Statistik und Datenanalyse: Aufbau

Übung 1 – 4. Sitzung

Benjamin Fretwurst

▶ PDF-Version der Folien

Inhalt

- Übung 1 a+b
- Take Home – Ausblick – Vokabeln

Orga

Orga

- Es gibt neue Studienteilnahmestudien!

Lernziele

Grundprinzipien der Regression

- Sinn und Zweck von Quarto
- Einübung Regressionsanalyse
- Übung zur Berechnung von b_2, um zu verstehen, wie die Zusammenhänge bivariat angelegt sind und die übrigen Dritteinflüsse “herausgerechnet” werden
- Vor- und Nachteile der visuellen Residualanalyse vs. Tests

Übung 1 a+b

Ü1.1 Erstellen Sie eine Quarto-Datei.qmd

1. Öffnen Sie R-Studio
2. In R-Studio ➪ File ➪ New File ➪ Quarto Document…
3. Klicken Sie unten links auf «Create Empty Document»
4. (Wählen Sie als `title` «Erste Regression»)
5. Fügen Sie einen r-Chunk hinzu mit diesem Schalter: 
6. speichern Sie an einem günstigen Ort
   (am besten in der Cloud + nicht auf Desktop)

Ü1.2 Installation und Setup

Die R-Befehle für die Installation von Paketen haben wir in die Datei “Installation.R” ausgelagert, weil man sie im Grunde jeweils nur einmal braucht. Ich habe Ihnen hier eine Installationsdatei gebaut, mit der Sie die Pakete mit höhrer Erfolgschance installieren können. Mit folgendem Befehl wird diese Datei automatisch von unserer Homepage heruntergeladen und im Unterordner “files” des Projekts gespeichert. Sie können Sie dann dort öffnen und (am besten Zeilenweise) ausführen, wenn Sie die Pakete noch nicht installiert haben.

```
1 # Prüfe, ob es in dem Ordner in der die Uebung_1_ab.qmd gespeichert ist, ei
2 if(dir.exists("files")){} else {dir.create("files")}
```

```
NULL
```

```
1 # Lade die Installations.R herunter und speichere sie im Unterordner des Pr
2 download.file("https://stat.ikmz.uzh.ch/Aufbau/Folien/Sitzung_04/files/Inst
```

Für generelle Grundeinstellungen haben wir eine “_common.R” angelegt, in der wir den Aufruf der Basispakete des tidyverse geschrieben haben und andere Optionen und Einstellungen für Designs (wie Farben). Die Datei kann man dann immer am Anfang seiner Quarto-Dateien aufrufen und braucht diese Generaleinstellungen nicht immer wieder neu kopieren. Das ist doch praktisch.

```
1 # Prüfe, ob es in dem Ordner in der die Uebung_1_ab.qmd gespeichert ist, ei
2 if(dir.exists("files")){} else {dir.create("files")}
```

```
NULL
```

```
1 download.file("https://stat.ikmz.uzh.ch/Aufbau/Folien/Sitzung_04/files/_comm
2
3 source("files/_common.R")
```

Ü1.3 Laden Sie die Daten

Laden Sie den Fragebogen hier runter und schauen ihn an.

Laden Sie die Daten und lassen Sie mal die Variablenlabel raus:

- R-Code anzeigen

Ü1.4 Rechnen Sie ein Regressionsmodell

```r
DATEN <- DATEN  |> haven::zap_formats()

Modell_1 <- lm(E201_10 ~ E102_02, data = DATEN)

summary(Modell_1)
##
## Call:
## lm(formula = E201_10 ~ E102_02, data = DATEN
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2726 -0.7684 -0.2641  0.7274  2.7274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>
## (Intercept)  1.28105    0.16912   7.575 2.60
## E102_02      0.49578    0.06668   7.435 5.73
```

Was sehen Sie?

1. Wie gross ist R^2?
2. Wie gross ist die bivariate Korrelation r? (selbst ausrechnen)
3. Ist der Zusammenhang positiv oder negativ?
4. Ist der Zusammenhang signifikant?

Ü1.5 Verändern Sie das Regressionsmodell

Kopieren Sie den r-Chunk der letzten Folie und setzen Sie andere Variablen ein: Nehmen Sie die Variablen für «Statistik Einführung hat mir viel Spass gemacht» und erklären Sie damit: «Ich freu mich auf Statistik Aufbau!».

