Statistik II - Sitzung 8

Lena Masch

Institut für Politikwissenschaft

28. November 2024

Statistik II - Sitzung 8

1. kurze Verortung
2. Die multivariate Regression - ein neues Beispiel
3. Die Logik der Drittvariablen
4. Die Voraussetzungen der multivariaten Regression
5. Ausblick

Verortung

- Wiederholung der Grundlagen der multivariaten Regression
- Regressionen und Interaktionen konzeptionell verstehen (Wann und wieso wird es genutzt?)
- Einführung in die Annahmen der Regression
- Ziel: Annahmen (verstehen), kennen und ggf. in eigener Forschung prüfen

Beispiel: populistische Einstellungen

- Fragestellung(en)
  - Wie lassen sich populistische Einstellungen erklären?
  - Welche Faktoren beeinflussen populistische Einstellungen?
  - Inwiefern beeinflusst Bildung populistische Einstellungen?
  - Hat das Geschlecht einen Einfluss auf populistische Einstellungen?
- Analyse anhand von Sekundärdaten (ALLBUS 2018)

Operationalisierung der Variablen

- Die abhängige Variable (DV): **Populistische Einstellungen**
  - Populistische Einstellungen werden als Mittelwert eines Index aus mehreren Items gemessen.
  - Der Index fasst Aussagen zur Unterstützung populistischer Ideologien zusammen (z. B. Anti-Establishment, einfacher Bürger*innen besser geeeignet).
- Die unabhängigen Variablen (UV):
  - **Geschlecht**: Binäre Variable (männlich = 0, weiblich = 1).
  - **Alter**: Metrische Variable, gemessen in Jahren.
  - **Bildungsniveau**: Kategorische Variable, die den höchsten Abschluss angibt in drei Kategorien (niedrig, mittel, hoch.
  - **Gesundheitsstatus**: Subjektive Einschätzung der Gesundheit (5-Punkt Skala von "schlecht" bis "gut").
  - **Einkommen**: Einkommen als metrische Variable (monatliches Nettoeinkommen in €).

Populistische Einstellungen

```
Call:
lm(formula = pop_index ~ gender + age + education + health +
    income, data = za)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05755 -0.47432 -0.00675  0.45873  2.20012

Coefficients:
                    Estimate  Std. Error t value             Pr(>|t|)
(Intercept)       3.52555105  0.07903522  44.607 < 0.0000000000000002 ***
genderweiblich   -0.02155719  0.02815767  -0.766                0.444
age              -0.00060240  0.00085327  -0.706                0.480
educationmittel   0.44279080  0.03189722  13.882 < 0.0000000000000002 ***
educationniedrig  0.56292502  0.03791083  14.849 < 0.0000000000000002 ***
health           -0.08043002  0.01399818  -5.746           0.0000000102 ***
income           -0.00010200  0.00001169  -8.728 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6756 on 2608 degrees of freedom
  (862 observations deleted due to missingness)
Multiple R-squared:  0.2004,    Adjusted R-squared:  0.1986
F-statistic:   109 on 6 and 2608 DF,  p-value: < 0.0000000000000022
```

Populistische Einstellungen: Erklärung

Regressionsgleichung mit den Namen der Variablen und dem Datensatz "za" : $\hat{y} \sim x1 + x2 + x3 + x4 + x5$

```
Call:
lm(formula = pop_index ~ gender + age + education + health +
    income, data = za)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05755 -0.47432 -0.00675  0.45873  2.20012

Coefficients:
                   Estimate  Std. Error t value             Pr(>|t|)
(Intercept)      3.52555105  0.07903522  44.607 < 0.0000000000000002 ***
genderweiblich  -0.02155719  0.02815767  -0.766                0.444
age             -0.00060240  0.00085327  -0.706                0.480
educationmittel  0.44279080  0.03189722  13.882 < 0.0000000000000002 ***
educationniedrig 0.56292502  0.03791083  14.849 < 0.0000000000000002 ***
health          -0.08043002  0.01399818  -5.746          0.0000000102 ***
income          -0.00010200  0.00001169  -8.728 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6756 on 2608 degrees of freedom
  (862 observations deleted due to missingness)
Multiple R-squared:  0.2004,    Adjusted R-squared:  0.1986
F-statistic:   109 on 6 and 2608 DF,  p-value: < 0.00000000000000022
```

