
Case-Study zur Arbeitslosigkeit in Deutschland

Organisatorische Hinweise

- Bis zum 08.05.2025 muss der Fragebogen für die Gruppenzusammenstellung ausgefüllt werden (alle die daran teilnehmen sind auch bei den Projekten dabei)
- Bis zum 12.05.2025 ist das 6. R-Tutor Problem Set auf Moodle hochzuladen (nur die .sub-Datei!)
- In KW 20 (ab 15.05.2025) gibt es die Kick-off Tutorien zum Kennenlernen ihrer Gruppen
- Freitag den 16.05.2025 wird das 4. Projekt (des Gesamtkurses) für jede Gruppe auf Github hochgeladen
- Freitag den 16.05.2025 wird das 4. Projekt besprochen (Vorstellung und Interpretationen)
- Bearbeitungszeit des 4. Projekts: Bis 22.05.2025
- Freitag den 23.05.2025 gibt es die Probeklausur im Hörsaal TTU ab 10:15 Uhr. Teil der Vorleistung, mind. 20% der Punkte müssen bestanden werden, keine Hilfmittel!
- Freitag 23.05.2025 wird das 5. Projekt vorgestellt (nach der Probeklausur)

Recap der Vorlesungsinhalte

- Wir hatten die Wahrscheinlichkeitstheorie und die Normalverteilung besprochen
- Wir hatten über die Stichprobenvarianz, Standardfehler und Konfidenzintervalle gesprochen
- Wir hatten einen Hypothestest durchgeführt
- Wir hatten die Korrelation von zwei Variablen und die lineare Regression betrachtet
- Anschließend sind wir in der multiplen linearen Regression auch auf Interaktionsterme eingegangen

Empirische Analyse unserer Case-Study

Induktive Statistik

- Interesse gilt nicht dem Datensatz selbst, sondern der Population
  - Sie haben keine Vollerhebung durchgeführt, sondern nur eine (zufällige) Stichprobe der Population gezogen
- **Beispiel:** Mikrozensus, d.h. eine Befragung von zufällig ausgewählten Haushalten in Deutschland
- Sie wollen aus der Stichprobe schätzen, wie sich die beobachtete Größe in der Population verhält
- Es gibt viele Arten der induktiven Statistik. Die zwei häufigsten:
  - Vorhersage
  - Erkennen kausaler Zusammenhänge

Bereiche der induktiven Statistik

- Stichprobentheorie
  - Güte der Stichprobe; Wichtig um repräsentative Ergebnisse zu erhalten
- Schätztheorie
  - Punktschätzer und Konfidenzintervalle
- Testtheorie
  - Hypothesentest, lineare Regression

Wie sieht die induktive Statistik in der Praxis aus?

Dritter Teil der Case Study

Daten aus der Case-Study, welche wir im vorherigen Semester eingelesen und deskriptiv analysiert haben wollen wir nun mittels der induktiven Statistik näher untersuchen.

- Erster Schritt: Kurzer Recap mittels bivariater deskriptiver Statistik um den Zusammenhang verschiedener Variablen darzustellen
- Zweiter Schritt: (Multiple) lineare Regression der Daten um herauszufinden, welche Faktoren die Arbeitslosenquote in den deutschen Landkreisen treibt
  - Darstellung mit dem Paket `modelsummary` + `kableExtra`
  - Sehr gute Vignette des Pakets `modelsummary`

Dritter Teil der Case Study

Daten aus der Case-Study, welche wir im vorherigen Semester eingelesen und deskriptiv analysiert haben wollen wir nun mittels der induktiven Statistik näher untersuchen.

- Erster Schritt: Kurzer Recap mittels bivariater deskriptiver Statistik um den Zusammenhang verschiedener Variablen darzustellen
- Zweiter Schritt: (Multiple) lineare Regression der Daten um herauszufinden, welche Faktoren die Arbeitslosenquote in den deutschen Landkreisen treibt
  - Darstellung mit dem Paket `modelsummary` + `kableExtra`
  - Sehr gute Vignette des Pakets `modelsummary`

Ziele des dritten Teils der Case Study:

- (Multiple) lineare Regression und Interpretation der Koeffizienten
- Interaktionsterme
- Besprechen der Kausalität

Dritter Teil der Case Study

Daten aus der Case-Study, welche wir im vorherigen Semester eingelesen und deskriptiv analysiert haben wollen wir nun mittels der induktiven Statistik näher untersuchen.

- Erster Schritt: Kurzer Recap mittels bivariater deskriptiver Statistik um den Zusammenhang verschiedener Variablen darzustellen
- Zweiter Schritt: (Multiple) lineare Regression der Daten um herauszufinden, welche Faktoren die Arbeitslosenquote in den deutschen Landkreisen treibt
  - Darstellung mit dem Paket `modelsummary` + `kableExtra`
  - Sehr gute Vignette des Pakets `modelsummary`

Ziele des dritten Teils der Case Study:

- (Multiple) lineare Regression und Interpretation der Koeffizienten
- Interaktionsterme
- Besprechen der Kausalität

Im vierten RTutor Problem Set beschäftigen Sie sich auch mit der linearen Regression zu einzelnen Ländern auf europäischer Ebene und im 5. und 6. Problem Set geht es um die Kausalität.

