

# Multiple Regression II: Signifikanztests, Gewichtung, Multikollinearität und Kohortenanalyse

## Statistik II

Wiederholung  
Inferenz und standardisierte Koeffizienten  
Gewichtung  
Kollinearität und Kohorten  
Zusammenfassung

## Übersicht

Wiederholung  
  Literatur  
  Kausalität und Regression  
Inferenz und standardisierte  
Koeffizienten  
  Inferenz für multivariate Modelle  
  Standardisierte Koeffizienten  
Gewichtung  
Kollinearität und Kohorten  
Zusammenfassung

## Literatur für heute

- ▶ Agresti ch. 11
- ▶ Mason/Wolfinger: Cohort Analysis.
- ▶ Kish: Weighting: Why, When, and How?
- ▶ (alle im ReaderPlus)

## Literatur für nächste Woche

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

## Daten/Kommandos für heute

- ▶ regression2.do

## Kausalität

- ▶ Im strengen Sinne nur hypothetisch-kontrafaktische Prüfung
- ▶ Real:
  - ▶ Experimentaldesign
  - ▶ Ex post facto Design
- ▶ Voraussetzung, um von Kausalität sprechen zu können (**kein endgültiger Beweis**):
  1. Theoretisch plausibel
  2. Statistische Assoziation (Korrelation)
  3. Zeitliche Reihenfolge
  4. Drittvariablen ausschließen

## Multiple Regression

- ▶ Bivariate Regression mit zusätzlichen unabhängigen Variablen
- ▶ Unabhängige Variablen wirken linear-additiv zusammen
- ▶ Regressionskoeffizienten sind „partiell“ (Wert aller anderen Variablen wird „konstant gehalten“)
- ▶ Schätzung der Koeffizienten mit OLS (minimiert die quadrierten Abweichungen zwischen gemessenen/erwarteten Werten)
- ▶ Fit des Regressionsmodells wie im bivariaten Fall mit  $R^2$  und RMSE

## Voraussetzungen für Inferenz

1. *Konditionale* Verteilung von  $y$  (Residuen) normal (robust für große Fallzahlen)
2. Varianz der Residuen konstant für alle Kombinationen von Werten von  $x_1, x_2, x_3, \dots$  (Homoskedastizität)
3. Zufallsstichprobe
  - ▶ Konfidenzintervalle für Koeffizienten
  - ▶ Konfidenzintervall für einen vorhergesagten Wert von  $y$
  - ▶ Hypothesentests für
    - ▶ alle Koeffizienten gemeinsam (F-Test)
    - ▶ individuelle Koeffizienten (t-Test)

## Erinnerung: Logik Hypothesentests

- ▶  $H_0$  und  $H_A$
- ▶ Annahme: in der Grundgesamtheit gilt  $H_0$ ; wenn  $H_0$  gilt, ist Testgröße gleich null
- ▶ Wenn  $H_0$  in Grundgesamtheit gilt, weicht Testgröße in Stichprobe wg. zufälliger Fehler trotzdem von 0 ab
- ▶ Sehr große Abweichungen sehr unwahrscheinlich, wenn  $H_0$  tatsächlich gilt
- ▶ Festlegen einer akzeptablen Irrtumswahrscheinlichkeit (z. B. 5%)
- ▶ „Irrtum“:  $H_0$  wird zu Unrecht aufgegeben, obwohl tatsächlich gültig
- ▶ Wenn Wahrscheinlichkeit  $< \alpha$ , „signifikantes“ Ergebnis

## Erinnerung: Bestimmung empirische Irrtumswahrscheinlichkeit

- ▶ Vergleich der Prüfgröße (z. B. empirischer t-Wert)
- ▶ Mit seiner theoretischen Verteilung
  - ▶ Normalverteilung
  - ▶ t-Verteilung
  - ▶  $\chi^2$ -Verteilung
  - ▶ ...
- ▶ Dichte der Verteilung (Parameter beachten)
- ▶ Kumulierte Verteilung (Integral)
- ▶ Wie wahrscheinlich ist empirischer Wert, wenn  $H_0$  gilt?

