Inhaltsverzeichnis

Kapitel 3

Stochastische Regressionsanalyse: Eine erste Intuition

"What we observe is not Nature itself but Nature exposed to our method of questioning." (Werner Heisenberg)

3.1 Deskriptive versus stochastische Regressionsanalyse

Bisher haben wir die Regressionsanalyse einzig und allein dazu verwendet, um eine gegebene Datenmenge kompakt zu beschreiben. Im einführenden Beispiel mit den Gebrauchtautos haben wir z.B. gezeigt, dass der Zusammenhang zwischen Alter und Preis von 40 Gebrauchtwagen relativ gut durch die Regressionsgleichung

$$\widehat{\text{Preis}}_i = 23\,056 - 2\,636\,\text{Alter}_i \qquad (n = 40, R^2 = 0.868) \tag{3.1}$$

beschrieben werden kann (siehe Abbildung 2.1).

Jeder Forscher, der die OLS-Formel auf die 40 Beobachtungen anwendet, wird zum exakt gleichen Resultat kommen, in dieser Beschreibung ist kein Zufallselement enthalten!

Wann immer die Regressionsanalyse ausschließlich dazu verwendet wird, um den Zusammenhang zwischen Variablen für eine fix gegebene Anzahl von Beobachtungen kompakt zu beschreiben, und wir uns nur für diese beobachteten Daten interessieren, spricht man von einer *deskriptiven Regressionsanalyse*.

Tatsächlich wird die Regressionsanalyse eher selten für deskriptive Zwecke eingesetzt. In den meisten Fällen interessieren wir uns nicht für die konkret beobachteten Einzelfälle, sondern wir interpretieren diese Beobachtungen lediglich als Stichprobe aus einer *unbeobachtbaren Grundgesamtheit*, und unser eigentliches Interesse gilt den Zusammenhängen in dieser Grundgesamtheit. Der Zweig der Statistik, der sich mit

© herbert.stocker@uibk.ac.at

Schlüssen von einer beobachtbaren Stichprobe auf eine unbeobachtbare Grundgesamtheit beschäftigt, wird *induktive Statistik* genannt. Wenn wir die Regressionsanalyse als Instrument für induktive Schlussfolgerungen einsetzen sprechen wir von einer induktiven oder *stochastischen Regressionsanalyse*.

Ob eine Regressionsanalyse deskriptiv oder stochastisch ist hängt nicht von den Daten ab, sondern von unserem Erkenntnisinteresse! Die gleichen Beobachtungen können mit Hilfe einer deskriptiven Regressionsanalyse einfach beschrieben werden, oder als Stichprobe aus einer größeren Grundgesamtheit interpretiert werden. Im zweiten Fall wird mit Hilfe der stochastischen Regressionsanalyse versucht, die Information aus der Stichprobe für Rückschlüsse auf die Grundgesamtheit zu nützen.

In diesem Kapitel werden wir versuchen eine erste Intuition für das stochastische Regressionsmodell zu entwickeln. Wir werden uns dabei auf sehr einfache Beispiele beschränken und uns darauf konzentrieren, ein erstes intuitives Verständnis für die teilweise etwas 'tieferen' Konzepte zu vermitteln. In späteren Kapiteln werden wir viele der Begriffe präziser definieren und einige dieser Konzepte verallgemeinern. Aber wir werden sehen, dass die einfache Intuition manchmal erstaunlich weit trägt.

3.2 Grundgesamtheit und Stichprobe

In der stochastischen Regressionsanalyse gehen wir davon aus, dass die Grundgesamtheit unbeobachtbar ist, andernfalls wären wir ja im Bereich der deskriptiven Regressionsanalyse. Wir interessieren uns also für Zusammenhänge in dieser unbeobachtbaren Grundgesamtheit und *vermuten*, dass dieser Zusammenhang durch eine lineare Funktion

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

zumindest approximiert werden kann.

Man beachte, dass wir nicht wie früher in der deskriptiven Analyse $y_i = b_1 + b_2 x_i + e_i$ schreiben, sondern dass wir nun griechische Symbole verwenden. Der Grund ist einfach, in der deskriptiven Analyse konnten wir die Koeffizienten b_1 und b_2 berechnen, da alle interessierenden Beobachtungen verfügbar waren.

Hingegen ist die Grundgesamtheit nicht beobachtbar, deshalb können wir die beiden Koeffizienten β_1 und β_2 *nicht* berechnen! Trotzdem wissen wir, dass die unbekannten Koeffizienten β_1 und β_2 existieren, und dass sie fixe Zahlen sind. Solche unbekannte Größen der Grundgesamtheit werden häufig *'Parameter'* genannt.

Das Wort *'para'-'meter'* verweist aber auf etwas, das über das Messen hinausgeht (wie die Parapsychologie auf etwas verweist, was über die Psychologie hinausgeht). In der Mathematik versteht man darunter spezielle Variablen, die im gegenständlichen Fall als konstant angenommen werden, in anderen Fällen aber variiert werden können (gewissermaßen 'beliebig, aber fest' sind). In diesem Sinne verwenden wir hier den Begriff 'Parameter' für Werte, die in einer unbeobachtbaren Grundgesamtheit als konstant – aber unbeobachtbar – angenommen werden. Eine typische Aufgabe der Statistik ist es solche Parameter aus einer Stichprobe zu schätzen.[1]

[1] Der Gebrauch des Begriffs *Parameter* unterscheidet sich hier übrigens von dem, wie er üblicherweise in der ökonomischen Literatur gebraucht wird. Dort werden unter Parametern häufig exogene Einflussgrößen verstanden, die entweder bekannt (z.B. Steuersätze) oder unbekannt (z.B. Zeitpräferenzrate) sein können.

Die unbekannte Regressionsfunktion, die den Zusammenhang in der (unbeobachtbaren) *Grundgesamtheit* beschreibt, wird im Englischen **'Population Regression Function'** (PRF) genannt. Die deutsche Übersetzung 'Regressionsfunktion der Grundgesamtheit' (oder noch schlimmer, 'Populationsregressionsfunktion') klingt leider etwas holprig, deshalb werden wir häufig das englische Akronym 'PRF' verwenden.

Für die Koeffizienten der PRF verwenden wir generell griechische Symbole ($y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$). Dies signalisiert, dass dies unbeobachtbare Parameter sind.

Gehen wir nun einen Schritt weiter, nehmen wir an, dass wir eine *Zufallsstichprobe*[2] aus der Grundgesamtheit vorliegen haben. Das Beste, was wir in diesem Fall tun können, ist die OLS-Methode auf diese Stichprobenbeobachtungen anzuwenden, und das Resultat als *Schätzung* für die unbekannte PRF (*'Population Regression Function'*) zu verwenden. Genau dies passiert bei der stochastischen Regressionsanalyse.

Eine Regressionsfunktion, die man durch Anwendung der OLS Methode auf Stichprobendaten erhält, wird 'Stichprobenregressionsfunktion' (*'Sample Regression Function'*, SRF) genannt.

Die Unterscheidung zwischen PRF und SRF ist für alles Folgende von zentraler Bedeutung, und weil der Unterschied derart wichtig ist, werden für die beiden unterschiedliche Symbole verwendet.

Aber bevor wir aber auf die Details der Notation eingehen wollen wir zuerst die grundlegenden Ideen anhand eines Beispiels verdeutlichen.

3.2.1 Ein Gedankenexperiment

Wir kehren noch einmal zu dem Beispiel mit den Gebrauchtautos zurück, aber im Unterschied zu früher verwenden wir nun 60 Beobachtungen.

Wir stellen uns vor, dass diese 60 Beobachtungen die interessierende Grundgesamtheit darstellen, z.B. alle Autos, die in einer bestimmten Region in einer bestimmten Zeitperiode zum Verkauf angeboten wurden. Prinzipiell ist die Grundgesamtheit unbeobachtbar, aber im Gedankenexperiment nehmen wir an, dass wir – gleichsam mit überirdischem Wissen ausgestattet – diese Grundgesamtheit kennen.

Mit diesem exquisiten Wissen können wir die PRF nach der OLS Methode berechnen

$$\text{PRF:} \qquad \text{Preis}_i = 23\,081 - 2\,630\,\text{Alter}_i + \varepsilon_i$$

Diese PRF ist in Abbildung 3.1 dargestellt. Den armen 'irdischen' Forschern sind diese 'wahren' Parameter $\beta_1 = 23\,081$ und $\beta_2 = 2\,630$ unbekannt, aber sie interessieren sich brennend dafür.

Angenommen ein potentieller Autokäufer interessiert sich für diesen Zusammenhang in der Grundgesamtheit, hat aber keinen Zugang zu den Daten. Deshalb erhebt er sieben zufällig ausgewählte Beobachtungen (diese Stichprobe mit den 7 Beobachtungen finden Sie in Tabelle 3.1 links). Das beste, was er mit seinen sieben Beobachtungen machen kann, ist darauf die OLS Methode anzuwenden, und zu hoffen, dass der Unterschied zu den 'wahren' Werten der PRF nicht allzu groß sein wird.

[2]Bei einer Zufallsstichprobe hat jedes Element der Grundgesamtheit eine angebbare Wahrscheinlichkeit in die Stichprobe zu gelangen.

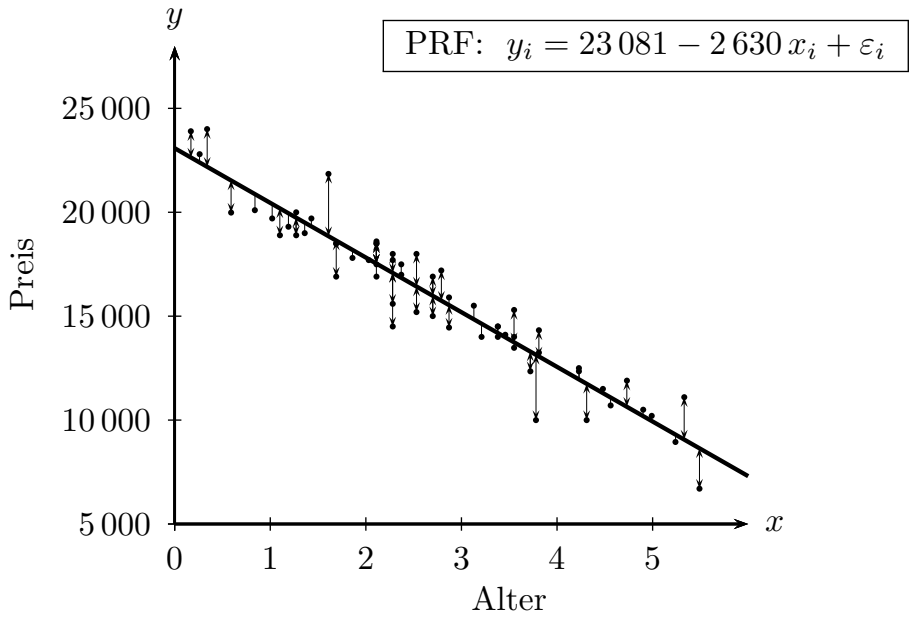


Abbildung 3.1: Die 'Population Regression Function' (PRF) $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ ist für die Forscher unbeobachtbar. β_1 und β_2 sind unbekannte Parameter. Die ebenfalls unbeobachtbaren ε_i werden Störterme genannt.
Daten: `http://www.uibk.ac.at/econometrics/data/auto60.csv`

Da er die OLS Methode auf die Stichprobe anwendet erhält er eine Stichprobenregressionsfunktion (SRF), die OLS Methode liefert

$$\text{SRF 1:} \qquad \text{Preis}_i = 19\,996 - 1\,158\text{Alter}_i + e_i, \quad R^2 = 0.47,\ n = 7$$

Diese SRF 1 ist in Abbildung 3.2 dargestellt.

Da die Koeffizienten dieser SRF fixe Zahlen sind können wir dafür wie früher in der deskriptiven Regressionsanalyse schreiben $\text{Preis}_i = b_1 + b_2\text{Alter}_i + e_i$.

Nun stellen Sie sich vor, eine zweite Forscherin möchte ebenfalls den Zusammenhang in der Grundgesamtheit untersuchen, und auch sie erhebt sieben zufällig ausgewählte Beobachtungen (siehe 'Stichprobe zu SRF 2' in Tabelle 3.1). Die Anwendung der OLS Methode auf diese zweite Stichprobe liefert natürlich andere *Schätzungen* für die Koeffizienten

$$\text{SRF 2:} \qquad \text{Preis}_i = 26\,775 - 4\,022\text{Alter}_i + e_i, \quad R^2 = 0.93,\ n = 7$$

Die Darstellung dieser SRF 2 finden Sie in Abbildung 3.2.

Schließlich sammelt noch ein dritter Interessierter eine Zufallstichprobe SRF 3 (Tabelle 3.1 rechts), und erhält wiederum andere Schätzungen. Abbildung 3.4 zeigt diese SRF (mit den beiden vorhergehenden SRFs).

In Unkenntnis des wahren Wertes von β_2 würde vermutlich jeder der drei Interessierten seine Schätzung b_2 aus der SRF verwenden.

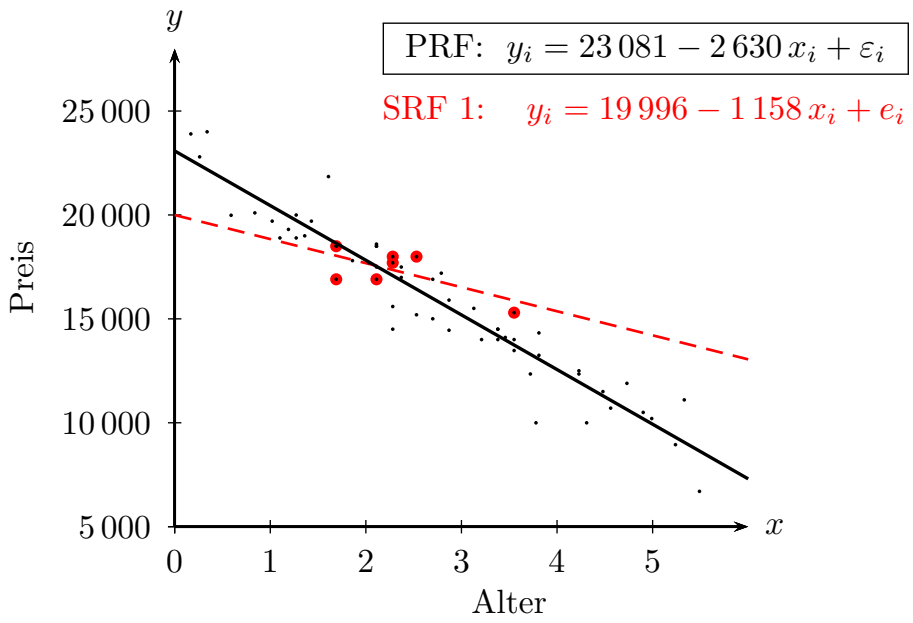


Abbildung 3.2: Eine 'Sample Regression Function' (SRF) für eine beobachtete Stichprobe mit $n = 7$.

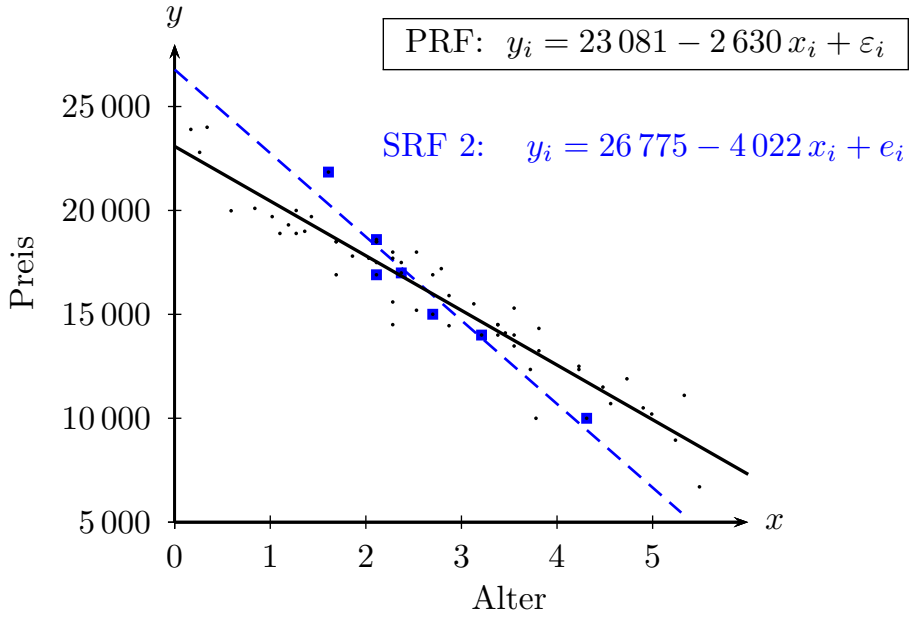


Abbildung 3.3: Eine andere 'Sample Regression Function' (SRF) für eine andere Stichprobe ($n = 7$).

Tabelle 3.1: Die drei den Abbildungen 3.2, 3.3 und 3.4 zugrunde liegenden Stichproben mit $n = 7$.

Stichprobe zu SRF 1

Obs.	Preis	Alter
3	18000	2.28
7	18000	2.53
14	17700	2.28
21	16900	1.69
31	15300	3.55
56	16900	2.11
59	18500	1.69

Stichprobe zu SRF 2

Obs.	Preis	Alter
1	16990	2.37
16	15000	2.70
25	10000	4.31
29	21850	1.61
35	18600	2.11
56	16900	2.11
60	14000	3.21

Stichprobe zu SRF 3

Obs.	Preis	Alter
11	11100	5.33
25	10000	4.31
30	13250	3.81
40	18500	2.11
47	11500	4.48
50	10500	4.90
51	19300	1.19

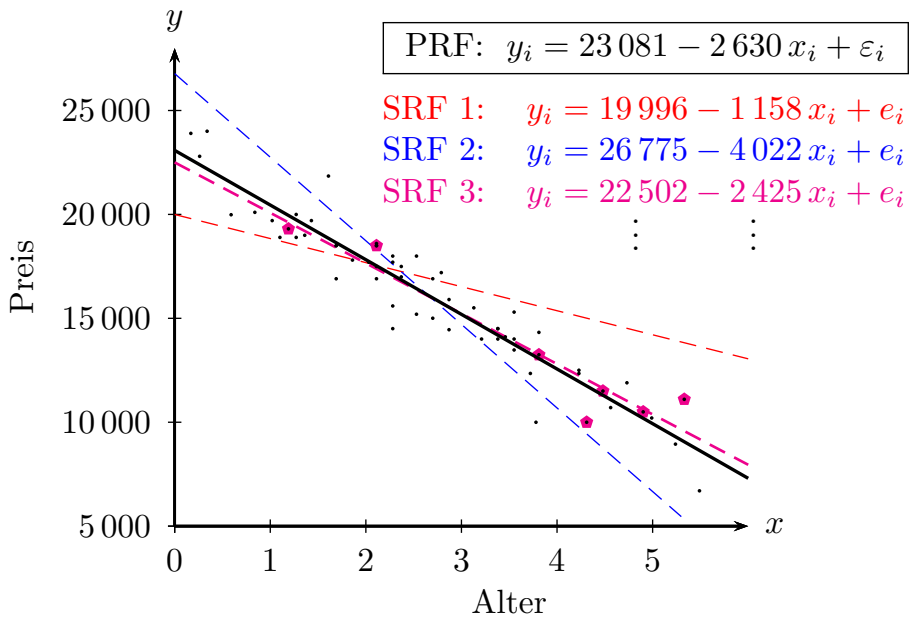


Abbildung 3.4: Und noch eine 'Sample Regression Function' (SRF) für eine andere Stichprobe, $n = 7$.

