Von der Korrelation zur Regression

Zusammenhang zwischen Korrelation und Regression

Gegeben Sie wollen eine Zufallsvariable Y erklären, wobei Sie die Werte von $X = x$ auf einer Regressionslinie kennen:

- Für jede Standardabweichung σ_X, für die x über dem Durschnittswert μ_X liegt, steigt Y um ρ Standardabweichungen σ_Y über dem Durchschnittswert μ_Y.
- ρ ist die Korrelation von X und Y.

Die Formel für die Regressionsgeraden ergibt sich dann folgendermaßen:

$$\left(\frac{Y - \mu_Y}{\sigma_Y}\right) = \rho\left(\frac{x - \mu_X}{\sigma_X}\right)$$

Zusammenhang zwischen Korrelation und Regression

Die Formel für die Regressionsgerade können wir wie folgt umschreiben:

$$Y = \mu_Y + \rho \left(\frac{x - \mu_X}{\sigma_X} \right) \sigma_Y$$

- Bei einer Korrelation von 1 würden wir eine Steigerung um eine Standardabweichung von Y vorhersagen, gegeben, dass sich X um eine Standardabweichung verändert
- Bei einer Korrelation von 0 würden wir den Durchschnitt μ_Y vorhersagen
- Bei einer Korrelation <0 würden wir eine Reduktion anstatt einer Steigerung vorhersagen

Zusammenhang zwischen Korrelation und Regression

Auf unser Beispiel übertragen:

- Korrelationskoeffizent positiv, aber < 1
- Lehrevaluationsergebnisse liegen näher an den durchschnittlichen μ_X und sind nicht nur abhängig von der Attraktivität des Dozenten/der Dozentin x
- Regression zur Mitte: Die individuellen, möglicherweise sehr extremen Lehrevaluationsergebnisse werden im Gesamtverlauf ausgeglichen und bewegen sich zu den durchschnittlichen Lehrevaluationsergebnissen.

Zusammenhang zwischen Korrelation und Regression

Wenn Sie die Regressionsgerade auf Grundlage der Korrelation und ihren bisherigen Erkenntnissen ihrem Streudiagramm hinzufügen möchten, so müssen Sie die Formel des linearen Modells etwas umschreiben:

$$y = \alpha + \beta x$$

$$\text{mit der Steigung } \beta = \rho \frac{\sigma_y}{\sigma_x}$$

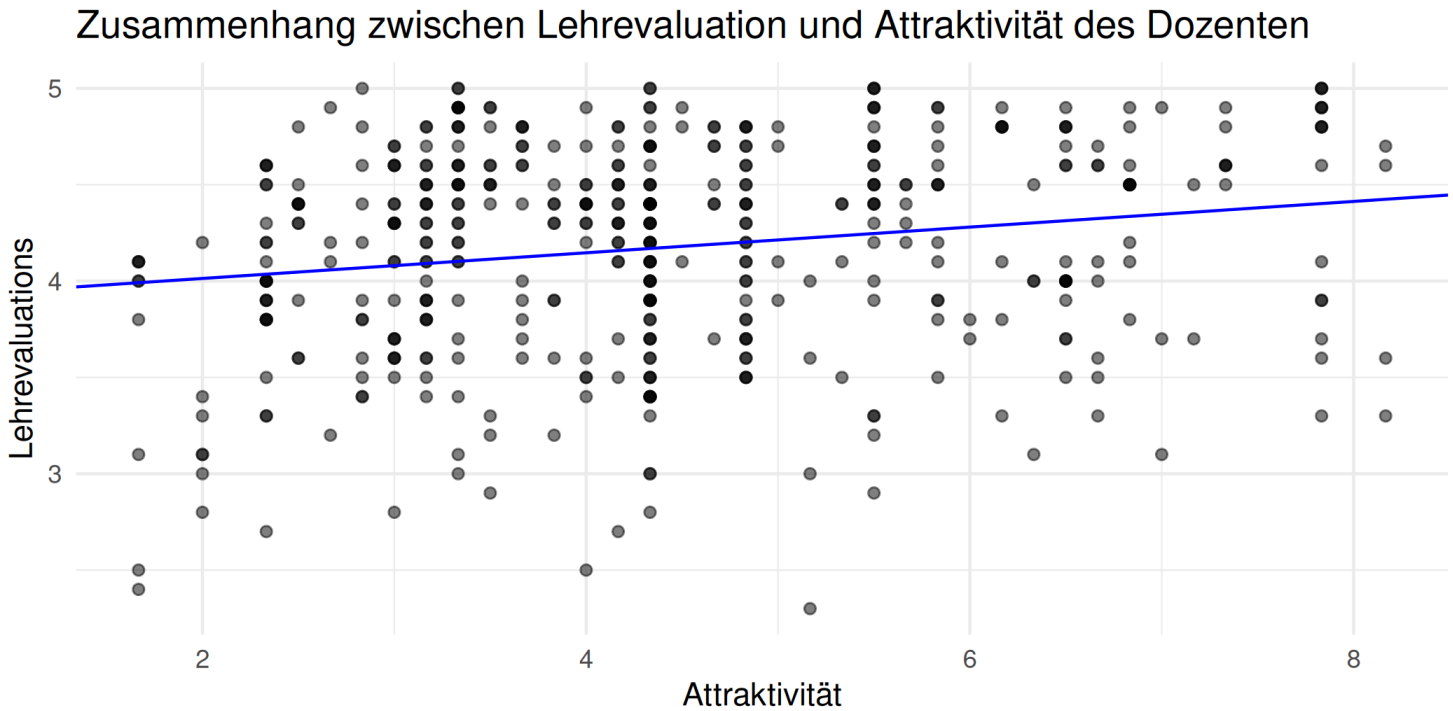
$$\text{und dem Achsenabschnitt } \alpha = \mu_y - \beta \mu_x$$

