient gibt an, wie viele Standardabweichungen sich die abhängige Variable ändert, wenn sich die unabhängige Variable um eine Standardabweichung ändert. Ein $\beta$-Wert von 0.5 bedeutet beispielsweise, dass eine Erhöhung der unabhängigen Variable um eine Standardabweichung zu einem Anstieg der abhängigen Variable um 0.5 Standardabweichungen führt.
- da die Koeffizienten standardisiert sind, kann das $\beta$ über verschiedene Modelle hinweg direkt verglichen werden, was hilfreich ist, wenn man wissen möchte, welcher Prädiktor den stärksten Einfluss hat (insbesondere in einer multiplen Regressionsanalyse).
- während der $\beta$-Koeffizient die Stärke und Richtung der Beziehung zwischen zwei Variablen beschreibt, impliziert er *keine Kausalität*.

## Residuen

Eigenschaften von Residuen:

- Die $\varepsilon_i$ enthalten Anteile der Kriteriumsvariablen $y$, die durch die Prädiktorvariable $x$ nicht erfasst/erklärt werden.
- Diese Anteile bestehen aus
  - Messfehler
  - Anteile, die evtl. durch weitere Variablen erklärt werden, die aber mit dem verwendeten Prädiktor nichts zu tun haben, also mit $r(x, \varepsilon) = 0$ korrelieren.

Für den sogenannten **Standardfehler der Residuen** gilt:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{1}{n-2} \cdot \sum_{i=1}^{n} e_i^2}$$

## Standardschätzfehler

Der **Standardschätzfehler*}** ($s_e$) (engl. standard error of estimate) bei einer Regression kennzeichnet die Streuung der $y$-Werte um die Regressionsgerade und ist damit ein Gütemaßstab für die Genauigkeit der Regressionsvorhersagen. Je kleiner dieser Fehler, desto besser die Regressionsvorhersage.

Dieser Fehler ist ein eigenes Modellgütemaß, wird aber weniger häufig als das Bestimmtheitsmaß, bzw. als Determinationskoeffizient angegeben, obwohl der Standardfehler der Residuen bei der Bewertung Anpassungsgüte möglicherweise aussagekräftiger ist (siehe Bemerkungen beim Bestimmtheitsmaß).

*Bemerkung:*

Mithilfe des Standardfehlers können Konfidenzintervalle konstruiert werden.

## Determinationskoeffizient

Das Bestimmtheitsmaß, auch als $R^2$ bekannt, quantifiziert, wie gut die unabhängigen Variablen in einem linearen Regressionsmodell die Streuung der abhängigen Variable erklären, indem es den Anteil der Gesamtvarianz darstellt, der durch das Modell erklärt wird[2].

Es ist wichtig für die Bewertung der Modellanpassung, da ein höherer $R^2$-Wert auf eine bessere Erklärungsfähigkeit des Modells hinweist. Bei der Verwendung ist jedoch unbedingt auch auf die nachfolgend diskutierten Einschränkungen der (sinnvollen) Interpretierbarkeit diese Maßes zu achten!

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

[2]Wikipedia

in Worten:

$$R^2 = \frac{\text{erklärte Variation}}{\text{gesamte Variation}} = 1 - \frac{\text{unerklärte Variation}}{\text{gesamte Variation}}$$

**Wertebereich**: das Maß nimmt den Wert 1 an, wenn $\sum(y_i - \hat{y}_i)^2 = 0$, oder $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$ ist, bzw. den Wert 0, wenn $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 0$!

