Analysen politikwissenschaftlicher Datensätze mit Stata

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Sitzung 7: Komplexe Fehlerstrukturen, Interaktionen, Hypothesen über Koeffizienten, ordinale und nominale abhängige Variablen

Vorbereitung

- Bitte laden Sie den Datensatz z:\daten\pi-gesamt-77-01.dta

Regression des Anteils der Parteiidentifizierer auf die Zeit

- Welche Entwicklung erwarten Sie?
- graph twoway scatter piwest zeit
- reg piwest zeit
  - predict linpred
  - predict linres,resid
- Bedeutung der Koeffizienten?

Regression des Anteils der Parteiidentifizierer auf die Zeit

- Welche Probleme gibt es hier?
  - Wertebereich der Variablen beschränkt
  - Heteroskedastizität wg.
    - Höhere Varianz von Anteilswerten im mittleren Bereich
    - unterschiedlich große Stichproben
  - serielle Korrelation der Störvariable
  - Informationsverlust durch Aggregation

Regression des Anteils der Parteiidentifizierer auf die Zeit

- Konsequenzen?
  - unplausible Vorhersagen (hier unproblematisch: graph twoway linpred zeitp
  - zu kleine Standardfehler wg.
    - (Heteroskedastizität)
    - *Autokorrelation*
  - relativ uninteressante Modelle

Heteroskedastizität

- Plot der Residuen gegen die unabhängige Variable
  - graph twoway scatter linres zeitp *oder einfach*
  - rvpplot zeitp
- Hier sollte kein Muster zu sehen sein
  - summ linres if zeitp < 359,det
  - summ linres if zeitp >=359,det
- Varianz nicht konstant, formal: hettest

Autokorrelation

- Wert der *Störvariable* zum Zeitpunkt t nicht unabhängig von Werten bei t-1, t-2, …
- Zeigt sich in den *Residuen*:
  - auf große positive Residuen folgen tendenziell wieder große positive R.
  - auf große negative Residuen folgen tendenziell wieder große negative
  - sinus-artiges Muster beim Plot gegen die Zeit
  - Korrelation der Residuen zum Zeitpunkt t mit Zeitpunkt t-1
  - Durbin-Watson-Test oder ähnliches

Autokorrelation

- Daten müssen als Zeitreihe deklariert werden: tsset zeitpunkt
- Autokorrelation der Residuen:
  - gen laglinres=L.linres
  - list linres laglinres
  - corr linres laglinres
- dwstat oder durbina
- Wie geht es besser?
  - prais piwest zeitp, robust
- In diesem speziellen Fall kaum Unterschiede

Inhaltliches Problem

- Immenser Informationsverlust durch Aggregation
  - wg. Fallzahlen
  - wg. Wegfall der Verknüpfung von unabhängiger und abhängiger Variable auf Individualebene
- Wenn möglich, disaggregierte Daten analysieren, z.B. mit Logit

Alternativen

- Als Vergleichsbasis Zeitreihenregression ab April 1991 (doedit z:\preg.do)
  - drop if zeitpunkt <374
  - prais piwest zeitpunkt,robust
  - lincom _b[zeitpunkt]*12
  - lineare Rückgang von 0,5 Prozentpunkten, K-Intervall -0.6 bis
    -0.3
  - predict ppred
  - predict error,stdp
  - gen upper=ppred+1.96*error
  - gen lower=ppred-1.96*error
  - graph twoway (line ppred zeitpunkt) (line upper zeitpunkt) (line lower zeitpunkt)
  - graph save "u:\StataSeminar\ppred.gph", replace

PI über die Zeit: Logit

- use z:\daten\pi-disagg,replace
- PI, Zeitpunkt etc. auf Individualebene
- logit pi zeitpunkt if ost==0
- Wie groß ist die Zeitspanne: summ zeitp
- Vorhergesagte Werte für Variation über die Zeit

PI über die Zeit: Logit

- Vorhergesagte Werte über die Zeit
  - Weg über predict ist möglich aber unpraktisch wg. großer Fallzahl (evtl. Daten aggregieren)
  - prgen zeitpunkt,gen(logpred) from(375) to(503) funktioniert, liefert aber keine Standardfehler für die Prognose
  - Am besten mit einer Schleife: doedit z:\logpred.do
- Schätzung im Prinzip sehr ähnlich (linearer Bereich der S-Kurve), Konfidenzband deutlich schmäler
- Evtl. immer noch Korrelation der Fehlerterme innerhalb einer Umfrage: logit pi zeitp if ost==0, robust cluster(zeitpunkt)

