

§ 12 Lineare Regression

Erinnerung In der letzten Vorlesung (VL 11) ging es um Schätzer. Wir hatten ein statistisches Modell abhängig von einem Parameter Θ gegeben. Das Ziel war es, Θ mit Hilfe von Daten zu schätzen.

Heute wollen wir die Parameter im Linearen Modell schätzen.

Setting Gegeben seien Zufallsvariablen Y, X_1, \dots, X_n . Wir nehmen an, dass Parameter a_1, \dots, a_n, b existieren, so dass

$$Y = a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b + \varepsilon \quad (L)$$

$\overset{\uparrow}{\text{Zufallsvariable mit}}$

Das Modell (L) heißt das Lineare Modell.

$$\begin{aligned} E\varepsilon &= 0 \\ \text{Var } \varepsilon &\text{ klein.} \end{aligned}$$

Gegeben seien Daten z_1, \dots, z_n , wobei $z_i = (y^{(i)}, x_1^{(i)}, \dots, x_n^{(i)})$.

Ziel: Schätzung des Parameters $\Theta = (a_1, \dots, a_n, b)$ durch die Daten z_1, \dots, z_n .

Wur Daten nie ganz exakt sind, sondern Messfehler etc. haben, modellieren wir in (L) die Störungen als Zufallsvariable ε .

Die Strategie den Parameter Θ zu schätzen heißt Methode der kleinsten Quadrate.

Idee: Finde $\Theta = (\alpha_1, \dots, \alpha_n, b)$, so dass

$$\sum_{i=1}^N (y_i - (\alpha_1 x_1^{(i)} + \alpha_2 x_2^{(i)} + \dots + \alpha_n x_n^{(i)} + b))^2 \quad (\text{LS})$$

minimiert wird für gegebene Daten $z_i = (y_i, x_1^{(i)}, \dots, x_n^{(i)})$.

(d.h. wir wählen Θ so, dass die Summe der quadratischen Differenzen zwischen y_i und der Vorhersage $\alpha_1 x_1^{(i)} + \dots + \alpha_n x_n^{(i)} + b$ minimiert wird)

Wie lösen wir LS?

Erinnerung: Die Datupunkte waren z_1, \dots, z_N mit $z_i = (y_i, x_1^{(i)}, \dots, x_n^{(i)})$.

Definition

Die Designmatrix der Daten ist

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} & 1 \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_n^{(N)} & 1 \end{bmatrix} \in \mathbb{R}^{N \times (n+1)}$$

(erste Zeile gehört zu z_1 , etc.).

Wir definieren außerdem $\gamma = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$

(LS) ist dann äquivalent zu: $\min_{\Theta} \|Y - X \cdot \Theta\|^2$ (wobei
 $\|\cdot\|^2 :=$ Summe der Quadrate der Einträge, also $\|(k_1, \dots, k_N)\|^2 = k_1^2 + \dots + k_N^2$)

Die Lösung von (LS) ist dann wie folgt:

$$\Theta = (X^T X)^{-1} \cdot X^T \cdot Y$$

ist der LS-Schätzer (least-squares-Schätzer) für Θ .

Wann ist der LS-Schätzer ein guter Schätzer?

Probleme, die auftreten können:

- (1) $X^T X$ ist nicht invertierbar.
- (2) Y hängt nicht linear von X_1, \dots, X_n ab.

Zu (1): • Damit $X^T X$ invertierbar sein kann, muss $N \geq n+1$ sein.

D.h. wir brauchen mindestens so viele Daten wie Parameter.

- Außerdem sollten die X_i möglichst unkorreliert sein.

Definition

Sind X_1, X_2 Zufallsvariablen. Dann ist die Korrelation von X_1 und X_2

$$\text{cor}(X_1, X_2) := \frac{\mathbb{E} (X_1 - \mathbb{E} X_1)(X_2 - \mathbb{E} X_2)}{\sqrt{\text{Var}(X_1)} \cdot \sqrt{\text{Var}(X_2)}}$$

Die Kovarianz von X_1, X_2 ist $\text{cov}(X_1, X_2) = \mathbb{E} (X_1 - \mathbb{E} X_1)(X_2 - \mathbb{E} X_2)$.

Wir nennen x_1, x_2 korreliert, falls $\text{cor}(x_1, x_2) \approx \pm 1$.

