Einführung in die Empirische Wirtschaftsforschung

Basierend auf dem Textbuch von RAMANATHAN: *Introductory Econometrics*

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

September 27, 2011

Outline

Empirische Wirtschaftsforschung

Die *empirische Wirtschaftsforschung* stellt die Verbindung zwischen ökonomischer Theorie und der realen ökonomischen Welt dar.

Informationen über die Ökonomie liegen in Form von *Daten* vor.

Diese Daten (*quantitative* eher als qualitative) sind die eigentlichen Studienobjekte.

Die Methoden zur statistischen Analyse der Daten (nicht zur Erhebung der Daten) werden heute meistens Ökonometrie genannt.

Ökonometrie

Begriff entsteht um 1900: Parallelbildung zu Biometrie etc.

Gründung der *Econometric Society* und ihrer Zeitschrift *Econometrica* (1930, Ragnar Frisch u.a.): *mathematische und statistische* Methoden in der Ökonomie.

Mathematische Methoden in der Ökonomie werden heute nicht mehr als Ökonometrie bezeichnet, spielen aber weiter eine Rolle in Econometric Society und Econometrica.

Schwerpunkte der Ökonometrie

Erhebung ökonomischer Daten spielt zunächst eine große Rolle (*national accounts*).

Cowles Commission: genuin eigenständige Entwicklungen von Verfahren zum Schätzen linearer dynamischer Modelle für aggregierte Variablen mit simultanem Feedback (privater Konsum $\rightarrow$ BIP $\rightarrow$ Einkommen $\rightarrow$ privater Konsum): in der Statistik sonst ungewöhnlicher Rahmen.

In den vergangenen Jahrzehnten stärkere Bedeutung mikroökonometrischer Analysen, geringere des makroökonometrischen Modells: Mikroökonometrie verwendet in anderen Bereichen der Statistik üblichen Rahmen.

Einige Textbücher zur Ökonometrie

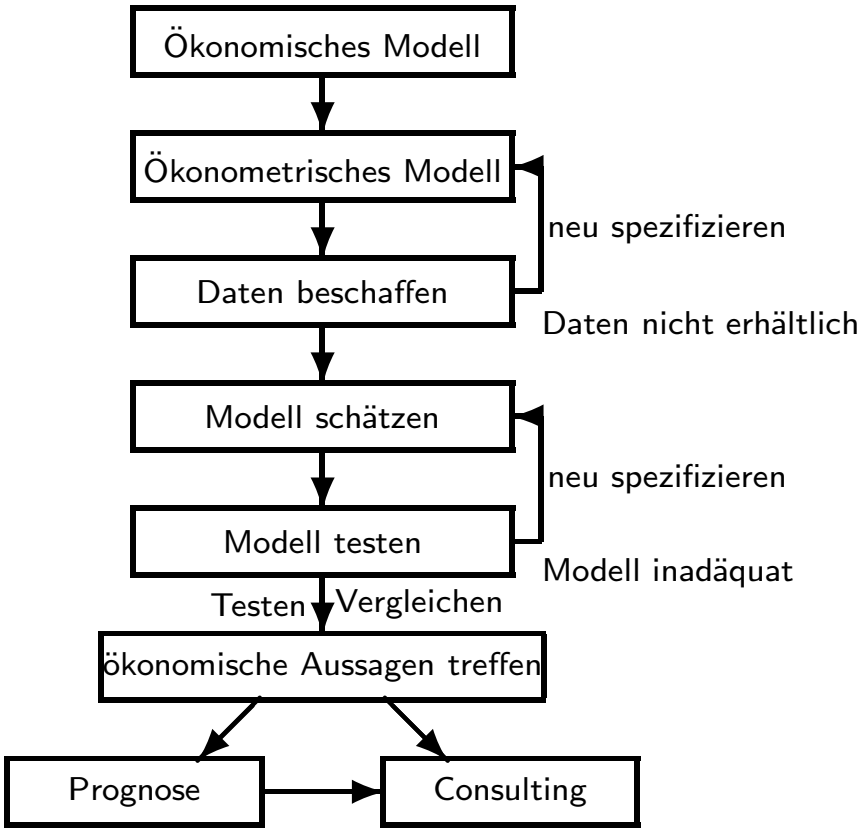
- Ramanathan, R. (2002). *Introductory Econometrics with Applications*. 5th edition, South-Western.
- Berndt, E.R. (1991). *The Practice of Econometrics*. Addison-Wesley.
- Greene, W.H. (2008). *Econometric Analysis*. 6th edition, Prentice-Hall.
- Hayashi, F. (2000). *Econometrics*. Princeton.
- Johnston, J. and DiNardo, J. (1997). *Econometric Methods*. 4th edition, McGraw-Hill.
- Stock, J.H., and Watson, M.W. (2007). *Introduction to Econometrics*. Addison-Wesley.
- Wooldridge, J.M. (2009). *Introductory Econometrics*. 4th edition, South-Western.

Ziele ökonometrischer Arbeit

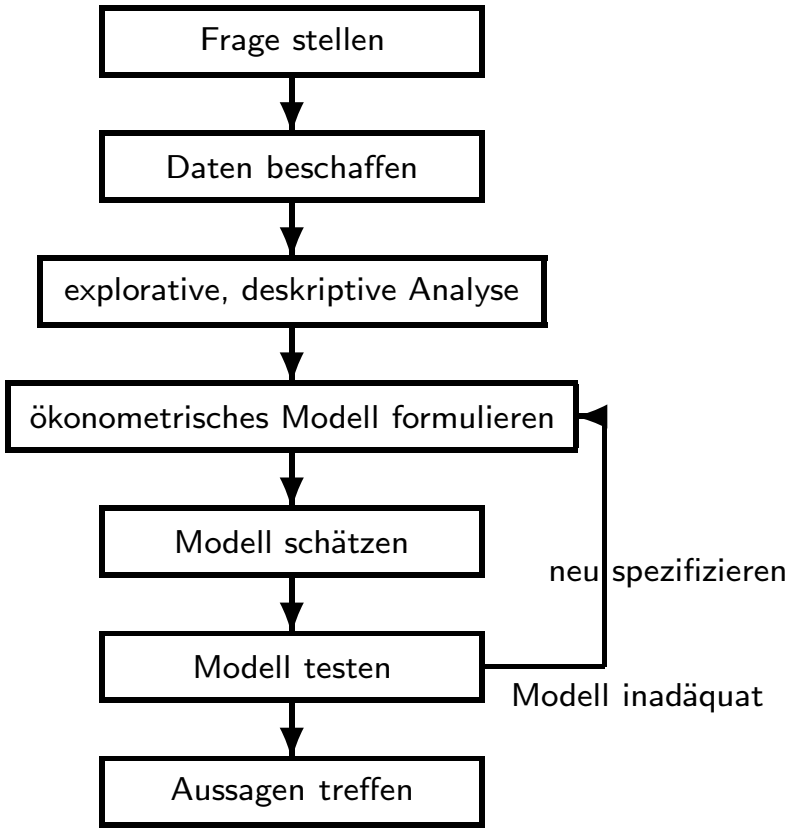
- Überprüfung ökonomischer Hypothesen;
- Quantifizierung ökonomischer Parameter;
- Prognose;
- Gewinnung neuer Erkenntnisse aus statistischer Evidenz (z.B. Wechselwirkungen von Variablen).

theory-driven oder *data-driven*

Flowchart nach Ramanathan



Modifiziertes Flowchart



Klassifikation von Daten nach dem Index

1. **Querschnittsdaten**: meist in der Mikroökonometrie. Index oft i steht für eine Person, Firma etc.;
2. **Zeitreihendaten**: meist in der Makroökonometrie und in Finance. Index t steht für einen Zeitpunkt (Jahr, Quartal, Monat, Tag, ...);
3. **Paneldaten**: zweidimensional, ein Index steht für 'Individuum' i, einer für den Zeitpunkt t;
4. mehrdimensionale Daten, z.B. mit räumlicher Dimension.

Im Folgenden werden die Beobachtungen immer mit t indiziert, auch wenn dieses t nicht zeitlich interpretiert ist: t ist *t*ypische Beobachtung.

Zeitreihen von Flows und Stocks

1. *Flows* sind die meisten makroökonomischen Account-Variablen: BIP für das Jahr 2005;
2. *Stocks* sind Mengen und Preise: Arbeitslose im Jänner 2009, können sich auf Stichzeitpunkte beziehen oder auf Mittelwerte über Zeiträume.

Unterschiede bei zeitlicher Aggregation: bei Flows entstehen Jahresdaten aus Monatsdaten durch Summieren; bei Stocks können Mittelwerte von Monaten gebildet werden, oft jedoch andere Konventionen.

Klassifikation von Daten nach Entstehung

1. *Experimentelle* Daten sind eher selten;
2. *Stichproben*: z.B. 100 Personen von 3 Mio. zufällig gewählt: mikroökonomische Situationen;
3. *Beobachtungsdaten*: BIP 2005 existiert nur einmal und wird von OeStat in einer Datenbank bereit gestellt: häufigster Fall.

Beobachtungsdaten können nicht vermehrt werden, ihre Grundgesamtheit (Population) ist ein gedankliches Konstrukt.

Typischer makroökonomischer Datensatz: aggregierter Konsum und verfügbares Haushaltseinkommen

Jahr	Konsum C	Einkommen YD
1976	71.5	84.2
1977	75.9	86.6
1978	74.8	88.9
...	...	...
2007	133.5	154.0
2008	134.5	156.7

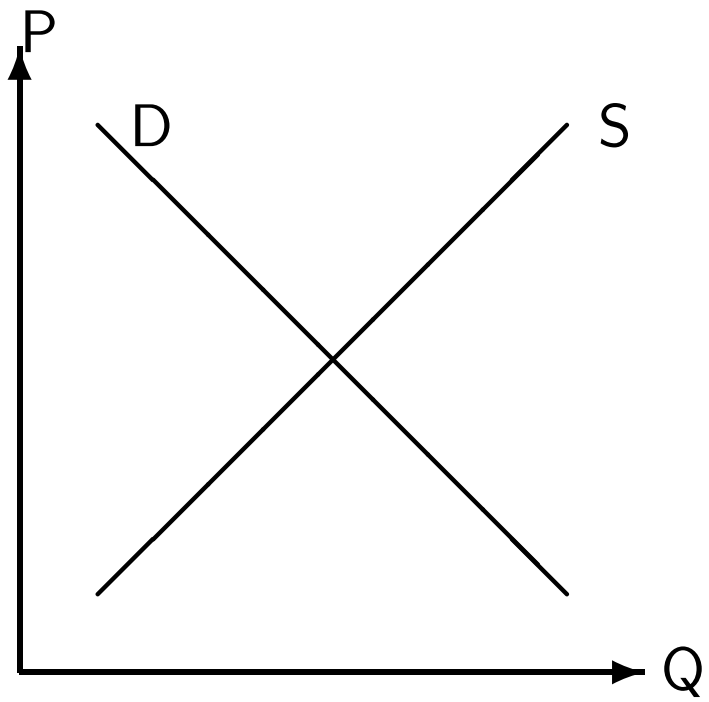
Was ist ein Modell?

