Vorlesung: Empirische Bildungsforschung

Session 02: Regression Bootcamp

Dr. Edgar Treischl
Last update: 2022-04-25

Ablauf

- Grundidee der linearen Regression

- Interpretation und KQ Schätzung

- Multivariate Analysen

- Güte der Vorhersage: R^2

- Regression App

Die Online App dient also der Vertiefung der linearen Regression. Ist die App auf der nächsten Seite nicht sichtbar, kannst du Sie auch in einem neuen Fenster laden und an den enstprechenden Stellen im Online Modul darauf zurückgreifen.

Online Linear Regression App

Lineare Regression in a Nutshell

Datensatz:

Cars

X

mpg

Y

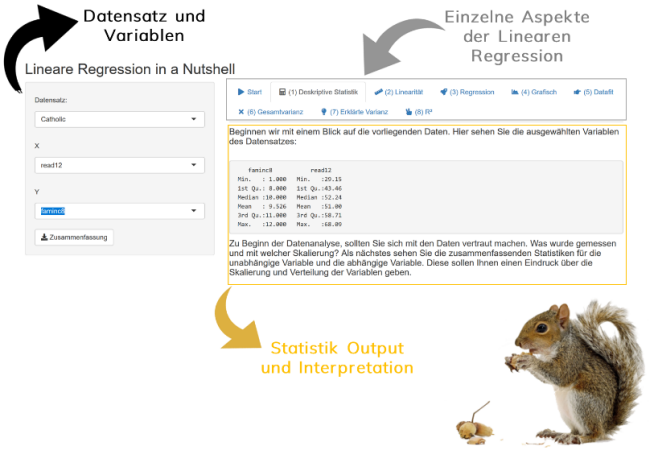
mpg

Zusammenfassung

Source Code

Start | (1) Deskriptive Statistik | (2) Linearität | (3) Regression | (4) Grafisch | (5) Datafit | (6) Gesamtvarianz | (7) Erklärte Varianz | (8) R^2

Berechne die Ergebnisse einer linearen Regression, ganz ohne eigene Programmierkenntnisse! Wähle einen Beispieldatensatz, eine unabhängige Variable (X) und eine abhängige Variable (Y). Schätze zum Beispiel mit dem Catholic Datensatz, ob das Einkommen einer Familie (faminc8) einen Einfluss auf die Leseleistung (read12) von Kinder hat.

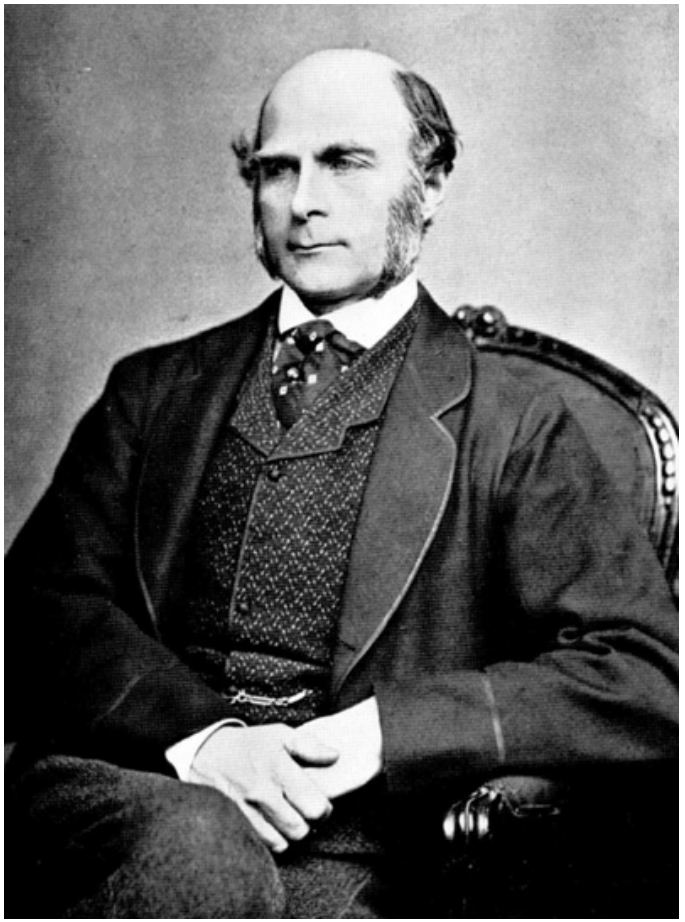


Die Grundidee der Regression

Was untersucht die EBF?

- Einfluss von X auf die Noten (oder Kompetenzen) von Schülerinnen/Schüler (SuS)
- Einfluss von X auf den Übergang (Ja/Nein) / Erreichen eines Schulabschlusses (Ja/Nein) von SuS
- Einfluss der Ausgestaltung des Bildungssystems (Klassengröße, G8/G9) auf Y
- (Kausale) Effekt von X auf Y
- Anwendungsbeispiel der Sitzung: Haben große Eltern auch große Kinder?

