

ANOVA und Transformationen

Statistik II

Wiederholung
ANOVA
Gemischte Modelle
Transformationen
Zusammenfassung

Wiederholung
 Literatur
 Zusammenfassung
ANOVA
Gemischte Modelle
Transformationen
Zusammenfassung

Zum Nachlesen

- ▶ Agresti ch. 12 (**nur bis Seite 381**)
- ▶ Agresti ch. 13 (**nur bis Seite 428**)

Literatur für nächste Woche

- ▶ Berry (1993, S. 10-45, 51, 67-83)

Zusammenfassung

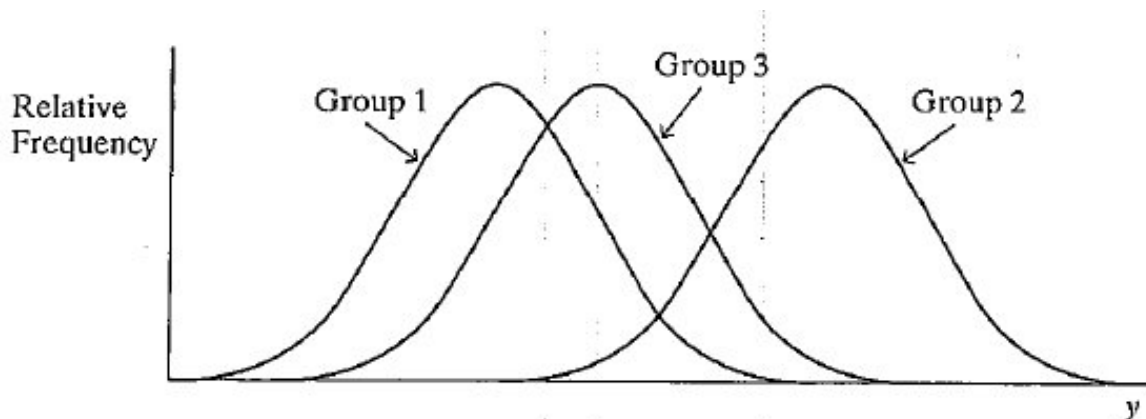
- ▶ Inferenz für das multivariate Modell: F- und t-Test
- ▶ Standardisierte Koeffizienten: oft keine Verbesserung
- ▶ Gewichtung: nicht unbedingt clever
- ▶ Kohortenanalyse: Extremfall von Kollinearität

Was bedeutet ANOVA?

- ▶ Analysis of Variance (Varianzanalyse)
- ▶ Eng verwandt mit drei bekannten Verfahren
 - ▶ Berechnung von η^2
 - ▶ t -Test
 - ▶ F -Test
- ▶ Grundfrage: Unterscheidet sich die Mittelwerte einer kontinuierlichen Variablen ...
- ▶ über verschiedene Gruppen (kategoriale Variable) hinweg?

F-Test

- ▶ Struktur
 - ▶ $H_0 : \mu_1 = \mu_2 = \mu_3 \dots$
 - ▶ $H_A : \mu_j \neq \mu_{j'}$
- ▶ Annahmen (niemals völlig korrekt)
 1. Zufallsstichprobe(n)
 2. Merkmal innerhalb der Gruppen normalverteilt
 3. Mit konstanter Varianz



Was sind nochmal „Freiheitsgrade“?

- ▶ Wieviele *unabhängige* Informationen haben wir für die Schätzung eines statistischen Parameters?
- ▶ Generell:
 - ▶ Zahl der (voneinander unabhängigen) Fälle N **minus**
 - ▶ Zahl der als Zwischenschritte benötigten Parameter k (Restriktionen)
- ▶ Grundidee: Durch *wiederholte* Schätzung auf Basis derselben Daten wird Information verbraucht
- ▶ Schätzungen für Parameter zweiter, dritter etc. Ordnung mit größerer Unsicherheit
- ▶ Beispiel Schätzung Mittelwert in Grundgesamtheit
 - ▶ Auf Basis einer Stichprobe von $N = 100 \rightarrow N = 100$ Freiheitsgrade
 - ▶ Zufällige Verteilung der Schätzungen über viele Stichproben
- ▶ Beispiel Schätzung Varianz in Grundgesamtheit

Verbindung Freiheitsgrade/Zufallsverteilungen?

