Lineare Regression am Beispiel des Datensatzes TIPS

Thomas Müller

Albert-Ludwigs-Universität Freiburg

Inhaltsverzeichnis

Einleitung

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

	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: Darstellung des Datensatzes

Inhaltsverzeichnis

Lineare 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.

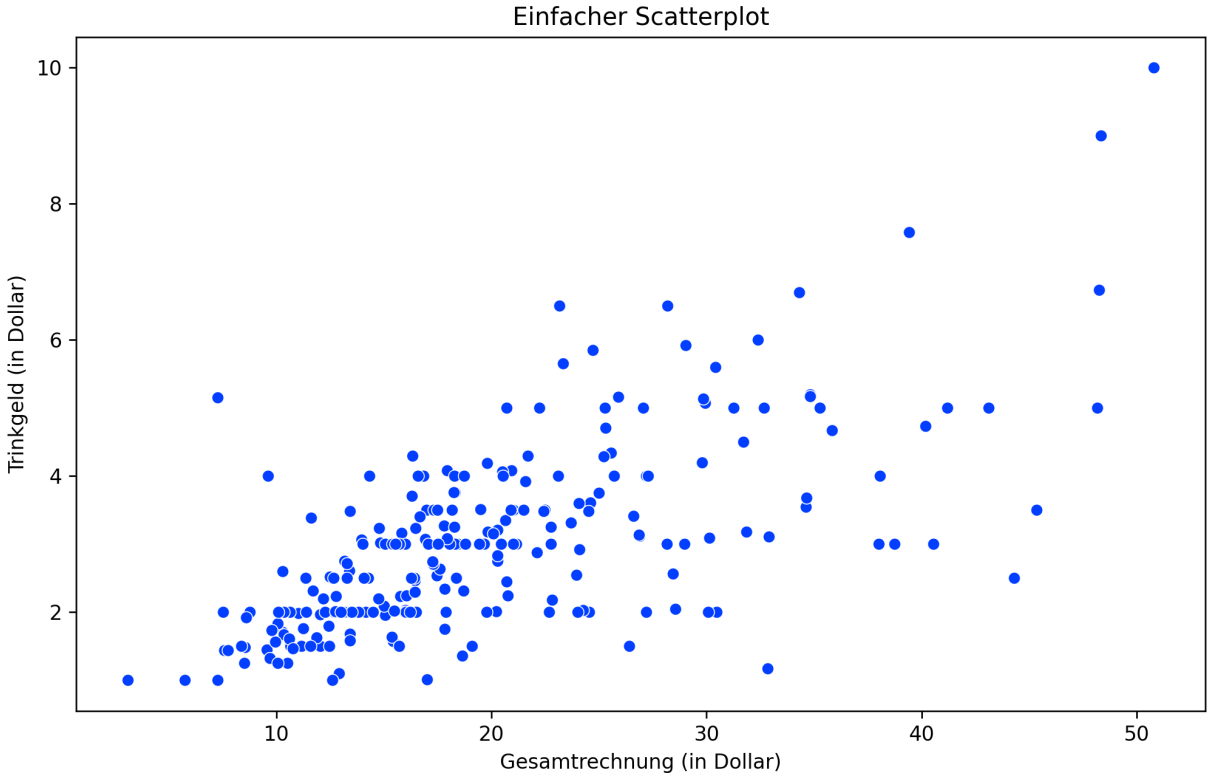

Abbildung: Simpler Scatterplot

Lineare 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.

