Einführung in die Empirische Wirtschaftsforschung

Übungsaufgaben 7

KQ Methode in der linearen Einfachregression - Lineare Mehrfachregression - Multikollinearität - Ausgelassene Prediktoren

1. Schreiben Sie die Zielfunktion der KQ-Methode für die lineare Einfachregression auf. Charakterisieren Sie die Bedingungen erster Ordnung für ein Minimum.
`Optional`: Leiten Sie die Schätzer für den Achsenabschnitt und die Steigung her.

2. Wenn eine in einem Duopol operierende Firma den Preis um einen Franken senkt, und der Preis der Konkurrenz gleich bleibt, dann steigt der Umsatz um 200. Wenn die Konkurrenz den Preis ebenfalls um einen Franken senkt, dann bleibt der Umsatz unverändert. In 2 von 3 Fällen senk die Firma den Preis um einen Franken (in den Anderen Fällen lässt sie den Preis unverändert). Im Falle einer Preissenkung, kopiert die Konkurrenz diese Preissenkung in 50% der Fälle. Was sind dann b_1, b_2 und g_1 in den zwei Gleichungen

$$\widehat{Umsatz} = b_0 + b_1 Preis_{eigen} + b_2 Preis_{Konkurrenz}$$
$$\widetilde{Umsatz} = g_0 + g_1 Preis_{eigen} \qquad ?$$

`Hinweis`: R kann bei der Lösung der Aufgabe helfen.

3. Betrachten Sie das Datenset "miete". Wir betrachten die Variable "rent" als Zielvariable und die Variabeln "size" und "rooms" als Prediktoren.

a) Regressieren Sie in einer linearen Einfachregression "rent" auf "rooms". Interpretieren Sie die Koeffizienten.

b) Erweitern Sie die linearen Einfachregression um den Prediktor "size". Was fällt Ihnen im Vergleich zu Aufgabe a) im Bezug auf b_1 auf? Wie kann es sein, dass b_1 nun nicht mehr signifikant ist?

c) In der Regression aus b), um wie viel ändert sich die erwartete Miete ceteris paribus, wenn bei einer Wohnung ein zusätzliches Schlafzimmer hinzukommt?
Macht es Sinn, in diesem Beispiel eine ceteris paribus Aussage zu treffen?

d) Gibt es weitere Variablen, welche man dem Regressionsmodel hinzufügen sollte?

4. Wir betrachten eine lineare Mehrfachregression mit p "interessanten" Prediktoren (abgesehen von der Konstanten, die auch im Modell enthalten ist). Die zugehörige R^2-Statistik sei R_p^2. Jetzt nehmen wir zusätzlich noch einen weiteren Regressor hinzu, d.h., es gibt jetzt $p+1$ "interessante" Variablen. Die zugehörige R^2-Statistik sei R_{p+1}^2. Zeigen Sie mathematisch, dass notwendigerweise $R_{p+1}^2 \geq R_p^2$.

5. Nehmen Sie an, (Y, X_1, X_2) genüge den Annahmen der linearen Mehrfachregression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

wobei $E(\epsilon) = 0$ (beziehungsweise $E(\epsilon X_1) = E(\epsilon X_2) = 0$). β_1, β_2 sind die interessierenden Effekte von X_1, X_2 auf Y. Alternativ wird auch das folgende lineare Einfachregressionsmodell geschätzt