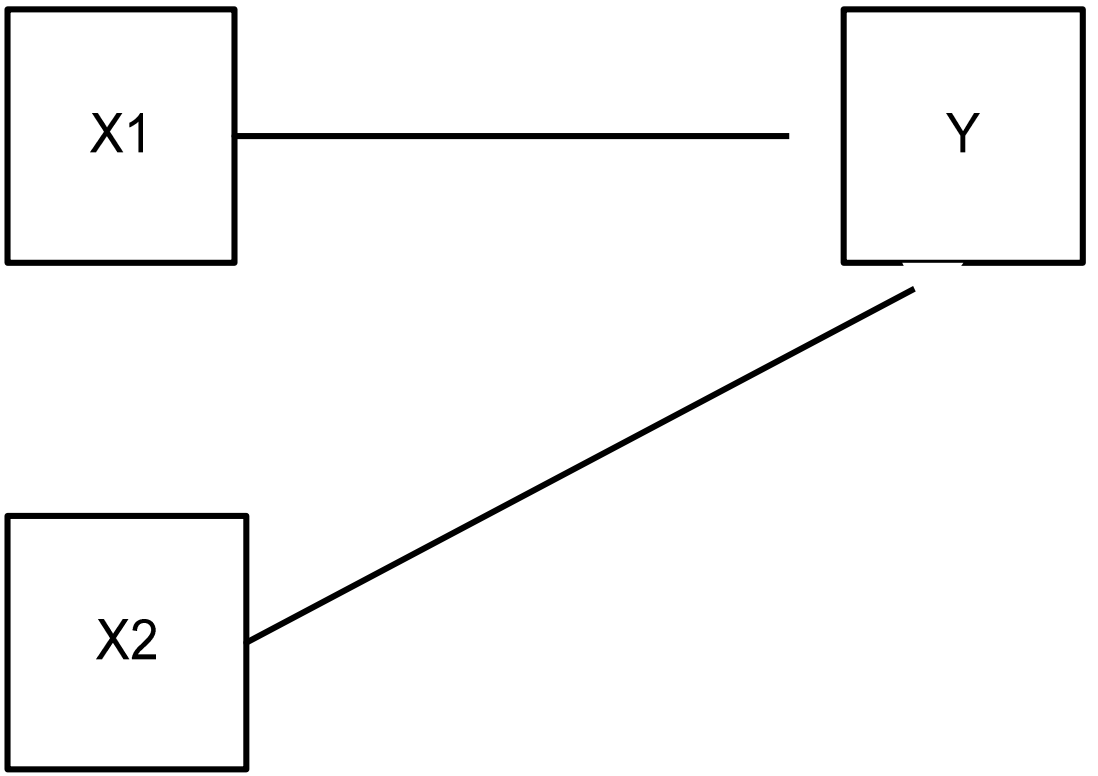
PI über die Zeit: Interaktionen

- PI dürfte sich in alten und neuen Bundesländern unterschiedlich entwickelt haben. Wie?
- Neue Länder...
  - niedrigeres Startniveau
  - anderer Trend
- Lösung für diese und komplexere Fragen: multiplikative Interaktionsterme

Interaktionen

- Lineares Modell geht grundsätzlich davon aus, daß unabhängige Variablen additiv (unabhängig voneinander) zusammenwirken
- y wird in Abhängigkeit von X1, X2 … modelliert
- X1, X2 etc. können untereinander korreliert sein, solange Zusammenhang nicht perfekt
- Koeffizient x1 beschreibt erwartete Veränderung in y, wenn X1 um eine Einheit zunimmt und alle anderen X konstant gehalten werden
- Niveau der anderen X spielt dabei keine Rolle

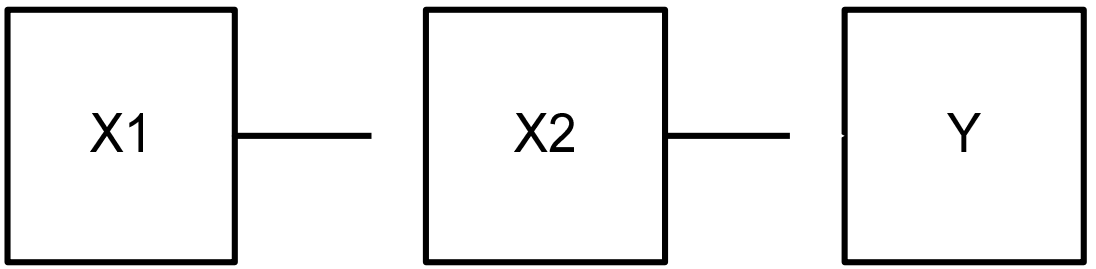
„Normales“ multivariates lineares Regressionsmodell