Francis Galton



Die Daten

Ignore the R Code, but not the Output

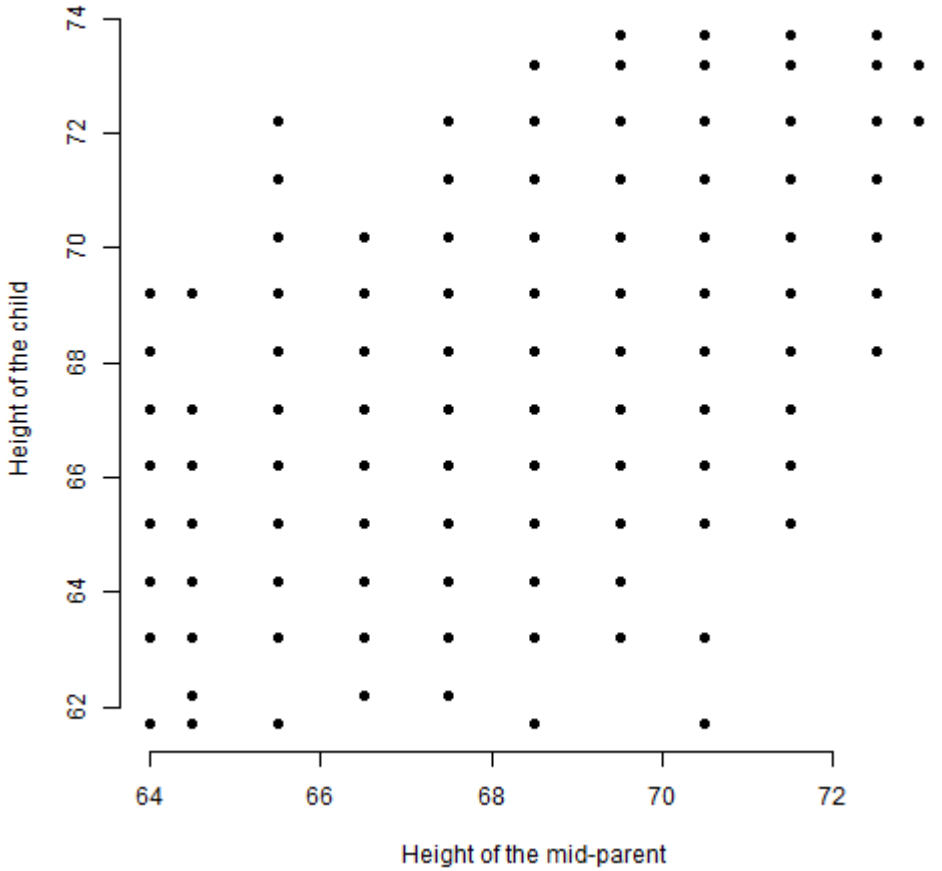
```
# code chunk here
library(HistData)
data(Galton)
head(Galton)
```

```
##   parent child
## 1   70.5  61.7
## 2   68.5  61.7
## 3   65.5  61.7
## 4   64.5  61.7
## 5   64.0  61.7
## 6   67.5  62.2
```

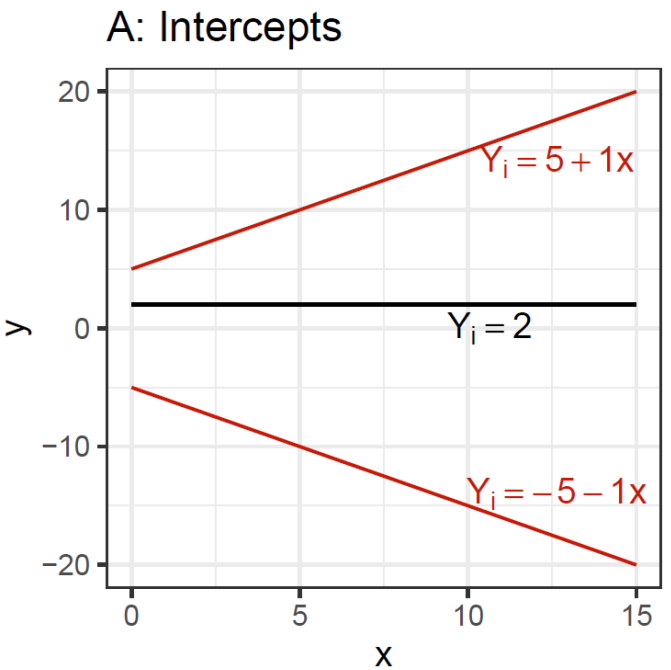
Regressieren heißt zurückführen: Eine abhängige metrische Variable (AV wie bspw. Körpergröße, die Schulleistung, etc.) wird auf eine oder mehrere unabhängige Variable (UV wie die Körpergröße der Eltern) zurückgeführt (sprich: regressiert).