Daten und Pakete laden

Wir laden die aus Teil 1 der Case-Study erstellten Datensätze:

```
library(tidyverse)
library(modelsummary)
library(kableExtra)
library(corrr)
```

```
# Daten einlesen
bip_zeitreihe <- readRDS("../case-study/data/bip_zeitreihe.rds")
gesamtdaten <- readRDS("../case-study/data/gesamtdaten.rds")

# Zuerst wollen wir die Arbeitslosenquote, einen Dummy für Ostdeutschland und die Verschuldung im Verhältnis zum BIP pro Landkreisberechnen
gesamtdaten <- gesamtdaten |>
  mutate(alo_quote = (total_alo / (erw+total_alo))*100,
         ost = as.factor(ifelse(bundesland_name %in% c("Brandenburg", "Mecklenburg-Vorpommern", "Sachsen", "Sachsen-Anhalt", "Thüringen"), 1, 0)),
         ost_name = ifelse(ost == 1, "Ostdeutschland", "Westdeutschland"),
         anteil_schulden = (Schulden_gesamt / bip)*100)

bip_wachstum <- bip_zeitreihe |>
  filter( nchar(Regionalschluessel) == 5) |>
  group_by(Regionalschluessel) |>
  arrange(Jahr) |>
  mutate( bip_wachstum = 100*(bip - lag(bip)) / bip ) |>
  ungroup() |>
  filter( Jahr == 2022 ) |>
  select(Regionalschluessel, bip_wachstum, Jahr)

gesamtdaten <- left_join(gesamtdaten, bip_wachstum, by = "Regionalschluessel")
```

Bivariate deskriptive Analysen (Korrelationen)

Korrelation zwischen den einzelnen Variablen

Wir hatten uns im letzten Semester bereits die Korrelation der einzelnen Variablen angeschaut und wollen diese Korrelationen noch einmal aufgreifen:

Korrelation zwischen den einzelnen Variablen

Wir hatten uns im letzten Semester bereits die Korrelation der einzelnen Variablen angeschaut und wollen diese Korrelationen noch einmal aufgreifen:

> Bevor wir uns der Regressionsanalyse zuwenden schauen wir uns den Zusammenhang der unterschiedlichen Variablen erst visuell noch einmal an.

- Wie hoch ist die Korrelation zwischen Arbeitslosenquote und BIP Wachstum?
- Wie hoch ist sie zwischen Arbeitslosenquote und dem Anteil der Schulden?
- Und schlussendlich: Wie hoch ist die Korrelation zwischen dem BIP Wachstum und dem Anteil der Schulden?

Korrelation zwischen den einzelnen Variablen

Wir hatten uns im letzten Semester bereits die Korrelation der einzelnen Variablen angeschaut und wollen diese Korrelationen noch einmal aufgreifen:

Bevor wir uns der Regressionsanalyse zuwenden schauen wir uns den Zusammenhang der unterschiedlichen Variablen erst visuell noch einmal an.

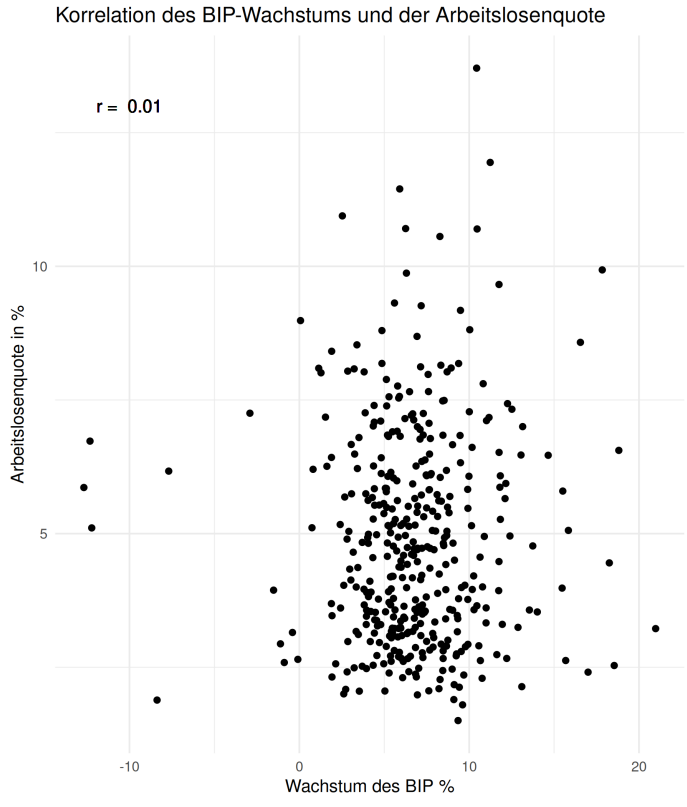
- Wie hoch ist die Korrelation zwischen Arbeitslosenquote und BIP Wachstum?
- Wie hoch ist sie zwischen Arbeitslosenquote und dem Anteil der Schulden?
- Und schlussendlich: Wie hoch ist die Korrelation zwischen dem BIP Wachstum und dem Anteil der Schulden?

