Analysen politikwissenschaftlicher Datensätze mit Stata

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Sitzung 5: Lineare Regression

Vorbereitung

- Stata durch z:\profile.do starten
- Datensatz z:\daten\rpstrukt laden

- Achtung: Ab dieser Sitzung werden Stata-Befehle nicht mehr durch Schreibmaschinenschrift hervorgehoben

Modelle

- Modelle sind eine extreme Vereinfachung der Wirklichkeit
- Statistische Modelle modellieren die Zusammenhänge zwischen Variablen
- Lineare Regression ist das einfachste und für die Politikwissenschaft wichtigste Modell

Lineare Regression

- Statistische Modelle beschreiben,
  - wie der erwartete Wert einer abhängigen Variablen
  - mit dem Wert einer oder mehrerer unabhängiger Variablen zusammenhängt
- Im Fall der linearen Regression ist dieser Zusammenhang *linear*, d.h. bei einem positiven Zusammenhang werden
  - für höhere/niedrigere Werte der unabhängigen Variablen
  - *proportional* höhere/niedrigere Werte der abhängigen Variablen erwartet
  - Zusammenhang kann durch eine Gerade veranschaulicht werden

Konkret:

- Geben Sie nochmals ein
  - summ pwbcdu71 if pkathv70>54
  - summ pwbcdu71 if pkathv70<54
- Dieser Zusammenhang kann durch die Formel CDU71=a+b*KATH70 beschrieben werden
- Allgemein y=a+b*x [systematischer Teil]

Erwartungswert

- Erwartungswert $\cong$ Mittelwertwert
  - Wenn wir sukzessive rheinland-pfälzische Wahlkreise betrachten, „erwarten“ wir für 1971 einen CDU-Anteil von 39,2 Prozent (Mittelwert)
  - „zentrale Tendenz“
  - „bester Tipp“ (geringste quardrierte Abweichung)
- In einem „leeren“ Regressionsmodell entspricht die Konstante dem Mittelwert der abhängigen Variablen: reg pwbcdu71

Erwartungswert

- In einem Regressionsmodell *mit einer unabhängigen Variablen* kann durch Einsetzen für jeden Wert der unabhängigen Variablen ein erwarteter Wert für die abhängige Variable bestimmt werden
- Auch dieser Wert ist ein *Erwartungswert*:
  - der konditionale (d.h. von x abhängige) Mittelwert für den Wert von y,
  - der sich nach unserem Modell ergeben würde, wenn wir viele Fälle mit dem entsprechenden x-Wert untersuchen würden

Erwartungswerte

- reg pwbcdu71 pkathv70
- Wir erwarten, daß der mittlere CDU-Anteil mit jedem Prozentpunkt Katholikenanteil um 0,3 Prozentpunkte zunimmt
- Die Konstante entspricht jetzt dem erwarteten CDU-Anteil, wenn der Katholikenanteil = 0
- graph twoway (scatter pwbcdu71 pkathv70) (lfit pwbcdu71 pkathv70)
- Transformation der x-Variablen erleichtert evtl. die Interpretation der Konstante
  - summ pkathv70
  - gen zpkathv70=pkathv70-r(mean)
  - reg pwbcdu71 zpkathv70

