Case Study für Deutschland

Alexander Rieber - 07 Mai 2020

Dritter Teil der Case Study

Im dritten Teil der Case Study untersuchen Sie mögliche Gründe für die regionalen Unterschiede innerhalb Deutschlands. Mit den ihnen zur Verfügung stehenden Daten zum BIP und der Verschuldung der einzelnen Landkreise wollen Sie die Arbeitslosenquoten in den einzelnen Regionen Deutschlands erklären. Ziele des dritten Teils der Case Study:

- Regressionen in R durchführen
- Interpretation von Regressionskoeffizienten

Sie lernen, wie Sie eine lineare Regression dazu Nutzen können um mögliche Zusammenhängen zwischen der Arbeitslosigkeit und anderen Faktoren näher zu beleuchten. Jedoch lernen Sie auch die Grenzen der lineraren Regression kennen, insbesondere im Hinblick auf die Interpetation der Koeffizienten der Regression. Ergänzend hierzu erhalten Sie im 4. RTutor Problem Set Einblicke in die Zusammenhänge verschiedener gesamtwirtschaftlicher Faktoren und der Arbeitslosigkeit in den einzelnen Ländern der europäischen Union. Im 5. RTutor Problem Set werden Sie zusätzlich erfahren, welche Möglichkeiten wir in den Wirtschaftswissenschaften haben, um kausale Schlüsse ziehen zu können.

Daten und Pakete laden

Nachdem wir uns im ersten Teil der Case Study alle Daten aus verschiedenen Datenquellen zusammengetragen und in R eingelesen haben, wurden diese im zweiten Teil visualisiert. In diesem dritten und letzten Teil der Case-Study wollen wir den Zusammenhang verschiedener Variablen untersuchen.

Hierzu laden wir uns die in Teil 1 erstellten Datensätze und die in Teil 2 gemachten Anpassungen:

```
library(tidyverse)
library(skimr)
library(stargazer)
library(corrr)
```

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

# Zuerst wollen wir die Arbeitslosenquote, einen Dummy für Ostdeutschland und die Verschuldung im Verhä
gesamtdaten <- gesamtdaten %>%
  mutate(alo_quote = (total_alo / (erw+total_alo))*100,
         ost = as.factor(ifelse(bundesland_name %in% c("Brandenburg", "Mecklenburg-Vorpommern", "Sachse
         ost_name = ifelse(ost == 1, "Ostdeutschland", "Westdeutschland"),
         anteil_schulden = (Schulden_gesamt / bip)*100)
```

Korrelation

Wir hatten uns im letzten Teil der Case Study die Wachstumsrate des BIP berechnet und deren Verlauf seit 2000 durch eine Grafik visualisiert. In diesem Teil der Case Study wollen wir uns anschauen, wie die Arbeitslosenqutoe mit dem BIP Wachstum und dem Anteil der öffentlichen Schulden am BIP im Jahr 2017 für die einzelnen Landkreise in Deutschland zusammenhängen.

Durch die bisherigen deskriptiven Analysen haben wir eine große Streuung des BIP pro Kopf innerhalb Deutschlands festgestellt. Insbesondere gab es einen deutlichen Unterschied zwischen ost- und westdeutschen Landkreisen. Im BIP Wachstum waren sich ost- und westdeutsche Landkreise im Durchschnitt recht ähnlich, doch auch hier gab es beträchtliche Unterschiede zwischen den Landkreisen. Um etwas darüber Aussagen zu können, wie die Arbeitslosenquote mit dem BIP-Wachstum oder der öffentlichen Verschuldung zusammenhängt, wollen wir uns die Korrelation der einzelnen Variablen untereinander näher anschauen.

Arbeitslosenquote und BIP-Wachstum

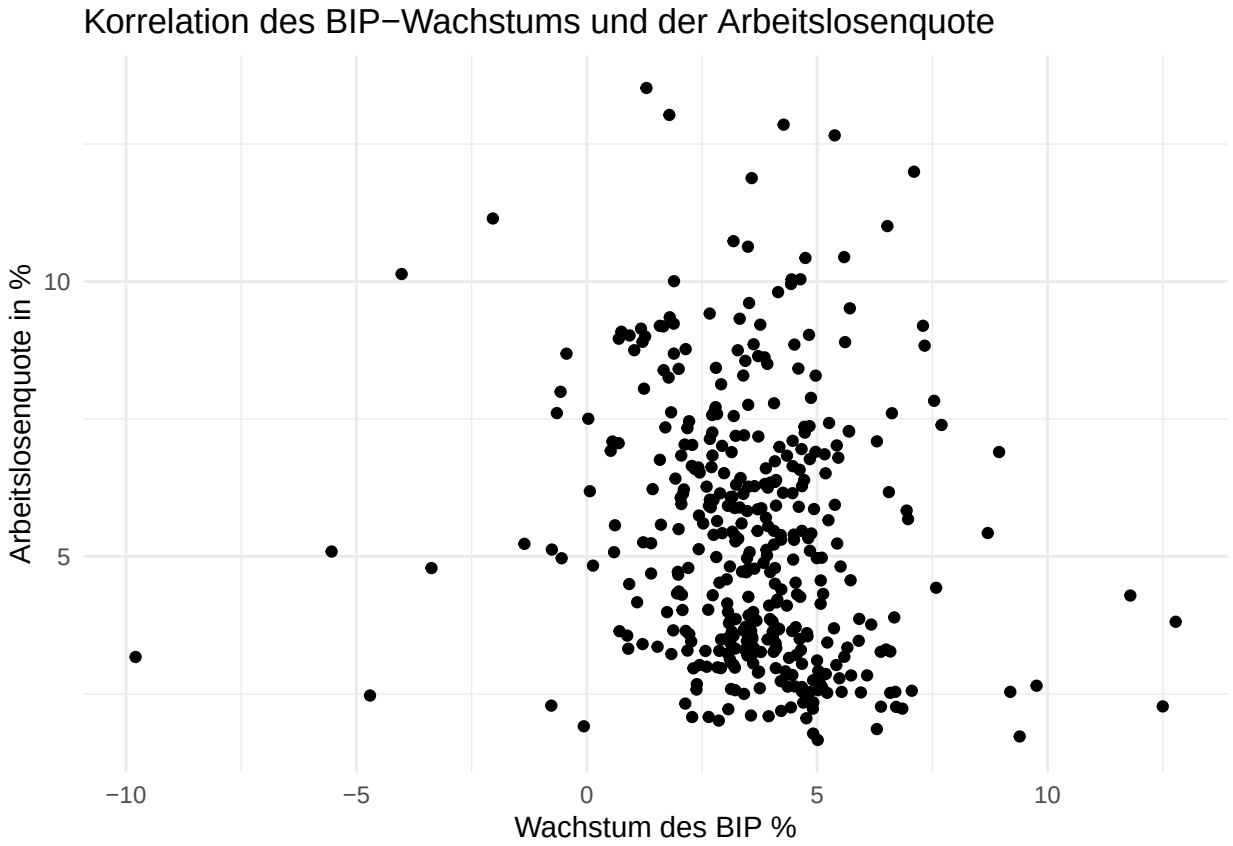
Zuerst berechnen wir uns die Variable `bip_wachstum`, wie im zweiten Teil der Case Study, und mergen diese zu unseren `gesamtdaten` für das Jahr 2017.

