Methoden Verkehrsökonometrie für Master-Studierende

Winter semester 2021/22, Solutions to Tutorial No. 4

Lösungsvorschlag zu Aufgabe 4.1: Über das Fitten von Elefanten

(a) Die einfache lineare Regression ist (der Name sagt's schon) linear, während die polynomische Regression quasilinear bzw. parameterlinear ist.

(b) Die einfache lineare Regression ist wohlspezifiziert: Die Daten lassen nicht erkennen, dass der Zufallsanteil heteroskedastisch oder nicht gaußförmmig ist oder Abhängigkeiten aufweist. Außerdem gibt es mehr als $J + 1 = 2$ Datenpunkte und die Daten sind nicht kolinear.

Die polynomischeRegression ist *nicht* wohlspezifiziert, da $n = J + 1 = 6$, es also genauso viele Datenpunkte wie Parameter gibt.

(c) Das polynomiale Modell hat die Fehlerquadratsumme null. Dies ist auch klar: Es gibt sechs Unbekannte β_j und sechs Systemgleichungen $y_i = \sum_{j=0}^{5} \beta_j x_i^j + \epsilon_i$, so dass das (lineare!) Gleichungssystem exakt lösbar für $\epsilon_i = 0$ ist. Dennoch beschreibt eindeutig das lineare Modell den Sachverhalt besser. In der Tat kann sind aufgrund der Fehlspezifikation dieses Modells induktive Fehlermaße der β-Schätzer undefiniert. Dieses Modell fittet maximal bei Null Aussagekraft. Daher auch der berühmte Spruch "mit 5 Parametern kann man einen Elefanten fitten und mit sechs das Wackeln seines Schwanzes"[1]

Lösungsvorschlag zu Aufgabe 4.2: Regression mit nominalskalierten Variablen

(a)
- Skalierung: x_1 und y kardinal bzw metrisch, x_2 nominal bzw. qualitativ.
- keine überflüssigen Var, mögliche fehlende Faktoren z.B. Haltestellendichte
- Linearität OK
- Möglicher Strukturbruch zwischen 1930 und 1960, da es vorher nur eine Geschwindigkeitsklasse gab

(b) Die Zugklasse und das Jahr sind offensichtlich keine Zufallsvariablen. Ferner kann man die Zugklasse nicht als lineare Funktion des Jahres schreiben, also gibt es keine Multikolinearität bei den exogenen Daten. Schließlich ist der Stichprobenumfang $n = 8$ zwar klein, aber nicht nur gleich sondern sogar größer als der formale Mindestumfang $J + 2 = 4$.

[1] Für Ökonometriker können sechs oder auch mehr Parameter bei geeigneter Datenlage durchaus OK sein.

(c) β_0: erwartete Geschwindigkeit der langsameren Zugklasse im Jahr 1900, β_1: Anstieg der erwarteten Geschwindigkeit jeder der beiden Zugklassen pro Jahr (anhand der Daten positiv erwartet), β_2: Erwartete Geschwindigkeitsdifferenz zwischen schnellen und langsameren Zügen (aufgrund der Definition ("schneller ist schneller als langsamer") und der Daten wird ein positiver Wert erwartet.

(d) Jahr 2020, also $x_1 = 120$: Langsame Züge:

$$\hat{y}(120, 0) = \hat{\beta}_0 + 120\hat{\beta}_1 = 99.1\,\text{km/h}$$

Schnelle Züge:

$$\hat{y}(120, 1) = \hat{\beta}_0 + 120\hat{\beta}_1 + \hat{\beta}_2 = 127.3\,\text{km/h}$$

(e) Allgemein berechnet sich Erwartungswert und Varianz des Prognoseschätzers

$$\hat{y}(\vec{x}) = \sum_j x_j \hat{\beta}_j$$

mit festen (da vorgegebenen) Werten $\vec{x}$ und Zufallsvariablen $\hat{\beta}_j$ durch die allgemeinen Regeln für die Varianz einer Linearkombination von Zufallsvariablen. Seien X und Y Zufallsvariable und a und b Konstante, so gilt

$$\begin{aligned} E(aX + bY) &= aE(X) + bE(y), \\ V(aX + bY) &= a^2V(X) + b^2V(y) + 2ab\text{Cov}(X, Y), \end{aligned}$$

