Analysen politikwissenschaftlicher Datensätze mit Stata

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Sitzung 6: Inferenzstatistik und Regression

Vorbereitung

- Bitte laden Sie den Datensatz z:\daten\infsim.dta

Inferenzstatistik

- Grundproblem der Inferenzstatistik:
  - Wie kann man von einer Stichprobe einen gültigen Schluß auf die Grundgesamtheit ziehen
  - Bzw.: Wie groß sind die Fehler, die man dabei macht
- Stichprobenparameter: $\bar{x}, s, s^2$

- Parameter der Grundgesamtheit: m, s, s^2
- Weitere Parameter: Anteilswerte, Zusammenhangsmaße, Regressionskoeffizienten etc.

Punktschätzungen

- Sind bereits bekannt:
- Punktschätzung für : $\hat{m} \quad \bar{x}$
- Punktschätzung für : $\hat{s}^2 \quad s^2 \times \frac{n}{n-1}$
- Punktschätzung für : $\hat{s} \quad \sqrt{\hat{s}^2}$
- Problem:
  - Schätzung beruht auf (Zufalls-)Stichproben
  - *Stichprobenwerte sind Ausprägungen einer Zufallsvariable*
  - Deshalb entsprechen sie praktisch niemals exakt dem Wert in der Grundgesamtheit
  - Aber: Schwankungen um Erwartungswert buchstäblich berechenbar

Zufallsvariablen

- Im Einzelfall weiß man nicht, welchen Wert die Variable annimmt
- Aber: Ausprägungen von Zufallsvariablen sind nicht willkürlich, sondern höchst regelmäßig verteilt
- Die *Verteilung* der Werte einer Zufallsvariablen ist in der Regel bekannt
- Zufallsvariablen (und ihre Verteilungen) können diskret oder stetig sein

Grenzwertsätze

- ZGWS: Verteilung von Stichprobenmittelwerten um wahren Mittelwert in der GG folgt bei hinreichenden großen Fallzahlen näherungsweise einer Normalverteilung
- ➜ Mittelwerte in der Nähe des wahren Mittelwertes treten relativ häufig auf
- Streuung dieser Normalverteilung („Standardfehler des Mittelwertes“) hängt ab
  - vom Quadrat des Stichprobenumfang
  - von der Streuung des interessierenden Merkmals, die ebenfalls geschätzt werden muß

Konfidenzintervalle

- Bei wiederholter Stichprobenziehung 95% der Stichprobenmittelwerte in einem Intervall von ±1,96 Standardabweichungen/Standardfehler vom wahren Mittelwert der GG
- ➜ Ein Intervall von ±1,96 Standardfehlern um den Stichprobenmittelwert schließt in 95% aller Fälle den wahren Mittelwert ein
- ➜ aufgrund der Stichprobenwerte kann ein „Konfidenzintervall“ berechnet werden, das mit einer Wahrscheinlichkeit von 95% den wahren Wert beinhaltet (Intervallschätzung)

Simulation

- Die Daten im Speicher repräsentieren eine hypothetische Grundgesamtheit von 100.000 Bürgern
- hist alter
- summ alter, det
- Wahren Mittelwert notieren
- Wie groß ist der zu erwartende Standardfehler für Stichproben mit n=250?
- display sqrt(r(Var)/250)
- Entspricht der Standardabweichung einer Normalverteilung, mit der Stichprobenmittelwerte um den wahren Mittelwert in der Grundgesamtheit streuen

Simulation

- Simulation von 100 Stichprobenziehungen (bitte in Zweiergruppen)
- doedit z:\infexperiment.do
  - Am Anfang wird eine Ergebnisdatei im Heimatverzeichnis vorbereitet
  - In der Schleife werden 100 unabhängige Stichproben gezogen
  - Der Mittelwert wird jeweils in die Ergebnisdatei geschrieben
  - Ausführen mit ctrl-d

Simulation

- Ergebnisdatei mit use ergebnisse,replace öffnen
- Jeder Fall entspricht einem Stichprobenmittelwert
- hist erg,norm bin(15) ➜ Mittelwerte näherungsweise normalverteilt
- summ erg, det
  - „Mittelwert der Mittelwerte“ (Erwartungswert der Verteilung) entspricht fast exakt dem wahren Mittelwert
  - Standardabweichung (Streuung um Erwartungswert) entspricht ungefähr dem errechneten Standardfehler
  - Grenzwertsatz funktioniert auch heute morgen
  - Ermöglicht Konfidenzintervalle und Hypothesentests

Konfidenzintervalle in Stata

- Konfidenzintervall für den Mittelwert
- use z:\daten\allbus1980-2000,replace
- summ v929
- ci v929
- Konfidenzintervall wg. großer Fallzahl sehr eng, sehr sichere Schätzung möglich

Konfidenzintervalle in Stata

- Auch für Anteilswerte können Konfidenzintervalle berechnet werden, wenn die Variable als Dummy (0/1) kodiert ist
- Geschlecht der Interviewer
  - numlabel v928,add
  - tab v928
  - recode v928 (1=1 m) (2=0 w),gen(male)
  - recode v928 (1=0 m) (2=1 w),gen(female)
  - ci male female,bin

Konfidenzintervalle in Stata

- Bei Variablen mit mehr Ausprägungen erzeugt man für jede interessante Ausprägung einen Dummy
- Dieser Prozeß läßt sich mit dem Kommando xi automatisieren
- Komfortabler ist das Zusatzkommando desmat (von mir installiert)

