Statistik und Datenanalyse: Aufbau

Regression – Voraussetzungen für BLUE

Benjamin Fretwurst

▶ PDF-Version der Folien

Inhalt

Orga

Orga

- Syllabus
- Abmeldung vom Leistungsnachweis via Antrags-Cockpit möglich
- Prüfungsrelevanz von R
- Pakete laden in R für die Übung
- Ihre Fragen? Kritiken? Hinweise?

Lernziele

Grundprinzipien der Regression

- Die Regressionsgleichung
- Das OLS-Prinzip
- Die Voraussetzungen für BLUE
    1. Fixe X und Y
    2. keine perfekte Multikollinearität
    3. keine hohe Multikollinearität
    4. erschöpfende Modellspezifikation
    5. Homoskedastizität
    6. Linearität der Zusammenhänge

1 Die bivariate Regressionsgleichung

$$\text{GG: } Y_i = \beta_1 + \beta_2 X_{i2} + U_i \tag{1}$$

$$\text{Stichprobe: } Y_i = b_1 + b_2 X_{i2} + e_i \tag{2}$$

Modell und Schätzung

Das Regressionsmodell für die Zusammenhänge in der GG wird durch die Berechnung der b's in der Stichprobe geschätzt. Alles was ein Subscript «i» hat, ist in der Stichprobe fix. Es bleiben nur die b's zu schätzen.

$$b_2 = r_{Y2} \frac{S_Y}{S_2} \tag{3}$$

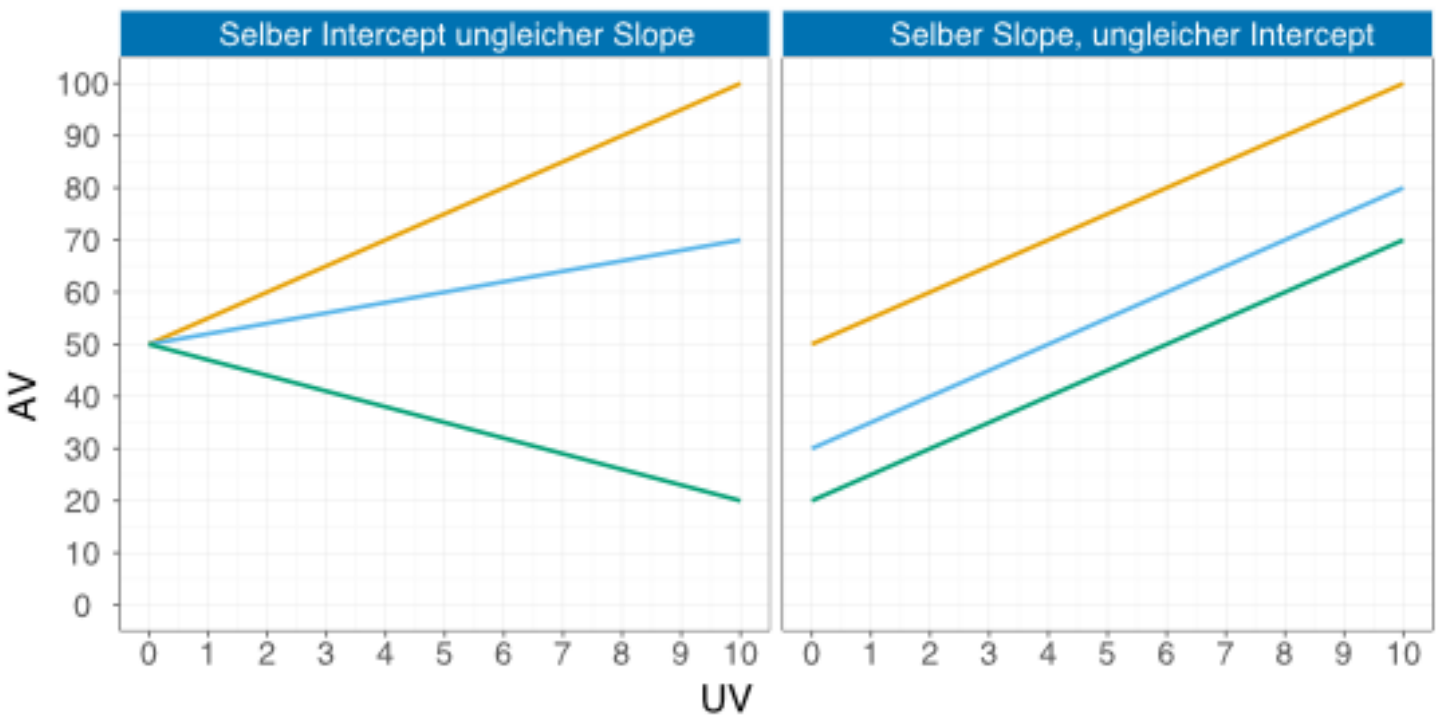
1.1 Die Schätzer streuen je nach Stichprobe

Die b's sind Stichprobeneigenschaften, wobei b_1 und b_2 an der Realisation der Stichprobe hängen, also «schwanken».

```
RuntimeError: e.fileOf is n
RuntimeError: e.fileOf is n
RuntimeError: e.fileOf is n
```

1.2 Parameter einer Regressionsgeraden

- Gerade mit unterschiedlichen Anstiegen, aber gleichem Schnittpunkt in Y.
- Gerade mit gleichem Anstieg, aber unterschiedlichen Schnittpunkten in Y.