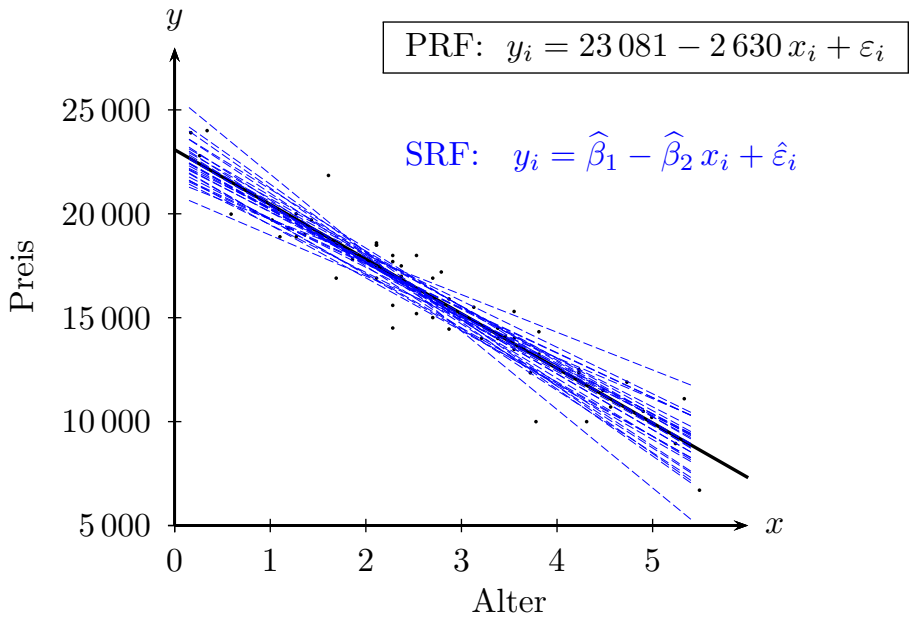


Abbildung 3.5: Viele 'Sample Regression Functions' (SRF) für Stichproben mit $n = 7$. Die *Schätzfunktionen* $\widehat{\beta}_1$ und $\widehat{\beta}_2$ beschreiben als Zufallsvariablen das Ergebnis für alle möglichen Stichproben.

Aber warum sollten wir nach drei Ziehungen stoppen? Prinzipiell können wir diesen Prozess beliebig oft wiederholen, wir können z.B. tausend Zufallsstichproben ziehen, und für jede dieser Stichproben die dazugehörige SRF berechnen. Nachdem dies händisch etwas mühsam wäre lassen wir den Computer die Arbeit machen (den entsprechenden R Programmcode finden Sie im Appendix, Abschnitt 3.A.1, Seite 44).

Tabelle 3.2 zeigt einen kleinen Ausschnitt der 1000 verschiedenen Schätzungen für die Koeffizienten, und Abbildung 3.5 zeigt gemeinsam mit der PRF die ersten 30 SRFs.

Es gibt zwar ein paar extreme Schätzungen[3], aber *im Durchschnitt* liegen wir offensichtlich gar nicht so schlecht, wie die Mittelwerte in Tabelle 3.2 verdeutlichen. Dies ist natürlich kein Zufall, sondern die Konsequenz eines der wichtigsten 'Gesetze' der Statistik, des *Gesetzes der großen Zahl*. Etwas salopp könnten wir es folgendermaßen formulieren: wenn wir mehr und mehr Stichproben ziehen, und jeweils die Mittelwerte der Koeffizienten über mehr Schätzungen berechnen, dann wird sich diese Folge von Mittelwerten schließlich dem 'wahren' Wert nähern.[4]

Auch das zweite große Gesetz der Statistik, der *zentrale Grenzwertsatz*, lässt sich anhand dieses Beispiels demonstrieren. Dazu konzentrieren wir uns auf den Steigungskoeffizienten und zeichnen ein Histogramm der tausend Schätzungen für β_2.

Abbildung 3.6 zeigt dieses Histogramm, und zusätzlich als strichlierte Linie die Gaußsche Glockenkurve, die Normalverteilung. Dass die Normalverteilung der durch

[3] Tatsächlich zeigen Abbildungen 3.2 und 3.3 die beiden extremsten der insgesamt 1000 Schätzungen, um den Unterschied grafisch hervorzuheben.

[4] Etwas genauer gesagt, wenn wir viele Stichproben ziehen, *und jede folgende Stichprobe größer ist als die vorhergehenden*, dann wir die gegen Unendlich gehende Folge der aus diesen Stichproben berechneten Mittelwerte gegen den Mittelwert der Grundgesamtheit konvergieren.

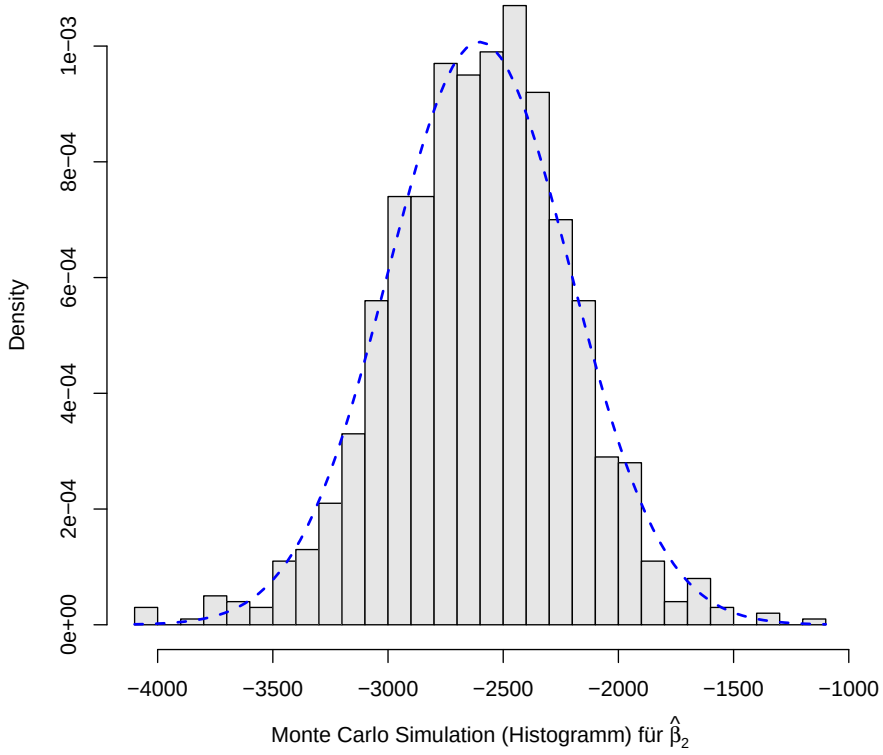


Abbildung 3.6: Histogramm für Steigungskoeffizienten b_2 auf Grundlage von tausend 'Sample Regression Functions' (1000 Stichproben mit jeweils $n = 7$). Die strichlierte Linie zeigt eine Normalverteilung.

Tabelle 3.2: Monte Carlo Simulation: für 1000 Stichproben ($n = 7$) werden Realisationen der Koeffizienten $\widehat{\beta}_1$ und $\widehat{\beta}_2$ berechnet (die extremsten Realisationen, die in Abbildungen 3.2 und 3.3 eingezeichnet sind, sind die Stichproben 33 & 369).
Die 'wahren Werte der Grundgesamtheit sind $\beta_1 = 23081$ und $\beta_2 = 2630$.

Stichprobe	b_1	b_2
1	22476	-2513
2	23525	-2658
3	22502	-2425
⋮	⋮	⋮
33	19996	-1158
⋮	⋮	⋮
369	26775	-4022
⋮	⋮	⋮
999	23327	-2714
1000	23598	-2875
Mittelwert:	22976	-2603

das Histogramm dargestellten empirischen Verteilung so nahe kommt ist natürlich kein Zufall, sondern eine Konsequenz des *zentralen Grenzwertsatzes*.

3.2.2 Notation: Parameter, Zufallsvariablen, und Realisationen

Wenn die Ökonometrie manchen Anfängern schwierig erscheint ist dies nicht zuletzt auf die ungewohnte Notation und häufig mehrdeutig gebrauchten Begriffe zurückzuführen.

Wie **?**, xvi ausführt kann allein der Begriff *'mean'* vier verschiedene Bedeutungen haben, je nachdem, ob er sich auf die Grundgesamtheit oder die Stichprobe bezieht, und ob damit eine Zufallsvariable oder eine Realisation gemeint ist; ein Zustand, den bereits **?** in den 20-iger Jahren des letzten Jahrhunderts beklagte.[5]

Selbst in Lehrbüchern hat sich bisher kein eindeutiger Standard für die Notation herausgebildet, was gerade Einsteiger beim Querlesen in verschiedenen Büchern ziemlich verwirren kann. In diesem Manuskript folge ich in der Notation weitgehend einem Vorschlag von **?**, da dieser meines Erachtens einen guten Kompromiss zwischen interner Konsistenz und guter Lesbarkeit bietet.

Bevor wir uns in dieses Thema stürzen fassen wir noch einmal kurz zusammen: Wir

[5]"in statistics a purely verbal confusion has hindered the distinct formulation of statistical problems; for it is customary to apply the same name, *mean*, *standard deviation*, *correlation coefficient*, etc., both to the true value which we should like to know, but can only estimate, and to the particular value at which we happen to arrive by our methods of estimation" (**?**, S. 311).

haben angenommen, dass wir den Zusammenhang in der Grundgesamtheit durch eine lineare Funktion $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ zumindest approximieren können.

Wir interessieren uns vor allem für die unbekannten Parameter β_1 und β_2, von denen wir nur wissen, dass sie existieren und fixe reelle Zahlen sind.

Dann haben wir aus dieser Grundgesamtheit eine Zufallsstichprobe gezogen. Hier kam das erste Mal der Zufall ins Spiel, *bevor* wir die Stichprobe gezogen hatten konnten wir nicht sagen, welche Beobachtungen die Stichprobe enthalten wird, und welche Koeffizienten wir aus dieser Stichprobe erhalten würden.

Eine solche Stichprobenziehung ist ein Beispiel für ein *Zufallsexperiment*, also ein prinzipiell beliebig oft wiederholbarer Vorgang mit unsicherem Ausgang, dessen mögliche Ausgänge bekannt sein sollen. Bei Zufallsexperimenten gibt es immer ein *vorher* (ex ante) und ein *nachher* (ex post).

Sobald wir die Stichprobe gezogen haben – also nachher – haben wir wieder fixe Zahlen vorliegen; aus einer gegebenen Stichprobe können wir die Koeffizienten berechnen (z.B. die SRF 1: $\text{Preis}_i = 19\,996 - 1\,158\text{Alter}_i + e_i$, Abbildung 3.2).

Nachdem wir diese Zahlen als Resultat der Durchführung *eines* Zufallsexperiments erhalten haben, nennen wir sie *Realisationen*. Diese Realisationen sind wieder feste fixe reelle Zahlen, auf Ebene der Realisationen existiert kein Zufall mehr!

Wenn wir betonen wollen, dass wir von Realisationen sprechen, verwenden wir für die Koeffizienten lateinische Buchstaben, z.B. $y_i = b_1 + b_2 x_i + e_i$.[6]

In der obigen Simulation haben wir tausend verschiedene Realisationen berechnet (vgl. Tabelle 3.2).

Bei Zufallsexperimenten gibt es aber auch ein *vorher*. Bevor wir die Zufallsstichprobe gezogen haben können wir nicht mit Sicherheit sagen, welches Resultat wir erhalten werden. Aber wir könnten – zumindest hypothetisch – für *jede mögliche* Stichprobe den entsprechenden Koeffizienten berechnen.

Eine Funktion, die jeder möglichen Stichprobe einen reellen Zahlenwert zuordnet, nennen wir (etwas salopp) eine *Zufallsvariable*. Zufallsvariablen sind ziemlich komplexe mathematische Gebilde, aber vorerst genügt es zu wissen, dass wir eine Zufallsvariable als eine Abbildung aller möglichen Ausgänge eines Zufallsexperiments in die reellen Zahlen interpretieren können.

In unserem Fall interessieren wir uns für den Steigungskoeffizienten β_2 der PRF, und das Zufallsexperiment ist die Ziehung einer Stichprobe. Die entsprechende Zufallsvariable können wir dann als Funktion interpretieren, die jeder möglichen Stichprobe den dazugehörigen Steigungskoeffizienten zuordnet.[7]

Die Regressionkoeffizienten der SRF sind in dieser *ex ante* Betrachtungsweise (also *vor* der Stichprobenziehung) Zufallsvariablen; in einer *ex post* Betrachtungsweise (also *nach* der Stichprobenziehung) handelt es sich um Realisationen.

[6] Darin weichen wir von der üblichen Notation ab, die häufig für Realisationen und Zufallsvariablen das gleiche Symbol verwendet.

[7] Für eine Zufallsvariablen können wir darüber hinaus angeben, mit welcher Wahrscheinlichkeit eine Ausprägung kleiner ist als eine beliebige reelle Zahl, aber das spielt im Moment noch keine Rolle.

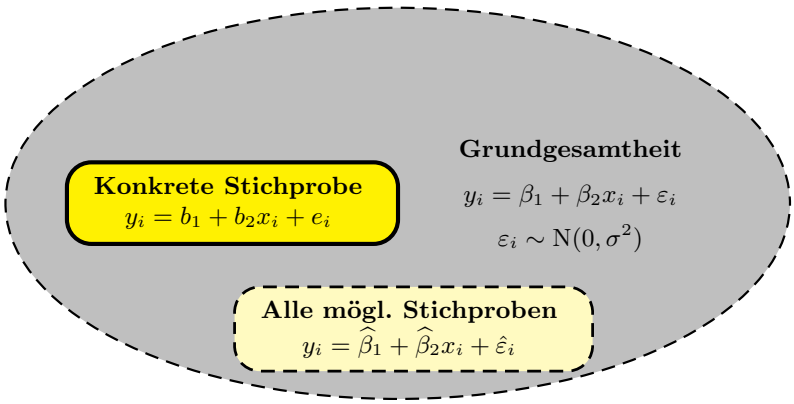


Abbildung 3.7: Grundgesamtheit und Stichprobe

In der Statistik ist es üblich Zufallsvariablen mit Großbuchstaben zu bezeichnen, und Realisationen mit Kleinbuchstaben. Diese Notation hat sich in der Ökonometrie nicht durchgesetzt, unter anderem, weil dies für griechische Symbole manchmal schwierig wäre.

Dagegen werden in der Ökonometrie Zufallsvariablen häufig durch ein Dach über dem entsprechenden griechischen Symbol gekennzeichnet, z.B.

$$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \hat{\varepsilon}_i$$

In dieser Interpretation als Zufallsvariablen werden $\widehat{\beta}_1$ und $\widehat{\beta}_2$ auch *'Schätzfunktionen'* oder kürzer *'Schätzer'* (*estimators*) genannt.

Man beachte den Unterschied:

$\boldsymbol{\beta}_h$ (mit $h = 1, \ldots, k$) sind unbekannte Zahlen (also deterministisch), für die wir uns eigentlich interessieren, aber die wir nicht berechnen können, weil wir die Grundgesamtheit nicht beobachten können.

$\widehat{\boldsymbol{\beta}}_h$ sind spezielle Zufallsvariablen (also stochastisch), die das Ergebnis für alle möglichen Stichprobenziehungen abbilden. Da sie jeder möglichen Stichprobe einen Zahlenwert zuordnen werden sie *Schätzfunktionen* (auch Schätzer oder *'estimator'*) genannt.

$\boldsymbol{b}_h$ sind Realisationen (also deterministisch) und sind das Ergebnis einer Stichprobenziehung. Dies sind fixe Zahlen, denen eine konkrete Stichprobe zugrunde liegt.

Diese Realisationen nennt man auch *Schätzungen* (*'estimates'*).

Wenn wir das Resultat einer empirischen Analyse vorliegen haben handelt es sich dabei um eine Schätzung, um fixe Zahlen, die aus einer konkreten Stichprobe berechnet wurden.

Abbildung 3.7 zeigt die Unterschiede.

Achtung: Einer ebenso alten wie verwirrenden Tradition folgend wird in der Literatur häufig für die Zufallsvariable und die Realisation das gleiche Symbol $\widehat{\beta}$ verwendet,

man muss dann aus dem Zusammenhang selbst erschließen, ob die Zufallsvariable oder eine Realisation gemeint ist.

Diese Tradition ist zwar verwirrend, aber manchmal doch ziemlich praktisch. Warum? Wie wir gleich zeigen werden gelten viele Aussagen sowohl für die Zufallsvariablen als auch für die Realisationen. Da es etwas umständlich wäre die gleichen Aussagen mit unterschiedlichen Symbolen doppelt zu tätigen, wird häufig die Dach-Notation für Zufallsvariable *und* Realisation verwendet.

Wenn die Gefahr eines Missverständnisses nicht allzu groß ist werden wir auch hier die Dach-Notation für Zufallsvariablen *und* Realisationen verwenden, obwohl es sich dabei natürlich um grundlegend verschiedene Dinge handelt!

Wir haben oben betont, dass wir diese Dach-Notation nur für die Koeffizienten und Residuen verwenden, für die Variablen x und y ist es leider noch etwas komplizierter. In der deskriptiven Regressionsanalyse haben wir $\widehat{y}$ für die gefitteten Werte (die systematische Komponente) verwendet. Da in der deskriptiven Regressionsanalyse kein Zufall existiert handelte es sich dabei natürlich um deterministische Größen.

In der stochastischen Regressionsanalyse sind die $\widehat{y}_i$ Zufallsvariablen, wann immer auf der rechten Seite der Regressionsgleichung mindestens eine Zufallsvariable vorkommt, und der Störterm ist in der induktiven Regressionsanalyse immer eine Zufallsvariable.

Obwohl wir am Anfang annehmen werden, dass die x_i deterministisch sind, werden wir später sehen, dass die x_i auch stochastisch (also Zufallsvariablen) sein können. Dies macht eine präzise Notation schwierig, häufig bleibt es der Leserin überlassen aus dem Kontext zu erschließen, ob die Variablen im jeweiligen Zusammenhang deterministisch oder stochastisch sind. Eine Möglichkeit, die sich in der Literatur aber kaum durchgesetzt hat, besteht darin deterministische Variablen zu unterstreichen, z.B. $\underline{x}$ oder $\underline{y}$. Nur wenn wir später explizit darauf hinweisen wollen, dass es sich um deterministische Variablen handelt, werden wir auf diese Notation zurückgreifen.