Hierdurch bekommen wir einen ersten Eindruck der Daten und werden auf mögliche Probleme aufmerksam, wie z.B. Multikolliniarität.

Korrelation zwischen der Arbeitslosenquote und dem BIP Wachstum

```
cor_alo_bip <- cor(gesamtdaten$alo_quote,
    gesamtdaten$bip_wachstum,
    use = "pairwise.complete.obs")

gesamtdaten |>
  ggplot(aes(x = bip_wachstum, y = alo_quote)) +
  geom_point() +
  labs( x = "Wachstum des BIP %",
        y = "Arbeitslosenquote in %",
        title = "Korrelation des BIP-Wachstums und
  theme_minimal() +
  geom_text(x = 0.02, y =13, label = paste("r = ",
```



Korrelation zwischen der Arbeitslosenquote und dem Anteil der Schulden

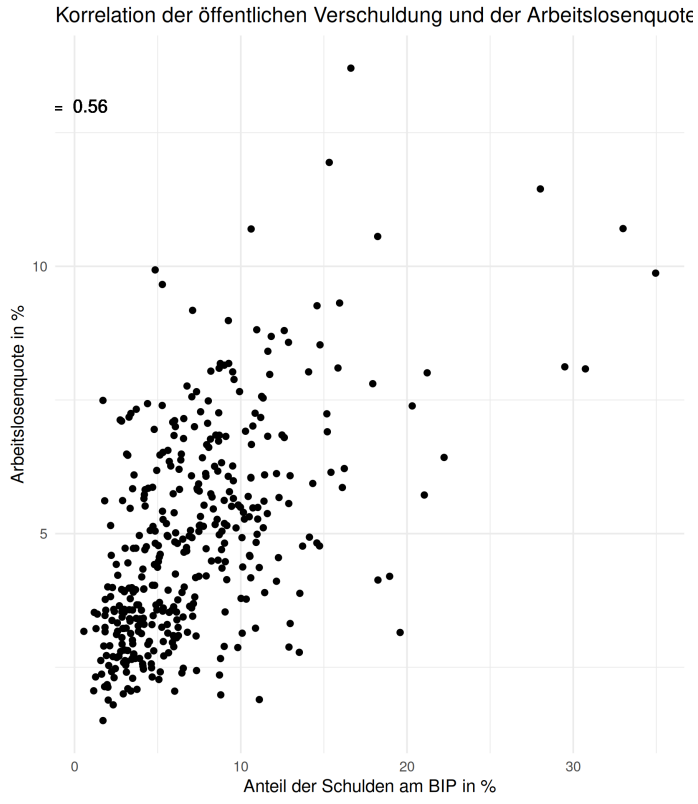
```
cor_alo_verschuldung <- cor(gesamtdaten$alo_quote,

gesamtdaten |>
  ggplot(aes(x = anteil_schulden, y = alo_quote)) +
  geom_point() +
  labs( x = "Anteil der Schulden am BIP in %",
        y = "Arbeitslosenquote in %",
        title = "Korrelation der öffentlichen Versc
  theme_minimal() +
  geom_text(x = 0.02, y =13, label = paste("r = ",
```

Korrelation zwischen der Arbeitslosenquote und dem Anteil der Schulden

```
cor_alo_verschuldung <- cor(gesamtdaten$alo_quote,

gesamtdaten |>
  ggplot(aes(x = anteil_schulden, y = alo_quote)) +
  geom_point() +
  labs( x = "Anteil der Schulden am BIP in %",
        y = "Arbeitslosenquote in %",
        title = "Korrelation der öffentlichen Versc
  theme_minimal() +
  geom_text(x = 0.02, y =13, label = paste("r = ",
```



Korrelationsmatrix

```
korrelationen <- gesamtdaten |>
  select(bip_wachstum, anteil_schulden, alo_quote) |>
  correlate() |> # Korrelationen erzeugen
  rearrange() |>  # Sortieren nach Korrelation
  shave() # Oberen Teil der Tabelle abschneiden

# Formatieren und anzeigen mit kableExtra
korrelationen |>
  fashion() |>     # Schönes Format mit zwei Nachkommastellen
  kbl(caption = "Korrelationsmatrix ausgewählter Variablen",
      align = "lrrr",
      col.names = c("Variable", "bip_wachstum", "alo_quote", "anteil_schulden")) |>
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover", "condensed"))
```

Korrelationsmatrix ausgewählter Variablen

Variable	bip_wachstum	alo_quote	anteil_schulden
bip_wachstum			
alo_quote	.01		
anteil_schulden	-.17	.56	

Interpretation der Korrelation

- Hat an sich keine intuitive quantitative Interpretation
- Ist eine univariate Repräsentation des Zusammenhangs zweier Variablen
- Kann dabei helfen stark korrelierte Variablen im Datensatz aufzuzeigen
  - Dies ist für eine spätere lineare Regression wichtig
  - Stichwort Multikollinearität