Interaktion

- Oft ist es plausibler, anzunehmen daß die Wirkung einer X-Variablen durch eine zweite Variable „moderiert“ werden
- Das bedeutet: Die Wirkung von X1 hängt vom Niveau von X2 ab
- Man sagt auch X1 hat eine „konditionale“ Wirkung auf y

Interaktion



Interaktion: Modellierung

- Interaktionen werden durch „Produktterme“ modelliert: man multipliziert die beiden X-Variablen miteinander und nimmt diese neue Variable X1*X2 ins Modell auf
- Dadurch ändert sich die Interpretation der Koeffizienten!
- Haupteffekte müssen jetzt *konditional* interpretiert werden:
  - x1: Effekt der Variable X1 wenn X2=0
  - x1*x2: *Veränderung des Koeffizienten x1*, wenn X2 um eine Einheit zunimmt
- D.h. Wirkung von X1 hängt vom Niveau von X2 ab; Berechnung erwarteter Werte und Interpretation der Koeffizienten etwas komplizierter

Interaktion: Modellierung

- Welche Variable theoretisch welche moderiert, spielt für Modellierung keine Rolle! ➜
  - x2: Effekt der Variable X2 wenn X1=0
  - x1*x2: *Veränderung des Koeffizienten x2*, wenn X1 um eine Einheit zunimmt
  - bzw. x1: Effekt der Variable X1, wenn X2=0 ...
- Signifikanz von Interaktionstermen hängt von der Modellierung ab (Zentrierung der Haupteffekte)
- Zusätzliche Komplikation bei Logit
  - Interaktion wird auf der linear-additiven Ebene der Haupteffekte modelliert
  - auf der Ebene der Wahrscheinlichkeiten sowieso nicht linear-additiv

Interaktion

- Im konkreten Beispiel
  - gehen wir davon aus, daß es zu Beginn (Frühjahr 1991) einen Unterschied im Niveau der PI zwischen alten und neuen Ländern gibt
  - daß es in beiden Landesteilen unterschiedliche Trends gibt
  - daß die Unterschiede infolgedessen wachsen, schrumpfen oder konstant bleiben können

Modellierung

- Zeitvariable, die am Anfang des Zeitraums den Wert 0 hat: gen nullzeit=zeitpunkt-375
- Regionalvariable schon vorhanden (ost)
- Interaktion gen ostXnullzeit=ost*nullzeit (alternativ: desmat ost*@nullzeit)
- logit pi ost nullzeit ostXnullzeit

Interpretation

- Konstante:
  - Logit, wenn alle Variablen = 0 (geschätzter Wert alte Länder für Frühjahr 1991)
- Ost:
  - Ost-Effekt, wenn Haupteffekt Zeit und Interaktion Ost*Zeit =0
  - (geschätzte Abweichung vom Wert der alten Länder für Frühjahr 1991)
  - *in diesem Monat* hochsignifikant
- ostXnullzeit:
  - Wert, um den sich Ost-West-Unterschied jeden Monat verringert

Interpretation 2

- nullzeit:
  - Monatliche Veränderung, wenn ost=0 und Interaktion ost*zeit=0
  - Trend für West
  - hochsignifikant
- ostXnullzeit:
  - Betrag, um den der Ost-Trend vom West-Trend abweicht (hochsignifikant)
- Ost-Trend:
  - Summe aus nullzeit+ostXnullzeit

Interpretation 3

- Auch hier am besten wieder geschätzte Wahrscheinlichkeiten plotten
- drop logpred-plotzeit
- Vorgehensweise wie vorher, aber mit Interaktion und zwei unabhängigen doedit z:\logpred2

Hypothesen über Koeffizienten

- Ist der Anstieg der PI im Osten signifikant?
  - entweder umkodieren
    - gen west=!ost
    - gen westXnullzeit=west*nullzeit
    - logit pi nullzeit west westXnullzeit
  - oder
    - lincom _b[nullzeit]+_b[ostXnullzeit] bzw.
    - lincom nullzeit+ostXnullzeit

Hypothesen über Koeffizienten

- Ist der Ost-West-Unterschied auch noch im Dezember 2001 (nullzeit==128) signifikant?
  - lincom ost+128*ostXnullzeit
- Wann wäre dem Modell zufolge der Ost-West-Unterschied komplett verschwunden?
  - display _b[ost]/_b[ostXnullzeit]
  - gen dummy=375+311
  - format %tm dummy
  - list dummy in 1
  - drop *lower *upper dummy
  - prvalue,x(ost 0 nullzeit 311 ostXnullzeit 0)
  - prvalue,x(ost 1 nullzeit 311 ostXnullzeit 311)