Zusammenhang zwischen Korrelation und Regression

In R sieht dies folgendermaßen aus:

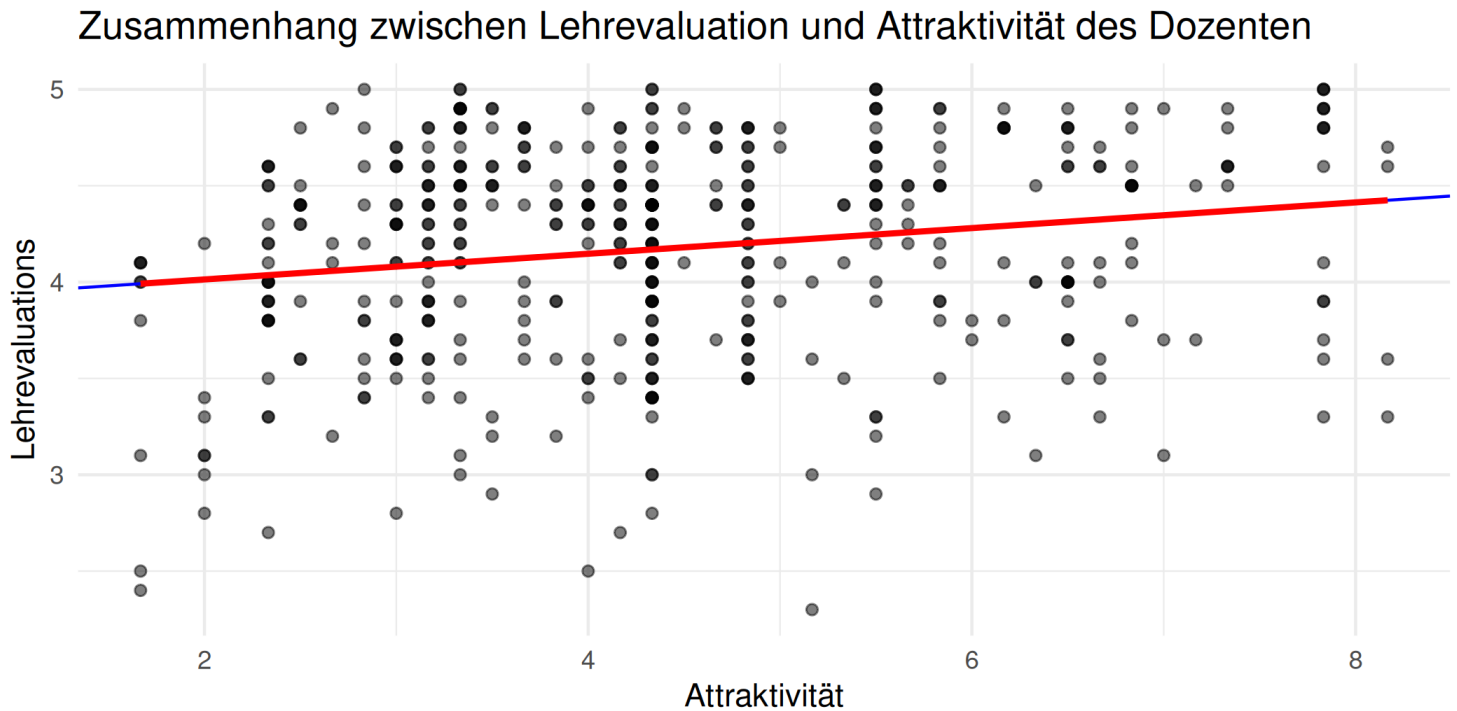
```
mu_x <- mean(used_evals$bty_avg)
mu_y <- mean(used_evals$score)
sd_x <- sd(used_evals$bty_avg)
sd_y <- sd(used_evals$score)
rho <- cor(used_evals$bty_avg, used_evals$score)

beta <-  rho * sd_y / sd_x
alpha <- mu_y - beta*mu_x
```

Zusammenhang zwischen Korrelation und Regression



Zusammenhang zwischen Korrelation und Regression

Zum selben Ergebnis gelangen wir mit der Regressionsgeraden berechnet nach der Methode der kleinsten Quadrate (in rot dazu):



Zusammenhang zwischen Korrelation und Regression

- Korrelation und Steigung der Regressionsgeraden haben immer das gleiche Vorzeichen
- **Jedoch:** Diese müsssen nicht immer den gleichen Wert haben (siehe Berechnung 2 Folien vorher)
- Die Regressionsgerade ist der beste lineare erwartungstreue Schätzer
  - Was bedeutet dies genau?

Lineare Regression

Lineares Modell

- Durch die lineare Regression können wir Zusammenhänge zwischen verschiedenen Variablen aufdecken und gleichzeitig für andere Faktoren *kontrollieren*
- "Lineare Modelle" werden so genannt, weil der bedingte Erwartungswert einer Variablen Y sich ergibt aus einer Linearkombination bekannter Größen

Wenn wir die Lehrevaluationsergebnisse verwenden können wir die N verfügbaren Attraktivitätseinschätzungen als $x_1, \ldots, x_n$ schreiben und dann die N Evaluationsergebnisse durch folgendes Modell erklären:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, N$$

- x_i ist hierbei die Einschätzung der Attraktivität der Dozenten/Dozentinnen
- Y_i ist das (zufällige) Lehrevaluationsergebnis, welches wir erklären wollen
- Annahmen:
  - Die ε_i sind unabhängig voneinander mit Erwartungswert 0
  - ε ist normalverteilt
  - Die Standardabweichung σ hängt nicht von i ab.