Interpretation der Korrelation

- Hat an sich keine intuitive quantitative Interpretation
- Ist eine univariate Repräsentation des Zusammenhangs zweier Variablen
- Kann dabei helfen stark korrelierte Variablen im Datensatz aufzuzeigen
  - Dies ist für eine spätere lineare Regression wichtig
  - Stichwort Multikollinearität

In empirischen Arbeiten wird meist auf die lineare Regression zurückgegriffen und nicht auf die Analyse von Korrelationen:

- Schätzer aus der linearen Regression sind BLUE (best linear unbiased estimator)
- Wir können auf mehrere Variablen kontrollieren in der linearen Regression

Einfache lineare Regression

Lineare Regression

Zur weiteren Analyse wollen wir uns der linearen Regression bedienen:

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

Wobei wir die Arbeitslosenquote (y_i) auf das BIP Wachstum (x_i) regressieren.

Arbeitslosenquote auf das BIP Wachstum regressieren

```
bip <- lm(alo_quote ~ bip_wachstum, data = gesamtdaten)

modelsummary(
  list(bip, bip, bip),
  type = "html",
  fmt = 2,
  estimate = c(
    "{estimate}",
    "{estimate}{stars}",
    "{estimate} [{conf.low}, {conf.high}]"
  ),
  coef_rename = c(
    "bip_wachstum" = "BIP-Wachstum",
    "(Intercept)" = "Konstante"
  ),
  gof_omit = "DF|Deviance|Log.Lik.|F|RMSE|AIC|BIC",
  title = "Arbeitslosigkeit auf BIP-Wachstum"
)
```

Arbeitslosenquote auf das BIP Wachstum regressieren

Arbeitslosigkeit auf BIP-Wachstum

	(1)	(2)	(3)
Konstante	4.84	4.84***	4.84 [4.44, 5.24]
	(0.20)	(0.20)	(0.20)
BIP-Wachstum	0.01	0.01	0.01 [−0.04, 0.06]
	(0.03)	(0.03)	(0.03)
Num.Obs.	398	398	398
R2	0.000	0.000	0.000
R2 Adj.	−0.002	−0.002	−0.002

Arbeitslosenquote auf das BIP Wachstum regressieren

```
bip <- lm(alo_quote ~ bip_wachstum, data = gesamtdaten)

modelsummary(bip,
          type = "html",
          statistic = 'conf.int',
          conf_level = .99,
          fmt = 2,
          gof_omit = 'DF|Deviance|Log.Lik.|F|RMSE|AIC|BIC',
          coef_rename = c("bip_wachstum" = "BIP-Wachstum", "(Intercept)" = "Konstante")
          )
```

Arbeitslosenquote auf das BIP Wachstum regressieren

	(1)
Konstante	4.84
	[4.32, 5.37]
BIP-Wachstum	0.01
	[−0.06, 0.08]
Num.Obs.	398
R2	0.000
R2 Adj.	−0.002

Erkenntnisse aus der Regressionstabelle

- 398 Beobachtungen
- R^2 mit 0.000 sehr klein
  - R^2 kann künstlich nach oben getrieben werden, darum besser *adjusted* R^2 anschauen
- R^2 ist irrelevant wenn wir unsere Schätzer kausal interpetieren wollen
  - R^2 misst die Variation in y, diese wollen wir aber gar nicht erklären, sondern ob x einen kausalen Einfluss auf y hat!
- R^2 ist wichtiger bei Vorhersagen
  - Bei Vorhersagen möchten wir nach Möglichkeit y so gut es geht erklären.
- Bei Zeitreihendaten ist das R^2 tendenziell immer höher als bei Querschnitts- oder Paneldaten

Bitte fixieren Sie sich in ihrer Interpretation nicht auf das R^2!

Erkenntnisse aus der Regressionstabelle

Interessanter: Der geschätze Koeffizient zum `BIP-Wachstum` in Höhe von 0,01.

> Wie kann dieser Koeffizient interpretiert werden?

Erkenntnisse aus der Regressionstabelle

Interessanter: Der geschätze Koeffizient zum `BIP-Wachstum` in Höhe von 0,01.

> Wie kann dieser Koeffizient interpretiert werden?

> Eine um 1 Prozentpunkt höheres BIP Wachstum korrespondiert im Durchschnitt mit einer um 0,01 Prozentpunkte niedrigeren Arbeitslosenquote.

Erkenntnisse aus der Regressionstabelle

Interessanter: Der geschätze Koeffizient zum `BIP-Wachstum` in Höhe von 0,01.

Wie kann dieser Koeffizient interpretiert werden?

> Eine um 1 Prozentpunkt höheres BIP Wachstum korrespondiert im Durchschnitt mit einer um 0,01 Prozentpunkte niedrigeren Arbeitslosenquote.

Wie kann die Konstante interpretiert werden?

