Lineare Regression am Beispiel des Datensatzes tips

Thomas Müller

Zusammenfassung

Anhand des Built-In Datensatzes aus `Seaborn` behandeln wir die Grundsätze der Linearen Regression. Dabei betrachten wir die Idee von linearen Modellen. Wir erarbeiten die Grundlagen des t-Tests im Falle der Linearen Regression, um den p-Wert zu berechnen und berechnen ein Konfidenzband um die Regressionsgerade.

1 Einleitung

Als Beispiel betrachten wir einen sehr simplen und weitverbreiteten Datensatz `tips`. Dieser ist unter anderem in dem Package `Seaborn` vorinstalliert.

Der Datensatz beschäftigt sich mit der Thematik des Trinkgeld Gebens. Wie in der Tabelle 1 zu sehen, besteht jeder der 244 Datenpunkte aus der Höhe der Gesamtrechnung, der Höhe des Trinkgelds, beides in Dollar angegeben, Geschlecht, (Nicht-)Raucher*in, Wochentag, Essenszeit und Gruppengröße.

2 Linearen Regression

Die Idee hinter der Linearen Regression ist, dass zwischen zwei Variablen in einem Datensatz ein Zusammenhang existiert, den man bis auf einen Fehler durch eine lineare Gleichung darstellen kann.

Falls eine solcher Zusammenhang gegeben ist spricht man von einem linearen Modell.

Definition 2.1 (Lineares Modell)**.** Ein n-dimensionaler Zufallsvektor Y genügt einem klassischen linearen Modell, falls es eine Darstellung

$$Y = \beta_0 + X^\top \beta + \epsilon \tag{2.1}$$

	Rechnung	Trinkgeld	Geschlecht	Raucher	Tag	Zeit	Anzahl
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

Tabelle 1: Darstellung des Datensatzes

mit Erwartungswert $\mathbb{E}[\epsilon] = 0$ und Kovarianzmatrix $\Sigma_\epsilon = \sigma^2 I_n$.
Gilt zusätzlich $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$, heißt das Modell **klassisches lineares Modell**.

Dabei nennt man den Vektor $Y = [Y_1, \dots, Y_n]^\top \in \mathbb{R}^n$ Regressand oder abhängige Variable und die Matrix $X = [X_1, \dots, X_n] \in \mathbb{R}^{k \times n}$ Regressor oder unabhängige Variable.

Wir betrachten hier nur den Fall des klassischen linearen Modells.

Zur Vereinfachung können wir die Verschiebungskonstante β_0 und die Steigung β zusammenfassen. Das erlangen wir, indem wir eine 1 vor den Regressoren X_i anhängen $\tilde{X}_i := (1, X_i) \in \mathbb{R}^{k+1}$.

Dadurch erhalten wir

$$Y = \tilde{X}^\top \tilde{\beta} + \epsilon.$$

Hierbei ist Y gegeben X ein Zufallsvektor, da der Fehler nicht deterministisch ist.

Wenn wir einen Datensatz $\mathcal{D} = \{(X_1, y_1), \dots, (X_n, y_n)\}$ haben, dann betrachten wir Realisierungen von Y. Somit ist y deterministisch und wir erhalten den Zusammenhang

$$y_i = \tilde{\beta}^\top \tilde{X}_i + \epsilon_i \quad \text{bzw.} \quad y = \tilde{X}^\top \tilde{\beta} + \epsilon.$$

Das Ziel der linearen Regression ist es für eine gegebene Realisierung y von Y den Vektor $\tilde{\beta}$ schätzen.

Der gleichmäßig beste erwartungstreue Schätzer (UMVU) der linearen Regression ist durch den Kleinste-Quadrate-Schätzer gegeben.

2.1 Kleinste-Quadrate-Schätzer

Der KQ-Schätzer minimiert den quadratischen Abstand der Realisierungen y_i und der Schätzungen $\hat{y}_i = \hat{\beta}^\top \tilde{X}_i$, in Formeln

$$\hat{\beta} = \operatorname*{argmin}_{\tilde{\beta} \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} \left| y_i - \tilde{\beta}^\top \tilde{X}_i \right|^2. \tag{2.2}$$

Diesen Minimierer kann explizit angegeben werden.

Satz 2.2 (KQ-Schätzer in linearen Modellen)**.** *Sei $Y = X + \epsilon$ ein lineares Modell mit einer $(k+1) \times n$-Matrix X und $rg(X) = k + 1$, dann ist für eine beobachtete Realisierung y von Y der KQ-Schätzer $\hat{\beta}$ gegeben durch*

$$\hat{\beta} = (\tilde{X}\tilde{X}^\top)^{-1} \tilde{X} y. \tag{2.3}$$

$\hat{\beta}$ ist ein erwartungstreuer Schätzer für β mit Kovarianzmatrix $\Sigma_{\hat{\beta}} = \sigma^2 (\tilde{X}\tilde{X}^\top)^{-1}$.