Jede Wissenschaft hat ihre eigene Modellvorstellung.

- Ein ökonomisches Modell enthält eine Vorstellung über Wirkungszusammenhänge zwischen Variablen, die Variablen müssen nicht beobachtet sein;
- Ein ökonometrisches Modell enthält statistische Verteilungen der (potenziell) datengenerierenden Mechanismen für beobachtete Variablen.

Die Übersetzung zwischen beiden Modellwelten ist ein häufiger Schwachpunkt empirischer Arbeiten.

Beispiel: ökonomisches aber kein ökonometrisches Modell



Angebot und Nachfrage: Variable nicht beobachtet, Modell statistisch ungeeignet.

Beispiel: ökonometrisches Modell mit ökonomischem Interpretationsproblem

$$\begin{aligned} y_t &= \alpha + \beta x_t + \varepsilon_t \\ \varepsilon_t &\sim N(0, \zeta_0 + \zeta_1 \varepsilon_{t-1}^2) \end{aligned}$$

Regressionsmodell mit ARCH-Fehlern: was ist ζ_0? Untere Schranke für die Varianz von Schocks in sehr ruhigen Episoden?

Stichprobe (*sample*)

Notation: Es liegen n Beobachtungen über die Variable X vor:

$$X_1, X_2, \ldots, X_n,$$

oder

$$X_t, t = 1, \ldots, n.$$

Die *Stichprobengröße* ist n (der 'Umfang').

Das Sample wird in statistischer Analyse als *Realisierung einer Zufallsvariable* angesehen. Die statistische Notation, in der Großbuchstaben Zufallsvariablen und Kleinbuchstaben Realisierungen andeuten ($X = x$), wird in der Ökonometrie nicht eingehalten.

'Stichprobe' bei Beobachtungsdaten

Beispiel: Daten über Konsum C und (verfügbares) Einkommen YD 1976 bis 2008. Es existiert das Denkmodell einer 'Konsumfunktion'

$$C_t \approx \alpha + \beta YD_t.$$

Haben wir 33 Realisierungen von C und YD oder *eine* Realisierung von $(C_1, \ldots, C_{33})$ und $(YD_1, \ldots, YD_{33})$?

Praktisch oder ehrlich?

- 33 Realisierungen *einer* Zufallsvariable: C und YD wachsen über den Zeitbereich, zeitliche Abhängigkeit (Konjunktur, Trägheit der Haushaltsreaktion), Änderungen ökonomischen Verhaltens: unglaubwürdig. Parameter der Konsumfunktion schätzbar.
- Eine Realisierung: Interpretation der nicht beobachteten Realisierungen problematisch (Parallelwelten?). Parameter der Konsumfunktion nicht schätzbar.

Ausweg: nicht Variable, sondern unbeobachteter Fehlerterm ist die Zufallsvariable: 33 Beobachtungen, 'Schocks stationär', Konsumfunktion schätzbar.

Definition des Begriffes 'Parameter'

Ein Parameter ('jenseits des Messbaren') ist eine unbekannte Konstante. Sie beschreibt die statistischen Eigenschaften der Variablen, denen die Beobachtungen entstammen.

Beispiel: Im Regressionsmodell

$$C_t = \alpha + \beta YD_t + u_t, \quad u_t \sim N(0, \sigma^2),$$

sind α, β, σ^2 Parameter. Nicht jeder Parameter ist ökonomisch interessant.

Definition des Begriffes 'Schätzer'

Ein *Schätzer* ist eine aus den Daten gewonnene Statistik, die einen Parameter annähern soll. Das Wort 'Schätzer' wird sowohl für die Vorschrift der Berechnung dieser Statistik (Schätzfunktion) als auch für den Wert dieser Statistik im konkreten Beispiel (Schätzwert) verwendet.

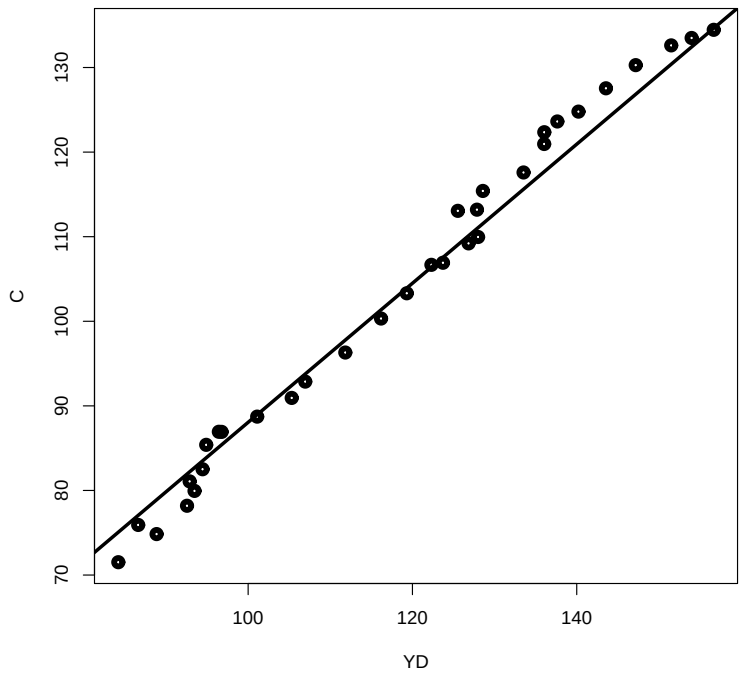
Schätzer sind Statistiken. Statistiken sind beobachtet. Daten sind Realisierungen von Zufallsvariablen, daher sind Statistiken Zufallsvariablen. Parameter sind unbeobachtet.

Notation für Schätzer und Parameter

Parameter werden gerne mit griechischen Buchstaben bezeichnet: α,β,θ,...

Zu diesen gehörige Schätzer werden mit Hut (Dach) bezeichnet: $\hat{\alpha}, \hat{\beta}, \hat{\theta}, \ldots$ Lateinische Buchstaben (*a* schätzt α etc.) sind weniger deutlich.

Beispiel für Schätzer



Konsum C in Abhängigkeit vom Einkommen YD als Konsumfunktion (Linie) und als Datenpunkte im Streudiagramm: Intercept $\hat{\alpha}$ schätzt autonomen Konsum, Neigung $\hat{\beta}$ schätzt marginale Konsumneigung der Haushalte.

Definition des Hypothesentests

Ein Test ist eine statistische Entscheidungsregel. Es wird eine Entscheidung gesucht, ob die Daten es nahelegen, eine vorgegebene Hypothese zu verwerfen oder anzunehmen ('nicht zu verwerfen').

Meistens wird aus den Daten zunächst eine *Teststatistik* berechnet, eine reelle Zahl, die eine Funktion der Daten ist. Fällt diese in den *Verwerfungsbereich* (kritischen Bereich), so verwirft der Test.

Tests: Häufige Fehler

- **Ein Test ist keine Teststatistik**. Die Teststatistik ist eine Zahl, der Test gibt nur Verwerfung oder Annehmen ('Nichtverwerfung') an, man kann diese Aussagen mit 0/1 codieren. Ein Test kann nicht 1.5 sein.
- **Ein Test ist keine Nullhypothese**. Ein Test kann nicht verworfen werden. Verworfen wird die getestete Hypothese. Der Test verwirft.

Gilt für ökonomische Theorien Unschuldsvermutung?

Die *Nullhypothese* ist eine i.a. eng gefasste Aussage über einen Parameter. Beispiel: $\beta = 1$. Dies entspricht oft einer ökonomischen Theorie.

Die *Alternativhypothese* oder *Alternative* ist oft die Verneinung der Nullhypothese im Rahmen einer allgemeinen Hypothese (engl. *maintained hypothesis*). Beispiel: $\beta \in (0, 1) \cup (1, \infty)$. Dies entspricht oft der Ungültigkeit einer ökonomischen Aussage.

Notation H_0 für die Nullhypothese und z.B. H_A für die Alternative. $H_0 \cup H_A$ ist die allgemeine Hypothese. Beispiel: allgemeine Hypothese $\beta > 0$ wird nicht getestet, sondern vorausgesetzt.

Hypothesen: Häufige Fehler

- **Eine Hypothese kann keine Aussage über einen Schätzer sein**. "$H_0 : \hat{\beta} = 0$" ist sinnlos, denn $\hat{\beta}$ ist exakt bestimmbar.
- **Alternativen werden nicht verworfen**. Der klassische Hypothesentest ist asymmetrisch. Nur die H_0 kann verworfen oder 'angenommen' werden. (Im Bayes-Test ist dies anders)
- **Hypothesen haben keine Wahrscheinlichkeiten**. H_0 ist nicht nach dem vorliegenden Testresultat zu 90% korrekt. (Im Bayes-Test ist dies anders)

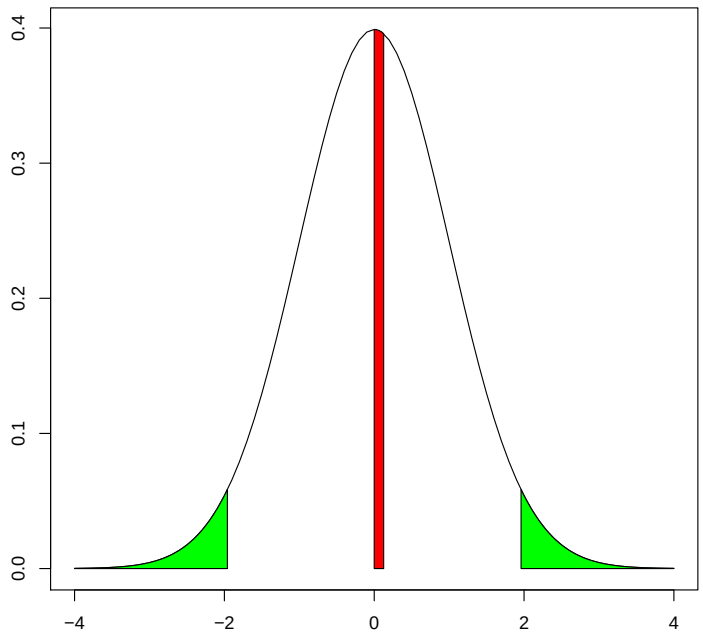
Fehler 1. und 2. Art

Da die Parameter nicht beobachtet werden und aus der Stichprobe nur approximativ bestimmt werden können, kommt es zu Fehlentscheidungen:

- Verwirft der Test, obwohl die Nullhypothese korrekt ist, dann ist dies ein **Fehler 1. Art** (*type I error*);
- Nimmt der Test die Nullhypothese an, obwohl sie falsch ist, dann ist dies ein **Fehler 2. Art** (*type II error*).