```
bip_wachstum <- bip_zeitreihe %>%
  filter( nchar(Regionalschluessel) == 5) %>%
  group_by(Regionalschluessel) %>%
  arrange(Jahr) %>%
  mutate( bip_wachstum = 100*(bip - lag(bip)) / bip ) %>%
  ungroup() %>%
  filter( Jahr == 2017 ) %>%
  select(Regionalschluessel, bip_wachstum, Jahr)

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

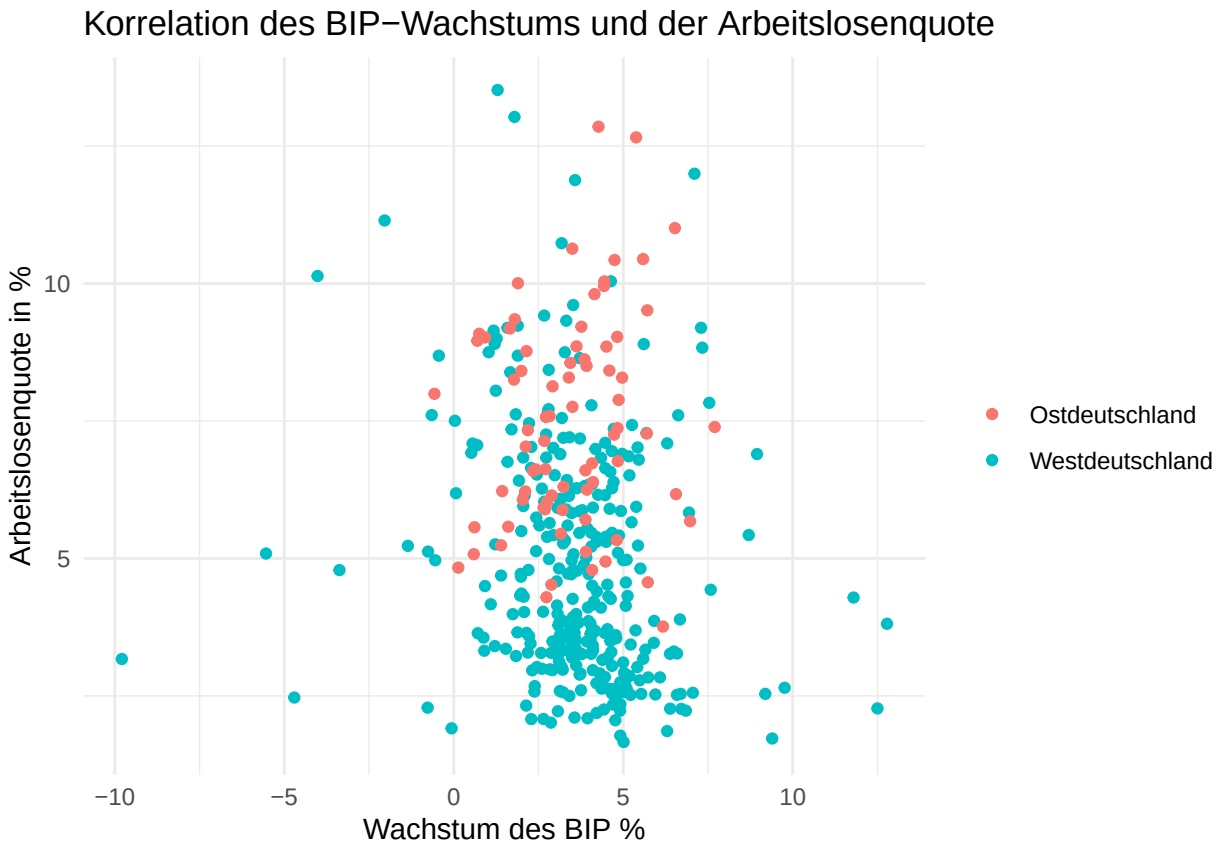
Anschließend erstellen wir ein Streudiagramm, welches die Arbeitslosenquote und das BIP Wachstum in den einzelnen Landkreisen einander gegenüberstellen. Hier erhalten wir erste Einblicke in die Korrelation der zwei Variablen.

```
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 der Arbeitslosenquote") +
  theme_minimal()
```



