Korrelation

Einführung

In der Datenanalyse möchten wir den Zusammenhang zwischen

- der Ergebnisvariablen *y* (auch abhängige Variable genannt)
- und einer erklärenden/vorhersagenden Variablen *x* (auch unabhängige Variable genannt)

modellieren.

In der Mathematik wird dies oft beschrieben als: Modellierung der abhängigen Variablen *y* als Funktion der unabhängigen Variablen *x*.

Zielsetzungen

- **Inferenz:** Beschreibung und Erklärung des Zusammenhangs von *y* und *x*. Wenn möglich den kausalen Zusammenhang zwischen den zwei aufdecken.
- **Vorhersage:** Ergebnisvariable *y* möglichst genau mit der verfügbaren Information aus *x* vorhersagen. Hier ist das Zusammenspiel unterschiedlicher Variablen nicht so wichtig, hauptsache es wird die Vorhersagegenauigkeit gesteigert.

Zielsetzungen

- **Inferenz:** Beschreibung und Erklärung des Zusammenhangs von *y* und *x*. Wenn möglich den kausalen Zusammenhang zwischen den zwei aufdecken.
- **Vorhersage:** Ergebnisvariable *y* möglichst genau mit der verfügbaren Information aus *x* vorhersagen. Hier ist das Zusammenspiel unterschiedlicher Variablen nicht so wichtig, hauptsache es wird die Vorhersagegenauigkeit gesteigert.

> Wir beschäftigen uns im folgenden mit der Inferenz und klammern die Vorhersage aus

Datensatz

Das Konzept der Regressionsanalyse soll an einem anschaulichen Beispieldatensatz verdeutlicht werden

Datensatz

> Das Konzept der Regressionsanalyse soll an einem anschaulichen Beispieldatensatz verdeutlicht werden

Der Datensatz, welcher im folgenden verwendet wird beschäftigt sich mit der Lehrevaluation von 463 Vorlesungen an der University of Austin, Texas aus dem Jahr 2005.

Die Daten sind von Openintro.org.

```
evals <- read_csv("data/evals.csv")
```

Datensatz

Neben der Evaluation der Veranstaltung sind unter anderem *Attraktivität*, *Alter*, *Geschlecht* und noch einige weitere Faktoren zum Dozent/der Dozentin erfasst worden.

Sie sollen sich im folgenden auf *Evaluation*, *_Attraktivität*, *Alter* und *Geschlecht* beschränken.

```
used_evals <- evals %>%
  mutate(ID = rownames(evals),
         gender = as.factor(gender)) %>%
  select(ID, score, bty_avg, gender, age)
```

Einführung

Zuerst sollten Sie sich einen Überblick über den Datensatz verschaffen

```
glimpse(used_evals)
```

```
Rows: 463
Columns: 5
$ ID      <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"…
$ score   <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, 4…
$ bty_avg <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, 3.333…
$ gender  <fct> female, female, female, female, male, male, male, male, male,…
$ age     <dbl> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, 4…
```

Einführung

Zuerst sollten Sie sich einen Überblick über den Datensatz verschaffen

```
glimpse(used_evals)
```

```
Rows: 463
Columns: 5
$ ID      <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"…
$ score   <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, 4…
$ bty_avg <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, 3.333…
$ gender  <fct> female, female, female, female, male, male, male, male, male,…
$ age     <dbl> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, 4…
```

- Im ersten Schritt sollten Sie sich **immer** die Rohdaten anschauen

Einführung

Zuerst sollten Sie sich einen Überblick über den Datensatz verschaffen

```
glimpse(used_evals)
```

```
Rows: 463
Columns: 5
$ ID      <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"…
$ score   <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, 4…
$ bty_avg <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, 3.333…
$ gender  <fct> female, female, female, female, male, male, male, male, male,…
$ age     <dbl> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, 4…
```

- Im ersten Schritt sollten Sie sich **immer** die Rohdaten anschauen
- Im zweiten Schritt deskriptive Analysen erstellen

