Vorlesung: Statistik II für Wirtschaftswissenschaft

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

Sommersemester 2017

Lineare Regressionsmodelle

Deskriptive Statistik:

Gegeben Datenpunkte (Y_i, X_i) schätze die beste Gerade

$$Y_i = \beta_0 + \beta_1 X_i, \ i = 1, \ldots, n.$$

(mit der Methode der kleinsten Quadrate)

Statistisches Modell

- Linearer Zusammenhang
- Im Folgenden:
  Probabilistische Modelle in Analogie zu den deskriptiven Modellen aus Statistik I

Lineare Einfachregression

Zunächst Modelle mit nur *einer* unabhängigen Variable.

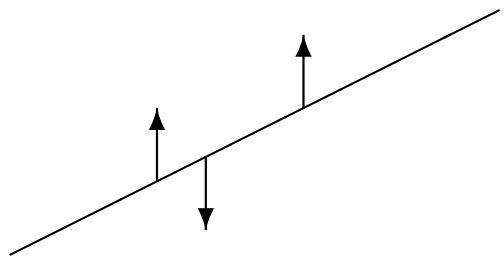
Statistische Sichtweise:

- Modell

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

β_1 Wirkung der Änderung von X_i um eine Einheit auf Y

- gestört durch zufällige Fehler ε_i



Modellannahmen

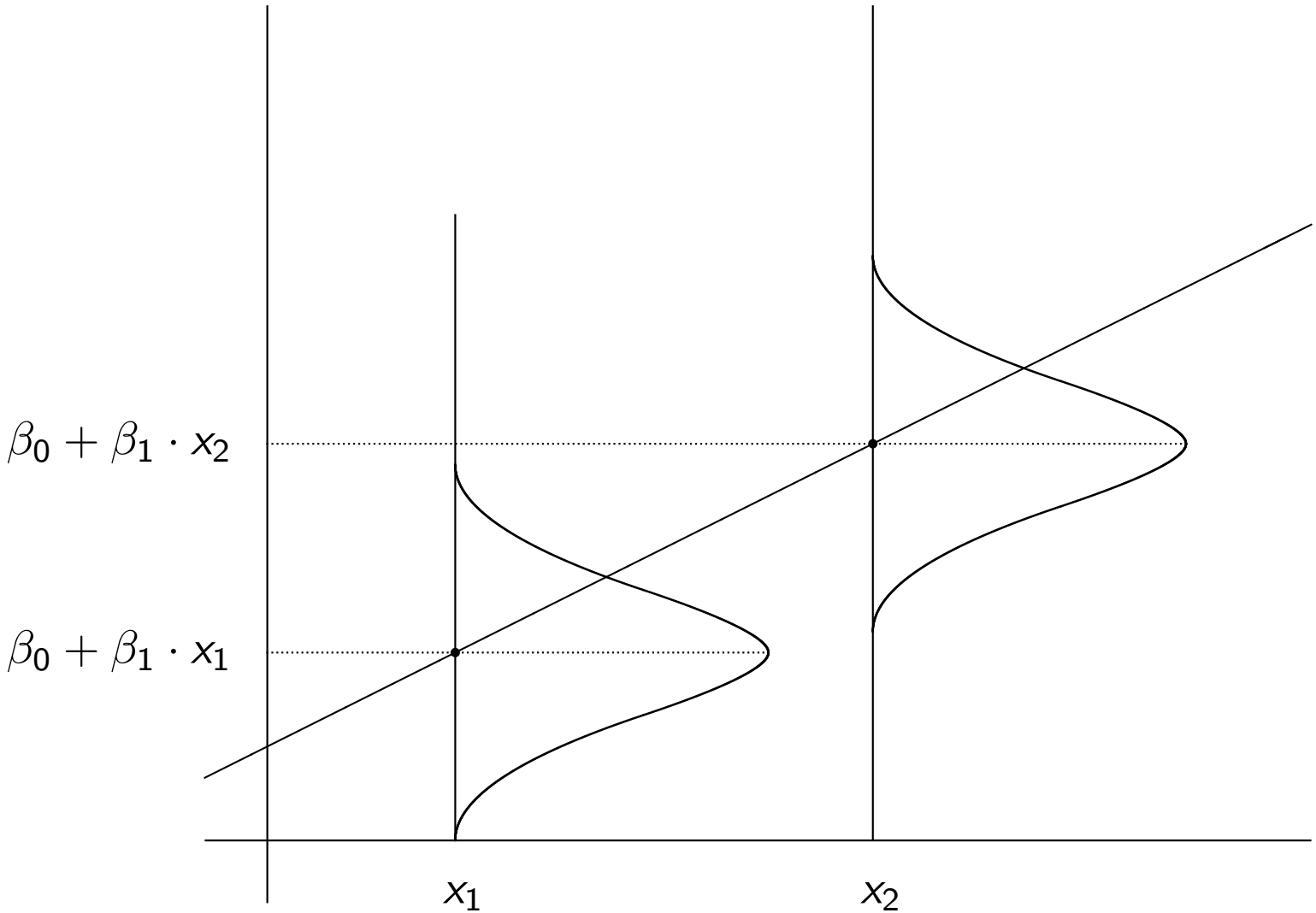
Beobachtung von Datenpaaren (X_i, Y_i), $i = 1, \ldots, n$ mit