Störterme und Residuen

Auf eine wichtige Unterscheidung haben wir bisher noch nicht explizit hingewiesen, den Unterschied zwischen Residuen und Störtermen. Den Begriff Residuen haben wir bereits in der deskriptiven Regressionsanalyse definiert und verwendet. Dort handelte es sich um Realisationen, wir konnten sie mit Hilfe der geschätzten Koeffizienten aus den Daten berechnen: $e_i = y_i - b_1 - b_2 x_i$, deshalb haben wir das lateinische e_i als Symbol verwendet.

In der PRF $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ verwenden wir das griechische Symbol ε_i. Weil die Parameter β_1 und β_2 unbekannt sind können auch die ε_i nicht berechnet werden, diese sind ebenso wie β_1 und β_2 unbeobachtbare Größen der Grundgesamtheit.

Um den Unterschied zu den Residuen der SRF deutlich zu machen erhalten die ε_i der PRF einen eigenen Namen, wir nennen die ε_i *Störterme* (*'errors'*).

Im vorhergehenden Beispiel mit einer fix gegebenen Grundgesamtheit sind die Störterme unbeobachtbare Zahlen, also deterministisch, aber in den meisten späteren Spezifikationen werden wir annehmen, dass die ε_i (mit $i = 1, \ldots, n$) Zufallsvariablen sind.

Der Grund ist einfach, wir werden uns später vorstellen, dass die beobachteten Daten das Ergebnis eines *Datengenerierenden Prozesses* (DGP) sind. Wir stellen uns vor, dass die systematische Komponente dieses DGP wieder durch $\widehat{y}_i = \beta_1 + \beta_2 x_i$ beschrieben wird, und dass diese systematische Komponente durch eine nichtsystematische Komponente ε_i *gestört* wird, daher die Bezeichnung *Störterme*.

Kommen wir zurück zur Stichprobe, sobald wir eine Stichprobe gezogen haben – also *nachher* – können wir daraus wie in der deskriptiven Statistik die Realisationen der Residuen e_i berechnen. Dabei handelt es sich natürlich um einfache reelle Zahlen.

Aber wie früher bei den Koeffizienten $\widehat{\beta}$ können wir uns mental wieder in den Zustand *vor* der Ziehung der Zufallsstichprobe versetzen. In dieser *ex-ante* Betrachtungsweise wird jeder möglichen Stichprobe ein Residuenvektor zugeordnet, in diesem Fall handelt es sich um Zufallsvariablen (da die Stichprobe n Beobachtungen enthält handelt sind es um einen Vektor mit n Zufallsvariablen). Für diese speziellen Zufallsvariablen verwenden wir wieder die Dach Notation und schreiben $\hat{\varepsilon}_i$ (mit $i = 1, \ldots, n$).

Es ist unglücklich, dass für die (*ex post*) Realisationen e_i und für die (*ex ante*) Zufallsvariablen $\hat{\varepsilon}_i$ der gleiche Begriff *Residuen* verwendet wird. Da in der Literatur für beide häufig auch das gleiche Symbol $\hat{\varepsilon}_i$ verwendet wird, muss die Leserin meist aus dem Kontext erschließen, ob es sich dabei um die Realisationen oder Zufallsvariablen handelt.

Abbildung 3.8 zeigt eine PRF (*'population regression function'*) und eine SRF (*'sample regression function'*). Die Störterme der PRF $\varepsilon_i = y_i - \widehat{y}_i := y_i - \beta_1 - \beta_2 x_i$ sind ebenso unbeobachtbar wie die Parameter β_1 und β_2 der Grundgesamtheit.

Bei dem Residuum müssen wir unterscheiden: wenn die dargestellte SRF eine Realisation ist wie z.B. die SRF 1 in Abbildung 3.2 (Seite 5), dann ist das Residuum ebenfalls eine Realisation e_i.

Wenn wir uns aber vorstellen, dass die dargestellte SRF nur symbolisch für eine mögliche SRF aus einer große Anzahl von SRFs steht, wie z.B. in Abbildung 3.5 (Seite 7), also in einer ex ante Perspektive, dann ist das Residuum eine Zufallsvariable $\hat{\varepsilon}_i$.

Dies gilt auch für die Berechnung der Regressionskoeffizienten. Wir haben im Kapitel zur deskriptiven Regressionsanalyse bereits die OLS Formeln hergeleitet.

Genau das gleiche können wir auch für die Zufallsvariablen machen, wir minimieren die Quadratsumme der 'ex ante' Residuen

$$\min_{\widehat{\beta}_1, \widehat{\beta}_2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{\widehat{\beta}_1, \widehat{\beta}_2} \sum_{i=1}^{n} (y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i)^2$$

In diesem Fall gelten die Restriktionen der Bedingungen erster Ordnung

$$\begin{aligned}
\frac{\partial \sum_i \hat{\varepsilon}_i^2}{\partial \widehat{\beta}_1} &= 2 \sum_i \underbrace{\left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i\right)}_{\hat{\varepsilon}_i} (-1) = 0 \quad \Rightarrow \quad \sum_i \hat{\varepsilon}_i = 0 \\
\frac{\partial \sum_i \hat{\varepsilon}_i^2}{\partial \widehat{\beta}_2} &= 2 \sum_i \underbrace{\left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i\right)}_{\hat{\varepsilon}_i} (-x_i) = 0 \quad \Rightarrow \quad \sum_i x_i \hat{\varepsilon}_i = 0
\end{aligned}$$

nur für die Zufallsvariablen $\hat{\varepsilon}_i$ der SRF, aber nicht notwendigerweise für die Störterme ε_i der PRF! Deshalb können wir nicht länger garantieren, dass auch die Summe

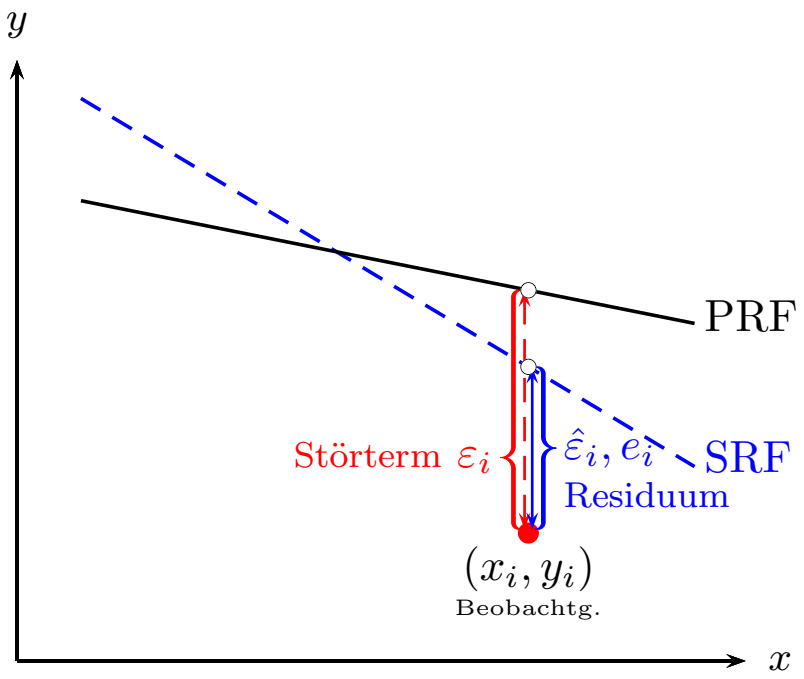


Abbildung 3.8: Die Störterme ε_i der PRF sind unbeobachtbar. Für eine Realisation einer SRF können die Residuen e_i berechnet werden, auch diese sind Realisationen. In einer ex ante Betrachtungsweise sind die Residuen $\hat{\varepsilon}_i$ Zufallsvariablen.

der Störterme der Grundgesamtheit gleich Null ist, und noch wichtiger, dass auch die Störterme der Grundgesamtheit unkorreliert sind mit der erklärenden x-Variable. Dies wird später noch von Bedeutung sein.

Als Lösung dieses Minimierungsproblems erhalten wir natürlich die gleichen Formeln wie früher, nur dass sie in dieser ex ante Perspektive Zufallsvariablen sind und als *Schätzfunktionen* interpretiert werden.

$$\widehat{\beta}_2 = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(x)}, \qquad \widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x} \tag{3.2}$$

wobei das Dach über dem 'cov' bzw. 'var' Operator ausdrücken soll, dass es sich – im Unterschied zur Varianz der Grundgesamtheit – um die Stichprobenkovarianz bzw. -varianz handelt, die in einer ex ante Betrachtungsweise ebenfalls Zufallsvariablen sind.

Die Schätzfunktion $\widehat{\beta}_2 = \widehat{\mathrm{cov}}(x,y)/\widehat{\mathrm{var}}(x)$ ordnet als Zufallsvariable wieder jeder möglichen Stichprobe eine reelle Zahl zu, während die Realisation $b_2 = \mathrm{cov}(x,y)/\mathrm{var}(x)$ einer konkreten Stichprobe genau eine Zahl zuordnet.

Welche der beiden Sichtweisen gerade zutreffend ist hängt vom Kontext ab, aber in den allermeisten Fällen interessieren wir uns für die ex ante Perspektive, also die Zufallsvariablen. Deshalb wird – wenn die Aussagen für die Zufallsvariablen *und* für die Realisationen gelten – meist die Notation für die Zufallsvariablen ($\widehat{\beta}$, $\hat{\varepsilon}_i$) als 'Default Symbol' gewählt.

Fassen wir noch einmal zusammen: Zufallsvariablen beziehen sich immer auf einen interessierenden Aspekt eines zugrunde liegenden Zufallsexperiments; solche Zufallsvariablen nennen wir stochastisch. Im Gegensatz dazu sind deterministische Größen fixe Zahlen, hinter denen kein Zufallsexperiment steht.

1. PRF (*'population regression function'*):

   $$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

   Unser Interesse gilt den Parametern β_1 und β_2, diese sind feste, aber unbeobachtbare reelle Zahlen. Die *Störterme* ε_i können je nach Spezifikationen fixe Zahlen (also deterministisch) oder Zufallsvariablen sein, und y_i ist eine Zufallsvariable wann immer ε_i (oder x_i) eine Zufallsvariable ist.

2. SRF *vor* Durchführung des Zufallsexperiments (ex ante):

   $$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \hat{\varepsilon}_i$$

   $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\hat{\varepsilon}_i$ und y_i (für $i = 1, \ldots, n$) sind Zufallsvariablen; x_i nehmen wir vorläufig deterministisch an, könnte aber auch eine Zufallsvariable sein.

3. SRF *nach* Durchführung des Zufallsexperiments (ex post):

   $$y_i = b_1 + b_2 x_i + e_i$$

   b_1, b_2, e_i und y_i (für $i = 1, \ldots, n$) sind ebenso wie die x_i Realisationen, also deterministische Größen (auf Ebene der Realisationen existiert kein Zufall!).

Wie mehrfach erwähnt unterscheiden viele Lehrbücher in der Notation auf Stichprobenebene nicht zwischen Zufallsvariablen und Realisationen, sondern verwenden für diese beiden völlig verschiedene Dinge das gleiche Symbol.

Monte Carlo Simulationen und Stichprobenkennwertverteilungen

Nachdem wir uns jetzt ziemlich ausführlich mit der Notation beschäftigt haben kommen wir noch einmal zurück zu unserem Gedankenexperiment mit den Gebrauchtautos. Wir haben dort wiederholt Stichproben aus einer gegebenen Grundgesamtheit gezogen, und für jede dieser Stichproben eine SRF berechnet.

Tatsächlich war dies bereits eine sehr einfache Monte Carlo Simulation, wir haben den Computer sehr oft ein *Zufallsexperiment* durchführen lassen (in unserem Fall Stichprobenziehungen), und für jede dieser Stichproben einen interessierenden *Stichprobenkennwert* (in unserem Fall den Steigungskoeffizienten) berechnet.

Diese tausend Schätzungen des Steigungskoeffizienten b_2^r (das hochgestellte $r = 1, \ldots, 1000$ bezeichnet die r-te Stichprobenziehung) für den interessierenden 'wahren' Parameter β_2 haben wir schließlich in einem Histogramm dargestellt (Abbildung 3.6, Seite 8).

Dieses Histogramm zeigt uns, dass von den 1000 Schätzungen nur wenige einen kleineren Wert als -3500 oder einen größeren Wert als -1500 lieferten. Daraus

könnten wir schließen, dass es eher *unwahrscheinlich* ist, dass der 'wahre' Wert β_2 noch extremer ist.

Dieses Histogramm der Stichprobenkennwerte kann man als empirische Simulation einer theoretischen *Stichprobenkennwertverteilung* (oder einfacher 'Stichprobenverteilung', *'sampling distribution'*) interpretieren.

Unter einer Stichprobenkennwertverteilung verstehen wir hier die theoretische Verteilung einer Schätzfunktion, in diesem Fall der Zufallsvariable $\widehat{\beta}_2$. Den Vorgang der wiederholten Stichprobenziehungen nennt man *'repeated sampling'* (die Vorgangsweise ist in Abbildung 3.9 symbolisch dargestellt).

Die Monte Carlo Simulation haben wir hier nur zur Illustration einer Stichprobenkennwertverteilung herangezogen, als pädagogisches Instrument. In der Realität haben wir natürlich meist nur *eine einzige* Stichprobe zur Verfügung, aus der wir *eine* Schätzung für die Parameter der Grundgesamtheit berechnen. Aber das Gedankenexperiment mit den wiederholten Stichprobenziehungen zeigt uns, dass wir unsere Schätzung als eine Realisation aus einer Stichprobenkennwertverteilung interpretieren können, und auf Grundlage dieser Stichprobenkennwertverteilung können wir später statistische Hypothesentests entwickeln.

Stichprobenkennwertverteilungen haben meist, d.h. unter wenig strengen Annahmen, zwei ganz erstaunliche Eigenschaften, die bereits im Histogramm (Abbildung 3.6) ersichtlich sind:

1. Offensichtlich liegt der Mittelwert der vielen Schätzungen sehr nahe beim 'wahren' Wert der Grundgesamtheit. Dies ist kein Zufall, sondern kann als eine Folge des Gesetzes der großen Zahl interpretiert werden. Das **Gesetz der großen Zahl** besagt sehr vereinfacht, dass unter sehr allgemeinen Bedingungen der Mittelwert einer großen Zahl von Zufallsvariablen sich *mit steigendem Stichprobenumfang* an den wahren Wert der Grundgesamtheit annähert.[8]

2. Außerdem erkennt man, dass die Verteilung der Schätzwerte einer Glockenform ähnelt. Auch dies ist kein Zufall, sondern eine Folge des **Zentralen Grenzwertsatzes**. Der zentrale Grenzwertsatz besagt vereinfacht, dass die Summe einer großen Zahl von unabhängigen, identisch verteilten, zentrierten und normierten Zufallsvariablen gegen die Standardnormalverteilung konvergiert, unabhängig von der Verteilung der Grundgesamtheit. Dies erklärt u.a. die Sonderstellung der Normalverteilung.

Im nächsten Abschnitt werden wir zeigen, dass man Stichprobenkennwertverteilungen häufig allgemein aus theoretischen Überlegungen ohne Zuhilfenahme von Monte Carlo Simulationen herleiten kann.

Trotzdem spielen Monte Carlo Simulationen in der Forschung auch heute noch eine wichtige Rolle, vor allem wenn es darum geht, die Eigenschaften komplizierterer Schätzfunktionen zu bestimmen.

[8] In diesem Fall stellen wir uns vor, dass die Zahl der Stichprobenziehungen gegen unendlich geht, und aus jeder Stichprobe ein gewünschter Koeffizient (z.B. b_2) berechnet wird, wie z.B. in Tabelle 3.2 (Seite 9). Für das Gesetz der Großen Zahl stellen wir uns vor, dass fortlaufend neue Durchschnitte über alle früher gezogenen Koeffizienten berechnet werden. Die Folge dieser Durchschnitte sollte sich mit steigender Anzahl der Ziehungen dem wahren Wert annähern. Das heißt, wenn wir z.B. eine Million Stichproben gezogen hätten, und über diese Million realisierter b_2 den Durchschnitt berechnet hätten, würden wir eine genauere Schätzung erwarten als der in Tabelle 3.2 angegebene Durchschnitt über die ersten 1000 Stichproben.

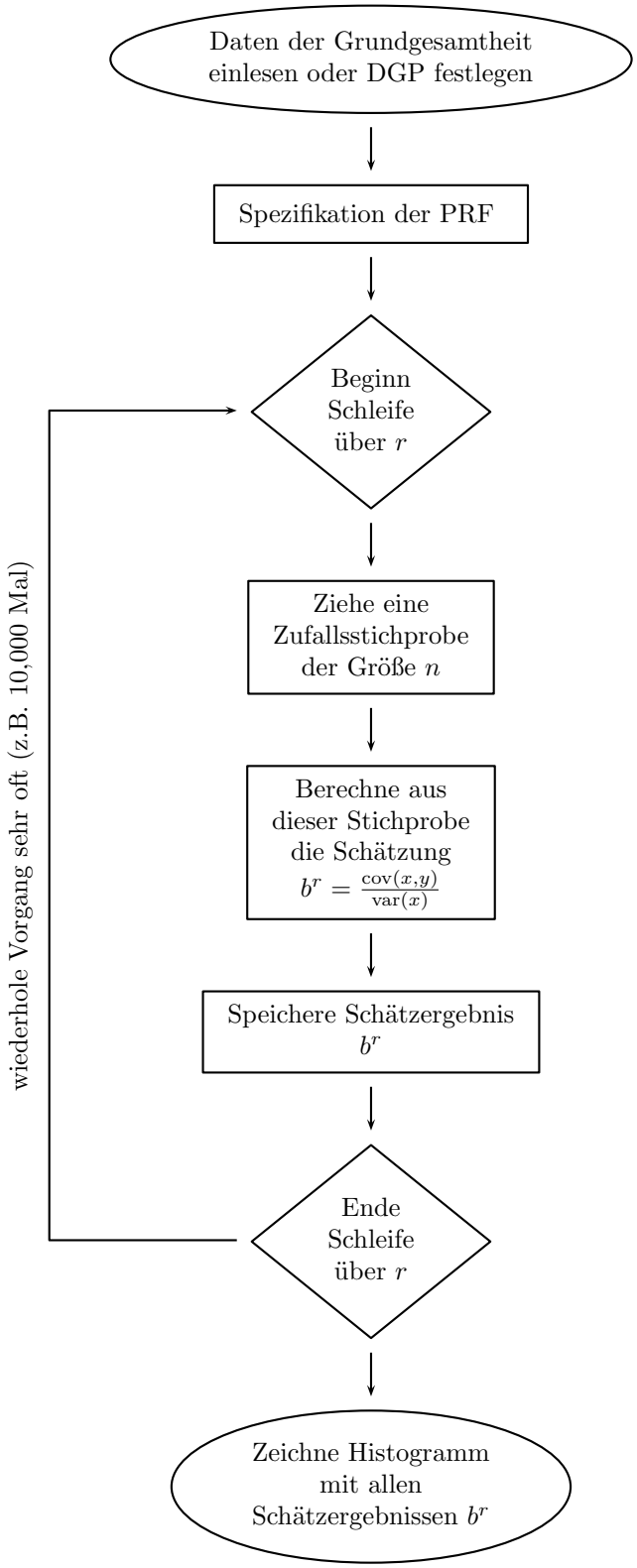


Abbildung 3.9: Wiederholte Stichprobenziehungen aus einer Grundgesamtheit (Monte Carlo Simulation).