Für Linear Regression heißt das: Falls zwei der x_i korreliert sind, ist Θ nicht mehr eindeutig, (bzw. die Matrix $X^T X$ ist nicht mehr invertierbar)

Definition

Seien $x_{1,-}, x_n$ Zufallszahlen von der Zufallsvariable X und $y_{1,-}, y_n$ Zufallszahlen von der Zufallsvariable Y . Die **empirische Korrelation** der Daten ist dann,

$$\hat{\text{cor}} := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

$\hat{\text{cor}}$ wird auch **Korrelationskoeffizient** genannt.
(In R: $\text{cor}(x, y)$).

Zu (2): Um beurteilen zu können wie gut das Lineare Modell (L) die Daten erklärt, können wir den R^2 -Koeffizienten betrachten:

Definition

$$R^2 := 1 - \frac{s_{\text{res}}}{s_{\text{total}}} \quad , \text{ wobei}$$

$$S_{\text{total}} := \sum_{i=1}^N (\gamma_i - \bar{\gamma})^2 \quad \text{und} \quad S_{\text{res}} := \sum_{i=1}^N (\gamma_i - (a_1 x_1^{(i)} + \dots + a_n x_n^{(i)}) + b)^2$$

Interpretation: $R^2 \approx$ Anteil der Daten, die gut durch das Modell (L) erklärt werden.

hängt von Fall zu Fall ab.

In manchen Anwendungen ist $R^2 = 0.4$ "gut" in anderen "schlecht".

Um zu beurteilen, ob (L) ein gutes Modell ist, können wir auch den Residualplot betrachten:

Definition

$$e_i := \gamma_i - \left(\sum_{j=1}^n a_j x_j^{(i)} + b \right) \text{ heißt } i\text{-tes Residual.}$$

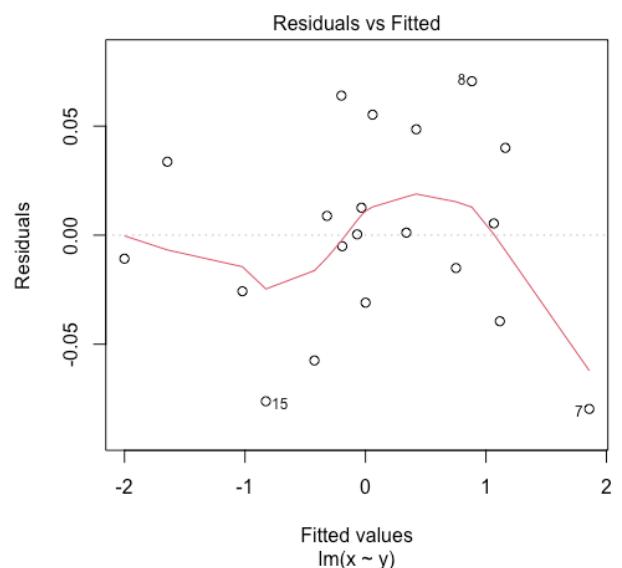
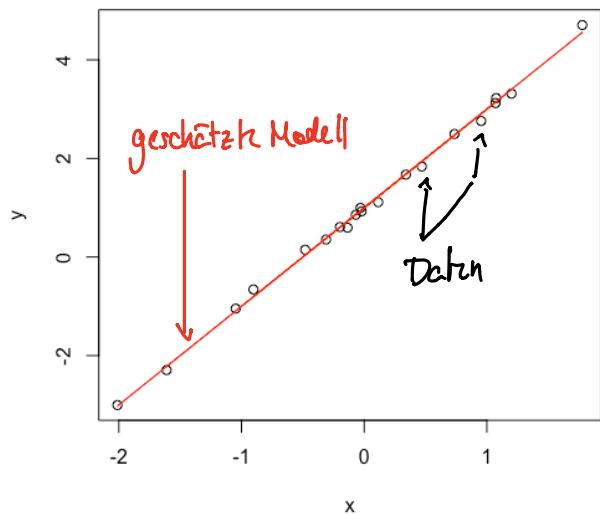
Idee: Wenn das Modell (L) die Daten $\gamma_1, \dots, \gamma_N$ gut erklärt, dann ist $e_i \sim E$. D.h., wenn wir die e_i in einem Plot betrachten, sollten sie zufällig sein.

Anders gesagt, wenn der Plot der Residuals Muster aufweist, ist das ein Zeichen, dass (L) kein gutes Modell für die Daten ist.

Beispiel $Y = 2 \cdot X + 1 + \varepsilon$.

Im linken Plot sehen wir X gegen Y geplottet.

Im rechten Plot sehen wir die Residuals.



Die Residuals scheinen zufällig um 0 verfüllt zu sein. Wir schließen, dass das lineare Modell eine gute Wahl sein könnte.