Beantworten wieder die Fragen:

1. Wie gross ist R^2?
2. Wie gross ist die bivariate Korrelation r? (selbst ausrechnen)
3. Ist der Zusammenhang positiv oder negativ?
4. Ist der Zusammenhang signifikant?

▶ Code

Ü1.6 b_2 aus Korrelationen und SDs berechnen

Note

Lassen Sie die Korrelationen durchlaufen, schauen Sie sich an, wo was steht und setzen Sie es in die Formel für $b_2 = \frac{r_{Y2} - r_{23} r_{Y3}}{(1 - R^2_{2.3})} \frac{s_Y}{s_2}$, um es zu berechnen.

Also: $b_2 = \frac{.50 - (-.32 \cdot -.23)}{(1 - .32^2)} \frac{1.05}{1.07}$

▶ R-Code anzeigen

Man kann natürlich auch R nutzen

- R-Code anzeigen

Ü1.7 Berechnen Sie b_2 mit Hilfe einer Regressionsanalyse

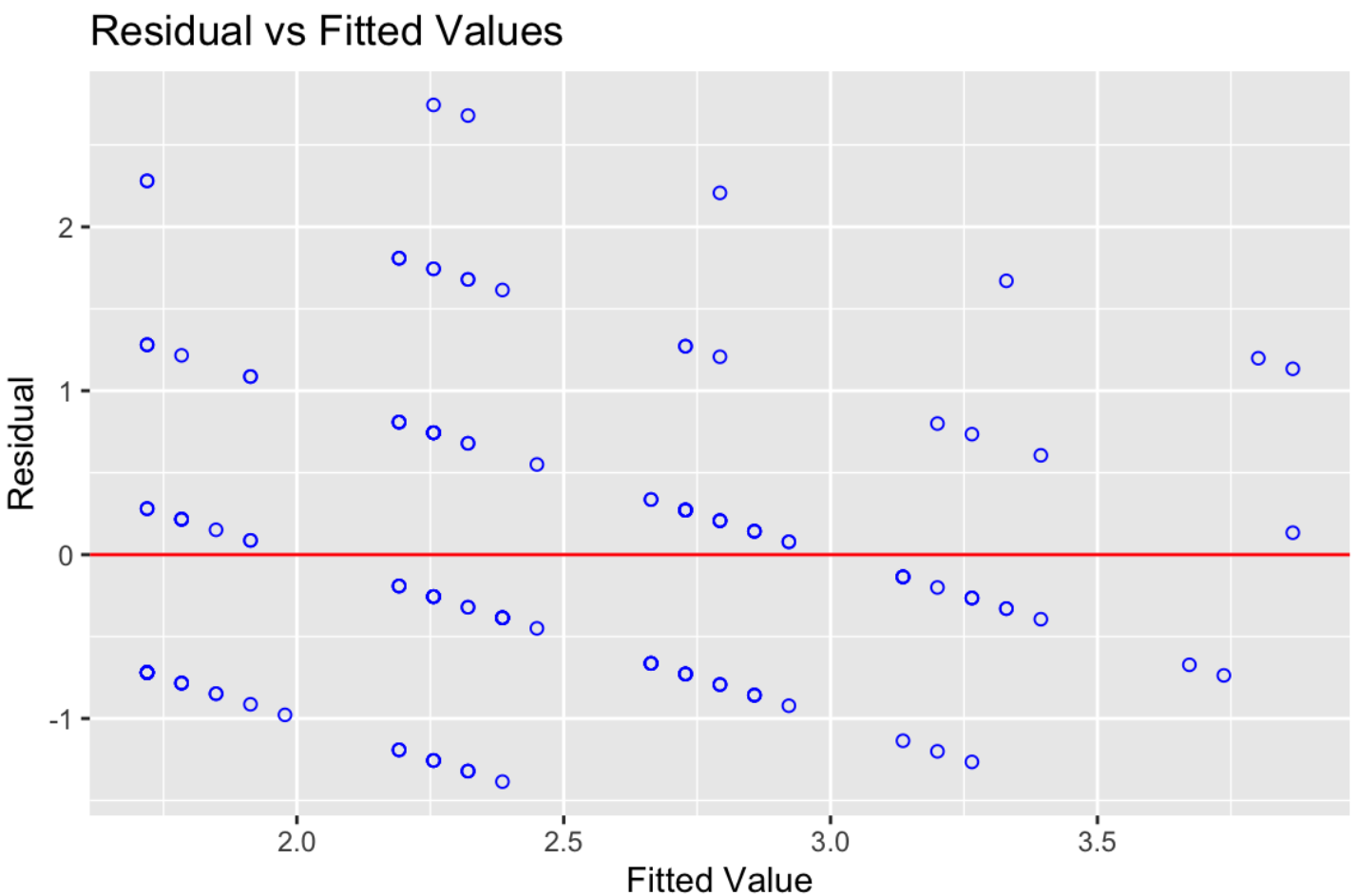
```r
Modell1 <- lm(E201_10 ~ E102_02 + E102_04, data = DATEN)

Modell_1_beta <- lm.beta::lm.beta(Modell1)

summary(Modell_1_beta, digits = digits, maxsum = maxsum)
##
## Call:
## lm(formula = E201_10 ~ E102_02 + E102_04, data = DATEN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3852 -0.7194 -0.2281  0.5642  2.7439
##
## Coefficients:
##             Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)  1.57016           NA    0.32129   4.887 2.46e-06 ***
## E102_02      0.47210      0.48031    0.07031   6.714 3.08e-10 ***
```

