

Exkurs: Einfache lineare Regression

Grundlagen

- ▶ Die lineare Regression ist eine Methodik zur Bestimmung des Einflusses einer oder mehrerer Variablen auf eine Zielgröße.
- ▶ Wir sprechen von einfacher linearer Regression, falls es genau eine unabhängige Variable gibt.
- ▶ In der einfachen linearen Regression wird versucht die beobachteten Werte durch eine Gerade abzubilden.
- ▶ Falls die einfache lineare Regression als Prognosemodell dient, liegen die Prognosen auf dieser Geraden.
- ▶ Beispiele:
 - ▶ Größe und Gewicht
 - ▶ Lernzeit und Punkte im Abschlusstest

Einfache lineare Regression

Sei X die Einflussgröße (auch unabhängige Variable genannt) und Y die gesuchte Zielgröße (auch abhängige Variable genannt). Mithilfe der einfachen linearen Regression wird eine Gerade durch die Punktwolke $\{(x_1, y_1), \dots, (x_n, y_n)\}$ gelegt, sodass der lineare Zusammenhang zwischen X und Y möglichst genau beschrieben wird.

Die Gleichung der einfachen linearen Regression ist gegeben durch:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i \in \{1, \dots, n\}.$$

ϵ_i beschreibt den Fehler zwischen dem wahren Wert y und dem Schätzwert \hat{y} , β_0 die Regressionskonstante (auch intercept genannt) und β_1 das Gewicht der unabhängigen Variable X .

Falls alle Punkte auf einer Geraden liegen würden, könnte die einfache lineare Regression die abhängige Variable perfekt erklären und es würden keine Fehler existieren. Dies ist in den seltensten Fällen der Fall.

Wie wird die Gerade durch die Punktwolke gelegt?

Das Ziel ist die Fehler zwischen y_i und \hat{y}_i zu minimieren. Da sich positive und negative Fehler nicht aufheben sollen, können beispielsweise die absoluten oder quadrierten Fehler betrachtet werden.

Zielfunktion

Aus diesem Grund minimiert die einfache lineare Regression die quadrierten Abweichungen zwischen y_i und \hat{y}_i . Die Zielfunktion lautet:

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^n \epsilon_i^2.$$

Diese Methode wird Methode der kleinsten Quadrate sowie OLS (engl. *ordinary least squares*) genannt. Demnach werden diejenigen β_0 sowie β_1 gesucht, die folgendes Problem lösen:

$$\min Q = \min \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Um dieses Minimierungsproblem zu lösen, leiten wir nach β_0 und β_1 ab und setzen gleich Null. Wir erhalten

$$\frac{\partial Q}{\partial \beta_1} \stackrel{!}{=} 0$$

$$-2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) x_i = 0$$

$$\sum_{i=1}^n x_i y_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2$$

sowie

$$\frac{\partial Q}{\partial \beta_0} \stackrel{!}{=} 0$$

$$-2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\sum_{i=1}^n y_i = \beta_0 n + \beta_1 \sum_{i=1}^n x_i$$

.

Dies können wir wie folgt darstellen:

$$\begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Wir wissen, dass eine Matrix wie folgt invertiert wird:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Daraus ergibt sich:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

Nun wäre noch zu zeigen, dass es ein Minimum ist. Dies lässt sich über die zweite Ableitung zeigen.

Interpretation von β_0

Die Regressionskonstante gibt den Schnittpunkt der Regressionsgeraden mit der y-Achse bei $x = 0$ an. Falls $\beta_0 = 0$ ist, geht die Regressionsgerade durch den Ursprung.

Interpretation von β_1

Der Vorhersagewert des Modells erhöht sich um β_1 Einheiten, falls die unabhängige Variable sich um eine Einheit erhöht. Das Vorzeichen von β_1 gibt demnach die Richtung des Effekts an.

Multiple lineare Regression

Falls eine beobachtete abhängige Variable durch mehrere unabhängige Variablen erklärt wird, sprechen wir von multipler linearer Regression. Dies ist eine Verallgemeinerung der einfachen linearen Regression.

DataCamp

Im DataCamp Kurs *Introduction to Regression with statsmodels in Python* lernen wir die lineare Regression besser kennen und wenden diese an gewissen Datensätzen an.