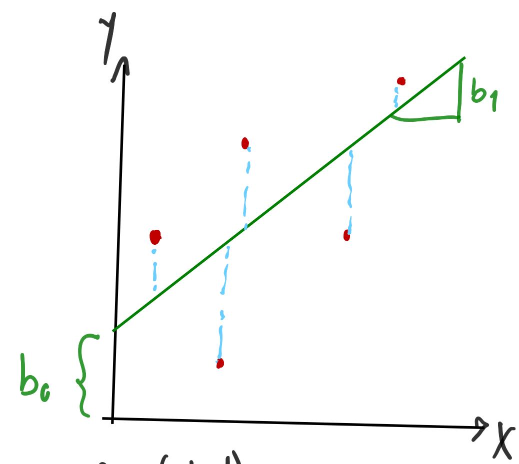
$$Y = \alpha_0 + \alpha_1 X_1 + u.$$

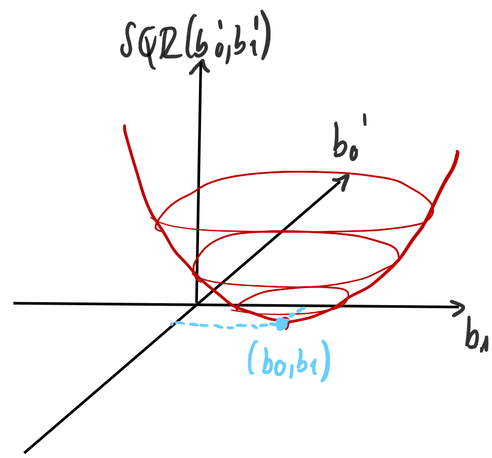
Zeigen Sie, das der Schätzer a_1 im kleinen Modell i.d.R. nicht unverzerrt (unbiased) ist für den Koeffizienten β_1 im grossen Modell: Leiten Sie eine Formel für den zugehörigen Omitted Variable Bias $E(a_1) - \beta_1$ her.

1. Schreiben Sie die Zielfunktion der KQ-Methode für die lineare Einfachregression auf. Charakterisieren Sie die Bedingungen erster Ordnung für ein Minimum.
`Optional`: Leiten Sie die Schätzer für den Achsenabschnitt und die Steigung her.

$$(b_0, b_1) = \underset{b_0', b_1'}{\arg\min}\; SQR(b_0', b_1') = \underset{b_0', b_1'}{\arg\min} \sum_{i=1}^{n} e_i^2$$

$$= \underset{b_0', b_1'}{\arg\min} \sum_{i=1}^{n} (y_i - b_0' - b_1' x_i)^2$$





- Wir minimieren diese Funktion im Bezug auf b_0, b_1:

$$\frac{\partial SQR(b_0', b_1')}{\partial b_0'} = 2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \overset{!}{=} 0$$

$$\frac{\partial SQR(b_0', b_1')}{\partial b_1'} = 2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(-x_i) \overset{!}{=} 0$$

- Gleichungssystem mit zwei Gleichungen und zwei Unbekannten (siehe Lösungen für Auflösung)
- Verbindung Mathematik: Wir können das obige schreiben als

$$\nabla SQR(b_0', b_1') = (0, 0)^T$$

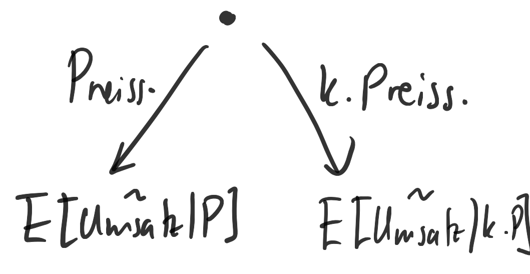
2. Wenn eine in einem Duopol operierende Firma den Preis um einen Franken senkt, und der Preis der Konkurrenz gleich bleibt, dann steigt der Umsatz um 200. Wenn die Konkurrenz den Preis ebenfalls um einen Franken senkt, dann bleibt der Umsatz unverändert. In 2 von 3 Fällen senk die Firma den Preis um einen Franken (in den Anderen Fällen lässt sie den Preis unverändert). Im Falle einer Preissenkung, kopiert die Konkurrenz diese Preissenkung in 50% der Fälle. Was sind dann b_1, b_2 und g_1 in den zwei Gleichungen

$$\widehat{Umsatz}_1 = b_0 + b_1 Preis_{eigen} + b_2 Preis_{Konkurrenz}$$
$$\widetilde{Umsatz}_2 = g_0 + g_1 Preis_{eigen} \qquad ?$$

`Hinweis:` R kann bei der Lösung der Aufgabe helfen.