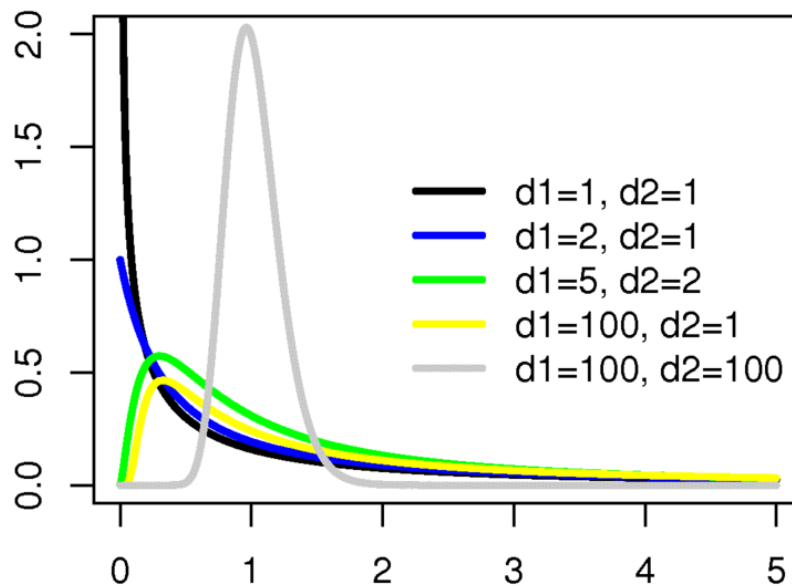
## Die F-Verteilung

- ▶ Eine der wichtigen Wahrscheinlichkeitsverteilungen
- ▶ Benannt nach R. A. Fisher
- ▶ F-Werte: Werte die sich ergeben
  - ▶ Wenn jeweils zwei Werte aus voneinander unabhängigen  $\chi^2$ -Verteilungen gezogen werden
  - ▶ Und der Quotient gebildet wird
- ▶ Modell für das Verhältnis zweier positiver Zufallszahlen
- ▶ Kontinuierliche Verteilung mit zwei Parametern (Freiheitsgrade)
- ▶ Nur positive Werte

## 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

## F-Verteilungen mit verschiedenen Freiheitsgraden



## Die F-Prüfstatistik

- ▶  $H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$
- ▶  $H_A : \text{Wenigstens ein } \beta_k \neq 0$  (unter Kontrolle aller anderen Variablen)
- ▶ Bzw. äquivalent:
  - ▶  $H_0 : R^2$  (in der Grundgesamtheit) = 0
  - ▶  $H_1 : R^2$  (in der Grundgesamtheit)  $\neq 0$
- ▶ Warum nicht  $\beta_0$ ?

$$F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} \quad (1)$$

- ▶ Wovon hängt  $F$  ab?
  1. „Erklärte Varianz“  $R^2$  – Erklärungsleistung über  $\bar{y}$  hinaus
  2.  $k$  Zahl der unabhängigen Variablen
  3.  $n$  Zahl der Fälle

# Wer mag Frau Merkel?

```
. reg polsympangelamerkel parteienrankingcdu parteienrankingcsu lrsselbstselb
> st galtanselbstselbst alter
```

Source	SS	df	MS			
Model	137.081922	5	27.4163844	Number of obs =	76	
Residual	290.115446	70	4.14450638	F( 5, 70) =	6.62	
Total	427.197368	75	5.69596491	Prob > F	= 0.0000	
				R-squared	= 0.3209	
				Adj R-squared	= 0.2724	
				Root MSE	= 2.0358	

polysympang-l	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parteienr-du	.8298748	.2011086	4.13	0.000	.4287763	1.230973
parteienr-su	-.3445445	.1639751	-2.10	0.039	-.6715826	-.0175064
lrsselbsts-t	-.376568	.2325599	-1.62	0.110	-.8403941	.0872582
galtanselb-t	.3165245	.1487547	2.13	0.037	.0198425	.6132064
alter	.048202	.1416287	0.34	0.735	-.2342675	.3306715
_cons	2.404232	3.221482	0.75	0.458	-4.02081	8.829275