Lineares Modell

Lineare Modelle werden häufig verwendet:

- Die Koeffizienten von linearen Modellen sind direkt interpretierbar
- Beispiel Lehrevaluationsergebnisse:
  - Das Lehrevaluationsergebnis steigt je attraktiver ein Dozent/eine Dozentin eingeschätzt wird
  - ε fängt hierbei die Varianz in den Lehrevaluationsergebnissen auf
  - In ε stecken alle zusätzlichen Faktoren, welche die Lehrevaluationsergebnisse mit beeinflussen, aber in unserem Modell nicht gesondert enthalten sind
  - Bspw: Das Geschlecht, Alter, Rethorikfähigkeiten, eingesetzte didaktische Mittel ...

Kleinste Quadrate Schätzer

In unserem Modell wollen wir die Lehrevaluationsergebnisse vorhersagen. Hierfür benötigen wir eine Abschätzung der βs.

Um dies zu erreichen verwenden wir die Methode der kleinsten Quadrate. Hierbei wird versucht eine Regressionsgerade zu finden, welche den Abstand zwischen den einzelnen Datenpunkten und der Regressionsgeraden minimiert. Wir können dies mathematisch wie folgt darstellen, wobei RSS für die residual sum squares (Residuenquadratsumme) steht:

$$RSS = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Kleinste Quadrate Schätzer

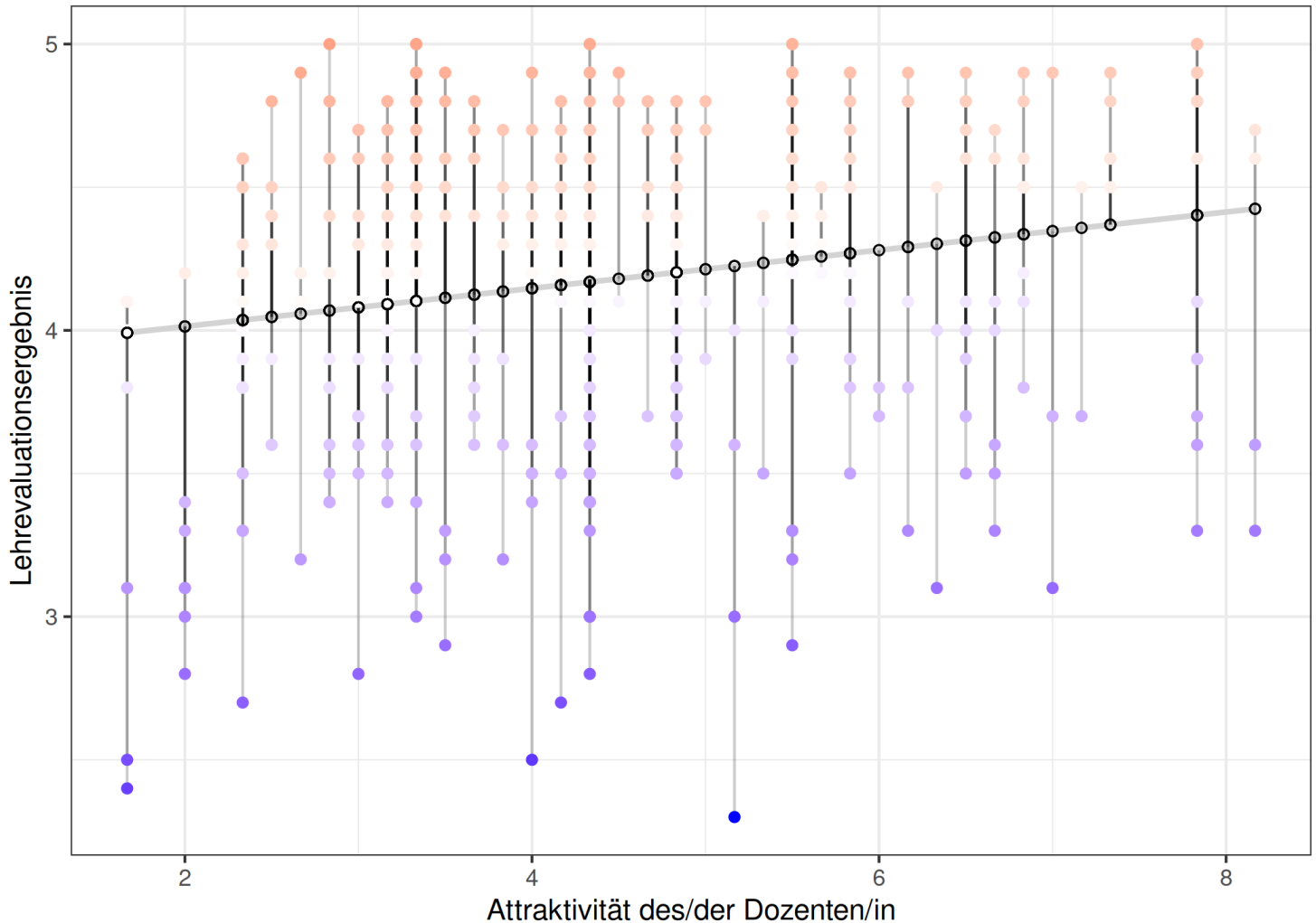
In unserem Modell wollen wir die Lehrevaluationsergebnisse vorhersagen. Hierfür benötigen wir eine Abschätzung der βs.