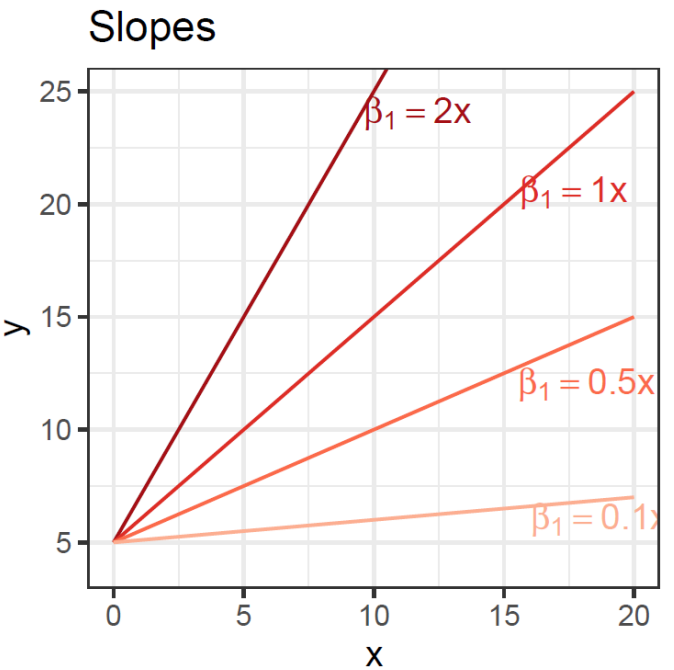
Der Zusammenhang zwischen der UV und der AV wird durch eine Funktionsgerade beschrieben. Wir tragen alle Werte der UV und der AV in einem Diagramm ab und legen eine Gerade in die Datenwolke, welche den Zusammenhang zwischen UV and AV möglichst präzise beschreibt.

Die Datenwolke



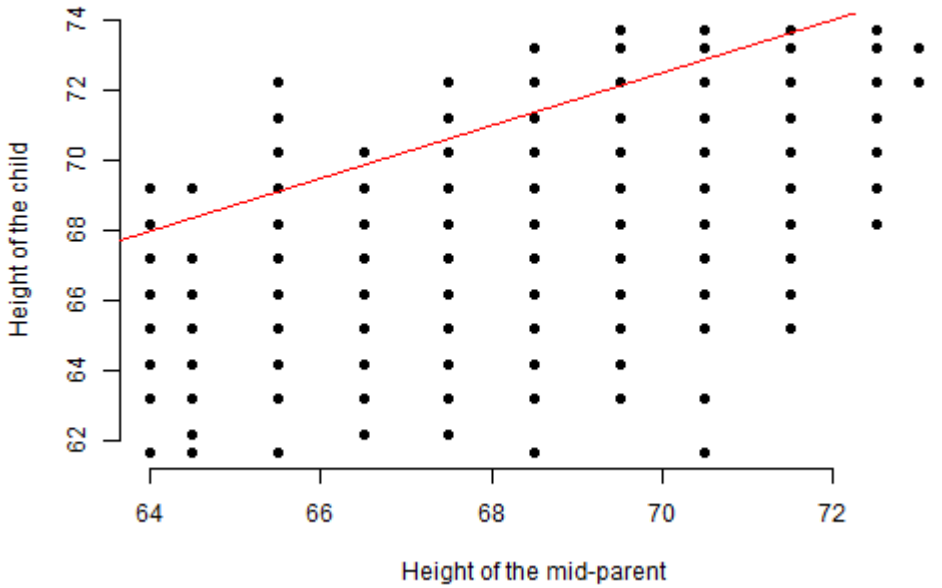
Welche Funktionsgerade?





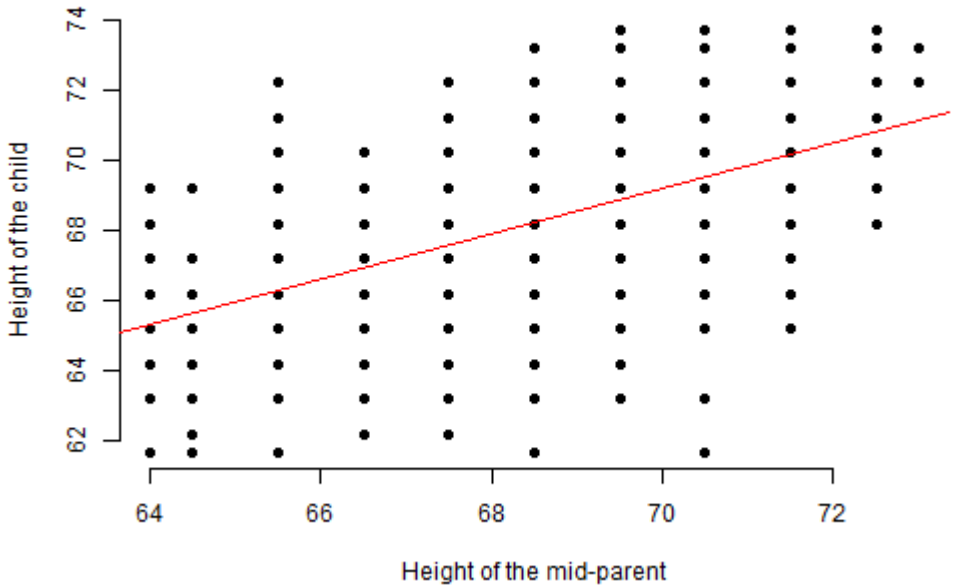
Welche Funktionsgerade darf's denn sei?

```r
#a = intercept, b = slope
plot(Galton$parent, Galton$child,
xlab = "Height of the mid-parent", ylab = "Height of the child",
pch = 19, frame = FALSE)
abline(a=20, b=0.75, col = "red")
```