- ▶ F-Wert = 6.62
- ▶ Hoch signifikant
- ▶  $H_0$  ablehnen



# Die F-Prüfstatistik

- ▶  $H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$
- ▶  $H_A : \text{Wenigstens ein } \beta_k \neq 0$  (unter Kontrolle aller anderen Variablen)
- ▶ Bzw. äquivalent:
  - ▶  $H_0 : R^2$  (in der Grundgesamtheit) = 0
  - ▶  $H_1 : R^2$  (in der Grundgesamtheit)  $\neq 0$
- ▶ Warum nicht  $\beta_0$ ?

$$F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} \quad (1)$$

- ▶ Wovon hängt  $F$  ab?

1. „Erklärte Varianz“  $R^2$  – Erklärungsleistung über  $\bar{y}$  hinaus
2.  $k$  Zahl der unabhängigen Variablen
3.  $n$  Zahl der Fälle



## Wer mag Frau Merkel?

```
. reg polsympangelamerkel parteienrankingcdu parteienrankingcsu lrsselbstselb
> st galtanselbstselbst alter
```

Source	SS	df	MS			
Model	137.081922	5	27.4163844	Number of obs =	76	
Residual	290.115446	70	4.14450638	F( 5, 70) =	6.62	
Total	427.197368	75	5.69596491	Prob > F	= 0.0000	
				R-squared	= 0.3209	
				Adj R-squared	= 0.2724	
				Root MSE	= 2.0358	

polsympang-l	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parteienr-du	.8298748	.2011086	4.13	0.000	.4287763	1.230973
parteienr-su	-.3445445	.1639751	-2.10	0.039	-.6715826	-.0175064
lrsselbsts-t	-.376568	.2325599	-1.62	0.110	-.8403941	.0872582
galtanselb-t	.3165245	.1487547	2.13	0.037	.0198425	.6132064
alter	.048202	.1416287	0.34	0.735	-.2342675	.3306715
_cons	2.404232	3.221482	0.75	0.458	-4.02081	8.829275

- ▶ F-Wert = 6.62
- ▶ Hoch signifikant
- ▶  $H_0$  ablehnen



Statistik II

Multiple Regression II (16/37)

## Hypothesentests für einzelne Koeffizienten

- ▶ t-Test
- ▶ Wiederholte Stichprobenziehung: Koeffizienten t-verteilt
- ▶  $H_0 : \beta_j = 0$
- ▶  $H_A : \beta_j \neq 0$
- ▶ Standardfehler = Standardabweichung einer theoretischen t-Verteilung
- ▶ Koeffizienten durch ihren Standardfehler teilen → empirischer t-Wert
- ▶ Wie wahrscheinlich ist empirischer t-Wert bei Gültigkeit von  $H_0$ ? → Signifikanz

## Was beeinflusst den Standardfehler?

- ▶ Beim Standardfehler des Mittelwertes: Streuung des Merkmals und  $\sqrt{n}$
- ▶ Im bivariaten Fall:  $V(b_1) = \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}$ 
  - ▶ Streuung der Fehler in der Population (geschätzt)
  - ▶  $SAQ_x$ : Fallzahl und Varianz der unabhängigen Variablen
- ▶ Im multivariaten Fall:

$$V(b_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \text{ mit } j \neq 0$$

beziehungsweise in Matrix-Schreibweise:

$$\mathbf{V} = \sigma_\epsilon^2 \times (\mathbf{X}'\mathbf{X})^{-1} \text{ mit } V(b_0) \dots V(b_k) \text{ als Hauptdiagonale}$$

