Einführung in die statistische Datenanalyse mit R

Grundbegriffe und erste Schritte

David Benček

Wintersemester 2015/16

Kurze Wiederholung

Warum analysieren wir Daten?

- Beantwortung wissenschaftlicher Fragestellungen
- Unterstützung von Hypothesen
- Entwicklung von Theorien/Modellen

Weshalb benötigen wir spezielle Analyseprogramme?

- Professionalisierung traditioneller empirischer Forschung
- Transparenz der Forschung (**Replizierbarkeit**)
- Effizienter Umgang mit Daten

Lernziele

- Gezielte Verarbeitung von Daten
- Zusammenführen unterschiedlicher Quellen
- Analyse von Daten gemäß einer Fragestellung
- Kommunizieren und visualisieren der Erkenntnisse
- Beispiele
  - Übersichtstabellen von Daten
  - Deskriptive Statistiken
  - Regressionstabellen
  - Abbildungen

Beispiele I

Table 1: Datenauszug

Kreis	Einwohner	AL-Quote	Schulabschlüsse	Einwohner_75+
01001	82258	10.5	1282	7590
01002	235782	10.1	2186	18772
01003	210305	10.1	2100	22059
01004	77249	10.9	1087	7432
01051	133900	7.4	1553	13381
01053	187137	6.0	1809	17514
01054	163665	6.4	1979	15441
01055	198413	6.4	2148	21452

Beispiele II

Beispiel III

```
##      crime          poverty         single     
##  Min.   :  82.0   Min.   : 8.00   Min.   : 8.40  
##  1st Qu.: 326.5   1st Qu.:10.70   1st Qu.:10.05  
##  Median : 515.0   Median :13.10   Median :10.90  
##  Mean   : 612.8   Mean   :14.26   Mean   :11.33  
##  3rd Qu.: 773.0   3rd Qu.:17.40   3rd Qu.:12.05  
##  Max.   :2922.0   Max.   :26.40   Max.   :22.10
```

Beispiel IV(a)

```
##
## Call:
## lm(formula = crime ~ poverty + single, data = cdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.14 -114.27  -22.44  121.86  689.82
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1368.189    187.205  -7.308 2.48e-09 ***
## poverty         6.787      8.989   0.755    0.454
## single        166.373     19.423   8.566 3.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 243.6 on 48 degrees of freedom
## Multiple R-squared:  0.7072, Adjusted R-squared:  0.695
```

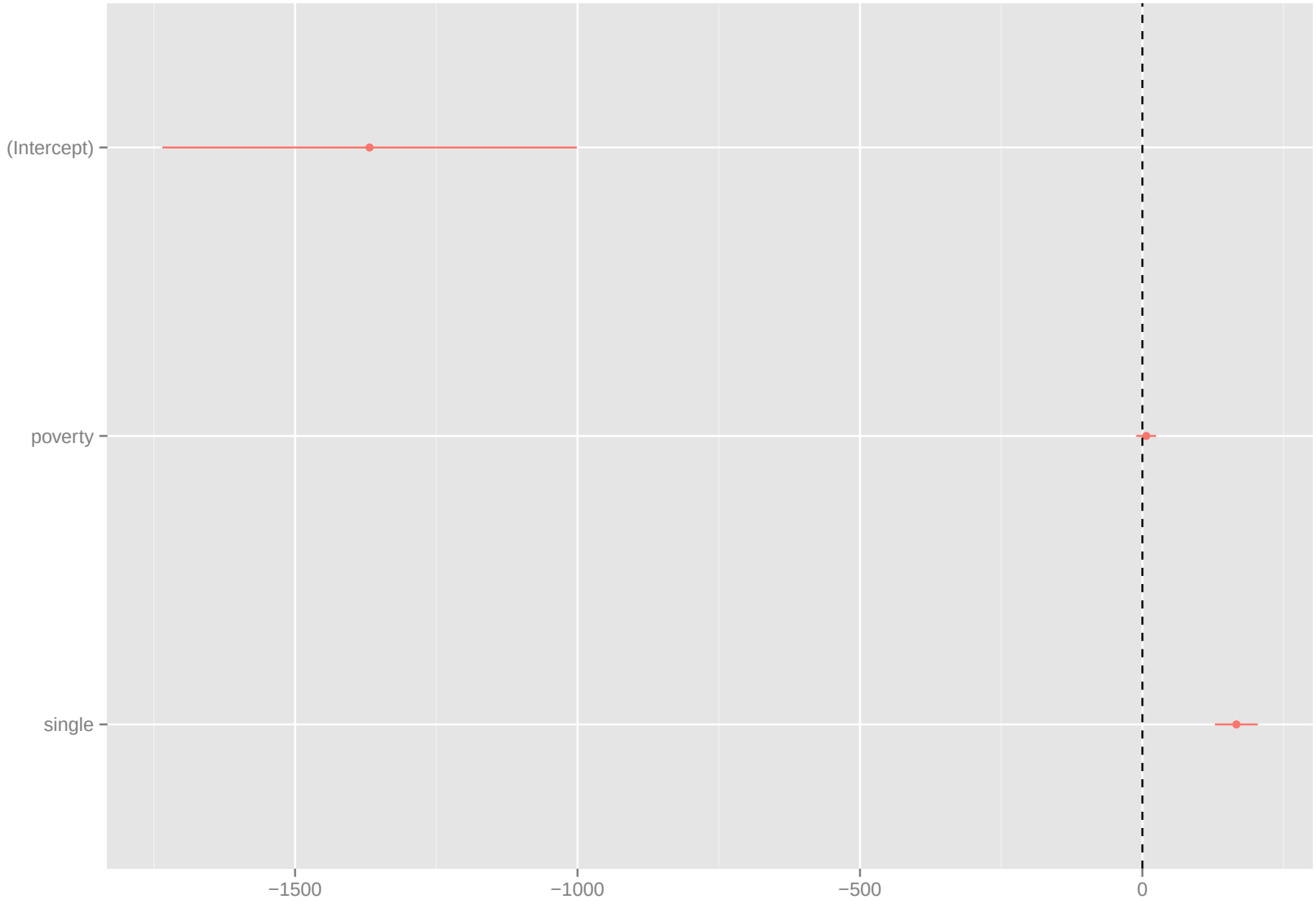
Beispiel IV(b)