Eine Funktionsgerade mit kleinem Fehler!

```
plot(Galton$parent, Galton$child,
xlab = "Height of the mid-parent", ylab = "Height of the child",
pch = 19, frame = FALSE)
abline(a=23.94153, b=0.64629, col = "red")
```



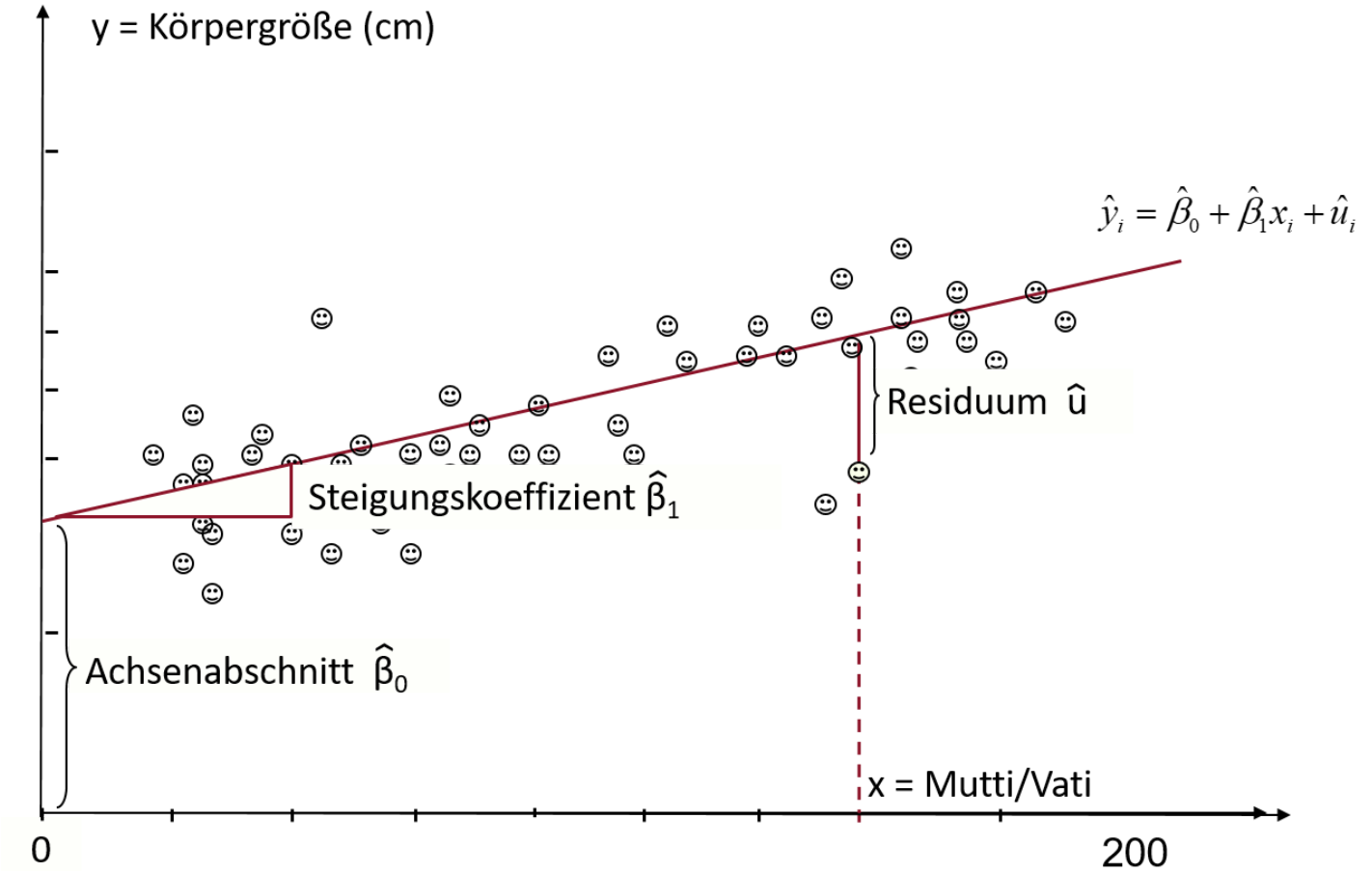
Zusammenfassung der Begrifflichkeiten

- Y Achsenabschnitt: Wo beginnt die Gerade?