## In Stata...

```
. estat vce,format(%9.3f)
Covariance matrix of coefficients of regress model
```

e(V)	partei~du	partei~su	lrsselb~t	galtans~t	alter
parteienr~du	0.040				
parteienr~su	-0.019	0.027			
lrsselbsts~t	-0.022	-0.003	0.054		
galtanselb~t	-0.002	0.001	-0.017	0.022	
alter	0.004	-0.001	-0.003	0.001	0.020
_cons	-0.094	0.035	-0.000	-0.029	-0.446
e(V)	_cons				
_cons	10.378				

```
. corr parteienrankingcdu parteienrankingcsu lrsselbstselbst galtanselbstselbs
> t alter
(obs=76)
```

	parte~du	parte~su	lrssel~t	galtan~t	alter
parteienr~du	1.0000				
parteienr~su	0.7525	1.0000			
lrsselbsts~t	0.7551	0.5928	1.0000		
galtanselb~t	0.5437	0.3975	0.6769	1.0000	
alter	-0.0896	-0.0231	-0.0075	-0.0495	1.0000

## Zurück zum t-Test

```
. reg polsympangelamerkel parteienrankingcdu parteienrankingcsu lrsselbstselb
> st galtanselbstselbst alter
```

Source	SS	df	MS			
Model	137.081922	5	27.4163844	Number of obs = 76		
Residual	290.115446	70	4.14450638	F( 5, 70) = 6.62		
Total	427.197368	75	5.69596491	Prob > F = 0.0000		
				R-squared = 0.3209		
				Adj R-squared = 0.2724		
				Root MSE = 2.0358		

polysympang-l	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parteienr-du	.8298748	.2011086	4.13	0.000	.4287763	1.230973
parteienr-su	-.3445445	.1639751	-2.10	0.039	-.6715826	-.0175064
lrsselbsts-t	-.376568	.2325599	-1.62	0.110	-.8403941	.0872582
galtanselb-t	.3165245	.1487547	2.13	0.037	.0198425	.6132064
alter	.048202	.1416287	0.34	0.735	-.2342675	.3306715
_cons	2.404232	3.221482	0.75	0.458	-4.02081	8.829275

## CSU-Effekt: bivariat

```
. reg polsympangelamerkel parteienrankingcsu
```

Source	SS	df	MS			
Model	17.6908381	1	17.6908381	Number of obs = 79		
Residual	415.39777	77	5.39477623	F( 1, 77) = 3.28		
Total	433.088608	78	5.55241805	Prob > F = 0.0741		
				R-squared = 0.0408		
				Adj R-squared = 0.0284		
				Root MSE = 2.3227		

polysympang-l	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parteienr-su	.2193309	.121119	1.81	0.074	-.0218479	.4605096
_cons	4.914498	.4986422	9.86	0.000	3.921575	5.907422

## Warum standardisieren?

- ▶ Manchmal große Unterschiede in Skalierung der unabhängigen Variablen
- ▶ Einheiten oft arbiträr (monatliches Einkommen in K-Euro, monatliches Einkommen in Euro, Jahreseinkommen in Euro)
- ▶ Umrechnung von  $x$  (und evtl.  $y$ ) in Standardabweichungen
- ▶ Standardisierung macht Effekte unterschiedlicher Werte scheinbar vergleichbar

## Wie standardisieren?

- ▶ Vorab alle Variablen durch jeweilige Standardabweichung teilen
- ▶ Oder ex-post Koeffizienten durch  $s_y$  teilen und mit  $s_x$  multiplizieren

$$y = 2 + 5x$$

$$5 \Leftrightarrow 1$$

$$(s_y = 10) \quad (s_x = 20)$$

$$5 \Leftrightarrow \frac{1}{20}$$

$$\frac{1}{2} \Leftrightarrow \frac{1}{20}$$

$$10 \Leftrightarrow 1$$

## In Stata...

```
. reg polsympangelamerkel parteienrankingcdu parteienrankingcsu lrsselbstselb
> st galtanselbstselbst alter,beta
```