2 Regression multivariat mit 2 UV's

Regressionsgleichung

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i \quad (4)$$

$$Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i \quad (5)$$

2.1 OLS

Grundidee OLS

Wir suchen die b's. Die gesuchten b's sollen eine Regressionsgerade ergeben, die «optimal» in der Punktwolke der gemessenen Werte liegt. Wir suchen also die b's, die die kleinsten quadrierten Abweichungen zwischen den vom Modell vorhergesagten und den gemessenen Werten ergibt. Das «Prinzip der kleinsten Quadrate» wird als OLS bezeichnet (Ordenary Least Squares).

$$\sum_{i=1}^{n} e_i^2 \rightarrow minimal \tag{6}$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \rightarrow minimal \tag{7}$$

2.2 Die Formel für b nach OLS

$$b_2 = (r_{Y2} - r_{23} r_{Y3}) \frac{1}{1 - R^2_{2.3}} \frac{S_y}{S_2} \tag{8}$$

In Worten

Der Anstieg der «Regressionsgeraden» für X_2 ergibt sich aus der Korrelation r_{Y2} zwischen X_2 und Y, die um den vermittelten Zusammenhang über die Drittvariable, also das Produkt aus r_{23} und r_{Y3} reduziert wird. Der Rest sind Korrekturen damit, wie stark X_2 von den übrigen Variablen erklärt wird $\frac{1}{1-R^2_{2.3}}$ und quasi die Umkehr der Standardisierung $\frac{S_Y}{S_2}$.

3 BLUE – Best Linear Unbiased Estimators

«Linear Estimator»

Die «Linear Estimator» sind die b's, also b_1, b_2, ... b_k.

«Unbiased»

«Unbiased» bedeutet, dass wir unverzerrte Schätzer, also unverzerrte b's haben wollen. Die b's scätzen ihre βs unverzerrt, wenn die Streuung der b's um die wahren βs herum liegen (man sagt auch: «erwartungstreu»).

«Best» bezeichnet die Effizienz der Schätzer b

Die besten Schätzer erhalten wir, wenn die Standardfehler der b's (se_b) minimal sind.

3.1 «Unbiased», also Unverzerrtheit der b's

Die Variablen (X und Y) müssen fix sein

Wir müssen also davon ausgehen, dass die erhobenen Variablen bei einer nächsten Ziehung nicht ganz anders aussehen würden.

im U_i darf nur Rauschen sein

Es darf im Unbekannten U_i keine Variable stecken, die mit den UV's korreliert. Der Erwartungswert dieser Covarianz muss 0 sein: $E(C_{2U}) = 0 = E(C_{3U})$.

Modellspezifikation

Wir sollten aus der Theorie und in der Operationalisierung keine Variable vergessen, die mit den UVs zusammenhängt! Theoriearbeit besteht in der Suche nach der vollen Modellspezifikation! Die perfekte Modellspezifikation wäre das Ende der Forschung zu einem fixen Phänomen.

3.1.1 Unterspezifikation

Die Grösse des Bias bei Unterspezifikation

Gibt es eine X_{i4} mit einem wahren β_4 und ist dieses mit X_2 sowie Y korreliert, dann ist b_2 ein verzerrter Schätzer für β_2.

Verzerrung von b_2 wenn eine Einflussgrösse X_4 nicht mitgeschätzt wird

$$\text{wahr: } Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + U_i \tag{9}$$

$$\text{geschätzt: } Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i^\star \quad \text{wobei} \quad U_i^\star = \beta_4 X_{i4} + U_i \tag{10}$$

$$\rightarrow E(b_2) = \beta_2 + \beta_4 b_{42} \tag{11}$$

$$\text{mit: } b_{42} = \frac{r_{42} - r_{32} r_{43}}{1 - r_{32}^2} \sqrt{\frac{V_4}{V_2}} \tag{12}$$

3.2 Multikollinearität

Definition

Multikollinearität bedeutet, dass die Varianz einer Variablen durch eine oder mehrere übrige UVs teilweise aufgeklärt wird.

herausgerechnete Erklärungskraft

Wird einer Variablen viel Erklärungsvarianz ($R_{2.34...}$) weggerechnet, dann hat sie kaum noch welche, um die AV zu erklären.

Wann ein Problem

- Der Grund für Regressionsanalysen
- Problem hoher Multikollinearität (TOL < .5)
- Standardfehler ➪ Schätzqualität schlecht (VIF > 2)

3.2.1 Steigende Fehlerstreuung bei Multikollinearität

Fehlervarianz von b_2

$$s^2_{b_2} = \frac{s^2_e}{n} \cdot \frac{1}{V_2} \cdot \frac{1}{1-R^2_{2.34...}}$$

Die Fehlerstreuung des Regressionskoeffizienten b ist proportional zur Streuung der Fehler e_i und umgekehrt proportional zur Fallzahl n, der Varianz V_2 (also von X_2) und zu Multikollinearität bzw. Toleranz TOL = $1 - R^2_{2.34...}$.