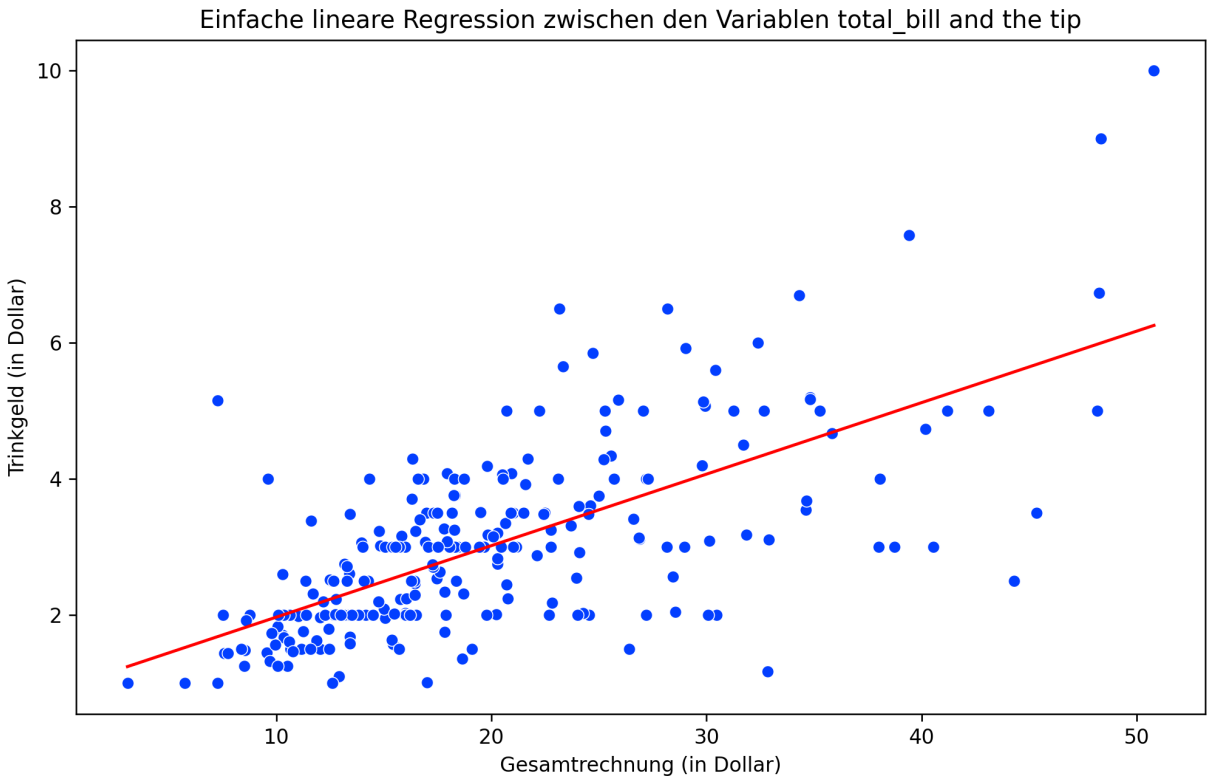

Abbildung: Simpler Scatterplot mit Regressionsgeraden

Lineare 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.