- ▶ Freiheitsgrade als Parameter für die *Form* von Zufallsverteilungen
- ▶ Normalverteilung
 - ▶ *Keine* Freiheitsgrade als Parameter
 - ▶ Modelliert Schätzung auf Basis **sehr vieler** voneinander unabhängige Informationen
- ▶ *t*-Verteilung
 - ▶ Freiheitsgrade als Parameter
 - ▶ Je weniger Freiheitsgrade (unabhängige Informationen), desto mehr extreme Werte
 - ▶ Modell für unsicherere Schätzungen

Beispiel: Ideologie und Parteibindung

TABLE 12.1: Political Ideology by Political Party ID, for Subjects Age 18–30

Group (Party ID)	Political Ideology							Sample Size	Mean	Standard Deviation
	1	2	3	4	5	6	7			
Democrat	9	20	17	36	4	5	0	91	3.23	1.28
Independent	7	11	17	48	12	11	5	111	3.90	1.43
Republican	0	2	7	23	23	17	2	74	4.70	1.10

Note: For political ideology, 1 = extremely liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = extremely conservative.

- ▶ Nicht wirklich intervallskaliert
- ▶ Unterscheiden sich Gruppen in ihrer zentralen Tendenz?

Logik: Varianz innerhalb und zwischen Gruppen

- ▶ Viel Varianz zwischen, wenig innerhalb der Gruppen:
 - ▶ η^2 : Gruppenmitgliedschaft wichtig
 - ▶ F -Test: Starker Hinweis darauf, daß sich Gruppenmittelwerte in Grundgesamtheit unterscheiden
- ▶ Was bedeutet „in“ / „zwischen“ den Gruppen?
 - ▶ „Gesamtvarianz“: Varianz aller Werte um den Gesamtmittelwert
 - ▶ „In den Gruppen“: Varianzen der Meßwerte in den Gruppen (z. B. Republikaner) um die *Gruppenmittelwerte*
 - ▶ „Zwischen den Gruppen“: Varianz der *Gruppenmittelwerte* um den Gesamtmittelwert
- ▶ Wenn H_0 gilt: Alle drei Varianzen unverzerrte Schätzungen für wahre Gesamtvarianz in Grundgesamtheit
- ▶ F : $\frac{\text{Schätzung } \sigma^2 \text{ zwischen Gruppen}}{\text{Schätzung } \sigma^2 \text{ innerhalb Gruppen}}$
- ▶ Wenn H_0 :

unverzerrte Schätzung für Gesamtvarianz in Grundgesamtheit

Varianz *innerhalb* der Gruppen

- ▶ Abweichungen vom jeweiligen Gruppenmittelwert → Summe der SAQ → Summe der SAQ gesamt
- ▶ Freiheitsgrade (Nenner)?
 - ▶ Zahl der Fälle N **minus**
 - ▶ Anzahl der Parameter, auf denen Varianzschätzung basiert (g Gruppenmittelwerte) =
 - ▶ $N - g$
- ▶ Unverzerrte (da korrigierte) Schätzung für die Gesamtvarianz σ^2 (unabhängig von H_0)

Varianz *zwischen* den Gruppen

- ▶ Quadrierte Abweichungen der Gruppenmittelwerte vom Gesamtmittelwert
- ▶ Gewichten mit Gruppengröße
 - ▶ Berücksichtigt im Mittelwert enthaltene Information
 - ▶ Erzeugt wieder SAQ
- ▶ Freiheitsgrade:
 - ▶ Zahl der „Fälle“ (Gruppenmittelwerte) g **minus**
 - ▶ Anzahl der Parameter, auf denen Varianzberechnung basiert (Gesamtmittelwert, 1) =
 - ▶ $g - 1$
- ▶ Unverzerrte (da korrigierte) Schätzung für Gesamtvarianz σ^2 (*wenn H_0 gilt, sonst höher*)

Berechnung von F

- ▶ Beide Varianzschätzungen
 - ▶ χ^2 -verteilt
 - ▶ mit entsprechenden df (basieren auf Quadrierung normalverteilter zufälliger Abweichungen)
- ▶ Quotient beider Schätzungen 1 wenn H_0
- ▶ Zufällige Abweichungen davon F -verteilt mit df für Zähler und Nenner
 - ▶ Signifikanztest
 - ▶ Zerlegung Gesamt-SAQ in erklärte/nicht-erklärte SAQ