Um dies zu erreichen verwenden wir die Methode der kleinsten Quadrate. Hierbei wird versucht eine Regressionsgerade zu finden, welche den Abstand zwischen den einzelnen Datenpunkten und der Regressionsgeraden minimiert. Wir können dies mathematisch wie folgt darstellen, wobei RSS für die residual sum squares (Residuenquadratsumme) steht:

$$RSS = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 x_i)\}^2$$

- Die geschätzten Werte, welche die RSS minimieren bezeichnen wir mit $\hat{\beta}_0$ und $\hat{\beta}_1$

Kleinste Quadrate Schätzer



Schätzung

Wir können in R das lineare Modell mittels der Funktion `lm` berechnen:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

wobei Y_i das Lehrevaluationsergebnis des Dozenten und x_i dessen Attraktivität ist:

```
schätzer <- lm(score ~ bty_avg, data = used_evals)
```

- Mit der Tilde ~ (Alt Gr + +-Taste) zeigen wir der Funktion `lm`:
  - Links der ~: Variable, die wir vorhersagen wollen
  - Rechts der ~: Variable(n), die wir für die Vorhersage verwenden
  - R fügt automatisch einen Achsenabschnitt β_0 hinzu (falls Sie ein Modell *ohne* Achsenabschnitt berechnen möchten müssen Sie folgendes schätzen:
    - lm(score ~ bty_avg + 0, data = used_evals)

Schätzung

Um mehr über unsere Schätzung zu erfahren können wir die Funktion `summary` verwenden:

```
summary(schätzer)
```

```
Call:
lm(formula = score ~ bty_avg, data = used_evals)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9246 -0.3690  0.1420  0.3977  0.9309 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.88034    0.07614   50.96  < 2e-16 ***
bty_avg      0.06664    0.01629    4.09 5.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5348 on 461 degrees of freedom
Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293 
F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.083e-05
```

Interpretation der geschätzen Koeffizienten

Das Modell:

$$\hat{y} = \beta_0 + \beta_1 x$$
$$\hat{Lehrevaluation} = \beta_0 + \beta_1 * btyavg$$
$$\hat{Lehrevaluation} = 3.880 + 0.067 * btyavg$$

Interpretation der geschätzen Koeffizienten

Das Modell:

$$\hat{y} = \beta_0 + \beta_1 x$$
$$\hat{Lehrevaluation} = \beta_0 + \beta_1 * btyavg$$
$$\hat{Lehrevaluation} = 3.880 + 0.067 * btyavg$$

Interpretation der Koeffizienten:

- Achsenabschnitt (β_0) ist das durchschnittliche Lehrevaluationsergebnis, bei der die Attraktivität auf 0 eingeschätzt wurde
  - Mathematische Interpretation *jedoch* keine *praktische* Interpretation, da Attraktivität nicht mit 0 bewertet werden kann (Skala von 1 bis 10)
- Der Koeffizient für die Attraktivität (β_1) ist 0.067
  - Positiver Zusammenhang zwischen Attraktivität und Lehrevaluationsergebnis
  - Gleiches Vorzeichen wie Korrelation, jedoch unterschiedliche Werte
  - Korrelation $\rightarrow$ Stärke des linearen Zusammenhangs

Interpretation der geschätzen Koeffizienten

Dozenten mit einer Einheit höheren Attraktivität haben im Durchschnitt eine um 0.067 Einheiten bessere Lehrevaluation

- Wir sprechen bei der Interpretation des Koeffizienten der Attraktivitätsvariablen nur von einer Assoziation zwischen Attraktivität und Lehrevaluationsergebnis, **nicht** von einer kausalen Interpretation
- Folgende Einschätzung wäre **falsch**: Eine um eine Einheit höhere Attraktivität **führt** zu einer um 0.067 Einheiten besseren Lehrevaluation
- Es könnte durchaus sein, dass es weitere Variablen gibt, die sowohl die Attraktivität des Dozenten, als auch die Lehrevaluation beeinflussen, z.B. das Alter.
  - Nur weil zwei Variablen stark miteiander korrelieren bedeutet dies nicht, dass eine zur anderen führt.

→ Korrelation ist nicht gleich Kausalität

Interpretation der geschätzen Koeffizienten

Dozenten mit einer Einheit höheren Attraktivität haben im Durchschnitt eine um 0.067 Einheiten bessere Lehrevaluation

- Wir sprechen bei der Interpretation des Koeffizienten der Attraktivitätsvariablen nur von einer Assoziation zwischen Attraktivität und Lehrevaluationsergebnis, **nicht** von einer kausalen Interpretation
- Folgende Einschätzung wäre **falsch**: Eine um eine Einheit höhere Attraktivität **führt** zu einer um 0.067 Einheiten besseren Lehrevaluation
- Es könnte durchaus sein, dass es weitere Variablen gibt, die sowohl die Attraktivität des Dozenten, als auch die Lehrevaluation beeinflussen, z.B. das Alter.
  - Nur weil zwei Variablen stark miteiander korrelieren bedeutet dies nicht, dass eine zur anderen führt.