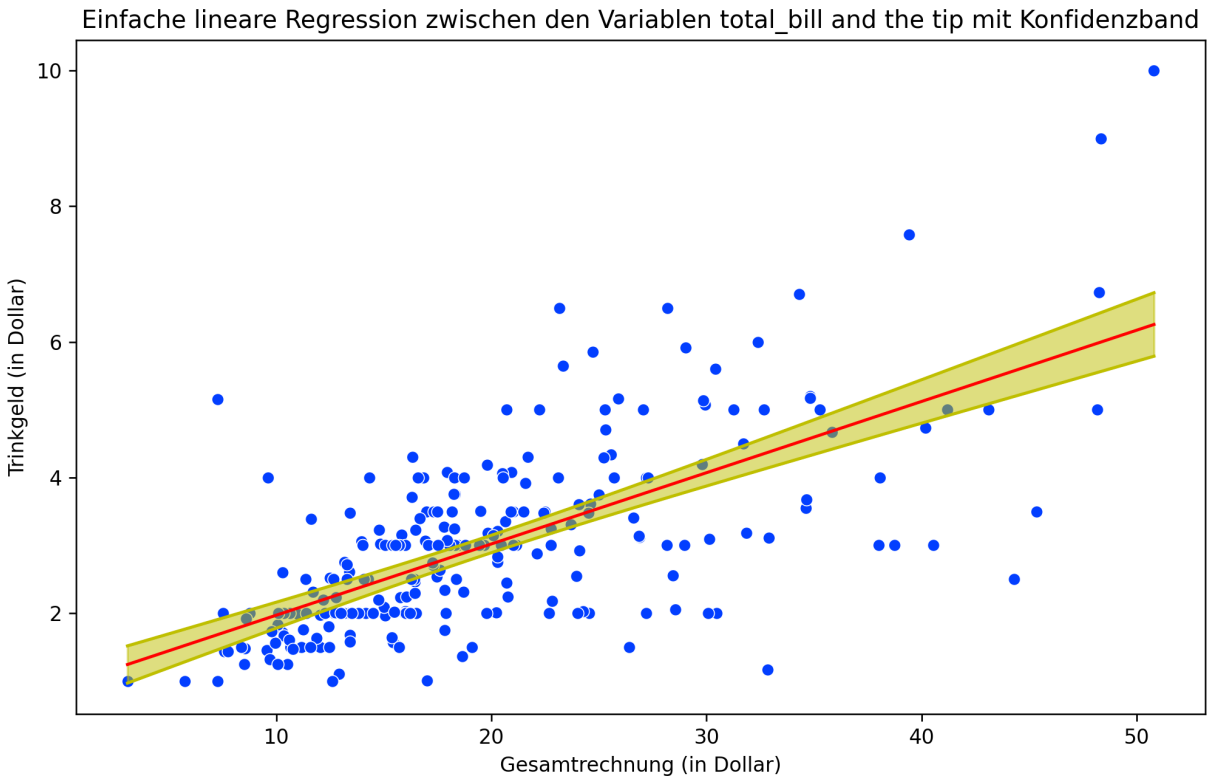

Abbildung: Simpler Scatterplot mit Regressionsgeraden und Konfidenzband

Lineares Modell

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

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

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

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

Lineares Modell

Zur Vereinfachung können wir die Verschiebungskonstante β_0 und die Steigung β zusammenfassen. Das erlangen wir, indem wir eine 1 zu dem Vektor der X_i hinzufügen $\tilde{X}_i = (1, X_i) \in \mathbb{R}^{k+1}$.
Dadurch erhalten wir

$$Y_i = \tilde{\beta}^\top \tilde{X}_i + \epsilon_i,$$

wobei $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. In Matrix Schreibweise ergibt das

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

KQ-Schätzer

Sei $\mathcal{D} = \{(X_1, y_1), \ldots, (X_n, y_n)\}$ einen Datensatz. Durch die lineare Regression wollen wir den Vektor $\tilde{\beta}$ schätzen.
Der gleichmäßig beste erwartungstreue Schätzer (UMVU) bei der linearen Regression ist durch den Kleinste-Quadrate-Schätzer gegeben, welcher folgende Gleichung minimiert.

KQ-Schätzer

Sei $\mathcal{D} = \{(X_1, y_1), \ldots, (X_n, y_n)\}$ einen Datensatz. Durch die lineare Regression wollen wir den Vektor $\tilde{\beta}$ schätzen.
Der gleichmäßig beste erwartungstreue Schätzer (UMVU) bei der linearen Regression ist durch den Kleinste-Quadrate-Schätzer gegeben, welcher folgende Gleichung minimiert.

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

KQ-Schätzer

Satz 2.2 (KQ-Schätzer in linearen Modellen)

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

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

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

Beweis von Satz 2.2

Beweis.

...

□

Implementierung des KQ-Schätzers

Die Darstellung des KQ-Schätzers aus Satz 2.2 haben wir verwendet, um den Schätzer für $\tilde{\beta}$ zu implementieren.