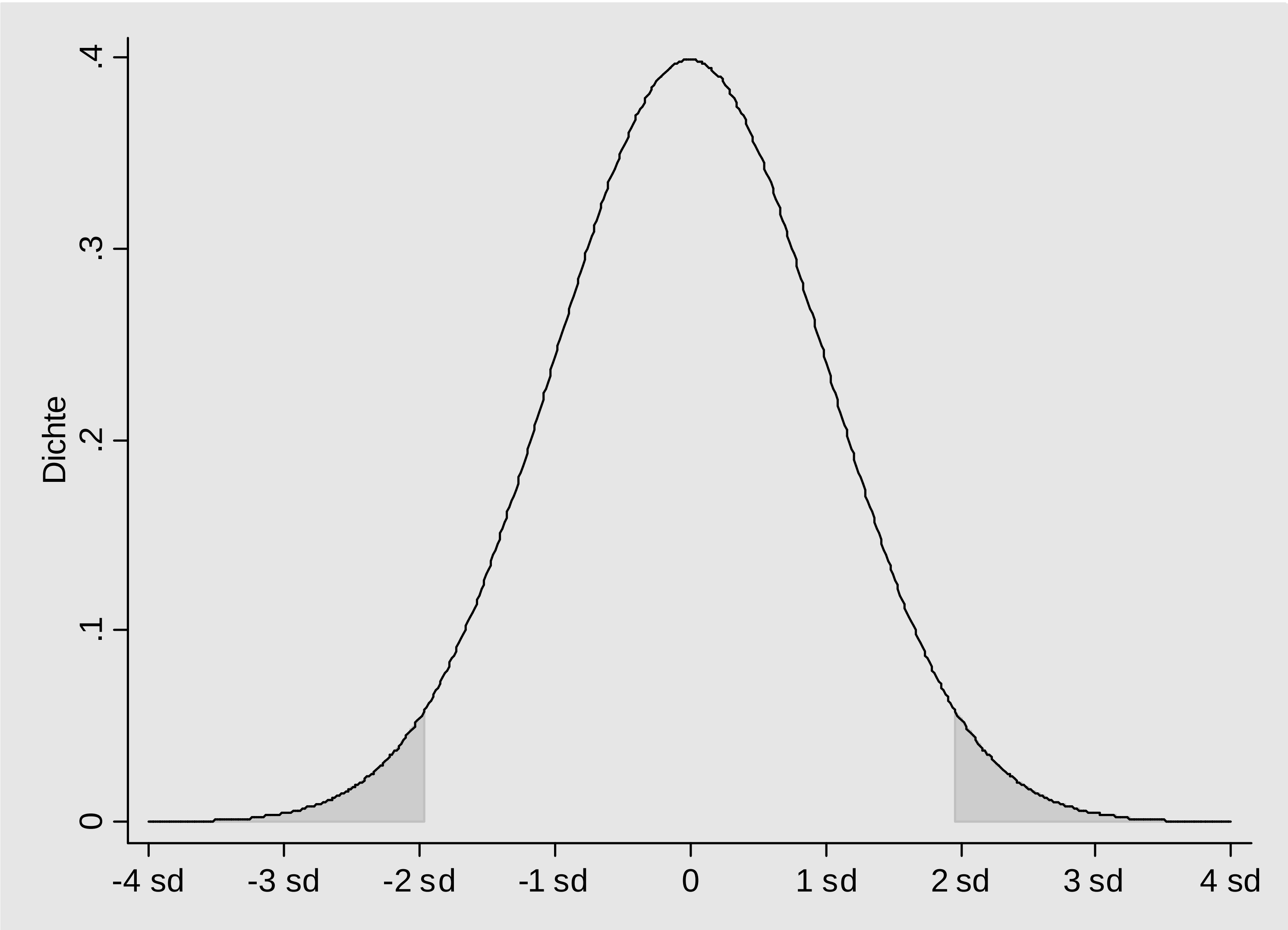
Stochastische Komponente

- graph twoway (scatter pwbcdu71 pkathv70) (lfit pwbcdu71 pkathv70), xline(85.1)
- Reale Werte weichen vom erwarteten Wert ab
- Ergänzung des Modells um eine Zufallsvariable: $y=a+b*x+e$
- Erfaßt *alle* unsystematischen (nicht modellierten) Einflüsse auf y

Stochastische Komponente

- Zufallsvariable $\cong$ Ziehung (mit Zurücklegen) aus einer Verteilung
- Jedes e wird gezogen aus einer Normalverteilung
  - mit beliebiger Streuung
  - und einem Mittelwert von null
  - Die Ziehungen sind unabhängig voneinander und nicht vom Wert von x beeinflußt
- Normalverteilung
  - enthält viele Werte nahe dem Mittelwert
  - wenige Werte, die stark vom Mittelwert abweichen
  - sinnvolles Modell für das additive Zusammenwirken von vielen zufälligen Einflüssen