Zuerst fallen uns die Ausreißer ins Auge. Insbesondere im positiven Bereich gibt es Landkreise, welche ein BIP-Wachstum von über 10% von 2016 nach 2017 aufweisen. Aber es gibt auch einen Landkreis, welcher eine Schumpfung des BIP um über 9% aufweist. Solche Ausreißer können daher stammen, dass es in den jeweiligen Landkreisen im Vorjahr ein sehr niedriges BIP Wachstum (bspw. Landkreis Forchheim) oder ein sehr hohes BIP Wachstum (bspw. Landkreis Aschaffenburg) gab, d.h. wir sehen hier für vereinzelte Landkreise einen Aufholprozess aus dem vorherigen Jahr. Diese hohen oder niedrigen Wachstumsraten könnten natürlich durch vereinzelte Projekte auf Landkreisebene bedingt sein, werden sich jedoch im gesamten Datensatz mit 401 Landkreisen und kreisfreien Städten wieder relativieren. Durch unseren umfangreichen Datensatz können wir solche einzelnen Ausreißer auffangen. Wenn wir jedoch eine kleine Stichprobe aus diesem Datensatz extrahieren (z.B. nur alle ostdeutschen Landkreise), dann könnte schon ein deutlich anderes Bild des Zusammenhangs der beiden Variablen entstehen. Hier haben z.B. alle ostdeutschen Landkreise eine relativ große Streuung in den Arbeitslosenquoten, aber nur eine geringe Bandbreite beim BIP Wachstum.

```
gesamtdaten %>%
  ggplot(aes(x = bip_wachstum, y = alo_quote, color = ost_name)) + geom_point() +
  labs( x = "Wachstum des BIP %",
        y = "Arbeitslosenquote in %",
        title = "Korrelation des BIP-Wachstums und der Arbeitslosenquote",
        color = "") +
  theme_minimal()
```



Anhand der Punktewolke könnten wir abschätzen das der Zusammenhang zwischen der Arbeitslosenquote und dem BIP-Wachstum leicht negativ ist. D.h. Landkreise mit einer niedrigen Wachstumsrate haben tendenziell eine höhere Arbeitslosenquote in 2017. Jedoch scheint der Zusammenhang nicht besonders stark zu sein.

Wenn wir uns die Korrelation zwischen den zwei Variablen ausgeben lassen, so erhalten wir eine negative Korrelation von:

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

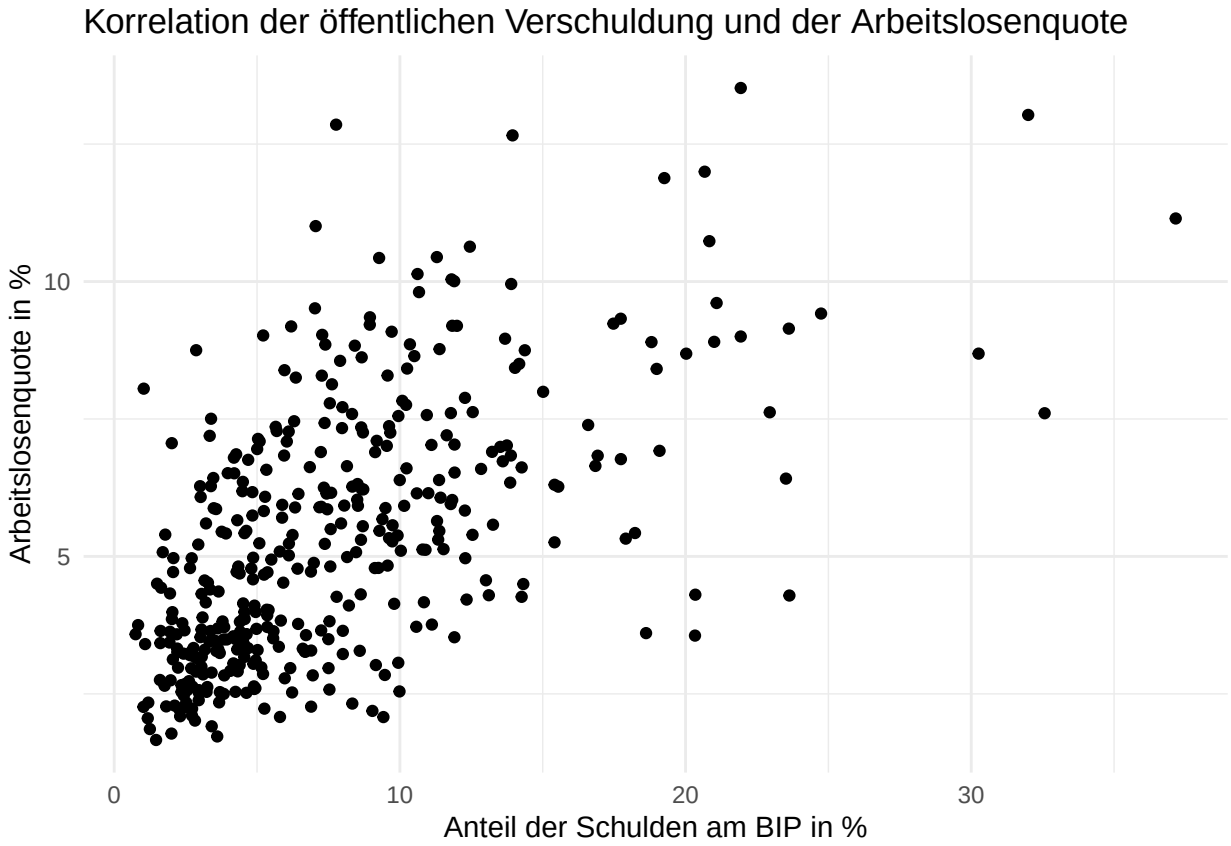
```
## [1] -0.1521478
```

D.h. der von uns in der Grafik vermutete negative Zusammenhang zwischen der Arbeitslosenquote und dem BIP-Wachstum wird durch die Analyse der Korrelation bestätigt.

Arbeitslosenquote und Verschuldung pro Person

Die Korrelation der Arbeitslosenquote und dem Anteil der öffentlichen Schulden möchten wir nun auch über ein Streudiagramm untersuchen.

```
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 Verschuldung und der Arbeitslosenquote") +
  theme_minimal()
```



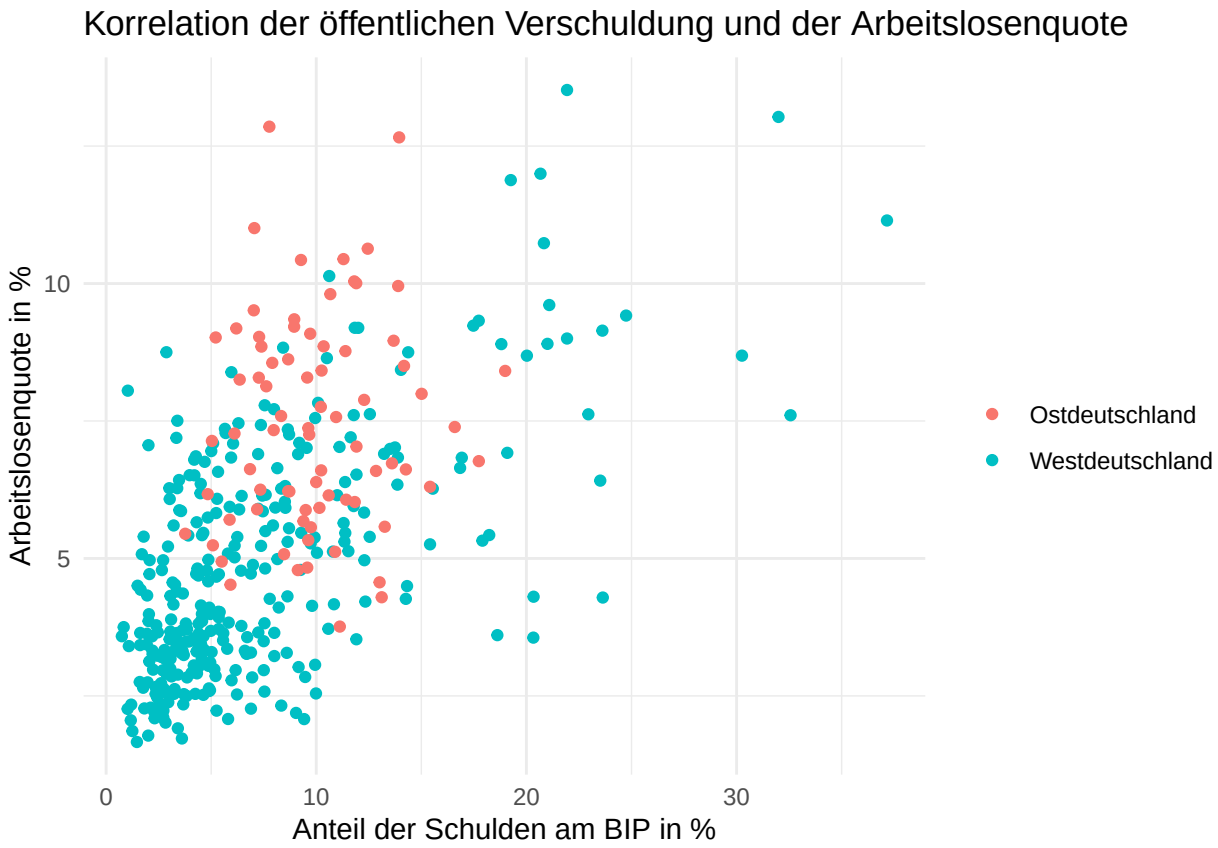
In diesem Schaubild scheint es einen positiven Zusammenhang zwischen der öffentlichen Verschuldung und der Arbeitslosenquote zu geben. D.h. Landkreise mit einer hohen Verschuldung im Verhältnis zu ihrem BIP haben tendenziell auch eine höhere Arbeitslosenquote. Jedoch ist es auch hier schwer zu sagen, wie groß die Korrelation tatsächlich ist. Wenn wir diese berechnen, so liegt sie bei:

```
cor(gesamtdaten$alo_quote, gesamtdaten$anteil_schulden,use = "pairwise.complete.obs")
```

```
## [1] 0.5938476
```