3.2.3 Ein Beispiel aus der Statistik: Schätzung eines Anteils

Die grundlegenden Ideen unseres weiteren Vorgehens lassen sich am besten anhand eines einfachen Beispiels aus der Statistik erläutern, nämlich der Ziehung von schwarzen und weißen Kugeln aus einer Urne, wie sie in Abbildung 3.10 dargestellt ist.

Beginnen wir mit der Urne links oben in Abbildung 3.10. Die schwarzen Kugeln könnten z.B. für an einer bestimmten Krankheit erkrankten Personen stehen, und weiße Kugeln für gesunde Personen; oder die schwarzen Kugeln für Unternehmen, die von einer bestimmten Steuer betroffen sind, weiße Kugeln für andere Unternehmen, usw.

Wir interessieren wir uns für den Anteil der 'schwarzen' Gruppe in der Grundgesamtheit, und da wir mit Zurücklegen ziehen ändert sich der Anteil der schwarzen Kugeln nicht mit den Ziehungen. Wir können uns auch vorstellen, dass wir aus einer unendlich großen Grundgesamtheit ziehen.

Unsere Forscherin kennt diesen Anteil nicht, obwohl dieser Anteil gegeben und eine fixe Zahl ist, also ein Parameter, und sie möchte aus einer Zufallsstichprobe eine möglichst genaue Schätzung dieses Anteils ermitteln.

Falls es sich um Beobachtungsdaten handelt und wir keine (experimentelle) Kontrolle über die Urne haben müssen wir vorausschicken, dass wir implizit bereits einige sehr strenge Annahmen getroffen haben, denn woher konnten wir wissen, dass sich in der Urne tatsächlich nur weiße und schwarze Kugeln befinden, und dass nicht irgendein Spaghettimonster heimlich Kugeln in die Urne hinein schmuggelt oder daraus entfernt? Diese wichtigen Fragen betreffen die *Spezifikation*, wir haben *angenommen*, dass unser Problem durch eine Urne mit ausschließlich weißen und schwarzen Kugeln und einem konstant bleibenden Anteil beschrieben werden kann, und die weiteren Schlussfolgerungen sind nur korrekt, wenn diese Annahmen über den Datengenerierenden Prozess zutreffend sind.

Achtung: Unter einem Datengenerierenden Prozess versteht man in der Regel den Gesamtprozess von der Entstehung der Daten bis zum Mechanismus der dafür sorgt, dass die Beobachtungen in der Stichprobe landen (Selektion). Da wir hier nur echte Zufallsstichproben annehmen fokussieren wir hier v.a. auf den Mechanismus hinter der Entstehung der Daten. Die viel schwierigeren Selektionsprobleme wollen wir vorläufig ausklammern.

Unsere Forscherin zieht eine Zufallsstichprobe mit sechs Kugeln. Auch dies ist in praktischen Fällen keineswegs so einfach wie es klingt, echte Zufallsstichproben existieren hauptsächlich in Märchen- und Lehrbüchern, in vielen praktischen Fällen handelt es sich eher um *'convenience samples'* als um echte Zufallsstichproben. Aber sehen wir von diesen praktischen Problemen mal ab und nehmen an, dass es sich bei der Stichprobe mit den sechs Kugeln (in Abbildung 3.10 links unten) wirklich um eine echte Zufallsstichprobe handelt.

Wir können nun den Farben dieser Kugeln Zahlen zuordnen, z.B. 1 für schwarz und 0 für weiß, und *in der Stichprobe* den Anteil schwarzer Kugeln p für diese Stichprobe ausrechnen. Dieser Anteil in der Stichprobe ist eine *Realisation*, und wir können $p = 2/6$ als *Schätzung* für den unbeobachtbaren 'wahren' Anteil π in der Urne verwenden.

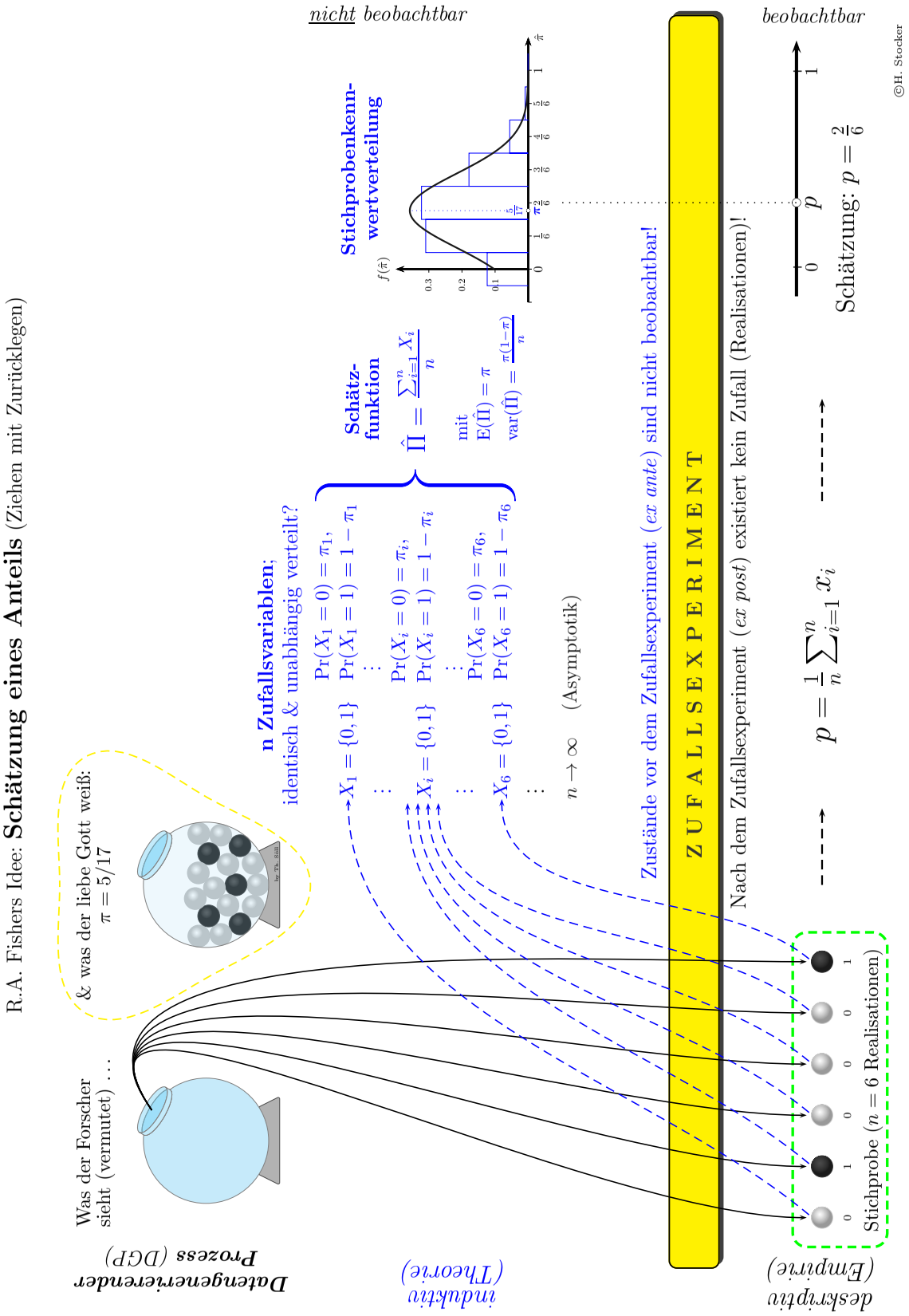


Abbildung 3.10: Ein Urnenmodell.

Die folgende Idee von R.A. Fisher (1890 – 1962) geht aber weit darüber hinaus.

Angenommen, wir ziehen die erste Kugel, verdecken aber die Farbe der Kugel. Mit welcher *Wahrscheinlichkeit* wird diese Kugel schwarz sein? Das können wir natürlich nicht wissen, bevor wir die Farbe gesehen haben, aber wir werden zu Recht vermuten, dass diese Wahrscheinlichkeit gleich dem unbekannten *Anteil* schwarzer Kugeln in der Urne π ist.

In anderen Worten, wir interpretieren die Farbe *jeder einzelnen* Kugel als Zufallsvariable. Hinter der ersten Kugel steht die Zufallsvariable X_1 mit

$$X_1 = \begin{cases} 1 & \text{wenn Kugel schwarz} \\ 0 & \text{sonst} \end{cases}$$

und die unbekannte Wahrscheinlichkeit für eine schwarze Kugel ist $\Pr(X_1) = \pi_1$.

Genau die gleichen Überlegungen gelten für X_2 und alle andern Kugeln der Stichprobe.[9] Eine Stichprobe der Größe n gibt uns also n Zufallsvariablen.

Mit dieser Überlegung treten wir also gewissermaßen hinter unser Zufallsexperiment 'Stichprobenziehung' zurück und untersuchen *alle möglichen* Realisationen. Deshalb werden diese Zufallsvariablen manchmal als *Stichprobenraum* bezeichnet, da sie gewissermaßen eine mathematische Abbildung aller möglichen Ergebnisse der Stichprobenziehungen symbolisieren.

Da jede einzelne Kugel nur schwarz oder weiß sein kann ist jede dieser sechs einzelnen Zufallsvariablen Bernoulli verteilt. Die erste Kugel ist mit einer unbekannten Wahrscheinlichkeit π_1 weiß, und mit der Gegenwahrscheinlichkeit $1 - \pi_1$ schwarz; die zweite Kugel mit einer ebenfalls unbekannten Wahrscheinlichkeit π_2 weiß, und mit $1 - \pi_2$ schwarz, usw. bis zur sechsten Kugel mit Wahrscheinlichkeit π_6.

i.i.d. **Annahmen:** Hier kommen zwei zentrale *Annahmen* ins Spiel. Aufgrund der Struktur des Zufallsexperiments, nämlich Ziehungen aus einer Urne *mit Zurücklegen*, erwarten wir, dass die Wahrscheinlichkeit *für jede Kugel gleich groß* ist, also $\pi_1 = \pi_2 = \cdots = \pi_6$, oder in anderen Worten, dass die sechs Zufallsvariablen *identisch verteilt* sind.

Die zweite wichtige Annahme ist die *Unabhängigkeit* (*independence*) der Ziehungen, das Resultat einer Ziehung soll z.B. nicht davon abhängen, welche Farben in früheren Ziehungen gezogen wurden. In diesem Beispiel ist diese Annahme vernünftig, weil wir mit Zurücklegen gezogen haben. Hätten wir ohne Zurücklegen gezogen, dann hätte sich das Farbenverhältnis in der Urne mit jeder Ziehung geändert, und die Annahme der Unabhängigkeit wäre verletzt. Man beachte, dass sich diese Annahmen auf die Grundgesamtheit, bzw. den Datengenerierenden Prozess, beziehen!

Fassen wir zusammen, anstelle der Stichprobe mit den sechs Realisationen betrachten wir jetzt die sechs 'dahinterliegenden' Zufallsvariablenn, von denen wir annehmen, dass sie identisch und unabhängig (i.i.d.) Bernoulli verteilt sind.

Im nächsten Schritt können wir für diese sechs Zufallsvariablen eine *Schätzfunktion* $\hat{\Pi}$ ermitteln, also eine neue Zufallsvariable, die wir in diesem Fall einfach erhalten,

[9] Wie in der Statistik üblich verwenden wir hier Großbuchstaben für Zufallsvariablen.

indem wir über die sechs Zufallsvariablen mitteln[10]

$$\hat{\Pi} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Jetzt wird es spannend. Aus der theoretischen Statistik wissen wir, dass die Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben, binomial verteilt ist. Das bedeutet, wir können die theoretische Verteilung unserer Schätzfunktion bestimmen!

Diese Verteilung der Schätzfunktion nennen wir eine *Stichprobenkennwertverteilung*. Im Unterschied zu den empirischen Simulationen der Monte Carlo Simulationen können wir – zumindest für einfache Fälle – diese Stichprobenkennwertverteilung allgemein herleiten sowie deren Momente (v.a. Erwartungswert und Varianz) bestimmen und die Eigenschaften dieser Momente (z.B. Erwartungstreue, Effizienz und Konsistenz) untersuchen. Mit diesen Fragen werden wir uns im nächsten Kapitel beschäftigen, und in weiterer Folge wird uns dieses Wissen die theoretischen Grundlagen zur Durchführung von Hypothesentests liefern.

Noch eine zentrale Einsicht ist wichtig. Die theoretische Statistik lehrt uns, dass die Binomialverteilung für $n \to \infty$ gegen die Normalverteilung konvergiert (Satz von Moivre-Laplace). Wenn wir mehr und mehr Kugel ziehen, wird bei der Schätzfunktion über mehr und mehr Zufallsvariablen gemittelt, und die Binomialverteilung nähert sich mehr und mehr einer Normalverteilung an. Solche Überlegungen gehören zur asymptotischen Analyse, die das Grenzverhalten von Funktionen untersucht und in der fortgeschritteneren Ökonometrie eine zentrale Rolle spielt.

Zur Veranschaulichung könnten wir wieder eine Monte Carlo Simulation durchführen, indem wir sehr viele Stichproben ziehen, für jede dieser Stichproben den *empirischen* Anteil p berechnen, und schließlich all diese Anteile in einem Histogramm darstellen. In Abbildung 3.10 würde dies bedeuten, dass wir rechts unten nicht die einzelne Schätzung $p = \frac{2}{6}$ auf der Zahlenlinie hätten, sondern ein Histogramm erhalten würden.

Dieses Histogramm wäre eine empirische Annäherung and die unbeobachtbare *theoretische* Stichprobenkennwertverteilung darüber.

Wozu benötigen wir diese *theoretische* Stichprobenkennwertverteilung? Tatsächlich beobachten wir fast immer nur eine einzige Stichprobe, die uns eine einzige Schätzung liefert. Aber die Theorie sagt uns, dass es sich dabei um eine Realisation aus dieser Stichprobenkennwertverteilung handelt, und dieses Wissen können wir für die Beurteilung unserer Schätzung nützen.

[10] Generell werden Schätzfunktionen mit Hilfe bestimmter Verfahren ermittelt, die wichtigsten sind die bereits bekannte OLS Methode, die Maximum Likelihood Methode und die Methode der Momente. Mehr dazu später.

3.3 Die PRF und die bedingte Erwartungswertfunktion

Für die bisherigen Erläuterungen des stochastischen Regressionsmodells haben wir auf einen aus der einführenden Statistik bekannten Ansatz zurück gegriffen, Ziehung von Zufallsstichproben aus einer gegebenen Grundgesamtheit. Im Rest dieses Kapitels versuchen wir nun eine Brücke zu einer eher ökonometrischen Sichtweise zu schlagen.

Üblicherweise wird die OLS Methode angewandt, um eine *stetige* abhängige Variable mit Hilfe einer oder mehrerer x Variablen zu erklären. Da die Mathematik für stetige Zufallsvariablen etwas anspruchsvoller ist und dies leicht den Blick auf das Wesentliche verstellen kann beginnen wir mit einem sehr einfachen Beispiel für diskrete Variablen.

Dazu kehren wir wieder zu unserem alten Beispiel mit den Gebrauchtautos zurück. Ähnlich wie wir es bereits in der deskriptiven Statistik gemacht haben runden wir das Alter wieder auf ganze Jahre, aber diesmal runden wir auch die Preise auf 5000 Euro um eine überschaubare Anzahl von Ausprägungen zu erhalten. Dadurch werden sowohl die abhängige Variable 'Preis' als auch die erklärende Variable 'Alter' zu diskreten Variablen.[11]

Wieder nehmen wir im Gedankenexperiment an, dass wir die in Tabelle 3.3 gegebene Grundgesamtheit mit 100 Beobachtungen kennen (der Einfachheit halber haben wir diese Daten mit dem Computer erzeugt).

Wenn wir die OLS Methode auf diese Grundgesamtheit mit 100 Beobachtungen anwenden erhalten wir die PRF

$$\text{Preis}_i = 24\,465 - 2\,484\,\text{Alter}_i + \varepsilon_i$$

Diese PRF zeigt den 'wahren' Zusammenhang in der Grundgesamtheit, der für 'normal Sterbliche' natürlich unbeobachtbar ist.

Wir wollen nun eine Intuition dafür vermitteln, wie man von einer solchen gegebenen Grundgesamtheit zu einer gemeinsamen Wahrscheinlichkeitsverteilung zweier Zufallsvariablen 'Preis' und 'Alter' kommen kann.

Dazu beginnen wir damit, die Daten von Tabelle 3.3 kompakter darzustellen, indem wir die *Häufigkeiten* der unterschiedlichen Ausprägungen der Variablen Alter und Preis zählen.

Wenn wir in Tabelle 3.3 nachzählen, wie oft z.B. die Kombination Alter = 2 *und* Preis = 20000 vorkommt, so finden wir diese Merkmalskombination 15 Mal; die Kombination Alter = 6 *und* Preis = 15000 kommt hingegen nur ein Mal vor (Beobachtung 67), die Kombination Alter = 1 *und* Preis = 5000 kommt überhaupt nicht vor.

Tabelle 3.4 zeigt diese Häufigkeiten.[12] Selbstverständlich könnten wir aus diesen Häufigkeiten wieder die Grundgesamtheit in Tabelle 3.3 rekonstruieren, die beiden Tabellen sind nur unterschiedliche Darstellungen der gleichen Grundgesamtheit.

[11] Für diskrete abhängige Variablen existieren geeignetere Schätzverfahren als OLS, aber dies spielt im Moment keine Rolle.

[12] In R erhalten Sie diese Tabelle mit `table(Alter, Preis)`, in Stata mit `tabulate Alter Preis`.