Erkenntnisse aus der Regressionstabelle

Interessanter: Der geschätze Koeffizient zum `BIP-Wachstum` in Höhe von 0,01.

Wie kann dieser Koeffizient interpretiert werden?

> Eine um 1 Prozentpunkt höheres BIP Wachstum korrespondiert im Durchschnitt mit einer um 0,01 Prozentpunkte niedrigeren Arbeitslosenquote.

Wie kann die Konstante interpretiert werden?

> Die erwartete Arbeitslosenquote bei einem Wachstum von 0% liegt im Durchschnitt bei 4,84 Prozent.

Erkenntnisse aus der Regressionstabelle

Weitere wichtige Erkenntnis aus der Tabelle:

- ✚ Der Koeffizient von `BIP-Wachstum` ist auf keinem gängigen Signifikanzniveau signifikant

> Woran kann dies gesehen werden?

Erkenntnisse aus der Regressionstabelle

Weitere wichtige Erkenntnis aus der Tabelle:

- ✚ Der Koeffizient von `BIP-Wachstum` ist auf keinem gängigen Signifikanzniveau signifikant

> Woran kann dies gesehen werden?

> Wie hoch ist die t-Statistik für unseren Koeffizienten `BIP-Wachstum`?

Erkenntnisse aus der Regressionstabelle

Weitere wichtige Erkenntnis aus der Tabelle:

- ✚ Der Koeffizient von `BIP-Wachstum` ist auf keinem gängigen Signifikanzniveau signifikant

> Woran kann dies gesehen werden?

> Wie hoch ist die t-Statistik für unseren Koeffizienten `BIP-Wachstum`?

Landkreise mit einem höheren BIP Wachstum könnten neue Unternehmen angesiedelt haben, welche neue Mitarbeiter brauchen. Daher würde ein entsprechend negativer Zusammenhang zwischen BIP-Wachstum und Arbeitslosenquote unseren Erwartungen entsprechen.

Arbeitslosenquote auf öffentliche Schulden regressieren

```
schulden <- lm(alo_quote ~ anteil_schulden, data=gesamtdaten)

modelsummary(schulden,
          type = "html",
          fmt = 2,
          statistic = 'conf.int',
          conf_level = .99,
          title = "Arbeitslosigkeit auf Anteil der Schulden pro Landkreis",
          gof_omit = 'DF|Deviance|Log.Lik.|F|RMSE|AIC|BIC',
          coef_rename = c("anteil_schulden" = "Anteil der Schulden", "(Intercept)" = "Konstante")
          )
```

Arbeitslosenquote auf öffentliche Schulden regressieren

Arbeitslosigkeit auf Anteil der Schulden pro Landkreis	
	(1)
Konstante	3.22
	[2.84, 3.60]
Anteil der Schulden	0.23
	[0.19, 0.27]
Num.Obs.	396
R2	0.316
R2 Adj.	0.314

Erkenntnisse aus der Regressionstabelle

Der geschätze Koeffizient zum `Anteil der öffentlichen Schulden` lieg bei 0,23.

> Wie kann dieser Koeffizient interpretiert werden?

Erkenntnisse aus der Regressionstabelle

Der geschätze Koeffizient zum `Anteil der öffentlichen Schulden` lieg bei 0,23.

Wie kann dieser Koeffizient interpretiert werden?

> Eine um 1 Prozentpunkt höhere Verschuldung korrespondiert im Durchschnitt mit einer um 0,23 Prozentpunkte höheren Arbeitslosenquote

Erkenntnisse aus der Regressionstabelle

Der geschätze Koeffizient zum `Anteil der öffentlichen Schulden` lieg bei 0,23.

> Wie kann dieser Koeffizient interpretiert werden?

> Eine um 1 Prozentpunkt höhere Verschuldung korrespondiert im Durchschnitt mit einer um 0,23 Prozentpunkte höheren Arbeitslosenquote

Die Interpretation der Konstante wäre dann wie folgt:

> Für einen Landkreis ohne Verschuldung wäre die erwartete Arbeitslosenquote im Durchschnitt bei 3,22 Prozent.

Erkenntnisse aus der Regressionstabelle

Der geschätze Koeffizient zum `Anteil der öffentlichen Schulden` lieg bei 0,23.

> Wie kann dieser Koeffizient interpretiert werden?

> Eine um 1 Prozentpunkt höhere Verschuldung korrespondiert im Durchschnitt mit einer um 0,23 Prozentpunkte höheren Arbeitslosenquote

Die Interpretation der Konstante wäre dann wie folgt:

> Für einen Landkreis ohne Verschuldung wäre die erwartete Arbeitslosenquote im Durchschnitt bei 3,22 Prozent.

Ein stark verschuldeter öffentlicher Haushalt hat potentiell weniger Gewerbeeinnahmen und da dort potentiell weniger Unternehmen vorhanden sind in denen Arbeitnehmer angestellt sein könnten.

Multiple lineare Regression

Multiple lineare Regression

- Sowohl das BIP Wachstum, als auch die öffentliche Verschuldung sind wichtige Faktoren zur Erklärung der Arbeitslosenquote
- Öffentliche Verschuldung schien wichtiger zu sein, doch können wir beide Variablen in EINE Regression aufnehmen?