→ Korrelation ist nicht gleich Kausalität

- Weiterhin sprechen wir von einer Erhöhung der Lehrevaluationsergebnisse von *im Durchschnitt* 0.067 Einheiten

Die Funktion `lm`

- Die geschätzten Koeffizienten sind Zufallsvariablen
- Diese Zufallsvariablen haben eine Verteilung
- Die t-Statistik (`t value`) und p-Werte (`Pr(>|t|)`) basieren auf der Annahme, dass ε normalverteilt ist
- Dadurch ergibt sich für die t-Statistik:
  - $\hat{\beta}_0 / \hat{\mathrm{SE}}(\hat{\beta}_0)$ und $\hat{\beta}_1 / \hat{\mathrm{SE}}(\hat{\beta}_1)$ folgen einer **t-Verteilung** mit $N - p$ Freiheitsgraden
  - p ist die Anzahl an Parametern in unserem Modell (in unserem Fall $p = 2$)
  - die p-Werte testen ob $\beta_0 = 0$ bzw. ob $\beta_1 = 0$
  - Für große N nähert sich die t-Verteilung der Normalverteilung an

Schätzer sind Zufallsvariablen

Für jedes Lehrevaluationsergebnis können wir eine Vorhersage treffen ($\hat{Y}$), gegeben unserer Regressionsgeraden und dem bekannten Wert der Attraktivität des Dozenten/der Dozentin (x):

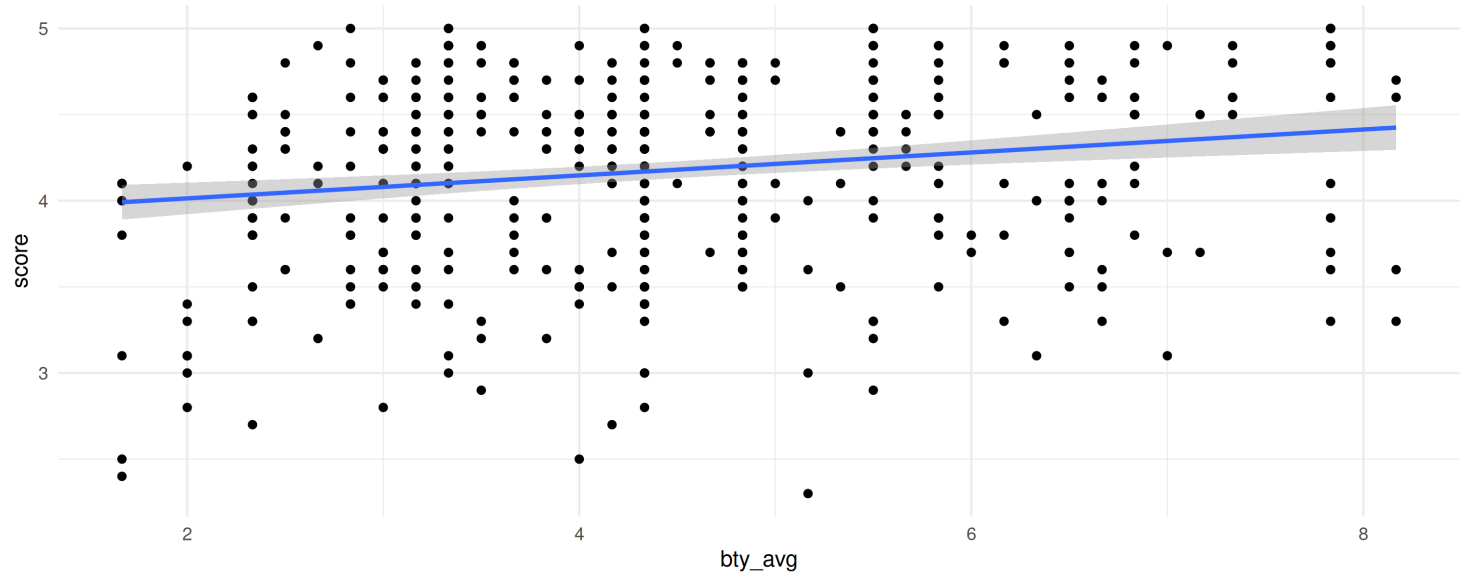
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Beachten Sie, dass $\hat{Y}$ eine Zufallsvariable ist, für welche Sie den Standardfehler bestimmen können. Wenn wir nun annehmen, dass die Standardfehler normalverteilt sind, so können wir Konfidenzintervalle für $\hat{Y}$ bilden.

In ggplot2 können wir diese Konfidenzintervalle um $\hat{Y}$ auch zeichnen (wir nutzen hier `geom_smooth(method = "lm")`)