F- und t-Test

- ▶ Für zwei Gruppen $F = t^2 \rightarrow$ identisches Ergebnis für beide Tests
- ▶ Für große Zahl von Gruppen (z. B. 10), $\frac{g \times (g-1)}{2} = 45$ paarweise Tests
- ▶ Für jeden Test Irrtumswahrscheinlichkeit α (z. B. 5%), H_0 zu Unrecht aufzugeben
- ▶ Irrtumswahrscheinlichkeit sehr viel größer als 5%
 - ▶ Korrekturen (Bonferroni: gewünschte Irrtumswahrscheinlichkeit α durch Zahl der Vergleiche teilen)
 - ▶ Oder F-Test

In Stata

```
. tabstat lrsselbstselbst,by(zweitstimme )
Summary for variables: lrsselbstselbst
by categories of: zweitstimme (zweitstimme )
```

zweitstimme	mean
FDP	6.428571
Grüne	3.652174
SPD	4.65
Piraten	3.857143
CDU/CSU	6.8
Linke	3
Total	4.842105

```
. anova lrsselbstselbst zweitstimme
```

Source	Partial SS	df	MS	F	Prob > F
Model	120.452158	5	24.0904315	15.10	0.0000
zweitstimme	120.452158	5	24.0904315	15.10	0.0000
Residual	111.653106	70	1.59504437		

ANOVA und Regression

- ▶ Statt ANOVA auch Regression mit $g - 1$ Dummy-Variablen möglich
- ▶ (ANOVA kann noch mehr, andere Forschungstraditionen)
- ▶ (Einfache) ANOVA ein Spezialfall des linearen Regressionsmodells
- ▶ **Regressionsmodell kann unabhängige kategoriale Variablen berücksichtigen**

In Stata

```
. tabstat lrsselbstselbst,by(zweitstimme )
Summary for variables: lrsselbstselbst
by categories of: zweitstimme (zweitstimme )
```

zweitstimme	mean
FDP	6.428571
Grüne	3.652174
SPD	4.65
Piraten	3.857143
CDU/CSU	6.8
Linke	3
Total	4.842105

```
. anova lrsselbstselbst zweitstimme
```

Source	Partial SS	df	MS	F	Prob > F
Model	120.452158	5	24.0904315	15.10	0.0000
zweitstimme	120.452158	5	24.0904315	15.10	0.0000
Residual	111.653106	70	1.59504437		

In Stata

```
. tabstat lrsselbstselbst,by(zweitstimme )
Summary for variables: lrsselbstselbst
by categories of: zweitstimme (zweitstimme )
```

zweitstimme	mean
FDP	6.428571
Grüne	3.652174
SPD	4.65
Piraten	3.857143
CDU/CSU	6.8
Linke	3
Total	4.842105

```
. anova lrsselbstselbst zweitstimme
```

Source	Partial SS	df	MS	F	Prob > F
Model	120.452158	5	24.0904315	15.10	0.0000
zweitstimme	120.452158	5	24.0904315	15.10	0.0000
Residual	111.653106	70	1.59504437		

Statistik II ANOVA und Transformationen (18/28)

- ▶ Freiheitsgrade Gesamtvarianz: $N - 1 = 75$

Kategoriale und kontinuierliche Variablen

- ▶ Regressionsmodell kann sowohl kategoriale als auch kontinuierliche Variablen enthalten
- ▶ Beispiel im Text:
 - ▶ Einkommen als Funktion von Bildung und Ethnizität
 - ▶ Interaktion?