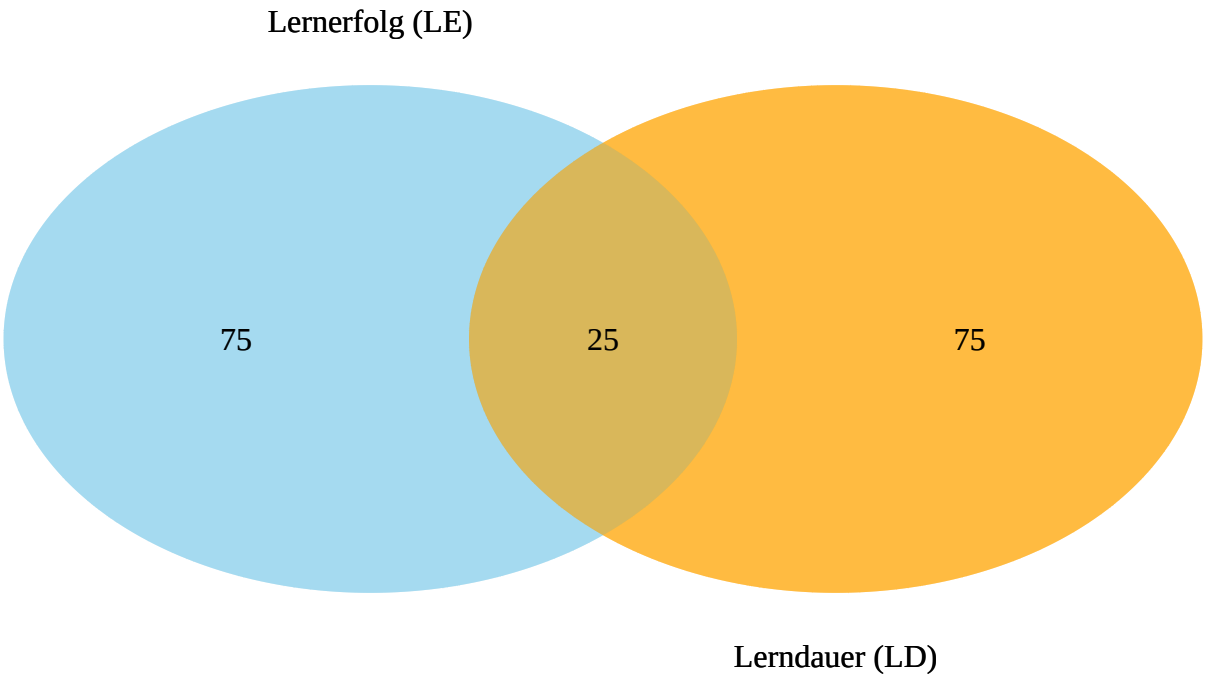
Wenn das Regressionsmodell kein Absolutglied enthält (es liegt ein homogenes Regressionsmodell vor), kann das Bestimmtheitsmaß negativ werden.[3]

## Venn Diagramm

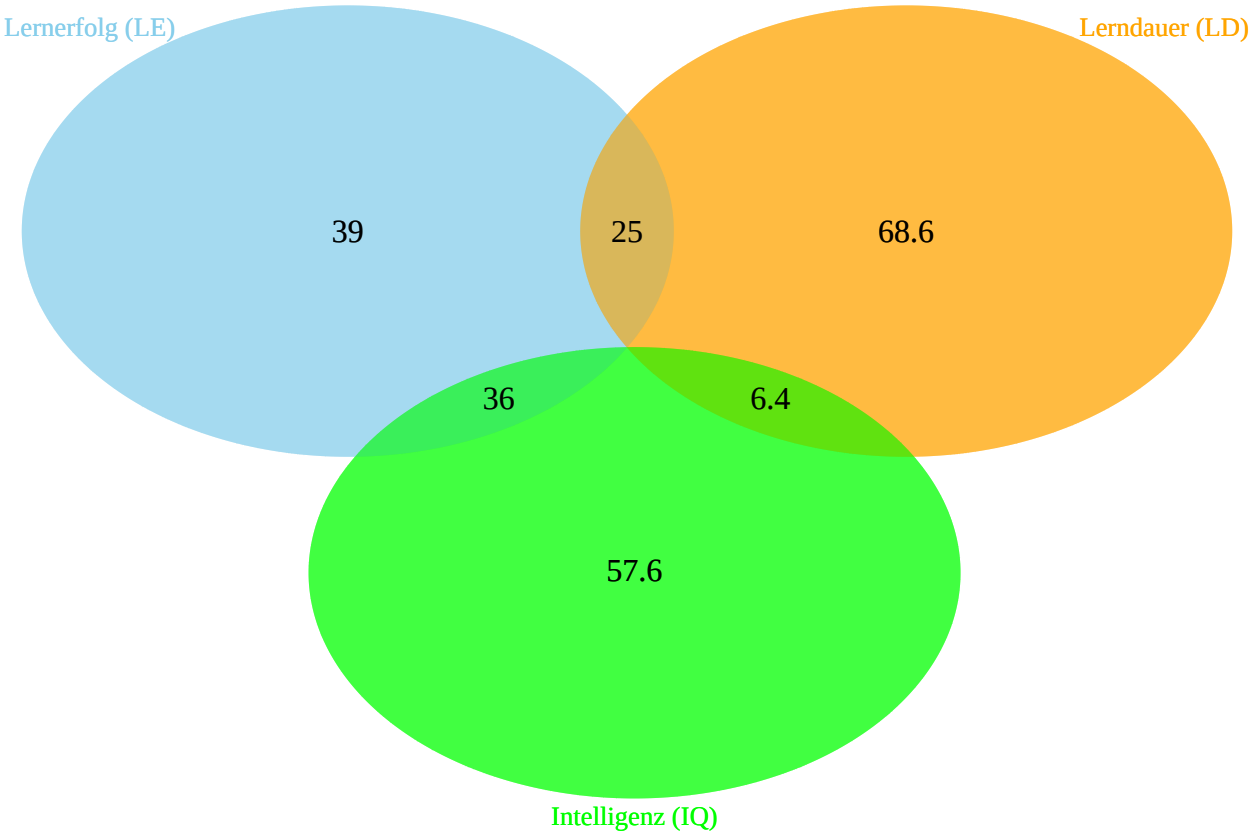
Cohen und Cohen (1975) und Kennedy (1981) konnten zeigen, dass sich das Bestimmtheitsmaß graphisch mittels Venn-Diagrammen veranschaulichen lässt.