Vorhergesagter Populismus Score für Person mit gender, age, educ(ation), health, income = 0

Oder: Mittelwert, wenn Mittelwertzentrierung der Variablen vorgenommen wurde

Unstandardisierte Koeffizienten

P-Werte der Koeffizienten

nicht signifikant

nicht signifikant

*** Signifikant mit 0,1% Irrtumswahrscheinlichkeit; (ebenso 1% und 5%)

Gängige Signifikanzniveaus werden ausgewiesen

19,9% der Varianz des Populismus-Indexes wird durch das Modell erklärt

Wenn die eingeschätzte Gesundheit um eine Einheit steigt, sinkt der vorhergesagte Populismus Score um 0.08

Für eine Person mit niedriger Bildung steigt der Populismus-Score um 0.563 im Vergleich zu Personen mit hoher Bildung (hier die Referenzkategorie)

F-Test und dazugehöriger p-Wert: Das Gesamtmodell ist signifikant und kann interpretiert werden.

Mittelwertzentrierung des Alters

- Das Alter wurde **mittelwertzentriert**, um die Interpretation der Regressionskoeffizienten zu erleichtern.
- Die Zentrierung erfolgt durch Subtraktion des Mittelwerts (**51,7 Jahre**) von jedem beobachteten Wert (für jede befragte Person im Datensatz).

Originalwert (Jahre)	**Mittelwert**	**Zentrierter Wert**
46	51,7	-5,7
48	51,7	-3,7
50	51,7	-1,7
51,7	51,7	0
53	51,7	1,3
55	51,7	3,3
58	51,7	6,3

Tabelle: Beispiel für Mittelwertzentrierung des Alters

Populistische Einstellungen: Erklärung

- Die metrischen unabhängigen Variablen wurden mittelwertzentriert

Mittelwertzentrierung
Regressionsgleichung mit den Namen der Variablen (zentriert um den Mittelwert/ Mittelwert = 0, ".m") und dem Datensatz "za" : y ~ x1 + x2 + x3 + x4 + x5
Hinweis: ".m" zeigt den Unterschied zur vorherigen Regression, die Koeffizienten bleiben unverändert, lediglich der Interzept ändert sich.

Vorhergesagter Populismus Score für Person mit gender, education = 0, d.h. männlich und hohe Bildung ; mittleren Alters, mittlerer Gesundheit und mittlerem Einkommen (jeweils mittelwertzentriert)

```
Call:
lm(formula = pop_index ~ gender + age.m + health.m + education +
    income.m, data = za)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05755 -0.47432 -0.00675  0.45873  2.20012

Coefficients:
                   Estimate  Std. Error t value             Pr(>|t|)
(Intercept)      3.01788380  0.02656509 113.603 < 0.0000000000000002 ***
genderweiblich  -0.02155719  0.02815767  -0.766                0.444
age.m           -0.00060240  0.00085327  -0.706                0.480
health.m        -0.08043002  0.01399818  -5.746           0.0000000102 ***
educationmittel  0.44279080  0.03189722  13.882 < 0.0000000000000002 ***
educationniedrig 0.56292502  0.03791083  14.849 < 0.0000000000000002 ***
income.m        -0.00010200  0.00001169  -8.728 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6756 on 2608 degrees of freedom
  (862 observations deleted due to missingness)
Multiple R-squared:  0.2004,    Adjusted R-squared:  0.1986
F-statistic: 109 on 6 and 2608 DF,  p-value: < 0.00000000000000022
```

Populistische Einstellungen: Erklärung

- Unstandardisierte Koeffizienten und Standardisierte Koeffizienten

Comparison of Unstandardized and Standardized Coefficients		
Predictor	**Unstandardized Coeff.**	**Standardized Coeff.**
(Intercept)	3.526	NA
genderweiblich	−0.022	−0.014
age	−0.001	−0.014
educationmittel	0.443	0.279
educationniedrig	0.563	0.323
health	−0.080	−0.109
income	0.000	−0.173

Standardisierte Koeffizienten zum Vergleich der UV bzgl. Effektstärke:
Bildung und genauer die Dummy-Variable niedrige Bildung (vs. hohe Bildung) hat den stärkten Einfluss auf den Populismus Score

Hinweis: Effektstärke als Betrag lesen

Wiederholung: Multivariate Regression

- Die multivariate Regression enthält im Gegensatz zur bivariaten Regression **mehr** als eine unabhängige Variable.