Normalverteilung



Annahmen des Regressionsmodells (Auswahl)

- x hat linearen Einfluß auf erwarteten Wert von y [systematische Komponente]
- y-Werte streuen um erwarteten Wert [stochastische Komponente]
- Dieser unsystematische Einfluß wird durch eine Zufallsvariable e modelliert
- e wird manchmal auch als „Fehler“ bezeichnet

Annahmen (Auswahl)

- Für e gilt
  - e ist normalverteilt
  - Für jede Ausprägung der unabhängigen Variablen ist der erwartete Wert von e=0
  - zwischen e und den unabhängigen Variablen besteht keine Korrelation
  - Die Varianz von e ist für jede Ausprägung der unabhängigen Variablen konstant (Homoskedastizität)
  - Die Ausprägung von e bei einem Fall hat keinen Einfluß auf die Ausprägung von e bei irgendeinem anderen Fall (keine Autokorrelation des Fehlers)

Lineare Regression

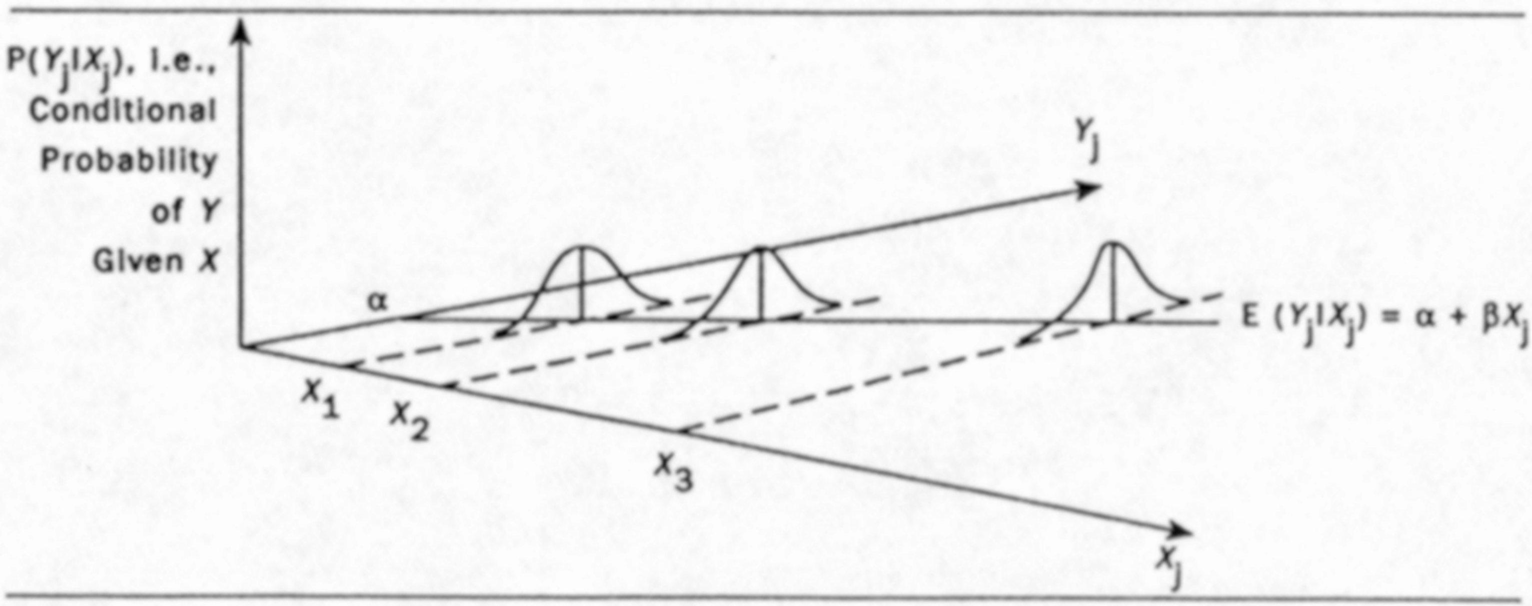


Figure 2.2. Regression Assumptions for a Bivariate Model

Bestimmung der Parameter

- Gesucht werden Parameter, die
  - die Summe der quadrierten Abweichungen zwischen beobachteten und erwarteten Werten minimieren
  - d.h. die Gerade optimal in die Punktewolke einpassen
  - Lineare Regression wird deshalb auch als OLS („Ordinary Least Squares“) bezeichnet
- Berechnung
  - von Hand: Abweichungsprodukte und Quadrate
  - Computer benutzt intern Matrix-Algebra
  - äquivalent, funktioniert auch mit mehreren unabhängigen Variablen
- Stata-Kommando: whelp reg

Lineare Regression

- Zur Veranschaulichung kann man sich vorstellen, daß die beobachteten y-Werte durch einen Daten Generierenden Prozeß (DGP) zustande kommen
- Ein y-Wert entsteht durch das additive Zusammenwirken eines gegebenen x-Wertes und eines zufälligen e-Wertes

