Elements of Statistics and Econometrics

Assignment 3

Problem 4: Linear regression analysis

A telephone service provider aims to decrease the churn rate and analyses the data and service usage of 1000 clients. The following variables are used in the study

tenure	-	month a client
age	-	age in years
marital status	-	marital status (1 - married, 0 - single)
address	-	years at the current address
income	-	household income in TEuro
ed	-	education (5 categories: Did not complete high school; High school degree; Some college; College degree; Post-undergraduate degree)
retire	-	retired (0 - no, 1 - yes)
gender	-	gender (0 - male, 1 - female)
longmon	-	long distance calls last month
wiremon	-	internet use last month
churn	-	1 if the contract was terminated last month and 0 else

The overall objective is to analyze the service usage using the long distance calls last month as the dependent variables and the remaining variables as explanatory.

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 variable `ed`. There are two possibilities how the variable `ed` 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. 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. The variable `wiremon` shows a very specific pattern. How would you take it into account?

3. After making up your decision about the above two problems run a simple linear regression. Pick up one of the regressors. Write down the corresponding hypothesis of the t-test. Provide the formula for the test statistics, explain the components of the formula and give the values for this components. Evaluate the goodness of the model. Explain in your own words the difference between R^2 and adjusted R^2.

4. Compute manually the predicted values from the above regression and the residuals. Make two plots: residuals vs. true y's and predicted y's vs. true y's. What do you expect in both cases and why? Do the obtained figures support your expectations?

5. If you wish to argue that education 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 `ed`. Run a test for general linear hypothesis and conclude about the significance of `ed`. Write down the matrix and the vector needed in the hypothesis.

6. Provide an economic interpretation for the parameters of `address`, `ed`, and `retire`. Neglect the possible insignificance and keep in mind possible transformations of the variables.

7. Compute the 95% confidence intervals for the parameters of `address` and `income` and provide its economic meaning. Relate the CIs to the tests of significance, i.e. how would you use these intervals to decide about the significance of the corresponding explanatory variables? 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.

8. 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.

9. 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?

10. Frequently data is missing. Pick up 5 rows in the data set and delete the value for `address`. 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 `retire` is missing? Implementation is not required.

11. We consider now the model you have worked with so far and the model with original y if you applied some transformation OR the model with $log(y)$ if you have not transformed y. Run an appropriate test to decide which of the models is superior. Explain, the idea of the test and why you cannot make a similar decision using AIC/BIC, etc.

12. We compare the predictive ability of the estimated regression. Consider the model you worked so far and the original model with the same y but without transformation and selection of features. Compare the two models using leave-one-out CV and 5-fold CV. Explain the idea of this technique with formulas and draw a conclusion about the predictive ability of the models.

Problem 5: further issues

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?