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i,\ i = 1, \ldots, N$$

- Durch die multiple lineare Regression können wir den Effekt einer unabhängigen Variablen auf die abhängige Variable untersuchen und zusätzlich auf den Effekt anderer Variablen **kontrollieren**.
- Konkret: BIP-Wachstum und öffentliche Verschuldung in eine Regression packen!

Multiple lineare Regression

```
multi <- lm(alo_quote ~ anteil_schulden + bip_wachstum, data=gesamtdaten)

modelsummary(multi,
          type = "html",
          fmt = 2,
          statistic = 'conf.int',
          conf_level = .99,
          title = "Arbeitslosigkeit auf Anteil Schulden und BIP-Wachstum",
          gof_omit = 'DF|Deviance|Log.Lik.|F|RMSE|AIC|BIC',
          coef_rename = c("anteil_schulden" = "Anteil der Schulden", "bip_wachstum" = "BIP-Wachstum", "(Inter
          )
```

Multiple lineare Regression

Arbeitslosigkeit auf Anteil Schulden und BIP-Wachstum	
	(1)
Konstante	2.72
	[2.15, 3.30]
Anteil der Schulden	0.24
	[0.19, 0.28]
BIP-Wachstum	0.06
	[0.01, 0.12]
Num.Obs.	396
R2	0.331
R2 Adj.	0.327

Multiple lineare Regression

Arbeitslosigkeit auf Anteil Schulden und BIP-Wachstum	
	(1)
Konstante	2.72
	[2.15, 3.30]
Anteil der Schulden	0.24
	[0.19, 0.28]
BIP-Wachstum	0.06
	[0.01, 0.12]
Num.Obs.	396
R2	0.331
R2 Adj.	0.327

- Varianz wird zum Größten Teil durch die öffentlichen Schulden erklärt
- Schätzer für die Verschuldung bleibt in Höhe und Signifikanz bestehen
- BIP-Wachstum ist nun signifikant auf dem 1% Signifikanzniveau

Sample Splits und Interaktionsmodell

Sample Splits und Interaktionsmodell

Durch die deskriptive Analyse wissen wir, dass es große Unterschiede zwischen ost- und westdeutschen Landkreisen gibt (und das in allen untersuchten Dimensionen).

> Gilt der dokumentierte Zusammenhang zwischen dem Anteil der öffentlichen Verschuldung und der Arbeitslosenquote für Ost- und Westdeutschland gleichermaßen?

Sample Splits und Interaktionsmodell

Durch die deskriptive Analyse wissen wir, dass es große Unterschiede zwischen ost- und westdeutschen Landkreisen gibt (und das in allen untersuchten Dimensionen).

> Gilt der dokumentierte Zusammenhang zwischen dem Anteil der öffentlichen Verschuldung und der Arbeitslosenquote für Ost- und Westdeutschland gleichermaßen?

Um dieser Frage auf den Grund zu gehen wollen wir im ersten Schritt die Variable `Ostdeutschland` in unserer Regression hinzufügen:

```
schulden <- lm(alo_quote ~ anteil_schulden + ost, data=gesamtdaten)

modelsummary(schulden,
          type = "html",
          fmt = 2,
          statistic = 'conf.int',
          conf_level = .99,
          title = "Arbeitslosigkeit mit Interaktionstermen",
          gof_omit = 'DF|Deviance|Log.Lik.|F|RMSE|AIC|BIC',
          coef_rename = c("anteil_schulden" = "Anteil der Schulden", "ost1" = "Ostdeutschland", "(Intercept)"
          )
```

Sample Splits und Interaktionsmodell

	Arbeitslosigkeit mit Interaktionstermen
	(1)
Konstante	3.03
	[2.67, 3.40]
Anteil der Schulden	0.22
	[0.18, 0.26]
Ostdeutschland	1.47
	[0.96, 1.99]
Num.Obs.	396
R2	0.399
R2 Adj.	0.396

Sample Splits und Interaktionsmodell

	Arbeitslosigkeit mit Interaktionstermen
	(1)
Konstante	3.03
	[2.67, 3.40]
Anteil der Schulden	0.22
	[0.18, 0.26]
Ostdeutschland	1.47
	[0.96, 1.99]
Num.Obs.	396
R2	0.399
R2 Adj.	0.396

- `Ostdeutschland` ist eine Dummyvariable, welche 1 ist für alle ostdeutschen Landkreise
- In ostdeutschen Landkreisen ist die Arbeitslosigkeit im Durchschnitt um 1.47 Prozentpunkte höher als in westdeutschen Landkreisen
- Koeffizient signifikant auf dem 1%-Signifikanzniveau
- Höheres R^2 (Varianz in der Alo-quote kann besser erklärt werden)
- Keine Auswirkung auf den Koeffizienten der öffentlichen Verschuldung

Sample Splits und Interaktionsmodell

Diese Regression beantwortet jedoch nicht genau unsere Frage!