Hypothesen über Koeffizienten

- test, testparm, lincom, lrtest und testnl erlauben das Testen komplexer Hypothesen über Koeffizienten
- z.B. läßt sich prüfen, ob nullzeit und ostXnullzeit signifikant voneinander verschieden sind oder ob ost von -0.4 signifikant verschieden ist
  - test nullzeit=ostXnullzeit
  - test ost=-0.4
  - beides im Kontext dieses Modells inhaltlich nicht notwendig; häufig aber sinnvolle Anwendungen
  - whelp test

Modellgüte

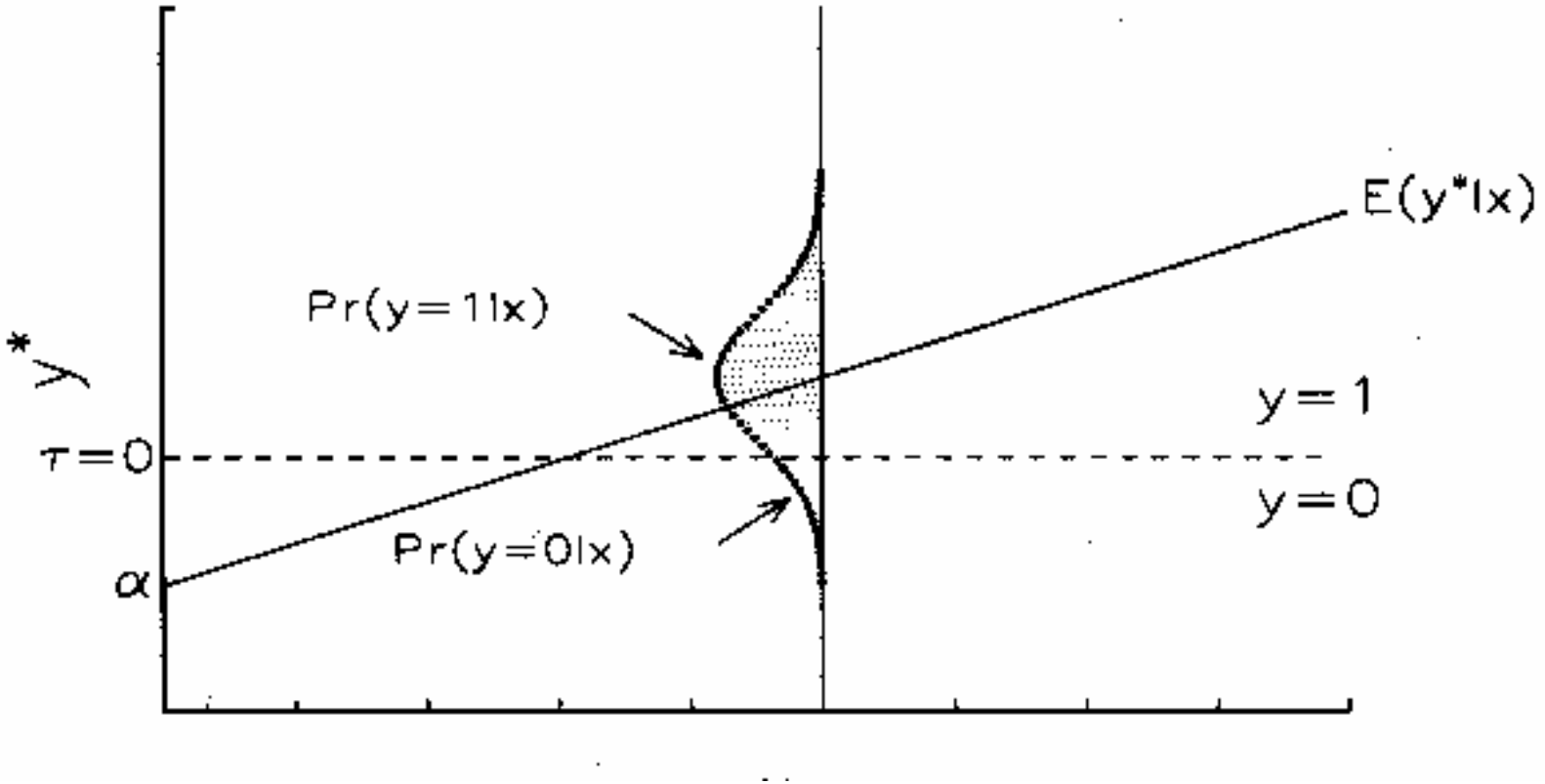
- Wie beim linearen Modell können diverse Maße der „Modellgüte“ berechnet werden
- ~ äquivalent zu R^2: Pseudo R^2 (verschiedene Varianten)
  - meist erheblich niedriger als R^2
  - besonders wenn abhängige Variable sehr schief verteilt
  - modell- und datenabhängig
  - Suche nach möglichst großem R^2 führt nicht notwendigerweise zu besseren Modellen
- Vergleich verschiedener Modelle
  - fitstat,save
  - logit pi ost nullzeit
  - fitstat,diff