Konfidenzintervalle in Stata

- numlabel v24,add
- tab v24
- desmat v24,full erzeugt zehn Dummies (einen für jede Kategorie)
- Namen der Dummies beginnen mit _x_
- ci _x_*,bin

Konfidenzintervalle in Stata

- „Immediate“ Variante ermöglicht Berechnung von Konfidenzintervallen auf Grund publizierter Ergebnisse (Tabellen)
- Politbarometer: 1007 Befragte, SPD bei 25,2 Prozent, FDP bei 6,7 Prozent
  - cii 1007 round(1007*0.252)
  - cii 1007 round(1007*0.067)

Hypothesentests

- Befassen sich mit der Frage, wie Hypothesen über die Grundgesamtheit mit einer Stichprobe überprüft werden können
- Basieren auf gleicher Logik wie Konfidenzintervalle
  - Stichprobenwerte sind Realisierungen einer Zufallsvariablen
  - Die Verteilung dieser Variable ist bekannt

Hypothesentests

- Ausgangspunkt: Nullhypothese, bestimmter Parameter (z.B. Zusammenhangsmaß) sei in der GG gleich null
- Alternativhypothese: Parameter in GG *ungleich* null
- Berechnung der Verteilung von Stichprobenwerten um Wert in der GG unter der Annahme, daß dieser tatsächlich gleich null ist

Hypothesentests

- Vergleich des konkreten Stichprobenwertes mit theoretischer Verteilung
- ➜ Berechnung der Wahrscheinlichkeit, daß Stichprobenwert diesen oder höheren Betrag erreicht, wenn Parameter in GG tatsächlich null
- entspricht der Wahrscheinlichkeit, aufgrund des Stichprobenergebnisses die Nullhypothese zu unrecht aufzugeben (Irrtumswahrscheinlichkeit, a)
- Bei sehr kleinen Irrtumswahrscheinlichkeiten (<1% oder <5%) gelten die Ergebnisse als „signifikant“ (mit großer Sicherheit von 0 verschieden)

Vorsicht

- Nullhypothese in aller Regel unplausibel, da immer irgendwelche Zusammenhänge/Unterschiede bestehen
  - mit sehr großen Stichproben erweisen sich auch triviale Zusammenhänge als signifikant
  - selbst sehr starke Zusammenhänge können mit zu kleinen Stichproben nicht sicher erkannt werden („Power“)
- Interessanter als die Frage nach der Signifikanz ist in der Politikwissenschaft oft
  - die Frage nach der Bedeutsamkeit eines Parameters
  - die Frage nach dem Konfidenzintervall
- Prinzipiell ist es möglich, die Nullhypothese inhaltlich zu modifizieren
  - inhaltlich interessant seien nur Parameter in der GG >0,1
  - „Wie wahrscheinlich ist das Ergebnis, wenn Parameter in der GG <=0,1?“
  - Wird in der Praxis leider nur sehr selten gemacht