Table 2: Statistical models

	Model 1
(Intercept)	-1368.19^{***}
	(187.21)
poverty	6.79
	(8.99)
single	166.37^{***}
	(19.42)
R^2	0.71
Adj. R^2	0.69
Num. obs.	51
RMSE	243.61

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Beispiel IV(c)



Variablen

- Variablen sind “Veränderliche”
- variieren je nach Analyseeinheit z.B. zwischen
  - einzelnen Personen,
  - unterschiedlichen Orten oder
  - Zeitpunkten

Beispiele?

Abhängige und unabhängige Variable

Eine wissenschaftliche Fragestellung enthält immer etwas, das Sie erklären möchten, und etwas, das Sie zu Erklärung anbieten.

$\Rightarrow$ Wirkung hervorgerufen durch einen Ursache

$$y_i = a + bx_i + e_i$$

Häufige Bezeichnungen:

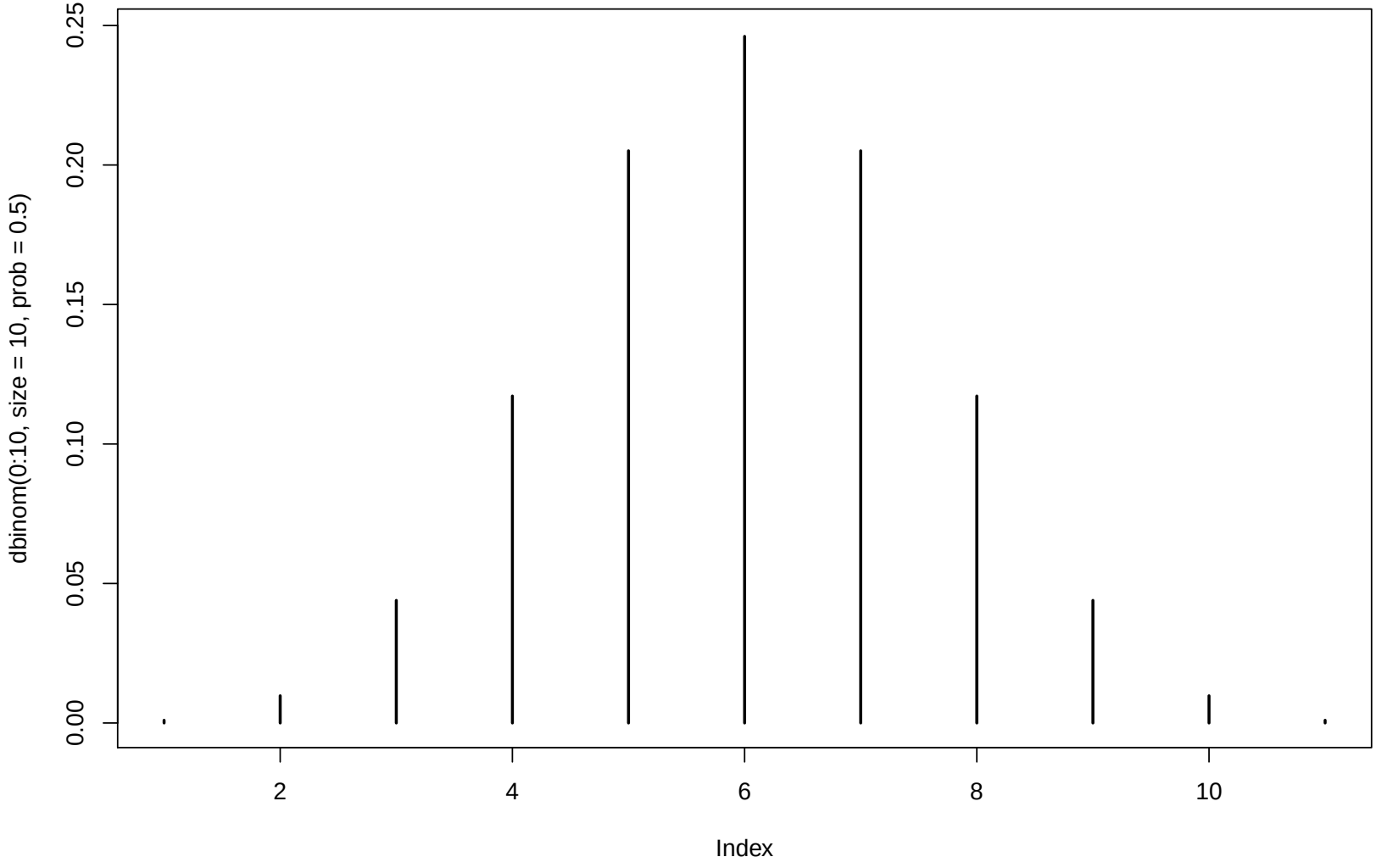
y_i	x_i
abhängige Variable	unabhängige Variable
erklärte Variable	erklärende Variable
endogene Variable	exogene Variable
Regressand	Regressor