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

wobei sich die Annahmen auf den zufälligen Störterm beziehen:

- $E(\varepsilon_i) = 0$
- $Var(\varepsilon_i) = \sigma^2$ für alle i gleich
- $\varepsilon_{i1}, \varepsilon_{i2}$ stochastisch unabhängig für $i_1 \neq i_2$
- $\varepsilon_i \sim N(0, \sigma^2)$ (zusätzlich, bei großen Stichproben nicht erforderlich)

Lineare Einfachregression



Schätzung der Parameter

Die Schätzwerte werden üblicherweise mit $\hat{\beta}_0, \hat{\beta}_1$ und $\hat{\sigma}^2$ bezeichnet.
In der eben beschriebenen Situation gilt:

- Die (Maximum Likelihood) Schätzer entsprechen den KQ-Schätzer aus Statistik 1

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}, \\
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2
\end{aligned}$$

mit den Residuen

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

Konstruktion von Testgrößen

Mit

$$\hat{\sigma}_{\hat{\beta}_0} := \frac{\hat{\sigma}\sqrt{\sum_{i=1}^n X_i^2}}{\sqrt{n\sum_{i=1}^n (X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t^{(n-2)}$$

und analog mit

$$\hat{\sigma}_{\hat{\beta}_1} := \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

gilt

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t^{(n-2)}.$$

Konfidenzintervalle

- $\hat{\beta}_0$ und $\hat{\beta}_1$ sind die *KQ*-Schätzer aus Statistik I. Unter Normalverteilung fällt hier das *ML*- mit dem *KQ*-Prinzip zusammen.
- Man kann unmittelbar Tests und Konfidenzintervalle ermitteln (völlig analog zum Vorgehen, das bei den t- Tests verwendet wurde

  Konfidenzintervalle zum Sicherheitsgrad γ:

$$\text{für } \beta_0 : \quad [\hat{\beta}_0 \pm \hat{\sigma}_{\hat{\beta}_0} \cdot t_{\frac{1+\gamma}{2}}^{(n-2)}]$$

$$\text{für } \beta_1 : \quad [\hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} \cdot t_{\frac{1+\gamma}{2}}^{(n-2)}]$$

Tests für die Parameter des Modells

Mit der Teststatistik

$$T_{\beta_1^*} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}_{\hat{\beta}_1}}$$

ergibt sich

	Hypothesen			kritische Region
I.	$H_0 : \beta_1 \leq \beta_1^*$	gegen	$\beta_1 > \beta_1^*$	$T \geq t_{1-\alpha}^{(n-2)}$
II.	$H_0 : \beta_1 \geq \beta_1^*$	gegen	$\beta_1 < \beta_1^*$	$T \leq t_{1-\alpha}^{(n-2)}$
III.	$H_0 : \beta_1 = \beta_1^*$	gegen	$\beta_1 \neq \beta_1^*$	$\lvert T \rvert \geq t_{1-\frac{\alpha}{2}}^{(n-2)}$

(analog für $\hat{\beta}_0$).

Von besonderem Interesse ist der Fall $\beta_1^* = 0$ (Steigung gleich 0): Hiermit kann man überprüfen, ob die $X_1, \ldots, X_n$ einen signifikanten Einfluss hat oder nicht.