Einführung

Zuerst sollten Sie sich einen Überblick über den Datensatz verschaffen

```
glimpse(used_evals)
```

```
Rows: 463
Columns: 5
$ ID      <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"…
$ score   <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, 4…
$ bty_avg <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, 3.333…
$ gender  <fct> female, female, female, female, male, male, male, male, male,…
$ age     <dbl> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, 4…
```

- Im ersten Schritt sollten Sie sich **immer** die Rohdaten anschauen
- Im zweiten Schritt deskriptive Analysen erstellen
- Im dritten Schritt sollten Sie *explorative Grafiken* erstellen

Einführung

- `ID`: ID für jeden Kurs
- `score`: Durchschnittliches Evaluationsergebnis für diesen Kurs (nummerisch). Dieses möchten wir erklären, ist somit unsere *y* Variable.
  - Bester Wert ist 5, schlechtester 1.
- `btg_avg`: Durchschnittliche Attraktivität des Dozenten/der Dozentin, wie dieser von den Studenten eingeschätzt wurde (nummerisch).
  - Höchster Wert ist 10, niegrigster 1.
- `age`: Alter des Dozenten (nummerisch).
- `language`: Muttersprachler oder nicht (String-Variable)

Einführung

Im nächsten Schritt sollten Sie erste deskriptive Analysen durchführen, um ihren Datensatz besser kennen zu lernen.

```
library(skimr)
used_evals %>%
  skim()
```

```
── Data Summary ────────────────────────
                           Values
Name                       Piped data
Number of rows             463
Number of columns          5
_____________________
Column type frequency:
  character                1
  factor                   1
  numeric                  3
_____________________
Group variables            None

── Variable type: character ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
1 ID                    0             1     1     3     0      463          0

── Variable type: factor ───────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 gender                0             1 FALSE          2 mal: 268, fem: 195

── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75
1 score                 0             1  4.17 0.544  2.3   3.8   4.3   4.6
2 bty_avg               0             1  4.42 1.53   1.67  3.17  4.33  5.5
3 age                   0             1 48.4  9.80  29    42    48    57
   p100 hist
1  5
2  8.17
3 73
```

Einführung

Die bisherigen Analysen waren ausschließlich *univariat*, d.h. Sie betrachteten bisher immer nur die Variable, welche Sie interessiert. Jedoch ist das Zusammenspiel der Variablen mit anderen Variablen auch wichtig.

Im nächsten Schritt sollten Sie sich die *Korrelation* zwischen unterschiedlichen Variablen anschauen, d.h. wie groß ist der lineare Zusammenhang zwischen zwei Variablen.

Korrelation

Den zuvor beschrieben Sachverhalt, dass attraktivere Dozenten/Dozentinnen bessere Lehrevaluationsergebnisse haben, können wir durch die Korrelation beschreiben. Die Korrelation gibt an, wie zwei Variablen sich zueinander verhalten, wenn z.B. eine Variable sich um eine Einheit erhöht.

Der Korrelationskoeffizient für zwei Variablen $(x_1, y_1), \ldots, (x_n, y_n)$ ist definiert als:

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_x}{\sigma_x} \right) \left(\frac{y_i - \mu_y}{\sigma_y} \right)$$

mit μ_x, μ_y als Mittelwerte von $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$. σ_x, σ_y sind die Standardabweichungen von diesem Mittelwert. ρ wird üblicherweise genutzt um den Korrelationskoeffizienten zu bezeichnen.

Korrelation

Um zu verstehen, wie die Korrelation die Verbindung zwischen zwei Variablen widerspiegelt können wir die Formel in ihre Bestandteile zerlegen:

- Gegeben wir schauen uns den i-ten Eintrag von x an. Dieser ist $\left(\frac{x_i-\mu_x}{\sigma_x}\right)$ Standardabweichungen entfernt vom Mittelwert von x.
- Weiterhin betrachten wir den i-ten Eintrag von y, welcher mit x_i verbunden ist. Dieser ist $\left(\frac{y_i-\mu_y}{\sigma_y}\right)$ Standardabweichungen vom Mittelwert von y entfernt.
- Gegeben x und y stehen in keiner Beziehung zueinander, dann ist das Produkt $\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right)$ im Durchschnitt 0. Das heißt bei unkorrelierten Zufallsvariablen ist der Korrelationskoeffizient 0.
- Wenn beide Variablen auf gleiche Weise variieren, dann ist der Korrelationskoeffizient positiv
- Wenn beide Variablen auf entgegengesetzte Weise variieren, dann ist der Korrelationskoeffizient negativ