Variablentypen

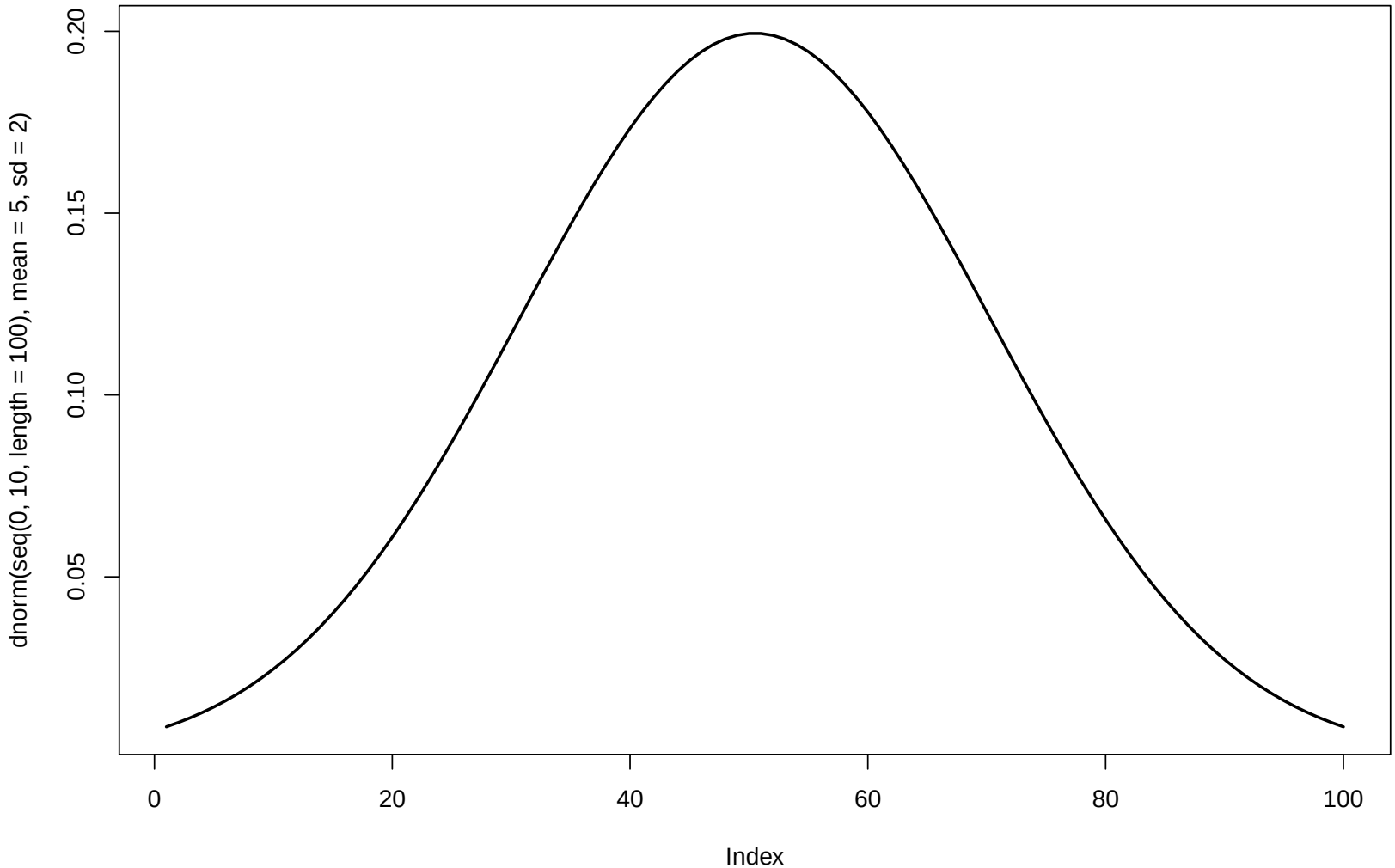
Je nach Messkonzept nutzen wir unterschiedliche Variablentypen (in R):

