

Institut für Soziologie  
Dipl.-Soz. Benjamin Gedon

# Methoden 2

Ausblick; Darstellung von Ergebnissen; Wiederholung



Ein (nicht programmierbarer) Taschenrechner kann in der Klausur hilfreich sein.





- Ausblick über weitere Analyseverfahren
- Wiederholung
- Darstellung von Ergebnissen in der Hausarbeit



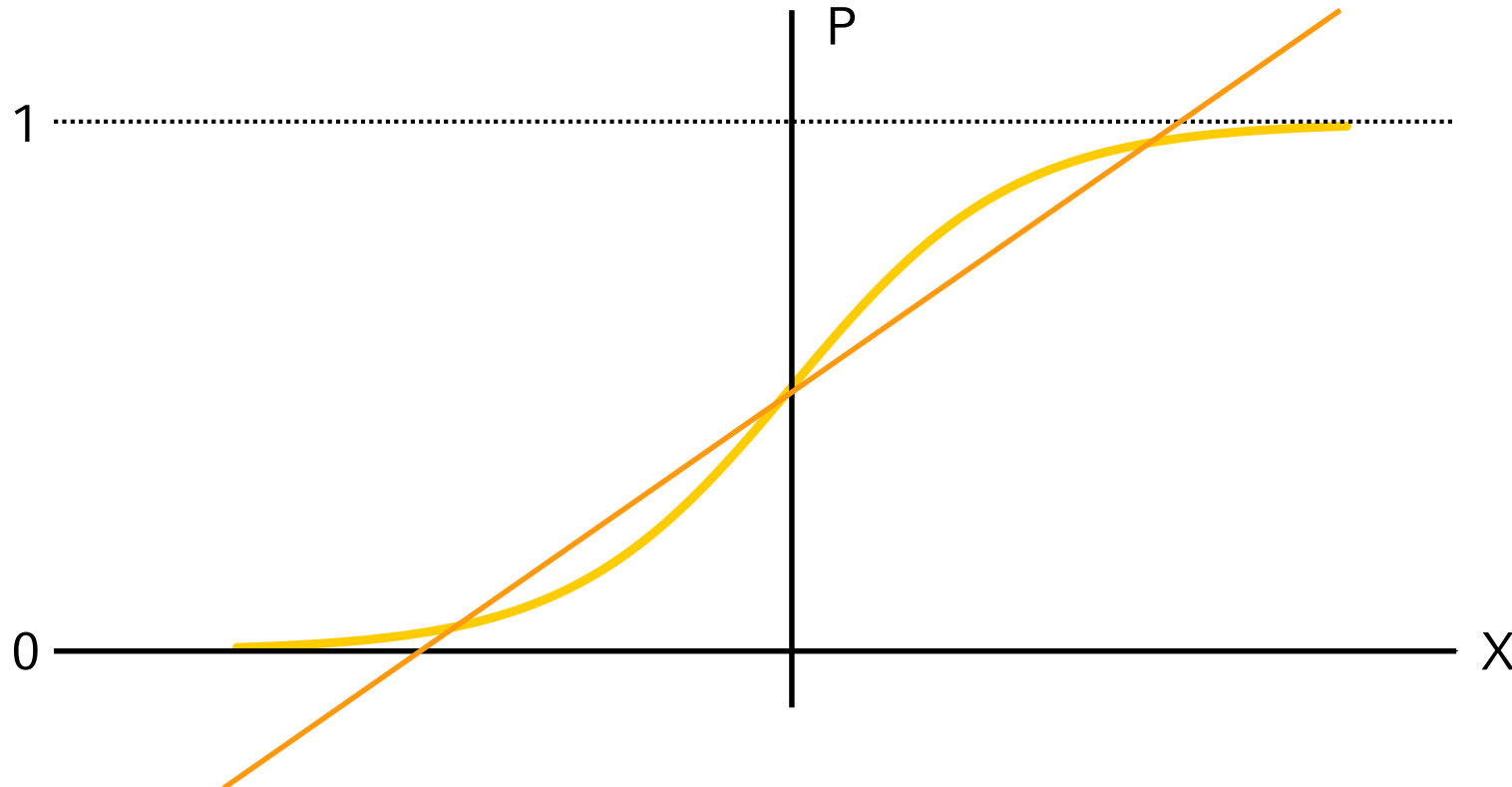
Lineare Regression: metrische abhängige Variable. Oft nicht zutreffend.

