Kapitel 8

Einfache Regression

Lernziele

- Lineares Regressionsmodell
- Anpassen des linearen Regressionsmodells, OLS
- Eigenschaften der Schätzer für das Modell
- Testen und Konfidenzintervalle
- Überprüfung der Modellannahmen
- Prognose mittels linearer Regression
- Multiples Regressionsmodell

Das lineare Regressionsmodell

Das einfache lineare Regressionsmodell lautet

$$y_t = \alpha + \beta x_t + \varepsilon_t \qquad t = 1, \ldots, n$$

- y_t ... abhängige Variable
- x_t ... unabhängige Variable
- ε_t ... Fehler, Residuum
- α ... Interzept, Konstante
- β ... Steigung

Modellvoraussetzung

Annahmen für das einfache lineare Regressionsmodell:

- Die Beziehung zwischen y und x ist linear.
- Die x_t sind deterministische (nicht-stochastische) Variablen. Ihre Werte sind genau bekannt.
- Die Residuen ε_t sind
  - unkorreliert,
    $$\mathrm{Cov}(\varepsilon_{t_i}, \varepsilon_{t_j}) = 0 \qquad \text{für } i \neq j$$
  - haben Erwartungswert 0 und konstante Varianz,
    $$\mathrm{E}(\varepsilon_t) = 0, \qquad \mathrm{V}(\varepsilon_t) = \sigma_\varepsilon^2, \qquad \text{für alle } t$$
  - und sind normalverteilt,
    $$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

Der bedingte Erwartungswert

Der bedingte Erwartungswert $\mathrm{E}(y_t|x_t)$ von y_t gegen x_t ist

$$\mathrm{E}(y_t|x_t) = \alpha + \beta\, x_t$$

da $\mathrm{E}(\varepsilon_t) = 0$ und x_t bekannt ist.
Das Mittel von y_t wird als Funktion von x_t angesehen.

- $\mathrm{E}(y_t)$ bezeichnet das univariate (**unbedingte**) Mittel von y_t, i.e., der Mittelwert aller y_t über die Zeit wenn die Information x_t nicht verwendet/bekannt ist.
- $\mathrm{E}(y_t|x_t)$ ist eine **bedingte Erwartung**. Die Information x_t wird zur Erklärung/Modellierung von y_t verwendet.

Die bedingte Varianz

- Die **unbedingte Varianz** ergibt sich als
  $$\mathrm{V}(y_t) = \mathrm{E}([y_t - \mathrm{E}(y_t)]^2)$$
- Die **bedingte Varianz** $\mathrm{V}(y_t|x_t)$ ist analog
  $$\mathrm{V}(y_t|x_t) = \mathrm{E}\left((y_t - \mathrm{E}(y_t|x_t))^2|x_t\right) = \sigma_\varepsilon^2$$

In unserem Modell ist die bedingte Varianz von y_t konstant:

$$y_t - \mathrm{E}(y_t|x_t) = (\alpha + \beta\, x_t + \varepsilon_t) - (\alpha + \beta\, x_t) = \sigma_\varepsilon^2$$

Und somit

$$\mathrm{V}(y_t|x_t) = \mathrm{E}(\varepsilon_t^2|x_t) = \mathrm{E}(\varepsilon_t^2) = \sigma_\varepsilon^2$$

Varianzreduktion

Zur Berechnung von $\mathrm{V}(y_t|x_t)$ wird die Information x_t verwendet.
Daher gilt immer (sofern das Modell einen Erklärungswert hat)

$$\mathrm{V}(y_t) > \mathrm{V}(y_t|x_t) = \sigma_\varepsilon^2$$

Die Regression reduziert die Varianz in den Daten.
Sie erklärt somit einen Teil der Variantion in y_t.

Anpassen eines Regressionsmodells

Die Parameter des Regressionsmodells werden so gewählt, daß die **Fehlerquadratsumme der Residuen minimal** wird.

$$\min_{\alpha,\beta} \sum_{t=1}^{n} \varepsilon_t^2 = \min_{\alpha,\beta} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 = \min_{\alpha,\beta} \sum_{t=1}^{n} (y_t - (\alpha + \beta\, x_t))^2$$

Das Minimierungsproblem wird mit der Methode der kleinsten Quadrate (***ordinary least squares***, **OLS**) gelöst.