Einkommen, Bildung und Ethnizität

```
. tabstat inc ,by(race)
Summary for variables: inc
by categories of: race
```

race	mean
b	27.75
h	31
w	42.48
Total	37.525

```
. tabstat educ ,by(race)
Summary for variables: educ
by categories of: race
```

race	mean
b	12.25
h	11.64286
w	13.12
Total	12.6875

Regression: Einkommen, Bildung und Ethnizität

```
. reg inc black hisp
```

Source	SS	df	MS
Model	3352.47	2	1676.235
Residual	30409.48	77	394.928312
Total	33761.95	79	427.366456

Number of obs = 80
F(2, 77) = 4.24
Prob > F = 0.0178
R-squared = 0.0993
Adj R-squared = 0.0759
Root MSE = 19.873

inc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
black	-14.73	5.708028	-2.58	0.012	-26.09614 -3.363864
hisp	-11.48	6.008971	-1.91	0.060	-23.44539 4.853897
_cons	42.48	2.810439	15.12	0.000	36.8837 48.0763

```
. reg inc black hisp educ
```

Source	SS	df	MS
Model	15597.7019	3	5199.23398
Residual	18164.2481	76	239.003264
Total	33761.95	79	427.366456

Number of obs = 80
F(3, 76) = 21.75
Prob > F = 0.0000
R-squared = 0.4620
Adj R-squared = 0.4408
Root MSE = 15.46

inc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
black	-10.87445	4.47302	-2.43	0.017	-19.78324 -1.965656
hisp	-4.933792	4.763206	-1.04	0.304	-14.42054 4.552954
educ	4.431669	.6191355	7.16	0.000	3.198553 5.664784
_cons	-15.66349	8.412141	-1.86	0.066	-32.41772 1.090741

Regression mit Interaktion

```

. gen beduc=black * educ
. gen heduc=hispanic * educ
. reg inc black hispanic educ beduc heduc

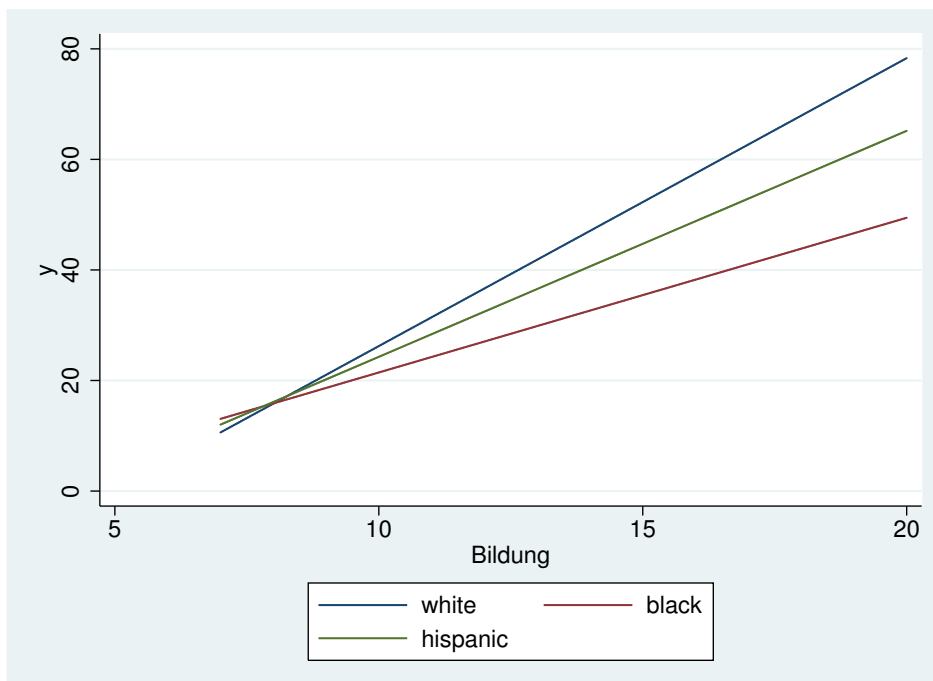
```

Source	SS	df	MS			
Model	16289.5385	5	3257.9077	Number of obs =	80	
Residual	17472.4115	74	236.113669	F(5, 74) =	13.80	
Total	33761.95	79	427.366456	Prob > F =	0.0000	
				R-squared =	0.4825	
				Adj R-squared =	0.4475	
				Root MSE =	15.366	

inc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
black	19.33327	18.29277	1.06	0.294	-17.11586	55.7824
hispanic	9.264024	24.27975	0.38	0.704	-39.11443	57.64248
educ	5.20951	.7828388	6.65	0.000	3.64967	6.76935
beduc	-2.410693	1.41773	-1.70	0.093	-5.235582	.4141956
heduc	-1.120759	2.006036	-0.56	0.578	-5.117872	2.876355
_cons	-25.86877	10.49822	-2.46	0.016	-46.78692	-4.95062

- ▶ Erwarteter Wert für Weiße: Konstante + Educ × Jahre Bildung
- ▶ Erwarteter Wert für Hispanics: Konstante + Educ × Jahre Bildung + Hispanic + (Hispanic × Educ) × Jahre Bildung

Graphisch



Nicht-lineare Effekte?