Implementierung des KQ-Schätzers

Die Darstellung des KQ-Schätzers aus Satz 2.2 haben wir verwendet, um den Schätzer für $\tilde{\beta}$ zu implementieren.

```
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 2: Simple lineare Regression aus der Datei `regression.py`

Statistik der linearen Regression

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: Zusammenfassung der OLS Regressions Statistik.

Darstellung der linearen Regression

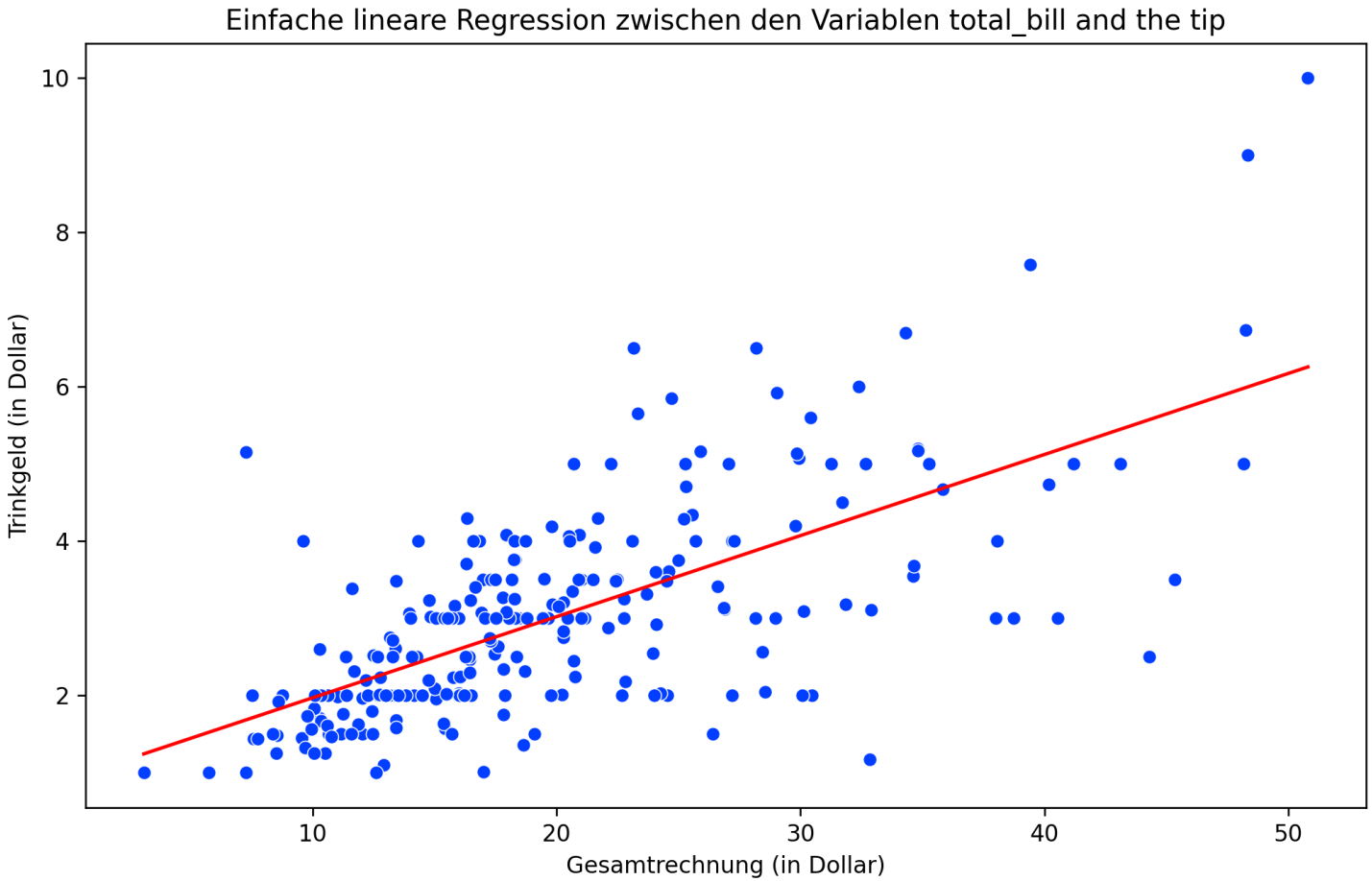

Abbildung: In einen Scatterplot eingezeichnete selbstimplementierte Regressionsgerade.

p-Wert

. . .

Konfidenzband für eine Lineare Regression

...

Inhaltsverzeichnis

Fazit

. . .

Danke für eure Aufmerksamkeit!

Quellen

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

[2] J. Brutsche und A. Rohde, „Sharp adaptive similarity testing with pathwise stability for ergodic diffusions,“ *arXiv preprint arXiv:2203.13776*, 2022.

[3] C. Czado und T. Schmidt, *Mathematische Statistik*. Springer-Verlag, 2011.

[4] G. Casella und R. L. Berger, *Statistical inference*. Cengage Learning, 2021.

[5] M. J. Schervish, *Theory of statistics*. Springer Science & Business Media, 2012.