Toleranz ist die exklusive Varianz einer UV

$$TOL_{b_2} = 1 - R^2_{2.34...}$$

Toleranz ist der Prozentsatz Varianz, der nicht durch die übrigen UVs rausgerechnet wird.

Der Varianz-Inflation-Factor VIF

$$VIF_{b_2} = \frac{1}{(1-R^2_{2.34...})} = \frac{1}{TOL_{b_2}}$$

3.3 Linearität

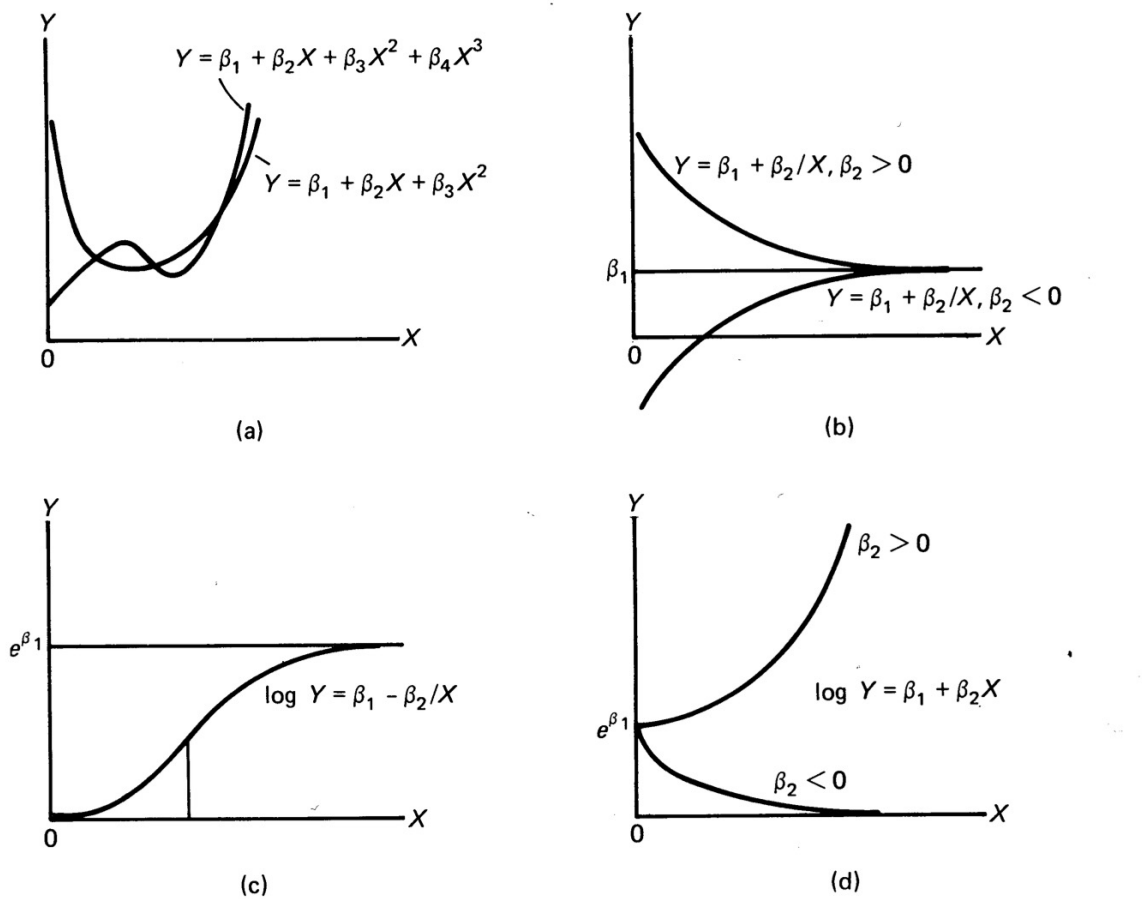
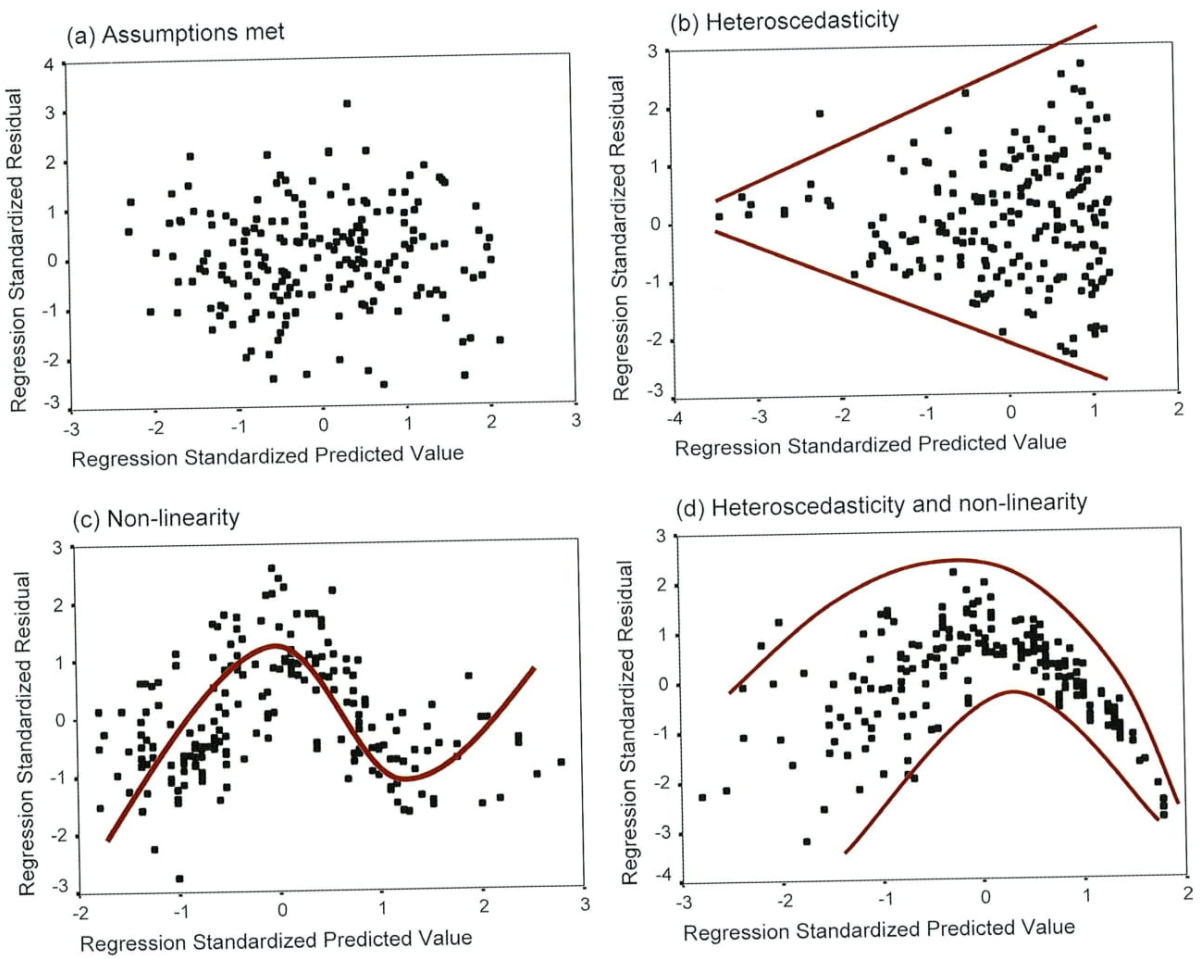