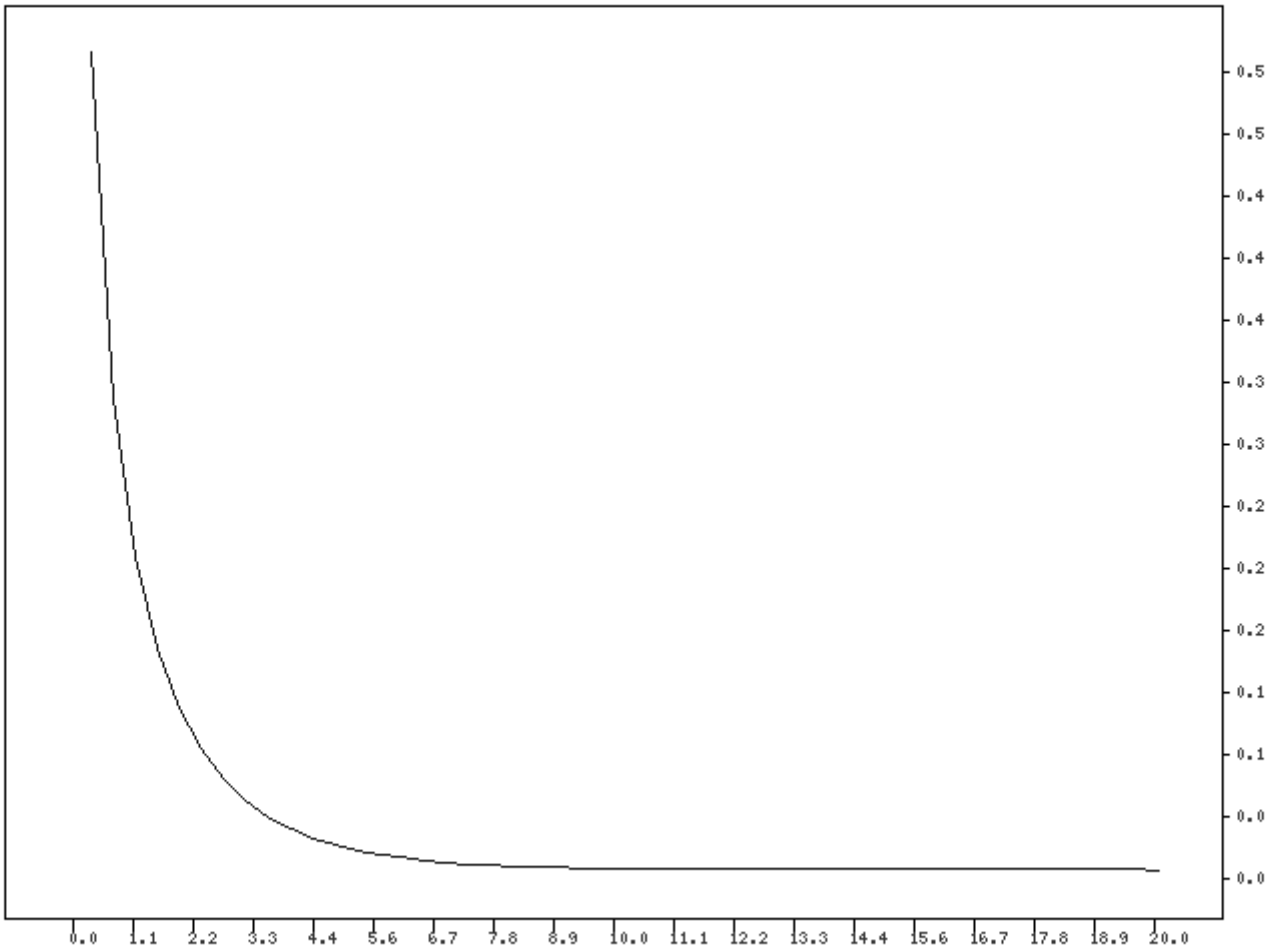
Beispiel Chi-Quadrat-Test

- Wenn in der GG kein Zusammenhang zwischen Geschlecht des Interviewers und des Befragten besteht, gilt die Indifferenztabelle
- tab v376 v928,exp
- Wenn empirische Daten Indifferenztabelle entsprechen, ergibt sich ein Chi-Quadrat von 0 (für alle Zellen Differenz zwischen erwarteten und beobachteten Werten bilden, durch erwarteten Wert teilen, über Zellen aufsummieren)

Beispiel Chi-Quadrat-Test

- Durch Stichprobenfehler können empirische Kreuztabellen von der Indifferenztabelle abweichen
- ➜ von null verschiedene Chi-Quadrat-Werte
- Diese folgen einer Chi-Quadrat-Verteilung mit einem Freiheitsgrad
- ➜ viele Werte im Bereich zwischen 0 und 1

Chi-Quadrat-Verteilung mit einem Freiheitsgrad



Beispiel Chi-Quadrat-Test

- Nur 5% aller Werte sind >=3,84 ➜ wenn sich in der Stichprobe ein empirischer Chi-Quadrat-Wert >=3,84 zeigt ist die Wahrscheinlichkeit, daß in der GG kein Zusammenhang besteht <=5%
- In diesem Fall ist der Zusammenhang „signifikant“
- tab v376 v928,chi V
- tab v376 v928,chi; exakter Wert der Irrtumswahrscheinlichkeit: display 1-chi2(1,r(chi2))

Beispiel t-Test

- Ist Mittelwertunterschied zwischen zwei Gruppen real, d.h. auch in der GG vorhanden?
- Ist das Einkommen der Männer in der GG höher als das der Frauen?
- Mittelwertunterschiede
  - folgen einer t-Verteilung
  - die bei größeren Fallzahlen in eine Normalverteilung übergeht
- Aus Standardfehler der Mittelwertdifferenz und t-Verteilung läßt sich ableiten, wie wahrscheinlich Stichprobenwert erreicht/überschritten wird, wenn realer Unterschied gleich null

Beispiel t-Test

- Wichtig: Beim t-Test muß man gerichtete (Männer verdienen mehr) und ungerichtete Hypothesen (es besteht ein Unterschied) unterscheiden
- recode v376 (1=0 m) (2=1 w),gen(weiblich)
- tabstat v495,by(weiblich)
- ttest v495,by(weiblich)

Zurück zur Regression

- Statt eines t-Test kann man auch eine Regression mit einer kategorialen unabhängigen Variablen rechnen
- reg v495 weiblich
  - Konstante entspricht dem Mittelwert der Männer
  - Effekt von „weiblich“ entspricht der Differenz
  - Standardfehler des Koeffizienten entspricht dem Standardfehler der Differenz
  - display _b[_cons]+_b[weiblich] zeigt den Mittelwert der Frauen

Zurück zur Regression

- t-Test/anova und Regression im Kern identisch
- Kategoriale Größen können in Form von Dummies als unabhängige Variablen fungieren
- Regressionskoeffizienten sind ebenfalls Schätzungen, die mit entsprechenden Standardfehlern behaftet sind

Standardfehler von Regressionskoeffizienten

- Vermitteln Eindruck von der Streuung der Koeffizienten um die wahren Koeffizienten in der GG
- Konfidenzintervalle / Hypothesentests
- Standardfehler der hängen ab von
  - der Varianz der Störgröße (die aus den Residuen geschätzt werden muß)
  - und der Summe der Abweichungsquadrate für die betreffende unabhängige Variable
- Bei konstanter Störvarianz erzielt man präzisere Schätzungen
  - durch größere Fallzahlen
  - durch größere Varianz der unabhängigen Variablen

Einkommen und Stichprobenfehler

- gen alter=v2-v372
- reg v495 weiblich alter
- Welches Einkommen erwarten wir für eine 36 Jahre alte Frau?
  - display _b[_cons]+_b[weiblich]*1+_b[alter]*36
  - Problem: alle Koeffizienten mit Standardfehler behaftet!