$$\left.\begin{aligned} b_1 &= -200 \\ b_2 &= 200 \end{aligned}\right\} \quad \widehat{Umsatz} = b_0 - 200\, Preis_{eigen} + 200\, Preis_{konkurrenz}$$

Was ist der erwartete Umsatz i.e $\widetilde{Umsatz}$?



$$E[\widetilde{Umsatz}] = E[\widetilde{Umsatz} \mid Preissenkung]\, P(Preissenkung)$$
$$+ E[\widetilde{Umsatz} \mid keine\ Preissenkung]\, P(keine\ Preissenkung)$$
$$= \underbrace{EE[\widetilde{Umsatz} \mid Preissenkung]}_{= 0.5 \cdot 0 + 0.5 \cdot 200} \underbrace{P(Preissenkung)}_{= 2/3}$$

, wobei 0.5 := Wahrscheinlichkeit, dass Konk. preis senkt

$$+ \underbrace{EE[\widetilde{Umsatz} \mid keine\ Preissenkung]}_{=0} \underbrace{P(keine\ Preissenkung)}_{= 1/3}$$

$\longrightarrow$ Der erwartete Umsatz ($\widetilde{Umsatz}$) liegt bei $2/3 \cdot (0.5 \cdot 0 + 0.5 \cdot 200)$, wobei der erwartete Umsatz nach einer eigenen Preissenkung bei $0.5 \cdot 0 + 0.5 \cdot 200 = 100$ liegt.

$\longrightarrow$ Somit haben wir:

$$\tilde{\text{Umsatz}} = 90 - 100 \text{Preis}_{\text{eigen}}$$

3. Betrachten Sie das Datenset “miete”. Wir betrachten die Variable “rent” als Zielvariable und die Variabeln “size” und “rooms” als Prediktoren.

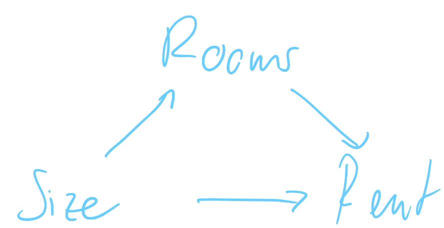
a) Regressieren Sie in einer linearen Einfachregression “rent” auf “rooms”. Interpretieren Sie die Koeffizienten.

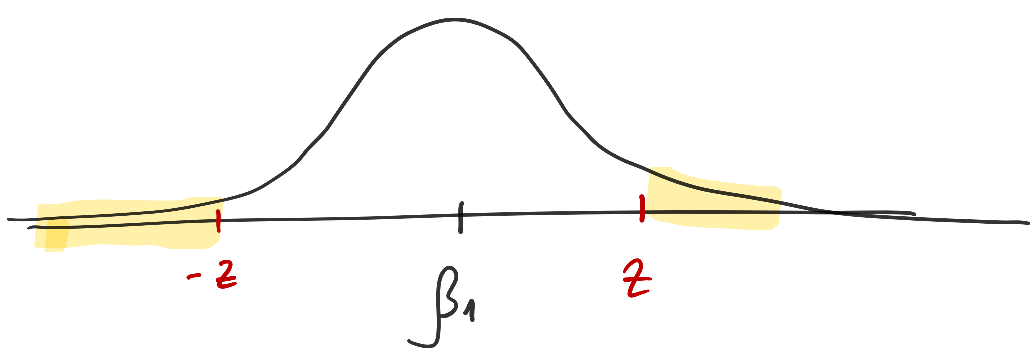
b) Erweitern Sie die linearen Einfachregression um den Prediktor “size”. Was fällt Ihnen im Vergleich zu Aufgabe a) im Bezug auf b_1 auf? Wie kann es sein, dass b_1 nun nicht mehr signifikant ist?

c) In der Regression aus b), um wie viel ändert sich die erwartete Miete ceteris paribus, wenn bei einer Wohnung ein zusätzliches Schlafzimmer hinzukommt?
Macht es Sinn, in diesem Beispiel eine ceteris paribus Aussage zu treffen?

d) Gibt es weitere Variablen, welche man dem Regressionsmodel hinzufügen sollte?

