Methods in Transportation Econometrics and Statistics (Master)

Winter semester 2023/24, Solutions to Tutorial No. 4

Solution to Problem 4.1: On the fitting of elephants

(a) As the name implies (simple linear regression), M1 is linear[1] The polynomial regression is nonlinear in $\tilde{x}$ but has six linear factors $x_0 = 1$, $x_1 = \tilde{x}$,..., $x_5 = \tilde{x}^5$. Thus, it is parameter linear.

(b) The simple linear model is well specified: There are more data points than parameters ($n = 6 > J + 1 = 2$) and the graph implies that no systematic influencing factors are ignored.

The polynomial model is *not* well specified since there are as many data points as there are parameters to be estimated.

(c) The polynomial model has a SSE $S = 0$ which can be implied from $n = J + 1 = 6$: There are six data points and six parameters, i.e., a linear system of 6 unknowns which (since there is no multi-collinearity and the system matrix is regular) can be exactly solved such that the fitting curve goes exactly through the data points. The SSE of the simple linear model is $S > 0$. Nevertheless, the linear model has a good predictive power (as indicated by the new open-bullet point which is rather well estimated) while the polynomial model has nil explanatory nor predictive power: There is the famous saying by John von Neumann that *with four parameters I can fit an elephant and with five I can make him wiggle its trunk* (and with six additional its tail)[2]

Solution to Problem 4.2: Qualitative exogenous variables

(a)
- (i) Scaling: Endogenous variable y (speed) and $\tilde{x}_1$ (year): metric. In contrast, the train category is qualitative (or ordinal which makes no difference for binary variables), so we need a pseudo-metric dummy variable x_2 as given in (c).
- (ii) obviously no superfluous factors; possible missing factors include the frequency of stations (a high station density lowers the average speed).
- (iii) Linearity: regarding year: OK as implied by the data; regarding the train category or, more specifically, its dummy: OK for principle reasons since two values 0 and 1 always can be connected by a straight line.

[1] Strictly speaking, M1 is only linear if setting $x_0 = 1$ and $x_1 = \tilde{x}$, so we have $y(\vec{x}) = \beta_0 x_0 + \beta_1 x_1 + \epsilon$ (linearity means if doubling all input variables, the output doubles as well). With only x, it is affine-linear.

[2] For econometrists, six or even more parameters are still OK provided there is enough data.

  - Structural change: Possibly between 1930 and 1960 since only one train category is listed 1930 and earlier, and two categories later on.

  Generally, the scaling of the variables is crucial for the model specification:

  - For a linear model, the endogenous variable must be *metric* (at least *interval scaled*, i.e., the operations + and − must make sense) and *continuous* since y is given as a sum of real-valued terms (real-valued because the parameters are real-valued): No problem here. Otherwise, one needs to use a different model class such as *logistic regression* or a *discrete-choice model*.
  - Both for the regression models and discrete-choice models to be discussed later, the exogenous factors, as they enter the model, must be at least interval scaled (no need to be continuous):
    * If this is already the case (as here $\tilde{x}_1$), they can be used directly, $x_1 = \tilde{x}_1$.[3]
    * If they are *ordinally scaled* (only the operations $=$, $>$, $\geq$, $<$, $\leq$ make sense but neither of the arithmetic operations +,−,*,/) or *nominally scaled* (only the operation "=" makes sense) and they can assume J values, one needs $J-1$ binary dummies ("*selectors*") taking on 1 for a certain value and 0, otherwise (the value without a dummy is the reference). This applies here for $\tilde{x}_2$.
    * If there is a justified assumption that the steps between neighboring values of an ordinal variable are equal (in a plausible sense), the *ordinal number* of the value (=0 for the lowest value and $= J-1$ for the highest) can be used as a direct exogenous factor.

(b)
  - (i) $n = 8$ data points and $J+1 = 3$ parameters, so there remain df=8-3=5>0 degrees of freedom: OK,
  - (ii) the train category is not perfectly correlated with the year (since, particularly, there are two train categories for most of the years: OK,
  - (iii) both the time and the train category are obviously deterministic, measurable properties.

(c) A dummy is needed because all factors in linear regression must be metric variables (see comments at (a)).
  - β_0: expected speed of the slower train class ($x_2 = 0$) in 1900 ($x_1 = 0$); $\beta_0 > 0$ expected,
  - β_1: expected rate of change of the speed over years [km/h per year]; $\beta_1 > 0$ expected,
  - β_2: Expected speed difference between fast and slow trains for each given year; $\beta_2 > 0$ expected

(d) Speed forecast in 2020,i.e., $x_1 = 120$:

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

[3] However, one can also use nonlinear transformations as, e.g., in the "fitting an elephant" problem.