Einkommen

- Lösung 1: lincom _b[_cons]+_b[weiblich]*1+_b[alter]*36
- Lösung 2: SPost (Scott/Freese)
  - prvalue, x(weiblich 1 alter 36)
  - prtab alter weiblich

Annahmen und ihre Verletzung

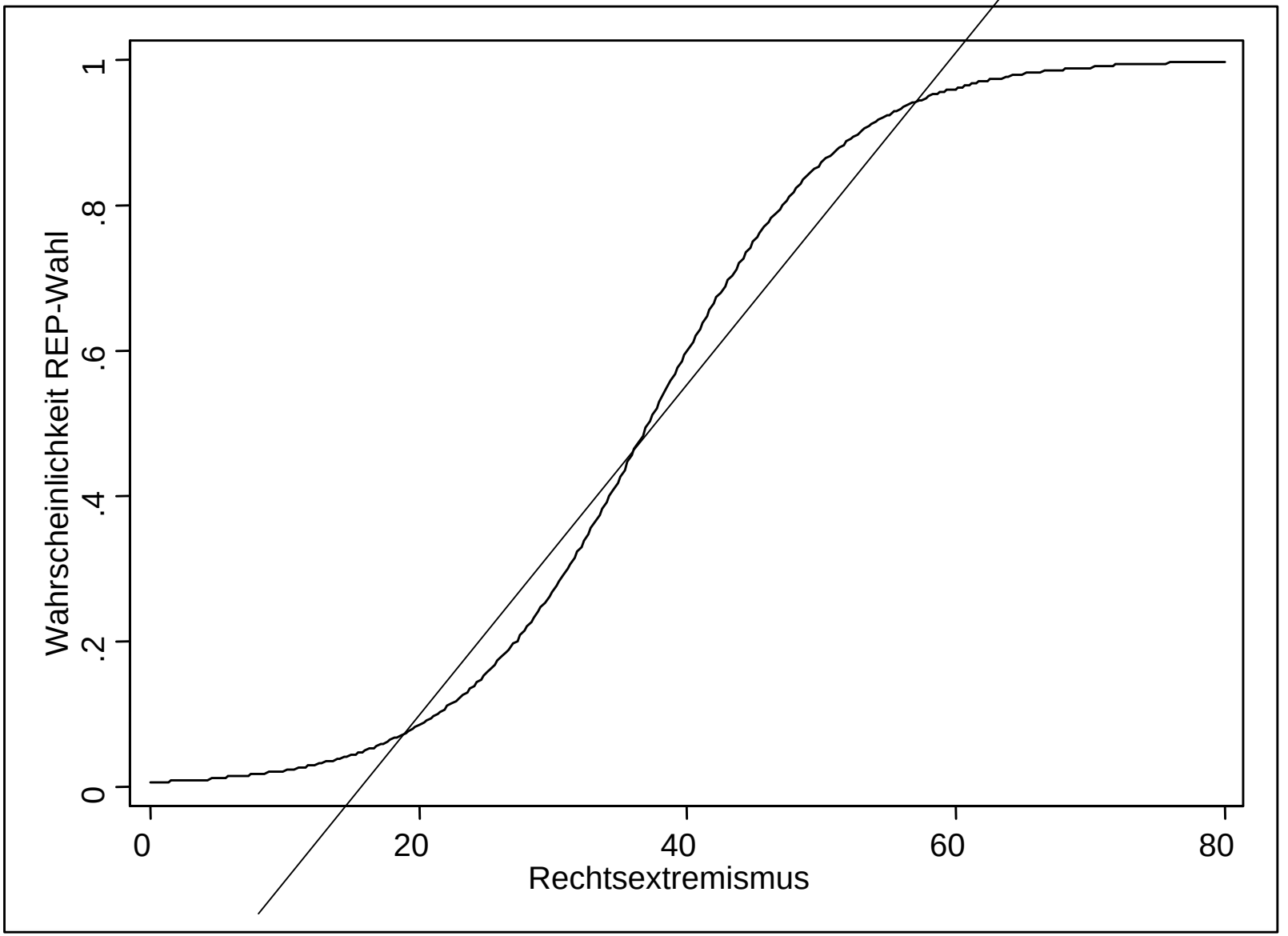
- Das klassische lineare Modell geht von einigen Annahmen aus
  - keine perfekte Multikollinearität (1)
  - Zufallsvariable e hat Mittelwert von null und ist für jede Kombination von x normalverteilt (2)
  - Zufallsvariable e nicht mit unabhängigen Variablen korreliert (3)
  - Zufallsvariable e hat konstante Varianz über alle Kombinationen von x hinweg (4, Homoskedaszität)
  - Ausprägungen von e bei zwei beliebigen Fällen unabhängig voneinander (5, Non-Autokorrelation)
  - abhängige Variable kontinuierlich und unbeschränkt (6)
- Gegenüber vielen Verletzungen dieser Annahmen ist das Modell relativ robust
- Drei Annahmen besonders häufig verletzt : 4-6