- ✚ Wir wollten wissen, ob der Zusammenhang zwischen öffentlicher Verschuldung und Arbeitslosenquote für alle ost- und westdeutschen Landkreise gleichermaßen gilt

Dafür müssen wir die Variable `Ostdeutschland` mit der Variablen `Anteil Schulden` **interagieren**!

Sample Splits und Interaktionsmodell

```
schulden <- lm(alo_quote ~ anteil_schulden + ost, data=gesamtdaten)
ost <- lm(alo_quote ~ anteil_schulden, data=filter(gesamtdaten, ost==1))
west <- lm(alo_quote ~ anteil_schulden, data=filter(gesamtdaten, ost==0))
interaktion <- lm(alo_quote ~ anteil_schulden*ost, data=gesamtdaten)

modelsummary(list(schulden, interaktion, west, ost),
          type = "html",
          fmt = 2,
          statistic = 'conf.int',
          conf_level = .99,
          title = "Arbeitslosigkeit mit Interaktionstermen",
          gof_omit = 'DF|Deviance|Log.Lik.|F|RMSE|AIC|BIC',
          coef_rename = c("anteil_schulden" = "Anteil der Schulden", "ost1" = "Ostdeutschland", "(Intercept)"
          )
```

Sample Splits und Interaktionsmodell

Arbeitslosigkeit mit Interaktionstermen

	(1)	(2)	(3)	(4)
Konstante	3.03	2.93	2.93	6.08
	[2.67, 3.40]	[2.56, 3.30]	[2.56, 3.30]	[4.77, 7.38]
Anteil der Schulden	0.22	0.23	0.23	0.03
	[0.18, 0.26]	[0.19, 0.28]	[0.19, 0.28]	[−0.12, 0.18]
Ostdeutschland	1.47	3.15		
	[0.96, 1.99]	[1.77, 4.53]		
Anteil der Schulden:Ostdeutschland		−0.21		
		[−0.36, −0.05]		
Num.Obs.	396	396	321	75
R2	0.399	0.416	0.379	0.003
R2 Adj.	0.396	0.412	0.377	−0.011

Darstellung der GOF

```
schulden <- lm(alo_quote ~ anteil_schulden + ost, data=gesamtdaten)
ost <- lm(alo_quote ~ anteil_schulden, data=filter(gesamtdaten, ost==1))
west <- lm(alo_quote ~ anteil_schulden, data=filter(gesamtdaten, ost==0))
interaktion <- lm(alo_quote ~ anteil_schulden*ost, data=gesamtdaten)

gm <- tibble::tribble(
  ~raw,        ~clean,          ~fmt,
  "nobs",      "N",             0,
  "adj.r.squared", "Adj. R<sup>2</sup>", 2)

modelsummary(list(schulden, interaktion, west, ost),
          type = "html",
          fmt = 2,
          statistic = 'conf.int',
          conf_level = .99,
          title = "Arbeitslosigkeit mit Interaktionstermen",
          gof_omit = 'DF|Deviance|Log.Lik.|F|RMSE|AIC|BIC',
          coef_rename = c("anteil_schulden" = "Anteil der Schulden", "ost1" = "Ostdeutschland", "(Intercept)"
          gof_map = gm
          )
```

Darstellung der GOF

	Arbeitslosigkeit mit Interaktionstermen			
	(1)	**(2)**	**(3)**	**(4)**
Konstante	3.03	2.93	2.93	6.08
	[2.67, 3.40]	[2.56, 3.30]	[2.56, 3.30]	[4.77, 7.38]
Anteil der Schulden	0.22	0.23	0.23	0.03
	[0.18, 0.26]	[0.19, 0.28]	[0.19, 0.28]	[−0.12, 0.18]
Ostdeutschland	1.47	3.15		
	[0.96, 1.99]	[1.77, 4.53]		
Anteil der Schulden:Ostdeutschland		−0.21		
		[−0.36, −0.05]		
N	396	396	321	75
Adj. R^2	0.40	0.41	0.38	−0.01

Sample Splits und Interaktionsmodell

> Wie können Sie den Interaktionsterm interpretieren?

Sample Splits und Interaktionsmodell

Wie können Sie den Interaktionsterm interpretieren?

- Spalte 2 repräsentiert das Interaktionsmodell
- In Spalte 3 und 4 wurden separate Regressionen für alle westdeutschen (Spalte 3) und ostdeutschen (Spalte 4) Landkreise durchgeführt
- Analyse von Spalte 2 im Zusammenspiel mit Spalte 3 und 4 erleichtert das Verständis für die Interaktionsvariable

Sample Splits und Interaktionsmodell

Wie können Sie den Interaktionsterm interpretieren?