$$y = \hat{y} + e = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ... + e$$

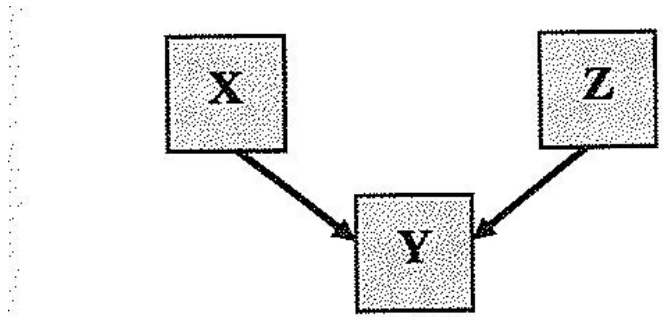
- Dadurch soll für die Möglichkeit der gleichzeitigen Effekte zweier unabhängiger Variablen ODER für die Effekte durch dritte (Kontroll-)Variablen auf einen bivariaten Zusammenhang kontrolliert werden.
- Die multivariate Regression ist also EINE Möglichkeit der **Drittvariablenkontrolle**

Die Drittvariablenkontrolle

- **Drittvariablenkontrolle** = Überprüfung des Einflusses einer Variablen auf einen bivariaten Zusammenhang
- Generell drei Modelle der und Begründungen für die Drittvariablenkontrolle
  - Grundlegend: Vorstellung von **Multikausalität** (Model 1)
  - Kontrolle: Vermeidung von **Scheinkausalität** (Model 2)
  - Einflussmediation: Kontrolle von **Interaktionseffekten** (Model 3)

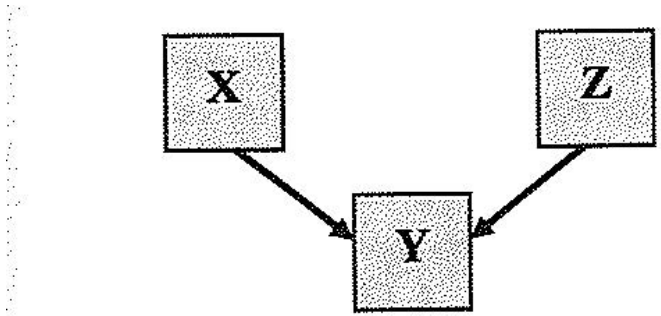
Die Drittvariablenkontrolle

- Drittvariablenkontrolle (Modell 1)
  - Multikausalität, d.h. die Erklärung einer abhängigen Variablen durch *mehrere* unabhängige Variablen, ist die Regel in den Sozialwissenschaften
  - Drittvariablenkontrolle = grundsätzliche Kontrolle, ob eine monokausale Beziehung zwischen unabhängiger und abhängiger Variable existiert



Die Drittvariablenkontrolle

- Drittvariablenkontrolle (Modell 1)
  - Durch Test für dritte Variable wird Effekt der unabhängigen auf die abhängige Variable oft kleiner => Messung eines bivariaten Zusammenhangs überschätzt die Wirkung einer einzelnen unabhängigen Variable
  - Korrektur der Überschätzung durch das Hinzufügen weiterer (theoretisch erklärungskräftiger Variablen) in das Modell!



Die Drittvariablenkontrolle

- Drittvariablenkontrolle (Modell 2)
  - Drittvariable kann sowohl die unabhängige als auch die abhängige Variable beeinflussen=> **Scheinkausalität**

