Statistik II - Sitzung 5

Lena Masch

Institut für Politikwissenschaft

Sitzung 6

Statistik II - Sitzung 6

1. Die multivariate Regression
   - Drittvariablenkontrolle: Multivariate Regression
   - Kurze Vertiefung: F-Teststatistik

Die multivariate Regression

- Forschungsfrage und Beispiel der heutigen Sitzung: Welche Faktoren beeinflussen die Zustimmmung zu Verschwörungstheorien?
- multikausal (mehr als ein Faktor)
- Fokus: auf eine zentrale UV möglich (unter Berücksichtigung sogenannter Kontrollvariablen)
- hier zunächst auf die Bildung

Wiederholung: ANOVA

- Hängt die Zustimmung zu Verschwörungstheorien von der Bildung ab?
- der Gesamtmittelwert liegt 2,46 (Skala von 1-7)
- die Mittelwerte für niedrige (2,82), mittlere (2,67) und hohe Bildung (2,16)
- Eine Varianzanalyse (ANOVA) kann mehrere Gruppenmittelwerte vergleichen

```
               Df Sum Sq Mean Sq F value             Pr(>F)
education       2    235  117.42   56.41 <0.0000000000000002 ***
Residuals    2853   5939    2.08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
27 observations deleted due to missingness
>
```

- Woher wissen wir, welche Gruppen sich unterscheiden? Was können wir machen?

Wiederholung: Bivariate Regresssion

- Wir könnten eine ANOVA und Posthoc Tests durchführen ODER eine Regression
- die Abbildung zeigt eine bivariate lineare Regression (AV = Verschwörungsglaube , UV = Bildung)

```
Call:
lm(formula = conspiracy ~ education, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8224 -1.1631 -0.4891  0.8369  4.8369

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)
(Intercept)      2.82242    0.06152  45.877 <0.0000000000000002 ***
educationmedium -0.14879    0.07720  -1.927               0.054 .
educationhigh   -0.65934    0.07299  -9.033 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.443 on 2853 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.03804,   Adjusted R-squared:  0.03736
F-statistic: 56.41 on 2 and 2853 DF,  p-value: < 0.00000000000000022
```

Wiederholung: Bivariate Regression

- die Abbildung zeigt eine Tabelle der bivariaten lineare Regression (AV = Verschwörungsglaube , UV = Bildung)
- üblich in Publikationen

	Dependent variable:
	consp_
educationmedium	-0.149
	(0.077)
educationhigh	-0.659***
	(0.073)
Constant	2.822***
	(0.062)
Observations	2,856
R^2	0.038
Adjusted R^2	0.037
Residual Std. Error	1.443 (df = 2853)
F Statistic	56.407*** (df = 2; 2853)
Note:	*p<0.05; **p<0.01; ***p<0.001

erste Erkenntnis(se)

- Voraussetzung der Skalenniveaus einer linearen Regression (OLS)
  - AV: metrisch (min. intervallskaliert)
  - UV: metrisch oder kategoriell
  - kategorielle Variablen müssen dummy-kodiert sein (k-1)
- Zusammenhang: F-Test (ANOVA) und Regression
- F-Test bestimmt die Güte des Modells
- nur signifikante Modelle (F-Test) werden interpretiert
- ein sig. F-Test bedeutet, dass min. eine UV einen signifkanten Einfluss hat

Die multivariate Regression

- Eine multivariate Regression kontrolliert, ob unser Effekt zwischen zwei Variablen tatsächlich weiter auftritt, wenn wir andere Variablen mit in die Regressionsgleichung (d.h. in unser theoretisches Modell) aufnehmen

Die multivariate Regression

- Die daraus entstehende multivariate Gleichung lautet dann:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * x_4 + e$$

- Die generelle multivariate Gleichung ist

$y = b_0 + b_1 * x_1 + b_k * x_k + e$ mit $k =$ Anzahl der Variablen

Die multivariate Regression

- Forschungsfrage und Beispiel der heutigen Sitzung: Welche Faktoren beeinflussen die Zustimmmung zu Verschwörungstheorien?
- multikausal (mehr als ein Faktor)
- Fokus: auf eine zentrale UV möglich (unter Berücksichtigung sogenannter Kontrollvariablen)
- für Drittvariablen sollte kontrolliert werden
- Was meinen Sie, was spielt noch eine Rolle?