Wenn wir wieder eine kleine Stichprobe aus diesem Datensatz extrahieren, dann haben auch hier alle alle ostdeutschen Landkreise eine relativ große Streuung in den Arbeitslosenquoten, aber nur eine geringe Bandbreite in der öffentlichen Verschuldung.

```
gesamtdaten %>%
  ggplot(aes(x = anteil_schulden, y = alo_quote, color=ost_name)) + geom_point() +
  labs( x = "Anteil der Schulden am BIP in %",
        y = "Arbeitslosenquote in %",
        title = "Korrelation der öffentlichen Verschuldung und der Arbeitslosenquote",
        color = "") +
  theme_minimal()
```



Interpretation der Korrelation

Die Korrelation als Zahl an sich hat keine intuitive quantitative Interpretation. Sie ist eine univariate Repräsentation des Zusammenhangs zweier Variablen. Sie kann uns jedoch dabei helfen stark miteinander korrelierte Variablen in unserem Datensatz ausfindig zu machen. Dies ist insbesondere bei einer späteren Regressionsanalyse hilfreich, wenn es sich um mehrere erklärenden Variablen handelt, welche wir alle in unsere Regression einfließen lassen möchten. Wenn wir beispielsweise eine sehr hohe positive oder negative Korrelation zwischen den erklärenden Variablen `bip_wachstum` und `anteil_schulden` finden würden, dann müssten wir uns überlegen, ob wir beide Variablen mit in unsere spätere Regression aufnehmen (bei einer multiplen linearen Regression). In unserem Fall ist die Korrelation der beiden Variablen jedoch gering:

```
cor(gesamtdaten$bip_wachstum, gesamtdaten$anteil_schulden,use = "pairwise.complete.obs")
```

```
## [1] -0.1323647
```

In einer empirischen Ausarbeitung werden Sie i.d.R. keine Schaubilder zur Korrelation der einzelnen Variablen sehen, sondern nur eine Tabelle in der die Korrelationen der Variablen untereinander abgetragen sind. Eine solche Tabelle können Sie auch einfach in R innerhalb des `tidyverse` erzeugen:

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

fashion(korrelationen)
```

```
##           rowname bip_wachstum anteil_schulden alo_quote
## 1    bip_wachstum
```

```
## 2 anteil_schulden         -.13
## 3       alo_quote         -.15             .59
```

In empirischen Ausarbeitungen wird nach der Analyse der Korrelation eine Analyse mittels Regression durchgeführt. Dies liegt daran, dass Regressionskoeffizienten, anders als Korrelationen eine quantitative Interpretation zulassen (und das wir den Zusammenhang zwischen mehreren Variablen betrachten können). Wir wollen uns nun auch mit der linearen Regression beschäftigen.

Einfache lineare Regression

In diesem Kapitel werden wir uns mit der (einfachen) linearen Regression beschäftigen. Durch die lineare Regression können wir mit der Methode der kleinsten Quadrate den Zusammenhang zweier Variablen in einer einzelnen Zahl zusammenfassen und interpretieren. Das Modell für unsere Regression kennen wir aus der Vorlesung:

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

Wobei y die abhängige und x die unabhängige Variable ist (auch erklärende Variable genannt).

Arbeitslosigkeit auf BIP-Wachstum regressieren

Wir wollen nun die Arbeitslosenquote auf die Wachstumsrate des BIP regressieren:

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

stargazer(bip, type = "html", header = FALSE, digits = 2)
```

Dependent variable:

alo_quote

bip_wachstum

-0.17***

(0.05)

Constant

5.93***

(0.23)

Observations

399

R2

0.02

Adjusted R2

0.02

Residual Std. Error

2.34 (df = 397)

F Statistic

9.41*** (df = 1; 397)

Note:

p<0.1; ***p<0.05;*** p<0.01

Zuerst betrachten wir die Anzahl an Beobachtungen in der Regression. Wir haben hier 399 Beobachtungen, jedoch wissen wir, dass es in Deutschland insgesamt 401 Landkreise und kreisfreie Städte gibt. Da wir allerdings für Berlin und Hamburg keine Information zu den Arbeitslosenquoten haben werden diese zwei kreisfreien Städte in unserer Regression nicht berücksichtigt.

Das "R2" (R^2) gibt wieder, wie hoch der Anteil der Varianz in unseren Daten ist, welchen wir durch unser Modell erklären können. Ein R^2 von 0.02 sagt uns, das die Varianz der Residuen 2% der Varianz unserer Responsevariable (hier der Arbeitslosenquote) ausmacht. Dies gilt jedoch immer auf der Grundlage unserer Daten, unseres Modells und unserer Annahmen. Das Problem mit dem R^2 ist nun, dass wir dieses allein durch die Hinzunahme von weiteren erklärenden Variablen in unser Modell erhöhen könnten (rein technisch bedingt). Daher sollten wir uns eher auf das "Adjusted R2" konzentrieren. Hier wird das R^2 um die Anzahl an Variablen in unserem Modell bereinigt. Insgesamt sollten wir dem R^2 und auch adjusted R^2 keine zu Große Bedeutung beimessen. Es ist eines von vielen Gütemaßen, kann jedoch nur unter bestimmten Voraussetzungen zum Vergleich mehrer Modelle herangezogen werden: Die Modelle müssen auf den gleichen Daten angewendet werden und unter den gleichen Annahmen. Bei Zeitreihenanalysen werden wir tendenziell sehr hohe Werte für das R^2 erhalten, bei Querschnitts- und Paneldaten niedrigere.