Was passiert wenn…

- Beispiele
  - Heteroskedaszität
    - die Varianz der zufälligen Fehler nimmt mit dem politischen Interesse ab
    - die Varianz der Fehler ist manchen Ländern größer als in anderen
  - Autokorrelation
    - länderspezifische Faktoren werden nicht modelliert (räumliche Korrelation)
    - In Zeitreihen wirken nicht-spezifizierte Einflüsse zum Zeitpunkt t auch noch bei t+1 (serielle Korrelation)
- Konsequenzen
  - Schätzung der Koeffizienten i.d.R. weiterhin unverzerrt, aber nicht mehr effizient ➜ evt. alternative Schätzverfahren
  - Schätzung der Standardfehler (nach unten) verzerrt ➜ Konfidenzintervalle und Tests zu optimistisch ➜ evtl. alternative Berechnung der Standardfehler notwendig

Abhängige kategoriale Variablen

- ordinale Variablen (nicht zufrieden, zufrieden, sehr zufrieden)
- nominale Variablen (Konfession, Wahlabsicht)
- Dichotome Variablen (0/1, Nichtwahl, Wahl einer rechten Partei)
  - Interpretation als Wahrscheinlichkeit / relative Häufigkeit für bestimmte Konstellationen ➜ lineares Wahrscheinlichkeitsmodell
  - Probleme: (Heteroskedaszität), vorhergesagte Wahrscheinlichkeiten <0, >1
  - S-förmiger Zusammenhang (nicht-linear)


Wahrscheinlichkeit REP-Wahl
0
.2
.4
.6
.8
1
0
20
40
60
80
Rechtsextremismus

Grundgedanke der logistischen Regression (Logit-Analyse)

- Wahrscheinlichkeiten haben Wertebereich 0-1
- Unterschiedliche Wahrscheinlichkeiten für verschiedene Gruppen
- Abhängige Variable wird deshalb transformiert:
- Zunächst werden statt Wahrscheinlichkeiten „Odds“ betrachtet, das sind Brüche aus einer Wahrscheinlichkeit und ihrer Gegenwahrscheinlichkeit: p ➔ p/1-p. Diese haben einen Wertebereich von 0 bis (fast) $+\infty$
- Diese Odds werden in einem zweiten Schritt logarithmiert:
  p/1-p ➔ ln(p/1-p) (natürlicher Logarithmus = Umkehrfunktion zur Exponentialfunktion auf Basis von e (2,718...) .

Grundgedanke der logistischen Regression (Logit-Analyse)

- Diese Größe wird auch als „logit“ bezeichnet. Logits haben einen Wertebereich von (fast) -∞ bis (fast) +∞ und werden häufig als latente kontinuierliche Variable betrachtet (fixierte Fehlervarianz)
- Als unabhängige Variablen kommen wie bei der linearen Regression intervallskalierte oder kategoriale Merkmale in Frage
- Modell ist ebenfalls multivariat darstellbar und linear in den Logits: $\text{logit}(y)=a+b_1x_1-b_2x_2\ldots$
- Wenn man sich für die abhängige Variable in der ursprünglichen Form interessiert, muß die Transformation rückgängig gemacht werden. Diese Beziehung ist nicht-linear