Wir betrachten die Variablen Lernerfolg (`LE`) und Lerndauer (`LD`). Angenommen, ihre Korrelation beträgt $r = 0.5$. Daraus folgt der Determinationskoeffizient mit $r^2 = 0.5^2 = 0.25$, was 25% der Varianz erklärt. Mit anderen Worten: 25% der Varianz im Lernerfolg wird durch die Lerndauer erklärt.

[3] Normalerweise wird $R^2$ in der Formel $R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$ definiert, wobei SSR (Sum of Squared Residuals) der Teil der Varianz ist, der nicht durch das Modell erklärt wird, und SST (Total Sum of Squares) die Gesamtvarianz der abhängigen Variablen darstellt. In einem gewöhnlichen linearen Regressionsmodell, das ein Absolutglied (Intercept) beinhaltet, wird das Modell der optimalen Anpassung dadurch erreicht, dass es den Datenmittelwert als Ausgangspunkt nutzt. Dadurch hat SSR im besten Fall einen geringeren Wert als SST, was zu einem positiven $R^2$ führt. Bei einem homogenen Modell ohne Absolutglied, also ohne den konstanten Term, wird das Modell gezwungen, durch den Ursprung (Nullpunkt) zu verlaufen. Dies kann die Anpassung des Modells an die Daten stark verschlechtern, insbesondere wenn die tatsächlichen Datenpunkte um einen von Null verschiedenen Mittelwert gruppiert sind. In solchen Fällen kann die Summe der quadrierten Residuen (SSR) größer werden als die Gesamtvarianz (SST), was zu einem negativen Wert von $1 - \frac{\text{SSR}}{\text{SST}}$ führt. Ein negatives $R^2$ ist eine Indikation dafür, dass das Modell schlechter abschneidet als ein einfaches Durchschnittsmodell, das lediglich den Mittelwert der abhängigen Variablen als Vorhersage nutzt.




Damit ergibt sich aber auch zwangsläufig die Erkenntnis, dass immerhin noch 75% der Variabilität vom Lernerfolg unerklärt sind! Man könnte also davon ausgehen, dass es weitere Variablen/Merkmale gibt, die weitere Anteile dieser noch unerklärten Varianz erklären können. Ideal wären dabei eine, oder mehrere Prädiktorvariablen, die mit der Kriteriumsvariablen hoch, mit den anderen unabhängigen Variablen aber wenig bis gar nicht korrelieren, wie im nachfolgenden Graph dargestellt wird:

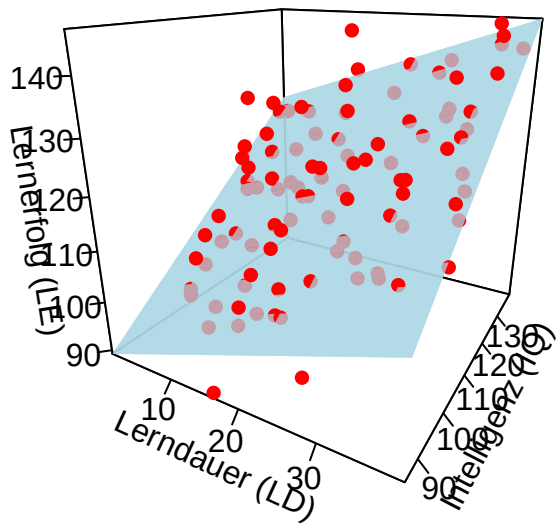



Wie man den Angaben zur aufgeklärten Varianz entnehmen kann, beträgt die Korrelation von $r(LE, LD) = 0.50$, $r(LE, IQ) = 0.60$ und der $r(IQ, LD) = 0.08$. Daraus kann man sich ganz einfach die nicht aufgeklärte Varianz von `LE` errechnen. Diese ergbit sich aus $100 - 25 - 36 = 39$.

## Multiple Regression

Die Multiple lineare Regression ist ein statistisches Verfahren, mit dem versucht wird, eine beobachtete abhängige Variable durch mehrere unabhängige Variablen zu erklären. Die multiple lineare Regression stellt eine Verallgemeinerung der einfachen linearen Regression dar. Das Beiwort „linear“ bedeutet, dass die abhängige Variable als eine Linearkombination (nicht notwendigerweise) linearer Funktionen der unabhängigen Variablen modelliert wird (siehe Wikipedia). Dieser Themenbereich wird im Rahmen der Methodenlehre und Statistik 3 im Detail besprochen.

Die Grundlegende Idee kann man jedoch bereits mit den gerade vorgestellten Venn-Diagrammen bereits gut erfassen. Für ein Modell mit 2 Prädiktoren eignet sich auch noch der Scatterplot gut, um die Eigenschaften und Zusammenhänge der Residuen und des Modells zu verstehen. Anstelle einer Linie, wird es bei 2 Prädiktoren einen Ebene. Ab 3 Prädiktoren ist der Scatterplot allerdings nicht mehr darstellbar.