Ü1.8 Geben Sie mit folgendem Befehl die Tolerance und VIF-Werte raus

```
olsrr::ols_regress(Modell1)
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.509       RMSE                 0.915
## R-Squared               0.260       Coef. Var           37.790
## Adj. R-Squared          0.250       MSE                  0.837
## Pred R-Squared          0.235       MAE                  0.736
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares          DF    Mean Square      F          Sig.
## --------------------------------------------------------------------
```

Ü1.9 Schauen Sie sich die Residualplotts an

- R-Code anzeigen



Plot der Residuen

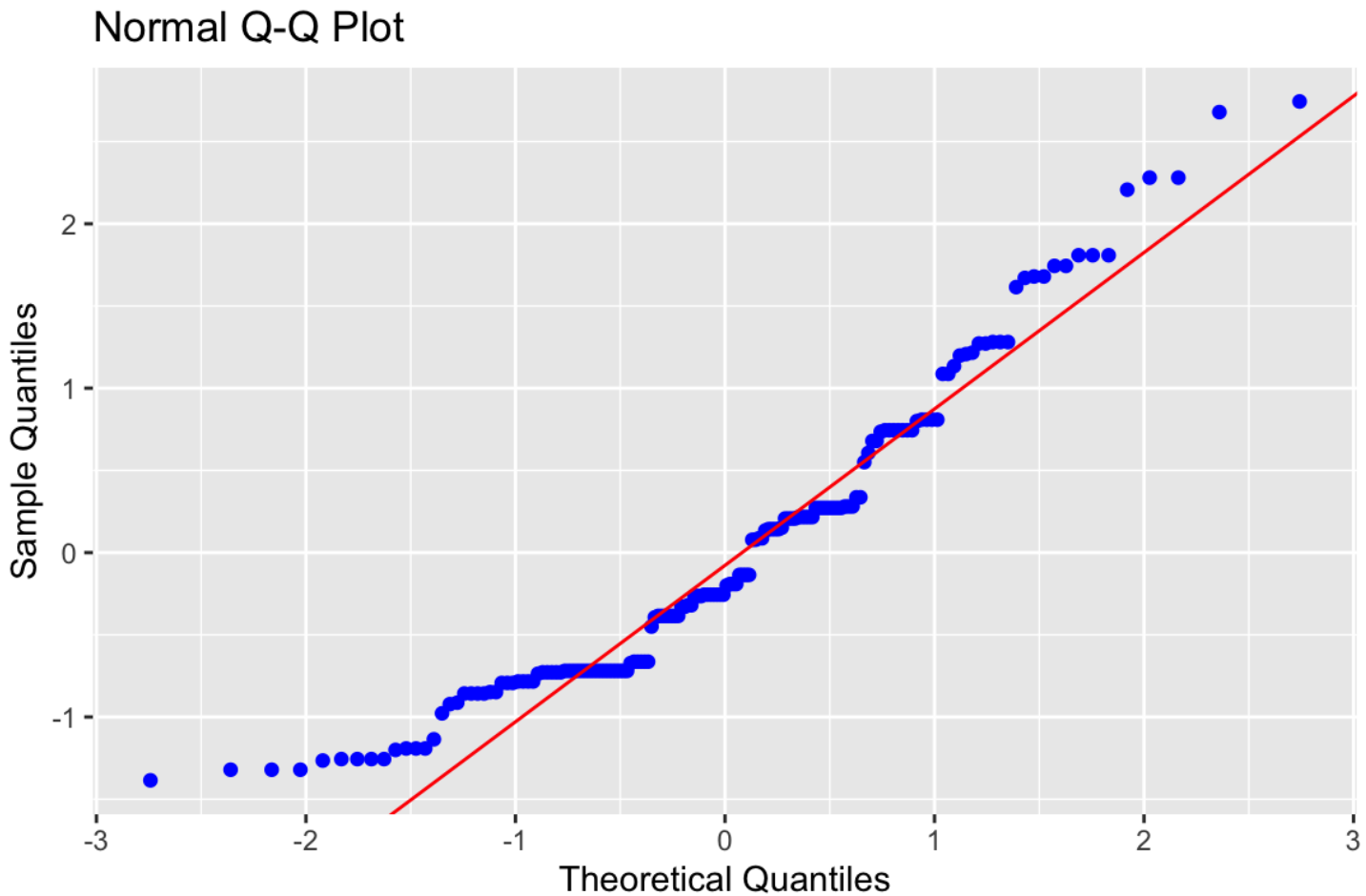
Ü1.10 Testen Sie auf Homoskedastizität

- R-Code anzeigen

Was sagt Ihnen das?