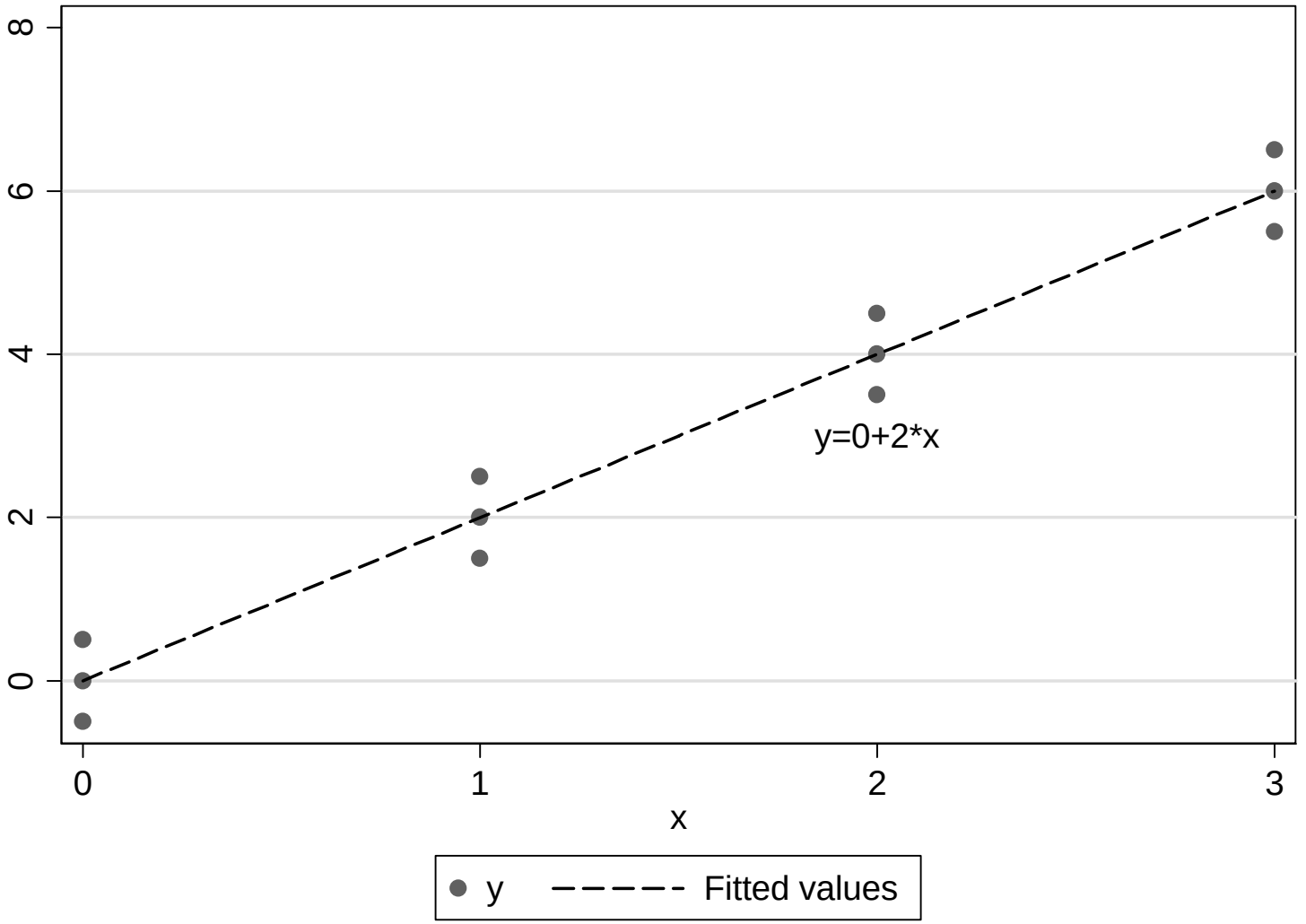
in Stata: invlogit() ←

$$y=\frac{e^{(a+b_1x_1+b_2x_2\ldots)}}{1+e^{(a+b_1x_1+b_2x_2\ldots)}}$$

Interpretation der Koeffizienten

- Bei der linearen Regression ist die Interpretation einfach
  - a ist der erwartete Wert von y, wenn $x_1 = 0$
  - b_1 entspricht der Veränderung des erwarteten Wertes von y, wenn x_1 um eine Einheit zunimmt.
  - Veränderungen von x und y sind proportional und vom Niveau der unabhängigen Variablen unabhängig
  - Positives Vorzeichen $\Leftrightarrow$ positiver Zusammenhang