### Eigenschaften des Determinationskoeffizienten

Der Determinationskoeffizient wird häufig als Gütemaß eines linearen Modells verwendet. Dabei sind jedoch folgende Eigenschaften unbedingt zu berücksichtigen:

- zeigt die *Qualität* der linearen Approximation, jedoch nicht, ob das Modell richtig spezifiziert wurde, also ob eine lineare Anpassung überhaupt die geeignete Modellvorstellung ist.
- sagt nichts über die kausale Ursache des Zusammenhangs aus. Der Schluss, dass die unabhängige Variable $x$ der Grund für die Änderungen in $y$ sind kann diesem Maß nicht entnommen werden!
- gibt keine Auskunft über die statistische Signifikanz des ermittelten Zusammenhangs!
- ein hohes Bestimmtheitsmaß ist kein Beweis für ein *gutes* Modell und ein niedriges Bestimmtheitsmaß bedeutet nicht, dass es sich um ein *schlechtes* Modell handelt. Veranschaulicht wird diese Eigenschaft durch das Anscombe-Quartett[4], siehe nächste Folie.

### Ascombe-Quartett

Bei Betrachgung der nachfolgenden Streudiagramme sieht man klar, dass die Daten und Zusammenhänge verschieden aussehen. Berechnet man die statistischen Kennzahlen, haben diese aber nahezu dieselben Werte!

## Signifikanztests

Wie bei den meisten Kennwerten der Statistisk, will man neben der Größe und Richtung des Effektes auch die statistische Signifikanz berechnen. Damit die Ergebnisse dieser Berechnung sinnvoll interpretiert werden dürfen, müssen bestimmte Voraussetzungen erfüllt sein.