- Steigungskoeffizient: Wie steil oder flach verläuft die Gerade?

- Fehlerterm (Residuum): Vorhersage ist nicht perfekt, wir machen also einen Fehler.

Zusammenfassung grafisch



Interpretation und KQ Schätzung

Kochrezept Interpretation

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \epsilon$$

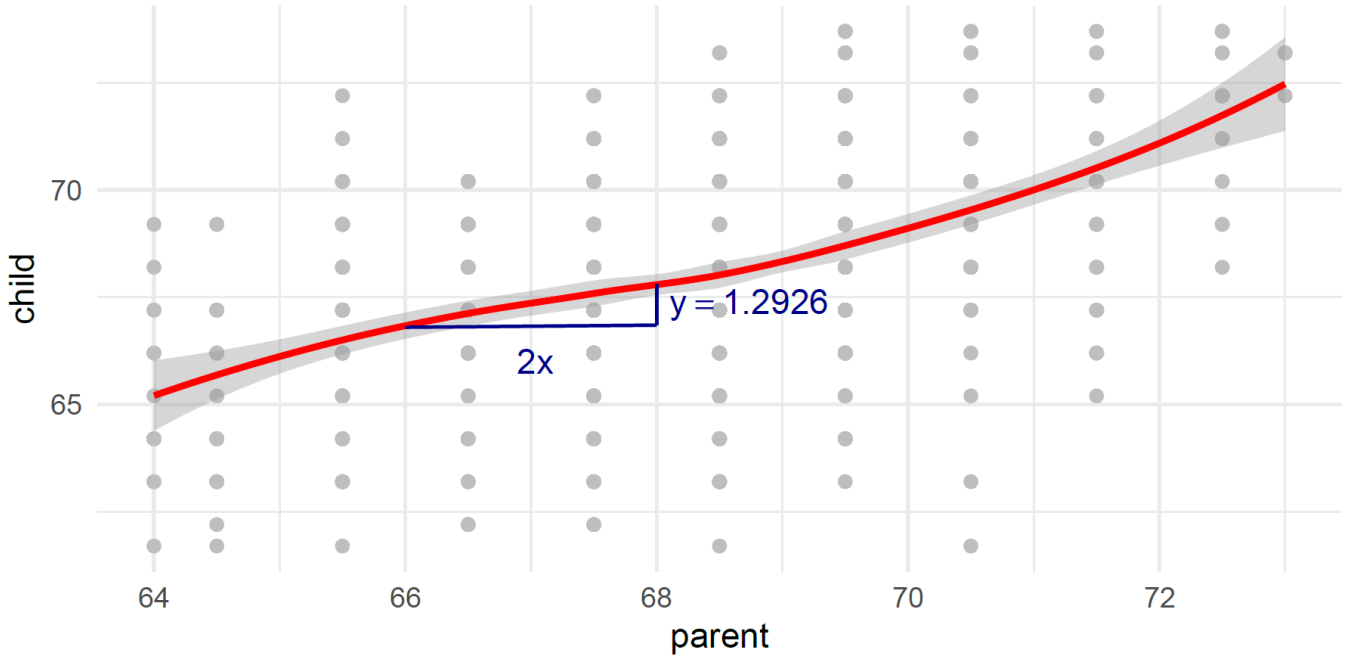
- β_0: Durchschnittlicher Wert der abhängigen Variable y, wenn alle anderen Einflussgrößen gleich 0 sind. (y-Achsensabschnitt oder Intercept)
- β_1: Verändert sich die unabhängige Variable X um eine Einheit, so verändert sich die abhängige Variable um $\beta 1$ Einheiten (Steigung)
- ϵ: Fehlerterm (Residuum)

Beispiel Regressionsoutput

```
library(HistData)
lm(child ~ parent, data = Galton)
```

```
##
## Call:
## lm(formula = child ~ parent, data = Galton)
##
## Coefficients:
## (Intercept)       parent
##     23.9415       0.6463
```

Interpretation grafisch



Sind die Effekte signifikant?

```r
library(HistData)

model.galton<-lm(child ~ parent, data = Galton)
summary(model.galton)
```

```
##
## Call:
## lm(formula = child ~ parent, data = Galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Bei der Regression soll die Gerade den Funktionszusammenhang so gut wie möglich beschreiben. Hierfür nutzen wir den KQ Schätzer.

Die Gerade wird so durch die Punktewolke gelegt, dass die Summe der quadrierten Residuen möglichst klein ist. Hierfür wird die Methode der kleinsten Quadrate (KQ Schätzer) eingesetzt, häufig auch Ordinary Least Squares (OLS) genannt.

Beim KQ Schätzer werden die Residuen quadriert, weil

...

- sich sonst positive und negative Abweichungen ausgleichen.

- stärkere Abweichungen stärker in die Berechnung einfließen sollen.

Der Erwartungswert der Residuen ist gleich 0, d.h. im Durchschnitt sind die Fehler 0: E(u)=0.

Math Wizardy: The Slope spell