Es ist ein Grundprinzip klassischer Hypothesentests, die Wahrscheinlichkeit der Fehler 1. Art zu beschränken (etwa durch 1%, 5%) und unter dieser Schranke die Wahrscheinlichkeit des Fehlers 2. Art gering zu halten.

Zwei Tests auf dem Niveau von 5%



Dichte einer Teststatistik unter der H_0. 2 Tests verwenden die gleiche Teststatistik. Der mit rotem Verwerfungsbereich ist i.a. schlechter als der mit grünem Verwerfungsbereich, weil er mehr Fehler 2. Art impliziert.

Weiteres Testvokabular

Die meist vorgegebene Wahrscheinlichkeit des Fehlers 1. Art heißt **Size** des Tests oder einfach Signifikanzniveau. Bei einfachen Nullhypothesen ist sie annähernd konstant unter H_0, ansonsten eine Oberschranke.

Die Wahrscheinlichkeit, einen Fehler 2. Art *nicht* zu machen, heißt **Power** (Macht, Mächtigkeit, Güte) des Tests. Sie ist auf der Alternative definiert, von der 'Distanz' zur H_0 abhängig, nahe H_0 gering (nahe der Size), weit weg von H_0 nahe 1.

Erreicht der Test Power 1 auf der gesamten Alternative, wenn die Stichprobengröße gegen ∞ geht, dann heißt der Test **konsistent**.

Beste Tests

Die Analogie zu effizienten Schätzern, ein Test, der maximale Power bei gegebener Size aufweist, existiert nur in wenigen Fällen. Für die meisten Testprobleme hat aber der *Likelihood-Ratio-Test* (LR–Test, Teststatistik ist Quotient der Maxima der Likelihood unter H_0 und H_A) sehr gute Eigenschaften.

Oft sind der *Wald-Test* und der *LM–Test* (Lagrange multiplier) 'billigere' Approximationen zu LR und haben für große n identische Eigenschaften.

Achtung: LR–, LM– und Wald-Test sind *Konstruktionsprinzipien* für Tests, keine speziellen Tests. Es ist sinnlos zu sagen, dass 'der Wald-Test verwirft', wenn nicht klar ist, was getestet wird.

Gilt manchmal Unschuldsvermutung für Ungültigkeit der Theorie?

Ökonomische Aussagen müssen nicht der Nullhypothese eines Hypothesentests entsprechen. Manchmal ist die Nullhypothese die Ungültigkeit der Theorie. Z.B. die Theorie sagt, die Exportquote hängt vom Wechselkurs ab.

Die Nullhypothese des Tests entspricht auch dann der restriktiveren Aussage. Die ökonomische Aussage wird zur Alternative. Die Ungültigkeit der ökonomischen Theorie kann verworfen werden. Z.B. H_0 sagt, der Koeffizient der Exportquote in Abhängigkeit vom Wechselkurs ist 0 und dies wird verworfen.

Wie findet man Signifikanzpunkte?

Tests sind meist durch Teststatistik und durch kritische Werte definiert. Kritische Werte (Signifikanzpunkte) sind meist Quantile der Verteilung der Teststatistik unter H_0.

Bestimmung kritischer Werte:

1. Tabellen: Quantile wichtiger Verteilungen wurden durch Simulation oder analytisch tabelliert.
2. Ausrechnen: Computer invertiert Verteilungsfunktion.
3. p–Werte: Computer berechnet marginale Signifikanzniveaus.
4. Bootstrap: Computer simuliert Verteilungsfunktion unter H_0 und berücksichtigt dabei Samplecharakteristika.

Definition des p–Wertes

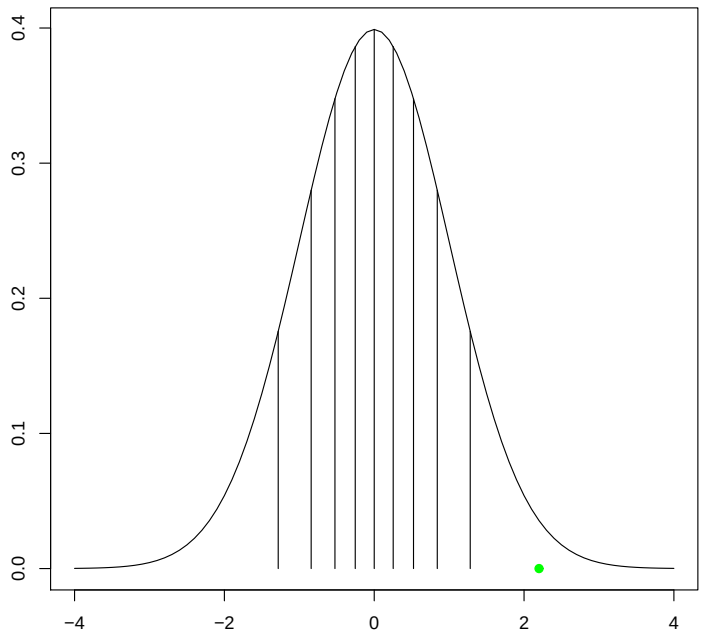
Richtige Definitionen:

- Der p–Wert ist jenes Signifikanzniveau, auf dem der Test für die vorliegende Stichprobe (Teststatistik) gerade noch verwirft = gerade nicht mehr verwirft;
- Der p–Wert ist die Wahrscheinlichkeit, unter der Nullhypothese noch ungewöhnlichere (atypischere, oft 'größere') Werte der Teststatistik zu produzieren als den vorliegenden.

Falsche Definition:

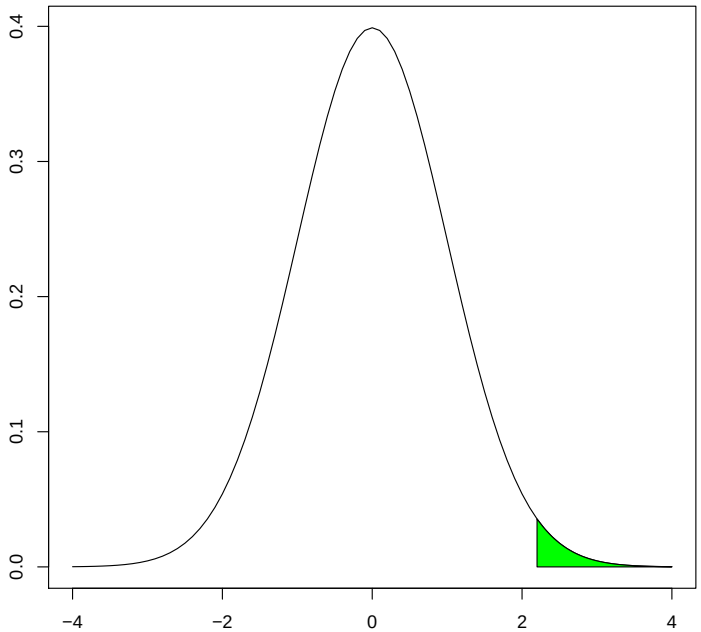
- Der p–Wert ist die Wahrscheinlichkeit der Nullhypothese für die vorliegende Stichprobe.

Test auf der Basis von Quantilen



10% bis 90%–Quantile der Normalverteilung. Der beobachtete Wert von 2.2 für die unter H_0 normalverteilte Teststatistik ist im einseitigen Test auf 10% signifikant.

Test auf der Basis von p–Werten



Die Fläche unter der Dichtekurve rechts vom beobachteten Wert 2.2 ist 0.014 und das ist auch der p–Wert. Der einseitige Test verwirft auf dem Niveau von 10%, 5%, aber nicht auf 1%.

Definition der Erwartung

Die *Erwartung* einer Zufallsvariable X mit Dichte $f(x)$ ist durch

$$\mathrm{E}X = \int_{-\infty}^{\infty} xf(x)dx$$

definiert. Bei diskreten Zufallsvariablen mit Wahrscheinlichkeiten $P(X = x_j)$ entspricht dies

$$\mathrm{E}(X) = \sum_{j=1}^{n} x_j P(X = x_j).$$

Linearität der Erwartung

Für zwei (auch voneinander abhängige) Zufallsvariablen X und Y und beliebiges $\alpha \in \mathbb{R}$ gilt

$$\mathrm{E}(X + \alpha Y) = \mathrm{E}X + \alpha \mathrm{E}Y,$$

und die Erwartung einer konstanten 'Zufallsvariable' a ist a.

Demgegenüber ist i.a. $\mathrm{E}(XY) \neq \mathrm{E}(X)\mathrm{E}(Y)$, und i.a. $\mathrm{E}g(X) \neq g(\mathrm{E}X)$ für eine allgemeine Funktion $g(.)$, insbesondere $\mathrm{E}(1/X) \neq 1/\mathrm{E}X$.

Schätzung des Erwartungswertes

Gebräuchlichster Schätzer für die Erwartung einer Zufallsvariable ist das *empirische Mittel*

$$\bar{X} = \sum_{t=1}^{n} X_t / n \quad .$$

Begriffspaar:

- Erwartung $\mathrm{E}X$: *population mean*, Parameter, unbeobachtet.
- Mittelwert $\bar{X}$: *sample mean*, Statistik, beobachtet.

Definition der Varianz

Die *Varianz* einer Zufallsvariable X mit Dichte $f(x)$ ist durch

$$\mathrm{var}X = \mathrm{E}(X - \mathrm{E}X)^2 = \mathrm{E}X^2 - (\mathrm{E}X)^2$$

definiert. Sie misst die Streuung einer Zufallsvariable um ihre Erwartung (zentriertes 2. Moment). Oft schreibt man $\sigma_X^2 = \mathrm{var}X$.

Die Wurzel der Varianz heißt Standardabweichung (*standard deviation*, eher bei beobachtetem X) oder Standardfehler (*standard error*, eher bei Statistiken) und wird auch σ_X geschrieben.

Eigenschaften der Varianz

1. $\mathrm{var}X \geq 0$;
2. $X \equiv \alpha \in \mathbb{R} \therefore \mathrm{var}X = 0$;
3. $\mathrm{var}(X + Y) = \mathrm{var}X + \mathrm{var}Y + \mathrm{cov}(X, Y)$ mit $\mathrm{cov}(X, Y) = \mathrm{E}\{(X - \mathrm{E}X)(Y - \mathrm{E}Y)\}$;
4. $\alpha \in \mathbb{R} \therefore \mathrm{var}(\alpha X) = \alpha^2 \mathrm{var}X$.

Der Varianzoperator ist nicht linear, $\mathrm{var}X$ kann auch $+\infty$ sein.