welches rekursiv auch auf Summen mit mehr als zwei Summanden angewandt werden kann. Hier spielen $x_0 = 1$ sowie x_1 und x_2 die Rolle der Konstanten und $\hat{\vec{\beta}}$ die Rolle der Zufallsvariablen. Also

$$\begin{aligned} E(\hat{y}) &= \sum_j x_j E(\hat{\beta}_j) = \sum_j x_j \beta_j, \\ V(\hat{y}) &= = \sum_j x_j^2 V(\hat{\beta}_j) + 2\sum_j \sum_{k>j} x_j x_k \text{Cov}(\hat{\beta}_j, \hat{\beta}_k) \end{aligned}$$

Hier ergibt sich mit $x_0 = 1$ in Abhängigkeit von x_1 und x_2:

$$\begin{aligned} E(\hat{y}) &= \hat{\beta}_0 + x_1\hat{\beta}_1 + x_2\hat{\beta}_2, \\ V(\hat{y}) &= \hat{V}_{00} + x_1^2\hat{V}_{11} + x_2^2\hat{V}_{22} + 2\left(x_1\hat{V}_{01} + x_2\hat{V}_{02} + x_1x_2\hat{V}_{12}\right) \end{aligned}$$

und korrekt für das Jahr 2020 ($x_1 = 120$) für die langsamen ($x_2 = 0$) und schnellen ($x_2 = 1$) Züge:

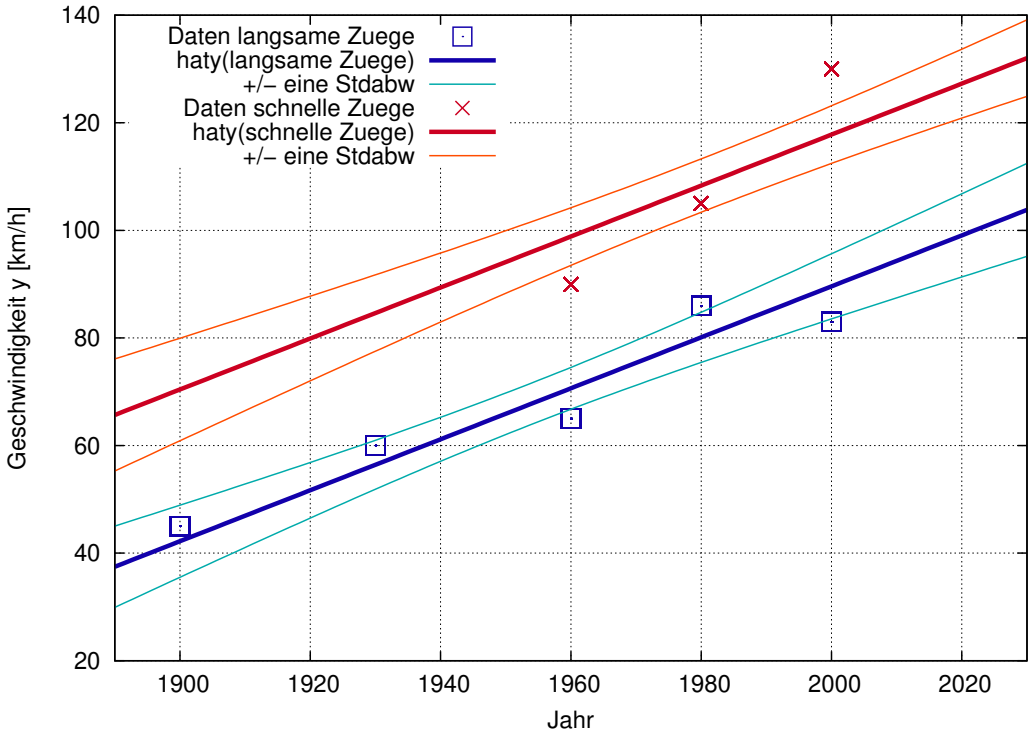
Langsame Züge:

$$\begin{aligned} E(\hat{y}) &= \hat{\beta}_0 + 120\hat{\beta}_1 = 99.1\,\text{km/h}, \\ V(\hat{y})) &= \hat{V}_{00} + 120^2\hat{V}_{11} + 2 * 120\hat{V}_{01} = 59.8\,\text{km}^2/\text{h}^2 \end{aligned}$$