Lineare Regression



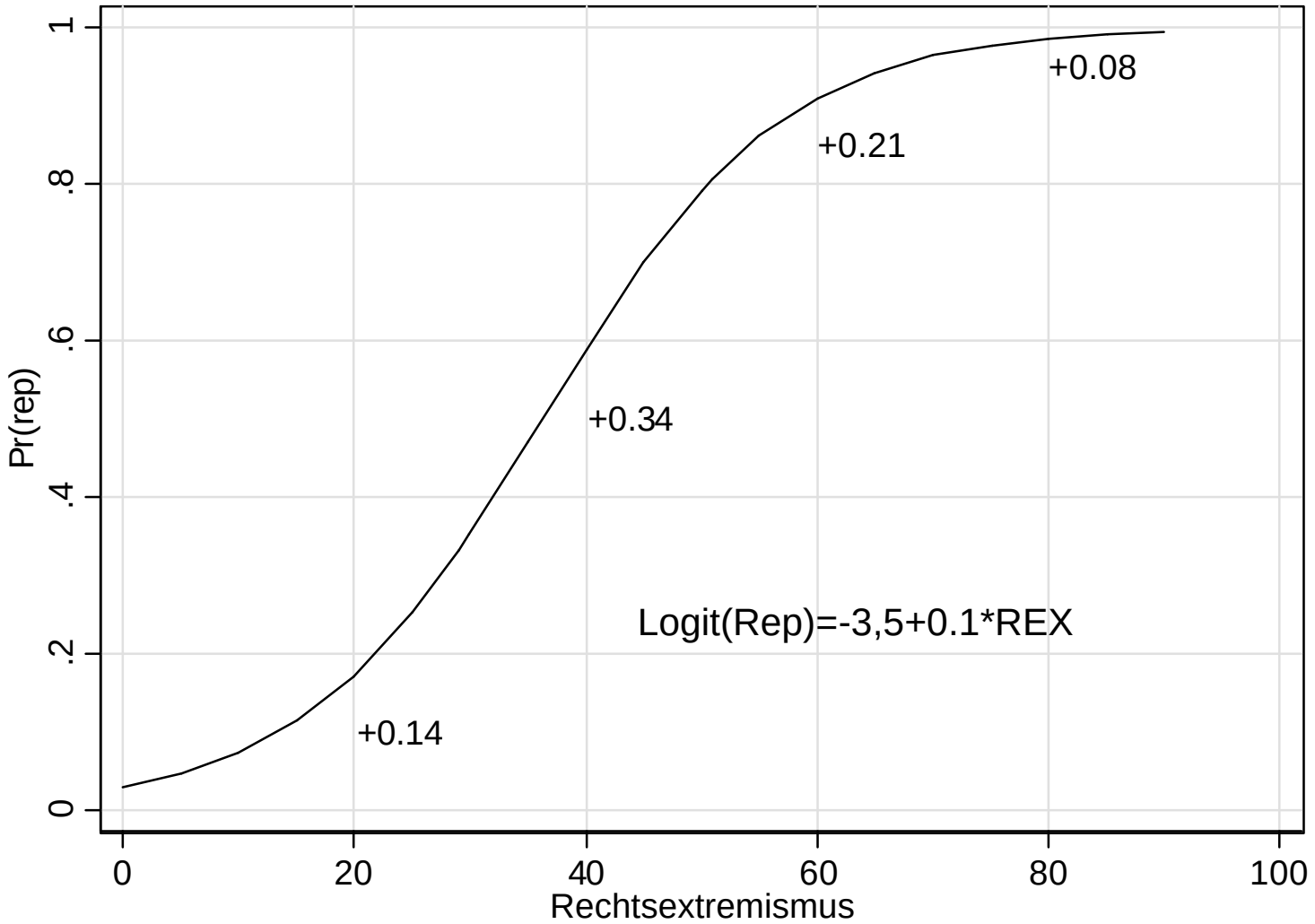
Interpretation der Koeffizienten

- Bei der logistischen Regression ist die Interpretation schwieriger:
  - Positives Vorzeichen $\Leftrightarrow$ positiver Zusammenhang (höhere Wahrscheinlichkeit)
  - Koeffizient b_1 beschreibt lineare Veränderungen des *Logits*, wenn x_1 um eine Einheit zunimmt. Bedauerlicherweise nicht intuitiv nachvollziehbar
  - *Etwas* anschaulicher ist e^{b1}: Dies ist der multiplikative *Faktor*, um den sich die Odds verändern, wenn x_1 um eine Einheit zunimmt

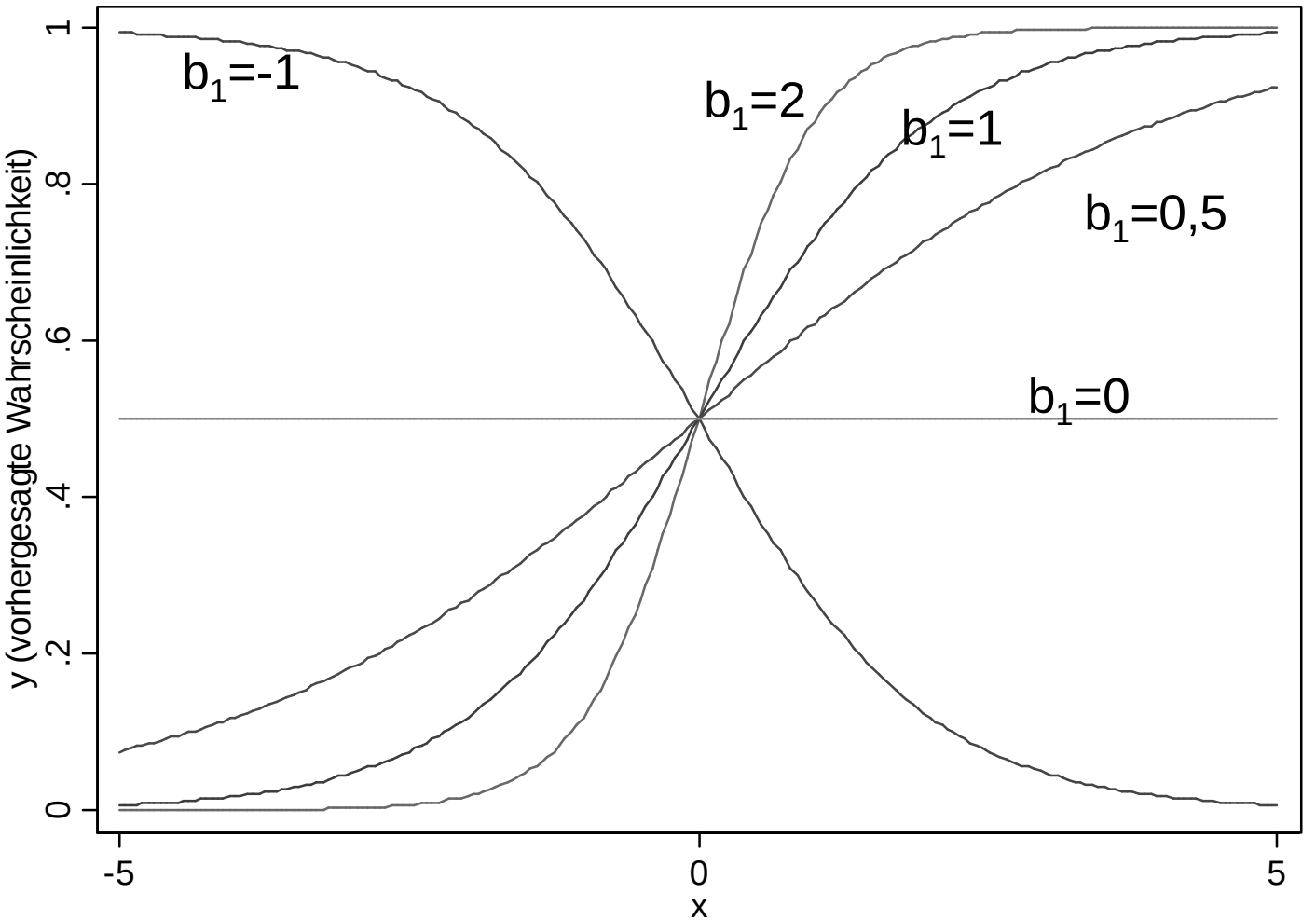
Interpretation der Koeffizienten

- Eigentlich interessant: geschätzten Wahrscheinlichkeiten
  - Wegen der S-Form der Kurve aber keine Proportionalität zwischen x und *Wahrscheinlichkeit*:
  - Steigung der Kurve (=Veränderung der Wahrscheinlichkeit) hängt vom Wert von x ab (geringe Steigung bei sehr hohen und sehr niedrigen Werten)
- Im multivariaten Logit-Modell hängt die Veränderung der geschätzten Wahrscheinlichkeit vom Wert *aller* unabhängigen Variablen ab
- Häufig ist es deshalb sinnvoll, sich typische bzw. interessante Konstellationen anzuschauen und für diese durch Einsetzen die geschätzte Wahrscheinlichkeit zu errechnen