- Binäre AV: verheiratet - ledig
- Nominale AV: Wahlentscheidung zwischen Linke – Grüne – SPD – CSU – FDP – AFD
- Ordinale AV: höchster erreichter Schulabschluss
- AV Zähldaten: Anzahl der Fernseher im Haushalt
- AV auf anderer Ebene als UV: Noten von Schülern auch über Klassen- und Schulmerkmale erklären
- AV Ereignisdaten: Dauer bis eine Ehe geschieden wird



## Binäre abhängige Variable – logistische Regression

- abhängige Variable: arbeitslos / nicht arbeitslos; Raucher ja/nein
- Geschätzt werden soll die Wahrscheinlichkeit ob ein Ereignis eintritt.





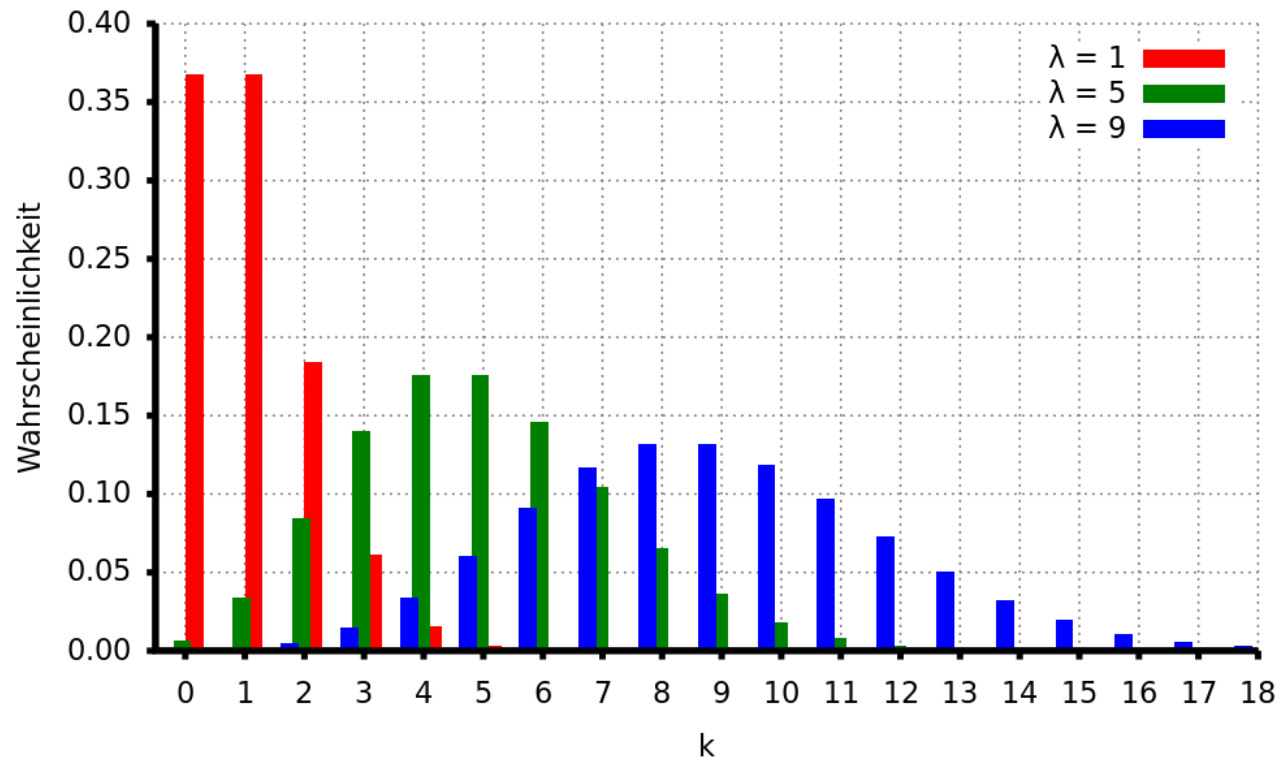
## Nominale abhängige Variable – multinomiale logistische Regression