Simulation

- use z:\daten\regsim,replace
- enthält 10.000 „Fälle“
- x: linkssteile Variable mit Mittelwert von 4 und sd von 2,8 (graph hist x)
- Normalverteilte Zufallsvariable:
  - gen e=20*invnorm(uniform())
  - summ e oder graph hist e,normal
- y erzeugen: gen y= 5 + 2.3*x+e

Simulation

- Parameter des Modells per Regression bestimmen: reg y x
- vorhergesagte Werte
  - predict yhat
  - summ yhat y
  - Vergleich mit Sum of Squares
- Residuen:
  - predict r, resid
  - summ r e

R-Quadrat

- Gibt an, wieviel Prozent der Gesamtvarianz von y auf die systematische Komponente des Modells zurückgehen
- Oft als „Maß der Modellgüte“ oder ähnliches bezeichnet
- Tatsächliche Bedeutung meist überschätzt
- Wert hängt von den Varianzen in der Stichprobe ab, kann deshalb nicht über Stichproben hinweg verglichen werden

R-Quadrat

- use z:\daten\regsim2,replace
- graph twoway (scatter y x) (lfit y x)
- reg y x
- graph twoway (scatter y x) (lfit y x) if x>14 & x<17
- reg y x if x>14 & x<17

R-Quadrat

- Mit Aggregatdaten läßt sich sehr leicht ein hohes R-Quadrat erreichen, da die nicht erklärte Varianz auf der Individualebene ignoriert wird
- Beispiel
  - use z:\daten\elecstudies,replace
  - reg ca relig
  - collapse ca relig,by(var003)
  - scatter ca relig
  - reg ca relig
  - Das Modell ist nicht besser (im Gegenteil), aber R^2 ist fast dreimal größer

Multiple lineare Regression

- Oft ist es plausibel anzunehmen, daß
  - mehrere unabhängige Variable parallel auf eine abhängige Variable wirken
  - diese Einflüsse additiv zusammenwirken
- $y = a + b_1 * x_1 + b_2 * x_2$
  z.B.
  SPD75=a + b1 * KATH70
  + b2 * ARB70