```
model.galton<-lm(child ~ parent, Galton)
model.galton$coefficients
```

```
## (Intercept)      parent
##  23.9415302   0.6462906
```

$$m = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$$

```
Galton %>%
  summarise(mean_x = mean(parent),
            mean_y = mean(child),
            cov  = sum((parent - mean_x)*(child-mean_y)),
            var_x = sum((parent - mean_x)^2),
            slope = cov / var_x
            )
```

```
##      mean_x   mean_y      cov    var_x     slope
## 1 68.30819 68.08847 1913.898 2961.358 0.6462906
```

Math Wizardy: The Intercept spell

$$t = \bar{y} - m\bar{y}$$

```
intercept <- function(x, y, m){
  t <- mean(y) - (m * mean(x))
  return(t)
}

Galton_slope <- 0.6462906


Galton_intercept <- intercept(Galton$parent,
                              Galton$child,
                              Galton_slope)
Galton_intercept
```

```
## [1] 23.94153
```

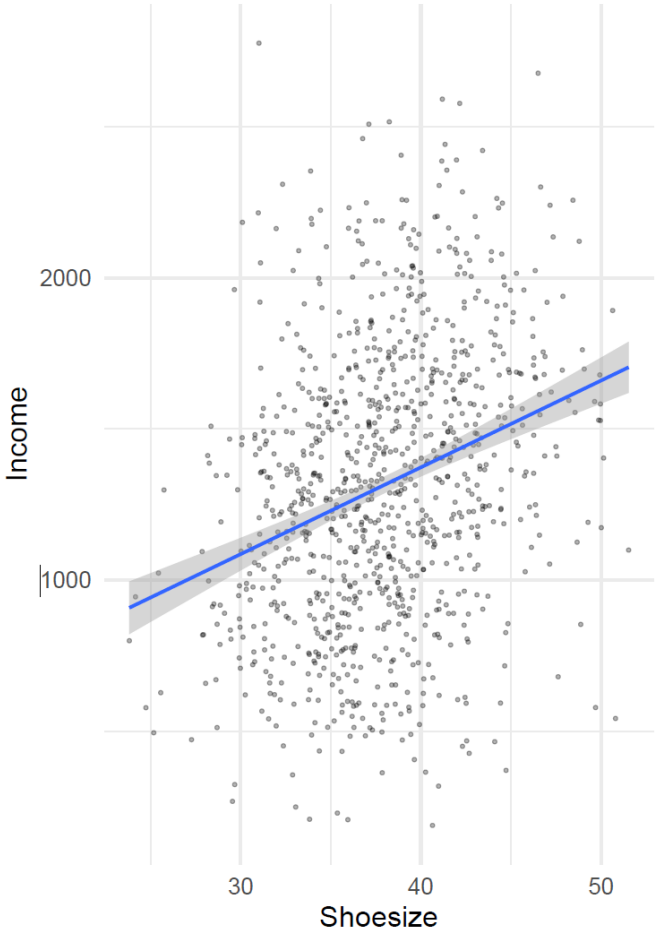
Multivariate Analysen

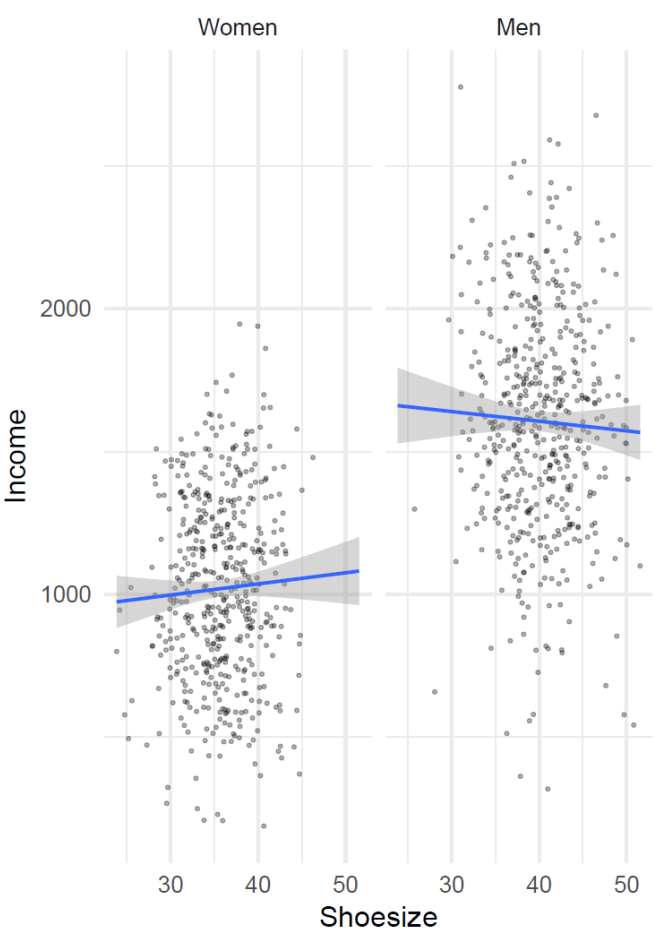
- Bivariate vs. multivariate Analysen

- Funktionsgleichung wird mit weiteren Koeffizienten (Variablen) erweitert

- Ceteris paribus (d.h. unter sonst gleichen Bedingungen): Der Einfluss von X auf Y kann unter Konstanthaltung weiterer Variablen berechnet werden

Spurious confounder





Identifikation eines kausalen Effekts....

- **mit Hilfe der Regression:** Berechnung eines Effekts durch Kontrolle beobachteter Variablen, d.h. Kontrolle aller Variablen die sowohl den Treatmentstatus beeinflussen und einen Einfluss auf die abhängige Variable haben.
- **Hoffnung**: Nach Kontrolle der Variablen ist die Zuweisung des Treatmentstatus Klassengröße „so gut wie zufällig“ (CIA)
- **Merke:** Konditionale Unabhängigkeit (conditional independence assumption, CIA): Bei der Regression muss die Annahme getroffen werden, dass nach der Kontrolle der anderen unabhängigen Variablen die Verteilung der Einheiten über die Treatment- und Kontrollgruppe in Hinblick auf die abhängige Variable so gut wie zufällig erfolgt.

Wie gut ist unsere Vorhersage?