[4]Details siehe Wikipedia

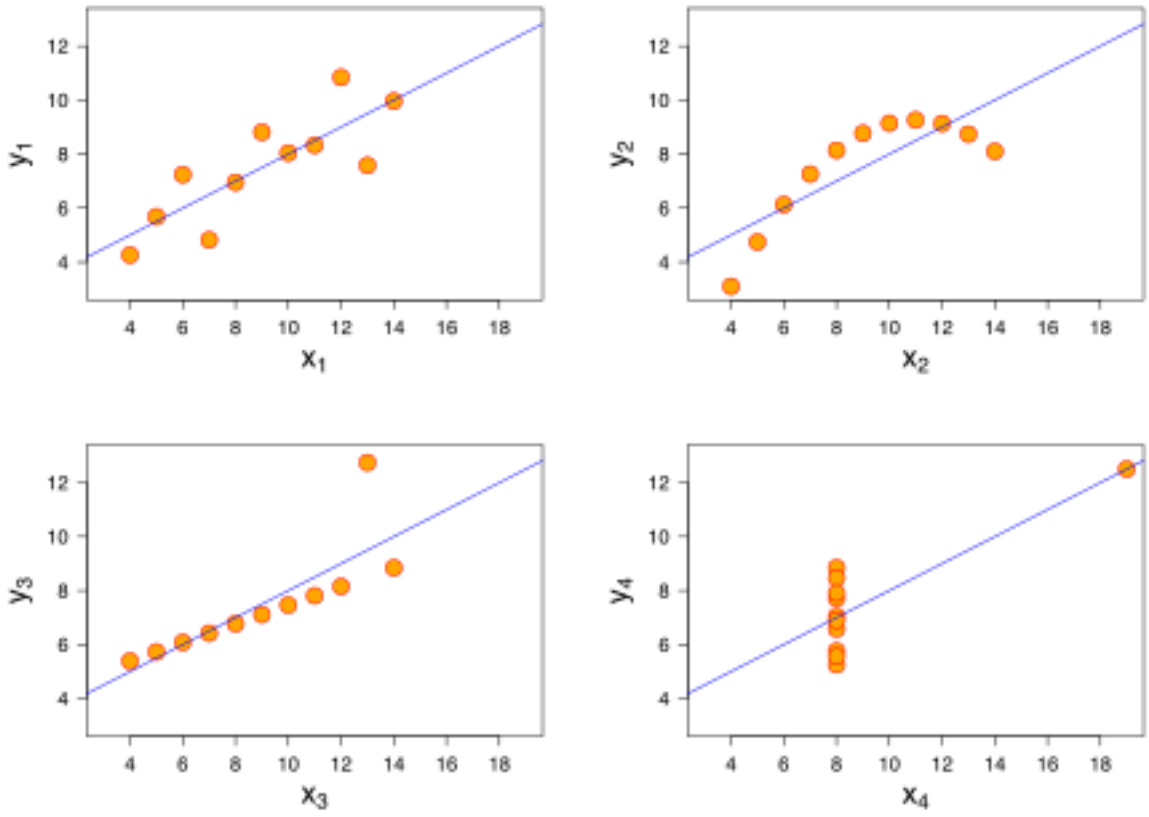



Figure 1: Alternativtext

### Voraussetzungen

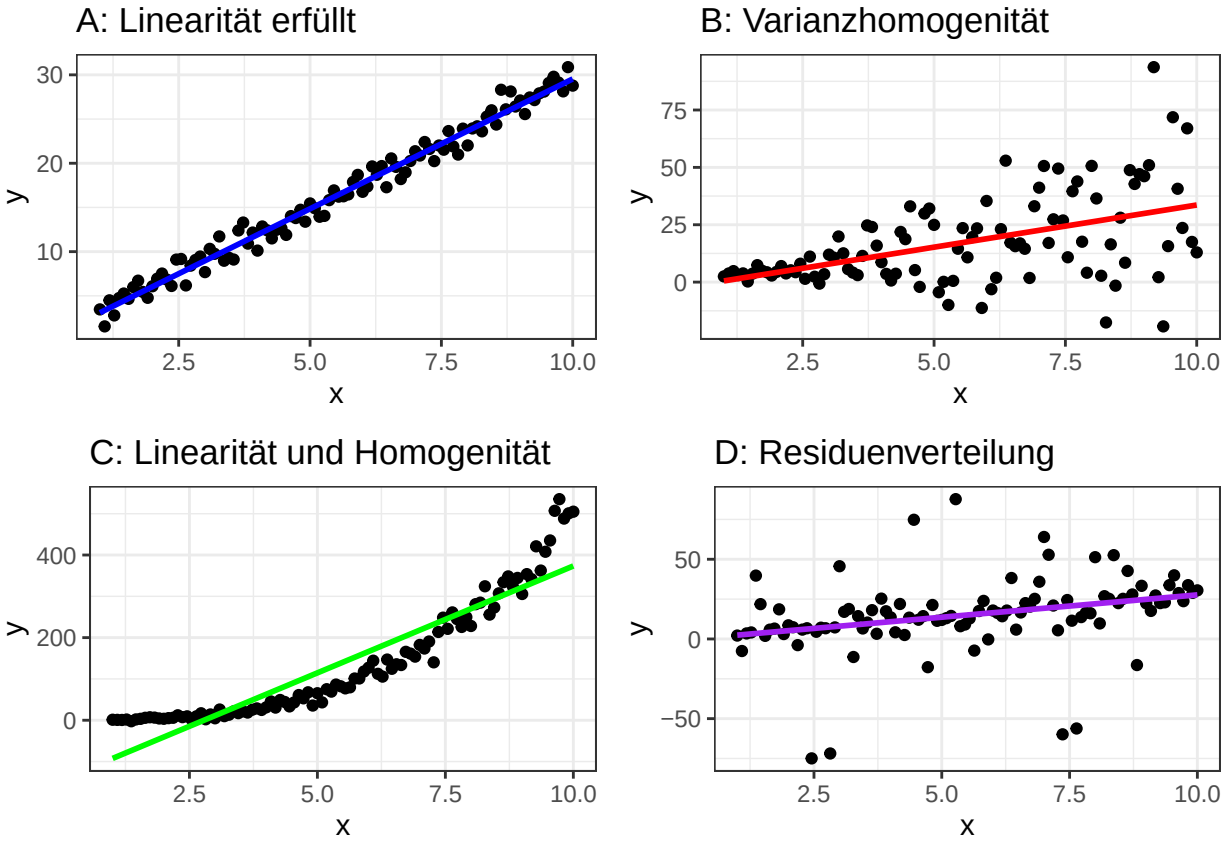
1. **Skalenniveau des Kriteriums**: für die hier besprochenen einfache lineare Regression muss die Kriteriumsvariable intervallskaliert sein.
2. **Linearität**: die in der Population vorliegende Abhängigkeit zwischen Prädiktor und Kriterium kann durch eine Gerade dargestellt werden.
3. **Homoskedastizität**: die Varianz der $y$-Werte, welche an einer bestimmten Stelle des Prädiktors vorliegt, ist für alle Prädiktorwerte gleich.
4. **Normalverteilung**: die Residuen ($\varepsilon_i$) sind normalverteilt ($\Rightarrow$ der Mittelwert des Fehlers $\bar{\varepsilon} = 0$).
5. **Unabhängigkeit der Daten und der Fehler** $\varepsilon$: Alle Daten sollten unabhängig voneinander sein, d.h. die Fälle sollten nicht untereinander korrelieren. Der Wert $x_i$ sollte also nicht einfach von $x_{i-1}$ (mit $1 \leq i \leq N$ ) abgeleitet werden können (gilt insbesondere für die Fehler oder Residuen $\varepsilon$).