Ü1.11 Gucken Sie sich den N-Q-Q-Plot an

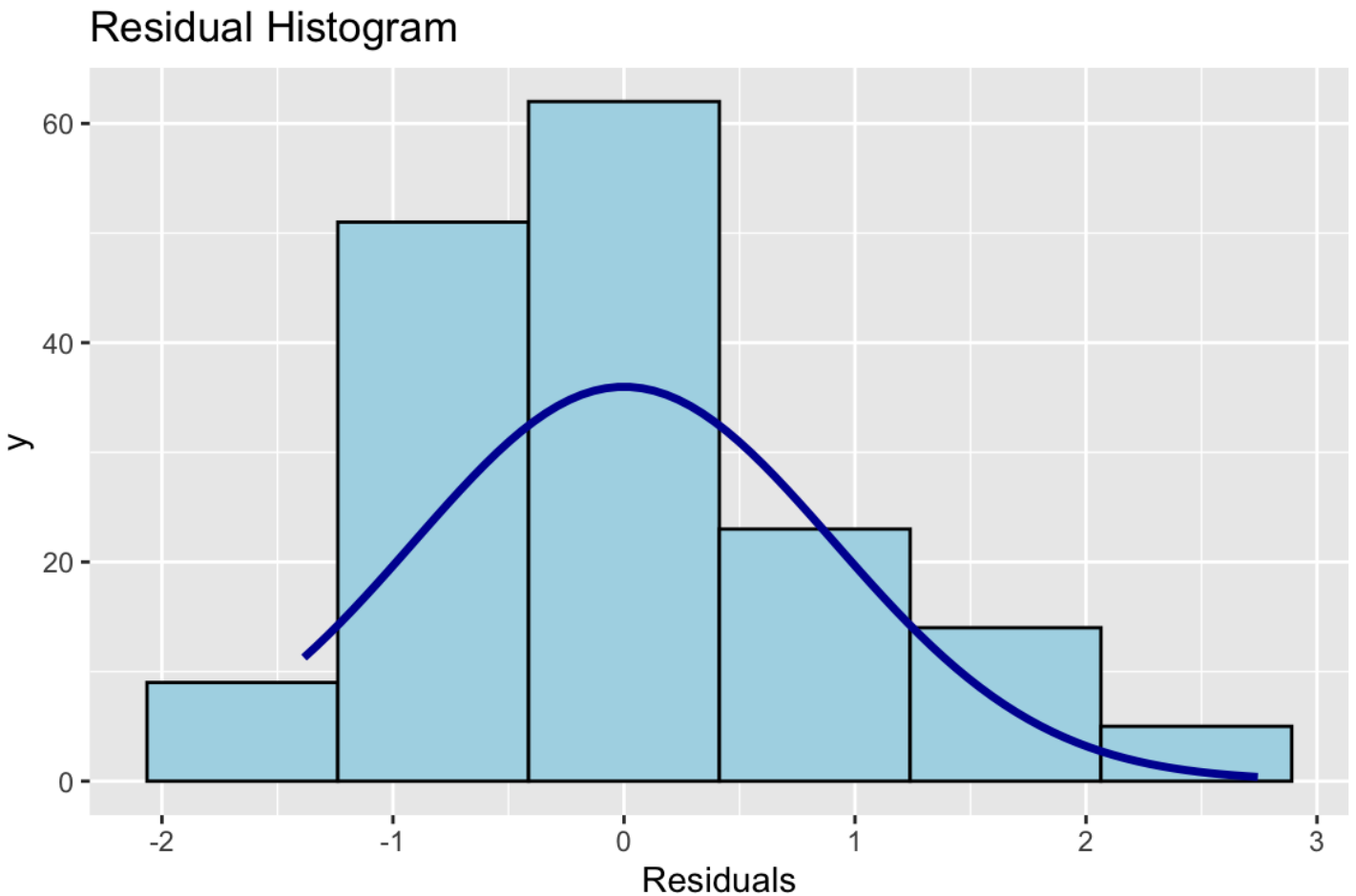
- R-Code anzeigen



Normal-Q-Q-Plot

Ü1.12 Und das Histogramm

- R-Code anzeigen



Histogramm der Residuen

Ü1.13 Jetzt auf Normalverteilung testen

```
# Führe Tests auf signifikante Verletzungen
# der Normalverteilungsannahme aus.

olsrr::ols_test_normality(Modell1)
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9349         0.0000
## Kolmogorov-Smirnov        0.127          0.0101
## Cramer-von Mises          15.381         0.0000
## Anderson-Darling          3.2969         0.0000
## -----------------------------------------------
```

Ü1.14 Fazit

Was ist Ihr Fazit aus der Regressionsrechnung?

Weiterführung

Predictors	B	BETA	std.err	t	p
(Intercept)	1.57		0.32	4.89	<.001
E102_02	0.47	.480	0.07	6.71	<.001
E102_04	-0.06	-.076	0.06	-1.06	.292

[a] R^2 = 0.25 (F = 28, df = 161, p = 161), $R^2_{adj.}$ = 0.25

Take Home – Ausblick – Vokabeln

Take Home

Note

- Sie können eine Regressionsanalyse in Quarto berechnen und die Ergebnisse interpretieren
- Ihnen ist klar, was es bedeutet, dass Dritteinflüsse herausgerechnet werden.
- Sie können Residualanalysen anschauen und erkennen, wann es Probleme gibt (mögliche Lösungen kommen später)
- Sie können den notwendigen Code für Regressionsanalysen in Ihre Projekte kopieren und an den richtigen Stellen anpassen.

Ausblick

Wie beschäftigen uns mit kategorialen Variablen in den UVs.

Vokabeln 🔗

Search:

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
All	All	All	All	All	All