Tabelle 3.3: Grundgesamtheit für Alter und Preis von Gebrauchtautos. (`https://www.uibk.ac.at/econometrics/data/auto100.csv`)

Obs	Alter (x)	Preis (y)	Obs	Alter (x)	Preis (y)
1	3	15000	51	2	20000
2	3	15000	52	1	20000
3	4	15000	53	5	15000
4	2	20000	54	6	10000
5	4	15000	55	4	15000
6	6	10000	56	2	20000
7	4	15000	57	4	15000
8	4	15000	58	4	15000
9	3	15000	59	4	15000
10	6	10000	60	5	10000
11	3	15000	61	4	15000
12	1	20000	62	5	10000
13	6	10000	63	4	15000
14	3	15000	64	2	20000
15	6	10000	65	5	10000
16	6	10000	66	4	10000
17	3	20000	67	6	15000
18	6	10000	68	2	15000
19	3	15000	69	2	20000
20	3	20000	70	3	20000
21	2	20000	71	2	20000
22	1	25000	72	5	10000
23	4	15000	73	4	10000
24	2	20000	74	4	15000
25	5	10000	75	2	20000
26	4	15000	76	2	20000
27	4	15000	77	4	15000
28	4	15000	78	4	15000
29	3	15000	79	4	15000
30	6	10000	80	1	20000
31	6	10000	81	6	10000
32	6	5000	82	6	10000
33	6	10000	83	5	15000
34	5	10000	84	5	15000
35	4	15000	85	5	10000
36	3	15000	86	6	10000
37	2	25000	87	2	20000
38	3	15000	88	2	20000
39	2	20000	89	3	15000
40	3	20000	90	6	10000
41	2	25000	91	4	15000
42	4	15000	92	3	20000
43	2	20000	93	4	15000
44	5	10000	94	4	15000
45	4	10000	95	4	15000
46	6	5000	96	5	15000
47	5	15000	97	6	10000
48	1	20000	98	2	15000
49	3	20000	99	5	10000
50	4	15000	100	2	20000

Tabelle 3.4: Darstellung der Grundgesamtheit mit 100 Beobachtungen in Form einer Häufigkeitstabelle.

		Preis (y)					
		5000	10000	15000	20000	25000	Summe
Alter (x)	1	0	0	0	4	1	5
	2	0	0	2	15	2	19
	3	0	0	10	6	0	16
	4	0	3	25	0	0	28
	5	0	9	5	0	0	14
	6	2	15	1	0	0	18
Summe		2	27	43	25	3	100

Tabelle 3.5: Darstellung der Grundgesamtheit als gemeinsame Verteilung (Wahrscheinlichkeitsfunktion) zweier Zufallsvariablen Alter und Preis.

		Preis (y)					
		5000	10000	15000	20000	25000	$\Pr(x)$
Alter (x)	1	0	0	0	0.04	0.01	0.05
	2	0	0	0.02	0.15	0.02	0.19
	3	0	0	0.10	0.06	0	0.16
	4	0	0.03	0.25	0	0	0.28
	5	0	0.09	0.05	0	0	0.14
	6	0.02	0.15	0.01	0	0	0.18
$\Pr(y)$		0.02	0.27	0.43	0.25	0.03	1

Nun dividieren wir die absoluten Häufigkeiten in Tabelle 3.4 durch die Anzahl der Beobachtungen (100) und erhalten als Ergebnis in Tabelle 3.5 die *relativen Häufigkeiten* (d.h. einfache Anteile); z.B. weisen 15% der Autos die Merkmalskombination Alter = 2 *und* Preis = 20000 auf.

Tabelle 3.5 zeigt die entsprechenden Anteile. Aber wir können die Perspektive wechseln und sie intuitiv als gemeinsame Wahrscheinlichkeitsverteilung zweier Zufallsvariablen 'Alter' und 'Preis' interpretieren. Für den einfachen Fall mit der gegebenen Grundgesamtheit ist dies ziemlich einleuchtend, wenn wir *zufällig* ein Auto aus dieser Grundgesamtheit ziehen würden ist die Wahrscheinlichkeit, dass dieses Auto zwei Jahre alt und 20000 Euro kosten wird, 0.15.

Wir können die Idee nun etwas verallgemeinern. Erinnern wir uns dazu an die *frequentistische Wahrscheinlichkeitsdefinition*: wenn ein Zufallsexperiment unter identischen Bedingungen beliebig oft wiederholt werden kann und wir die relative Häufigkeit eines Ereignisses A nach n Durchführungen des Experiments mit n_A/n bezeichnen, dann ist die Wahrscheinlichkeit der Grenzwert dieser relativen Häufigkeit, wenn die Anzahl der Experimente gegen Unendlich geht

$$\Pr(A) = \lim_{n\to\infty} \left(\frac{n_A}{n}\right)$$

Wir können uns vorstellen, dass die Anteile in Tabelle 3.5 das Resultat von zumindest sehr vielen Durchführungen eines Zufallsexperiments sind, und in diesem Sinne als (frequentistische) Wahrscheinlichkeiten interpretiert werden können.

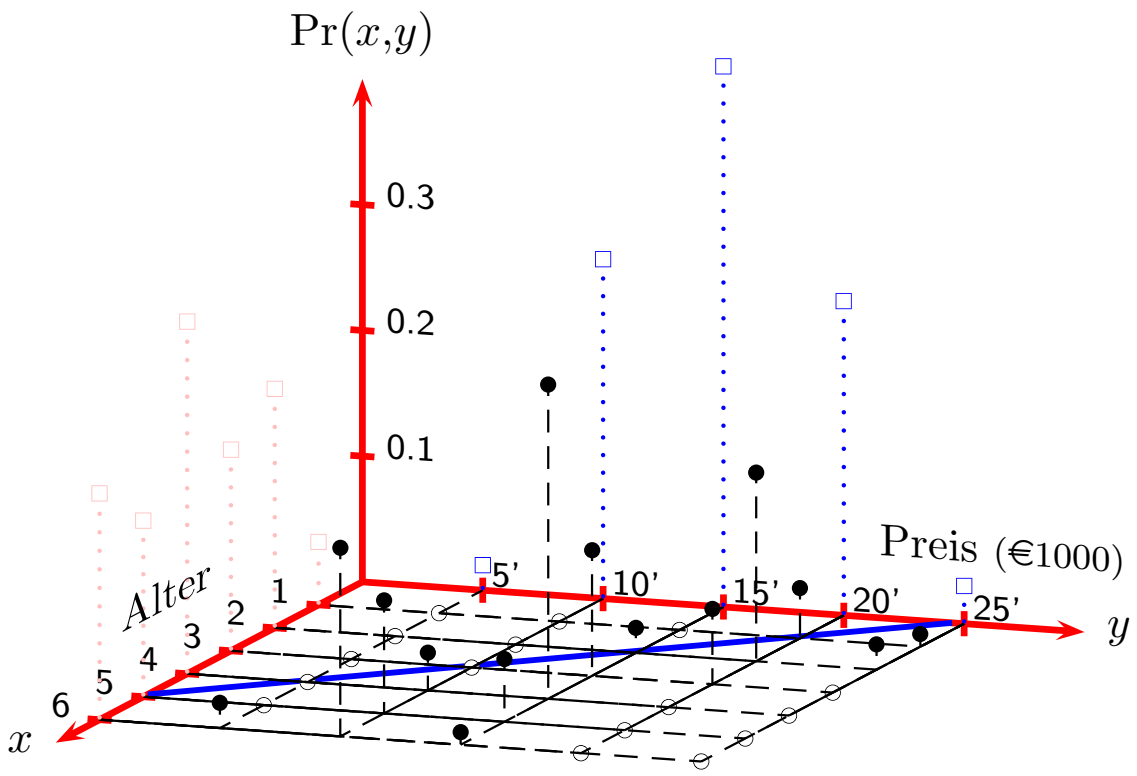


Abbildung 3.11: Grafische Darstellung der gemeinsamen Wahrscheinlichkeitsfunktion in Tabelle 3.5 (mit Randverteilungen). Blau eingezeichnet die PRF: $\widehat{\text{Preis}}_i = 24\,465 - 2\,484\,\text{Alter}_i$.

Anstelle der Vorstellung einer fix gegebenen Grundgesamtheit tritt dann die Idee eines *Datengenerierenden Prozesses* (DGP), und Tabelle 3.5 kann dann als gemeinsame Wahrscheinlichkeitsfunktion interpretiert werden, die eine kompakte Beschreibung dieses DGP ermöglicht. Alternativ kann man sich auch eine unendlich große Grundgesamtheit hinter Tabelle 3.5 vorstellen, und tatsächlich werden beide Formulierungen in der Literatur häufig synonym verwendet. Wichtig ist, dass die gemeinsame Wahrscheinlichkeitsfunktion die gesamte relevante Information des zugrunde liegenden Zufallsexperiments enthält, und dass wir uns deshalb nur noch um diese kümmern müssen.

Abbildung 3.11 zeigt eine grafische Abbildung dieser gemeinsamen Wahrscheinlichkeitsfunktion; auf der Grundebene sind Preis und Alter aufgetragen, auf der Höhenachse die entsprechenden Wahrscheinlichkeiten.

Die nächste Frage wird sein, ob wir die Information dieser gemeinsamen Wahrscheinlichkeitsfunktion ähnlich wie in der deskriptiven Statistik 'verdichten' können, und das Resultat wird sein, dass die *bedingte Erwartungswertfunktion* genau das gleiche für den DGP (oder die Grundgesamtheit) leistet, was die einfache OLS Regression in der deskriptiven Statistik geleistet hat. An die Stelle der bedingten Mittelwerte der Realisationen in der deskriptiven Statistik treten nun die bedingten Erwartungswerte der beiden Zufallsvariablen mit ihrer gemeinsamen Wahrscheinlichkeitsverteilung.

Wir werden nun zeigen, dass wir die *'population regression function'* (PRF) als lineare Approximation an die bedingte Erwartungswertfunktion interpretieren können (die PRF ist in Abbildung 3.11 als blaue Gerade eingezeichnet), ähnlich wie wir

in der deskriptiven Statistik die OLS Regression als lineare Approximation an die bedingten Mittelwerte interpretiert haben.

Im Kern geht es nun um die Momente der gemeinsamen Wahrscheinlichkeitsfunktion, um die Erwartungswerte und Varianzen, vor allem aber um die *bedingten* Erwartungswerte und die *bedingten* Varianzen.

Zur Erinnerung, der **Erwartungswert** einer diskreten Zufallsvariable ist definiert als die Summe aller der mit den Wahrscheinlichkeiten gewichteten möglichen Ausgänge des Zufallsexperiments. Wenn eine diskrete Zufallsvariable x insgesamt m verschiedene Ausprägungen annehmen kann ist der Erwartungswert definiert als

$$\mathrm{E}(x) = \sum_{j=1}^{m} x_j \Pr(x_j)$$

Achtung: Beim Erwartungswert wird über *alle möglichen Ausprägungen* der Zufallsvariable aufsummiert, gewichtet mit deren Wahrscheinlichkeiten, nicht über Beobachtungen (d.h. Realisationen). Erwartungswerte beziehen sich immer auf Zufallsvariablen, niemals auf die Realisationen! Das Analogon zum Erwartungswert für Realisationen ist das arithmetische Mittel.

Den Erwartungswert der Zufallsvariable 'Alter' können wir unter Verwendung der Randwahrscheinlichkeiten einfach berechnen

$$\begin{aligned} \mathrm{E}(\text{Alter}) &= \sum_{j=1}^{6} \text{Alter}_j \Pr(\text{Alter} = j) \\ &= 1 \times 0.05 + 2 \times 0.19 + 3 \times 0.16 + 4 \times 0.28 + 5 \times 0.14 + 6 \times 0.18 \\ &= 3.81 \end{aligned}$$

Analog können wir den Erwartungswert der Zufallsvariable 'Preis' berechnen, E(Preis) = 15000. Die Erwartungswerte der Zufallsvariablen Preis und Alter sind fixe Zahlen, also deterministisch.

In diesem Beispiel mit den 100 Beobachtungen ist der Erwartungswert gleich dem Mittelwert der Grundgesamtheit. Aber man beachte, dass der Mittelwert $\bar{x} := \frac{1}{n}\sum_i x_i$ nur für eine endliche Zahl n von Beobachtungen berechnet werden kann, der Erwartungswert $\mathrm{E}(x) := \sum_j x_j \Pr(x_j)$ ist hingegen auch für unendlich große Grundgesamtheiten definiert.

Auch die Varianz $\mathrm{var}(x) := \mathrm{E}[x - \mathrm{E}(x)]^2 = \mathrm{E}(x^2) - [\mathrm{E}(x)]^2$ kann einfach berechnet werden, $\mathrm{var}(\text{Alter}) = \mathrm{E}(\text{Alter}^2) - [\mathrm{E}(\text{Alter})]^2 = 16.71 - 3.81^2 = 2.1939$, und analog $\mathrm{var}(\text{Preis}) = 1.8 * 10^7$.

Für uns sind allerdings die bedingten Erwartungswerte und Varianzen relevanter.

3.3.1 Bedingte Erwartungswerte

Die bedingten Erwartungswerte sind das stochastische Analogon zu den bedingten Mittelwerten der deskriptiven Statistik. Für diskrete Zufallsvariablen ist der bedingte Erwartungswert definiert als

$$\mathrm{E}(y|x = \underline{x}) = \sum_{j=1} y_j \Pr(y_j|x = \underline{x})$$

Tabelle 3.6: Bedingte Wahrscheinlichkeiten der Preise für gegebenes Alter.

Preis (y)		5000	10000	15000	20000	25000	Summe
	1	0.00	0.00	0.00	0.80	0.20	1.00
	2	0.00	0.00	0.11	0.79	0.11	1.00
Alter	3	0.00	0.00	0.62	0.38	0.00	1.00
(x)	4	0.00	0.11	0.89	0.00	0.00	1.00
	5	0.00	0.64	0.36	0.00	0.00	1.00
	6	0.11	0.83	0.06	0.00	0.00	1.00

d.h. wir berechnen die bedingten Erwartungswerte genau gleich wie die unbedingten Erwartungswerte, nur dass wir als Gewichte nun die *bedingten* Wahrscheinlichkeiten wählen.

Durch das 'Bedingen auf x' halten wir gewissermaßen die Variable x bei der Ausprägung $\underline{x}$ fest, zum Beispiel das Alter des Autos bei der Ausprägung Alter = 1. In diesem Moment sind die anderen Ausprägungen des Alters irrelevant, nur die erste Zeile der gemeinsamen Wahrscheinlichkeitsfunktion in Tabelle 3.5 zählt. Allerdings sind die Einträge der ersten Zeile dann keine Wahrscheinlichkeiten mehr, da die Summe nicht Eins ergibt.

Um die bedingten Wahrscheinlichkeiten zu erhalten müssen wir die Einträge zuerst 'normalisieren', d.h. durch die entsprechende Randwahrscheinlichkeit dividieren, damit die Summe über alle Ausprägungen von y Eins ergibt.

Zum Beispiel: $\Pr(\text{Preis} = 20000|\text{Alter} = 1) = 0.04/0.05 = 0.8$.

Indem wir dies für alle Zeilen wiederholen erhalten wir die bedingten Wahrscheinlichkeiten für die Preise bei gegebenem Alter, siehe Tabelle 3.6.

Mit Hilfe der bedingten Wahrscheinlichkeiten können wir die bedingten Erwartungswerte berechnen, z.B.

$$\text{E}(\text{Preis}|\text{Alter} = 1) = 20000 \times 0.8 + 25000 \times 0.2 = 21000$$

Die Bedingte Erwartungswertfunktion (CEF)

Wenn wir für jede Alterstufe den bedingten Erwartungswert berechnen, und diese als Funktion des Alters anschreiben, erhalten wir die bedingte Erwartungswertfunktion (*'Conditional Expectation Function'*, CEF). Die CEF ordnet jeder Ausprägung der erklärenden Variable 'Alter' den entsprechenden bedingten Erwartungswert der abhängigen Variable 'Preis' zu

$$\text{E}(\text{Preis}|\text{Alter} = \underline{\text{Alter}}) = \begin{cases} 21000.00 & \text{für Alter} = 1 \\ 20000.00 & \text{für Alter} = 2 \\ 16875.00 & \text{für Alter} = 3 \\ 14464.29 & \text{für Alter} = 4 \\ 11785.71 & \text{für Alter} = 5 \\ 9722.22 & \text{für Alter} = 6 \end{cases}$$

Die grafische Abbildung dieser bedingten Erwartungswertfunktion (CEF) sind die in Abbildung 3.12 eingezeichneten sechs Punkte (Achtung, nicht die ebenfalls eingezeichnete strichlierte Linie).

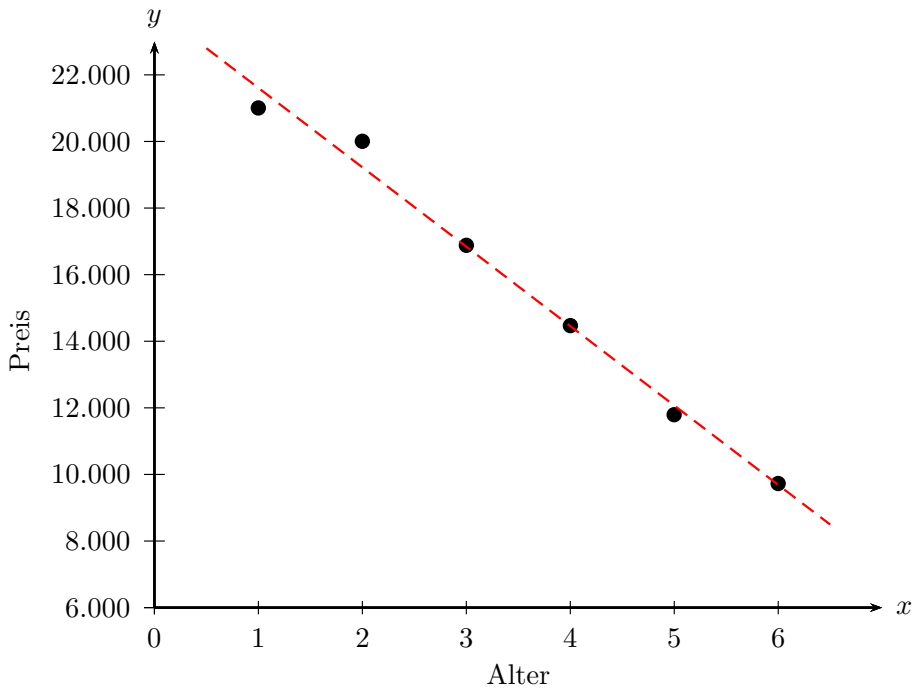


Abbildung 3.12: Bedingte Erwartungswertfunktion (CEF) für das Auto-Beispiel (Achtung: die CEF sind die eingezeichneten Punkte, die Line ist eine lineare Approximation an die CEF).

Für diskrete Variablen ist der bedingte Erwartungswert einfach der Erwartungswert der entsprechenden Untergruppe. In diesem einfachen Beispiel sind die bedingten Erwartungswerte die Mittelwerte der Grundgesamtheit für die entsprechenden Alters-Gruppen, der durchschnittliche Preis von Autos mit Alter = 1 in der Grundgesamtheit beträgt z.B. 21000 Euro.

Allgemeiner und für mehrere erklärende Variablen $x_1, x_2, \ldots, x_k$ wird dies manchmal geschrieben als

$$\mathrm{E}(y|x_1, x_2, \ldots, x_k) = m(x_1, x_2, \ldots, x_k)$$

Diese CEF ordnet jeder möglichen Kombination der x Variablen den entsprechenden bedingten Erwartungswert zu.