Schätzer sind Zufallsvariablen

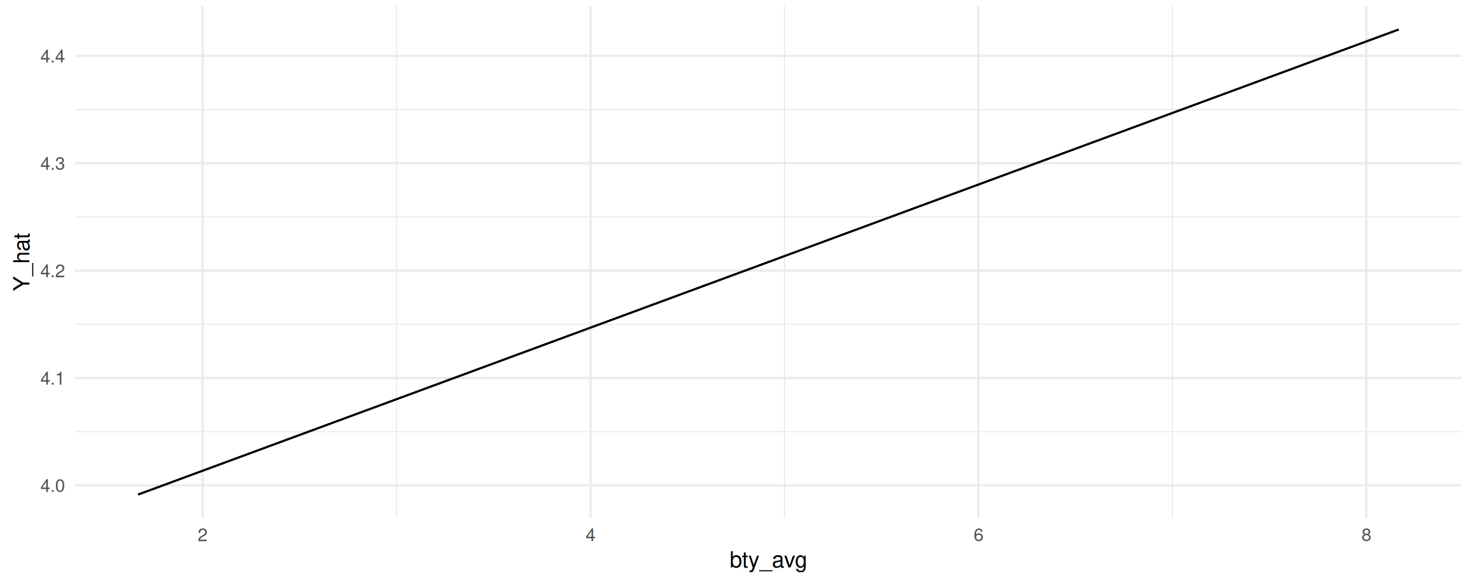
```
used_evals %>% ggplot(aes(bty_avg, score)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Schätzer sind Zufallsvariablen

Durch die R Funktion `predict` können die vorhergesagten Werte unserer Schätzung durch `lm` für jeden Punkt ausgegeben werden.

```
used_evals %>%
  mutate(Y_hat = predict(lm(score ~ bty_avg, data = .))) %>%
  ggplot(aes(bty_avg, Y_hat)) +
  geom_line()
```



Das Paket broom

In dieser Vorlesung nutzen wir Pakete des `tidyverse` Universums um unsere Datenanalyse durchzuführen. Jedoch sind sehr viele Funktionen in R nicht teil des `tidyverse`, wie z.B. die `lm` Funktion um eine lineare Regression durchzuführen.

Durch das Paket `broom` und deren Funktionen `tidy`, `glance` und `augment` können wir die Ergebnisse von Funktionen wie `lm` in das uns bekannte `tidyverse` überführen.

Das Paket broom

Die Funktion `tidy` gibt die Ergebnisse aus `lm` als Dataframe wieder:

```
library(broom)
library(janitor)
schätzer <- lm(score ~ bty_avg, data = used_evals)
tidy(schätzer) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 2 x 5
  term        estimate std_error statistic p_value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    3.88      0.076     51.0        0
2 bty_avg        0.067     0.016      4.09       0
```

Das Paket broom

Hier können wir auch andere Informationen wie Konfidenzintervalle ausgeben lassen:

```
tidy(schätzer, conf.int = TRUE) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 2 x 7
  term        estimate std_error statistic p_value conf_low conf_high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)    3.88      0.076     51.0        0    3.73      4.03
2 bty_avg        0.067     0.016      4.09       0    0.035     0.099
```

Multiple lineare Regression

Einführung

- Statt nur eine erklärende Variable ins Modell aufzunehmen können auch mehrere erklärende Variablen hinzugenommen werden, bspw. könnten wir ein Modell schätzen mit:
  - Alter (nummerisch)
  - Geschlecht (kategorisch)

> Bekommen ältere oder jüngere Dozenten/innen bessere Lehrevaluationen und unterscheidet sich dies nach Geschlecht?

Deskriptive Analysen

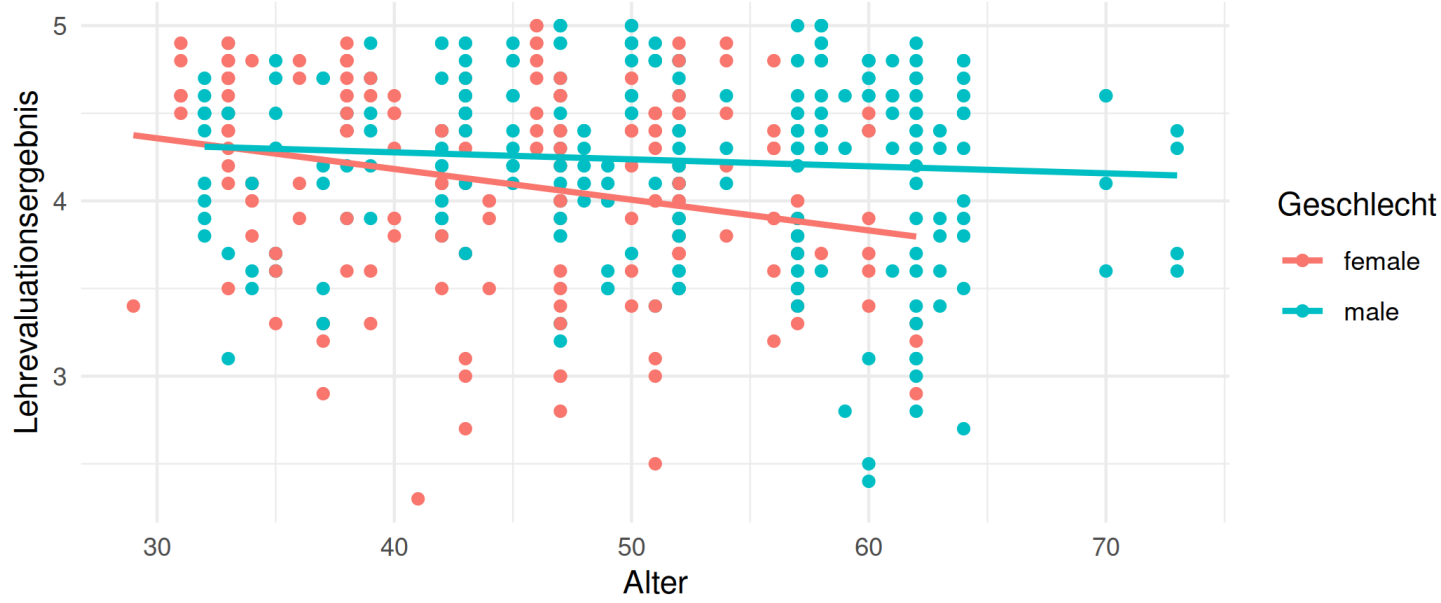
- Korrelationskoeffizient für Alter und Lehrevaluation