Die multivariate Regression

- Forschungsfrage und Beispiel der heutigen Sitzung: Welche Faktoren beeinflussen die Zustimmmung zu Verschwörungstheorien?
- Daten (GESIS Panel)
  - AV: consp Zustimmung zu drei Statements bzgl. COVID-19 Verschwörungstheorien (als Mittelwertindex)
  - UV: education /Bildung als Schulabschluss (niedrig, mittel, hoch)
  - UV: gender/ Gender (männlich, weiblich)
  - UV: trust/ Vertrauen in politische Institutionen (Index)
  - UV: income/ Einkommmen (monatliches Nettoeinkommen)
  - UV: lr/ Links-Rechts-Selbsteinstufung

Operationalisierungen I

- **Verschwörungstheorien**:
  - "Das Coronavirus ist eine biologische Waffe, die in geheimen staatlichen Laboren entwickelt wurde."
  - "Das Coronavirus wird genutzt, um die Bürgerrechte einzuschränken und eine anhaltende Überwachung der Bürger zu starten."
  - "Die Gefahr und die Verbreitung des Coronavirus werden absichtlich übertrieben."
  - Teilnehmer*innen wurden gebeten, die Wahrscheinlichkeit jeder Aussage auf einer Sieben-Punkte-Skala zu bewerten (1 = äußerst unwahrscheinlich; 7 = äußerst wahrscheinlich).
- **Politisches Vertrauen**:
  - Vertrauen in verschiedene politischer Institutionen: Justizsystem, Fernsehen, Zeitungen, Regierung, politische Parteien, Europäische Kommission
  - Vertrauen auf einer Skala von 1 (überhaupt nicht vertrauen) bis 7 (vollständig vertrauen).

Operationalisierungen II

- **Alter**:
  - Alter in Jahren
- **Geschlecht**:
  - Geschlecht (0 = männlich, 1 = weiblich)
- **Bildung**:
  - Bildungsstand im deutschen dreigliedrigen System:
    - ★ 1 = niedrig, bis 9 Jahre Schule oder weniger
    - ★ 2 = mittel, 10 Jahre Schule
    - ★ 3 = hoch, 12 oder 13 Jahre Schule mit Fachhochschul- oder Universitätszugangsberechtigung
- **Einkommen**:
  - Persönliches Einkommen, gemessen mit 15 Einkommenskategorien
- **Politische Selbstpositionierung**:
  - Links-Rechts-Selbsteinstufung auf einer 10-Punkte-Likert-Skala (1 = links, 10 = rechts)

Die multivariate Regression

- Die multivariate Regression im Output der Software

```
Call:
lm(formula = consp_     ~ age + gender + education + trust +
    income + lr, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5850 -0.7941 -0.1924  0.6405  6.4880

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)      7.076545   0.181287  39.035 < 0.0000000000000002 ***
age             -0.010695   0.001812  -5.902    0.000000004083759 ***
genderfemale    -0.074223   0.051701  -1.436              0.151231
educationmedium -0.150541   0.069274  -2.173              0.029866 *
educationhigh   -0.555662   0.070550  -7.876    0.000000000000005 ***
trust           -0.983767   0.030727 -32.016 < 0.0000000000000002 ***
income          -0.037005   0.007762  -4.767    0.000001974799125 ***
lr               0.044746   0.013206   3.388              0.000714 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.177 on 2485 degrees of freedom
  (390 observations deleted due to missingness)
Multiple R-squared:  0.3558,    Adjusted R-squared:  0.354
F-statistic: 196.1 on 7 and 2485 DF,  p-value: < 0.00000000000000022
```

Die multivariate Regression

- Die Koeffizienten in der Übersicht

Variable	coefficient	std.error	t-statistic	p.value
(Intercept)	7.077	0.181	39.035	< 0.001
age	-0.011	0.002	-5.902	< 0.001
genderfemale	-0.074	0.052	-1.436	0.151
educationmedium	-0.151	0.069	-2.173	0.03
educationhigh	-0.556	0.071	-7.876	< 0.001
trust	-0.984	0.031	-32.016	< 0.001
income	-0.037	0.008	-4.767	< 0.001
lr	0.045	0.013	3.388	< 0.001

Die multivariate Regression

- Die Konfidenzintervalls der Koeffizienten
- Vergleichen Sie, die zuvor berichtete Signfikanz (t-Werte) und die Konfidenzintervalle. Was fällt auf?

```
                          2.5 % 97.5 %
(Intercept)       7.077   6.721  7.432
age              -0.011  -0.014 -0.007
genderfemale     -0.074  -0.176  0.027
educationmedium  -0.151  -0.286 -0.015
educationhigh    -0.556  -0.694 -0.417
trust            -0.984  -1.044 -0.924
income           -0.037  -0.052 -0.022
lr                0.045   0.019  0.071
```

Die multivariate Regression

- Für die multivariate Regression gilt, dass die Herleitung und Interpretation des Determinationskoeffizienten R^2 und der Regressionskoeffizienten $b_1, b_2, b_3, ..., b_k$ gleich bleibt
- Zusätzlich gibt es aber
  - die standardisierten Regressionskoeffizienten (β oder Beta-Koeffizienten)
  - Probleme mit den Anwendungsvoraussetzungen der linearen Regression ($=>$ nächste Sitzung(en))