Man beachte, dass *für fixierte* x diese bedingten Erwartungswerte *keine* Zufallsvariablen sind, obwohl y eine Zufallsvariable ist. Wer immer auf Grundlage der gemeinsamen Verteilung (Tabelle 3.5) die bedingten Erwartungswerte berechnet wird das gleiche Ergebnis erhalten, es gibt dabei kein Element der Unsicherheit.

Die **Eigenschaften der CEF** haben wir bereits im 'Statistischen Intermezzo' diskutiert, deshalb zählen wir hier nur die wichtigsten auf:

1. Linearität: wenn x, y, z Zufallsvariablen und a, b beliebige Konstante sind gilt

$$\mathrm{E}(ay + bz|x = \underline{x}) = a\,\mathrm{E}(y|x = \underline{x}) + b\,\mathrm{E}(z|x = \underline{x})$$

2. Das Gesetz der iterierten Erwartungen: Für $\mathrm{E}(|y| < \infty$ gilt

$$\mathrm{E}_x[\mathrm{E}(y|x = \underline{x})] = \mathrm{E}(y)$$

Wenn wir den Erwartungswert über alle bedingten Erwartungswerte von y nehmen erhalten wir den unbedingten Erwartungswert. E_x soll bedeuten, dass wir den äußeren Erwartungswert über x bilden. Das Analogon dazu aus der deskriptiven Statistik ist, dass das mit den Anteilen gewichtete Mittel über die Mittelwerte mehrerer Untergruppen den Mittelwert über alle Beobachtungen gibt. Wenn wir z.B. die durchschnittliche Körpergröße von Frauen und Männern haben, können wir daraus die mittlere Körpergröße aller Personen berechnen, indem wir die Gruppenmittel mit deren Anteilen multiplizieren und aufsummieren.

Dieses Gesetz gilt auch allgemeiner $\mathrm{E}(\mathrm{E}(y|x_1 = \underline{x}_1, x_2 = \underline{x}_2)|x_1 = \underline{x}_1) = \mathrm{E}(y|x_1 = \underline{x}_1)$

3. Konditionierungstheorem (*'Taking out what is known property'*):

$$\mathrm{E}(g(x)h(y)|x = \underline{x}) = g(x)\,\mathrm{E}(h(y)|x = \underline{x})$$

Wann immer man auf eine Variable x konditioniert kann diese wie eine Konstante behandelt werden. Zum Beispiel ist $\mathrm{E}(x|x = \underline{x}) = \underline{x}$ oder $\mathrm{E}(xy|x = \underline{x}) = x\,\mathrm{E}(y|x = \underline{x})$.

Die CEF Störterme

Die CEF kann nun verwendet werden, um die y Variable in zwei Teile zu zerlegen, in einen systematischen Teil $\mathrm{E}(y|x = \underline{x})$ und in einen Störterm ε

$$y = \mathrm{E}(y|x = \underline{x}) + \varepsilon$$

und definieren so die CEF Störterme $\varepsilon = y - \mathrm{E}(y|x = \underline{x})$.

Für jedes mögliche (x, y) Paar existiert ein Störterm, z.B. im Autobeispiel für Preis $= 25000$, Alter $= 1$: $\varepsilon_{y=20000,x=1} = 20000 - 21000 = -1000$, $\varepsilon_{y=25000,x=1} = 25000 - 21000 = 4000$, usw.

Diese CEF Störterme werden direkt aus der gemeinsamen Verteilung von x, y hergeleitet, deshalb folgen deren Eigenschaften aus deren Konstruktion.

Eigenschaften der CEF Störterme: Für $\mathrm{E}(|y|) < \infty$ gilt

1. Die bedingten Erwartungswerte der CEF Störterme sind Null

$$\mathrm{E}(\varepsilon|x = \underline{x}) = 0$$

Dies folgt aus dem Konditionierungstheorem und dem Gesetz der iterierten Erwartungen

$$\begin{aligned}
\mathrm{E}(\varepsilon|x = \underline{x}) &= \mathrm{E}\left[(y - \mathrm{E}(y|x = \underline{x}))|x = \underline{x}\right] \\
&= \mathrm{E}(y|x = \underline{x}) - \mathrm{E}\left[\mathrm{E}(y|x = \underline{x}))|x = \underline{x}\right] \\
&= \mathrm{E}(y|x = \underline{x}) - \mathrm{E}(y|x = \underline{x}) \\
&= 0
\end{aligned}$$

Diese Eigenschaft wird manchmal auch *'mean independence'* genannt, da $\mathrm{E}(\varepsilon|x = \underline{x}) = 0$ impliziert, dass der bedingte Erwartungswert der Störterme unabhängig von x ist.

2. Aus der vorherigen Eigenschaft folgt mit dem Gesetz der iterierten Erwartungen, dass auch die unbedingten Erwartungswerte der CEF Störterme Null sind
$$\mathrm{E}(\varepsilon) = \mathrm{E}\left[\mathrm{E}(\varepsilon | x = \underline{x})\right] = 0$$

3. Die CEF Störterme sind mit dem Regressor x unkorreliert
$$\mathrm{E}(x\varepsilon) = 0$$
für $\mathrm{E}(|x\varepsilon|) < \infty$. Auch diese Eigenschaft folgt wieder aus dem Gesetz der iterierten Erwartungen und dem Konditionierungstheorem
$$\mathrm{E}(x\varepsilon) = \mathrm{E}\left[\mathrm{E}((x\varepsilon)|x = \underline{x})\right] = \mathrm{E}\left[x\,\mathrm{E}(\varepsilon|x = \underline{x})\right] = 0$$
weil $\mathrm{E}(\varepsilon|x = \underline{x}) = 0$.
Wie man einfach zeigen kann gilt dies auch allgemeiner für beliebige Funktionen von x, d.h.
$$\mathrm{E}(g(x)\varepsilon) = 0$$

Die Varianz der CEF Störterme

Eine wichtige Kennzahl für die unerklärte Streuung von y um die CEF ist die Varianz der CEF Störterme
$$\mathrm{var}(\varepsilon) := \sigma^2 = \mathrm{E}[(\varepsilon - \mathrm{E}(\varepsilon))^2] = \mathrm{E}(\varepsilon^2)$$

Diese Varianz σ^2 wird auf Englisch auch *'regression variance'* oder *'variance of the regression error'* genannt.

Die CEF als bester Prediktor: Angenommen, wir möchten mit Hilfe des Regressors x eine Vorhersage für y machen. Dann suchen wir eine (noch unbekannte) Vorhersagefunktion $g(x)$, die als Resultat eine 'möglichst gute' Vorhersage liefern soll. Aber was ist eine 'möglichst gute' Vorhersage? Eine Möglichkeit wäre die Funktion zu suchen, die den Erwartungswert des quadrierten Vorhersagefehlers $y - g(x)$ minimiert, also
$$\mathrm{E}[y - g(x)]^2$$
Es zeigt sich, dass die CEF genau die gesuchte Funktion ist.
Wir schreiben $\mathrm{E}(y|x = \underline{x}) = m(x)$, also $y = m(x) + \varepsilon$. Einsetzen gibt

$$\begin{aligned}
\mathrm{E}[y - g(x)]^2 &= \mathrm{E}[m(x) + \varepsilon - g(x)]^2 \\
&= \mathrm{E}[\varepsilon + (m(x) - g(x))]^2 \\
&= \mathrm{E}(\varepsilon^2) + 2\,\mathrm{E}[\varepsilon[m(x) - g(x)] + \mathrm{E}[m(x) - g(x)]^2 \\
&= \mathrm{E}(\varepsilon^2) + \mathrm{E}[m(x) - g(x)]^2 \qquad (\text{weil } \mathrm{E}[m(x)\varepsilon] = 0)
\end{aligned}$$

$\mathrm{E}[m(x)\varepsilon] = 0$ weil wir bei den Eigenschaften der CEF Störterme gezeigt haben, dass für jede Funktion $g(x)$ gilt $\mathrm{E}(g(x)\varepsilon) = 0$, also auch für $m(x)$.

Der letzte Ausdruck $\mathrm{E}[y - g(x)]^2 = \mathrm{E}(\varepsilon^2) + \mathrm{E}[m(x) - g(x)]^2$ wird minimiert, wenn $g(x) = m(x)$, also minimiert die CEF den Erwartungswert des quadrierten Vorhersagefehlers $y - g(x)$, und ist in diesem Sinne der beste Prediktor. Wir halten also fest

$$\mathrm{E}[y - g(x)]^2 \geq \mathrm{E}[y - m(x)]^2$$

mit $m(x) = \mathrm{E}(y|x = \underline{x})$.

Die **bedingte Varianz der CEF Störterme** ist für $\mathrm{E}(\varepsilon^2) < \infty$

$$\mathrm{var}(\varepsilon|x) = \mathrm{E}(\varepsilon^2|x) := \sigma^2(x)$$

wobei $\sigma^2(x)$ bedeuten soll, dass diese bedingte Varianz eine Funktion der x ist.

Wenn die Störterme von x abhängen, also $\mathrm{E}(\varepsilon^2|x) = \sigma^2(x)$, werden sie *heteroskedastisch* genannt.

Falls die Störterme *nicht* von x abhängen, also $\mathrm{E}(\varepsilon^2|x) = \sigma^2$, werden die Störterme *homoskedastisch* genannt.

Ein wichtiger Spezialfall ist die Normalverteilung. Wenn x und y gemeinsam normalverteilt sind ist die CEF linear und die Störterme sind homoskedastisch. Allerdings ist die Normalverteilung in dieser Hinsicht eine große Ausnahme. Im allgemeinen wird eher davon auszugehen sein, dass die CEF nichtlinear ist und dass die Störterme heteroskedastisch sind.

3.3.2 Die lineare CEF

Wir haben gerade erwähnt, dass die CEF zweier gemeinsam normalverteilter Zufallsvariablen linear ist, d.h. wir können die CEF schreiben als

$$\mathrm{E}(y|x = \underline{x}) = \beta_1 + \beta_2 x$$

Neben der gemeinsamen Normalverteilung führt auch das gesättigte Dummy Variablen Modell (d.h. wenn nur Dummy Variablen vorkommen und alle möglichen Interaktionen zwischen ihnen berücksichtigt werden) zu einer linearen CEF (die Koeffizienten sind exakt die bedingten Mittelwerte).

Für diese Spezialfälle haben wir im 'Statistischen Intermezzo' gezeigt, dass für die PRF $y = \beta_1 + \beta_2 x + \varepsilon$ gilt

$$\beta_2 = \frac{\mathrm{cov}(y, x)}{\mathrm{var}(x)} \qquad \text{und} \qquad \beta_1 = \mathrm{E}(y) - \beta_2 \, \mathrm{E}(x)$$

Diese beiden Fälle, d.h. die gemeinsam normalverteilten Zufallsvariablen und das gesättigte Dummy Variablen Modell, sind zwar wichtig, aber dennoch Ausnahmen.

In der Regel wird die CEF keine lineare Funktion sein. Die Punkte der CEF im Autobeispiel lagen zwar fast auf einer geraden Linie, siehe Abbildung 3.12 (Seite 28), aber eben nur fast.

Dieser Fall ist analog zu unserem Beispiel in der deskriptiven Statistik, wo wir die OLS Gerade als eine lineare Approximation an die bedingten Mittelwerte interpretierten.

Ebenso können wir in diesem Fall die lineare CEF als eine Approximation an die CEF ansehen.

Wir haben vorhin gesehen, dass die CEF $m(x)$ der beste Prediktor von allen möglichen Prediktor Funktionen $g(x)$ ist, weil der mittlere quadratische Vorhersagefehler $\mathrm{E}(y - g(x))^2$ nur minimal ist, wenn $g(x) = m(x)$.

Analog können wir eine *lineare* Funktion suchen, die *von allen linearen Funktionen* den minimalen mittleren quadratischen Vorhersagefehler liefert.

Das heißt, wir suchen die Werte von β_1 und β_2, welche die folgende Funktion minimieren:

$$\min_{\beta_1,\beta_2} \quad \mathrm{E}[y - \beta_1 - \beta_2 x]^2$$

Dies ist natürlich genau das OLS Problem, und wir erhalten – wenig überraschend – die Lösungen

$$\beta_2 = \frac{\mathrm{E}[x - \mathrm{E}(x)][y - \mathrm{E}(y)]}{\mathrm{E}[x - \mathrm{E}(x)]^2} = \frac{\mathrm{cov}(y, x)}{\mathrm{var}(x)} \qquad \text{und} \qquad \beta_1 = \mathrm{E}(y) - \beta_2 \, \mathrm{E}(x)$$

Man beachte, dass diese Parameter eine Funktion der ersten beiden Momente der zugrunde liegenden gemeinsamen Wahrscheinlichkeitsfunktion sind. Da sie direkt aus der gemeinsamen Wahrscheinlichkeitsfunktion von x und y berechnet werden sind sie auch für eine unendlich große Grundgesamtheit definiert.

Diese lineare Approximation an die CEF wird auch *lineare Regressionsfunktion* oder – aus Gründen, die erst später zu Tage treten werden – *lineare Projektion* genannt.

Diese lineare Regressionsfunktion ist in Abbildung 3.5 als blaue Linie eingezeichnet.

Für die Interpretation der CEF $\mathrm{E}(y|x_1, x_2, \ldots, x_k) = m(x_1, x_2, \ldots, x_k)$ werden häufig die *marginalen Effekte* herangezogen, für stetige Variablen sind dies einfach die partiellen Ableitungen

$$\frac{\partial}{\partial x_1} m(x_1, x_2, \ldots, x_k)$$

Falls x_1 eine Dummy Variable ist kann die Differenz berechnet werden

$$m(1, x_2, \ldots, x_k) - m(0, x_2, \ldots, x_k)$$

Diese messen den *ceteris paribus* Effekt, aber im Unterschied zur üblichen ceteris paribus Interpretation in der Ökonomik bezieht sich die ceteris paribus Interpretation hier *nur auf die berücksichtigten Variablen* $(x_2, \ldots, x_k)$. Wenn z.B. die log-Stundenlöhne einer Lohngleichung durch Bildungsjahre, Berufserfahrung und Geschlecht erklärt werden bezieht sich die ceteris paribus Interpretation nur auf die Bildungsjahre, Berufserfahrung und Geschlecht, aber nicht auf nicht berücksichtigte Variablen wie z.B. Intelligenz. Da wir selten davon ausgehen können, tatsächlich alle relevanten Regressoren berücksichtigt zu haben, erlaubt die Regressionsanalyse kaum Kausalaussagen.[13] Die marginalen Effekte hängen davon ab, welche Regressoren in die Berechnung der bedingten Erwartungswerte eingingen.

Natürlich beziehen sich diese marginalen Effekte auf eine Änderung der bedingten Erwartungswerte von y, nicht auf das beobachtete y eines Individuums.

[13] Wie wir schon früher einmal betont haben bieten randomisierte kontrollierte Experimente (*Randomized Controlled Trials*, RCT) hier einen entscheidenden Vorteil.

Beispiel: Für unser Autobeispiel mit der Wahrscheinlichkeitsfunktion in Tabelle 3.5 (Seite 24) erhalten wir die folgenden Kennwerte

$$\begin{aligned} \mathrm{E}(xy) &= 51700 \\ \mathrm{E}(x) &= 3.81 \\ \mathrm{E}(x^2) &= 16.71 \\ \mathrm{E}(y) &= 15000 \end{aligned}$$

Wir erinnern uns, dass

$$\begin{aligned} \mathrm{cov}(x,y) &= \mathrm{E}[x-\mathrm{E}(x)]\,\mathrm{E}[y-\mathrm{E}(y)] = \mathrm{E}(xy) - \mathrm{E}(x)\,\mathrm{E}(y) \\ \mathrm{var}(x) &= \mathrm{E}[x-\mathrm{E}(x)]^2 = \mathrm{E}(x^2) - [\mathrm{E}(x)]^2 \end{aligned}$$

und erhalten $\mathrm{cov}(x,y) = 51700 - 3.81 * 15000 = -5450$ und $\mathrm{var}(x) = 16.71 - 3.81^2 = 2.1939$.

Daraus folgt

$$\begin{aligned} \beta_2 &= -5450/2.1939 = -2484.161 \\ \beta_1 &= 15000 + 2484.161 * 3.81 = 24464.652 \end{aligned}$$

Dies sind natürlich die gleichen Werte die wir erhielten, als wir die OLS Methode direkt auf die Daten der Grundgesamtheit in Tabelle 3.3 anwandten (abgesehen von unterschiedlichen Rundungen).

In diesem (künstlichen) Beispiel war die lineare Regressionsfunktion eine relativ gute Approximation an die CEF, aber das wird natürlich nicht immer der Fall sein. Wenn der durch die CEF beschriebene Datengenerierende Prozess hochgradig nichtlinear ist, oder die lineare Regressionsfunktion falsch spezifiziert wurde, wird die lineare Regressionsfunktion eine sehr schlechte Approximation – oder im Fall von Fehlspezifikationen meist falsche Resultate – liefern.

3.3.3 Deterministische versus stochastische Regressoren

Bisher haben wir bei der Konditionierung die x gewissermaßen bei einem Wert $\underline{x}$ 'festgehalten', z.B. $\mathrm{E}(y|x = \underline{x})$.

Als R.A. Fisher in der landwirtschaftlichen Versuchsstation Rothamsted seine Feldexperimente durchführte, konnte er die Düngermenge *kontrollieren*. Aufgrund unterschiedlicher Boden- und Wetterbedingungen variierte die Erntemenge auch bei gleicher Düngermenge. In einem solchen experimentellen Setting macht die Annahme *deterministischer* Regressoren sehr viel Sinn, sie sind unter Kontrolle des Experimentators.

Im Gegensatz dazu haben Forscherinnen, die mit *Beobachtungsdaten* arbeiten (müssen) sehr viel weniger oder gar keine Kontrolle über ihre Regressoren. Warum sollten z.B. in einer Konsumfunktion die abhängigen Konsumausgaben stochastisch, das erklärende Einkommen aber deterministisch sein?

Bei den meisten ökonometrischen Fragestellungen stellen wir uns vor, dass sowohl die y als auch die x das Ergebnis eines Zufallsexperiments sind. Wenn wir – wie im Autobeispiel – aus einer gemeinsamen Verteilung Beobachtungen ziehen, halten wir

nicht die Zeile bei $x = \underline{x}$ fest und ziehen dazu wiederholte y, sondern wir ziehen zufällig (x_i, y_i)-Paare. Deshalb macht die Annahme deterministischer x in den meisten ökonometrischen Settings viel weniger Sinn.

Glücklicherweise kann man zeigen, dass das Gesetz iterierter Erwartungen und das Konditionierungstheorem auch für stochastische x gilt, deshalb gelten die vorhin angeführten Eigenschaften der CEF und der CEF Störterme auch für stochastische Regressoren.

Die dahinter stehende Mathematik ist etwas aufwändiger, doch es wurde gezeigt, dass man auch auf stochastische x konditionieren kann, und dass unter wenig strengen Annahmen eine *stochastische CEF* existiert und eindeutig ist. Die stochastische CEF wird häufig als $\mathrm{E}(y|\sigma(x))$ oder einfacher $\mathrm{E}(y|x)$ geschrieben.