Beispiel : Mietspiegel

Call:
lm(formula = nmqm wfl, data = mietsp2015)
Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	11.72	0.46	26.286	$< 2e-16$
wfl	-0.0226	0.005787	-3.905	< 0.00012

Multiples Regressionsmodell

Beispiel: Mietspiegel

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

mit

$$\begin{aligned} X_1 &= \begin{cases} 1 & \text{Gute Lage} \\ 0 & \text{schlechte Lage} \end{cases} \\ X_2 &= \text{Wohnfläche} \\ Y &= \text{Miete} \end{aligned}$$

Multiples Regressionsmodell: Interpretation

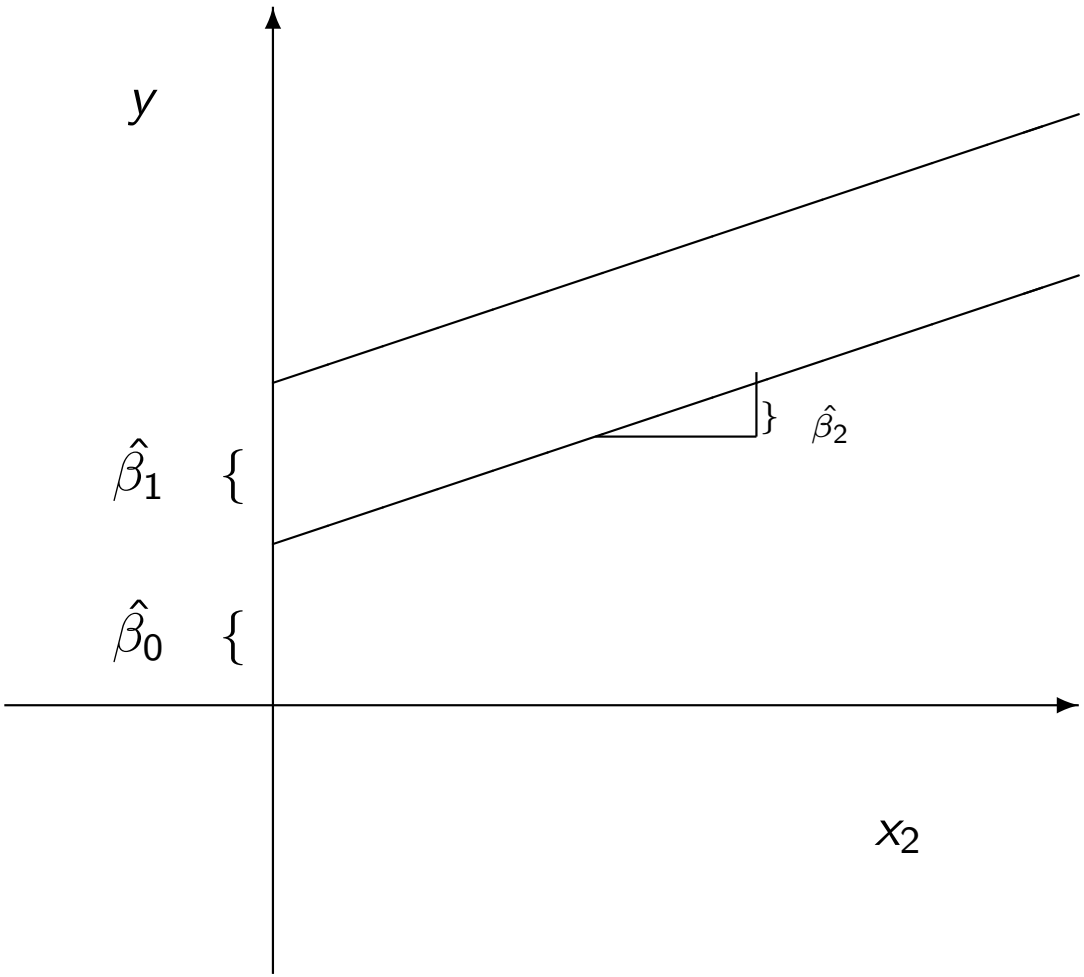
- Geschätzte Regressionsgerade für gute Lage

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot x_{2i}$$

- Geschätzte Regressionsgerade für die schlechte Lage :

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot x_{2i} \\ &= \hat{\beta}_0 + \hat{\beta}_2 \cdot x_{2i} \end{aligned}$$

Grundidee (ANCOVA)



Mehr als 2 Gruppen

Lösungsansatz

Hier ist eine direkte Lösung nicht sinnvoll.
Grundidee:

- aus einem nominalen Regressor mit k Merkmalsausprägungen
- $k - 1$ neue Regressoren (Dummys) gebildet werden.
- Eine Merkmalsausprägung des ursprünglichen Regressors wird zur *Referenzkategorie*.