FIGURE 4.2 Alternative relationships: (a) polynomial, (b) reciprocal. (c) log–reciprocal, (d) semilog.

Nichtlineare Zusammenhänge

3.4 Heteroskedastizität



Ursachen für Heteroskedastizität

3.4.1 Heteroskedastizitätsproblem und -lösung

Probleme

- Die Residuen hängen mit X zusammen.
- Standardfehler der b verzerrt
- nichtlineare Zusammenhänge unerkannt

Lösungen

- Gibt es
- Generalized least Squares (GLS)
- Kurvlineare Schätzungen

3.5 Annahmen zur Residualverteilung

Normalverteilung und Unabhängigkeit der Residuen

Schaut man sich visuell an. Wenn sie stark verletzt ist (z.B. bimodal) oder extrem schief, dann andere Methode.

Unabhängigkeit der Fehler

Die Fehler können nur voneinander abhängig sein, bei zeitlich geordneten Erhebungen, also Zeitreihenanalysen. Das braucht uns also erstmal nicht kümmern.

3.5.1 Multivariat normalvereilt

```
TypeError: Cannot read properties of null (reading 'get
ShaderPrecisionFormat')
```

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 Schlater: 
6. speichern Sie an einem günstigen Ort
   (am besten in der Cloud + nicht auf Desktop)

Ü1.2 Installieren Sie ein paar Pakete

Kopieren Sie in Ihre Datei:

```r
## die einfache Variante
install.packages("tidyverse")

## die Quelle mit angegeben und alle abhängigen Pakete mit installieren
install.packages(c("ggpubr", "corrr", "olsrr"),
   repos = "https://cloud.r-project.org/",
   dependencies = TRUE)

## damit auch Developer-Versionen installiert werden können:
install.packages("devtools")

## Versuch über die Developer-Versionen
devtools::install_github("kassambara/ggpubr", force = TRUE)

# wird nur installiert, wenn es nicht schon in der aktuellsten Version da is
devtools::install_github("strengejacke/sjmisc")

```

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

```
CASE

"Interview-Nummer (fortlaufend)"

QUESTNNR

                                                                "Fragebogen,
der im Interview verwendet wurde"

MODE

"Interview-Modus"

STARTED

                                                        "Zeitpunkt zu dem das
Interview begonnen hat (Europe/Berlin)"
```


Ü1.4 Rechnen Sie ein Regressionsmodell

```r
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(>|t|)
(Intercept)  1.28105    0.16912   7.575 2.60e-12 ***
E102_02      0.49578    0.06668   7.435 5.73e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.9151 on 162 degrees of freedom
```

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 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?