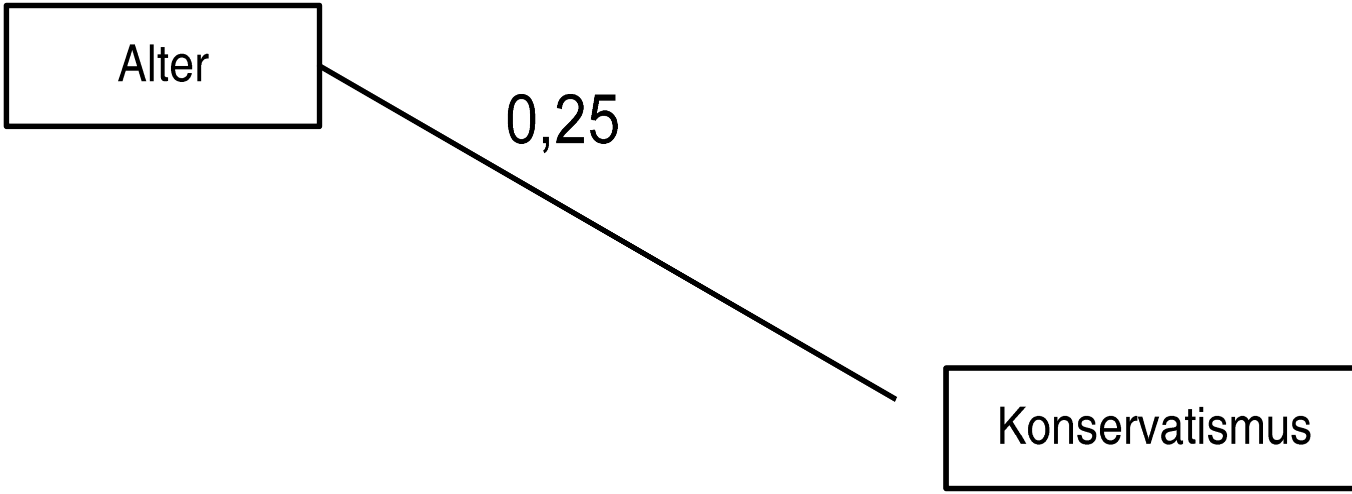
Multiple lineare Regression

- Koeffizienten (b_1, b_2 etc.) beschreiben, welche Veränderungen der abhängigen Variablen zu erwarten ist, wenn die entsprechende unabhängige Variable variiert und *alle anderen unabhängigen Variablen konstant gehalten werden*
- Konstante gibt den erwarteten Wert der abhängigen Variablen an, wenn alle unabhängigen Variablen gleich null sind (nicht immer sehr anschaulich; evtl. unabhängige Variablen zentrieren)

Warum multiple Regression?

- Soziale Phänomene lassen sich selten auf eine einzige Ursache zurückführen
- Bei bivariater Betrachtung können „Scheinkorrelationen“ auftreten
- use z:\daten\regsim3,replace
- reg kons alter
  - Mit jedem Lebensjahr nimmt der erwartete Konservatismuswert um 0.25 Punkte zu
  - je älter, desto konservativer
  - scatter kons alter