Siehe R





$$2P(Z > |z|)$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$\hookrightarrow \hat{X}_1 = a_0 + a_1 X_2 + a_2 X_3 \to R^2$ $\quad \hat{X}_3 = a_0' + a_1' X_2 + a_2' X_1 \to R'^2$

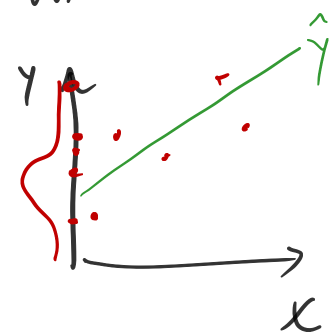
$\hat{X}_2 = \hat{a}_0 + \tilde{a}_1 X_1 + \tilde{a}_2 X_3 \to \tilde{R}^2$

$$VIF = \frac{1}{1-R^2}$$

4. Wir betrachten eine lineare Mehrfachregression mit p "interessanten" Prediktoren (abgesehen von der Konstanten, die auch im Modell enthalten ist). Die zugehörige R^2-Statistik sei R_p^2. Jetzt nehmen wir zusätzlich noch einen weiteren Regressor hinzu, d.h., es gibt jetzt $p+1$ "interessante" Variablen. Die zugehörige R^2-Statistik sei R_{p+1}^2. Zeigen Sie mathematisch, dass notwendigerweise $R_{p+1}^2 \geq R_p^2$.

Definition $R^2 := 1 - \frac{SQR}{TQS} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$

Wir sehen, dass $TQS = \sum_{i=1}^{N} (y_i - \bar{y})^2$ von p unabhängig ist. Wir müssen also zeigen, dass:



$$\sum_{i=1}^{N} (y_i - \hat{y}_{i,p})^2 \leq \sum_{i=1}^{N} (y_i - \hat{y}_{i,p+1})^2$$

Die zusätzliche Flexibilität durch die $(p+1)$-te Variable kann die Anpassungsgüte der Regression nur verbessern. Falls die $(p+1)$-te Variable rein gar nichts von y erklären kann, dann gilt $b_{p+1} = 0$ und somit $\hat{y}_{i,p} = \hat{y}_{i,p+1}$ und $R_p^2 = R_{p+1}^2$.

↳ Das deckt aber natürlich ein Problem von R^2 auf, da es mit zusätzlichen Regressoren nie schlechter wird. Deshalb gibt es grössen wie das adjusted R^2_{adj}:

$$R^2_{adj} := 1 - (1 - R^2)\frac{N-1}{N-p}$$

5. Nehmen Sie an, (Y, X_1, X_2) genüge den Annahmen der linearen Mehrfachregression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

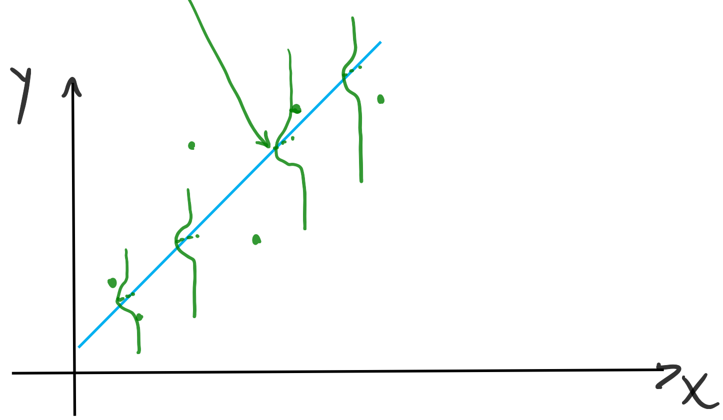
wobei $E(\epsilon) = 0$ (beziehungsweise $E(\epsilon X_1) = E(\epsilon X_2) = 0$). β_1, β_2 sind die interessierenden Effekte von X_1, X_2 auf Y. Alternativ wird auch das folgende lineare Einfachregressionsmodell geschätzt

$$Y = \alpha_0 + \alpha_1 X_1 + u.$$

Zeigen Sie, das der Schätzer a_1 im kleinen Modell i.d.R. nicht unverzerrt (unbiased) ist für den Koeffizienten β_1 im grossen Modell: Leiten Sie eine Formel für den zugehörigen Omitted Variable Bias $E(a_1) - \beta_1$ her.