In der früheren Sichtweise mit bei $\underline{x}$ fixierten x waren die bedingten Erwartungswerte $\mathrm{E}(y|x = \underline{x})$ deterministisch.

Im Gegensatz dazu sind die bedingten Erwartungswerte bei stochastischen x selbst Zufallsvariablen. Trotzdem gelten die meisten vorhergehenden Aussagen auch für die stochastische CEF, es werden aber zusätzliche Annahmen benötigt. Eine der wichtigsten dieser Annahmen ist, dass die ersten vier Momente der gemeinsamen Wahrscheinlichkeitsfunktion existieren und nicht unendlich sind.

Wir werden im Folgenden trotzdem häufig deterministische Regressoren annehmen, allerdings nicht, weil dies realistischer ist, sondern einzig und allein deshalb, weil dies einfacher ist und eine übersichtlichere Notation ermöglicht.

3.4 Das stochastische Modell für stetige Variablen

Bisher haben wir uns hauptsächlich auf Beispiele mit diskreten Zufallsvariablen beschränkt. Wir haben z.B. das Alter und den Preis der Autos gerundet, um eine überschaubare Anzahl von Ausprägungen zu erhalten. Der Grund dafür war rein didaktischer Natur, diskrete Variablen sind in der Regel einfacher vorstellbar, und sie erlauben v.a. einfachere Beispiele.

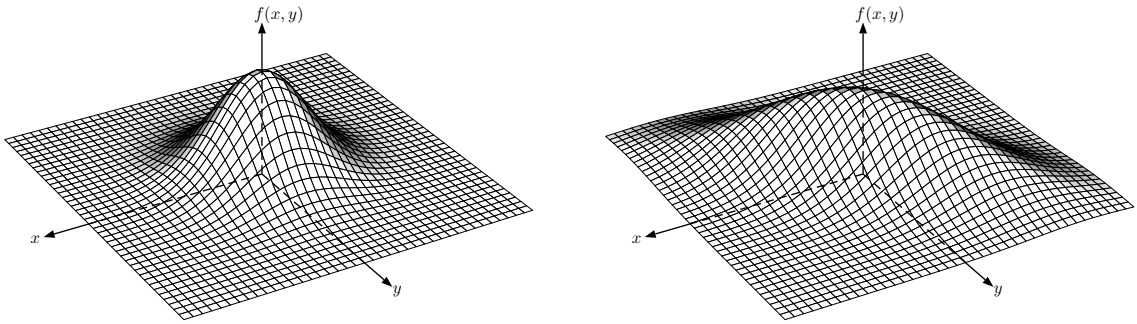
Was passiert, wenn wir weniger stark Runden, und im Extremfall, wenn die Zufallsvariablen (y, x) stetig sind und eine gemeinsame Dichtefunktion $f(y, x)$ haben?

Vorab, fast alles, was wir bisher gesagt haben, bleibt auch für stetige Zufallsvariablen gültig. Ein wichtiger Unterschied ist, dass der Wert der Dichtefunktion an einer gegebenen Stelle nicht mehr als Wahrscheinlichkeit interpretiert werden darf; die Wahrscheinlichkeiten sind in diesen Fällen als *Flächen* (oder Volumina) unter der Dichtefunktion definiert.

Wenn $f(y)$ eine Dichtefunktion ist, dann ist die Wahrscheinlichkeit dafür, dass die stetige Zufallsvariable y einen Wert in einem beliebigen Intervall $[a, b]$ (mit $a < b$ und $a, b \in \mathbb{R}$) annimmt, gleich

$$\Pr(a < y < b) = \int_a^b f(y)dy$$

Im Folgenden werden wir uns wieder auf einen eher intuitiven Zugang beschränken und die zentralen Ideen anhand einiger Grafiken erläutern. Der Einfachheit halber werden wir uns dabei auf die Normalverteilung beschränken.



$\text{corr}(x, y) = 0$ $\text{corr}(x, y) = -0.7$

Abbildung 3.13: Gemeinsame Normalverteilung $f(y, x)$.

Abbildung 3.13 zeigt die gemeinsame Dichte zweier normalverteilter Zufallsvariablen, im linken Panel sind die Variablen unkorreliert und stochastisch unabhängig.[14]

Stochastische Unabhängigkeit ist definiert als

$$f(y, x) = f_y(y) f_x(x)$$

Dies impliziert für $f_x(x) > 0$, dass die bedingte Dichte gleich der unbedingten Dichte ist, da

$$f_{y|x}(y|x) = \frac{f(y, x)}{f_x(x)} = \frac{f_y(y) f_x(x)}{f_x(x)} = f_y(y)$$

Das rechte Panel in Abbildung 3.13 zeigt die gemeinsame Dichte zweier negativ korrelierten Zufallsvariablen.

In Abbildung 3.14 ist die Randdichte von x und die bedingte Dichte $f_{y|x}(y|x)$ eingezeichnet.

Für die gemeinsame Dichte $f(y, x)$ ist die Randdichte (*marginal density*) definiert als

$$f_x(x) = \int_{\mathbb{R}} f(y, x) dy$$

Abbildung 3.14 zeigt die Randdichte von x als graue Linie an der 'Rückwand'. Da über alle y aufsummiert wurde ist sie natürlich höher als die gemeinsame Dichte und nur eine Funktion von x.

Die bedingte Dichte (*conditional density*) von y gegeben x ist für $f_x(x) > 0$ definiert als

$$f_{y|x}(y|x) = \frac{f(y, x)}{f_x(x)}$$

Wir erhalten die bedingte Dichte in Abbildung 3.14, indem wir die gemeinsame Dichte bei einem gegebenen x parallel zur y-Achse 'durchschneiden'. Allerdings ist die blau eingezeichnete Schnittlinie keine Dichte, da die Fläche darunter nicht Eins ist. Deshalb muss sie normalisiert werden, damit die Fläche darunter wieder gleich Eins ist. Dies wird erreicht, indem sie durch den Wert der Randdichte von x an

[14] Achtung, Variablen können unkorreliert, aber trotzdem stochastisch abhängig sein. Korrelationen messen nur lineare Abhängigkeiten.

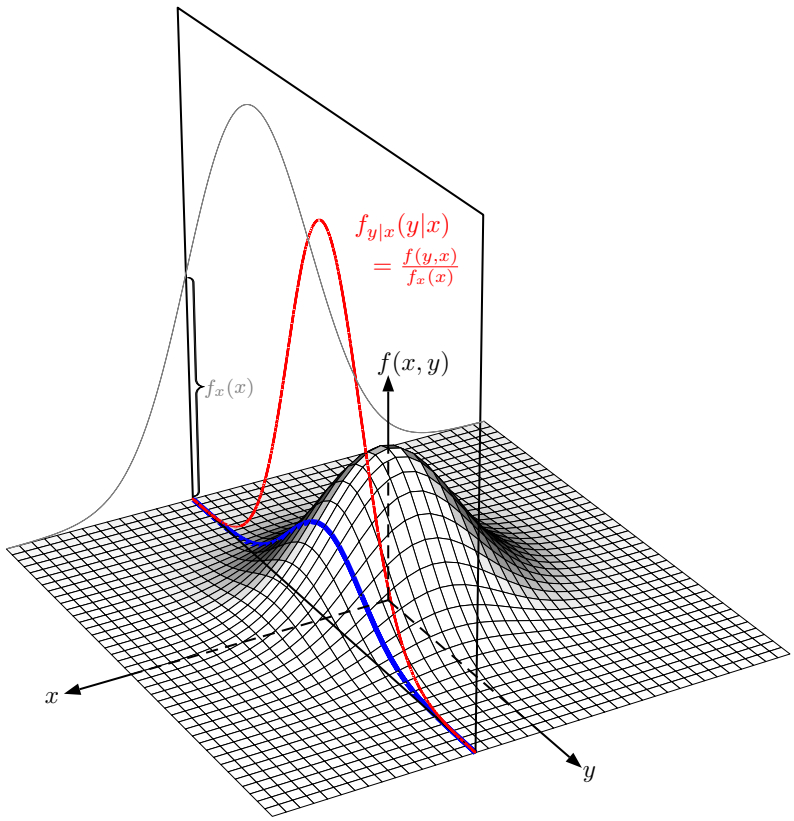


Abbildung 3.14: Gemeinsame Dichtefunktion zweier normalverteilter Zufallsvariablen mit Randdichte $f_x(x)$ und bedingter Dichte $f(y,x)/f_x(x)$.

dieser Stelle dividiert wird. Das Ergebnis ist die rot eingezeichnete bedingte Dichte $f_{y|x}(y|x)$.

Den (nicht eingezeichneten) *bedingten Erwartungswert* für ein gegebenes x erhalten wir, indem wir alle möglichen Ausprägungen von y aufsummieren, *gewichtet mit der bedingten Dichte*.

$$\mathrm{E}(y|x) = \int_{\mathbb{R}} y\, f_{y|x}(y|x)\, dy$$

Da in diesem Beispiel die bedingte Dichte eine Normalverteilungskurve ist erhalten wir diesen, indem wir – im übertragenen Sinn – ein Senkblei vom Maximum der bedingten Dichtefunktion auf die Grundfläche senken.

Dies können wir für andere x wiederholen, Abbildung 3.15 zeigt vier verschieden bedingte Dichtefunktionen. Man beachte, dass $y = \mathrm{E}(y|x) + \varepsilon$, bzw. $\varepsilon = y - \mathrm{E}(y|x)$. Deshalb zeigen die bedingten Dichtefunktionen zugleich die Verteilung der CEF Störterme. Dabei wird eine spezielle Eigenschaft der Normalverteilung sichtbar, die bedingten Dichtefunktionen sehen alle gleich aus, obwohl die blauen Schnittlinien mit zunehmender Entfernung von μ_y immer flacher werden. Dies folgt natürlich daraus, dass wir durch eine ebenfalls kleiner werdende Randdichte $f_x(x)$ dividieren. Darüber hinaus liegen alle bedingten Erwartungswerte auf einer Geraden, d.h. die CEF ist linear, und alle bedingten Dichtefunktionen haben die gleiche Varianz, d.h. die CEF Störterme sind homoskedastisch. Dies gilt allerdings *nur* für die Normalverteilung, nicht allgemein!

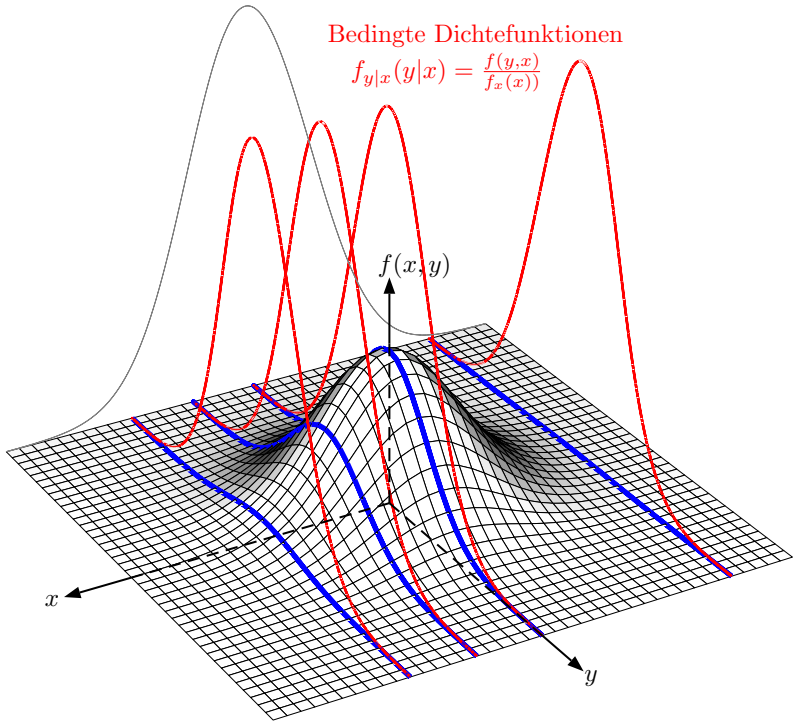


Abbildung 3.15: Bedingte Erwartungswertefunktionen von y für verschiedene x und bedingte Dichtefunktionen der Störterme ε_i (rot).

Da in diesem Beispiel (y, x) stochastisch unabhängig waren und $\mu_y = \mu_x = 0$ liegen die bedingten Erwartungswerte auf der x-Achse.

Abbildung 3.16 zeigt schließlich den Fall für zwei *positiv korrelierte* gemeinsam normalverteilte Zufallsvariablen (y, x). Die obere Ebene zeigt die gemeinsame Verteilung, in der unteren Ebene ist die CEF $\mathrm{E}(y|x) = m(x)$ mit drei bedingten Dichtefunktionen der dazugehörigen Störterme eingezeichnet. Diese CEF möchten Ökonometriker aus der beobachteten Stichprobe lernen.

3.4.1 Das 'wahre' Modell: Spezifikation und Identifikation

Die bedingte Erwartungswertfunktion (CEF) ist für den Forscher natürlich genauso wenig beobachtbar wie die gemeinsame Verteilung von (y, x) oder die Grundgesamtheit, bzw. der Datengenerierenden Prozess (DGP). Da die CEF eine kompakte Beschreibung der gemeinsamen Verteilung liefert versuchen Ökonometriker, die Parameter der CEF aus Stichprobendaten zu schätzen.

Im vorhergehenden Beispiel hatten wir eine sehr einfache lineare CEF mit nur einer erklärenden Variable x. Tatsächliche CEFs werden allerdings kaum eine so einfache Form haben, sondern hochkomplexe Funktionen von vielen erklärenden Variablen sein.

Schwieriger noch, die Forscherin kennt den Datengenerierenden Prozess 'hinter' der CEF nicht (vollständig), und kann daher auch nicht wissen, welche erklärenden Variablen und welche Funktionsform eine geeignete Schätzung liefern werden.

Wir haben eingangs den Physiker Werner Heisenberg zitiert, *"What we observe is not Nature itself but Nature exposed to our method of questioning."*

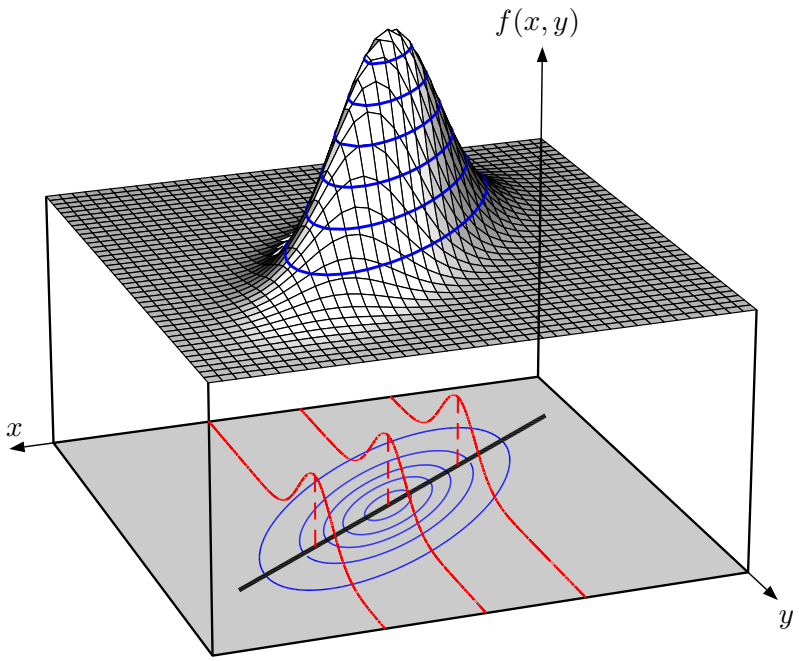


Abbildung 3.16: Bedingte Erwartungswertfunktion (CEF) für zwei gemeinsam normalverteilte und positiv korrelierte Zufallsvariablen (y, x). Im Falle gemeinsam normalverteilter Zufallsvariablen ist die CEF linear und die CEF Störterme sind homoskedastisch.

Bevor wir die Parameter der CEF schätzen können müssen wir *Annahmen* treffen, wie die CEF aussieht. Dies wird in der Ökonometrie **Spezifikation** genannt und gehört zu den schwierigsten Problemen, da es dafür wenige harte Regeln gibt. Das Problem erinnert an das bekannte Goethe Wort *"man sieht nur, was man weiß."* Erst wenn wir die Funktionsform und die erklärenden Variablen a priori festgelegt haben können wir die dazugehörigen Parameter aus den Daten schätzen. Aber wenn wir mit einer falschen Annahme zur CEF begonnen haben sind die Schätzungen von zweifelhaftem Wert, *'garbage in, garbage out'*.

Unterschiedlich gewählte Spezifikationen erklären zum Teil auch, weshalb verschiedene Forscher bei der Untersuchung der gleichen Fragen zu so unterschiedlichen Ergebnissen kommen.

Deshalb legen Ökonometriker in der Regel so viel Wert auf die theoretischen Grundlagen, die zumindest Hinweise auf die zu berücksichtigenden Variablen geben soll. Überhaupt starten Ökonometrikerinnen ihre Arbeit meist mit einem theoretischen Modell der interessierenden Zusammenhänge und entwickeln daraus die Spezifikation des ökonometrischen Modells. In anderen Worten, sie starten häufig mit einer theoretischen Vorstellung, wie die CEFs aussehen, und sehen die gemeinsame Verteilung eher als eine Folge.

Im Gegensatz dazu beginnen Statistiker häufig mit einer Analyse der gemeinsamen Verteilung, und suchen ausgehend von der gemeinsamen Verteilung eine möglichst adäquate Beschreibung der gemeinsamen Verteilung in Form der bedingten Momentfunktionen.

Aufgrund dieser unterschiedlichen Interessenslage wählen Ökonometriker häufig einfachere Spezifikationen, die eine bessere Interpretierbarkeit der Ergebnisse gewähr-

leisten, die aber auf Kosten der Genauigkeit der Abbildung der Daten geht, die manchmal von Statistikerinnen bevorzugt wird.

Ein mit der Spezifikation zusammenhängendes Problem ist die **Identifikation**. Im wesentlichen geht es bei der Identifikation darum, ob die Kenntnis der gemeinsamen Verteilung ausreicht, die interessierenden Parameter zu berechnen. Wohl gemerkt, dabei handelt es sich um kein Schätzproblem, sondern um ein logisches Problem.

Stellen Sie sich vor, dass zwei Ereignisse immer gemeinsam auftreten und mit einem dritten Ereignis zusammenhängen. Dann haben wir keine Chance festzustellen, wie groß der Einfluss eines einzelnen Ereignisses ist. Ein klassisches Beispiel ist perfekte Multikollinearität.