Nominale Regressoren

Dummykodierung

Nach Wahl der Referenzkategorie $j \in \{1, ..., k\}$ ergeben sich die Dummys $X_i, i = 1, ..., k$ und $i \neq j$ mit folgenden Werten:

$$x_i = \begin{cases} 1 & \text{falls Kategorie } i \text{ vorliegt,} \\ 0 & \text{sonst.} \end{cases}$$

Nominale Regressoren

Beispiel

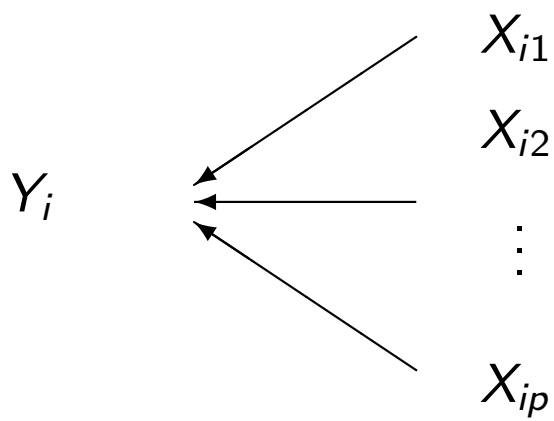
Gegeben seien folgende Daten:

lfd Nr.	Alter	Studienfach
1	19	BWL
2	22	Sonstige
3	20	VWL
⋮	⋮	⋮

Mit der Kodierung BWL = 1, VWL = 2, Sonstige = 3 erhalten wir bei Wahl der Referenzkategorie = 3 (Sonstige) zwei Dummys X_1 (für BWL) und X_2 (für VWL) gemäß folgendem Schema:

Ausprägung von X	Wert von X_1	Wert von X_2
1 BWL	1	0
2 VWL	0	1
3 Sonstige	0	0

Multiples Regressionsmodell



abhängige Variable	unabhängige Variablen
metrisch/quasistetig	metrische/quasistetige oder dichotome (0/1) Variablen (kategoriale Variablen mit mehr Kategorien $\rightarrow$ Dummy-Kodierung)

Multiple lineare Regression

- Analoger Modellierungsansatz, aber mit mehreren erklärenden Variablen:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

- Schätzung von $\beta_0, \beta_1, \ldots, \beta_p$ und σ^2 sinnvollerweise über Matrixrechnung bzw. Software.

  Aus dem R-Output sind $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ sowie $\hat{\sigma}_{\hat{\beta}_0}, \hat{\sigma}_{\hat{\beta}_1}, \ldots, \hat{\sigma}_{\hat{\beta}_p}$ ablesbar.

Schätzung im multiplen Modell

- Darstellung in Matrix-Form
- KQ- Methode und Maximum-Likelihood - Methode stimmen überein
- Berechnung effizient mit Matrix-Kalkül
- Zu den Parametern können jeweils die Standardfehler geschätzt werden.

Multiple lineare Regression

Es gilt für jedes $j = 0, \ldots, p$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t^{(n-p-1)}$$

und man erhält wieder Konfidenzintervalle für β_j:

$$[\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} \cdot t^{(n-p-1)}_{\frac{1+\gamma}{2}}]$$

sowie entsprechende Tests.

Multiple lineare Regression: Tests

Von besonderem Interesse ist wieder der Test

$$H_0 : \beta_j = 0, \;\; H_1 : \beta_j \neq 0.$$

Der zugehörige p-Wert findet sich im Ausdruck (Vorsicht mit Problematik des multiplen Testens!).
Man kann auch simultan testen, z.B.

$$\beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

Dies führt zu einem sogenannten F-Test ($\longrightarrow$ Software).