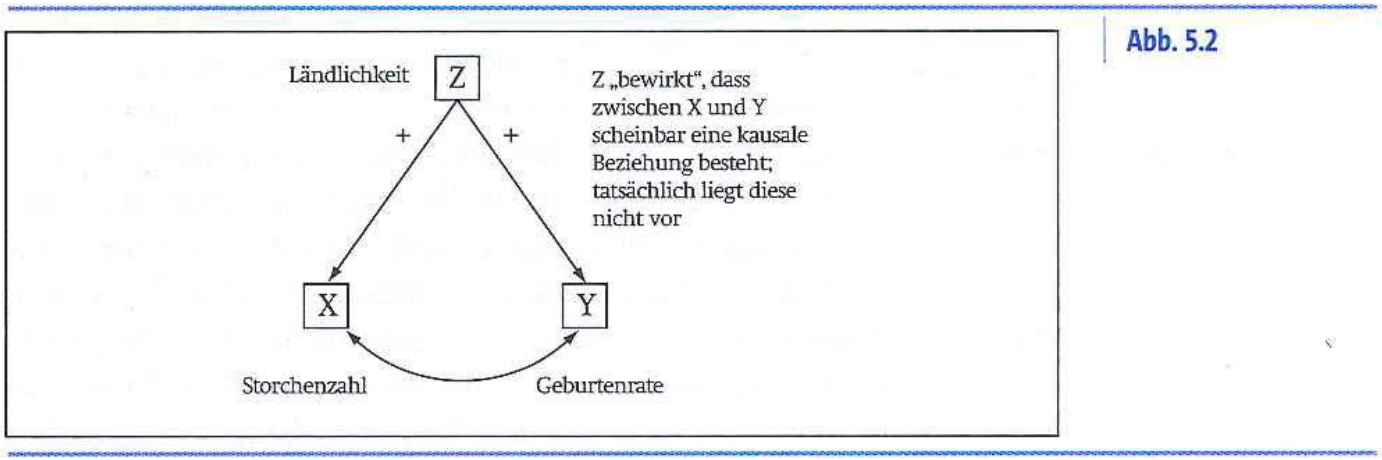


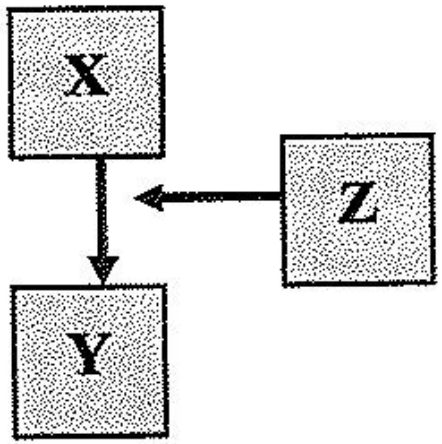
Abbildung: aus Diaz-Bone(2013): 115

Die Drittvariablenkontrolle

- Drittvariablenkontrolle (Modell 2)
  - ACHTUNG – selbst nach der Drittvariablenkontrolle kann ein statistischer Zusammenhang zwischen den kausal nicht verbundenen Variablen auftreten.
  - Dieser wird aber durch die Drittvariablenkontrolle deutlich geringer als vorher!

Die Drittvariablenkontrolle

- Drittvariablenkontrolle (Modell 3)
  - Die Drittvariable tritt als *interagierende* Variable auf



Die Drittvariablenkontrolle

- In diesem dritten Modell des Drittvariableneinflusses lassen sich zwei mögliche Formen unterscheiden
  - **Interaktion** = 'Eine Interaktion liegt vor, wenn je nach Ausprägung der Drittvariablen Z der statistische Zusammenhang [zwischen der abhängigen und der unabhängigen Variablen] verschieden ausfällt.' (Diaz-Bone 2006: 114)
  - **Suppression** = '[Die Suppression] besteht darin, dass eine Drittvariable Z den vorliegenden kausalen Zusammenhang zwischen zwei Variablen durch ihren Einfluss verdeckt. Erst nach Kontrolle des Drittvariableneinflusses von Z tritt der kausale Zusammenhang zwischen X und Y hervor.' (Diaz-Bone 2006: 117)

Voraussetzungen der multivariaten Regression

- Was könnten theoretische und methodische Probleme der Berechnung von multivariaten Regressionen sein?
  - Haben wir eine wichtige Erklärungsvariable übersehen?
  - Hängen die Variablen untereinander zu stark zusammen?
  - Üben die unabhängigen Variablen immer einen linearen Einfluss auf die abhängige Variable aus?
- Diese Fragen sind Teil der Anwendungsvoraussetzungen einer linearen Regression