Schätzung der Varianz

Gebräuchlichster Schätzer für die Varianz einer Zufallsvariable ist die *empirische Varianz*

$$\widehat{\mathrm{var}}(X) = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \bar{X})^2 = \frac{1}{n-1}\sum_{t=1}^{n}X_t^2 - \frac{n}{n-1}\bar{X}^2.$$

Begriffspaar:

- Varianz $\mathrm{var}X$: *population variance*, Parameter, unbeobachtet.
- Stichprobenvarianz $\widehat{\mathrm{var}}X$: *sample variance*, Statistik, beobachtet.

Bias des Schätzers

Es sei $\hat{\theta}$ ein Schätzer für den Parameter θ. Der Wert

$$\mathrm{E}\hat{\theta} - \theta$$

heißt der *Bias* ('Verzerrung') des Schätzers. Ist der Bias gleich 0, ist also

$$\mathrm{E}\hat{\theta} = \theta,$$

dann heißt der Schätzer *erwartungstreu* oder *unverzerrt* (*unbiased*). Das ist jedenfalls eine wünschenswerte Eigenschaft von Schätzern: die Verteilung des Schätzers ist im wahren Wert zentriert.

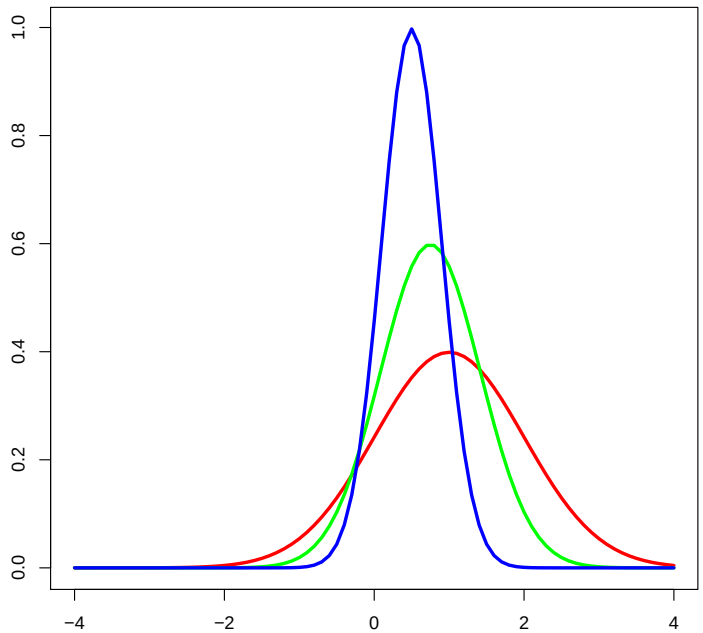
Konsistenz des Schätzers

Angenommen die Größe der Stichprobe geht gegen ∞. Für die Stichprobengröße n heiße der Schätzer $\hat{\theta}_n$. Gilt in einer vernünftigen Konvergenzdefinition

$$\hat{\theta}_n \to \theta, n \to \infty,$$

dann heißt der Schätzer *konsistent*. Das ist auch eine wünschenswerte Eigenschaft von Schätzern: die Verteilung des Schätzers schrumpft auf den wahren Wert zusammen, wenn die Stichprobe wächst.

Konsistenz ist wichtiger als Erwartungstreue



Konsistenter Schätzer mit Bias für den wahren Wert 0. Kleines Sample rot, mittleres grün, größtes blau.

Konsistenz und Erwartungstreue bedingen einander nicht

Beispiel: X_t seien unabhängig aus N(a,1) gezogen. Jemand glaubt *a priori*, b sei ein plausibler Erwartungswert. $\hat{a}_1 = \bar{X} + b/n$ ist konsistent, i.a. biased, und ein vernünftiger Schätzer.

Beispiel: Selbe Annahmen, $\hat{a}_2 = (X_1 + X_n)/2$ ist erwartungstreu, aber nicht konsistent und ein unvernünftiger Schätzer.

Beispiele konstruiert aber typisch: nicht konsistente Schätzer scheiden aus, verzerrte Schätzer können sinnvoll oder unvermeidlich sein.

Effizienz des Schätzers

Ein erwartungstreuer Schätzer $\hat{\theta}$ heißt effizienter als ein anderer erwartungstreuer $\tilde{\theta}$, wenn

$$\mathrm{var}\hat{\theta} < \mathrm{var}\tilde{\theta}.$$

Hat $\hat{\theta}$ minimale Varianz unter allen erwartungstreuen Schätzern, heißt er *effizient*.

Anmerkung: Ausschluss verzerrter konsistenter Schätzer nicht optimal, vorderhand reicht die Definition aber aus.

Die einfache lineare Regression

Das Modell erklärt eine Variable Y durch eine Variable X, wobei für beide Variablen *n* Beobachtungen vorliegen:

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1, \ldots, n,$$

wobei α und β zu bestimmende Parameter (Koeffizienten) sind. Die Regression heißt

- *einfach*, weil nur eine einzige Variable zur Erklärung herangezogen wird;
- *linear*, weil die Abhängigkeit durch eine lineare (eigentlich affine) Funktion modelliert wird.

Ausdrucksweise

Man sagt, Y wird auf X regressiert.

Y heißt die *abhängige Variable* der Regression, der Regressand, die erklärte Variable;

X heißt die *erklärende Variable* der Regression, der *Regressor*, die Designvariable oder Kovariate.

Der noch immer häufige Gebrauch des Ausdrucks 'unabhängige' Variable für X ist nicht sinnvoll, weil X nur in echten Experimenten unabhängig gesetzt werden kann und weil in der wichtigen multiplen Regression die Regressoren i.a. untereinander abhängig sind.

α und β

α heißt das *Intercept* der Regression oder die Regressionskonstante. Das ist der Wert auf der y–Achse für $x = 0$ auf der theoretischen Regressionsgerade $y = \alpha + \beta x$. Beispiel: in der Konsumfunktion ist α der autonome Konsum.

β heißt der *Regressionskoeffizient* von X und gibt die marginale Reaktion von Y auf Änderungen in X an. Das ist der Anstieg der theoretischen Regressionsgerade. Beispiel: in der Konsumfunktion ist β die marginale Konsumneigung.

Was ist u?

Im deskriptiven Regressionsmodell ist u einfach der unerklärte Rest.

Im statistischen Regressionsmodell ist das wahre u_t eine unbeobachtete Zufallsvariable, der *Fehler* oder Störterm der Regression.

Das Wort 'Residuen' bezeichnet den beobachteten Fehler, der nach Schätzung der Koeffizienten entsteht. Manche Autoren nennen aber den unbeobachteten Fehler u_t 'wahre Residuen'.

Ist die einfache Regression ein wichtiges Modell?

Das einfache Regressionsmodell bietet gegenüber Korrelationsanalyse zwischen zwei Variablen wenig Neues und ist eher uninteressant.

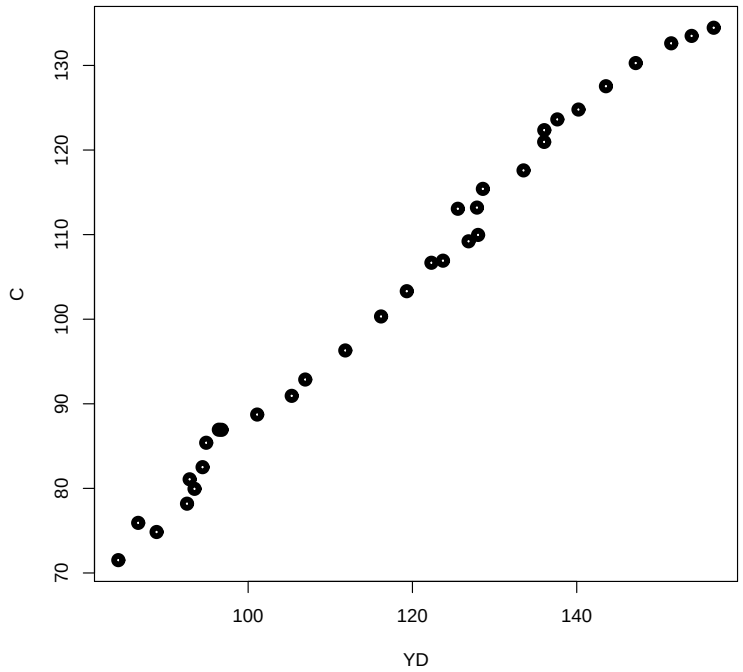
Es ist ein didaktisches Hilfsmittel, vieles erstmals zu verstehen, was dann im *multiplen Regressionsmodell* weiter gilt. Dieses multiple Modell ist das wichtigste Modell der empirischen Ökonomie.

Die deskriptive Regression

Im deskriptiven Problem werden keine Annahmen über die statistische Natur der Variablen und Fehler getroffen. Es sind keine statistischen Aussagen möglich.

Das deskriptive lineare Regressionsproblem besteht darin, eine Gerade durch eine Punktwolke zu legen, sodass diese möglichst gut passt.

Beispiel für Punktwolke: aggregierter Konsum

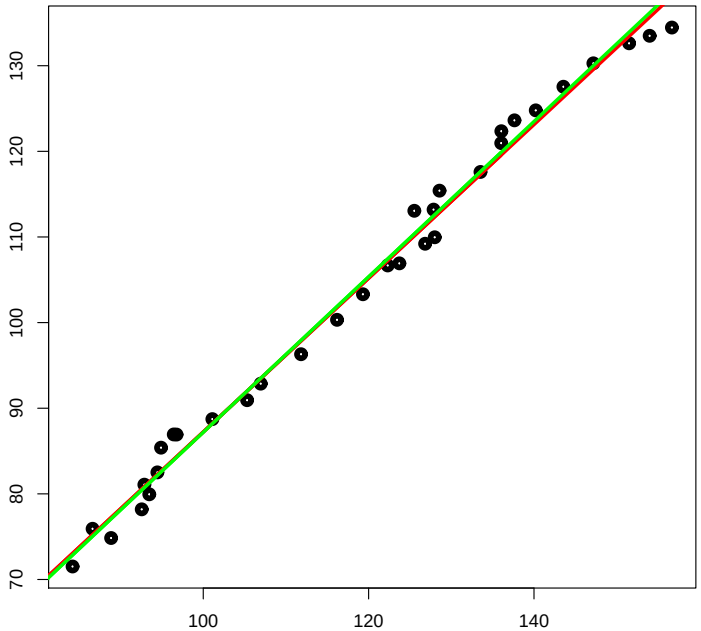


Konsum C und Einkommen YD für Österreich als Datenpunkte im Streudiagramm.