Man berechnet dazu die ersten partiellen Ableitungen und setzt diese gleich Null.

$$\hat{\alpha} = \frac{1}{n}\sum y_t - \hat{\beta}\,\frac{1}{n}\sum x_t \qquad \hat{\beta} = \frac{n\sum x_t\, y_t - \sum x_t \sum y_t}{n\sum x_t^2 - (\sum x_t)^2}$$

OLS-Schätzer

Man kann diese Schätzer auch anders darstellen:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{\mathrm{Cov}(x,y)}{\mathrm{V}(x)} = \mathrm{Corr}(x,y)\,\frac{s_y}{s_x}$$

Der Schätzer für β wird Null, falls x_t und y_t unkorrelliert sind. Dann wird die Variable x_t nicht in die Regressionsgleichung aufgenommen.

Nach der Schätzung lautet unser Modell:

$$y_t = \hat{\alpha} + \hat{\beta}\, x_t + \hat{\varepsilon}_t$$

$\hat{\varepsilon}_t$ sind die geschätzten Fehler.

Verteilung der Schätzer

Unter Annahme (Nullhypothese), daß unser Modell das *wahre Modell* ist, sind die OLS-Schätzer $\hat{\alpha}$ und $\hat{\beta}$ normalverteilt mit Erwartungswert α bzw. β:

$$\hat{\alpha} \sim N(\alpha, \sigma_\alpha^2) \qquad \hat{\beta} \sim N(\beta, \sigma_\beta^2)$$

mit

$$\sigma_\alpha^2 = \sigma_\varepsilon^2 \frac{\sum x_t^2}{n \sum (x_t - \bar{x})^2} = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_t - \bar{x})^2} \right)$$

$$\sigma_\beta^2 = \sigma_\varepsilon^2 \frac{1}{\sum (x_t - \bar{x})^2}$$

Verteilung der Schätzer / (2)

σ_ε^2 wird durch die **Stichprobenvarianz des Fehlers** s^2 ersetzt:

$$\hat{\sigma}_\varepsilon^2 = s^2 = \frac{1}{n-2} \sum \hat{\varepsilon}_t^2, \qquad \text{mit } \hat{\varepsilon}_t = y_t - (\hat{\alpha} + \hat{\beta}\, x_t)$$

Damit wird erhalten wir die Schätzer $\hat{\sigma}_\alpha^2$ und $\hat{\sigma}_\beta^2$ für σ_α^2 bzw. σ_β^2.

Eigenschaften der OLS-Schätzer

Die OLS-Lösung liefert den **BLUE**, den besten linearen unverzerrten Schätzer (***best linear unbiased estimator***):

- **Best**: der Schätzer hat die kleinste Varianz.
- **Linear**: das Modell ist linear.
- **Unbiased** oder **unverzerrt**: Erwartungswert des Schätzers ist der Wert des Parameters ist.
  (Ansonsten wäre er verzerrt.)
- **Estimator**: Schätzer

t-Test der geschätzten Parameter

Wir können obige Verteilung benutzen um die Hypothese

$$\mathsf{H}_0 \colon \beta = b \qquad \mathsf{H}_1 \colon \beta \neq b$$

zu testen. Die Prüfgröße (*Teststatistik*)

$$\frac{\hat{\beta} - b}{\hat{\sigma}_\beta} \sim t_{n-2}$$

ist dabei t-verteilt mit $n - 2$ Freiheitsgraden.

(Analog für α.)

Konfidenzintervalle

Das $(1 - \alpha)$-Konfidenzintervall für β ergibt sich aus

$$\left[\hat{\beta} - t_{n-2,1-\alpha/2}\,\hat{\sigma}_\beta, \hat{\beta} + t_{n-2,1-\alpha/2}\,\hat{\sigma}_\beta\right]$$

Das Bestimmtheitsmaß

Wir wollen messen, wie gut sich das Modell an die Daten anpasst und ob unsere Schätzung die Annahmen des Regressionsmodells erfüllt.

Das **Bestimmtheitsmaß** R^2 ist definiert durch als der Quotient aus der Summe der Abweichungsquadrate im geschätzten Modell und der Summe der Abweichungsquadrate der beobachteten Daten.

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2}$$

R^2 gibt den Anteil der durch die Regression erklärten Varianz an.