Standardisierte Regressionskoeffizienten

- In der multivariaten Regression unterscheidet man zwischen den unstandardisierten Regressionskoeffizienten $b_1, b_2, b_3, ..., b_k$ und den standardisierten Regressionskoeffizienten $\beta_1, \beta_2, \beta_3, ..., \beta_k$
- Die standardisierten Regressionskoeffizienten sind in ihrer Stärke untereinander vergleichbar
  - Über die Standardisierung wird die Einflussstärke der Koeffizienten auf den Mittelwert $= 0$ und eine Standardabweichung $= 1$ standardisiert.
  - Das bedeutet: Hat eine Variable X_1 einen höheren (positiven ODER negativen) Koeffizienten (β_1) als die Variable X_2, so übt Variable X_1 den stärkeren Einfluss auf die abhängige Variable (Y) aus ($X_1 > X_2$ weil $\beta_1 > \beta_2$)

Standardisierte Regressionskoeffizienten

- Die unstandardisierten Regressionskoeffizienten sind hingegen nicht direkt miteinander vergleichbar, drücken aber den Grad des individuellen Einflusses der Variable auf Y aus
  - Wenn b_1 für $X_1 = 0.5$, dann verändert sich Y für jede Einheit von X_1 um eine halbe (0.5) Einheit
  - Wenn b_1 für $X_1 = 0.3$, dann verändert sich Y für jede Einheit von X_1 um 0.3 Einheiten
  - Beispiel - Beide Variablen weisen $\%$ als Einheit auf: Wenn b_1 für $X_1 = 0.3$, dann verändert sich Y für jeden 1%-Anstieg von X_1 um 0.3%

Standardisierte Regressionskoeffizienten

- Geht es also in der Überprüfung eines theoretischen Arguments darum, die Stärke eines Einfluss einer bestimmten Variable zu überprüfen, so nutzt man den unstandardisierten Regressionskoeffizienten ($=>$ X-Zentrierung)
- Will man hingegen herausfinden, in welcher Rangfolge eine abhängige Variable Y durch viele verschiedene Variablen erklärt wird, so nutzt man die standardisierten Regressionskoeffizienten ($=>$ Y-Zentrierung)

Die multivariate Regression

- Die Darstellung der unstandardisierten (b) und standardisierten (ß) Koeffizienten

Variable	coefficient (b)	std_coefficient (ß)	std.error	t-statistic	p.value
(Intercept)	7.07654524	NA	0.181	39.035	< 0.001
age	-0.01069512	-0.101	0.002	-5.902	< 0.001
genderfemale	-0.07422340	-0.025	0.052	-1.436	0.151
educationmedium	-0.15054072	-0.048	0.069	-2.173	0.03
educationhigh	-0.55566213	-0.190	0.071	-7.876	< 0.001
trust	-0.98376691	-0.525	0.031	-32.016	< 0.001
income	-0.03700495	-0.088	0.008	-4.767	< 0.001
lr	0.04474619	0.055	0.013	3.388	< 0.001

Die multivariate Regression

- Die Darstellung einer multivariaten Regressionstabelle

	Dependent variable:
	consp_
age	-0.011***
	(0.002)
genderfemale	-0.074
	(0.052)
educationmedium	-0.151*
	(0.069)
educationhigh	-0.556***
	(0.071)
trust	-0.984***
	(0.031)
income	-0.037***
	(0.008)
lr	0.045***
	(0.013)
Constant	7.077***
	(0.181)
Observations	2,493
R^2	0.356
Adjusted R^2	0.354
Residual Std. Error	1.177 (df = 2485)
F Statistic	196.057*** (df = 7; 2485)
Note:	*p<0.05; **p<0.01; ***p<0.001

Ein Beispiel - Ausschnitt

Glaube an Verschwörungstheorien (consp)	b (unstandardisiert)	β (standardisiert)
Alter	-.011	-.101
Gender: weiblich	-.074	-.025
Bildung: mittel vs. niedrig	-.151	-0.048
Bildung: hoch vs. niedrig	-.556	-.071
Vertrauen pol. Institutionen	-.984	0.031
Einkommen	-.037	-.0.088
Links-Rechts-Selbsteinstufung	.045	0.055
Modellgüte $(adj.R^2)$	35.4%	
Quelle: GESIS Panel 2022. Eigene Berechnung		