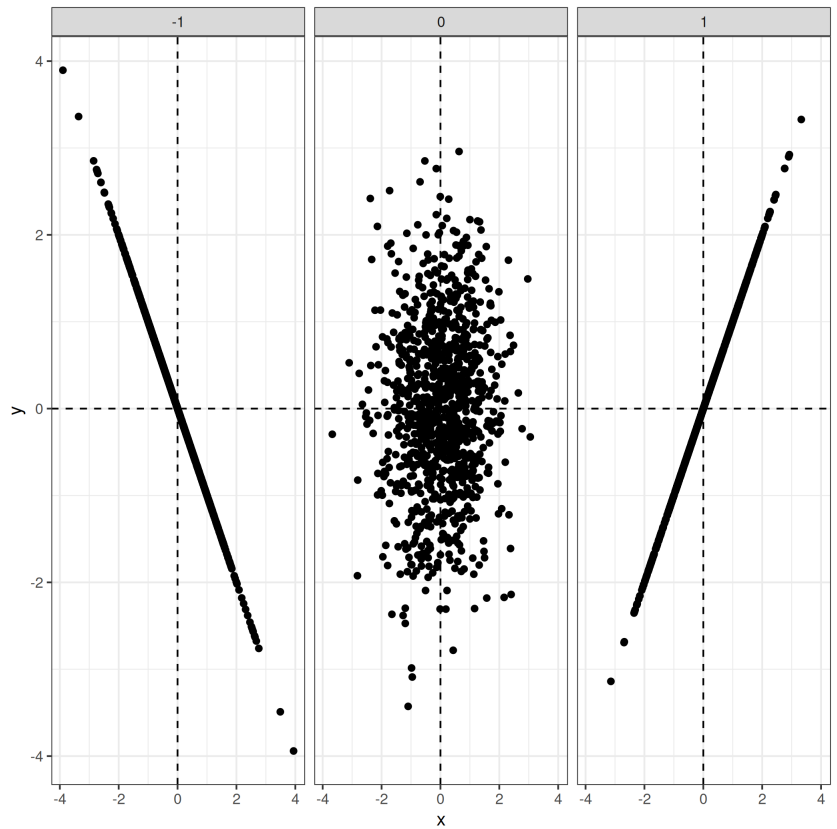
Korrelation

Im extrem variiert die Korrelation von zwei Variablen zwischen -1 und 1. Um dies zu sehen schauen wir uns den Fall von perfekter Korrelation an. D.h. wenn x um eine Einheit steigt, dann steigt gleichzeitig y um eine Einheit.

Hierbei ergibt sich die Korrelation als:

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x}\right)^2 = 1/\sigma^2 \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_x\right)^2 = 1$$

Korrelation - Simulation verschiedener Korrelationen



Korrelation

Die Korrelation zwischen der Attraktivität des Dozenten/der Dozentin und der Lehrevaluation liegt etwas über 18%. Der Zusammenhang ist schwach positiv.

```
used_evals %>% summarize(cor(bty_avg, score))
```

```
# A tibble: 1 x 1
  `cor(bty_avg, score)`
                  <dbl>
1                 0.187
```

Stichprobenkorrelation

In empirischen Arbeiten können wir leider sehr selten die Gesamtpopulation betrachten, sondern nur eine Stichprobe daraus. Deshalb sind die von uns berechneten Mittelwerte und Standardabweichungen auch immer Zufallsvariablen.

- Wir müssen in unseren Analysen die **Stichproben**korrelation ($\hat{\rho}$) als Schätzer für die Korrelation der Gesamtpopulation heranziehen
- Dadurch ergibt sich: Die **Stichprobenkorrelation** ist eine Zufallsvariable!

Stichprobenkorrelation

Das Verständnis der Stichprobenkorrelation kann gut durch die Stichprobengröße dargestellt werden.