Vorschläge zum Legen einer Gerade durch Punkte

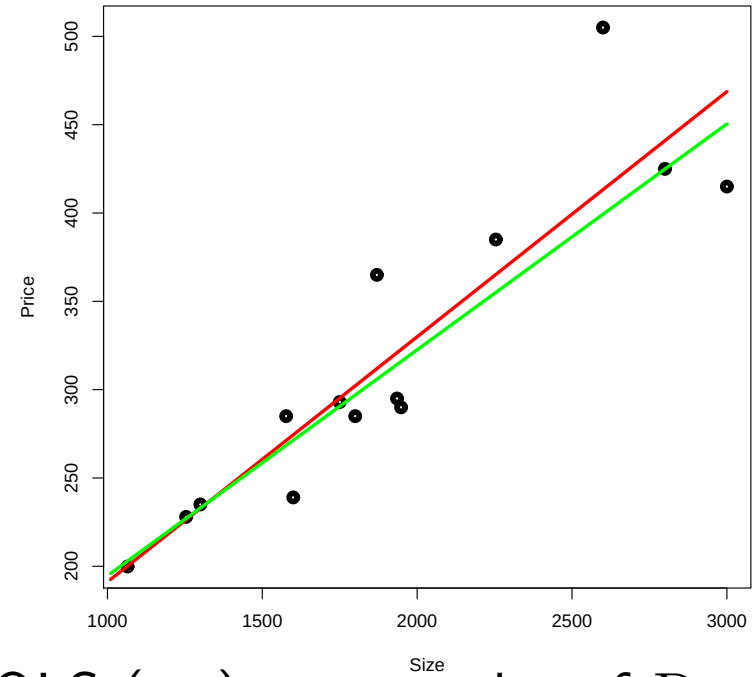
1. Gerade sollte möglichst viele Punkte exakt treffen: meistens lassen sich nur zwei Punkte auf die Gerade heften;
2. mit der Hand durchzeichnen ('Freihandmethode'): subjektiv;
3. Abstände (Gerade zu Punkten) in der Ebene minimieren: Orthogonalregression, wichtige Methode aber wenig verwendet. Variable sind gleich berechtigt, wie bei Korrelationsanalyse;
4. Summe vertikaler absoluter Abstände minimieren: *least absolute deviations* (LAD). Robustes Verfahren;
5. Summe vertikaler quadrierter Abstände minimieren: *ordinary least squares* (OLS). Gebräuchlichste Methode.

In guten Samples arbeiten viele Regressionsmethoden ähnlich



LAD–Regression (grün) und OLS–Regression (rot) auf die Konsumdaten angewendet.

Mehr Unterschiede zwischen Methoden in Samples moderater Qualität



LAD (grün) und OLS (rot) angewandt auf RAMANATHAN's Sample von 14 Häusern in San Diego, deren Größe und deren Preis.

Das OLS–Regressionsproblem ist leicht lösbar

Wir minimieren

$$Q(\alpha, \beta) = \sum_{t=1}^{n} (y_t - \alpha - \beta x_t)^2$$

in α and β, die Minima heißen $\hat{\alpha}$ und $\hat{\beta}$. Differenzieren nach α und Nullsetzen

$$\frac{\partial}{\partial \alpha} Q(\alpha, \beta) = 0$$

führt auf

$$-2 \sum_{t=1}^{n} (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \therefore \bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}.$$

Das ist die so genannte 1. Normalgleichung: die empirischen Mittel liegen exakt auf der Regressionsgerade.

Die 2. Normalgleichung

Differenzieren nach β und Nullsetzen

$$\frac{\partial}{\partial \beta} Q(\alpha, \beta) = 0$$

führt auf

$$-2 \sum_{t=1}^{n} (y_t - \hat{\alpha} - \hat{\beta} x_t) x_t = 0 \therefore \frac{1}{n} \sum_{t=1}^{n} y_t x_t = \hat{\alpha} \bar{x} + \hat{\beta} \frac{1}{n} \sum_{t=1}^{n} x_t^2.$$

Dies ist die 2. Normalgleichung. 2 Gleichungen in 2 Variablen $\hat{\alpha}$ und $\hat{\beta}$ sind lösbar.

Der OLS–Schätzer

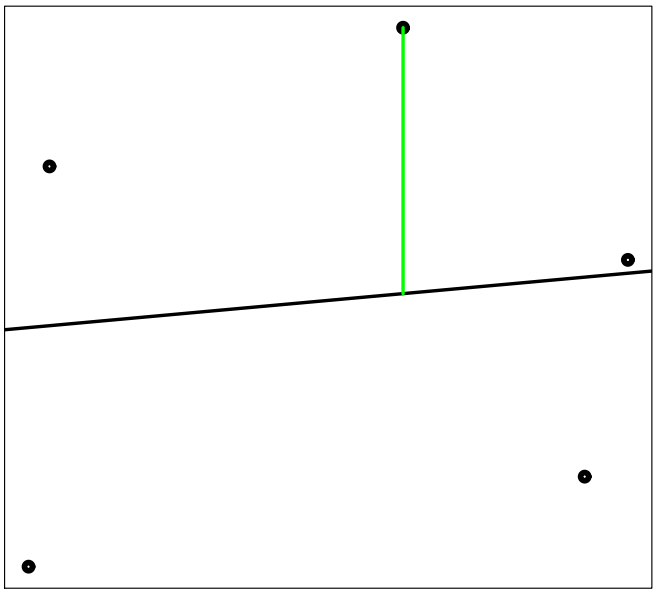
Aus den beiden Normalgleichungen gewinnt man die Lösung

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}.$$

Der Regressionskoeffizient ist eine empirische Kovarianz von X und Y dividiert durch eine empirische Varianz von X. Die Lösung für $\hat{\alpha}$ ist weniger leicht zu interpretieren:

$$\hat{\alpha} = \frac{\bar{y}\frac{1}{n}\sum_{t=1}^{n} x_t^2 - \bar{x}\frac{1}{n}\sum_{t=1}^{n} x_t y_t}{\frac{1}{n}\sum_{t=1}^{n} x_t^2 - \bar{x}^2}$$

OLS–Residuen

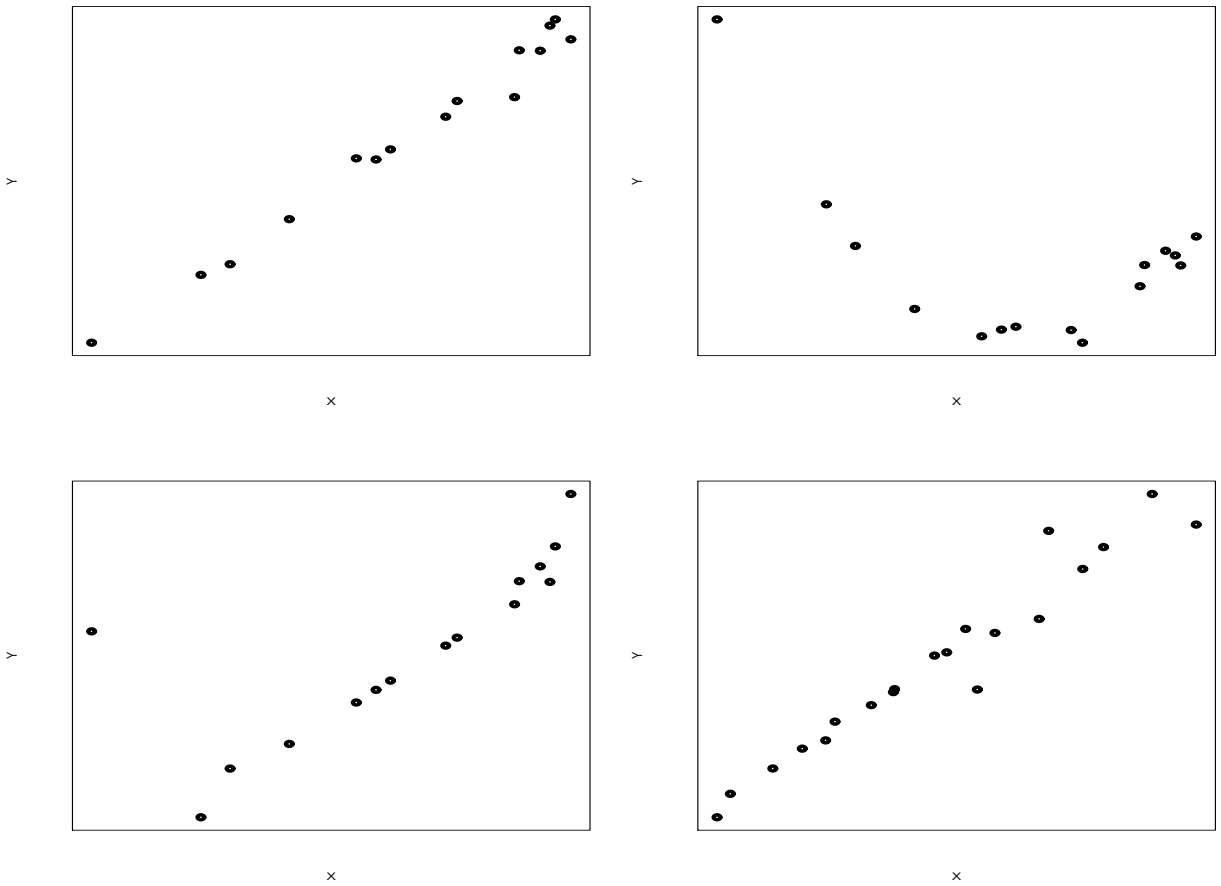


Der vertikale Abstand zwischen dem Wert y_t und der Höhe des Fußpunktes auf der eingepassten OLS–Gerade $\hat{\alpha} + \hat{\beta} x_t$ heißt *Residuum*.

Streudiagramme

Streudiagramme (*scatter plots*) mit abhängiger Variable auf der y–Achse und Regressor auf der x–Achse sind ein wichtiges grafisches Hilfsmittel in der einfachen Regression.

Da eine *lineare* Regressionsgerade eingepasst wird, sollte ein *linearer* Zusammenhang erkennbar sein.



Von links oben: schönes Diagramm, Verdacht auf Nichtlinearität, Ausreißer, Verdacht auf Heteroskedastie

Stochastische lineare Regression: Idee

Im klassischen Regressionsmodell

$$y_t = \alpha + \beta x_t + u_t$$

wirkt ein nicht zufälliger Input X auf einen Output Y ein und wird von der Zufallsvariable U gestört. Dadurch wird auch Y zur Zufallsvariable.

Annahmen über das 'Design' X und die statistischen Eigenschaften von U erlauben statistische Aussagen über Schätzer und Tests.

Stochastische lineare Regression: Kritik

Sowohl X als auch Y sind meistens Beobachtungsdaten, es liegt keine experimentelle Situation vor. Sollten nicht X und Y als Zufallsvariablen modelliert werden und über diese beiden Annahmen gesetzt werden?