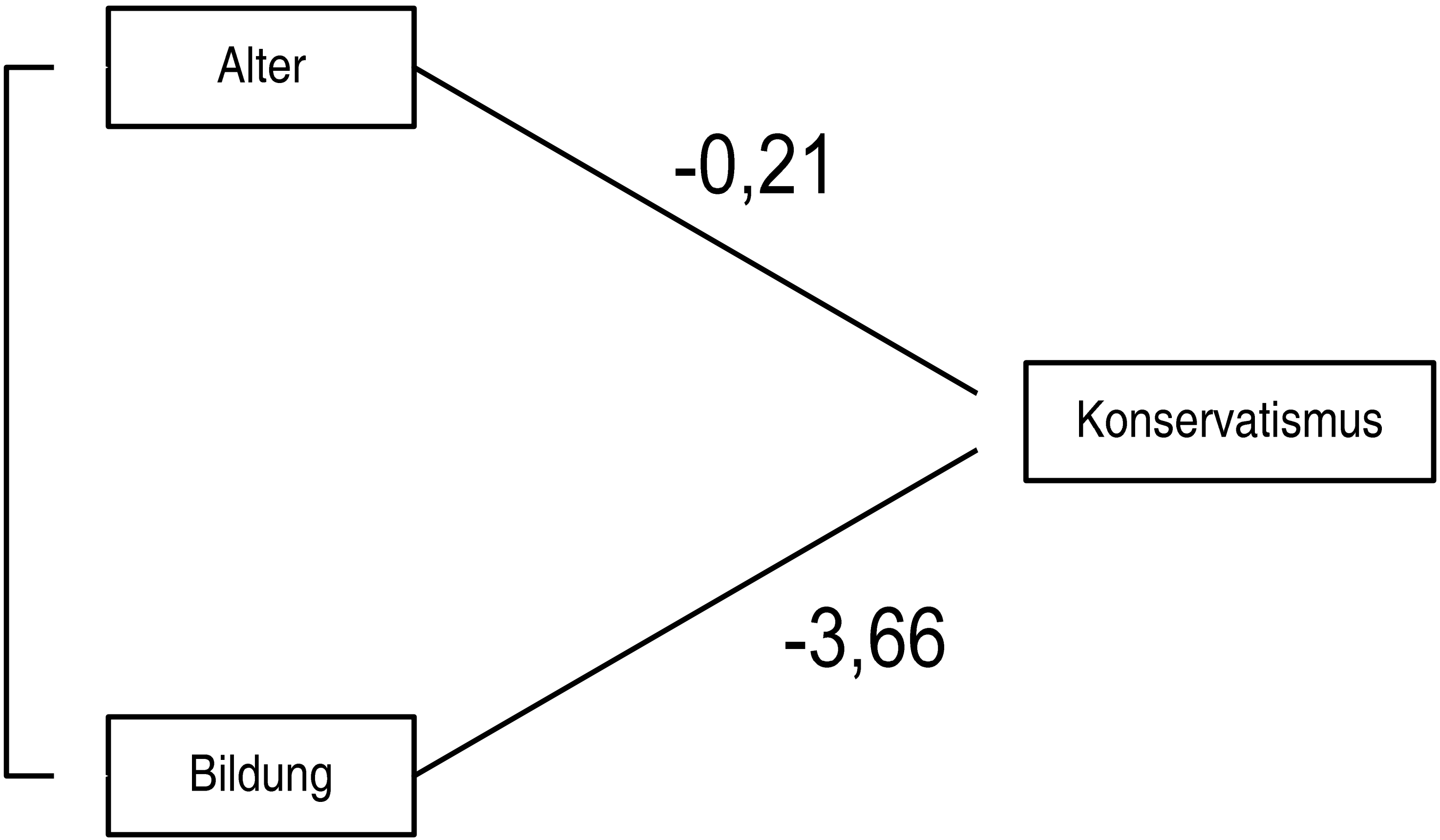
Modell 1



Warum multiple Regression?

- Zusammenhang zwischen Alter und Konservatismus kommt eventuell durch niedrigere Bildung der älteren Befragten zustande
  - scatter kons alter if bildung >10
  - reg kons alter if bildung >10
- Systematisch:
  - reg kons alter bildung
  - Simultane Schätzung unter wechselseitiger Kontrolle der Einflußfaktoren

Modell 2



Warum multiple Regression?

- Das bedeutet:
  - Bildung hat „in Wirklichkeit“ (d.h. bei Kontrolle des Alters) einen stärker negativen Einfluß auf Konservatismus als zunächst erkennbar
  - Für jede beliebige Altersgruppe nimmt der erwartete Konservatismus-Wert pro Punkt auf der Bildungsskala um 3,7 Punkte ab
  - Der Einfluß des Alters ist in „Wirklichkeit“ (bei Kontrolle der Bildung) schwach negativ!
  - Für jede beliebige Bildungsgruppe nimmt der erwartete Konservatismus-Wert pro Lebensjahr um 0,21 Punkte ab
  - Bei niedriggebildeten jungen sind die höchsten, bei hochgebildeten älteren Bürgern die niedrigsten Werte zu erwarten
  - Koeffizienten beziehen sich auf „natürliche“ Einheiten. Durch Standardisierung können Koeffizienten eventuell leichter vergleichbar gemacht werden (‚beta)

Warum multiple Regression?