Sind alle X_{ij} 0/1-wertig, so erhält man eine sogenannte *Varianzanalyse*, was dem Vergleich von mehreren Mittelwerten entspricht.

Varianzanalyse (Analysis of Variance, ANOVA)

- Vor allem in der angewandten Literatur, etwa in der Psychologie, wird die Varianzanalyse unabhängig vom Regressionsmodell entwickelt.
- Ziel: Mittelwertvergleiche in mehreren Gruppen, häufig in (quasi-) experimentellen Situationen.
- Verallgemeinerung des t-Tests. Dort nur zwei Gruppen.
- Hier nur *einfaktorielle Varianzanalyse* (Eine Gruppierungsvariable).

Varianzanalyse: Beispiel

Einstellung zu Atomkraft anhand eines Scores, nachdem ein Film gezeigt wurde.

3 Gruppen („Faktorstufen“):

- Pro-Atomkraft-Film
- Contra-Atomkraft-Film
- ausgewogener Film

Varianzanalyse: Vergleich der Variabilität in und zwischen den Gruppen

Beobachtungen: Y_{ij}

$$\begin{array}{ll} j = 1, \ldots, J & \text{Faktorstufen} \\ i = 1, \ldots, n_j & \text{Personenindex in der } j\text{-ten Faktorstufe} \end{array}$$

Modellformulierung

Modell (Referenzcodierung):

$$Y_{ij} = \mu_J + \beta_j + \varepsilon_{ij} \qquad j = 1, \ldots, J,\ i = 1, \ldots, n_j,$$

mit

μ_J	Mittelwert der Referenz
β_j	Effekt der Kategorie j im Vergleich zur Referenz J
ε_{ij}	zufällige Störgröße
ε_{ij}	$\sim N(0, \sigma^2)$, $\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{Jn_J}$ unabhängig.

Testproblem:

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots \beta_{j-1} = 0$$

gegen

$$H_1 \quad : \quad \beta_j \neq 0 \quad \text{für mindestens ein j}$$

Streuungszerlegung

Mittelwerte:

$\bar{Y}_{\bullet\bullet}$ Gesamtmittelwert in der Stichprobe
$\bar{Y}_{\bullet j}$ Mittelwert in der j-ten Faktorstufe

Es gilt (vgl. Statistik I) die Streuungszerlegung:

$$\sum_{j=1}^{J}\sum_{j=1}^{n_j}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2 = \sum_{j=1}^{J}\underbrace{n_j(\bar{Y}_{\bullet j}-\bar{Y}_{\bullet\bullet})^2}_{=SQE} + \underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{\bullet j})^2}_{=SQR}$$

Variabilität **der** Gruppen

Variabilität **in den** Gruppen

F-Test

Die **Testgröße**

$$F = \frac{SQE/(J-1)}{SQR/(n-J)}$$

ist geeignet zum Testen der Hypothesen

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots \beta_{j-1} = 0$$

gegen

$$H_1 \quad : \quad \beta_j \neq 0 \quad \text{für mindestens ein j}$$

Testprozedur

- **Kritische Region**: *große* Werten von F

  Also H_0 ablehnen, falls

  $$T > F_{1-\alpha}(J-1, n-J),$$

  mit dem entsprechenden $(1-\alpha)$-Quantil der F-Verteilung mit $(J-1)$ und $(n-J)$ Freiheitsgraden.

- (Je größer die Variabilität zwischen den Gruppen im Vergleich zu der Variabilität in den Gruppen, desto unplausibler ist die Nullhypothese, dass alle Gruppenmittelwerte gleich sind.)

- Bei Ablehnung des globalen Tests ist dann oft von Interesse, welche Gruppen sich unterscheiden.

  $\Rightarrow$ Testen spezifischer Hypothesen über die Effekte β_j. Dabei tritt allerdings die Problematik des multiplen Testens auf.

Zusammenfassung

- Testen von Regressionsmodellen wesentliches Werkzeug
- Gleichzeitige Berücksichtigung vieler Einflüsse möglich
- Viel Möglichkeiten zum Testen (F-Tests)
- Regressionsmodell Ausgangspunkt für viele neue Verfahren (Big Data, Algorithmen, KI)