Source	SS	df	MS			
Model	137.081922	5	27.4163844	Number of obs =	76	
Residual	290.115446	70	4.14450638	F( 5, 70) =	6.62	
				Prob > F =	0.0000	
				R-squared =	0.3209	
				Adj R-squared =	0.2724	
Total	427.197368	75	5.69596491	Root MSE =	2.0358	

polsympang-l	Coef.	Std. Err.	t	P> t	Beta
parteienr-du	.8298748	.2011086	4.13	0.000	.7693928
parteienr-su	-.3445445	.1639751	-2.10	0.039	-.3158423
lrsselbsts-t	-.376568	.2325599	-1.62	0.110	-.2802628
galtanselb-t	.3165245	.1487547	2.13	0.037	.2861729
alter	.048202	.1416287	0.34	0.735	.0339087
_cons	2.404232	3.221482	0.75	0.458	.

## Warum *nicht* standardisieren?

- ▶ *Ursprüngliche Einheit(en) geht/gehen verloren*
- ▶ Standardabweichung nicht extrem anschaulich
- ▶ **Standardisierte Koeffizienten nicht über Datensätze/Modelle vergleichbar** wg. unterschiedlicher Varianzen

## Was sind Gewichte?

### 1. Designgewichte

- ▶ Mehrstufige Zufallsauswahl
- ▶ Disproportionale Stichprobenziehung
- ▶ Kombination von Samples

### 2. Redressmentgewichte (Kompensation selektiver Ausfälle)

## Wie gewichtet man?

- ▶ Heute praktisch nur noch mit (kontinuierlichen) individuellen Gewichten
- ▶ Fall geht mit höheren (z. B. 2.4-fachem) oder niedrigerem (0.75-fachem) Gewicht in Berechnungen ein
- ▶ Mittleres Gewicht muß 1 sein wg. ursprünglicher Fallzahl
- ▶ Stata kennt vier Typen von Gewichten, für uns vor allem pweights interessant
- ▶ `regress y x1 x2 x3 [pweight=1/prob]`

## Warum kann Gewichtung problematisch sein?

- ▶ Extreme Unterschiede (15 vs. 0.5)
- ▶ Simultane Gewichtung nach verschiedenen (korrelierten) Merkmalen (Geschlecht, Alter, Region, Bildung)
- ▶ Welche Gewichtungsvariablen?
- ▶ Unter Umständen „schlechtere“ Standardfehler/Koeffizienten
- ▶ In Regressionsmodellen nicht notwendig, wenn für Gewichtungsvariablen kontrolliert wird (sofern keine Interaktion)

## Was ist Kollinearität?

- ▶ Bei ex post facto Design Korrelationen zwischen unabhängigen Variablen
- ▶ Größere Standardfehler, da weniger unabhängige Information in den Daten
- ▶ Bei sehr engen Beziehungen „Kollinearität“ → instabile Schätzungen und Standardfehler
- ▶ VIF = Variance Inflation Factor
  - ▶ Wie gut kann eine unabhängige Variable durch andere unabhängige Variablen erklärt werden?
  - ▶  $VIF(\beta_j) = \frac{1}{1-R_j^2}$ ;  $\sqrt{VIF(\beta_j)}$ : „Inflation“ des Standardfehlers
  - ▶  $VIF > 5$  oder  $10 \rightarrow$  Problem

## In Stata

```
. estat vif
```

Variable	VIF	1/VIF
parteienr~du	3.58	0.279069
lrsselbsts~t	3.09	0.323840
parteienr~su	2.33	0.429376
galtanselb~t	1.86	0.536365
alter	1.02	0.977348
Mean VIF	2.38	

