Methods in Transportation Econometrics and Statistics (Master)

Winter semester 2023/24, Tutorial No. 4

Problem 4.1: On fitting elephants

Given is the following data (black bullets in the figures)

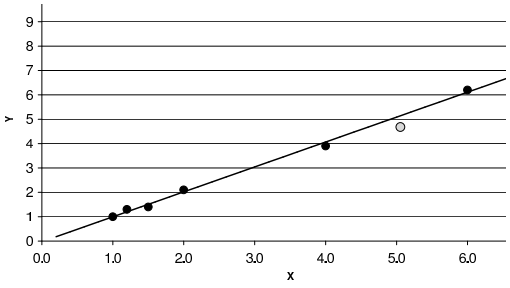
$\tilde{x}_i$	1.0	1.2	1.5	2.0	4.0	6.0
Y_i	1.0	1.3	1.4	2.1	3.9	6.2

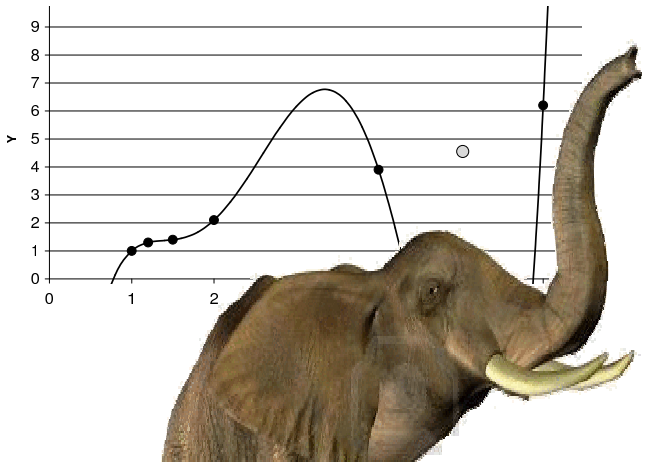
To analyse it, there are two model candidates: (i) M1: simple linear regression $Y = a + b\tilde{x} + \epsilon$, (ii) M2: polynomial regression $Y = \sum_{j=0}^{5} b_j \tilde{x}^j + \epsilon$.

(a) Discuss whether these models are linear, parameter linear, or nonlinear.

(b) Check if they are well specified (as far as can be deduced from the data). If this is not the case, which Gauß-Markow condition(s) are violated?

(c) The OLS estimators $\hat{y}_i$ of the two models are depicted in the following figure:





Which model has the lower SSE? Which model is better suited to describe the real relationship between x and y? Discuss, in particular, how an additional data point would change the parameterisation of either model.

Problem 4.2: Regression with qualitative exogenous variables

In order to analyze if, and to which extent, the trains have accelerated in the last decades, following data for the average speed of local (L) and intercity (IC) trains are available to the econometrist:

Year $\tilde{x}_1$ (seit 1900)	0	30	60	60	80	80	100	100
Train category $\tilde{x}_2$	L	L	L	IC	L	IC	L	IC
Avg speed [km/h] y	45	60	65	90	86	105	83	130

The change of the average speed for each category shall be described with the linear model

$$y(\vec{x}) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

with the factors $x_0 = 1$, $x_1 = \tilde{x}_1$ and $x_2 = g(\tilde{x}_2)$ to be discussed.

(a) Discuss the functional specification: (i) What is the scaling of the exogenous and endogenous variables? (ii) Does the model contain superfluous or obviously missing exogenous variables? (iii) Do you expect a structural change that is not contained in the model?

(b) Check the data specification for correctness, i.e. (i) sufficient number of data points, (ii) no multi-colinearity, (iii) the exogenous variables are deterministic and measurable.

(c) Discuss why, for the train category, a dummy variable x_2 is necessary taking on the value $x_2 = 0$ for $\tilde{x}_2 =$ "Local" and $x_2 = 1$ for $\tilde{x}_2 =$ "IC". Give the intuitive meaning of the three parameters and also discuss which sign you would expect for their estimated values.

(d) For this and the following subtasks, assume following OLS estimator and estimated error variance-covariance matrix:

$$\hat{\vec{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 42.2 \\ 0.474 \\ 28.2 \end{pmatrix}, \quad \underline{\underline{\hat{V}}} = \begin{pmatrix} 45 & -0.56 & -0.21 \\ -0.56 & 0.0104 & -0.27 \\ -0.21 & -0.27 & 46 \end{pmatrix}.$$

Give a forecast for the expected speeds of the slow and fast train category in the year 2020.

(e) Calculate the standard deviation of the forecast $\hat{y}$ made in the previous subtask.

Hint: Make use of general formulas for the variance of a linear combination of correlated random variables with a correlation structure given by $\underline{\underline{\hat{V}}}$.

(f) Calculate the confidence interval of β_1 at a significance level (alpha-error) $\alpha = 5\,\%$

Hint: Use the variance-covariance matrix given above and make sure to realize that their indices start at zero.

(g) The model discussed above does not reflect different rates of speed changes for the slower and faster train categories. Generalize the model (i.e., its functional specification) to include that.

(h) Consider now a common speed changing rate for all trains again but distinguish three train categories $\tilde{x}_2$ by splitting up the fast category into "intercities" (ICs) and "intercity expresses" (ICEs). Change the functional specification to include that.