- Parameter verzerrt, wenn relevante Variable nicht im Modell enthalten
- ➜Weitere Anwendungsvoraussetzungen
  - unabhängige Variablen müssen tatsächlich *additiv* zusammenwirken
  - alle unabhängigen Variablen, die einen nennenswerten systematischen Einfluß auf y haben, müssen im Modell enthalten sein
  - außer wenn die unabhängigen Variablen untereinander unkorreliert (orthogonal) sind

Orthogonalität

- Bei experimentellen Designs werden die unabhängigen Variablen von den Forschern *gesetzt*
- Bei zufälliger Aufteilung auf die Versuchsgruppen keine Korrelationen zwischen unabhängigen Variablen und keine Hintergrundvariablen, die deren Ausprägung beeinflussen

Orthogonalität

Horrorfilm

Alkohol	0	1	Total
0	10	10	20
1	10	10	20
Total	20	20	40

Kollinearität

- Bei Umfragedaten (ex-post-facto Design) in der Regel Korrelation zwischen unabhängigen Variablen
- Manche Kombinationen von Ausprägungen erstens nicht beobachtet, zweitens empirisch unplausibel/unmöglich (angelernter Arbeiter mit Hochschulabschluß)
- Lineare Beziehungen zwischen den unabhängigen Variablen werden als (Multi-) Kollinearität bezeichnet
- Typisch für Umfragedaten z.B. enge Beziehungen zwischen Schulabschluß, Beruf, Einkommen und politischen Einstellungen
- Moderate Kollinearität in Regressionsmodellen ist unproblematisch

Perfekte Kollinearität

- $x_1 = a + b * x_2 \mid R^2 = 1$
- Fehler
  - Intrinsische Beziehung zwischen zwei Variablen z.B. Geburtsjahr und Alter in Jahren (bei einer Querschnittsbefragung)
  - Bei Dummies: Referenzkategorie durch zusätzlichen Dummy repräsentiert
  - Interaktionseffekte
- Zahl der Fälle < Zahl der Variablen
- Konsequenz: Keine eindeutige Lösung für Regressionsgleichung: $y = a + 1 * x_1 + 0 * x_2 \Leftrightarrow y = a + 0 * x_1 + 1 * x_2$

Hohe Kollinearität

- $x_1 = a + b * x_2 \mid R^2 => 0.9$
- Interpretationsprobleme: Kann man sich überhaupt vorstellen, daß die übrigen Variablen konstant gehalten werden?
- Standardfehler werden sehr groß = Schätzung schwanken sehr stark über Stichproben hinweg
- Die Schätzungen für den Koeffizienten einer Variablen werden davon beeinflußt, welche anderen Variablen im Modell enthalten sind

Kollinearität

- Diagnose
  - Regression einer unabhängigen Variable auf eine (Kollinearität) oder *alle* (Multikollinearität) anderen unabhängigen Variablen
  - $1-R^2$=tolerance; Faustregel tol >0.1
  - 1/tol.=VIF; Faustregel VIF <10
  - In Stata:
    - corr alter bildung
    - vif (nach Regressionsbefehl)
- Maßnahmen
  - Alternative Kodierung (bei Interaktionseffekten)
  - Mehr Fälle, möglichst mit „ungewöhnlichen" Kombinationen der unabhängigen Variablen
  - (theoretisch begründeter) Ausschluß von unabhängigen Variablen
  - Zusammenfassung der hochkorrelierten Variablen zu einem Index/Faktor
  - Fortgeschrittene Methoden

Hausaufgabe

- Erzeugen Sie unter Verwendung von muster.do eine lauffähige Datei rpregression.do, die
  - den Datensatz z:\daten\rpstrukt.dta lädt
  - Ein Regressionsmodell für den Einfluß von Arbeiter- und Katholikenanteil (Volkszählung 1970) auf das Abschneiden der SPD 1975 rechnet
  - eine Grafik erzeugt, die – getrennt für Kreise mit niedrigem (<54%) und hohem(>=54%) Katholikenanteil einen Scatterplot und eine Schätzgerade für die Beziehung SPD 1975 vs. Arbeiter 1970 überlagert
- Schicken Sie die Datei bis zum 23.06. an do-files@politik.uni-mainz.de. Verwenden Sie das gewohnte Schema