Grundsätzlich berechtigte Kritik. Es ist aber didaktisch viel einfacher, mit diesem Modell zu beginnen, und man kann es leicht später erweitern. So genannte *voll stochastische* Modelle verwenden oft einschränkende Annahmen über die Unabhängigkeit der Beobachtungen.

Annahme 1: Linearität des Modells

Annahme 1. Die Beobachtungen entsprechen dem linearen Modell

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1, \ldots, n.$$

Theoretisch eine Leerannahme ohne weitere Annahmen über die Fehler u_t.

Empirischer Sinn: nichtlineare Zusammenhänge können oft durch Transformation von X und/oder Y linear gemacht werden.

Annahme 2: Vernünftiges Design

Annahme 2. Nicht alle beobachteten Werte für X_t sind gleich.

Experimente über Effekte von X auf Y sind nur sinnvoll, wenn man X variiert. Eine Beobachtung ist zuwenig. Vertikale Regressionsgeraden sind unerwünscht. Angenehm: diese Annahme lässt sich sofort an Hand der Daten (ohne Hypothesentest und exakt) überprüfen.

Annahme 3: Stochastische Fehler

Annahme 3. Die Fehler u_t sind Realisierungen von Zufallsvariablen U_t, wobei gelten möge, dass

$$\mathrm{E}U_t = 0 \quad \forall t.$$

Eigentlich 2 Annahmen: U ist unbeobachtete Zufallsvariable, und deren Erwartung ist 0. Erwartung ungleich 0 sinnlos, sonst ist die Regressionskonstante nicht definiert (nicht identifiziert).

Annahme 4: Natürliches Experiment

Annahme 4. Die erklärende Variable X_t ist nicht stochastisch. Die Werte $x_t, t = 1, \ldots, n$, sind gegebene Daten.

Wird das Experiment wiederholt, werden die gleichen X_t gesetzt, aber die Werte für u_t und damit y_t ändern sich. Technische Hilfsannahme.

Ist der Regressor theoretisch von Y abhängig oder ist er ein gelagtes Y, muss man diese Annahme aufgeben und anderweitig ersetzen.

Die Bedeutung von Annahmen 1–4

Unter Annahmen 1–4 ist das Modell bereits ein sinnvoll definiertes Modell.

Unter **Annahmen 1–4** sind die OLS–Schätzer $\hat{\alpha}$, $\hat{\beta}$ für die Koeffizienten α und β **erwartungstreu**;

Gemeinsam mit technischen Hilfsannahmen reichen Annahmen 1–4 auch aus, um die **Konsistenz** der OLS–Schätzer zu sichern.

Erwartungstreue von $\hat{\beta}$

Man erinnere, dass

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}.$$

Anwendung des Erwartungsoperators ergibt

$$\mathrm{E}\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})\mathrm{E}(y_t - \bar{y})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2},$$

weil der Nenner und alle x–Terme nicht stochastisch sind (Annahme 4). Annahme 2 garantiert, dass $\hat{\beta}$ berechnet werden kann (Nenner nicht 0).

Erwartungsterm auswerten durch Einsetzen von Annahme 1:

$$\mathrm{E}(y_t - \bar{y}) = \mathrm{E}(\alpha + \beta x_t + u_t - \alpha - \beta\bar{x} - \bar{u}),$$

wegen

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n} y_t &= \frac{1}{n}\sum_{t=1}^{n}(\alpha + \beta x_t + u_t) \\ &= \alpha + \beta\bar{x} + \bar{u}, \end{aligned}$$

mit der üblichen Notation für arithmetische Mittel. Wegen Annahme 3 ist die Erwartung der u–Terme 0, und man hat

$$\mathrm{E}(y_t - \bar{y}) = \beta(x_t - \bar{x}),$$

nicht stochastisch wegen Annahme 4.

Einsetzen ergibt nun sofort

$$\mathrm{E}\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})\beta(x_t - \bar{x})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2} = \beta.$$

Kann man das auch ohne Annahme 4 beweisen?

- Wenn X stochastisch, aber 'exogen' ist, kann man auf X bedingte Erwartungen auswerten und dann $\mathrm{E}\{\mathrm{E}(\hat{\beta}|X)\} = \mathrm{E}\hat{\beta}$ verwenden: OLS ist auch dann erwartungstreu;
- wenn X ein gelagtes Y ist, wirkt dieser Trick nicht, OLS ist dann i.a. nicht erwartungstreu.

Erwartungstreue von $\hat{\alpha}$

Wegen der 1. Normalgleichung

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

gilt

$$\begin{aligned} \mathrm{E}\hat{\alpha} &= \mathrm{E}(\bar{y}) - \mathrm{E}(\hat{\beta})\bar{x} \\ &= \alpha + \beta\bar{x} - \beta\bar{x} = \alpha, \end{aligned}$$

also ist auch der OLS–Schätzer für die Regressionskonstante erwartungstreu.

Konsistenz von $\hat{\beta}$

Durch Umformung erhält man

$$\hat{\beta} = \beta + \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(u_t - \bar{u})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}.$$

Für $n \to \infty$ konvergiert der Nenner gegen eine Art Varianz von X und der Zähler gegen 0, wenn die X von U unabhängig gesetzt werden. Dazu benötigt man schwer zu formulierende Hilfsannahmen über das 'Setzen' der nicht stochastischen X.

Die Konsistenz benötigt Annahme 4 nicht, wenn X und U unabhängig sind. Auch wenn X ein gelagtes Y ist, kann OLS konsistent sein (unter Annahme 6, siehe unten).

Annahme 5: Fehler homoskedastisch

Annahme 5. Alle Fehler U_t sind identisch verteilt mit endlicher Varianz σ_u^2.

U und nicht nur U_t ist also eine Zufallsvariable, ihre Varianz ist endlich. Bei Querschnittsdaten ist Annahme 5 oft verletzt. *Homoskedastie*=identische Varianz, *Heteroskedastie*=verschiedene Varianz.

Annahme 6: Fehler unabhängig

Annahme 6. Die Fehler U_t sind voneinander statistisch unabhängig.

Ist die Varianz von U_t endlich, sind die U_t damit auch unkorreliert (*keine Autokorrelation*). Bei Zeitreihendaten ist Annahme 6 oft verletzt. *Autokorrelation*=Korrelation einer Variable über die Zeit 'mit sich selbst' (*auto*).

Häufiger Fehler: Nicht die Residuen $\hat{u}_t$ sind unabhängig, sondern die Fehler u_t. OLS–Residuen sind i.a. (leicht) autokorreliert.

OLS wird linear effizient

Unter Annahmen 1–6 lässt sich zeigen, dass die OLS–Schätzer $\hat{\alpha}$ und $\hat{\beta}$ die linearen erwartungstreuen Schätzer mit der kleinsten Varianz sind. **BLUE**=*best linear unbiased estimator*.

Ein *linearer* Schätzer ist als Linearkombination der y_t darstellbar, und das gilt für OLS. OLS ist nicht unbedingt der beste Schätzer (effizient).

Diese (hier unbewiesene) BLUE–Eigenschaft unter Annahmen 1–6 heißt auch *Gauss-Markov–Theorem*.

Annahme 7: Stichprobe groß genug

Annahme 7. Die Stichprobe muss größer sein als die Anzahl der zu schätzenden Regressionskoeffizienten.

Das einfache Modell hat 2 Koeffizienten, also sollten mindestens 3 Beobachtungen vorliegen. Man will auch Eigenschaften der Fehler erheben und nicht eine Gerade durch 2 Punkte legen. Angenehme Annahme: ohne Hypothesentests etc. überprüfbar.

Annahme 8: Normalregression

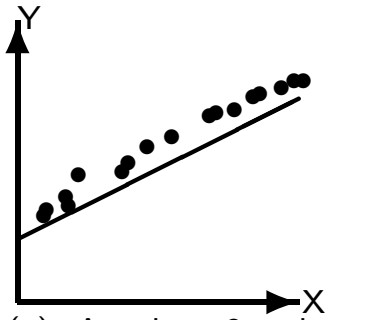
Annahme 8. Alle Fehler U_t sind normalverteilt.

Praktisch, wenn die Welt normalverteilt ist. In kleinen Stichproben nicht überprüfbar, in großen Stichproben oft falsch. Regression unter Annahme 8 heißt *Normalregression*.

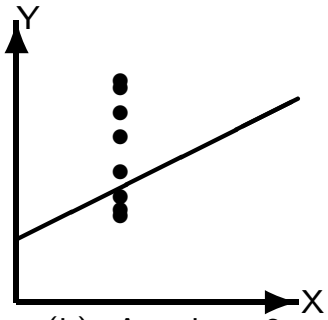
Bedeutung von Annahme 8

- Unter **Annahmen 1–8** ist OLS der **effiziente** Schätzer, er hat unter allen erwartungstreuen Schätzern kleinste Varianz. Meistens ist aber ohnehin die Suche nach nichtlinearen Schätzern bei kleinem n nicht ergiebig;
- Hypothesentests haben unter Annahme 8 exakte Verteilungen (t, F, DW), diese gelten ansonsten nur approximativ.

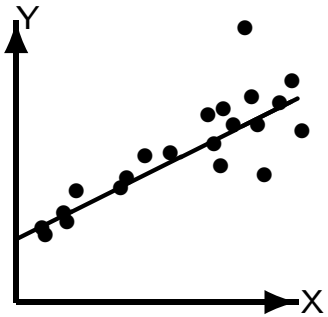
Verletzung einzelner Annahmen in Streudiagrammen



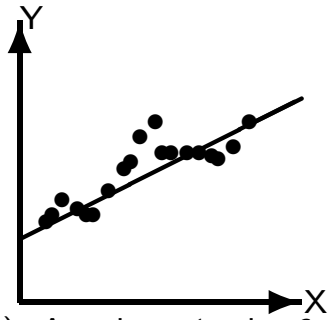
(a): Annahme 3 verletzt



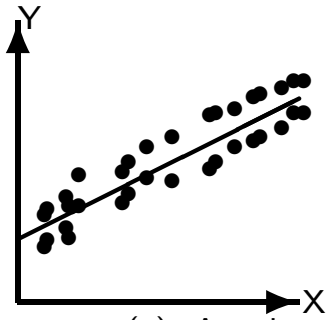
(b): Annahme 2 verletzt



(c): Annahme 5 verletzt



(d): Annahme 1 oder 6 verletzt



(e): Annahme 8 verletzt

Die Varianz von $\hat{\beta}$

Unter Annahmen 1–6 ist OLS der BLUE–Schätzer mit kleinster Varianz. Diese Varianz lässt sich für $\hat{\beta}$ als

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma_u^2}{n\widehat{\mathrm{var}}^*(X)}$$

angeben, mit der n–gewichteten empirischen X–Varianz

$$\widehat{\mathrm{var}}^*(X) = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2.$$

Beweis direkt durch Auswertung von Erwartungsoperationen. Annahmen 5 und 6 gehen ein.