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

(e) Generally, we have following rule for calculating expectation values and variances for a linear combination of correlated random variables X and Y (an extension to > 2 variables by recursion is straightforward):

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

Here, the exogenous factors $x_0 = 1$, x_1, and x_2 play the role of the constants, and the components of $\hat{\vec{\beta}}$ the role of the random variables, so we have

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

or more explicitely as a function of x_1 and x_2,

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

For the two predictions, this results in

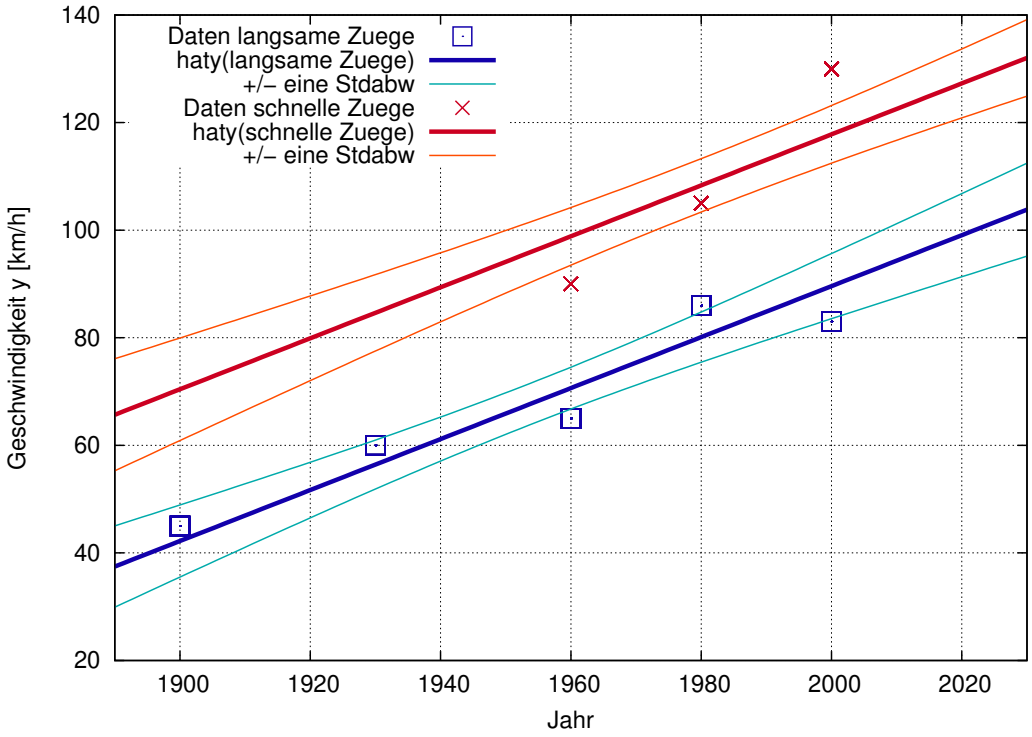
Slower trains in 2020:

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

Faster trains in 2020:

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

The following graphics shows that the forecast errors take on the form of hyperbolas which the "waist" (region of smallest extension) near the center of the data cloud used for calibration:



(f) With the error variance estimator of $\hat{\beta}_1$ given by

$$\hat{V}(\hat{\beta}_1) = \hat{V}_{11} = 0.0104,$$

we obtain the $\alpha = 5\,\%$ confidence interval by

$$\text{CI:} \quad \beta_1 \in [\hat{\beta}_1 - \Delta\hat{\beta}_1, \hat{\beta}_1 + \Delta\hat{\beta}_1]$$

with

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

resulting in

$$\text{CI}(\beta_1)^{\alpha=5\%} = \; [0.21, 0.74].$$

(g) If the rate of changes of the speed over time is different for the two categories, we need an own slope parameter for each:

$$y(\vec{x}) = \sum_{j=0}^{3} \beta_j x_j, \quad x_0 = 1, \quad x_1 = \tilde{x}_1 \delta_{\tilde{x}_2,\text{“Local”}}, \quad x_2 = \tilde{x}_1 \delta_{\tilde{x}_2,\text{“IC”}}, \quad x_3 = \tilde{x}_2,$$

where

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

is a selector dummy variable. The meanings of the parameters are now are given as follows:

- β_0 unchanged (expected speed of the slower trains in 1900),
- β_1: Speed changing rate for the slower trains,
- β_2: Speed changing rate for the faster trains,
- β_3: Speed difference between the faster ans slower trains *in 1900* (because of the different rates, the difference changes over time, now).

(h) Now, the qualitative (or ordinal) variable $\tilde{x}_2$ ("train category") is three-valued with the values "local train", "IC train" and "ICE train". As a rule, for J-valued nominal or ordinal variables, one needs $J-1$ binary dummy variables, i.e., two dummies (and two parameters) here (see comment below the solution to (a)).[4] There are several specification variants, e.g.,

$$y(\vec{x}) = \sum_j \beta_j x_j, \quad x_0 = 1, \quad x_1 = \tilde{x}_1, \quad x_2 = \delta_{\tilde{x}_2,\text{"IC"}}, \quad x_3 = \delta_{\tilde{x}_2,\text{"ICE"}}.$$

Notice that the slowest train category is the reference, here (The reference has no dummy and no parameter). With this specification, the meaning of the parameters are as follows:[5]

- β_2 gives the expected speed difference between the ICs and the reference local trains,
- β_3 gives the expected speed difference between the ICEs and the local trains

[4] The assumption that the difference between IC and local trains is the same as between ICEs and ICs is not satisfied in a plausible sense, so simplifying the model by using the ordinal number of the train category does not apply here.

[5] Notice that the meaning changes with the specification, it would, for example, be diffeent if one selected the IC or the ICE as reference category.