Anwendungsvoraussetzung I: Messung der Variablen

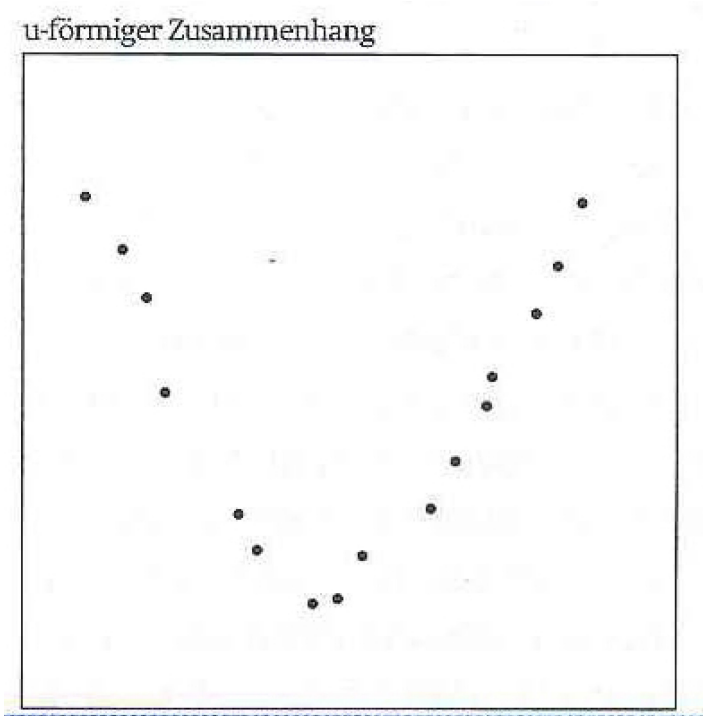
- Skalierung der Variablen
  - Die abhängige Variable muss metrisch skaliert sein!
  - Alle unabhängigen Variablen in einer OLS-Regression müssen entweder metrisch skaliert oder binär codiert (Dummy-Variablen) sein!
  - In Praxis werden ordinal skalierte Variablen mit $>= 5$ Ausprägungen oft als metrisch skaliert interpretiert $=>$ nicht ganz unproblematisch!
- Test der Anwendungsvoraussetzung I: Deskriptive Statistiken für alle Variablen

Voraussetzung II: Linearitätsannahme

- 'Der Einfluss der einzelnen unabhängigen Variablen auf die abhängige Variable soll jeweils linear sein.' (Diaz-Bone 2006: 198)
  - Problem - Ist der Einfluss einer unabhängigen Variablen nicht linear, so können die Regressionskoeffizienten in einer linearen Regressionsgleichung keine sinnvollen Ergebnisse liefern

Voraussetzung II: Linearitätsannahme

- 'Der Einfluss der einzelnen unabhängigen Variablen auf die abhängige Variable soll jeweils linear sein.' (Diaz-Bone 2006: 198)



Voraussetzung II: Linearitätsannahme

- 'Der Einfluss der einzelnen unabhängigen Variablen auf die abhängige Variable soll jeweils linear sein.' (Diaz-Bone 2006: 198)
  - Lösung - Recodierung der entsprechenden Variablen (wenn kurvilinearer Effekt einer Variablen) oder Nutzung anderer Regressionsmodelle
  - Beispiel kurvilinearer Effekt - EU-Skeptizismus unter radikalen Linken und Rechten. Effekt der Links-Rechts-Einstellung (von links nach rechts) nicht aussagekräftig, da nicht linear.
  - Stattdessen: Recodierung der Links-Rechts-Einstellung in zwei Dummy-Variablen: Links-radikal vs. alle anderen, Rechts-radikal vs. alle anderen. Beide neuen Variablen üben linearen Effekt auf Euroskeptizismus aus.

Voraussetzung II: Linearitätsannahme

- 'Der Einfluss der einzelnen unabhängigen Variablen auf die abhängige Variable soll jeweils linear sein.' (Diaz-Bone 2006: 198)
- 'Der Einfluss der einzelnen unabhängigen Variablen auf die abhängige Variable soll jeweils linear sein.' (Diaz-Bone 2006: 198)
  - Test - Diagnostik über die Verteilung der Residuen
  - Streudiagramm: Y-Residuen auf **Werte** der unabhängigen Variablen $=>$ Linearitätsannahme erfüllt, wenn Fälle gleich um Mittelwert von Y streuen

Voraussetzung III: Kein Omitted Variable Bias

- Haben wir eine wichtige Variable in unserem Erklärungsmodell übersehen?
  - Eigentlich theoretisches Problem, das aber statistische Effekte zeigt
  - Auslassung einer wesentlichen Variable könnte Koefffizienten verzerren
  - Messbar ist dieser bias darüber, dass die Residuen nicht normalverteilt sind.
- ABER: Außer in einem an Perfektion grenzenden Modell liegt dieser bias meistens vor (und die entsprechenden Tests sind signifikant). Enscheidung darüber, ob ein wirkliches Problem vorliegt, müssen Forschende aufgrund theoretischer und empirischer Überlegungen treffen!