Interpretation der Varianzformel für $\hat{\beta}$

$$\frac{\sigma_u^2}{n\widehat{\mathrm{var}}^*(X)}$$

hängt von drei Faktoren ab:

1. σ_u^2: wenn Fehler weniger streuen, dann sind die Datenpunkte näher an der Regressionsgerade und die Schätzung wird präziser;
2. $\widehat{\mathrm{var}}^*(X)$: wenn der Regressor stark streut, ist das ein gutes und informatives Design und die Schätzung wird präziser;
3. n: größere Stichprobe, bessere Schätzung (Konsistenz).

Die Varianz von $\hat{\alpha}$

Unter Annahmen 1–6 gilt für den OLS–Schätzer der Regressionskonstante

$$\operatorname{var}(\hat{\alpha}) = \frac{\sigma_u^2}{n} \frac{\frac{1}{n}\sum_{t=1}^{n} x_t^2}{\widehat{\operatorname{var}}^*(X)}.$$

Ist $\bar{x}$ nahe 0, also streuen die X um 0, dann ist der Quotient nahe 1 und das Intercept gut schätzbar. Ist das Zentrum der X weit weg von 0, wird die Schätzung weniger präzis.

Die Kovarianz der Koeffizientenschätzer

Unter Annahme 1–6 ergibt sich für die Kovarianz der Schätzer $\hat{\alpha}$ und $\hat{\beta}$

$$\operatorname{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma_u^2}{n} \frac{\bar{x}}{\widehat{\operatorname{var}}^*(X)}.$$

Bei positivem $\bar{x}$ also negativ: tendenziell größeres Intercept wiegt niedrigere Neigung auf. Bei $\bar{x} = 0$ sind die beiden Schätzer unkorreliert.

Schätzung der Fehlervarianz

Die Formeln für Varianz und Kovarianz der OLS–Schätzer sind nicht direkt empirisch anwendbar, weil die Fehlervarianz σ_u^2 nicht bekannt ist. Sollte man die beobachteten OLS–Residuen $\hat{u}_t^2$ zum Schätzen verwenden?

Die Residuen haben eine geringere Varianz als die Fehler, naive Schätzer wie

$$\frac{1}{n}\sum_{t=1}^{n}\hat{u}_t^2, \quad \frac{1}{n-1}\sum_{t=1}^{n}\hat{u}_t^2,$$

unterschätzen σ_u^2 systematisch.

Erwartungstreue Varianzenschätzer

Unter den Annahmen 1–7 ist

$$\hat{\sigma}_u^2 = \frac{1}{n-2} \sum_{t=1}^{n} \hat{u}_t^2$$

ein erwartungstreuer Schätzer für σ_u^2. (Annahme 7 geht ein, sonst Nenner 0.)

$n-2$ im Nenner durch Konzept der 'Freiheitsgrade' zu merken: n Beobachtungen, 2 Parameter werden in der OLS–Regression vorgeschätzt.

$\hat{\sigma}_u^2$ ist *kein* erwartungstreuer Schätzer für die Varianz der *Residuen*.

Die Wurzel $\hat{\sigma}_u$ wird oft 'Standardfehler der Regression' genannt (*standard error of regression*).

Standardfehler der Regressionskoeffizienten

Die geschätzte Varianz von $\hat{\beta}$ ist also

$$\hat{\sigma}^2_\beta = \frac{\hat{\sigma}^2_u}{n\widehat{\mathrm{var}}^*(X)}$$

und unter Annahmen 1–7 erwartungstreu für $\mathrm{var}\hat{\beta} = \sigma^2_\beta$. Analog für σ^2_α und die Kovarianz.

Der Standardfehler oder die geschätzte Standardabweichung ist die Wurzel aus diesem Wert. Dieser Schätzer ist nicht erwartungstreu für σ_β.

Homogene lineare Regression

Es kann attraktiv erscheinen (aus empirischen oder theoretischen Gründen), statt der *inhomogenen* Regression

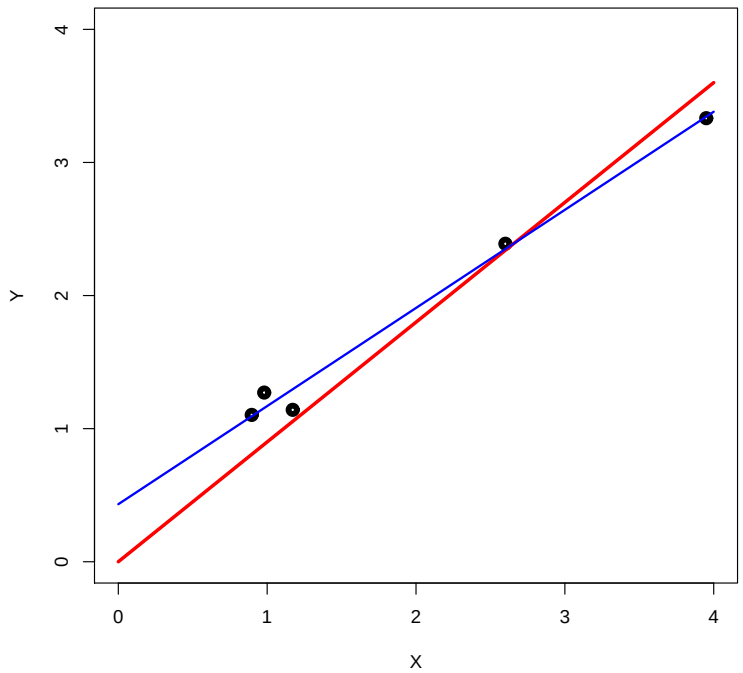
$$y_t = \alpha + \beta x_t + u_t$$

die *homogene* Regression

$$y_t = \beta x_t + u_t$$

zu betrachten, also die Regressionsgerade durch den Ursprung zu zwingen.

Homogene Regression: Beispiel



Blaue Gerade = OLS–Regressionsgerade $\hat{\alpha} + \hat{\beta}X$; rote Gerade = Ergebnis der homogenen OLS–Regression $\check{\beta}X$.

Der homogene OLS–Schätzer

Die Aufgabe, die Summe der vertikalen Abstandsquadrate unter der Bedingung zu minimieren, dass die Regressionsgerade durch den Ursprung geht, lautet mathematisch

$$\min Q_0(\beta) = \sum_{t=1}^{n} (y_t - \beta x_t)^2$$

und hat die Lösung

$$\check{\beta} = \frac{\sum_{t=1}^{n} x_t y_t}{\sum_{t=1}^{n} x_t^2}.$$

Inhomogenes OLS im homogenen Modell

Gelten die Annahmen 2–6 und hat man Annahme 1H

$$y_t = \beta x_t + u_t$$

statt Annahme 1, dann ist der übliche inhomogene OLS–Schätzer $\hat{\beta}$ erwartungstreu, aber nicht BLUE, weil er die Bedingung $\alpha = 0$ nicht berücksichtigt. Alle Varianzenschätzer bleiben weiter gültig, weil (1H) ein Spezialfall von Annahme 1 ist.

Homogenes OLS im inhomogenen Modell

Ist $\alpha \neq 0$ und schätzt man mit $\check{\beta}$, dann ist der Schätzer i.a. biased und inkonsistent wegen

$$\begin{aligned} \mathrm{E}\check{\beta} &= \frac{\sum_{t=1}^{n} \mathrm{E}\{x_t(\alpha + \beta x_t + u_t)\}}{\sum_{t=1}^{n} x_t^2} \\ &= \beta + \alpha \frac{\bar{x}}{\sum_{t=1}^{n} x_t^2}. \end{aligned}$$

Ist $\bar{x} = 0$, streuen die X um 0, ist der 'falsche' Schätzer erwartungstreu.

Ein Spiel gegen die Natur

		wahres Modell	
		Intercept	kein Intercept
geschätztes Modell	mit Intercept	sehr gut	nicht effizient
	ohne Intercept	biased, inkonsistent	sehr gut

Im Zweifelsfall ist das Risiko einer gewissen Ineffizienz geringer, und die inhomogene Regression ist vorzuziehen. Dies ist auch wegen der Kompatibilität und Vergleichbarkeit mit anderen Spezifikationen hier zu empfehlen.

Die t–Statistik

Um die Signifikanz der Regressionskoeffizienten zu eruieren, kann man den Wert von $\hat{\beta}$ mit seinem Standardfehler $\hat{\sigma}_\beta$ vergleichen. Wenn **Annahmen 1–8** gelten, so kann man zeigen, dass die t–Statistik

$$t_\beta = \frac{\hat{\beta}}{\hat{\sigma}_\beta}$$

unter ihrer Nullhypothese '$H_0 : \beta = 0$' t–verteilt ist mit $n - 2$ Freiheitsgraden.

Analog kann man auch '$H_0 : \alpha = 0$' testen mit dem entsprechenden Quotienten t_α, ebenfalls unter H_0 t_{n-2}–verteilt.

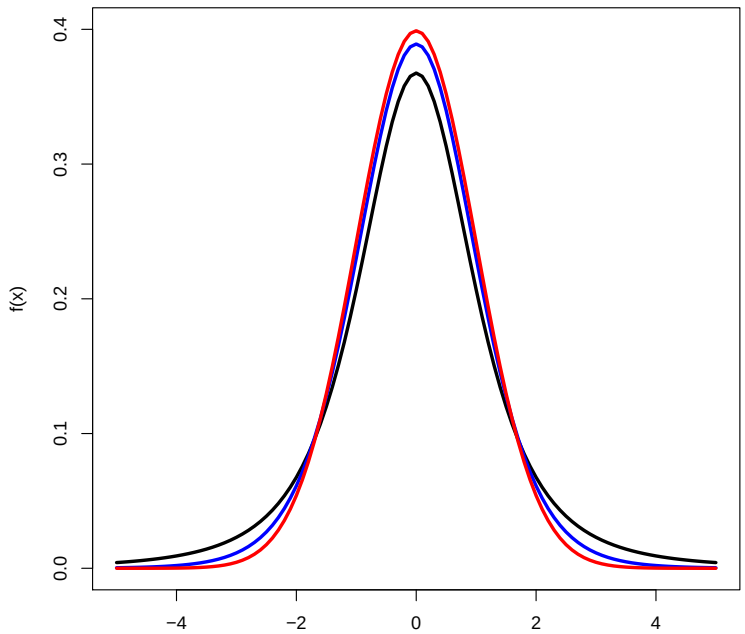
Häufige Fehler: '$H_0 : \hat{\beta} = 0$' und '$H_0 : \beta$ insignifikant' sind keine sinnvollen Nullhypothesen.