Ordinale abhängige Variablen

- use Z:\daten\allbus1980-2000.dta, replace
- Abhängige Variable in der Politikwissenschaft oft ordinal
- Meist wie metrisch behandelt
  - Abstände zwischen Kategorien nicht gleich
  - Vorhersagen außerhalb des Bereichs
- Schöner: ordinales Logit-Modell

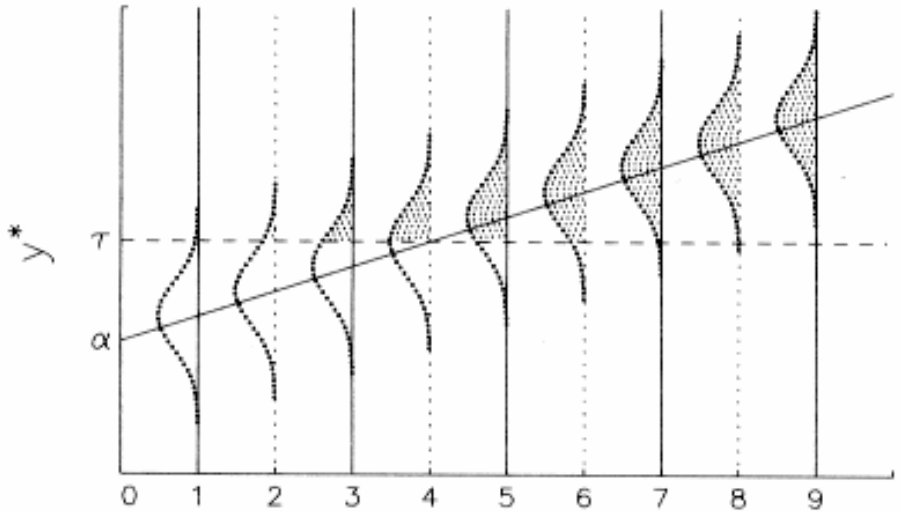
ordinale abhängige Variable

- Eine Interpretation des binären Logit-Modells:
  - hinter dichotomer beobachteter Variable (REP-Wahl ja/nein) steht latente kontinuierliche Variable im Sinne einer „Tendenz“
  - diese Variable ist der Logit
  - wenn Logit <=0 ➜ 0
  - wenn Logit >0 ➜ 1
  - bei gegebenem x-Wert resultiert aus der Streuung der Logits um ihren erwarteten Wert die geschätzte Wahrscheinlichhkeit von y=1

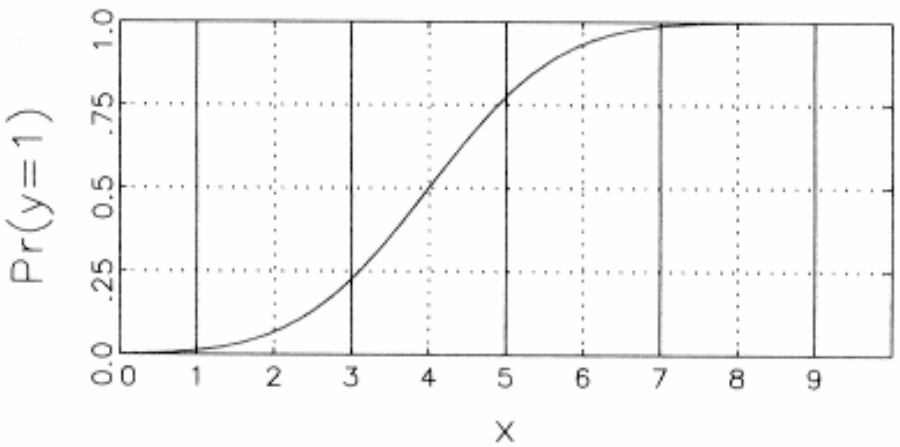


Quelle: Long/Freese 2001

Panel A: Plot of y*



Panel B: Plot of Pr(y=1|x)