### Graphisches prüfen der Voraussetzungen

```
read_chunk("Programme/Graphiken_Voraussetzungsverletzungen.R")
```

Einfach und effizient lassen sich bestimmte Voraussetzungen bereits in einem *Scatterplot* prüfen. Diese *Prüfmethode* sollte immer gekoppelt mit entsprechender Analyse von Kennwerten und gegebenenfalls mit statistischen Verfahren zur Prüfung der Signifikanz einer Verletzung gekoppelt werden.

Die statistischen Verfahren zur Prüfung auf Voraussetzungsverletzungen werden weiter unten noch besprochen.




## Berechnung des Modells in R

Die Berechnung eines linearen Modells mit R ist mit der Funktion:

$$lm(Y \sim X, data = df)$$

denkbar einfach. Die Funktion $lm()$ **l**inear **m**odel erfordert eine *Formel*, also die Angabe der Kriteriumsvariable und der Prädiktorvariable(n), sowie den Datensatz der zum Trainieren des Modells verwendet werden soll. In R steht die Tilde ($\sim$) in einer Formel für die Zuordnung zwischen der abhängigen und den unabhängigen Variablen.

Für das bereist verwendete Beispiel mit dem Lernerfolg und der Lerndauer würde die Berechnung der Koeffizienten in R konkret durch folgenden Aufruf durchgeführt werden:

```
lm_fit <- lm(LE ~ LD, data = df)
```

## Koeffizienten

Nachfolgende Tabelle zeigt die berechneten Regressionkoeffizienten:

```
## # A tibble: 2 x 6
##   term      estimate std.error statistic  p.value std.beta
##   <chr>        <dbl>     <dbl>     <dbl>    <dbl>    <dbl>
## 1 Intercept    34.9      1.57       22.3 3.99e-40   NA
## 2 LD            1.47     0.114      12.9 7.54e-23    0.793
```

### Formale Interpretation

Die Gerade schneidet die y-Achse bei einem Wert von 34.93. Der Standardfehler beträgt 1.57, die Tesstatistik $t = 22.3, p < .001$, womit sich der Interzept signifikant von 0 unterscheidet.[5]

Die Steigung beträgt 1.47, die Testatistik $t = 12.9, p < .001$ ist ebenfalls signifikant. Der standardisierte Steigungskoeffizient $B = .79$ entspricht der Korrelation der beiden Variablen. Der sich daraus berechenbare Determinationskoeffizient $R^2 = .63$. Damit klärt die Lerndauer 63% der Variabilität im Lernerfolg auf.

### Inhaltliche Interpretation

Die Inhaltliche Interpretation ist für den Bericht der Ergebnisse die wichtigere. Im vorliegenden Fall würde man bei einer Lerndauer von 0 Stunden 34.9% Prozent erreichen. Mit jeder zusätzlichen Lernstunde erhöht sich der Lernerfolg um 1.47%.

## Hypothesen

Von Bedeutung ist die statistische Absicherung der Steigung $b_1$. Bei jeder Wiederholung einer Untersuchung wird sich mit sehr großer Wahrscheinlichkeit der berechnete $b_1$ ändern. Sind die Annahmen des Regressionsmodels (Linearität, Homoskedastizität, Normalverteilung der Residuen) erfüllt, dann ist die Stichprobenverteilung der $b_1$ normalverteilt.

Der Mittelwert dieser Verteilung ist dann die unbekannte (wahre) Steigung $\beta$. Ebenfalls unbekannt ist die Standardabweichung der Stichprobenverteilung $\sigma_b$ (= Standardfehler der Steigung). Diese kann jedoch über den Standardfehler der Steigung $s_b$ geschätzt werden (Details siehe Literatur). Es gilt:

$$s_b = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Die Nullhypothese lautet:

$$H_0 : \beta = \beta_0$$

[5] in den meisten Fällen sind die Werte des Interzepts (also $b_0$), sowie dessen Signifikanz, nicht von Interesse!

## Bewertung des Modells - ANOVA

Um zu testen, ob das Modell sich im Vergleich zum *Mittelwertsmodell* signifikant unterscheidet, also ob es hinsichtlich der aufgeklärten Varianz der abhängigen Variablen bedeutsam besser ist, kann man die folgende ANOVA-Tabelle verwenden.

```
## # A tibble: 2 x 6
##   term         df  sumsq  meansq statistic   p.value
##   <chr>     <int>  <dbl>   <dbl>     <dbl>     <dbl>
## 1 LD            1 10009. 10009.       166.  7.54e-23
## 2 Residuals    98  5893.    60.1       NA  NA
```

Wenn die ANOVA ein signifikantes Ergebnis liefert, wie es in vorliegender Fall auch zutrifft ($F(1, 98) = 166, p < .001$), kann man von einer signifikant besseren Modell als das Mittelwertsmodell ausgehen.

