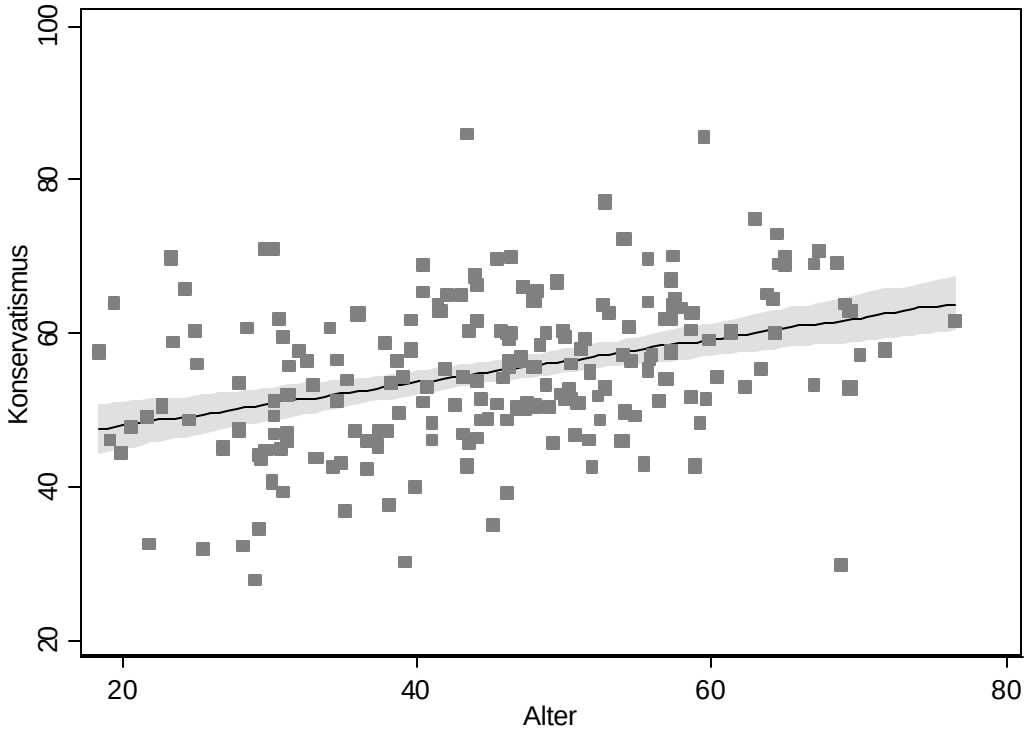
Bivariate Analyseverfahren

- Bivariate Verfahren beschäftigen sich mit dem Zusammenhang zwischen *zwei* Variablen
- Beispiel: Konservatismus/Alter
- Zusammenhangsmaße
  - beschreiben die Stärke eines Zusammenhangs durch eine Zahl zwischen 0 und 1 bzw. -1 und 1
  - Wahl des Zusammenhangsmaßes abhängig vom Skalenniveau (nominal, ordinal, intervall, ratio)
- Lineare Einfachregression: *Modelliert* eine abhängige Variable (bspw. Konservatismus) als lineare Funktion einer Konstanten und einer unabhängigen Variablen (bspw. Alter 18-80)

Lineare Einfachregression

- $y=a+bx \Leftrightarrow$ Konservatismus=a+bAlter
- Koeffizienten werden mit Stichprobe geschätzt
  - a: Konstante, geschätzter Wert für Konservatismus wenn Alter=0 (nicht immer anschaulich)
  - b: Steigung, Veränderung von Konservatismus, wenn Alter um ein Jahr zunimmt
- Voraussetzung: abhängige mindestens intervallskaliert, unabhängige intervallskaliert oder dichotom (0/1-kodiert)
- Modell ist linear, d.h. $\approx$ Zusammenhang kann durch eine gerade Linie beschrieben werden, auf der rechten Seite nur Addition/Multiplikation
- R^2 (0-1) beschreibt, wie gut die empirischen Daten vom Modell reproduziert werden können. $R^2=0 \Leftrightarrow$ kein Zusammenhang; $R^2=1 \Leftrightarrow$ alle Punkte liegen auf der Regressionsgeraden
- *Achtung: Die folgenden Beispiele beruhen auf künstlichen Daten un dienen nur der Veranschaulichung!*