Das Bestimmtheitsmaß / (2)

$$\underset{\substack{\uparrow\\ \text{totale}\\ \text{Abweichung}}}{(y_t - \bar{y})} = \underset{\substack{\uparrow\\ \text{nicht-erklärte}\\ \text{Abweichung}}}{(y_t - \hat{y}_t)} + \underset{\substack{\uparrow\\ \text{erklärte}\\ \text{Abweichung}}}{(\hat{y}_t - \bar{y})}$$

Quadrieren und Summieren ergibt

$$\underset{\substack{\uparrow\\ \text{SST}}}{\sum(y_t - \bar{y})^2} = \underset{\substack{\uparrow\\ \text{SSE}}}{\sum(y_t - \hat{y}_t)^2} + \underset{\substack{\uparrow\\ \text{SSR}}}{\sum(\hat{y}_t - \bar{y})^2}$$

SST ... *total sum of squares*

SSE ... *sum of squared errors*

SSR ... *sum of squares from regression*

(Der Term $2\sum(y_t - \hat{y}_t)(\hat{y}_t - \bar{y})$ ist Null.)

F-test

Das Bestimmtheitsmaß R^2 kann ebenfalls getestet werden.
Der Test $H_0: R^2 = 0$ ist dabei äquivalent dem Test $H_0: \beta = 0$.
Die Prüfgröße lautet

$$\frac{R^2}{1 - R^2}(n-2) \sim F_{1,n-2}$$

und ist F-verteilt mit 1 und $n-2$ Freiheitsgraden.

Der Test heißt kurz ***F*-Test**.

Überprüfung der Modellannahmen

- **Unkorreliertheit der Residuen**. Aus der Autokorrelation könnte man die Residuen prognostizieren. Damit wäre unser Modell nicht vollständig.
- **Konstanz der Varianz der Residuen**. Wenn die Varianzen in verschiedenen Teilperioden unterschiedlich sind, sind die geschätzten Varianzen von $\hat{\alpha}$ und $\hat{\beta}$ nicht korrekt.
- **Abweichungen von der Normalverteilung** kann mittels Histogrammen und Test auf Normalverteilung entdeckt werden.

Prognose mittels linearer Regression

Zur Prognose müssen die Werte der unabhängigen Variable $x_{t+\tau}$, $\tau = 1, \ldots, k$ bekannt sein.

- **Prognosefunktion**

$$\hat{y}_{t+\tau} = \hat{\alpha} + \hat{\beta}\, x_{t+\tau}$$

- **Standardfehler der Prognose**

$$\text{se}(\hat{y}_{t+\tau}) = \hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{t} + \frac{(x_{t+\tau} - \bar{x})^2}{\sum_{i=1}^{t}(x_i - \bar{x})^2}}$$

  Der Standardfehler der Prognose ist minimal, wenn $x_{\tau+t} = \bar{x}$ ist.

- **Prognoseintervall**

$$[\hat{y}_{t+\tau} - t_{t-2} \cdot \text{se}(\hat{y}_{t+\tau}) \;,\; \hat{y}_{t+\tau} + t_{t-2} \cdot \text{se}(\hat{y}_{t+\tau})]$$

Multiples Regressionsmodell

Wir erweitern das Modell auf mehrere erklärende Variable

$$y_t = \alpha + \beta_1\, x_{1t} + \beta_1\, x_{2t} + \cdots + \beta_m\, x_{mt} + \varepsilon_t$$

Die Koeffizienten α, β_1, β_2, …, β_m werden wieder mittels OLS geschätzt.

Es gibt eine Reihe von Verfahren um die Anzahl m der notwendigen Variablen zu bestimmen (so wenig wie möglich!).

Das multiple Bestimmtheitsmaß

Das **multiple Bestimmtheitsmaß** ist analog zum einfachen Bestimmtheitsmaß definiert.

Liegt eine Regression mit m unabhängigen Variablen, x_{jt}, mit $j = 1, \ldots, m$ vor, so ist der Test auf $R^2 = 0$ äquivalent dem Test, dass alle $\beta_j = 0$ sind (bzw. dass kein x_j einen Beitrag zur Varianzreduktion in y zu leisten im Stande ist).

***F*-Test**:

$$\begin{aligned} &\text{H}_0\text{:} \quad \beta_1 = \beta_2 = \ldots = \beta_m = 0 \qquad (R^2 = 0) \\ &\text{H}_1\text{:} \quad \text{zumindest ein } \beta_j \neq 0 \qquad (R^2 \neq 0) \end{aligned}$$

Die Prüfgröße

$$\frac{R^2}{1 - R^2}\, \frac{n - (m+1)}{m} \;\sim\; F_{m,n-(m+1)}$$

ist F-verteilt mit m und $n - (m+1)$ Freiheitsgraden.