Interessanter ist es nun den geschätzten Koeffizient zum `bip_wachstum` zu interpretieren. Dies können wir wie folgt formulieren:

> Eine um 1 Prozentpunkt höheres BIP Wachstum korrespondiert mit einer um 0,17 Prozentpunkte niedrigeren Arbeitslosenquote.

Weiterhin erhalten wir einen Wert für die Konstante in unserem Modell. Die Konstante kann folgendermaßen interpretiert werden:

> Die erwartete Arbeitslosenquote bei einem Wachstum von 0% liegt bei 5,93 Prozent.

Allerdings müssen wir bei der Interpretation der Koeffizienten auch immer deren Signifikanz berücksichtigen. Der Koeffizient von `bip_wachstum` ist signifikant auf dem 1%-Niveau. Damit können wir sagen, dass das BIP Wachstum ein signifikanter Faktor zur Erklärung der Arbeitslosenquote in einem Landkreis ist. Vermutlich ist es allerdings nicht der einzige wichtge Faktor, wie wir an dem R^2 von 0.02 sehen. Landkreise in denen sich bspw. im Jahr 2016 neue Unternehmen ansiedeln werden im Jahr 2017 tendenziell ein Wachstum des BIP verzeichnen, allein durch die zusätzlich produzierten Güter und Dienstleistungen in diesem neuen Unternehmen. Jedoch braucht das Unternehmen auch Mitarbeiter, welche sie aus dem Landkreis (und auch anderswo her) rekrutieren kann. Daher würden wir tendenziell erwarten, dass ein höheres BIP Wachstum mit einer niedrigeren Arbeitslosenquote korrespondiert.

Arbeitslosigkeit auf öffentliche Verschuldung regressieren

Im nächsten Schritt wollen wir anschauen ob die öffentlichen Schulden ihren Teil zur Erklärung der Arbeitslosenquote beitragen können, und wie hoch dieser Teil ist.

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

stargazer(schulden, type = "html" ,  header = FALSE, digits=2)
```

Dependent variable:

alo_quote

anteil_schulden

0.25***

(0.02)

Constant

3.37***

(0.16)

Observations

397

R2

0.35

Adjusted R2

0.35

Residual Std. Error

1.90 (df = 395)

F Statistic

215.18*** (df = 1; 395)

Note:

p<0.1; ***p<0.05;*** p<0.01

In dieser Regression haben wir nur noch 397 Beobachtungen. Dies liegt daran, dass in unserem Datensatz für Berlin, Hamburg, Bremen und Bremerhaven keine Information zu den Schulden vorliegen und diese deshalb in die Regression nicht aufgenommen werden können. Das R^2 ist mit 0.35 schon deutlich höher als bei der Regression mit dem BIP Wachstum. Dies lässt vermuten das die öffentlichen Schulden ein wichtiger Faktor zur Erklärung der Arbeitslosenquote ist, mit einem deutlich größeren Einfluss als das BIP Wachstum.

Die Interpretation der Schätzer könnte wie folgt lauten:

> Eine um 1 Prozentpunkt höhere Verschuldung korrespondiert mit einer um 0,25 Prozentpunkte höheren Arbeitslosenquote

Die Interpretation der Konstante wäre dann wie folgt:

> Für einen Landkreis ohne Verschuldung wäre die erwartete Arbeitslosenquote bei 3,37 Prozent.

Auch dies ist nachvollziehbar, da ein stark verschuldeter öffentlicher Haushalt in strukturschwachen Landkreisen weniger Gewerbeeinnahmen hat und daher auch weniger Unternehmen vorhanden sind in denen Arbeitnehmer angestellt sein könnten. Daher würden wir auch erwarten das eine höhere Verschuldung mit einer höheren Arbeitslosenquote korrespondiert.

Wir haben uns bisher nur einfachen linearen Regressionen gewidmet, jedoch können wir in die Regression auch mehrere erklärende Variablen aufnehmen. In diesem Fall sprechen wir dann von einer multiplen linearen Regression.

Multiple lineare Regression

Wir haben im vorherigen Abschnitt gesehen, dass sowohl das BIP Wachstum als auch die öffentliche Verschuldung wichtige Faktoren zur Erklärung der Arbeitslosenquote in den einzelnen Landkreisen sind. In diesem Abschnitt wollen wir beide Variablen zusammen in die Regression aufnehmen. Das Modell für unsere Regression kennen wir aus der Vorlesung:

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

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 bedeutet dies in unserem Fall: Wir vermuten nach den univariaten Regressionen, dass die öffentliche Verschuldung der Hauptfaktor für die Arbeitslosenquote in den Landkreisen darstellt, sind uns jedoch nicht sicher, ob nicht auch noch das BIP Wachstum einen erheblichen Anteil zur Erklärung beitragen könnte. In der multiplen linearen Regression können wir nun beide Variablen aufnehmen und so den Effekt der öffentlichen Schulden auf die Arbeitslosenquote, kontrolliert auf das BIP Wachstum, untersuchen.

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

stargazer(multi, type = "html" ,  header = FALSE, digits=2)
```