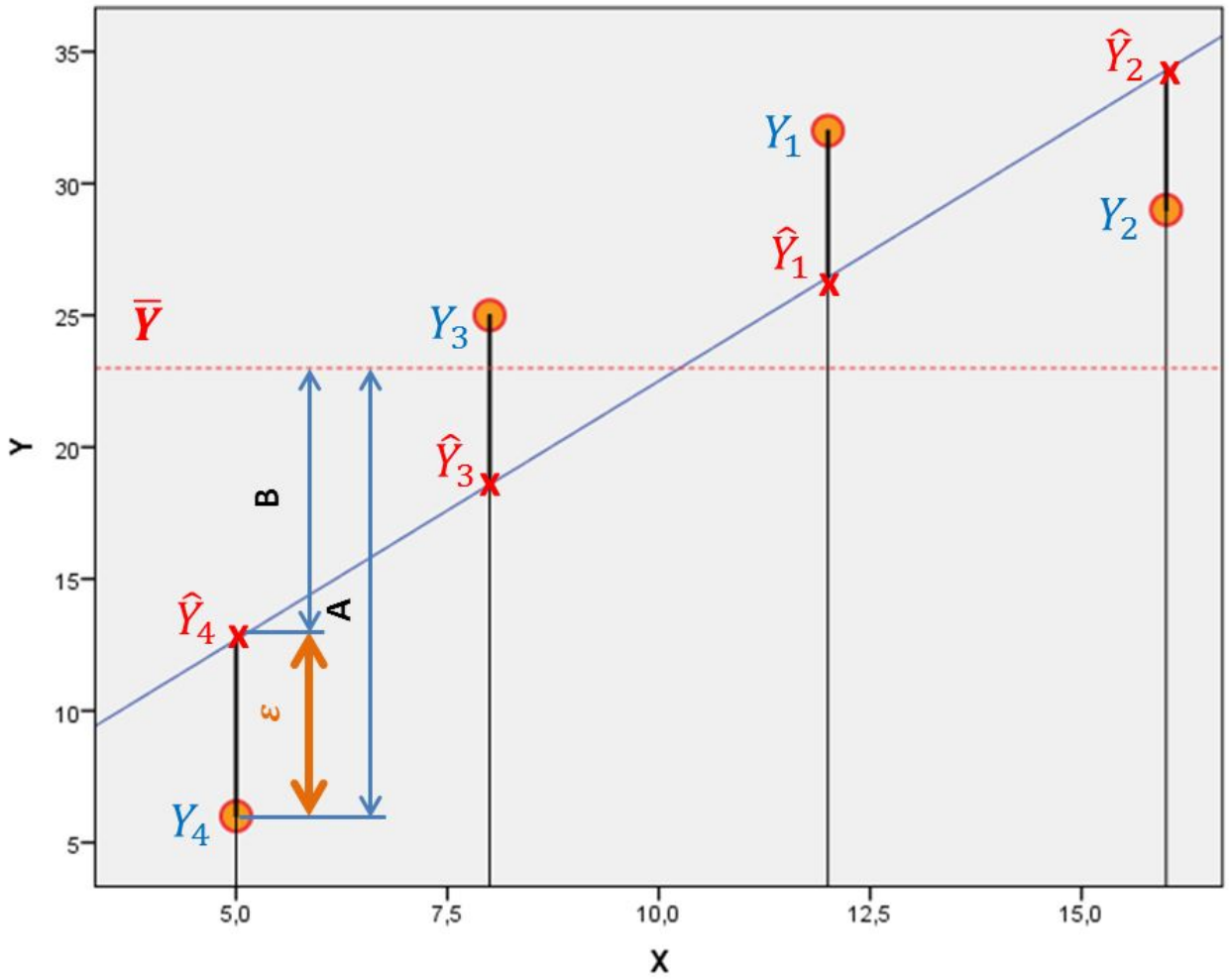



Figure 2: Quadratsummen

Erläuterung der Quadratsummen in der ANOVA-Tabelle:

- Totale-Quadratsumme (LD sumsq + Residuals sum_Sq): die quadrierte Summe der Differenzen zwischen beobachteten Werten und dem Mittelwert, also $QS_{Total} = \sum(y_i - \bar{y})^2$
- Fehler-Quadratsumme (Residual sumsq): die Summe der Differenzen zwischen beobachteten Werten und vorhergesagten Werten, also $QS_{Fehler} = \sum(y_i - \hat{y}_i)^2$

- Modell-Quadratsumme (LD sumsq): die Summe der Differenzen zwischen vorhergesagten Werten und dem Mittelwert, also $QS_{Modell} = \sum(\hat{y}_i - \bar{y})^2$

Mit den Ergebnissen der ANOVA lassen sich folgende Fragen beantworten:

- Welcher Anteil der Gesamtvariabilität $QS_{Total}$ wird durch das Modell $QS_{Modell}$ erklärt?
- In welchem Verhältnis steht die durch das Modell aufgeklärte Varianz $MQS_{Modell}$ (mittlerer Quadratsumme des Modells) zur nicht aufgeklärten Varianz $MQS_{Fehler}$ (mittlerer Quadratsumme des Fehlers)?