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, um es zu berechnen.

▶ R-Code anzeigen

```
Means, standard deviations, and correlations with confidence intervals


  Variable    M    SD   1           2
  1. E201_10 2.35 1.37

  2. E102_02 2.19 1.42 .68**
                       [.59, .76]

  3. E102_04 3.53 1.61 .25**       .23**
                       [.10, .39] [.08, .37]
```

Note. M and SD are used to represent mean and standard deviation, respectively.

Ü1.7 Berechnen Sie b_2 mit Hilfe einer Regressionsanalyse

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

summary(Modell1, 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 Std. Error t value Pr(>|t|)
(Intercept)  1.57016    0.32129   4.887 2.46e-06 ***
E102_02      0.47210    0.07031   6.714 3.08e-10 ***
E102_04     -0.06458    0.06102  -1.058    0.292
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9148 on 161 degrees of freedom
```

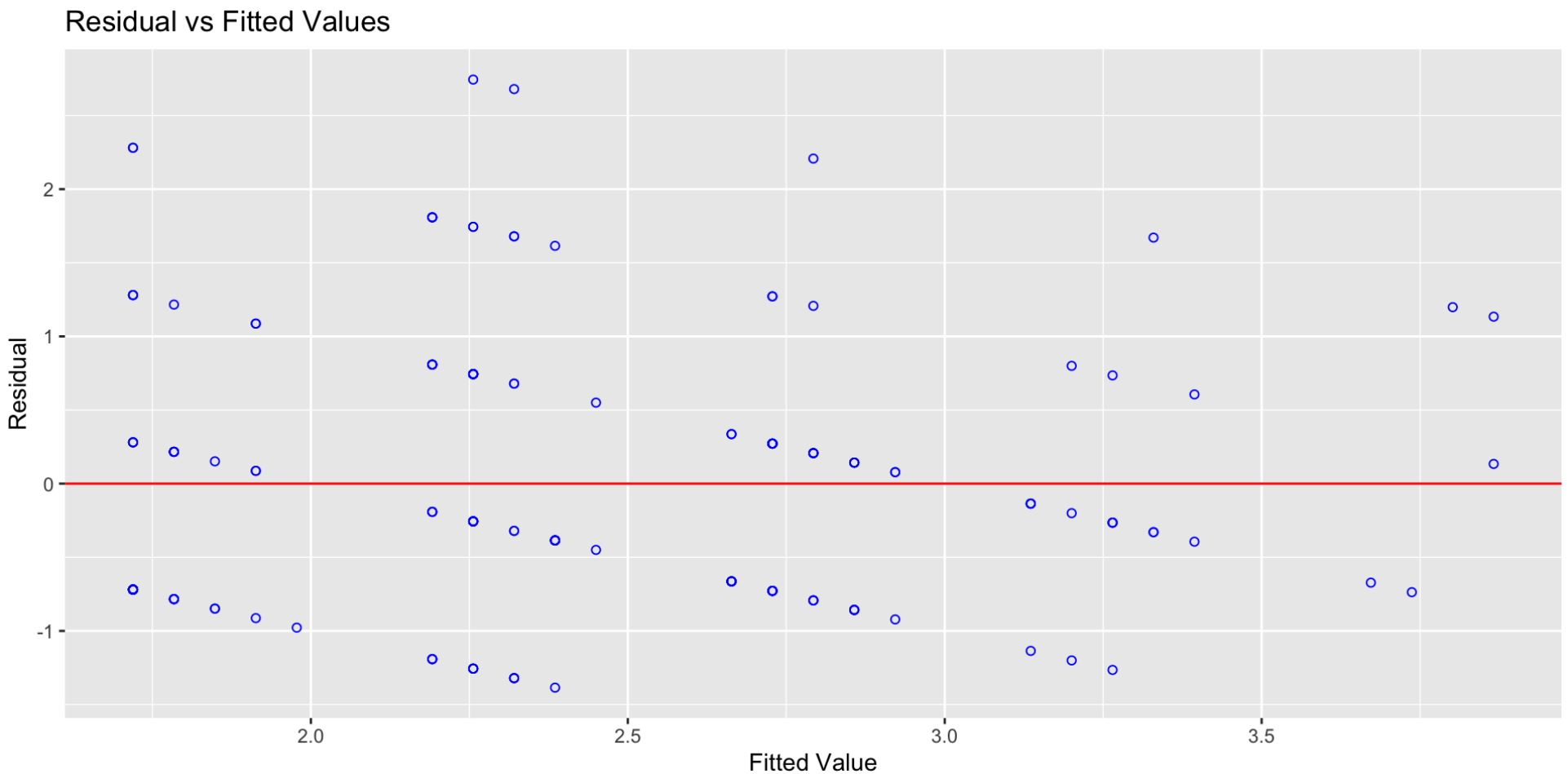
Ü1.8 Geben Sie mit folgendem Befehl die Tolerance und VIF-Werte raus