Beweis. Übungsaufgabe. □

Nach dem verallgemeinerten Gauß-Markov-Theorem, siehe [1, Satz 5.5.14], ist dieser Schätzer insbesondere der gleichmäßig beste Schätzer.

Die Darstellung des KQ-Schätzers aus Satz 2.2 haben wir verwendet, um unsere Schätzfunktion zu implementieren, siehe Auszug aus dem Code 1.

Dep. Variable:	tip	**R-squared:**	0.457
Model:	OLS	**Adj. R-squared:**	0.454
Method:	Least Squares	**F-statistic:**	203.4
Date:	Wed, 30 Aug 2023	**Prob (F-statistic):**	6.69e-34
Time:	15:50:13	**Log-Likelihood:**	-350.54
No. Observations:	244	**AIC:**	705.1
Df Residuals:	242	**BIC:**	712.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.9203	0.160	5.761	0.000	0.606	1.235
x1	0.1050	0.007	14.260	0.000	0.091	0.120

Omnibus:	20.185	**Durbin-Watson:**	2.151
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	37.750
Skew:	0.443	**Prob(JB):**	6.35e-09
Kurtosis:	4.711	**Cond. No.**	53.0

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tabelle 2: Zusammenfassung der OLS Regressions Statistik.

```
34      def umvu_beta(self):
35          """
36          Calculating beta hat based on given dataset.
37          Definition is in the main notebook.
38
39          Sets:
40              umvu-estimator for the linear model
41              beta @ [1, X] = y
42          """
43          # computes umvu-estimator
44          self.beta = np.linalg.inv(self.x_tilde@self.x_tilde.T)@self.x_tilde@self.y
```

Auszug aus dem Code 1: Simple lineare Regression aus der Datei `regression.py`

Wir wollen die Abhängigkeit der Gesamtrechnung und dem gegebenem Trinkgeld berechnen. Dabei betrachten wir die Gesamtrechnung als Regressor mit einer Dimension ($k = 1$) und das Trinkgeld als Regressand. Durch `statsmodels.api.OLS` können wir uns verschiedene statistische Werte des KQ-Schätzer ausgeben lassen, siehe Tabelle 2.

Dadurch könnte hier eine ausführliche Analyse dieser statistischen Werte stehen.

Über unsere eigens implementierten Funktionen haben wir die Daten in einem Scatterplot dargestellt und die Regressionsgerade einzeichnen, siehe Abb. 1

2.2 p-Wert

...

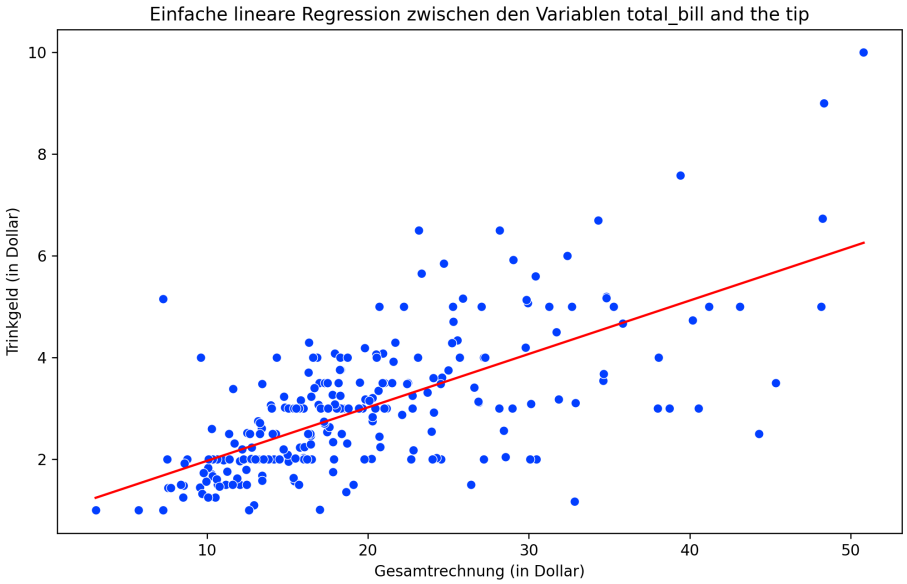

Abbildung 1: In einen Scatterplot eingezeichnete selbstimplementierte Regressionsgerade.

2.3 Konfidenzband für eine Lineare Regression

...

Literatur

[1] L. Rüschendorf. *Mathematische statistik*, volume 62. Springer, 2014.