Für das folgende Schaubild wurden aus einer Gleichverteilung zehn mal **unabhängig** zwei Zufallszahlen simuliert. Die Stichprobengröße wurde für jedes der Schaubilder variiert: Sie ist 10, 50, 100 und 1000 Beboachtungen groß.

Stichprobenkorrelation

Das Verständnis der Stichprobenkorrelation kann gut durch die Stichprobengröße dargestellt werden.

Für das folgende Schaubild wurden aus einer Gleichverteilung zehn mal **unabhängig** zwei Zufallszahlen simuliert. Die Stichprobengröße wurde für jedes der Schaubilder variiert: Sie ist 10, 50, 100 und 1000 Beboachtungen groß.

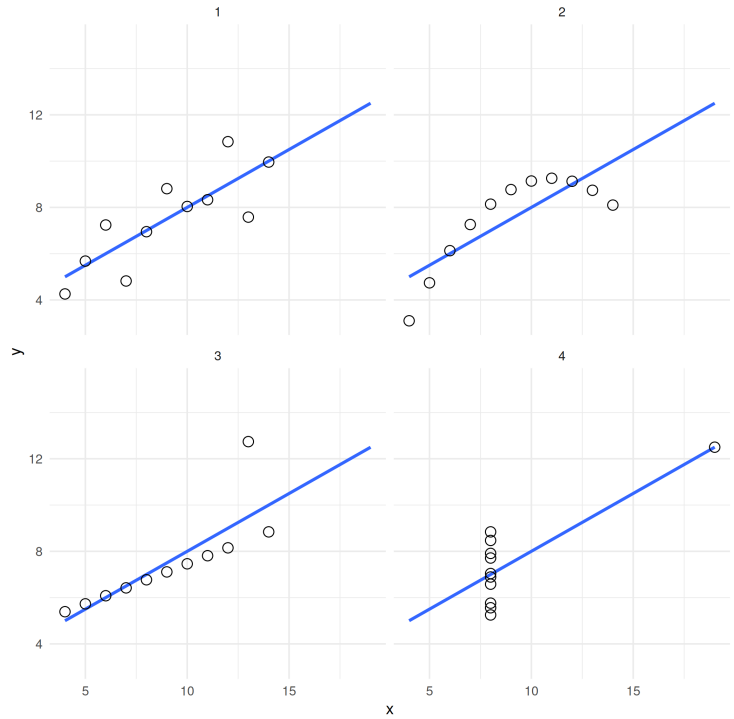
Insbesondere bei kleiner Stichprobenzahl scheint es öfter einen Zusammenhang zwischen den Variablen zu geben. Der Code auf der folgenden Folie wurde dazu genutzt die Stichprobenkorrelation als GIF zu veranschaulichen. Das Beispiel ist aus diesem hervorragenden Buch von Matthew J.C. Crump entnommen.

Stichprobenkorrelation

Anscombe-Quartett

Bei der Beschreibung ihres Datensatzes sollten Sie sich nicht nur auf deskriptive Statistiken, wie z.B. Mittelwert, Standardabweichung und Korrelation verlassen.

Anscombes Quartett:



Anscombe-Quartett

Vier verschiedene Stichproben mit gleicher Korrelation, Mittelwert und Standardabweichung.

Anscombe-Quartett

Vier verschiedene Stichproben mit gleicher Korrelation, Mittelwert und Standardabweichung.

> Machen Sie sich selbst ein Bild ob alle Datensätze gleich aussehen

Anscombe-Quartett

Vier verschiedene Stichproben mit gleicher Korrelation, Mittelwert und Standardabweichung.

> Machen Sie sich selbst ein Bild ob alle Datensätze gleich aussehen

> **Take-away:** Visualisieren Sie ihre Daten!

Explorative Grafik

```
used_evals %>%
  ggplot(aes(x = bty_avg, y = score)) +
  geom_point() +
  labs(x = "Attraktivität", y = "Lehrevaluations",
       title = "Zusammenhang zwischen Lehrevaluation und Attraktivität des Dozenten")
```