```
olsrr::ols_vif_tol(Modell1)
```

```
  Variables Tolerance      VIF
1   E102_02 0.8987008 1.112717
2   E102_04 0.8987008 1.112717
```

Ü1.9 Schauen Sie sich die Residualplotts an

- R-Code anzeigen



Plot der Residuen

Ü1.10 Testen Sie auf Homoskedastizität

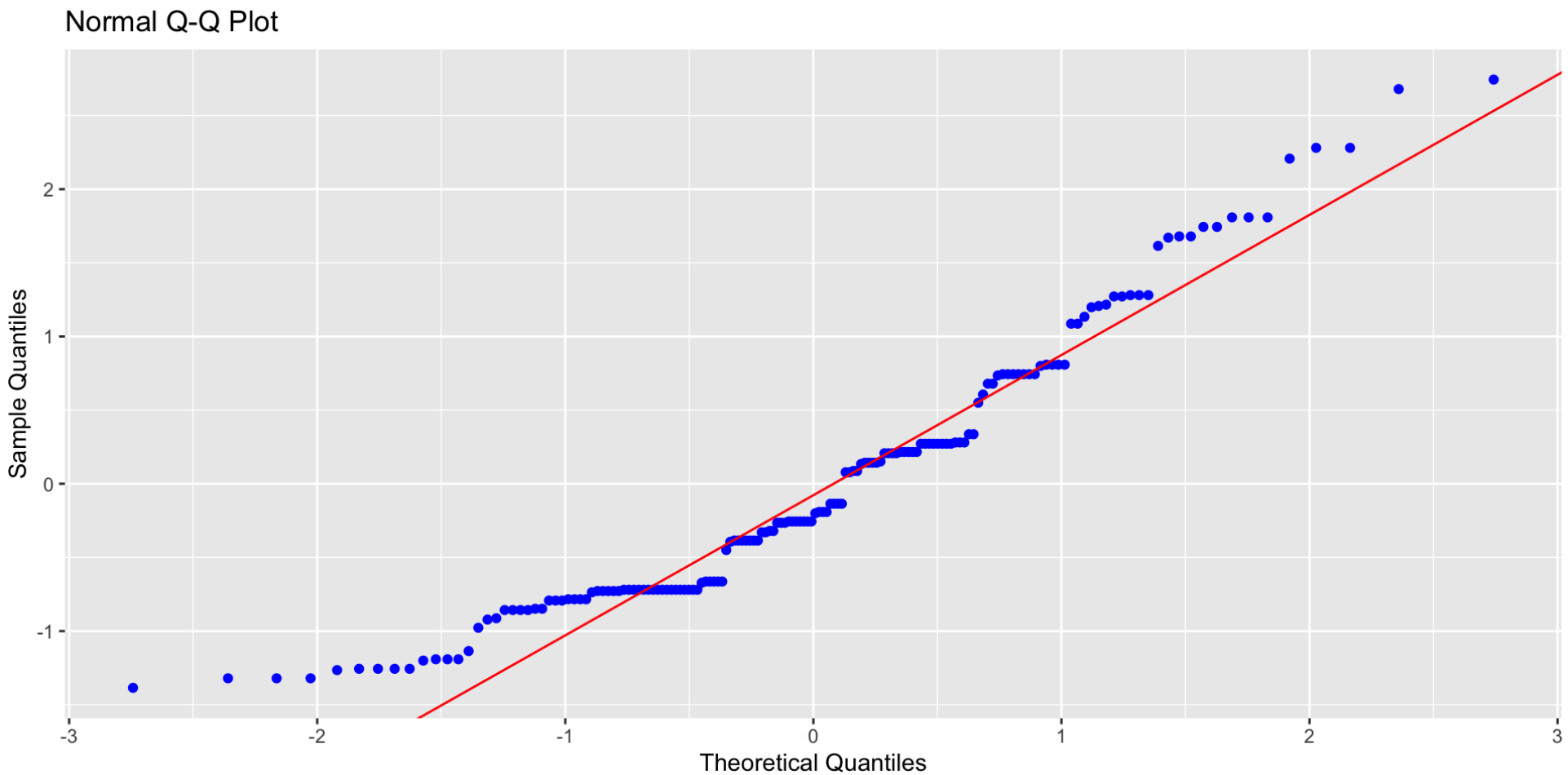
- R-Code anzeigen

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

               Data
-----------------------------------
Response : E201_10
Variables: fitted values of E201_10

       Test Summary
---------------------------
DF            =    1
Chi2          =    2.305095
Prob > Chi2   =    0.1289504
```

Was sagt Ihnen das?