Typ	Beschreibung
binary	0/1; positiv/negativ; ja/nein
integer	diskrete Skala: 1–10
numeric	stetige Skala (inkl. Zwischenwerten)
factor (unordered)	Kategorien: SPD/CDU/Grüne
factor (ordered)	geordnete Kategorien: kalt/lauwarm/heiß
character	Text
Date	Datum: “2015-10-26”

integer


dbinom(0:10, size = 10, prob = 0.5)
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
Index

numeric


dnorm(seq(0, 10, length = 100), mean = 5, sd = 2)
0.05
0.10
0.15
0.20
0
20
40
60
80
100
Index

Datenstrukturen

Struktur	Eigenschaft
vector	homogen
matrix	homogen
array	homogen
list	heterogen
data frame	heterogen

Daten einlesen

```
gdp <- read.csv(file = "https://github.com/davben/\n
```

Zahlreiche weitere Befehle, abhängig vom Dateityp der Datenquelle!

Im Optimalfall: **Rdata-files**

Daten begutachten

```
load(file = "./data/gdp_deu.csv")

head(gdp_deu)
tail(gdp_deu)

str(gdp_deu)
dim(gdp_deu)
summary(gdp_deu)

table(gdp$country)
```

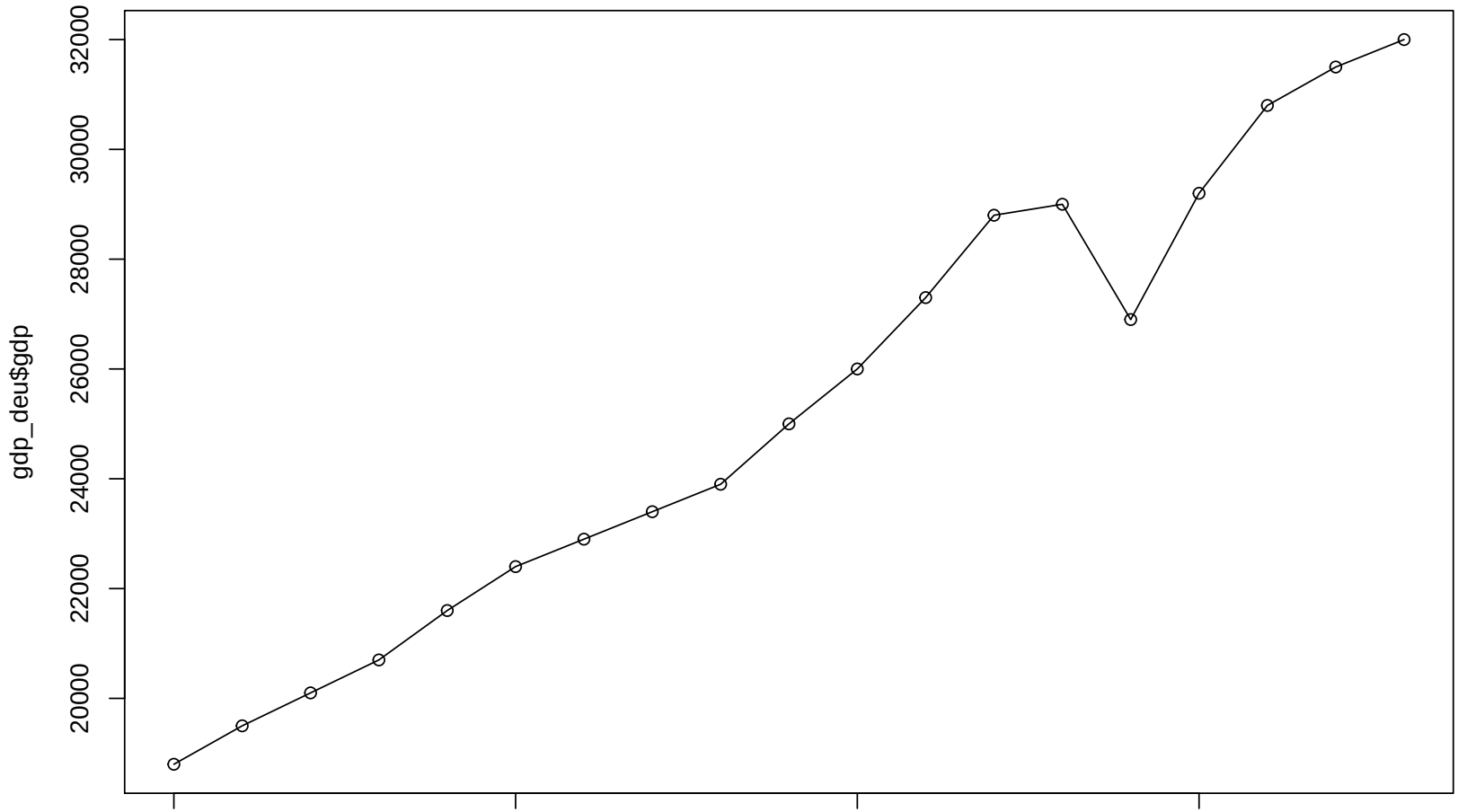
Style Guide

- aussagekräftige Namen von Objekten
- einheitliche Bezeichnung (R unterscheidet “Objekt” von “objekt”!)
- einheitliche Schreibweise (gdp_deu, gdp.deu, gdpDeu)
- Kommentare im Code! Sie sollten in zwei Monaten noch in der Lage sein, Ihren Code zu verstehen.
- saubere Formatierung, um Code übersichtlich zu halten.

Zum Nachlesen: `http://adv-r.had.co.nz/Style.html`

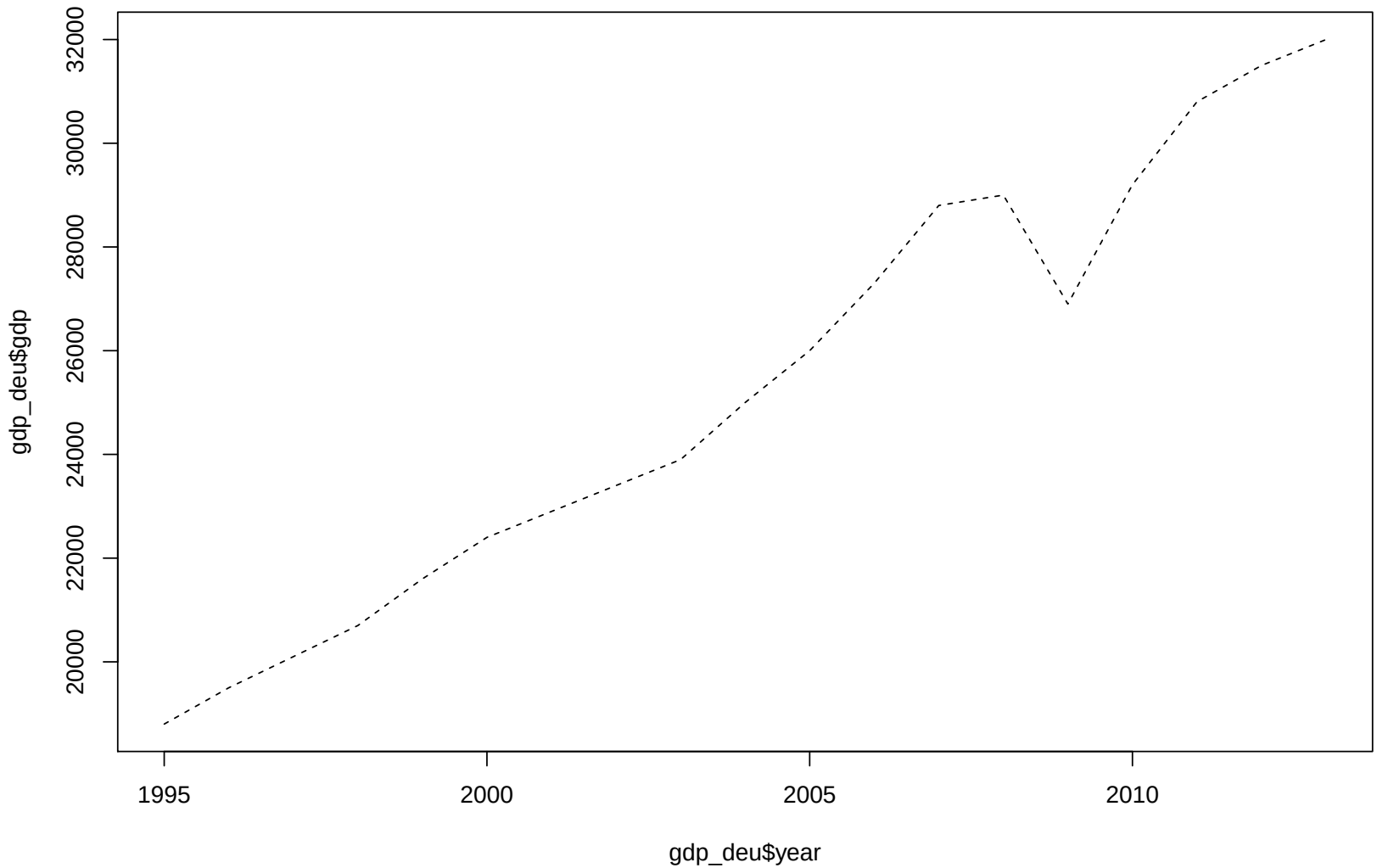
Erste Grafiken

```
load("./data/gdp_deu.Rdata")
plot(gdp_deu$year, gdp_deu$gdp)
lines(gdp_deu$year, gdp_deu$gdp)
```