Beispiel Einfachregression



- Konservatismus=42,3+0,28*Lebensalter: bei einem Altersunterschied von 10 Jahren ist ein um 2,8 Punkten höherer Skalenwert zu erwarten ➜ Je älter, desto konservativer
- R^2=0,12: Modell paßt nicht sehr gut

Dummy-Variablen

- Kategoriale Größen (Land, Geschlecht, Konfession u.ä.) können in Regressionsmodellen als unabhängige Variablen eingesetzt werden
- Dichotome Variablen (z.B. männliches Geschlecht) werden auf 0/1 kodiert: 1 = Merkmal liegt vor; 0= Merkmal liegt nicht vor
- Beispiel: Konservatismus= 55,9-2*männlich. 0/1 Variablen werden als „Dummies“ bezeichnet
- Nominalskalierte Merkmale mit mehr als zwei Ausprägungen (katholisch, protestantisch, keine Konfession) werden durch mehrere Dummies erfaßt. Dabei muß eine beliebige Kategorie ausgelassen werden („Referenzkategorie)
- Für ein Merkmal mit drei Ausprägungen werden deshalb nur zwei Dummies benötigt (der Wert des potentiellen dritten Dummies ist durch die ersten beiden schon festgelegt).
- Konservatismus=54,5+1,2*katholisch+0,6*protestantisch

Dummy-Kodierung: Konfession

Konfession	Dummy: Kath.	Dummy: Prot.
keine	0	0
Katholisch	1	0
Protestantisch	0	1
Logisch ausgeschlossen	1	1

Signifikanz

- Da Untersuchungen in der Politikwissenschaft meist auf Zufallsstichproben basieren, müssen die Ergebnisse darauf überprüft werden, ob sie durch Stichprobenfehler zustandegekommen sein könnten
- Für jeden Koeffizienten kann man errechnen, wie wahrscheinlich es ist, einen so großer Wert zu erhalten, falls der wahre Wert in der Grundgesamtheit gleich null ist
- Wenn diese Wahrscheinlichkeit sehr gering ist (<5% oder <1%) ist der Koeffizient statistisch signifikant von null verschieden
- Damit ist nicht ausgeschlossen, das sein Wert „in Wirklichkeit“ gleich null ist – es ist nur relativ unwahrscheinlich
- Statistische Signifikanz ≠ inhaltliche Relevanz