46	3	Voraussetzungen	Allgemeine kleinste Quadrate	Generalized Least Squares (GLS)	Schätzmetho OLS ersetzen wenn es Heteroskeda gibt, also die Residuen nic allen stellen streuen.
43	3	Voraussetzungen	BLUE	BLUE	Akronym für Linear Unbias Estimator
42	3	Voraussetzungen	Bias	Bias	Grad der Verzerrung

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
44	3	Voraussetzungen	Effizienz	efficiency	Die Genauigk eines Schätz also wie starl streut.
45	3	Voraussetzungen	Fehlervarianz	error variance	Streuung ein Kennwertes (b's).
47	3	Voraussetzungen	Heteroskedastizität	heteroscedasticity	Die Residuen streuen nicht gleichmässig nach grösser UV. Also häng Streubreite d Residuen mit Grösse einer zusammen.
48	3	Voraussetzungen	Homoskedastizität	homoscedasticity	Die Residuen streuen

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
					gleichmässig
49	3	Voraussetzungen	Modellspezifikation	model specification	Formulierung Modells, also welche UVs f AV wichtig si und wie dere Beziehung gestaltet ist.
50	3	Voraussetzungen	Multikollinearität	multicollinearity	Eine UV häng einer oder mehreren der übrigen UVs zusammen.
51	3	Voraussetzungen	Toleranz (TOL)	tolerance	Die übrige Va die eine Varia noch hat, wel gemeinsame Varianz mit a

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
					anderen UVs rausgerechne wurde.
53	3	Voraussetzungen	Unterspezifikation	under estimation	Zu wenige UV Modell (hat verzerrt b's zu Folge).
54	3	Voraussetzungen	Unverzerrtheit	unbiasedness	Eigenschaft e Methode vali Messungen d Kennwerte fü Parameter zu messen.
55	3	Voraussetzungen	Varianzinflationsfaktor (VIF)	variance inflation factor	Der Faktor, ur die Ungenaui (Fehlervarian einer UV steig wenn Multikollinea