- Spalte 2 repräsentiert das Interaktionsmodell
- In Spalte 3 und 4 wurden separate Regressionen für alle westdeutschen (Spalte 3) und ostdeutschen (Spalte 4) Landkreise durchgeführt
- Analyse von Spalte 2 im Zusammenspiel mit Spalte 3 und 4 erleichtert das Verständis für die Interaktionsvariable
- **Konstante:**
  - In Spalte 3 (für Westdeutsche) bei 2.93, was dem Wert aus Spalte 2 (Interaktionsmodell) entspricht.
  - In Spalte 4 (für Ostdeutsche) bei 6.08
  - Die durchschnittliche Arbeitslosenquote für einen unverschuldeten ostdeutschen Landkreis liegt deutlich höher als bei einem westdeutschen (2.93 Prozent vs. 6.08 Prozent)

Dieses Ergebnis bekommen wir auch aus dem Interaktionsmodell!

→ Dummy Variable `Ostdeutschland` und die Konstante aufaddieren: `Ostdeutschland + Constant` = `2.93 + 3.15` = `6.08` (Achtung: Rundung)!

Sample Splits und Interaktionsmodell

- **Anteil Schulden:**
  - In Spalte 3 (für Westdeutsche) bei 0.23, was dem Wert aus Spalte 2 (Interaktionsmodell) entspricht
  - In Spalte 4 (für Ostdeutsche) ist der Zusammenhang deutlich kleiner und insignifikant
  - Für alle westdeutschen Landkreise gibt es einen signifikanten Zusammenhang zwischen der öffentlichen Verschuldung und der Arbeitslosenquote
  - Direkt ersichtlich das der Zusammenhang für ostdeutsche Landkreise signifikant kleiner ist als für westdeutsche (um -0.21 Prozentpunkte, der Koeffizient von `Anteil Schulden * Ostdeutschland`

→ Wenn wir uns den Zusammenhang für alle ostdeutschen Landkreise berechnen möchten, dann ergibt sich dieser als
`Anteil Schulden + Anteil Schulden * Ostdeutschland = 0.23 + (−0.21) = 0.02`

Die westdeutschen Landkreise dienen uns hier überall als Basislevel!

Sample Splits und Interaktionsmodell

- **Anteil Schulden:**
  - In Spalte 3 (für Westdeutsche) bei 0.23, was dem Wert aus Spalte 2 (Interaktionsmodell) entspricht
  - In Spalte 4 (für Ostdeutsche) ist der Zusammenhang deutlich kleiner und insignifikant
  - Für alle westdeutschen Landkreise gibt es einen signifikanten Zusammenhang zwischen der öffentlichen Verschuldung und der Arbeitslosenquote
  - Direkt ersichtlich das der Zusammenhang für ostdeutsche Landkreise signifikant kleiner ist als für westdeutsche (um -0.21 Prozentpunkte, der Koeffizient von `Anteil Schulden * Ostdeutschland`

→ Wenn wir uns den Zusammenhang für alle ostdeutschen Landkreise berechnen möchten, dann ergibt sich dieser als
`Anteil Schulden + Anteil Schulden * Ostdeutschland` = `0.23 + (−0.21)` = `0.02`

Die westdeutschen Landkreise dienen uns hier überall als Basislevel!

Vorteil des Interaktionsmodells:

Durch das Interaktionsmodell nutzen wir **eine** Regression und verwenden den kompletten Datensatz, dadurch hat unsere Regression mehr Power um Effekte zu finden.

Sind diese Ergebnisse *kausal* zu interpretieren?

Sind diese Ergebnisse *kausal* zu interpretieren?

- ✚ Basieren auf Beobachtungsdaten
- ✚ Arbeitslosenquote könnte noch von vielen anderen Faktoren beeinflusst sein, welche wir hier nicht aufgenommen haben (z.B. der Bevölkerungszuwachs oder die Inflation)
- ✚ Um kausale Effekte messen zu können müssten wir entweder ein kontrolliert randomisiertes Experiment durchführen oder uns ein natürliches Experiment in den Daten zunutze machen

Kausale Antworten auf verschiedenste Fragen wollen wir in den folgenden Vorlesungseinheiten auf der Basis anderer Datensätze tätigen.

Zusammenfassung

> Was haben wir über die Arbeitslosenquote in Deutschland gelernt?

- ✚ Es gibt starke regionale Unterschiede in Deutschland
- ✚ Der Anteil der öffentlichen Schulden scheint ein wichtiger Faktor zur Vorhersage der Arbeitslosenquote zu sein
- ✚ Eine fundierte deskriptive Analyse schafft die Grundlage für eine spätere fundierte tiefergehende Analyse mittels linearer Regression

Übungsaufgaben

Im ersten Teil der Case Study hatten Sie sich noch die durchschnittlichen Einkommen auf Landkreisebene in R eingelesen und im zweiten Teil deskriptiv untersucht. Nun sollten Sie diese Tabelle mittels linearer Regression analysieren:

- Erstellen Sie eine Regressionstabelle mittels `modelsummary` in der Sie die Arbeitslosenquote auf die Einkommen für das Jahr 2022 regressieren.
  - Interpretieren Sie ihre Ergebnisse
- Erstellen Sie ein Interaktionsmodell incl. Sample Split mittels `modelsummary` und interpretieren Sie die Ergebnisse ihrer Regressionen.