***Frage 1:***

$$R^2 = \frac{QS_{Modell}}{QS_{Total}}$$

Das entspricht also dem bereits bekannten Determinationskoeffizienten, also der aufgeklärten Varianz.

***Frage 2***

$$MQS_{Modell} = \frac{QS_{Modell}}{df_{Modell}} \text{ und } MQS_{Fehler} = \frac{QS_{Fehler}}{df_{Fehler}}$$
$$df_{Modell} = 1 df_{Fehler} = N - p$$

mit \$p = \$ Anzahl der verwendeten Parameter $b_0$ und $b_1$, also 2.

***Teststatistik***

$$F = \frac{MQS_{Modell}}{MQS_{Fehler}}$$

### Konfidenzintervalle

Für den Interzept gilt:

$$\text{untere Grenze} = b_0 - t_{df;1-\alpha/2} \cdot s_b$$
$$\text{obere Grenze} = b_0 + t_{df;1-\alpha/2} \cdot s_b$$

Für die Steigung gilt:

$$\text{untere Grenze} = b_1 - t_{df;1-\alpha/2} \cdot s_b$$
$$\text{obere Grenze} = b_1 + t_{df;1-\alpha/2} \cdot s_b$$

***Erkenntnis:***

Der Signifikanztest $H_0 : \beta = 0$ führt bei zweiseitiger Testung dann zu einem nicht-signifikanten Ergebnis, wenn das $1 - \alpha$ Konfidenzintervall den Wert null enthält (z.B. $KI_{b_1} = [-0.2, 0.4]$).

## Kategorieller Prädiktor

Bei lineare Modellen ist häufig neben intervallskalierten Prädiktorvariablen auch die Verwendung von kategoriellen 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.

## Mehrstufige(r) kategorielle(r) Prädiktor(en)

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[6]. Darüber hinaus wurden die TeilnehmerInnen über ihre jeweilige Zuordnung zu eine bestimmten, persönlich bevorzugten Musikrichtung (*music*) befragt. Bei dem Fesitval 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 katetoriellen 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 ($N_{DummyVars} = N_{Stufen} - 1$)
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 Affliation | 0 | 0 | 0 |

Bei der linearen Modellierung in R werden kategorielle Daten im Modell automatisch Dummy-Kodiert. Will man jedoch eine spezielle Anordung 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
```

[6]Daten und Beispiel aus (A. Field 2017), Kapitel 7.12.1

```r
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"), digits = 3)
```

| | 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 |

## Modelle mit kategoriellen 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.

```r
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 |

```r
mod_dummy_1 <- lm(change ~ music, data = DF)
AllRes      <- summary(mod_dummy_1)
pander(anova(mod_dummy_1), digits = 3)
```

Table 5: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **music** | 3 | 4.65 | 1.55 | 3.27 | 0.0237 |
| **Residuals** | 119 | 56.4 | 0.474 | NA | NA |

```r
pander(summary.lm(mod_dummy_1), digits = 3)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -0.554 | 0.0904 | -6.13 | 1.15e-08 |
| **musiccrusty_v_NMA** | -0.412 | 0.167 | -2.46 | 0.0152 |
| **musicindie_v_NMA** | -0.41 | 0.205 | -2 | 0.0477 |
| **musicmetal_v_NMA** | 0.0284 | 0.16 | 0.177 | 0.86 |

Table 7: 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:

- $R^2 = 0.076$: d.h., dass 7.6% der Variabilität in der Änderung der Hygenewerte zwischen ersten und dritten Tag (*change*) durch die Zugehörigkeit zu einer Musikgruppe erklärt werden.
- $F(3, 119) = 3.27; p = .053$ 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 $crusty - NMA = -.966 - (-0.554) = -0.412$, stellt man fest, dass diese Differenz dem Estimate, also dem *b*-Koeffizienten entspricht. Offenbar ist die Änderung der Hygienewerte bei *crusty* höher als bei der *NMA* $\rightarrow$ *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.


Field, & Wilcox, A. P. 2017. "Robust Statistical Methods: A Primer for Clinical Psychology and Experimental Psychopathology Researchers." *Behaviour Research and Therapy, 98, 19-38.* https://doi.org/https://doi.org/10.1016/j.brat.2017.05.013.

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

Hurn, M. David. 2008. "Modern Robust Statistical Methods: An Easy Way to Maximize the Accuracy and Power of Your Research." *American Psychologist, Vol. 63, No. 7, 591–601.* https://doi.org/10.1037/0003-066X.63.7.591.

Mair, Wilcox R., P. 2020. "Robust Statistical Methods in r Using the WRS2 Package." *Behavioural Research Methods, 52(2), 464-488.* https://doi.org/10.3758/s13428-019-01246-w.

Wilcox, R. 2012. *Introduction to Robust Estimation & Hypothesis Testing.* 3rd ed. Amsterdam, The Netherlands: Elsevier.