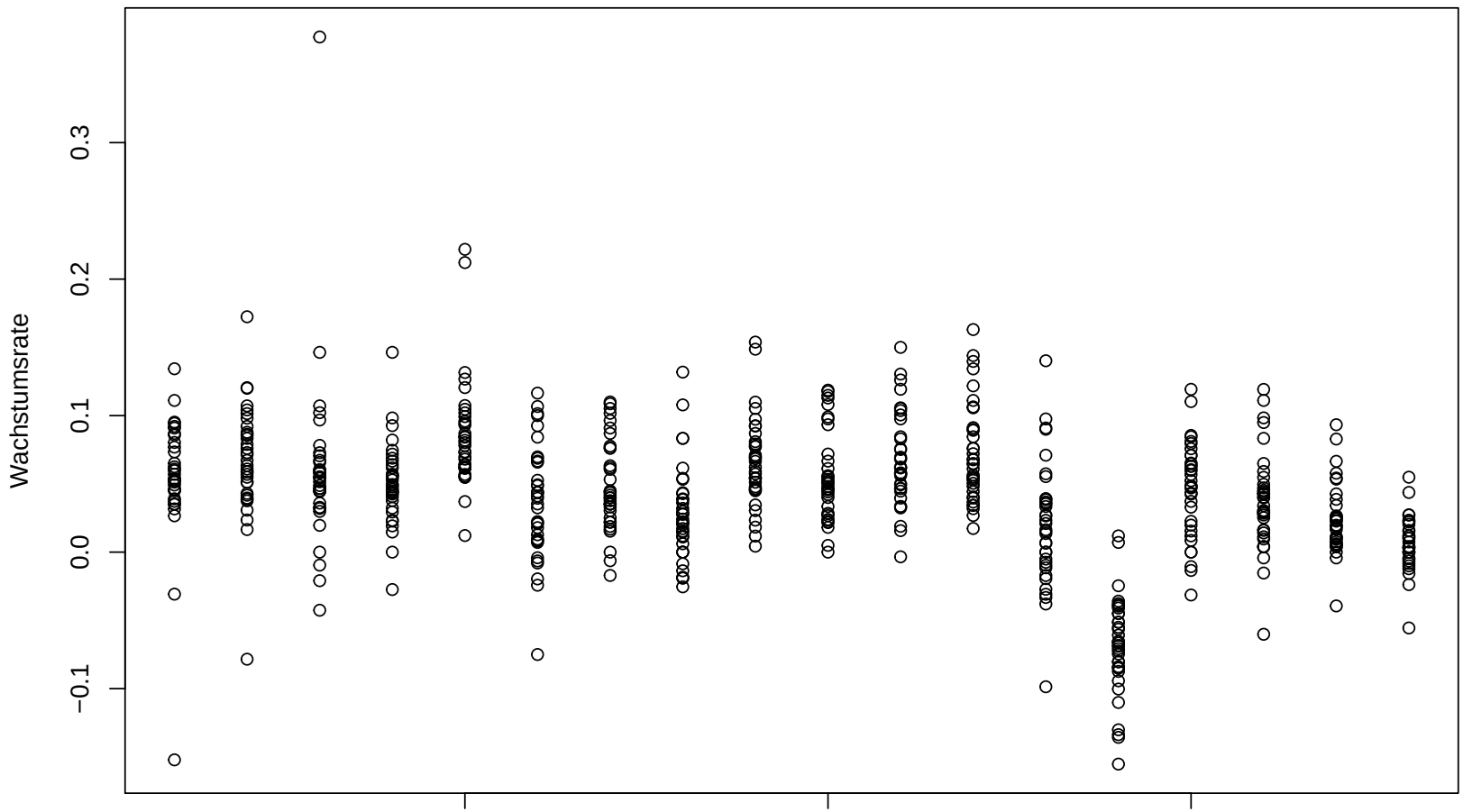
Erste Grafiken II

```
plot(gdp_deu$year, gdp_deu$gdp, type = "l", lty = 2)
```