Warum ist t_β t–verteilt?

Ein Beweis verläuft in folgenden Schritten:

1. Der Koeffizientenschätzer $\hat{\beta}$ ist normal verteilt $N(\beta, \sigma^2_\beta)$;
2. der normierte Varianzenschätzer $(n-2)\hat{\sigma}^2_\beta/\sigma^2_\beta$ ist χ^2–verteilt mit $n-2$ Freiheitsgraden;
3. Zähler und Nenner sind unabhängig; ist A normal N(0,1) verteilt und B χ^2_m–verteilt und sind A und B unabhängig, dann ist $A/\sqrt{B/m}$ t_m–verteilt.

t–Verteilung und Normalverteilung



Dichte der t_3–Verteilung (schwarz), der t_{10}–Verteilung (blau), und der N(0,1)–Normalverteilung (rot). Für $n > 30$ werden faktisch nur mehr Signifikanzpunkte der N(0,1) im t–Test verwendet.

t–Test wenn Annahme 8 nicht gilt

Wenn Annahme 8 nicht gilt und n 'groß', dann ist die t–Statistik approximativ normal N(0,1) verteilt. Oft wird $n > 30$ als Kriterium für große Stichproben angegeben.

Da die zweiseitigen 5%–Punkte der N(0,1) bei ± 1.96 liegen, betrachten manche ohnehin die t–Statistik nur unter dem Aspekt, ob sie absolut größer oder kleiner 2 ist: 'wenn $|t| > 2$ dann signifikant'.

Gilt Annahme 8 nicht und die Stichprobe ist klein, dann ist der t–Test unverlässlich.

t–Test auf '$H_0 : \beta = \beta_0$'

Man kann daran interessiert sein, ob β ein bestimmter vorgegebener Wert ist (Beispiel: ob die Konsumneigung gleich 1 ist). Die Nullhypothese '$H_0 : \beta = \beta_0$' wird mit der t–Statistik

$$\frac{\hat{\beta} - \beta_0}{\hat{\sigma}_\beta}$$

getestet, die unter H_0 wieder t_{n-2} verteilt ist.

Wird nicht automatisch ausgegeben, kann man aus $t_\beta - \beta_0/\hat{\sigma}_\beta$ auch selbst berechnen.

Statistische Regressionsprogramme und t–Test

- Oft werden $\hat{\beta}$, $\hat{\sigma}_\beta$, und der Quotient t_β ausgegeben, oft sogar der p–Wert des t–Tests;
- Die Software geht von Annahmen 1–8 aus und verwendet meist t–Verteilung, auch wenn Annahme 8 verletzt ist: Vorsicht;
- Der p–Wert bezieht sich auf den zweiseitigen Test gegen '$H_A : \beta \neq 0$': bei einseitigem Test (zum Beispiel '$H_A : \beta > 0$') umrechnen.

Bestimmtheitsmaß: Idee

Gesucht wird eine deskriptive Statistik, die wiedergibt, zu welchem Grade Y durch X linear beschreibbar ist. Der empirische Korrelationskoeffizient

$$\widehat{\text{corr}}(X,Y) = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2}}$$

leistet dies. Er schätzt für Zufallsvariable(!) X und Y die Korrelation

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}},$$

das Maß für lineare Abhängigkeit. Statt dessen verwendet man in der Regression das Quadrat und nennt es R^2.

OLS zerlegt die abhängige Variable

Die OLS–Schätzung zerlegt y_t in einen erklärten und einen unerklärten Anteil:

$$y_t = \hat{\alpha} + \hat{\beta} x_t + \hat{u}_t = \hat{y}_t + \hat{u}_t.$$

$\hat{u}_t$ heißt Residuum. $\hat{y}_t$ heißt systematischer, erklärter Wert für y_t oder Prädiktor.

Varianzzerlegung

Wegen der 2. Normalgleichung gilt

$$\sum_{t=1}^{n} x_t \hat{u}_t = 0,$$

und wegen der 1. Normalgleichung $\sum_{t=1}^{n} \hat{u}_t = 0$. Daher gilt

$$\widehat{\mathrm{cov}}(\hat{U}, X) = \frac{1}{n} \sum_{t=1}^{n} (\hat{u}_t - \bar{\hat{u}})(x_t - \bar{x}) = 0.$$

$\hat{y}_t$ ist aber nur eine Lineartransformation von x_t und daher ebenso mit den Residuen unkorreliert. Daraus folgt

$$\sum_{t=1}^{n} (y_t - \bar{y})^2 = \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{n} \hat{u}_t^2.$$

Notation für die Teile der Varianzzerlegung

Statt

$$\sum_{t=1}^{n}(y_t - \bar{y})^2 = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{n}\hat{u}_t^2$$

schreibt man kurz

$$TSS = ESS + RSS,$$

d.h. *total sum of squares* ist gleich *explained sum of squares* plus *residual sum of squares*. [**Achtung**: die Abkürzungen sind nicht standardisiert, und viele verwenden statt RSS auch SSR, ESS, SSE, statt ESS auch SSE, RSS, SSR.] Gesamtvarianz in erklärten und unerklärten Anteil zerlegt.

Definition des R^2

Das Bestimmtheitsmaß R^2 ist definiert als Quotient

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = 1 - \frac{RSS}{TSS}.$$

Offensichtliche Eigenschaften:

- Da alle Quadratsummen nichtnegativ sind, gilt $0 \leq R^2 \leq 1$;
- $R^2 = 0$ ist sehr 'schlecht', nichts wird erklärt;
- $R^2 = 1$ ist sehr 'gut', die Residuen sind 0 und alle Punkte liegen auf der Regressionsgerade.

Eigenschaften des R^2

- Das so definierte R^2 ist genau das Quadrat der empirischen Korrelation von Y und X (leicht zu zeigen);
- das R^2 ist das Produkt von $\hat{\beta}$ und dem OLS–Koeffizienten in der umgekehrten Regression von X auf Y;
- das R^2 ist als *deskriptive Statistik* gedacht, nicht als Schätzer für die Korrelation von X und Y, die ja wegen Annahme 4 nicht existiert;
- daher gibt es keine verbindliche Regel, ab welchem Wert des R^2 die Regression als 'akzeptabel' gilt (Querschnittsdaten: oft $R^2 \approx 0.3$, Zeitreihendaten: oft $R^2 \approx 0.9$).

Das korrigierte R^2

Fasst man R^2 entgegen der Intention doch als Schätzer einer Korrelation der Zufallsvariablen X und Y auf, dann ist es stark nach oben verzerrt. Die Statistik (nach Wherry und Theil)

$$\bar{R}^2 = 1 - \frac{n-1}{n-2}(1 - R^2)$$

ist auch verzerrt, aber Bias ist kleiner.

- Bedeutung des $\bar{R}^2$ als Hilfsmittel der Modellwahl im multiplen Modell mit k Regressoren, wobei der Nenner durch $n - k$ ersetzt wird;
- $\bar{R}^2$ kann negativ werden: Indiz für schlechte Regression.

Das R^2 als Teststatistik

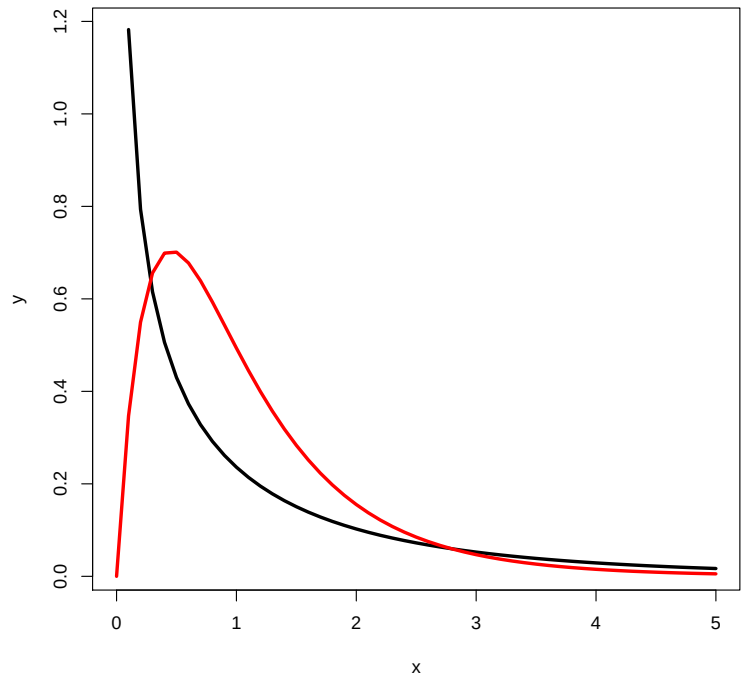
R^2 ist als Funktion der Daten, die Realisierungen von Zufallsgrößen sind nach Annahme 1–8 (zumindest y_t), eine *Statistik*, also eine Zufallsvariable mit Wahrscheinlichkeitsverteilung. Selbst unter '$H_0 : \beta = 0$' hat R^2 aber keine brauchbare oder gut tabellierbare Verteilung.

Die einfache Transformation

$$F_c = \frac{R^2 (n-2)}{1-R^2} = \frac{ESS}{RSS} (n-2)$$

ist aber unter dieser H_0 und Annahmen 1–8 F–verteilt mit $(1, n-2)$ Freiheitsgraden.

Die F–Verteilung



Dichte der Verteilungen $F(1, 20)$ (schwarz) und $F(4, 20)$ (rot). Im einfachen Regressionsmodell treten nur $F(1, .)$–Verteilungen für die Statistik F_c auf.

Warum ist F_c $F(1, n-2)$–verteilt?

- Die $F(m, n)$–Verteilung nach SNEDECOR (das 'F' zu Ehren von FISHER) ist die Verteilung einer Zufallsvariable $(A/m)/(B/n)$, wenn A und B voneinander unabhängig χ^2 verteilt sind mit m bzw. n Freiheitsgraden: daher Zählerfreiheitsgrade m und Nennerfreiheitsgrade n;
- der erklärte und der unerklärte Anteil ESS und RSS sind unabhängig (da unkorreliert und Annahme 8) und unter H_0 ist ESS $\chi^2(1)$ verteilt, RSS ist ohnehin $\chi^2(n-2)$;
- alle F–Verteilungen beschreiben nichtnegative Zufallsvariablen, Verwerfung der H_0 bei großen Werten (immer einseitiger Test).

Bedeutung von F_c im einfachen Regressionsmodell

- F_c wird von allen Regressionsprogrammen ausgegeben, oft mit p–Wert;
- der F-total–Test testet nicht nur die gleiche H_0 wie der t–Test auf β, es gilt auch einfach $F_c = t_\beta^2$, F_c hat also im einfachen Modell keinen Informationswert;
- im multiplen Modell wird F_c wichtig.