```
used_evals %>%
  summarize(cor(score,age)) %>%
  pull()
```

```
[1] -0.107032
```

Explorative Grafiken

```
used_evals %>%
  ggplot(aes(x = age, y = score, color = gender)) +
  geom_point() +
  labs(x = "Alter", y = "Lehrevaluationsergebnis", color = "Geschlecht") +
  geom_smooth(method = "lm", se = FALSE)
```



Multiple lineare Regression

```
basismodell <- lm(score ~ age + gender, data = used_evals)
tidy(basismodell) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 3 x 5
  term        estimate std_error statistic p_value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    4.48      0.125     35.8    0
2 age           -0.009     0.003     -3.28   0.001
3 gendermale     0.191     0.052      3.63   0
```

Multiple lineare Regression

```
basismodell <- lm(score ~ age + gender, data = used_evals)
tidy(basismodell) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 3 x 5
  term        estimate std_error statistic p_value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    4.48      0.125     35.8    0
2 age           -0.009     0.003     -3.28   0.001
3 gendermale     0.191     0.052      3.63   0
```

- Ein um ein Jahr älterer Dozent/Dozentin hat im Durchschnitt eine um 0.009 Einheiten schlechtere Lehrevaluation (`age`)
  - Signifikant auf dem 1% Signifikanzniveau
- Männliche Dozenten haben im Durchschnitt eine um 0.191 Einheiten bessere Lehrevaluation (`gendermale`)
  - Signifikant auf dem 1% Signifanzniveau

Multiple lineare Regression

```
basismodell <- lm(score ~ age + gender, data = used_evals)
tidy(basismodell) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 3 x 5
  term        estimate std_error statistic p_value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    4.48      0.125     35.8    0
2 age           -0.009     0.003     -3.28   0.001
3 gendermale     0.191     0.052      3.63   0
```

- Ein um ein Jahr älterer Dozent/Dozentin hat im Durchschnitt eine um 0.009 Einheiten schlechtere Lehrevaluation (`age`)
  - Signifikant auf dem 1% Signifikanzniveau
- Männliche Dozenten haben im Durchschnitt eine um 0.191 Einheiten bessere Lehrevaluation (`gendermale`)
  - Signifikant auf dem 1% Signifanzniveau

> Gibt es unterschiedliche Effekte für Männer und Frauen über das Alter hinweg? Wie könnte das gemessen werden?

Interaktionsmodell

Unsere explorative Grafik deutet zwei unterschiedliche Kurveverläufe an. Durch ein Interaktionsmodell können wir dem Phänomen Rechnung tragen

Interaktionsmodell

Unsere explorative Grafik deutet zwei unterschiedliche Kurveverläufe an. Durch ein Interaktionsmodell können wir dem Phänomen Rechnung tragen

```
interaktionsmodell <- lm(score ~ age * gender, data = used_evals)
tidy(interaktionsmodell, conf.int = TRUE) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 4 x 7
  term           estimate std_error statistic p_value conf_low conf_high
  <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)       4.88      0.205     23.8    0        4.48      5.29
2 age              -0.018     0.004     -3.92   0       -0.026    -0.009
3 gendermale       -0.446     0.265     -1.68   0.094   -0.968     0.076
4 age:gendermale    0.014     0.006      2.45   0.015    0.003     0.024
```

Interpretation der Koeffizienten

- Frauen bilden hier die Basisgruppe, da in unserer kategorischen Variable `gender` female vor male kommt und damit automatisch als Basisgruppe deklariert wird
- Der Achsenabschnitt ist hier *nur* für Frauen
  - Entspricht der roten Linie im vorherigen Schaubild
  - Steigung der roten Linie ist -0.018 im vorherigen Schaubild
- Männer werden hier als Vergleich zu den Frauen berechnet.
  - Achsenabschnitt für Männer $\rightarrow$ 4.833 - 0.446 = 4.437
  - Steigung im vorherigen Schaubild für Männer wäre dann entsprechend $\rightarrow$ -0.018 + 0.014 = -0.004

Interpretation der Koeffizienten

In einer Tabelle zusammengefasst bedeutet dies:

Geschlecht	Achsenabschnitt	Steigung
Frauen	4.833	-0.018
Männer	4.437	-0.004

Interpretation der Koeffizienten

In einer Tabelle zusammengefasst bedeutet dies:

Geschlecht	Achsenabschnitt	Steigung
Frauen	4.833	-0.018
Männer	4.437	-0.004

Das heißt die Lehrevaluationsergebnisse sind im Durchschnitt bei älteren Frauen pro Lebensjahr um -0.018 Einheiten schlechter, bei Männern nur um -0.004 Einheiten.

→ Das Alter ist bei Frauen im Durchschnitt mit einem **höheren negativen Effekt** auf die Lehrevaluationsergebnisse assoziiert.

Aufteilen der Stichprobe

Anstatt einen Interaktionsterm einzuführen können Sie die Stichprobe auch aufteilen:

```
split1 <- lm(score ~ age, data = filter(used_evals, gender=="female"))
split2 <- lm(score ~ age, data = filter(used_evals, gender=="male"))

#Alternativ
used_evals %>%
  group_by(gender) %>%
  do(tidy(lm(score ~ age, data = .), conf.int = TRUE)) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 4 x 8
# Groups:   gender [2]
  gender term        estimate std_error statistic p_value conf_low conf_high
  <fct>  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 female (Intercept)    4.88      0.21      23.2    0         4.47      5.30
2 female age           -0.018     0.005     -3.82   0        -0.027    -0.008
3 male   (Intercept)    4.44      0.165     26.9    0         4.11      4.76
4 male   age           -0.004     0.003     -1.25   0.212    -0.01      0.002
```

Aufteilen der Stichprobe

Anstatt einen Interaktionsterm einzuführen können Sie die Stichprobe auch aufteilen:

```
split1 <- lm(score ~ age, data = filter(used_evals, gender=="female"))
split2 <- lm(score ~ age, data = filter(used_evals, gender=="male"))

#Alternativ
used_evals %>%
  group_by(gender) %>%
  do(tidy(lm(score ~ age, data = .), conf.int = TRUE)) %>%
  mutate_if(is.numeric, round, digits = 3) %>%
  clean_names()
```

```
# A tibble: 4 x 8
# Groups:   gender [2]
  gender term        estimate std_error statistic p_value conf_low conf_high
  <fct>  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 female (Intercept)    4.88      0.21      23.2    0         4.47      5.30
2 female age           -0.018     0.005     -3.82   0        -0.027    -0.008
3 male   (Intercept)    4.44      0.165     26.9    0         4.11      4.76
4 male   age           -0.004     0.003     -1.25   0.212    -0.01      0.002
```

Die Koeffizienten sind die Selben wie wir sie beim Interaktionsmodell erhalten haben

Das Paket broom

Falls Sie mit den Werten aus der Regression weiterarbeiten möchten können Sie auch für unterschiedliche Gruppen einzelne Regressionen durchführen lassen. Hier hilft ihnen der `do`-Befehl aus dem `broom` Paket:

```
used_evals %>%
  group_by(gender) %>%
  do(tidy(lm(score ~ age, data = .), conf.int = TRUE))
```

```
# A tibble: 4 x 8
# Groups:   gender [2]
  gender term        estimate std.error statistic  p.value conf.low conf.high
  <fct>  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 female (Intercept)  4.88      0.210       23.2  9.90e-58   4.47     5.30
2 female age         -0.0175    0.00459     -3.82 1.79e- 4  -0.0266  -0.00848
3 male   (Intercept)  4.44      0.165       26.9  9.09e-78   4.11     4.76
4 male   age         -0.00399   0.00319     -1.25 2.12e- 1  -0.0103   0.00229
```

Das Paket broom

Diesen Dataframe können wir anschließend nach den Regressionskoeffizienten von Interesse filtern und uns nur die Spalten ausgeben lassen, welche uns interessieren:

```
used_evals %>%
  group_by(gender) %>%
  do(tidy(lm(score ~ age, data = .), conf.int = TRUE)) %>%
  filter(term == "age") %>%
  select(gender, estimate, conf.low, conf.high)
```

```
# A tibble: 2 x 4
# Groups:   gender [2]
  gender estimate conf.low conf.high
  <fct>     <dbl>    <dbl>     <dbl>
1 female -0.0175   -0.0266  -0.00848
2 male   -0.00399  -0.0103   0.00229
```

Multiple lineare Regression

Es besteht auch die Möglichkeit unterschiedliche Regressionsspezifikationen mit dem Paket `stargazer` einander gegenüberzustellen.

Code für die Verwendung von `stargazer`, wobei die jeweiligen Modelle auf den vorherigen Folien berechnet und unter den entsprechenden Namen abgespeichert wurden.

```
library(stargazer)
stargazer(basismodell, interaktionsmodell, split1, split2, type="html")
```

	Dependent variable:			
	score			
	(1)	(2)	(3)	(4)
age	-0.009***	-0.018***	-0.018***	-0.004
	(0.003)	(0.004)	(0.005)	(0.003)
gendermale	0.191***	-0.446*		
	(0.052)	(0.265)		
age:gendermale		0.014**		
		(0.006)		
Constant	4.484***	4.883***	4.883***	4.437***
	(0.125)	(0.205)	(0.210)	(0.165)
Observations	463	463	195	268
R^2	0.039	0.051	0.070	0.006
Adjusted R^2	0.035	0.045	0.066	0.002
Residual Std. Error	0.534 (df = 460)	0.531 (df = 459)	0.545 (df = 193)	0.521 (df = 266)
F Statistic	9.338*** (df = 2; 460)	8.288*** (df = 3; 459)	14.598*** (df = 1; 193)	1.564 (df = 1; 266)
Note:				*p<0.1; **p<0.05; ***p<0.01