Voraussetzung III: Kein Omitted Variable Bias

- Haben wir eine wichtige Variable in unserem Erklärungsmodell übersehen?
  - Test über Streudiagramm (Histogramm /Dichtefunktion) oder Q-Q-Plot der Residuen, um (Nicht-)Normalverteilung festzustellen
  - Zusätzliche Überprüfung durch sogenannte Tests der Normalverteilung, z.B. Shapiro-Wilk-Test

Voraussetzung IV: Keine Multikollinearität

- Hängen die Variablen untereinander zu stark zusammen?
- Problem der Multikollinearität
  - Korrelieren unabhängige Variablen untereinander zu stark (= **Multikollinearität**), werden die Standardfehler in der Berechnung der Regressionskoeffizienten als Schätzer ungenau (zu groß!)
  - Ist Multikollinearität zu groß, so wird durch das Modell auch der Einfluss der einzelnen Variablen verzerrt dargestellt => Interpretation der Ergebnisse entspricht nicht Realität

Voraussetzung IV: Keine Multikollinearität

- Hängen die Variablen untereinander zu stark zusammen?
- Überprüfung der Multikollinearität
  - Test über den sogenannten **Varianz-Inflations-Faktor (VIF)** oder dessen Kehrwert **Toleranzwert T**
  - Wenn VIF > 10 (oder T < 0.1) höchst problematische Multikollinearität
  - Wenn VIF > 4 (oder T < 0.25) problematische Multikollinearität
- Korrelationsmatrix überprüfen (v.a. auf r > 0.8)
- theoretische Zusammenhänge reflektieren und i.d.R. eine Variable auswählen oder einen Index aus mehreren Indikatoren bilden

Voraussetzung V: Homoskedastizität

- Annahme - Varianzhomogenität (oder: **Homoskedastizität von Residuen**)
  - Die Erklärungsleistung des Regressionsmodells hinsichtlich der abhängigen Variable muss einheitlich sein
  - D.h., die Residuen sollen für alle Werte der geschätzten abhängigen Variable in gleichem Umfang variieren => auch: Gleichverteilung des Fehlers
- Problem: Ist Annahme nicht erfüllt (= **Heteroskedastizität**), werden Standardfehler und damit die Berechnung der Signifikanztests verzerrt!

Voraussetzung V: Homoskedastizität

- Test der Homoskedastizitäts-Annahme - Streudiagramm und weitere Tests
  - Streudiagramm, das Variation der Residuen über unterschiedliche Werte der abhängigen Variablen abbildet
  - Weitere Testverfahren, die Homoskedastizität testen (z.B. Cook-Weisberg-Test)

Voraussetzung VI: Normalverteilung der Residuen

- die Residuen des Modells sind normalverteilt
- Problem: Ist Annahme nicht erfüllt, können Standardfehler und damit die Berechnung der Signifikanztests verzerrt sein
- v.a. für kleinere Stichproben bei gravierenden Abweichungen relevant, bei größeren Stichproben weniger problematisch (vgl. Diaz-Bone 2023: 231)