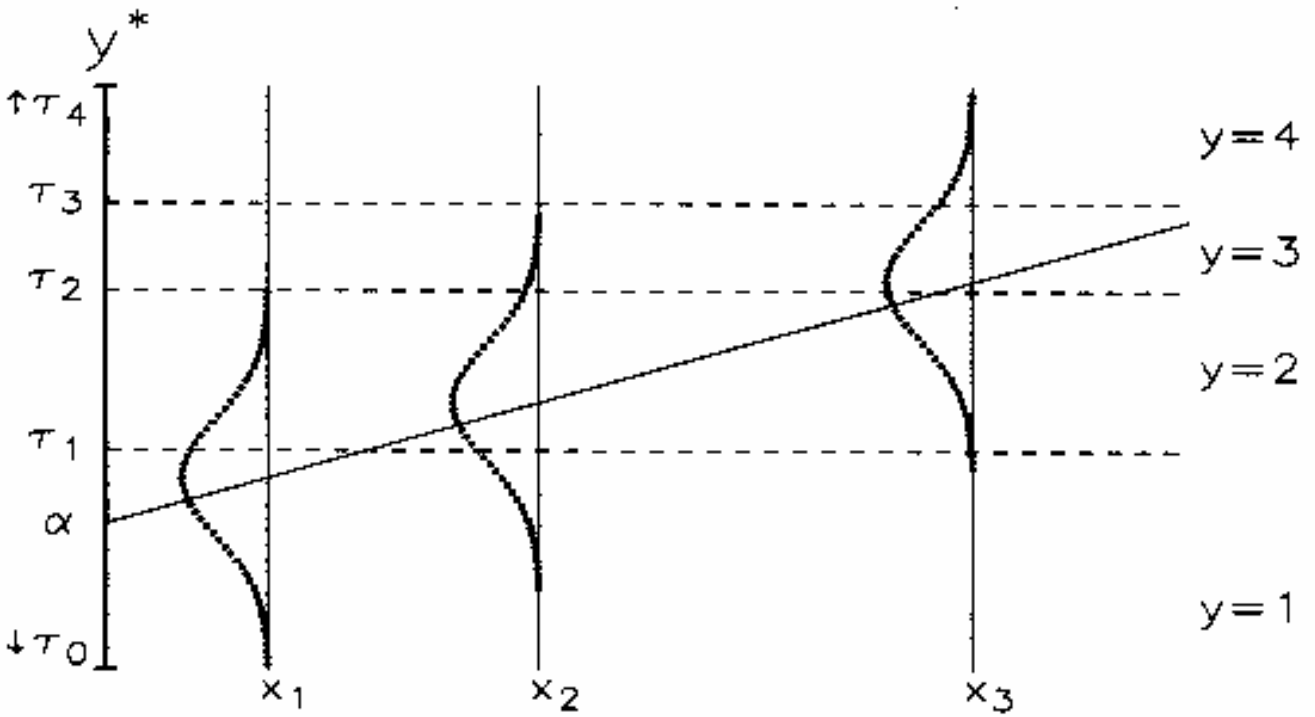


Quelle: Long/Freese 2001

ordinale abhängige Variable

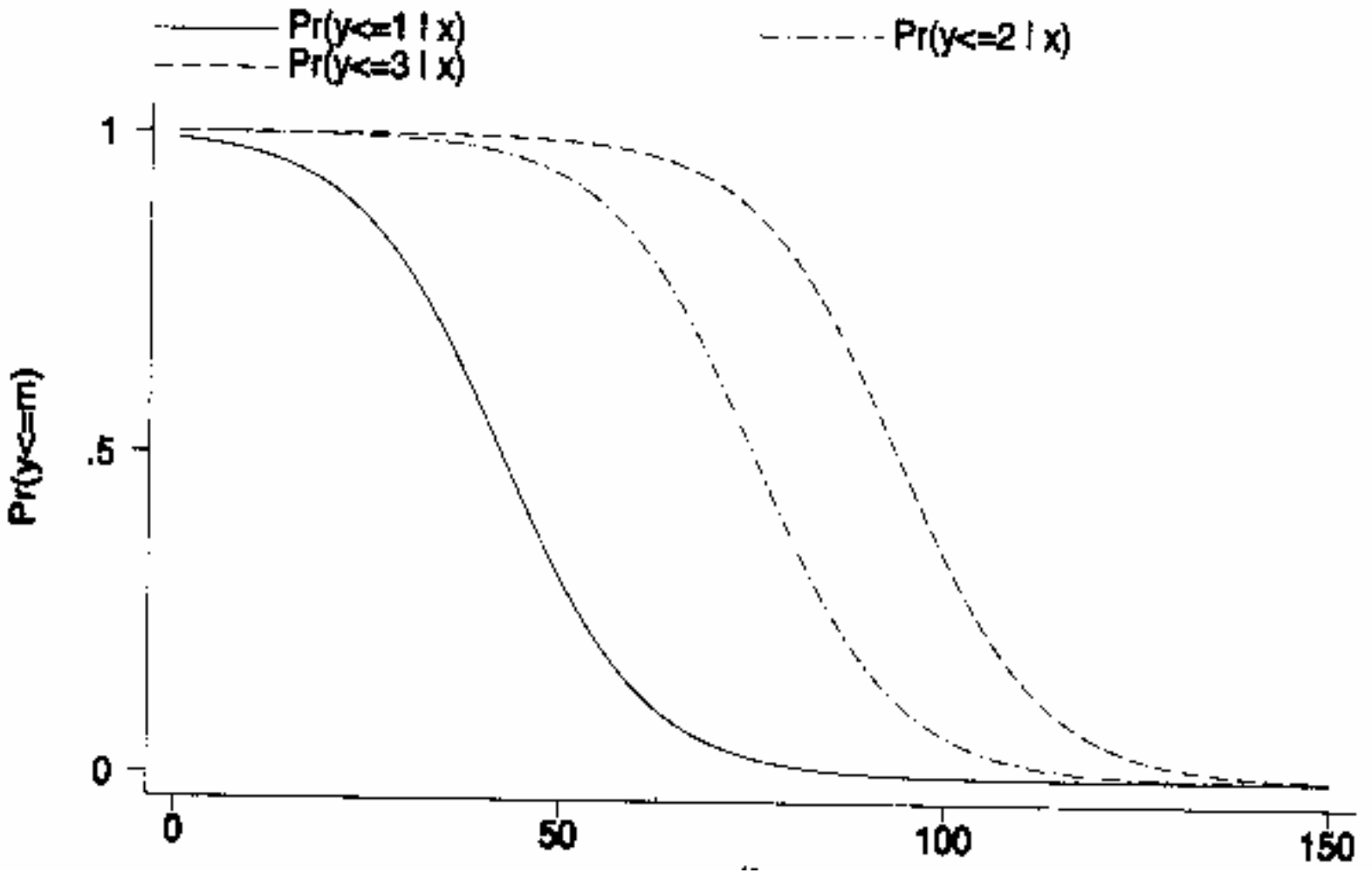
- Generalisierung auf ordinales Modell
  - Logit repräsentiert latente „Tendenz“, einem Item zuzustimmen
  - Den Kategorien der abhängigen Variablen entsprechen „cutpoints“, deren Abstand voneinander nicht identisch sein muß
  - Wenn der unbeobachtete Logit zwischen zwei cutpoints liegt, wird entsprechende Kategorie gewählt
  - Nicht alle Fälle mit gleichem x-Wert haben gleichen Logit (probabilistisches Modell)
  - deshalb wählen nicht alle die gleiche Kategorie
  - Wahrscheinlichkeit einer bestimmten Kategorie hängt von x-Wert ab
  - Modell hat keine Konstante; binäres Logit Modell entspricht ordinalem Logit-Modell mit einem cut-point
    - logit pi alter
    - ologit pi alter

Quelle: Long/Freese 2001



ordinale abhängige Variable

- tab interesse
- ologit interesse schuljahre
  - Die Zahl der Schuljahre hat einen signifikant positiven Effekt auf das politische Interesse
  - Es gibt nur einen Parameter für Schuljahre
  - ➜ die (kumulierte) Wahrscheinlichkeitskurve hat für alle kumulierten Kategorien die gleiche Steigung (Form), ist aber auf der x-Achse mehr oder minder verschoben (parallele Regression)