Stellen Sie sich die PRF $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ vor, und es gelte $x_3 = cx_2$, d.h. die Regressoren sind linear abhängig. Aufgrund der vorhergehenden linearen Abhängigkeit zwischen den Regressoren, d.h. perfekter Multikollinearität, können wir diese Regression nicht schätzen, aber wir können einsetzen und die PRF schreiben als

$$\begin{aligned} y &= \beta_1 + (\beta_2 + c\beta_3)x_2 + \varepsilon \\ &= \beta_1 + (\frac{1}{c}\beta_2 + \beta_3)x_3 + \varepsilon \end{aligned}$$

Mit Hilfe dieser Regression können wir zwar die gewichtete Summe der Koeffizienten $\beta_2 + c\beta_3$ aus den Daten schätzen, aber es gibt keine Möglichkeit β_2 oder β_3 einzeln zu schätzen! Die Koeffizienten β_2 oder β_3 sind einzeln nicht identifiziert.

Ein Beispiel zum Identifikationsproblem: Angenommen, sie möchten eine Nachfragefunktion schätzen und verfügen über Daten mit den in der Vergangenheit beobachteten Mengen und Preisen.

Nichts scheint nahe liegender als eine einfache Regressionsgerade in die Punktewolke hineinzulegen. Panel a) in Abbildung 3.17 zeigt das Ergebnis.

Aber wer sagt Ihnen, dass es sich dabei tatsächlich um eine Nachfragefunktion handelt, und nicht um eine Angebotsfunktion wie in Panel b)?

Tatsächlich ist jeder einzelne beobachtete Punkt ein Schnittpunkt einer Nachfrage- und Angebotsfunktion (siehe Panel c)), und man kann zeigen, dass die mit OLS geschätzten Koeffizienten ein gewichtetes Mittel der Koeffizienten von Angebots- und Nachfragefunktion sind, die Gewichte hängen nur von den Varianzen von Menge und Preis ab.

Dies ist ein klassisches Beispiel für das Identifikationsproblem; wenn wir nur die Daten für Mengen und Preise zur Verfügung haben gibt es keine Möglichkeit, aus diesen Daten die Koeffizienten von Angebots- und Nachfragefunktion zu schätzen, selbst wenn wir deren gemeinsame Verteilung kennen.

Die Daten enthalten in diesem Fall einfach nicht genügend Information um die Steigungen von Angebots- und Nachfragefunktion einzeln zu 'identifizieren'; eine unendliche Anzahl von Angebots- und Nachfragefunktionen sind mit diesen Daten kompatibel.

Um z.B. die Nachfragefunktion identifizieren zu können ist zusätzliche Information erforderlich, und um zu erkennen welche Information erforderlich ist, benötigt man

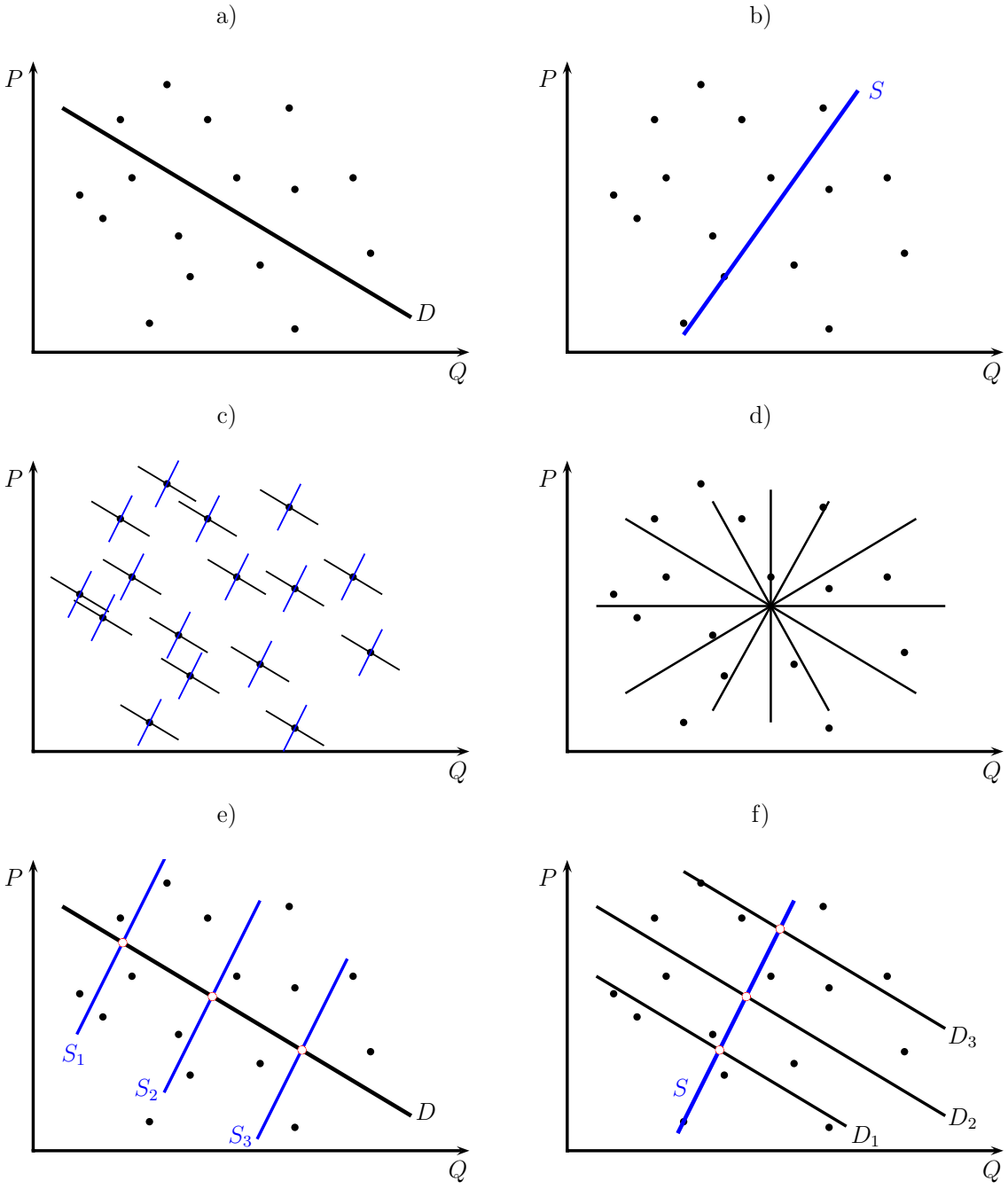


Abbildung 3.17: Das Identifikationsproblem.

Theorie. Stellen wir uns vor, bei dem untersuchten Markt handelt es sich um einen Kartoffelmarkt. Wir wissen, dass das Kartoffelangebot u.a. von den Wetterbedingungen abhängt, und dass die Wetterbedingungen exogen sind. Außerdem vermuten wir, dass die Wetterbedingungen nur einen vernachlässigbar kleinen Einfluss auf die Kartoffelnachfrage haben. Das bedeutet, dass Wetterschwankungen zwar die Angebotsfunktion verschieben, aber keinen Einfluss auf die Nachfragefunktion haben.

Wenn zusätzliche Daten für die Wetterbedingungen zur Verfügung stehen wird die Nachfragefunktion identifizierbar, und damit schätzbar, denn die Verschiebung der Angebotsfunktion erlaubt einen Rückschluss auf die Lage und Steigung der Nachfragefunktion (siehe Abbildung 3.17, Panel e)). Diese zusätzliche Information über die Angebotsfunktion hilft aber nicht für die Identifikation der Angebotsfunktion, diese ist trotzdem nicht schätzbar.

Wenn aber z.B. zusätzlich Daten über das Einkommen der Konsumenten verfügbar wären, und das Einkommen nur die Nachfragefunktion verschiebt, aber keinen Einfluss auf die Angebotsfunktion hat, so wird dadurch auch die Angebotsfunktion identifizierbar.

Die Wetterbedingungen und das Einkommen sind Beispiele für sogenannte *Instrumentvariablen*, durch ihren Einsatz können die Koeffizienten von Angebots- und Nachfragefunktion identifizierbar werden. Solche Identifikationsprobleme und die Entwicklung spezieller Schätzverfahren für simultane Gleichungssysteme standen an der Wiege der Ökonometrie als eigene Wissenschaft.

Identifikationsprobleme nehmen in der Ökonometrie einen zentralen Stellenwert ein, insbesondere wenn es um Fragen wie Kausalität geht. Wir werden diese Fragen in einem späteren Kapitel ausführlicher diskutieren!

3.4.2 Eine Übersicht

Nun haben wir die wichtigsten Bausteine beisammen und wir können anhand von Abbildung 3.18 die stochastische Regressionsanalyse zusammenfassen.

Am Anfang jeder ökonometrischen Arbeit steht natürlich eine Forschungsfrage. Durch Nachdenken, Literaturstudium, Diskussion mit Kolleginnen und häufig etwas Rechnerei sollte es gelingen, die Forschungsfrage etwas enger zu fassen und idealerweise in ein theoretisches Modell zu kleiden.

Aufgrund dieser Überlegungen entscheiden wir, welche verfügbaren Daten für unsere Forschungsfrage relevant sind. Auf Grundlage der theoretischen Überlegungen zur Forschungsfrage und der Verfügbarkeit von Daten versuchen wir eine geeignete Spezifikation zu finden, von der wir glauben, dass sie den zugrunde liegenden Datengenerierenden Prozess adäquat beschreibt und die eine Identifikation der interessierenden Parameter gestattet. Ist eine solche Spezifikation gefunden können wir überlegen, welche Schätzfunktion am ehesten geeignet ist.

Schätzfunktionen werden von theoretischen Ökonometrikern entwickelt. Dazu wird in der Regel jede einzelne Realisation der Stichprobe als Zufallsvariable interpretiert, man geht also gewissermaßen in den ex ante Zustand vor Durchführung des Zufallsexperiments zurück. So erhalten wir n verschiedenen (Paare von) Zufallsvariablen (y_i, x_i). Gemeinsam mit spezifischen Annahmen über den Datengenerierenden Prozess werden daraus Schätzfunktionen ermittelt.

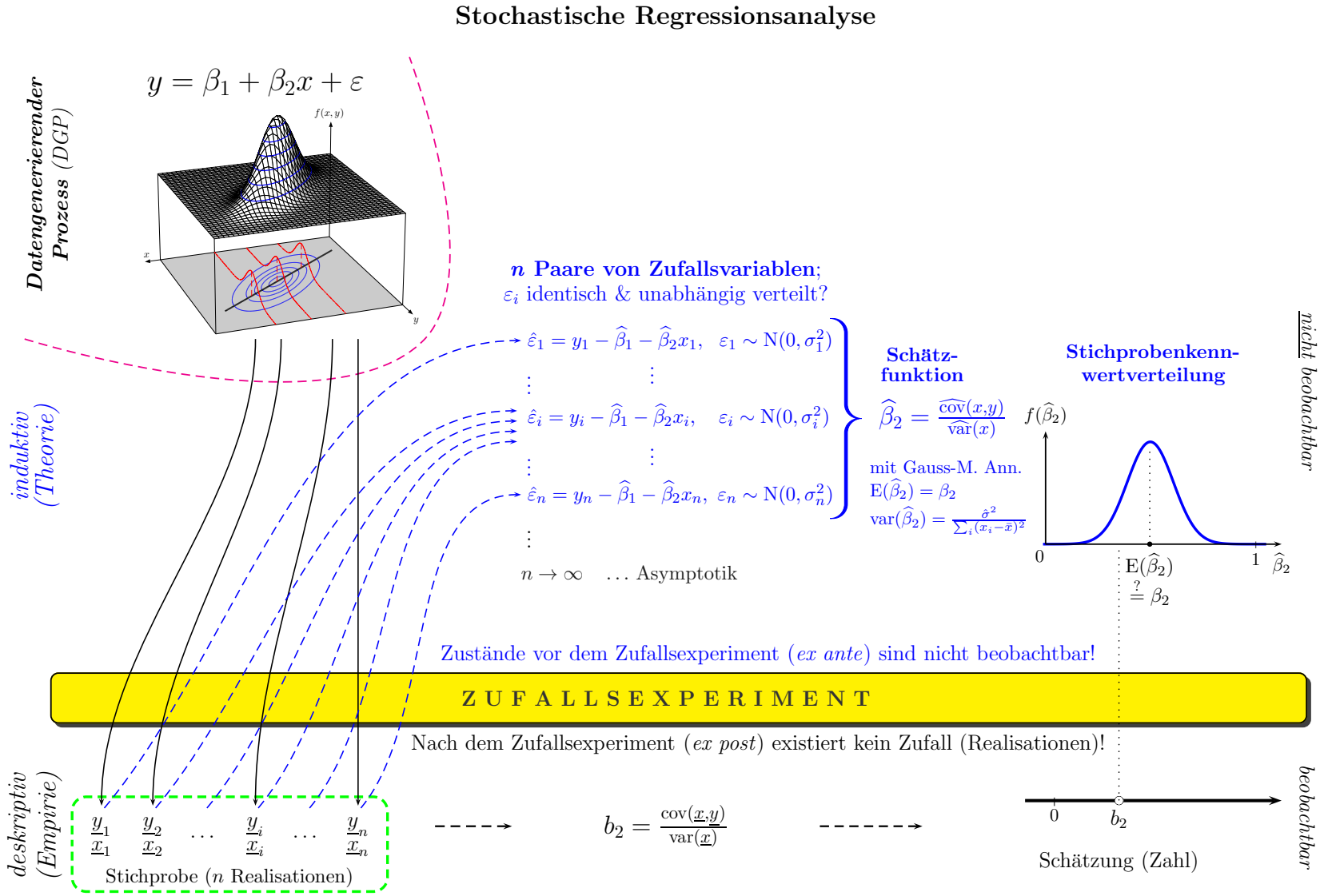


Abbildung 3.18: Stochastische Regressionsanalyse.

Fast alle in der Ökonometrie gebräuchlichen Schätzer beruhen auf drei Methoden zur Gewinnung von Schätzfunktionen

1. OLS Methode,
2. Maximum Likelihood Methode, und
3. Methode der Momente sowie deren Verallgemeinerung GMM.

Mit Hilfe einer dieser drei Methoden wird aus diesen Zufallsvariablen eine Schätzfunktion ermittelt, die selbst wieder eine Zufallsvariable ist.

Die OLS Methode liefert z.B. mit einem spezifischen Annahmenset (unter anderem, dass die einzelnen Störterme identisch und unabhängig verteilt sind, $\varepsilon_i \sim$ i.i.d.$(0, \sigma^2)$), den üblichen OLS Schätzer $\widehat{\beta}_2 = \widehat{\text{cov}}(y, x)/\widehat{\text{var}}(x)$.

Unterschiedliche Annahmen und Methoden führen zu unterschiedlichen Schätzfunktionen. Als Zufallsvariablen haben Schätzfunktionen eine Verteilung, die wir *Stichprobenkennwertverteilung* (oder einfacher 'Stichprobenverteilung' bzw. *'sampling distribution'*) nennen. Auf Grundlage dieser Stichprobenkennwertverteilungen versucht die theoretische Ökonometrie auch die allgemeinen Eigenschaften der Schätzfunktionen zu bestimmen, wie z.B. Erwartungstreue, Effizienz oder Konsistenz. Darüber hinaus werden von der theoretischen Ökonometrie auch geeignete Teststatistiken entwickelt.

Die anwendungsorientierte Ökonometrikerin muss sich aufgrund spezifischer Annahmen über den Datengenerierenden Prozess, der Datenlage und der gewünschten Eigenschaften für eine Schätzfunktion entscheiden. Selbst wenn man keine eigenen Schätzfunktionen entwickelt ist für die Entscheidung, welche der vorhandenen Schätzfunktionen für das spezifische Problem am besten geeignet ist, in der Regel einiges an theoretischem Wissen erforderlich.

Wenn wir uns schließlich für eine Spezifikation und eine Schätzfunktion entschieden haben können wir diese auf die beobachtete Stichprobe anwenden und erhalten als Ergebnis eine Schätzung (z.B. b_2).

Aufgrund der theoretischen Überlegungen wissen wir, dass wir diese Schätzung als eine Realisation aus einer Stichprobenkennwertverteilung interpretieren können. Diese Sichtweise wird uns in weiterer Folge u.a. auch die Durchführung von Hypothesentests erlauben.

Zuvor werden wir aber im nächsten Kapitel die Momente der Stichprobenkennwertverteilung des OLS Schätzers ermitteln und deren Eigenschaften bestimmen, insbesondere, inwieweit diese Schätzfunktion *erwartungstreu*, *effizient* und *konsistent* ist.

3.A Appendix

3.A.1 R Programmcode für Monte Carlo Simulation

Das folgende Programm zieht aus einer gegebenen Grundgesamtheit von 60 Gebrauchtautos tausend Stichproben mit je 7 Beobachtungen, berechnet für jede dieser 1000 Stichproben den Koeffizienten $\widehat{\beta}_2$ und zeichnet ein Histogramm für die 1000 $\widehat{\beta}_2$. Das Histogramm in Abbildung 3.6 (Seite 8) wurde mit diesem Programm erzeugt.

```
# Eine einfache Monte Carlo Simulation
# Beispiel mit Gebrauchtautos
# 27.10.2016

rm(list=ls(all=TRUE))
#setwd("C:/mydirectory/")

d <- read.csv2("http://www.hsto.info/econometrics/data/auto60.csv")
eq.POP <- lm(Preis ~ Alter, data = d)

set.seed(1234567)      # Zufallgenerator initialisieren
reps <- 1000           # Replikationen
sz <- 7  # sample size

# Matrix mit Namen R anlegen um Resultate b1 und b2 zu speichern
R <- matrix(rep(NA, 2*reps), ncol = 2)
Rsmpl <- matrix(rep(NA,reps*sz), ncol=sz) # Matrix für Ziehungen
d$OBS <- 1:nrow(d)  # Index für Beobachtungen anlegen

for (r in 1:reps) {
  s <-  d[sample(1:nrow(d), sz, replace=FALSE), ]
  eq.SRF <- lm(Preis ~ Alter, data = s)
  R[r,1] <- summary(eq.SRF)$coefficients[1,1]  # b1
  R[r,2] <- summary(eq.SRF)$coefficients[2,1]  # b2
  Rsmpl[r, ] <- s[ , "OBS"]  # Index der Beobachtungen für das Sample
} # Ende der Schleife mit Laufindex r (Reps.)

# Histogramm für b2
b2 <- R[,2]
X11()
hist(b2, breaks = 30, freq = FALSE, col=gray(0.9),
     main=expression(paste("Histogramm (empir. Verteilung) von ",
           hat(beta)[2], " und Normalverteilung (theoret. Verteilung)")),
     cex.main=0.9,
     xlab = expression(paste("Monte Carlo Simulation ", hat(beta)[2])))
curve(dnorm(x,mean = mean(b2),sd = sd(b2)), add = TRUE, col="blue", lwd=2, lty=2)
#dev.off()

# Sample Regression Functions:
eq.SRF1 <- lm(Preis ~ Alter, data = subset(d, OBS %in% Rsmpl[which.max(b2),]))
eq.SRF2 <- lm(Preis ~ Alter, data = subset(d, OBS %in% Rsmpl[which.min(b2),]))
eq.SRF3 <- lm(Preis ~ Alter, data = subset(d, OBS %in% Rsmpl[3,]))
```