$E(\varepsilon) = 0$

$E(\varepsilon X_1) = E(\varepsilon X_2) = 0$, da $Cov(\varepsilon, X_i) = E(\varepsilon X_i) - \underbrace{E(\varepsilon)}_{=0} E(X_i)$, $i = 1,2$



Wie berechnen wir a_1 (kleines Modell):

$$a_1 = \frac{\widehat{Cov}(X_1, Y)}{\widehat{Var}(X_1)} = \frac{\cancel{\frac{1}{N-1}} \sum_{i=1}^{N} (X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\cancel{\frac{1}{N-1}} (X_{i1} - \bar{X}_1)^2} \overset{(*)}{=} \frac{\sum_{i=1}^{N} (X_{i1} - \bar{X}_1) Y_i}{\underbrace{\sum_{i=1}^{N} (X_{i1} - \bar{X}_1)^2}_{:= S_X^2}}$$

$$= \frac{\sum_{i=1}^{N} (X_{i1} - \bar{X}_1)(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i)}{S_X^2}$$

$$= \frac{\beta_0 \sum_{i=1}^{N} (x_{i1} - \bar{x}_1) + \beta_1 \sum_{i=1}^{N} \cancel{(x_{i1} - \bar{x}_1)} x_{i1} + \beta_2 \sum_{i=1}^{N} (x_{i1} - \bar{x}_1) x_{i2} + \sum_{i=1}^{N} (x_{i1} - \bar{x}_1) \varepsilon_i}{S_x^2 = \sum_{i=1}^{N} \cancel{(x_{i1} - \bar{x}_1)} x_{i1}}$$

$$= \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1)^2 x_{i2}}{\sum (x_{i1} - \bar{x}_1) x_{i1}} + \frac{\sum (x_{i1} - \bar{x}_1) \varepsilon_i}{\sum (x_{i1} - \bar{x}_1) x_{i1}}$$

↳ Wie schätzen wir a_1 in Erwartung, also $E(a_1)$:

$$E(a_1) = \beta_1 + \beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)} + \underbrace{\frac{\sum (x_{i1} - \bar{x}_1) E(\varepsilon_i)}{\sum (x_{i1} - \bar{x}_1) x_{i1}}}_{=0} \neq \beta_1$$

(*) Alternative form um folgendes zu schreiben: $\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$:

$$= \sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i - \bar{x} \sum_{i=1}^{N} y_i + N \bar{x} \bar{y}$$

$$= \sum_{i=1}^{N} x_i y_i - \bar{x} \sum_{i=1}^{N} y_i - N \bar{x} \bar{y} + N \bar{x} \bar{y}$$

$$= \sum_{i=1}^{N} (x_i - \bar{x}) y_i$$

↳ Using the same trick, we can write s_x^2 as $\sum_{i=1}^{N} (x_{i1} - \bar{x}_1) x_{i1}$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Einfacher:

$$a_1 = \frac{\widehat{Cov}(X_1, Y)}{\widehat{Var}(X_1)} = \frac{\widehat{Cov}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon)}{\widehat{Var}(X_1)}$$

$$= \frac{\widehat{Cov}(X_1, \beta_0) + \beta_1 \widehat{Cov}(X_1, X_1) + \beta_2 \widehat{Cov}(X_1, X_2) + \widehat{Cov}(X_1, \varepsilon)}{\widehat{Var}(X_1)}$$

$$= 0 + \beta_1 \frac{\widehat{Var}(X_1)}{\widehat{Var}(X_1)} + \beta_2 \frac{\widehat{Cov}(X_1, X_2)}{\widehat{Var}(X_1)} + 0$$

$$= \beta_1 + \beta_2 \frac{\widehat{Cov}(X_1, X_2)}{\widehat{Var}(X_1)}$$

$\beta_2 :=$ Populationsparameter von Size