Pr(y<=1 | x)
Pr(y<=2 | x)
Pr(y<=3 | x)
1
.5
0
Pr(y<=m)
0
50
100
150

Interpretation

- prtab schuljahr
- prgen schuljahr,from(7) to(13) gen(intpred)
- set scheme s1mono
- graph twoway line intpredp* intpredx
- graph twoway line intpreds* intpredx
- Die Annahme paralleler Regression ist leider wie so häufig verletzt: brant
- In der Praxis machen ordinale Logit- / Probit-Modelle oft Probleme

ordinale unabhängige Variable

- Eigentlich ist Bildung auch nur ordinal (wenn überhaupt)
- Zerlegung in Dummies (drei Kategorien ➜ 2 Dummies)
- desmat bildung
- ologit interesse _x*
  - Ist der Effekt beider Bildungsdummies gleich stark? test _x_1=_x_2
  - Ist der Effekt höherer Bildung wirklich nur 2,5 mal stärker als der Effekt mittlerer Bildung? test 2.5*_x_1 = _x_2
  - um wieviel stärker ist der Effekt höherer Bildung?
    - display _b[_x_2]/_b[_x_1]
    - nlcom _b[_x_2]/_b[_x_1]

nominale abhängige Variable

- Viele interessante Variablen nur nominal (z.B. Wahlabsicht)
- Häufig analysiert mit multinomialem Logit-Modell
- Logik des multinomialen Logit-Modells
  - für eine abh. Variable mit z.B. drei Ausprägungen (z.B. SPD/CDU/AND) können drei dichotome Logit-Modelle geschätzt werden
    - SPD vs. CDU
    - CDU vs. AND
    - SPD vs. AND
  - diese drei Modelle sind redundant da z.B. $b_{1\text{SPD vs. CDU}} - b_{1\text{CDU vs. AND}} = b_{1\text{SPD vs. AND}}$
  - Im multinomialen Modell werden alle nicht-redundanten Koeffizienten *simultan* geschätzt
  - Das Modell beinhaltet deshalb mehrere Gleichungen
  - Eine Kategorie der abhängigen Variablen dient als Vergleichsgruppe
  - Das ganze kann sehr verwirrend werden

Ein Beispiel: Wahlverhalten

- Wie hängt das Wahlverhalten mit dem Alter zusammen?
  - tab wabs
  - mlogit wabs,base(3)
  - mlogit wabs alter,base(3)
- Ist der Effekt des Alters für CDU vs. andere und SPD vs. andere identisch?
  - test[1]alter=[2]alter

Ein Beispiel

- Was bedeuten die Effekte inhaltlich?
  - prgen alter,from(18) to (90) gen(pwabs)
  - graph twoway line pwabsp* pwabsx
- Welchen Einfluß hat die Konfession?
  - tab konf
  - desmat konf,base(3)
  - d _*
  - mlogit wabs _x*,base(3)
- Ist der Unterschied zwischen den Koeffizienten für prot/kath im Falle des SPD-Wahl (vs. andere) signifikant?
  - test [2]_x_1=[2]_x_2
  - lincom [2]_x_1-[2]_x_2
- Geschätzte Wahrscheinlichkeiten z.B. so:
  - predict wahlcdu wahlspd wahlandere
  - tabstat wahl*,by(konf)

Einen haben wir noch…

- Wie wirken Alter und Konfession gemeinsam auf das Wahlverhalten?
- Gibt es Interaktionen?
  - desmat @alter*konf,base(3)
  - mlogit wabs _*,base(3)
  - destest alter.konf
  - test [2]_x_4 [2]_x_5
  - Likelihood-Ratio-Test:
    - lrtest,saving(0)
    - constraint 1 [2]_x_4=0
    - constraint 2 [2]_x_5=0
    - mlogit wabs _*,base(3) cons(1-2)
    - lrtest

Geschätzte Wahrscheinlichkeiten

- Am einfachsten so:
  - predict wahlcdu wahlspd wahlandere
  - lab var wahlcdu "Pr(CDU)„
  - lab var wahlspd "Pr(SPD)„
  - lab var wahlandere "Pr(AND)„
  - lab var konf "Konfession"
  - sort alter
  - graph twoway (line wahl* alter),by(konf)

Hausaufgabe

- Schreiben Sie eine Datei ordinalnominal.do, die
  - den Allbus-Datensatz lädt
  - ein *ordinales* Logit-Modell für den Einfluß der kategorialen Bildungsvariable auf den Inglehartindex (Variable ingle) schätzt
  - eine Tabelle für die geschätzten Anteile der drei Wertetypen in den drei Kategorien des Index ausgibt
  - ein multinominales Logit-Modell für den Einfluß der kategorialen Bildungsvariable auf den Inglehartindex schätzt
  - auch hierfür eine entsprechende Tabelle ausgibt
  - Fügen Sie am Ende einen Kommentar ein, in dem Sie Ihre Befunde interpretieren (zwei bis drei Sätze genügen)
  - Wie üblich bis nächsten Mittwoch im gewohnten Format an die gewohnte Adresse