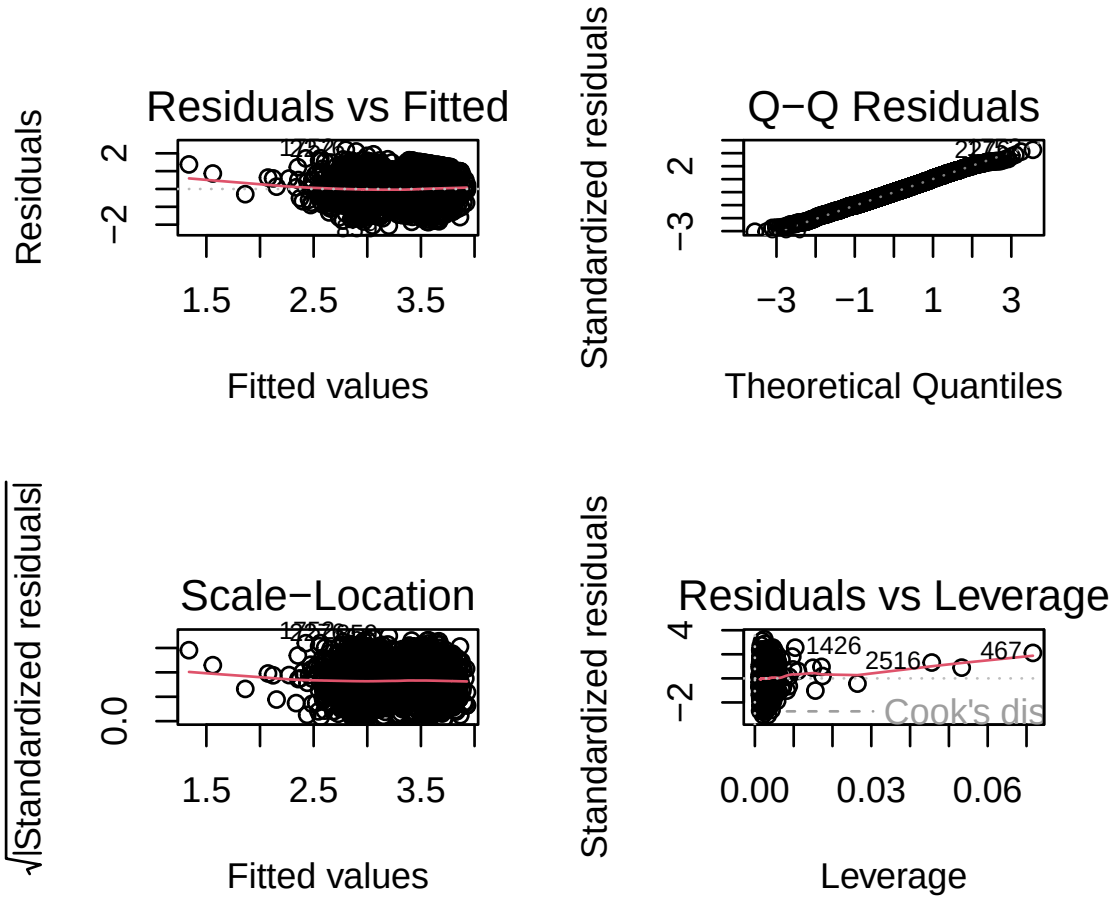
Voraussetzung VII: Abwesenheit einflussreicher Fälle

- Voraussetzung VII: Abwesenheit einflussreicher Fälle
  - Einflussreiche Fälle (bestimmte Ausreißer und Extremwerte) verzerren das gesamte Modell durch einen überproportional starken Einfluss auf die Regressionsgerade
  - Problemdiagnostik und -lösung möglich, z.B. Cook's Distances

Voraussetzung VIII: Statistische Unabhängigkeit der Residuen

- ausführliche Behandlung für BA-Studium zu weit gehend
  - Die Residuen dürfen untereinander nicht zusammenhängen
  - Problem bei wiederholter Messung der gleichen Fälle über die Zeit hinweg (Panel-Daten, Zeitreihen) oder bei Klumpenstichproben
  - Lösung: Berechnung spezifischer Regressionsmodelle (Zeitreihen-, Panel-, Mehrebenen-Analysen)
  - Gegenstand vertiefender Seminare und Kurse (Modellierung und Datenerhebung reflektieren)

Beispiel einer Residuen-Analyse



Take-Home-Messages

- Voraussetzungen für lineare Regression müssen erfüllt sein, damit sinnvolle Interpretation der Ergebnisse möglich
- Überprüfung der Voraussetzungen z.T. erst nach Durchführung der Regression über anschließende Berechnung der Residuen-Streudiagramme und der weiteren Tests möglich => *Post-Estimation Tests*
- Nur wenn Anwendungsvoraussetzung nicht gegeben, muss nach Lösung gesucht werden
  - Ob Anwendungsvoraussetzung problematisch verletzt ist, hängt von Interpretation der ForscherIn ab!
  - Achtung - auch verbesserte Modelle müssen nochmals auf Anwendungsvoraussetzungen überprüft werden => sonst möglicherweise Verschlimmbesserung!

Ausblick

- Annahmen der Regression (Regressionsdiagnostiken und Residualanalysen)
  - Selbststudium ratsam: Lektüre der Grundlagenliteratur, z.B. Diaz-Bone Kapitel 8
- die logistische Regression
  - Gedankenexperiment: Versuchen Sie die Wahl einer Partei (AfD) zu erklären. Welche Variablen würden Sie berücksichtigen?