- ▶ In manchen (wenigen) Fällen ist die Linearitätsannahme offensichtlich unplausibel
- ▶ Z. B. kurvilinearere Zusammenhang zwischen Alter und Rechtsextremismus
- ▶ Wenn gute theoretische Begründung vorhanden, können Transformationen von y und/oder x sinnvoll sein, die den Zusammenhang zwischen beiden linearisieren
- ▶ In diesem Fall ist OLS unproblematisch
- ▶ Verwendet werden normalerweise das Quadrat, die Quadratwurzel, deren Kehrwerte und der natürliche Logarithmus („ladder of powers“)
- ▶ Tendenziell: Vorsicht

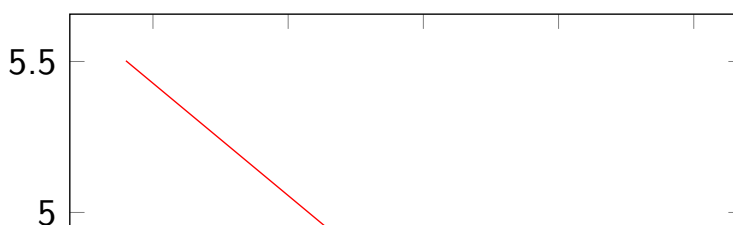
Beispiel: Alter und Ideologie

- ▶ Linearer Effekt des Lebensalters auf Links-Rechts-Selbsteinstufung?
- ▶ Daten: GLES 2009, Nachwahl, Ostdeutschland

```
. reg linksrechts alter
```

Source	SS	df	MS			
Model	61.599907	1	61.599907	Number of obs =	575	
Residual	2107.68009	573	3.67832477	F(1, 573) =	16.75	
Total	2169.28	574	3.77923345	Prob > F	= 0.0000	
				R-squared	= 0.0284	
				Adj R-squared	= 0.0267	
				Root MSE	= 1.9179	

linksrechts	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
alter	-.0186283	.0045521	-4.09	0.000	-.027569	-.0096875
_cons	5.809317	.2500042	23.24	0.000	5.318281	6.300354



Kurvilinearer Zusammenhang?

- ▶ Modell paßt sehr schlecht
- ▶ Zusammenhang möglicherweise U-förmig?
- ▶ Ideologie abhängig von
 - ▶ Konstante
 - ▶ Alter
 - ▶ Quadrat des Alters
- ▶ `gen altersq= alter * alter`

```
. reg linksrechts alter altersq
```

Source	SS	df	MS			
Model	63.487205	2	31.7436025	Number of obs =	575	
Residual	2105.7928	572	3.68145594	F(2, 572) =	8.62	
Total	2169.28	574	3.77923345	Prob > F	= 0.0002	
				R-squared	= 0.0293	
				Adj R-squared	= 0.0259	
				Root MSE	= 1.9187	

linksrechts	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
alter	.0008306	.0275562	0.03	0.976	-.0532931	.0549542
altersq	-.0001921	.0002683	-0.72	0.474	-.000719	.0003348
_cons	5.376145	.6546531	8.21	0.000	4.090328	6.661963

Statistik II ANOVA und Transformationen (26/28)

Stata-Skript für heute

- ▶ `net from`
`http://www.kai-arzheimer.com/Statistik-II/stata/`
- ▶ `net get anova`

Zusammenfassung

- ▶ Gewichtung – oft nicht notwendig
- ▶ Kohortenanalyse – oft problematisch
- ▶ ANOVA/Varianzanalyse traditionelles Verfahren für Experimentaldaten
 - ▶ Spezialfall des linearen Regressionsmodells
 - ▶ Regressionsmodell um nominale unabhängige Variablen erweiterbar
 - ▶ Interaktionen zwischen diesen und kontinuierlichen Variablen möglich
- ▶ Nächste Woche: Voraussetzungen für Anwendung Regressionsmodell → noch viel mehr Erweiterungen