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

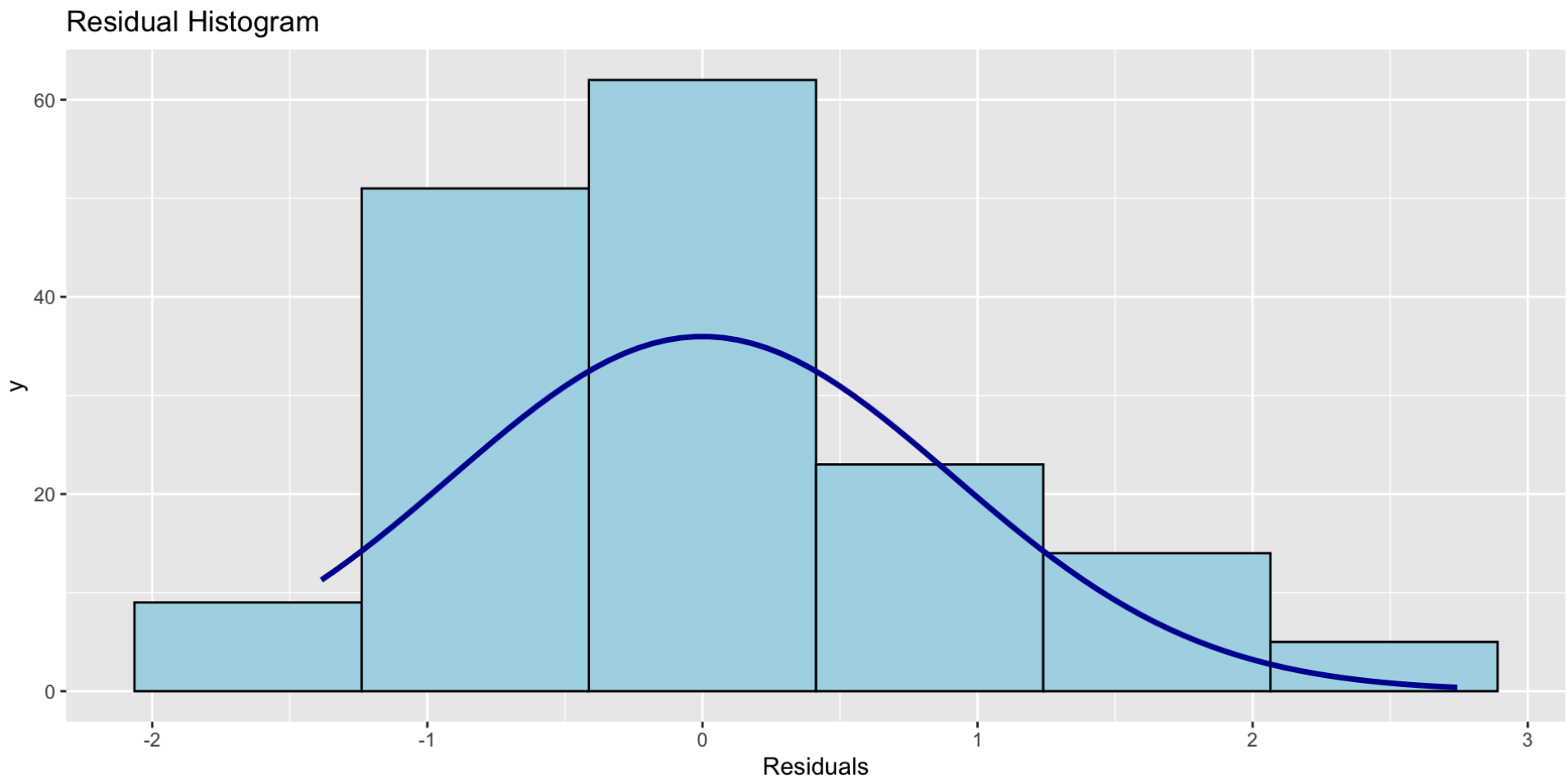
- R-Code anzeigen



Normal-Q-Q-Plot

Ü1.12 Und das Histogramm

- R-Code anzeigen



Histogramm der Residuen

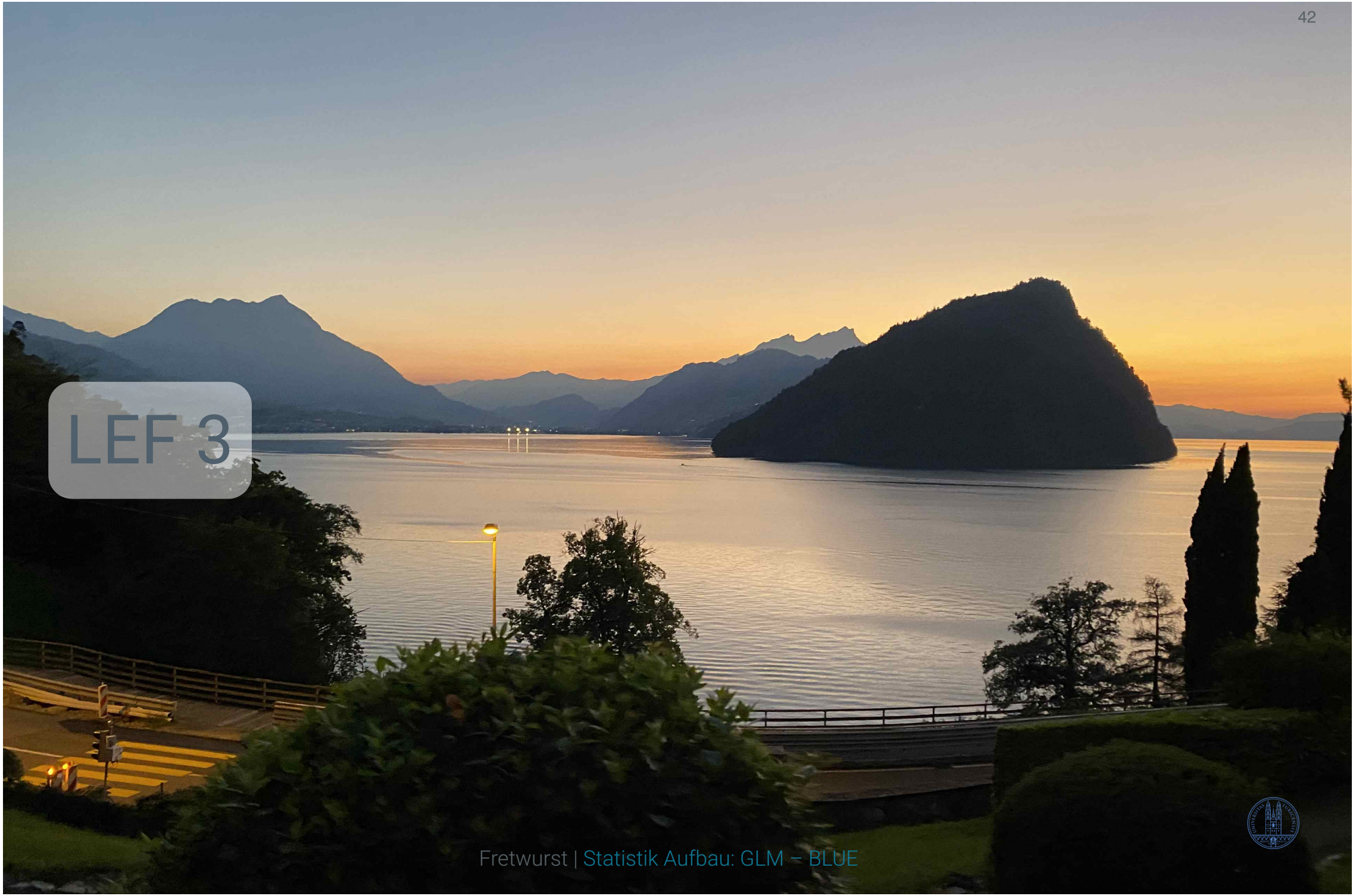
Ü1.13 Jetzt auf Normalverteilung testen

```
1 # Führe Tests auf signifikante Verletzungen
2 # der Normalverteilungsannahme aus.
3
4 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
-----------------------------------------------
```

3.5.2 Ü1.14 Fazit

Was ist Ihr Fazit aus der Regressionsrechnung?


LEF 3

Essayfragen 3

E3.1 Welches sind die Voraussetzungen für die Schätzung von Regressionen?

E3.2 Was bedeutet «Bias»?

E3.3 Was sagt Ihnen der Toleranzwert TOL?

E3.4 Was bedeutet Multikollinearität?

E3.5 Welche Kennwerte kennen Sie, mit denen Sie Multikollinearität abschätzen können?

E3.6 Wie reagieren a) p-Werte und b) Konfidenzintervalle auf Multikolliniearität?

E3.7 Warum kann man die volle Modellspezifikation nicht überprüfen?

E3.8 Was haben Theoriearbeit und Modellspezifikation miteinander zu tun?

E3.9 Was bedeutet es, dass die Variablen fix sein sollen?

MC-Fragen 3

MC 3.1.

MC 3.1: Sind folgende Aussagen richtig oder falsch?

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

Punkte:

OJS Runtime Error

invalid module

MC 3.2.

MC 3.2: Sind folgende Aussagen richtig oder falsch?

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

Punkte:

OJS Runtime Error

invalid module

MC 3.3.

MC 3.3: Sind folgende Aussagen richtig oder falsch?