Erste Grafiken III

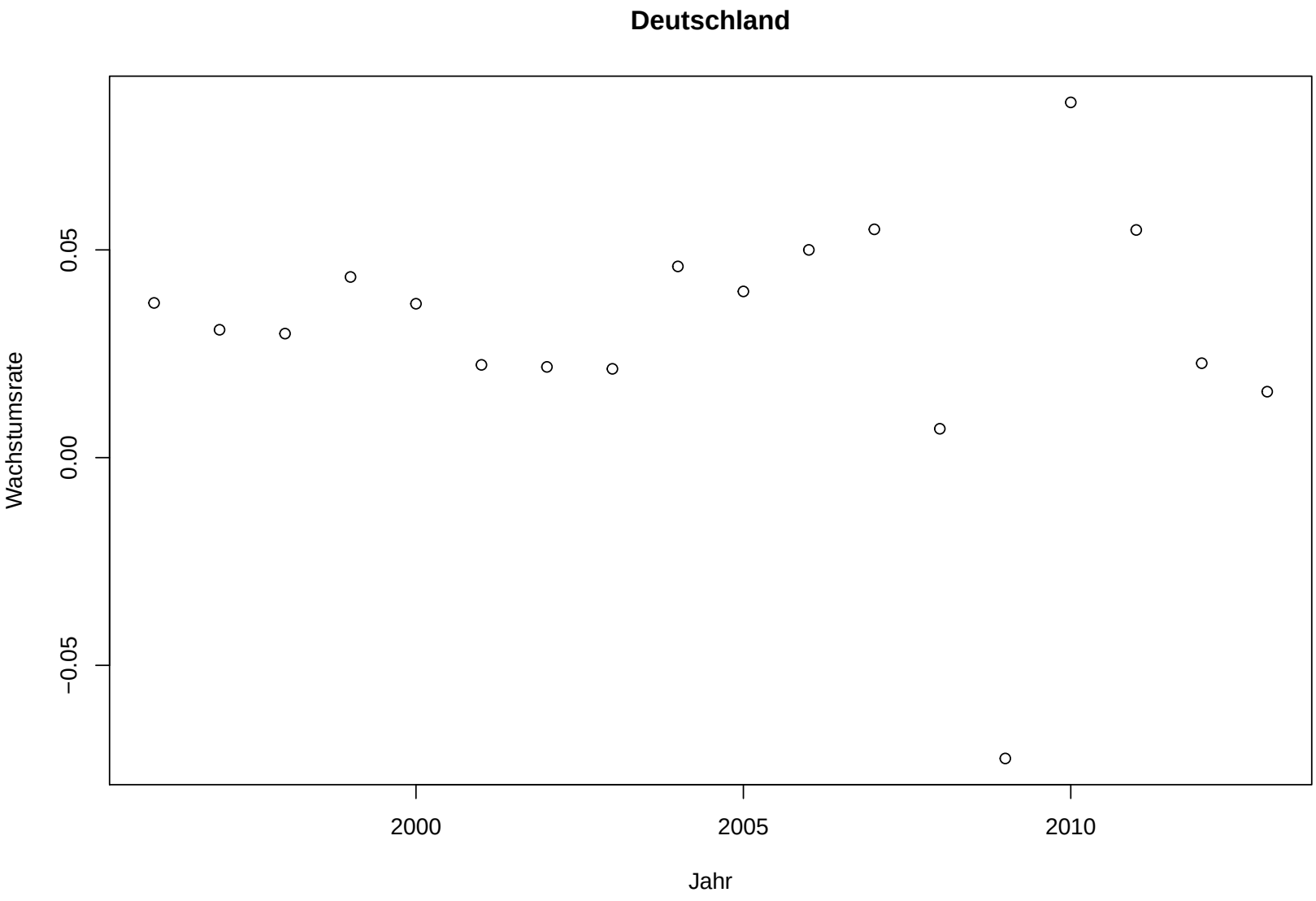
```
load("./data/gdp_growth.Rdata")
plot(gdp_growth$year, gdp_growth$growth, xlab = "Jahr",
     ylab = "Wachstumsrate")
```