Schnelle Züge:

$$\begin{aligned} E(\hat{y})) &= \hat{\beta}_0 + 120\hat{\beta}_1 + \hat{\beta}_2 = 127.3\,\mathrm{km/h}, \\ V(\hat{y})) &= \hat{V}_{00} + 120^2\hat{V}_{11} + \hat{V}_{22} + 2\left(120\hat{V}_{01} + \hat{V}_{02} + 120\hat{V}_{12}\right) = 41.1\,\mathrm{km^2/h^2} \end{aligned}$$

Die Grafik zeigt, dass die Prognoseintervalle Hyperbeln darstellen, wobei die "Taille" in der Nähe des jeweiligen Schwerpunkts der relevanten Punktewolke liegt:



(f) Gegeben: Varianzschätzer von $\hat{\beta}_1$:

$$\hat{V}(\hat{\beta}_1) = 0.0104$$

Damit ergibt sich das 5%-Konfidenzintervall zu

$$\beta_1 \in [\hat{\beta}_1 - \Delta\hat{\beta}_1, \hat{\beta}_1 + \Delta\hat{\beta}_1]$$

mit

$$\Delta\hat{\beta}_1 = \sqrt{\hat{V}(\hat{\beta}_1)}t^{(8-3)}_{1-0.05/2} = \sqrt{\hat{V}(\hat{\beta}_1)}t^{(5)}_{0.975} = 0.102 * 2.57 = 0.262$$

also

$$\mathrm{KI}(\beta_1)^{\alpha=5\%} = [0.21, 0.74].$$

(g) Spezifizierung, falls die Geschwindigkeitsänderungen über die Zeit bei schnellen und langsamen Zügen unterschiedlich sind:

$$\begin{aligned} y(\vec{x}) &= \beta_0 + \beta_1 x_1 \begin{cases} 1 & x_2 = 0 \\ 0 & \text{sonst.} \end{cases} + \beta_2 x_1 \begin{cases} 0 & x_2 = 0 \\ 1 & \text{sonst.} \end{cases} + \beta_3 x_2 + \epsilon \\ &= \beta_0 + \beta_1 x_1 (1 - x_2) + \beta_2 x_1 x_2 + \beta_3 x_2 + \epsilon \end{aligned}$$

mit

- β_0 unverändert (erwartete Geschwindigkeit der langsameren Zugklasse im Jahr 1900),
- β_1: Anstiegsrate der erwarteten Geschwindigkeit der langsameren Zugklasse pro Jahr,
- β_2: Anstiegsrate der erwarteten Geschwindigkeit der schnelleren Zugklasse pro Jahr,
- β_3: Differenz der erwarteten Geschwindigkeiten der beiden Zugklasse im Jahr $x_1 = 0$, also 1900.

Man könnte das Modell völlig äquivalent sogar einfacher formulieren:

$$y(\vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2 + \beta_3 x_2 + \epsilon$$

Die Parameter β_0, β_1 und β_3 behalten ihre Bedeutung während $\beta_2 = \beta_2^{\text{alt}} - \beta_1^{\text{alt}}$ nun die *Differenz* der Geschwindigkeitsanstiegsraten der schnelleren bezüglich der langsameren Zugklasse bedeutet.

(h) Nun hat die nomimalskalierte Variable x_2: "Zugkategorie" drei Asusprägungen: $x_2 = 0$: langsame Züge, $x_2 = 1$: schnell, aber kein ICE, $x_2 = 2$: ICE. Nach der allgemeinen Regel benötigt man bei k Ausprägungen $k - 1$ Dummy-Variablen bzw. Selektoren, also hier zwei:

$$y(\vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 \delta_{x_2,1} + \beta_3 \delta_{x_2,2}$$

mit dem "δ-Selektor"

$$\delta_{i,j} = \begin{cases} i & i = j \\ 0 & \text{sonst.} \end{cases}$$

Man beachte, dass hier die langsamen Züge ($x_2 = 0$) die Referenz darstellen:

- β_2 gibt den mittleren Geschwindigkeitsunterschied der schnellen Nicht-ICE-Züge bezüglich der langsamen an,
- β_3 gibt den mittleren Geschwindigkeitsunterschied der ICE-Züge bezüglich der langsamen an.