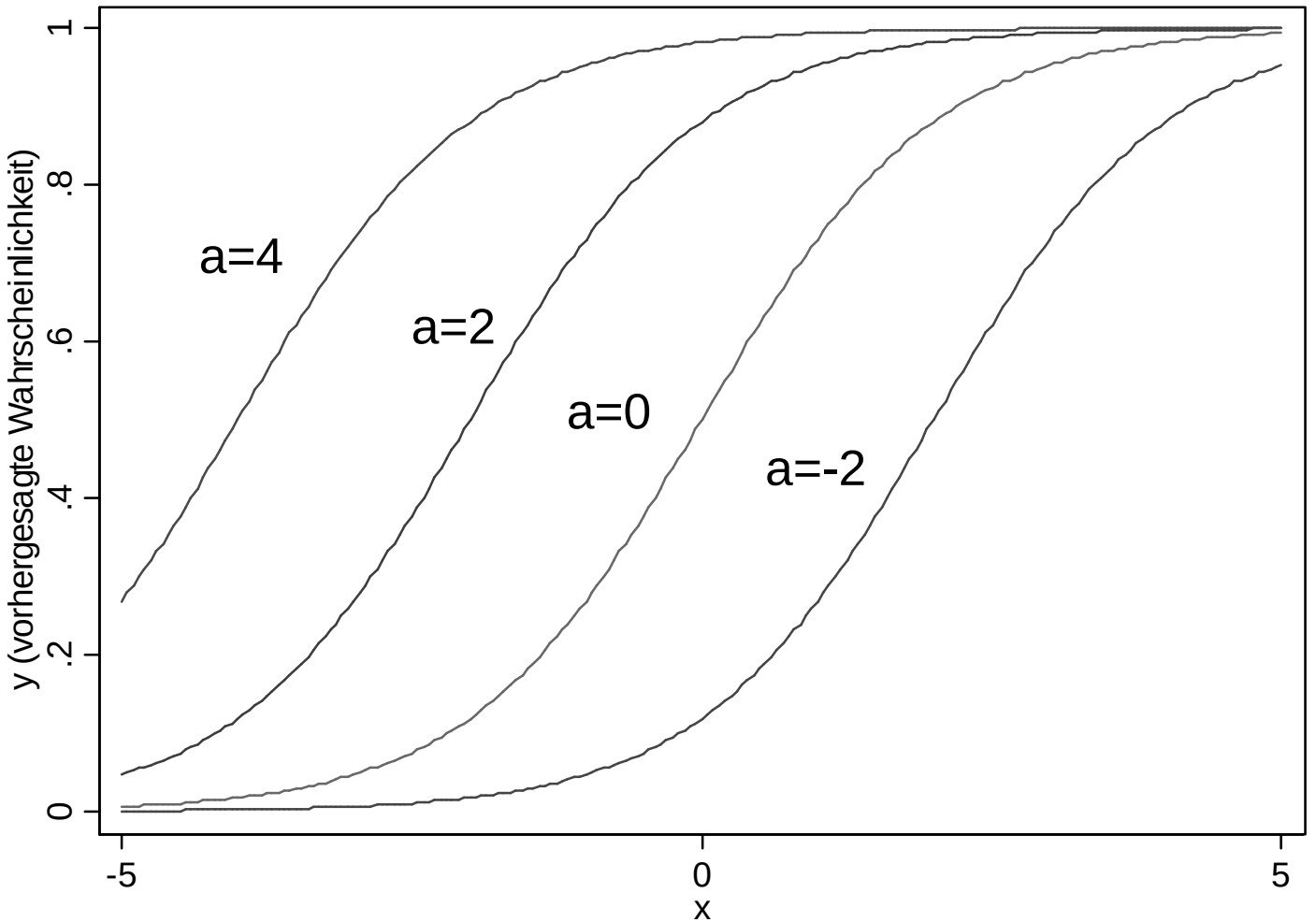
Logistische Regression



Wirkung von b_1 auf Wahrscheinlichkeiten (a=0)



Wirkung von a auf Wahrscheinlichkeit ($b_1 = 1$)



Implementation in Stata / Hausaufgabe

- Vollständig analog zu linearer Regresssion
  - logit y x1 x2 x3..., Optionen
  - Zugriff auf Koeffizienten, geschätzte Werte etc. wie bei linearer Regression
- Hausaufgabe
  - Übung zu Logit durcharbeiten (http://www.politik.uni-mainz.de/kai.arzheimer/Lehre-Stata/Logit-Uebung.html)
  - do-file
    - Allbus laden
    - Variable „PI vorhanden“ (v22) sinnvoll auf 0/1 umkodieren
    - Politisches Interesse sinnvoll umkodieren
    - Logistische Regression der PI auf Interesse
    - Geschlecht sinnvoll umkodieren
    - Logistische Regression PI auf Interesse und Geschlecht
    - Geschätzte Wahrscheinlichkeiten (=Anteile mit PI) für alle zehn Interesse*Geschlecht Gruppen (mit prtab)
    - bis zum 30. Juni an die bekannte Adresse