```
##
## Call:
## lm(formula = child ~ parent, data = Galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

R^2

Die Güte der Vorhersage der Regression. Ein kurzer Auszug zum Nachlesen.

“In vielen Fällen ist die Vorhersage der Regression nicht perfekt, d.h. es werden Fehler gemacht. Es handelt dabei sich um die Differenz zwischen dem vorhergesagten Wert (der Gerade) und den beobachteten Werten.”

“R-Quadrat ist ein Indikator, der uns hilft zu beurteilen, wie groß der Fehler ist oder wie gut das Modell die Zielvariable (Outcome) erklärt.”

“Um R-Quadrat zu verstehen, müssen wir zuerst die Gesamtvarianz (Streuung) von X und Y uns vorstellen. Nehmen wir an, dass X gar keinen Einfluss auf Y hat, dann wäre der Beta-Koeffizient und damit die Steigung der Geraden gleich 0.”

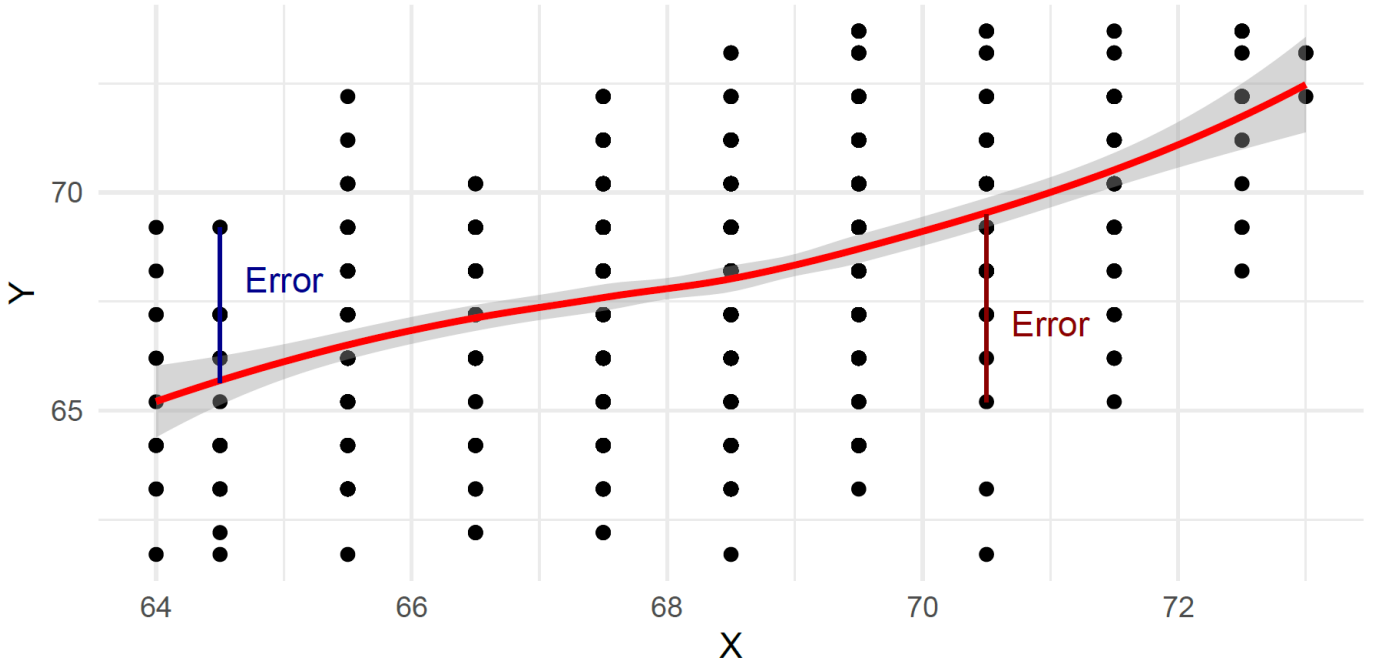
“D.h. es ist egal welcher X-Wert vorliegt, wir würden immer den gleichen Y-Wert beobachten. Die Gerade wäre flach. Das einzige zur Verfügung stehende Mittel um Y vorherzusagen, wäre beispielsweise das arithmetische Mittel von X und dies ist der durchschnittliche Fehler. ”