Probleme der Einfachregression

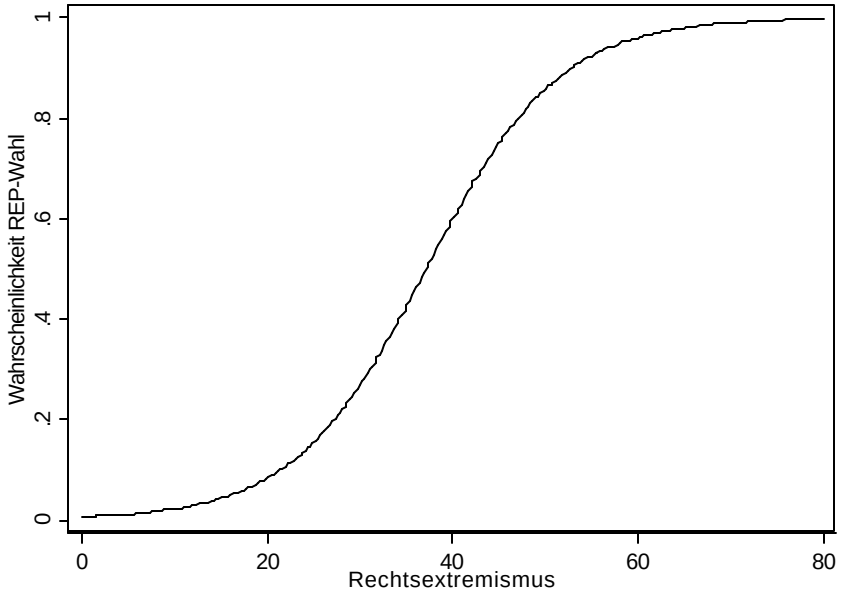
- Soziale Phänomene lassen sich selten auf eine einzige Ursache zurückführen
- Bei bivariater Betrachtung können „Scheinkorrelationen“ auftreten
  - In unserem Beispiel: Möglicher Zusammenhang zwischen Alter, Bildung (0-12) und Konservatismus
  - Zusammenhang zwischen Alter und Konservatismus kommt eventuell durch niedrigere Bildung der älteren Befragten zustande
- Lösung 1: Statistische Kontrolle durch Gruppenvergleich: Aufwendig und unübersichtlich
- Lösung 2: Multiple (multivariate) Regression:
  - $y=a+b_2x_1+b_2x_2$ ... $\Leftrightarrow$ $Konservatismus=a+b_1Alter+b_2Bildung$
  - Simultane Schätzung für zwei oder mehr Einflußfaktoren!
  - Wechselseitige Kontrolle der Einflußfaktoren

Beispiel multivariate lineare Regression

- Bivariat-1: Konservatismus=42,3+0,28*Lebensalter / R^2=0,12
- Bivariat-2: Konservatismus=73,1-3*Bildung / R^2=0,56
- Multivariat: Konservatismus=84,4-0,17*Lebensalter-3,6*Bildung ! /R^2=0,59
- 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 Konservatismus statistisch gesehen pro Punkt auf der Bildungsskala um 3,6 Punkte ab
  - Der Einfluß des Alters ist in „Wirklichkeit“ (bei Kontrolle der Bildung) schwach negativ!
  - Für jede beliebige Bildungsgruppe nimmt der Konservatismus pro Lebensjahr um 0,17 Punkte ab
  - Bei niedrigebildeten 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
- Bivariate Analysen führen oft in die Irre. Generell müssen alle relevanten unabhängigen Variablen berücksichtigt werden, damit man zu validen Ergebnissen kommt. Problem: Welches sind die relevanten Variablen?

Probleme der linearen Regression

- Diverse Anwendungsvoraussetzungen ➜ meist unproblematisch, da oft erfüllt und Verfahren robust
- Viele interessante abhängige Variablen in Politikwissenschaft dichotom oder dichotomisierbar (Nichtwahl, Wahl einer rechten Partei etc.)
- In diesem Fall häufig 0/1-Kodierung der Abhängigen und Interpretation als *Wahrscheinlichkeit*



- Im Prinzip akzeptabel, aber
  - Wahrscheinlichkeiten müssen zwischen 0 und 1 liegen
  - oft ist S-förmige Beziehung plausibler als lineare

Grundgedanke der logistischen Regression (Logit-Analyse)