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

Punkte:

OJS Runtime Error

invalid module

MC 3.4.

MC 3.4: Sind folgende Aussagen richtig oder falsch?

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

Punkte:

OJS Runtime Error

invalid module

MC 3.5.

MC 3.5: Sind folgende Aussagen richtig oder falsch?

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

Punkte:

OJS Runtime Error

invalid module

MC 3.6.

MC 3.6: Sind folgende Aussagen richtig oder falsch?

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

Punkte:

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

Insgesamt

OJS Runtime Error

invalid module

von 12 Punkten, was

OJS Runtime Error

invalid module

% und etwa einer

OJS Runtime Error

invalid module

entspricht.

OJS Runtime Error

invalid module

OJS Runtime Error

invalid module

OJS Error

RequireError: invalid module

Take Home – Ausblick – Vokabeln

Take Home

Note

- Sie kennen die Voraussetzungen für BLUE
- Schätzer sind unverzerrt, wenn die Modelle voll spezifiziert sind
- Schätzt man nicht lineare Zusammenhänge linear, macht man falsche Schlüsse
- Hängt die Streuung der Fehler mit den UVs zusammen, schätzt man die Standardfehler falsch (damit t-Wert, p-Wert, KI)
- Bei perfekter Multikollinearität können exklusive Effekte nicht geschätzt werden
- Sind die Fehler nicht unabhängig, verschätzt man sich in den Standardfehlern
- Etwas Multikollinearität ist der Grund für multivariate Analysen

Ausblick

Übung 1 (a und b)

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