Elements of Statistics, Econometrics and Time Series Analysis (2017)

Assignment 2

Problem 3: Linear regression analysis

In the light of the hot discussion on the educational system in Ukraine we consider here a data set on student performance at two schools in Portugal. The data set can be downloaded at
`http://archive.ics.uci.edu/ml/datasets/Student+Performance`

The zip-file contains some personal information and the grades in Mathematics (`student-mat.csv`) and Portuguese language (`student-por.csv`). Pick up the file for Math. The variable of interest is the final grade coded as `G3`. The remaining variables are used as explanatory variables. Hereafter we exclude G1 and G2 from the discussion and the analysis!

1. Have a closer look at the definitions of the variables and analyze which of them might require a separate treatment. Consider for example the variables `Mjob` or `goout`. There are two possibilities how the variables can be included into the model (one with dummy variables, the other one without dummies). Think about these two approaches and suggest which approach is more appropriate for each of the variables `MjOb` or `goout`. Motivate your decision.

2. Consider now the dependent variable and the interval (metric) scaled explanatory variables. Plot these data and decide if you wish to transform these x-variables and if there is a need to transform the y variable. You can also use some measure of skewness to decide about y.

3. After making up your decision about the above two problems run a simple linear regression. If you wish to argue that farther's job is insignificant and use the model with dummies than you have to check the simultaneous insignificance of all dummies which stem from the factor variable `Fjob`. Run a test for general linear hypothesis and conclude about the significance of `Fjob`.

4. Provide an economic interpretation for the parameters of `age`, `Fjob`, and `goout`. Neglect the possible insignificance and keep in mind possible transformations of the variables.

5. Compute the 95% confidence intervals for the parameters of `absences` and `famsup` and provide its economic meaning. The CIs are computed relying on the assumption, that the residuals follow normal distribution. Is this assumption fulfilled? Run an appropriate goodness-of-fit test.

6. Many of the variable appear insignificant and we should find the smallest model, which still has a good explanatory power. Choose this model using stepwise model selection (either based on the tests for R^2 or using AIC/BIC). Pick up the last step of the model selection procedure and explain in details how the method/approach works (or is implemented in your software). Work with this model in all the remaining steps.

7. Sometimes data contains outliers which induces bias in the parameter estimates. Check for outliers using Cook's distance and leverage. Have a closer look at the observation with the highest leverage (regardless if it is classified as an outlier or not). What makes this observation so outstanding (you may have a look at Box-plots for interval scaled variables or at the frequencies for binary/ordinal variables?

8. Frequently data is missing. Pick up 5 rows in the data set and delete the value for `age`. Implement at least two approaches to fill in these values. Write down the corresponding formulas/model and give motivation for your approach. If you use standard routines then check how exactly the data imputation is implemented. How would you proceed if the value of the binary variable `higher` is missing? Implementation is not required.

9. Now we look at the model assumptions. The variable `goout` seems to be very significant. However, if we look at the residuals we observe that the variance of the residuals is rather different for different values of `goout`. Run the Bartlett's test and compute the FGLS estimators assuming groupwise heteroscedasticity. Compare the results with the original model. Explain the advantages of the (F)GLS estimation.

10. Compute the White estimator of covariance matrix of the OLS estimators. Run the t-tests and compare the results with the original model. Explain the advantages of the White estimator for the variance.

11. Write a short summary with the pedagogical and political interpretation of the estimated model.

Problem 4: further issues + some theory

Monte-Carlo simulation: asymptotic properties of the OLS estimators

We postulate the true regression model in the form

$$y_t = 1 + 2 \cdot x_{1t} + 3 \cdot x_{2t} + u_t.$$

.

Consider two setups of the simulation study.

(A) Draw x_{1t} from $\mathcal{N}(0, 0.4)$ and x_{2t} from $\mathcal{N}(0, 0.8)$ in such way that $Corr(x_{1t}, x_{2t}) = \rho$ (for example $\rho = 0.2$). $u_t \sim \mathcal{N}(0, 1)$ and $Corr(u_t, u_s) = 0$ of $t \neq s$

(B) Draw x_{1t} from $\mathcal{N}(0, 0.4)$ and x_{2t} from $\mathcal{N}(0, 0.8)$ in such way that $Corr(x_{1t}, x_{2t}) = 0)$. $u_t \sim \mathcal{N}(0, 1)$ and $Corr(u_t, u_{t-1}) = \rho$

1. Simulate the data using the described algorithms A and B for a fixed sample size of $T = 100$. Estimate the linear regression model. Compute the covariance matrix of the estimated parameters. Compare the parameters with the true values. Manually run a test (GLH with the true values taken as target values) to verify if the differences are significant.

2. Consider the expression for the classical covariance matrix of the OLS estimators. Which problems might arise in A and B? Give a detailed discussion here.

3. Repeat (a) for a larger value of ρ. What happens to the variances and covariances of $\hat{\boldsymbol{\beta}}$? Test if the parameters significantly deviate from zero.

4. To illustrate the unbiasedness simulate and estimate $M = 100$ samples. Compute the average estimators over the M samples. What do you expect to happen if we increase M?

5. To get some feeling for the consistency of the OLS estimators increase the sample size and estimate several sets of parameters. What do you observe? How would you change the model to obtain an inconsistent estimator? Give motivation.

6. Imagine now that you use setup A with $\rho = 0.5$, but estimate the model without X_2 what do you expect from the estimation and why?

Adding a new observation

(Greene, 2008, p. 40, Ex. 6) A data set consists of n observations in $\boldsymbol{X}_n$ and $\boldsymbol{y}_n$. The OLS estimator based on these observations is $\hat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n'\boldsymbol{y}_n$. Another observation $\boldsymbol{x}_s$ and y_s are added to the sample. Prove that the OLS estimator computed using this additional observation is given by

$$\hat{\boldsymbol{\beta}}_{n,s} = \hat{\boldsymbol{\beta}}_n + \frac{1}{1 + \boldsymbol{x}_s'(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_s}(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_s(y_s - \boldsymbol{x}_s'\hat{\boldsymbol{\beta}}_n).$$

Conclude that the new data change the results of OLS only if the new observation on y cannot be perfectly predicted using the information already in hand.

Shifts of the variables, demeaned regression

(Davidson and MacKinnon, 2004, p. 121, Ex. 3.22) Consider a linear regression model for a dependent variable y_t that has a sample mean of 17.21. Suppose that we create a new variable $y_t^* = y_t + 10$ and run the same linear regression using y_t^* instead of y_t as a regressand.

1. How are R^2 and the estimate of the constant term related in the two regressions? What if we use $y_t^* = y_t - 10$ instead?

2. What if we do the same with one or all of the regressors?

3. Consider a demeaned regression, i.e. center the regressors and the regressand to have zero mean. How does it influence the estimates?