Ein Beispiel - Interpretation

- Zur Interpretation ist es maßgeblich, die Codierungen (Operationalisierungen) der Variablen zu kennen
- Die Variable 'Alter' hat einen unstandardisierten Koeffizienten von -.011. Der Effekt dieser Variable auf die AV beträgt also -.011. Was bedeutet das?
  - Für jede Einheit, um die sich die X-Variable verändert, verändert sich der Y-Wert um -.010 Einheiten
  - Alter ist gemessen in Jahren, d.h. für jeden Anstieg von x um einen Punkt (ein Jahr) sinkt der Glaube an Verschwörungstheorien durchschnittlich um -.011.
  - als standardardisierter Koeffizient zeigt sich der Effekt in Standardabweichungen: Steigt das Alter um eine Standardabweichung (sd $= 13.94$), so beträgt der Effekt durchschnittlich -.101

Ein Beispiel - Interpretation

- Zur Interpretation ist es maßgeblich, die Codierungen (Operationalisierungen) der Variablen zu kennen(!)
- Die Variable "Bildung: hoch vs. niedrig" hat einen unstandardisierten Koeffizienten von -.556. Der Effekt dieser Variable auf die AV beträgt also -.556. Was bedeutet das?
  - Für jede Einheit, um die sich die X-Variable verändert, verändert sich der Y-Wert um -.556 Einheiten
  - kategorielle Variablen können nur mit Bezug auf die Referenzkategorie interpretiert werden
  - die Referenzkategorie ist die ausgelassene Kategorie (k-1), hier "Bildung: niedrig"
  - der Effekt einer kategoriellen Variable kann nur einmal auftreten (zutreffen/ nicht zutreffen), eine Interpretation der standardisierten Koeffizienten ist nicht sinnvoll
  - Durchschnittlich stimmen Personen mit einem hohen Bildungsabschluss im Vergleich zu Personen mit einem niedrigen Bildungsabschluss weniger stark Verschwörungstheorien zu und zwar um -.556 Punkte.

Vertiefung: F-Statistik in der Regression

Die F-Statistik wird verwendet, um zu testen, ob das gesamte Regressionsmodell signifikant ist. Sie wird wie folgt berechnet:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\text{SSR}}{k}}{\frac{\text{SSE}}{n-k-1}} \tag{1}$$

Erklärungen:

- **MSR (Mittlere Quadratsumme der Regression)**: Der durchschnittliche Anteil der durch das Modell erklärten Varianz
- **MSE (Mittlere Quadratsumme der Fehler(Residuen)**: Der durchschnittliche Anteil der Varianz innerhalb der Residuen (nicht erklärte Varianz)
- **SSR (Summe der Quadrate der Regression)**: Varianz, die durch das Modell erklärt wird.
- **SSE (Summe der Quadrate der Residuen)**: Nicht erklärte Varianz (Residuen).
- **k**: Anzahl der unabhängigen Variablen.
- **n**: Gesamtzahl der Beobachtungen.

Beispiel: Vertiefung F-Statistik in der Regression

Betrachten wir die ANOVA-Ausgabe aus dem R-Skript:

```
               Df Sum Sq Mean Sq F value               Pr(>F)
education       2    235  117.42   56.41 <0.0000000000000002 ***
Residuals    2853   5939    2.08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
27 observations deleted due to missingness
>
```

Ergebnisse:

- **Df (Freiheitsgrade)**: 2 und 2853
- **Sum Sq (Summe der Quadrate)**: **SSR** $= 235$, **SSE** $= 5939$
- **Mean Sq (Mittlere Quadratsumme)**: **MSR** $= 117.42$, **MSE** $= 2.08$
- **F-Wert** 56.41

Beziehung zwischen F-Statistik und R^2

Die F-Statistik lässt sich auch über R^2 ausdrücken. Die Formel lautet:

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} \tag{2}$$

Erläuterung mit dem Beispiel:

- R^2 (Multiple R-squared): 0.03804
- k: 2 (Anzahl der unabhängigen Variablen)
- n: 2856 (Gesamtzahl der Beobachtungen)

Vertiefung F-Statistik in der Regression

- Berechnung der F-Statistik aus R^2

$$F = \frac{\frac{0.03804}{2}}{\frac{1-0.03804}{2853}} = 56.41 \tag{3}$$

Zusammenfassung

- die multivariate lineare Regression kann kategorielle und metrische als UV Variablen aufnehmen
- die AV muss metrisch sein
- die F-Statistik zeigt, ob eine Variable einen signifikanten Einfluss auf die AV hat
- R^2 zeigt den Anteil erklärter Varianz und wird in multivariaten Analysen als adjustiertes (adj.) oder korrigiertes (korr.) R^2 berichtet, um für die Aufnahme weiterer Variablen zu korrigieren
- die Interpretation wird in den nächsten Wochen eingeübt
- Lernziel ist es, Output und Regressionstabellen lesen zu können

Ausblick

OLS Vertiefung zu Annahmen und Interaktionen