- abhängige Variable: Wahlentscheidung zwischen verschiedenen Parteien bspw. Partei A, B, C
- Entscheidungsmöglichkeit zwischen A und B; A und C; B und C
- Wie bei Dummykodierung eine Referenzkategorie bspw. A
- Es wird nun der Einfluss der UV darauf untersucht ob es wahrscheinlicher ist B als A zu wählen und ob es wahrscheinlicher ist C als A zu wählen.
- Zu jeder UV gibt es also in unserem Beispiel 2 Koeffizienten.
- Grundlage ist logistisches Regressionsmodell.



## Abhängige Variable Zähldaten – Poissonregression

- abhängige Variable bspw. Kinderzahl
- Verteilung der AV bei seltenen Ereignissen sehr schief





AV auf anderer Ebene als UV – Mehrebenenmodelle

- abhängige Variable Lernerfolg einzelner Schüler
- unabhängige Variablen Klassenmerkmale
- Problem bei linearer Regression:
  - Daten sind nicht unabhängig
  - Standardfehler werden falsch berechnet – Signifikanzen sind falsch
- Mehrebenenmodell berücksichtigt die Datenstruktur





## AV Ereignisdaten – Ereignisdatenanalyse

- abhängige Variable: Dauer einer Ehe bis zur Scheidung
- Problem bei linearer Regression:
  - Nicht alle Ehen wurden bis zur Scheidung beobachtet.
  - Ehedauer bis zur Scheidung ist für viele Ehen also unbekannt.
  - Tendenziell ist sie für die Ehen unbekannt (zensiert), die erst nach längerer Zeit geschieden werden.
  - Die Ehedauer bis zur Scheidung wird also verzerrt geschätzt.
- Ereignisdatenmodelle können mit diesem Problem umgehen

Univariate  
Deskription

## Arbeitsstunden real

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	20	0	Obs	113
25%	31	0	Sum of Wgt.	113
50%	40		Mean	35.80973
		Largest	Std. Dev.	12.25607
75%	42	50		
90%	45	50	Variance	150.2112
95%	50	55	Skewness	-1.450213
99%	55	60	Kurtosis	4.996826

## Interpretation:

- **arithmetisches Mittel (=Mittelwert):**  
36 Stunden
- **Median:** Zentrum der Verteilung bei  
40 Stunden
- **Standardabweichung:** Durchschnittliche  
Abweichung vom Mittelwert: 12 Stunden
- **Schiefe:**  $< 0 \rightarrow$  linksschiefe Verteilung
- **Kurtosis (Wölbung):**  $> 0 \rightarrow$  Verteilung  
spitzer als Normalverteilung
- **Minimum/Maximum:**  
 $\rightarrow$  Minimum: 0 Stunden  
 $\rightarrow$  Maximum: 60 Stunden



## Kreuztabelle

Nebenfach Statistik?	Frau ja/nein		Total
	0. nein	1. ja	
0. nein	50 89.29	99 88.39	149 88.69
1. ja	6 10.71	13 11.61	19 11.31
Total	56 100.00	112 100.00	168 100.00

Pearson chi2 (1) = 0.0297 Pr = 0.863



## Korrelation

	einkom~n	arbstd~l
einkommen	1.0000	
arbstdreal	0.4495	1.0000
	0.0000	



## Mittelwertvergleich

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
männlich	34	2621.941	162.0211	944.7371	2292.307	2951.576
weiblich	67	2338.224	108.4392	887.6132	2121.718	2554.73
combined	101	2433.733	90.79818	912.5104	2253.592	2613.874
diff		283.7173	190.9928		-95.25395	662.6885

diff = mean(männlich) - mean(weiblich)