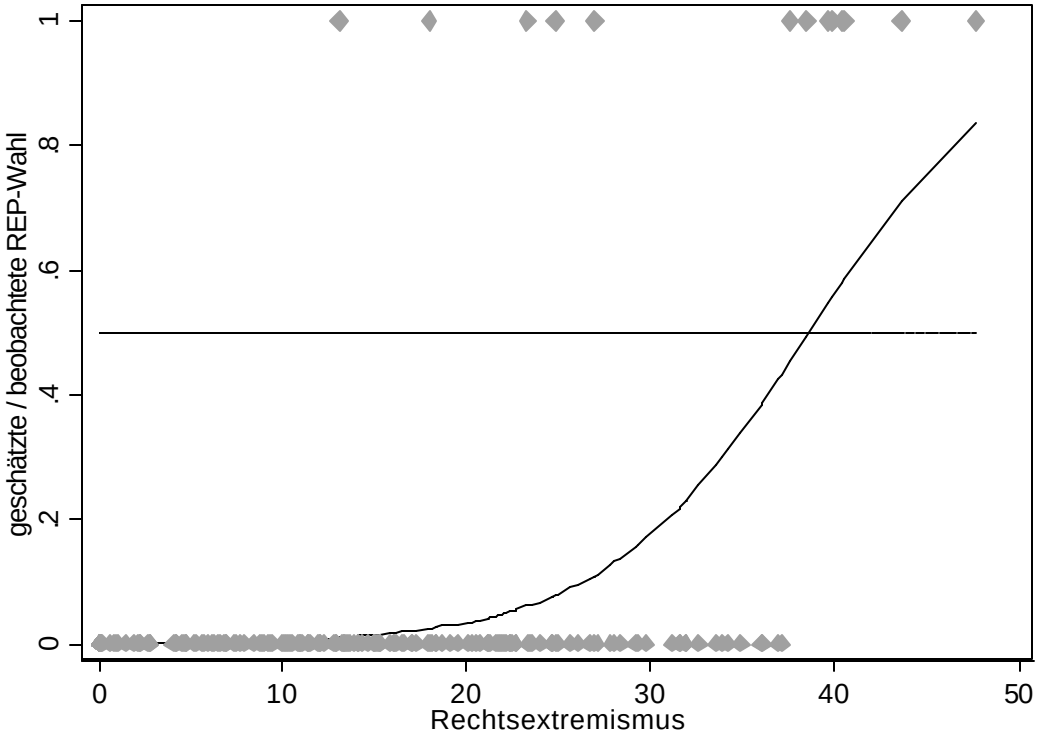
- Wahrscheinlichkeiten haben Wertebereich 0-1
- Abhängige Variable wird deshalb transformiert:
- Zunächst werden statt Wahrscheinlichkeiten „Odds“ betrachtet, das sind Brüche aus einer Wahrscheinlichkeit und ihrer Gegenwahrscheinlichkeit: p ➔ p/1-p. Diese haben einen Wertebereich von 0 bis (fast) +∞
- 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...) . Diese Größe wird auch als „logit“ bezeichnet. Logits haben einen Wertebereich von (fast) -∞ bis (fast) +∞
- 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

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

Interpretation der Koeffizienten

- Bei der linearen Regression ist die Interpretation einfach
  - b_1 entspricht der Veränderung 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 ⇔ positiver Zusammenhang
- Bei der logistischen Regression ist die Interpretation schwierig:
  - Positives Vorzeichen ⇔ positiver Zusammenhang (höhere Wahrscheinlichkeit)
  - Koeffizient b_1 beschreibt lineare Veränderungen des *Logits*, wenn x_1 um eine Einheit zunimmt. Bedauerlicherweise sind Veränderungen in den Logits für niemanden 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. Auch Odds sind aber nicht sehr anschaulich
  - Eigentlich interessant sind die 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 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

Beispiel logistische Regression



- Logit(REP-Wahl)=-6,93+0,18*Rechtsextremismus
- Multivariat: Logit(REP-Wahl)=-3,27+0,18*Rechtsextremismus-0,07*Alter
- Weitere unabhängige Variablen (z.B. Geschlecht) möglich
- Pseudo-R^2=0,39/0,44: Modell mit zwei Unabhängigen paßt vergleichsweise gut

Wahrscheinlichkeit REP-Wahl

REX	30 Jahre	50 Jahre	70 Jahre
3	0,7	0,2	0,0
12	3,6	0,8	0,2
22	18,8	5,0	1,1
32	59,3	24,7	6,9