Dependent variable:	
alo_quote	
anteil_schulden	
0.25***	
(0.02)	
bip_wachstum	
-0.09*	
(0.04)	
Constant	
3.71***	
(0.24)	
Observations	
397	
R2	
0.36	
Adjusted R2	
0.36	
Residual Std. Error	
1.89 (df = 394)	
F Statistic	
110.18*** (df = 2; 394)	
Note:	
p<0.1; ***p<0.05;*** p<0.01	

Hier sehen wir bereits, dass der größte Teil der Varianz in unserem Modell durch die öffentlichen Schulden erklärt wird (siehe Regression vorheriger Abschnitt mit dem R^2 von 0.35). Weiterhin bleibt der Schätzer für die Verschuldung signifikant auf dem 1%-Niveau und in seiner Höhe gleich (vorherige univariate Regression war der Schätzer auch bei 0.25). D.h. auch kontrolliert auf das BIP Wachstum ist die öffentliche Verschuldung ein signifikanter Faktor zur Erklärung der Arbeitslosenquote in den einzelnen Landkreisen. Jedoch ist das BIP Wachstum in dieser Regression nur noch auf dem 10%-Niveau signifikant und in seiner Höhe auch deutlich kleiner als bei der univariaten Regression aus dem vorherigen Abschnitt (vorher -0.17).

Dies legt nahe, dass die Verschuldung des öffentlichen Haushalts besser zur Erklärung der Arbeitslosenquote in den einzelnen Landkreisen dient als das BIP Wachstum.

Sample Splits und Interaktionsmodell

Im letzten Abschnitt wollen wir uns noch mit dem Interaktionsmodell beschäftigen.

Wir hatten in der deskriptiven Analyse schon herausgefunden, dass es deutliche Unterschiede zwischen Ost- und Westdeutschland gibt, was Arbeitslosenquote, Verschuldung und auch BIP anbelangt. Nun wäre es interessant zu wissen, ob der Zusammenhang zwischen dem Anteil der öffentlichen Verschuldung am BIP und der Arbeitslosenquote sowohl für ostdeutsche als auch westdeutsche Landkreise gilt.

In einem ersten Schritt haben wir hierzu die Dummyvariable `ost` der Regression hinzugefügt. Da es vielfältige Einflüsse geben könnte, warum ost- und westdeutsche Landkreise unterschiedlich sein könnten wollen wir durch eine Dummyvariable `ost` darauf kontrollieren. Der Vorteil einer solchen Dummyvariablen ist, dass hiermit alle beobachtbaren und unbeobachtbaren Unterschiede zwischen ost- und westdeutschen Landkreisen Rechnung getragen werden kann.

```
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)

stargazer(schulden, interaktion, west, ost, type = "html" ,  header = FALSE, digits=2)
```

Dependent variable:

alo_quote

(1)

(2)

(3)

(4)

anteil_schulden

0.22***

0.24***

0.24***

0.05

(0.02)

(0.02)

(0.02)

(0.07)

ost1

2.02***

3.82***

(0.23)

(0.68)

anteil_schulden:ost1				
-0.18***				
(0.07)				
Constant				
3.20***	3.12***	3.12***	6.94***	
(0.15)	(0.15)	(0.15)	(0.75)	
Observations				
397	397	321	76	
R2				
0.46	0.47	0.41	0.01	
Adjusted R2				
0.46	0.47	0.41	-0.01	
Residual Std. Error				
1.73 (df = 394)				
1.72 (df = 393)				
1.66 (df = 319)				
1.95 (df = 74)				
F Statistic				
169.06*** (df = 2; 394)				
117.33*** (df = 3; 393)				
220.33*** (df = 1; 319)				

0.53 (df = 1; 74)

Note:

p<0.1; ***p<0.05;*** p<0.01

Die Variable ist in unserer Tabelle als `ost1` enthalten, d.h. der Schätzer, den wir hier erhalten, gilt für alle Beobachtungen in denen der Wert der Variable 1 ist. Die Fälle in denen die Variable 0 ist werden als Basislevel herangezogen. Konkret bedeutet der Schätzer, dass es in Ostdeutschland eine durchschnittlich um 2.02 Prozentpunkte höhere Arbeitslosenquote gibt als in Westdeutschland. Der Schätzer ist signifikant auf dem 1%-Signifikanzniveau. Durch die Hinzunahme der Dummyvariable `ost` können wir die Varianz innerhalb der Arbeitslosenquote in Deutschland besser erklären als in der Regression ohne `ost`, was sich an dem erhöhten R^2 zeigt (zuvor 0.35, nun 0.46). Der Koeffizient für die öffentliche Verschuldung verändert sich nur marginal von 0.25 auf 0.22 und bleibt auf dem 1%-Signifikanzniveau signifikant.

Nun trifft diese erste Regression jedoch nicht den Punkt, welchen wir eigentlich untersuchen wollten: Wir wollten wissen, ob der Zusammenhang zwischen der öffentlichen Verschuldung und der Arbeitslosenquote für alle ostdeutschen und westdeutschen Landkreise gleichermaßen gilt. Hier reicht es nicht aus darauf zu kontrollieren ob ein Landkreis als ost- oder westdeutsch klassifiziert wird. Zur Beantwortung unserer Frage müssen wir die Variable `ost` mit der Variablen `anteil_schulden` *interagieren.* Erst dann erhalten wir einen Schätzer für die öffentliche Verschuldung in Ost- und Westdeutschland. Diese können wir miteinander vergleichen und so beantworten, ob der Zusammenhang zwischen der Verschuldung und der Arbeitslosigkeit in Ost- und Westdeutschland gleich stark ist.