## Was ist perfekte Kollinearität?

- ▶ Zwei unabhängige Variablen sind identisch bzw. lineare Transformation
- ▶ Keine eindeutige Lösung für Normalgleichungen bzw. kein eindeutiges Minimum für SAQ
- ▶  $VIF = \frac{1}{1-1} = \frac{1}{0}$
- ▶ Z. B.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{mit} \quad x_1 = x_2$$

$$y = 5 + 3x_1 + 0x_2$$

$$y = 5 + 0x_1 + 3x_2$$

$$y = 5 + 1.5x_1 + 1.5x_2$$



## Typisches Kohortendesign

- ▶ Kohorte: Prägendes Ereignis zum gleichen Zeitpunkt (z. B. Geburt, Wahlrecht)
- ▶ Verhalten/Einstellungen geprägt von
  - ▶ Alter (in Jahren, Lebenszyklus)
  - ▶ Periode (Jahreszahl, Meßzeitpunkt)
  - ▶ Kohortenzugehörigkeit (Geburtsjahr, Prägung durch Ereignisse)
- ▶ Beispiel Sympathie für die Grünen
  - ▶ Unabhängig von Zeitpunkt und Generation höher bei jüngeren Menschen (weniger konservativ)
  - ▶ Unabhängig von Alter und Generation höher in 1987 (wegen Tschernobyl)
  - ▶ Unabhängig von Alter und Zeitpunkt höher bei 1968 Geborenen (wegen Sozialisation in 1980er Jahren)

## Problem beim Kohortendesign

- ▶ Grundproblem
  - ▶ Alter in Jahren ist eine Funktion von Zeitpunkt und Geburtsjahr
  - ▶ Zeitpunkt ist eine Funktion von Geburtsjahr und Alter
  - ▶ Generation/Geburtsjahr ist eine Funktion von Alter und Zeitpunkt
- ▶ In der Theorie haben alle drei Variablen unabhängige Effekte
- ▶ Typischerweise Dummy-Variablen
- ▶ In der Praxis perfekte Multikollinearität: niemand, der 1968 geboren ist und 1990 befragt wird, ist *nicht* 22 Jahre alt
- ▶ Modell nicht schätzbar (unabhängig von Datenstruktur)

## Lösungsstrategien: Restriktionen

- ▶ Eine Variable weglassen und/oder Gleichheitsrestriktionen für Kategorien einführen
- ▶ Problem: nimmt nicht vorhandenen/identischen Effekt an
- ▶ Wenn Annahme falsch, verzerrte Schätzung für *alle* Effekte
- ▶ Aus Daten läßt sich *nicht* ableiten, ob Effekte vorhanden sind

## Lösungsstrategien: inhaltliche Variablen

- ▶ Alter, Geburtsjahr, Jahreszahl haben keine *direkten* politischen Konsequenzen
- ▶ Sondern stehen für soziale Prozesse
  - ▶ Involvierung in Friedens/Umweltbewegung während prägender Phase
  - ▶ Exposition gegenüber politischen Stimuli zum Zeitpunkt der Messung
  - ▶ Vorliegen von Lebensereignissen, die konservativer machen
- ▶ Inhaltliche Variablen
  - ▶ Theoretisch relevanter
  - ▶ Nicht perfekt mit Zeitvariablen kontrolliert (nicht jeder war 68 auf den Barrikaden)
- ▶ Zeit- durch inhaltliche Variable(n) ersetzen → Modell wird (vielleicht) schätzbar

## Beispiel: Logistisches Modell Wahlbeteiligung, ALLBUS 1980-2002

Periode		
GETAS	0,240**	(0,080)
IPSOS	0,115	(0,073)
INFAS	0,336**	(0,090)
Wahljahr	0,283**	(0,058)
Alter		
18-24	-0,005	(0,131)
25-29	0,120	(0,121)
30-34	0,089	(0,114)
(Referenz: 35-39)		
45-49	-0,166	(0,142)
50-54	-0,336*	(0,160)
55-59	-0,424*	(0,178)
60-64	-0,660**	(0,189)
65-69	-0,684**	(0,210)
70-	-1,043**	(0,221)
Kohorte		
(Referenz: -1921)		
1922-1934	-0,061	(0,121)
1935-1945	-0,375*	(0,167)
1946-1953	-0,665**	(0,208)
1954-1964	-1,109**	(0,233)
1965-1975	-1,487**	(0,264)
1976-1983	-1,173**	(0,327)

## Zusammenfassung

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