Erste Grafiken IV(a)

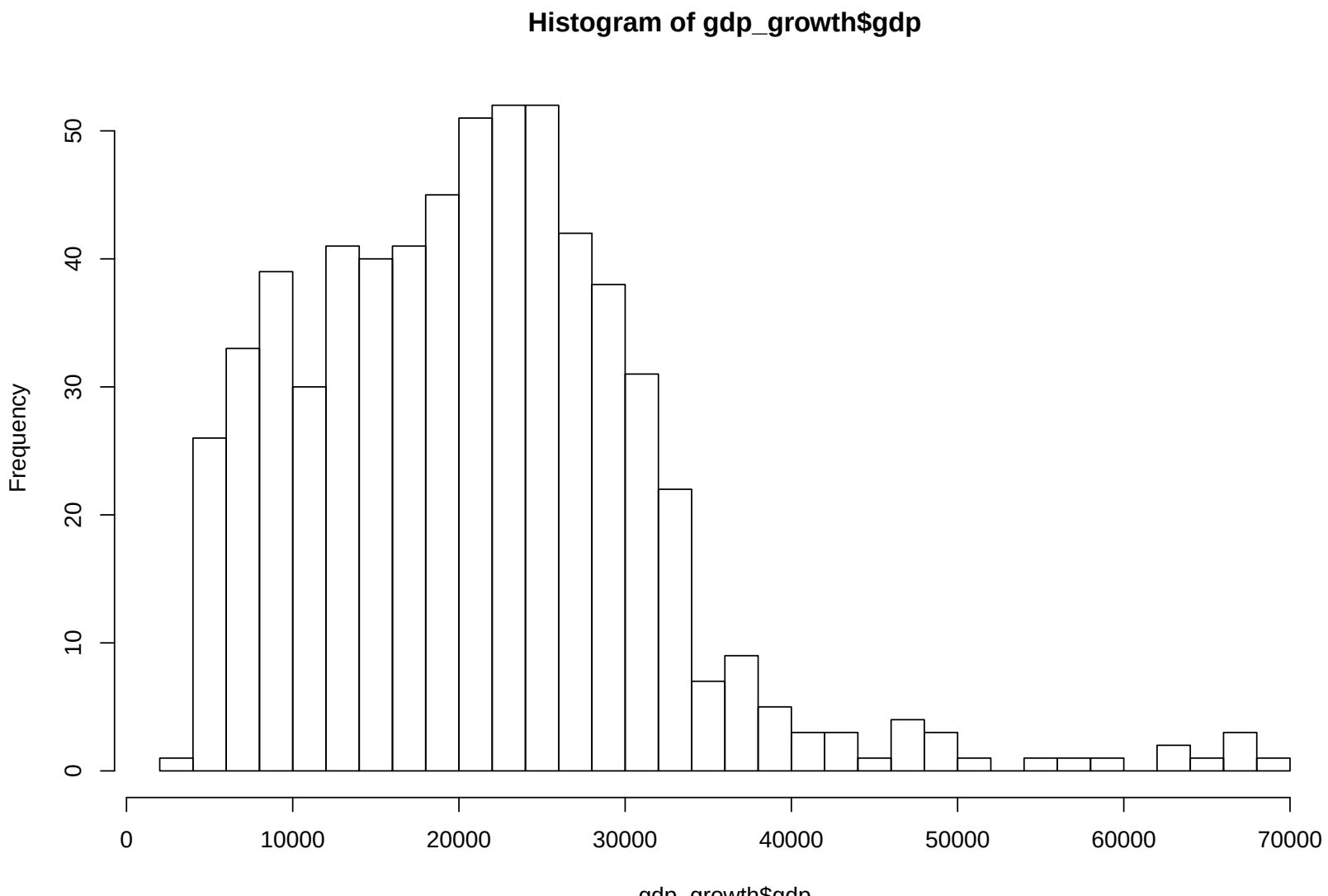
```
# nur Deutschland
plot(gdp_growth[gdp_growth$country == "Germany", ]$year,
     gdp_growth[gdp_growth$country == "Germany", ]$growth,
     xlab = "Jahr",
     ylab = "Wachstumsrate",
     main = "Deutschland")
```

Erste Grafiken IV(b)



Erste Grafiken V

```
# Verteilung
hist(gdp_growth$gdp, breaks=30)
```



Erste Grafiken VI

```
hist(gdp_growth$growth, breaks=30)
```