Um besser zu veranschaulichen was die Regression mit der interagierten Variable genau macht, bzw. wie diese zu interpretieren ist, haben wir zusätzlich einen Sample Split gemacht. In diesem Sample Split unterteilen wir unsere Stichprobe nach ost- und westdeutschen Landkreisen und wenden unser Modell zum einen nur auf die westdeutschen Landkreise an und zum anderen nur für die ostdeutschen Landkreise (Spalte 3 und 4 der Tabelle).

Analysieren wir Spalte 3 und 4: Dafür beginnen wir mit der Konstanten: Diese ist für die Westdeutschen bei 3.12 (Spalte 3), was dem Wert der Konstanten aus unserem Interaktionsmodell (Spalte 2) entspricht. Bei den Ostdeutschen liegt diese bei 6,94 (Spalte 4), d.h. die durchschnittliche Arbeitslosenquote für einen unverschuldeten ostdeutschen Landkreis liegt deutlich höher als bei einem westdeutschen (3.12 Prozent vs. 6.94 Prozent).

Können wir dies auch aus unserem Interaktionsmodell (Spalte 2) ablesen? Ja! In unserem Interaktionsmodell (Spalte 2) entspricht die erhalten wir genau die gleiche Arbeitslosenquote wie im Sample Split für Ostdeutschland: Hierfür müssen wir die Dummy Variable `ost` und die Konstante aufaddieren: `ost1 + Constant` = `3.82 + 3.12` = `6.94`!

Gleiches gilt auch für die jeweiligen Schätzer von `anteil_schulden` und dessen Interaktion mit `ost`. Der Schätzer für die öffentlichen Schulden liegt bei 0.24, sowohl im Interaktionsmodell (Spalte 2) als auch in der Regression rein nur für westdeutsche Landkreise (Spalte 3). Dies bedeutet für alle westdeutschen Landkreise gibt es einen signifikanten Zusammenhang zwischen der öffentlichen Verschuldung und der Arbeitslosenquote. Bei den ostdeutschen Landkreisen ist dieser Zusammenhang deutlich kleiner und insignifikant (Spalte 4). Auch in unserem Interaktionsmodell können wir sehen, dass der Einfluss der öffentlichen Verschuldung für ostdeutsche Landkreise signifkant kleiner ist als für westdeutsche (um -0.18 Prozentpunkte, der Koeffizient von `anteil_schulden:ost1`). Wenn wir uns den Zusammenhang für alle ostdeutschen Landkreise berechnen möchten, dann ergibt sich dieser als `anteil_schulden + anteil_schulden:ost1` = `0.24 + (-0.18)` = `0.06`. Durch Rundungsfehler können hier kleinere Abweichungen zwischen dem Koeffizienten aus dem Interaktionsmodell (Spalte 2) und dem Sample Split (Spalte 4) entstehen.

Vorteil des Interaktionsmodells gegenüber dem Sample Split: Durch das Interaktionsmodell nutzen wir **eine** Regression und verwenden den kompletten Datensatz, dadurch hat unsere Regression mehr Power um Effekte zu detektieren. Wenn wir einen Sample Split durchführen und unsere Stichprobe dadurch sehr klein wird (76 Beobachtungen ist schon recht wenig), dann ist es schwerer signifikante Ergebnisse zu finden, auch wenn diese eventuell vorhanden sind.

Korrelation ist nicht gleich Kausalität

Die in dieser Case Study vorgestellten Ergebnisse sind leider nicht kausal interpretierbar! Wir müssen dies auch bei der Interpretation der Schätzer immer berücksichtigen. Es gibt sehr viele andere Faktoren, welche die Arbeitslosenquote beeinflussen können und die wir in unserer Analyse aktuell nicht berücksichtigt haben. Beispielsweise könnte es sein, dass Städte mit Universitäten Innovationszentren sind und viele konkurrenzfähige Unternehmen hervorbringen, welche viele Arbeitskräfte anheuern. Wenn die Arbeitslosenquote dadurch getrieben würde, so haben wir dies nicht in unserer Analyse berücksichtigt und ziehen dadurch falsche Schlüsse über den eigentlichen Treiber der Arbeitslosenquote.

Um tatsächliche kausale Effekte messen zu können müssten wir entweder ein kontrolliert randomisiertes Experiment durchführen, oder ein natürliches Experiment nutzen (z.B. eine nicht antizipierte Gesetzesänderung o.ä.). Es gibt auch in Deutschland mögliche Kandidaten für solche natürlichen Experimente, welche wir heranziehen könnten, doch dies würde den Rahmen dieser Case-Study sprengen.

Ingesamt gibt uns diese Case Study schon einen tiefen Einblick in die regionalen Unterschiede innerhalb Deutschlands und sie deckt wichtige Faktoren auf, welche für die Arbeitslosenquote wichtig sind. Ihr Cousin in Spanien hatte recht mit der Aussage, dass die Arbeitslosenquote in Deutschland deutlich geringer ist als in Spanien, auch über die verschiedenen Regionen in Deutschland hinweg. Jedoch haben Sie in dieser Case Study einige Determinanten der Arbeitslosenquote kennen gelernt und können nun untersuchen, ob diese Determinanten, wie die öffentliche Verschuldung oder das BIP-Wachstum auch einen großen Teil der Arbeitslosenquote in Spanien erklären können.