t = 1.4855

Ho: diff = 0

degrees of freedom = 99

Ha: diff < 0

Pr(T < t) = 0.9297

Ha: diff != 0

Pr(|T| > |t|) = 0.1406

Ha: diff > 0

Pr(T > t) = 0.0703



## Regression

Source	SS	df	MS	Number of obs	=	99
Model	22544121.2	3	7514707.05	F(3, 95)	=	12.29
Residual	58069450.5	95	611257.373	Prob > F	=	0.0000
Total	80613571.6	98	822587.466	R-squared	=	0.2797
				Adj R-squared	=	0.2569
				Root MSE	=	781.83

einkommen	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
frau	-137.7868	171.1093	-0.81	0.423	-477.4818	201.9082
nfstat	830.2007	306.697	2.71	0.008	221.3302	1439.071
arbstdreal	35.68307	7.012726	5.09	0.000	21.76105	49.60509
_cons	1189.38	306.0525	3.89	0.000	581.7895	1796.972



## Regression mit quadratischem Effekt

Source	SS	df	MS	Number of obs	=	99
Model	23153482	4	5788370.51	F(4, 94)	=	9.47
Residual	57460089.6	94	611277.549	Prob > F	=	0.0000
Total	80613571.6	98	822587.466	R-squared	=	0.2872
				Adj R-squared	=	0.2569
				Root MSE	=	781.84

einkommen	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
frau	-151.8831	171.6936	-0.88	0.379	-492.7849	189.0186
nfstat	756.1185	315.5496	2.40	0.019	129.5873	1382.65
arbstdreal	11.38279	25.32866	0.45	0.654	-38.90785	61.67343
arbstd2	.4448372	.4455362	1.00	0.321	-.4397854	1.32946
_cons	1446.039	399.6902	3.62	0.000	652.4452	2239.634



## Regression mit quadratischem Effekt

	Einkommen	
	Modell 1	Modell 2
Frau	-137,8 (0,423)	-151,9 (0,379)
Nebenfach Statistik?	830,2 (0,008)	756,1 (0,019)
Arbeitsstunden real	35,68 (0,000)	11,38 (0,654)
Arbeitsstunden quadriert		0,445 (0,321)
Konstante	1,189 (0,000)	1,446 (0,000)
Fallzahl	99	99
R <sup>2</sup>	0,280	0,287

p-Werte in Klammern





## Regression mit Interaktion Dummy \* Dummy

Source	SS	df	MS	Number of obs	=	99
				F(5, 93)	=	8.17
Model	24613098.2	5	4922619.64	Prob > F	=	0.0000
Residual	56000473.5	93	602155.629	R-squared	=	0.3053
				Adj R-squared	=	0.2680
Total	80613571.6	98	822587.466	Root MSE	=	775.99

einkommen	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
frau	-82.97161	176.0622	-0.47	0.639	-432.5962	266.653
nfstat	1495.267	568.7485	2.63	0.010	365.8448	2624.688
intfraustat	-1079.701	693.4876	-1.56	0.123	-2456.83	297.4275
c_arbstdreal	46.82076	10.4971	4.46	0.000	25.9756	67.66591
c_arbstd2	.6160425	.4556671	1.35	0.180	-.2888221	1.520907
_cons	2357.446	152.6179	15.45	0.000	2054.377	2660.515



Die folgenden Hinweise sind nicht erschöpfend.

- Keine Stata-Outputs übernehmen!
- Alle Werte die interpretiert werden in Tabellen
- Nur die Maßzahlen, die interpretiert werden in Tabellen
- Modellvergleiche übersichtlich nebeneinander
- Tabellen selbsterklärend (Legende, Referenzkategorie in Fußzeile)



- Die Klausur findet am 24.7.2017 um 10.00 Uhr ausschließlich im Hörsaal A 240 am Geschwister-Scholl-Platz 1 statt.
- Tutorien zur Klausurvorbereitung finden am Mittwoch von 16 bis 20 Uhr im Raum 409 in der Konradstraße 6 statt.
- Ein Taschenrechner kann in der Klausur hilfreich sein.