“Aber wir haben im Online Tool ja bereits eine Regressionslinie angepasst. Auf Grundlage der beobachteten Werte von X können wir Y also bis zu einem gewissen Grad erklären.”

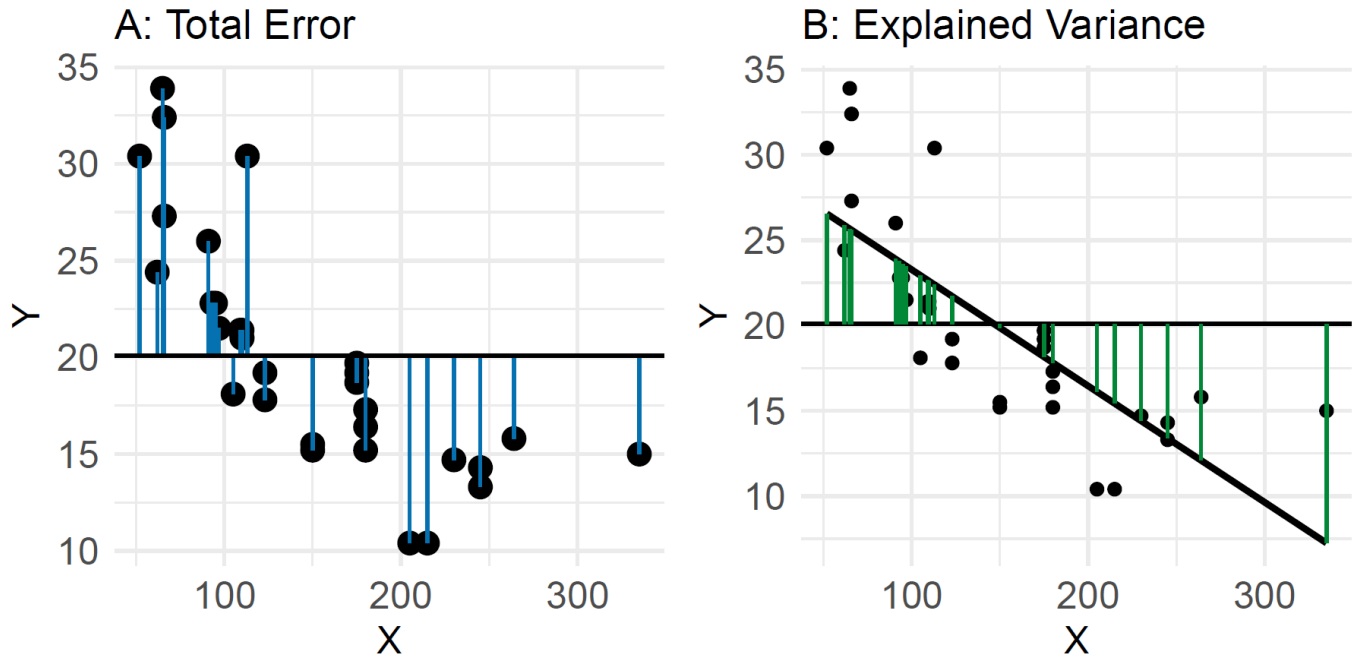
“Wir kennen demnach die Gesamtvarianz und die erklärte Varianz. Mit diesen zwei Werten können wir R^2 berechnen und somit den Fehler des Modells bewerten. R^2 ist der Anteil (%) der Varianz von Y, der durch die Regression vorhersagbar ist. Also die erklärte Varianz geteilt durch die Gesamtvarianz.”

“Grafisch und an einem Beispiel lässt sich dies leichter nachvollziehen.”

ERROR



Gesamtvarianz vs. Erklärte Varianz



Merke:

> R^2 ist der Anteil der erklärten Variation an der Gesamtvariation und liegt zwischen 0 und 1.

Interpretation:

- $R^2 = 0$: Die Regression (X) trägt nichts zur Erklärung von Y bei.
- $R^2 = 1$: Die Regressionsfunktion erklärt Y perfekt (alle Punkte liegen genau auf der Geraden).
- $R^2 = 0{,}2$: 20% der Varianz wird durch die unabhängige(n) Variable(n) erklärt.

Zusammenfassung

- Grundidee und Logik (multivariater) Modelle

- Linear regression as the working horse of the social sciences

- Interpretation der Koeffizienten und Modellgüte