Elements of Statistics and Econometrics - 2021

Assignment 3

Problem 5: linear regression

The overall objective is to explain the sale prices by a set of variables using the data set from the previous two assignments.

1. Before fitting a model to the data, first we have a closer look at the definitions of the variables and decide if some of them might require a separate treatment. Consider, for example, the first variable `MSSubClass`. It takes textual values and obviously cannot be included into the linear model as it is. There are two approaches how this can be handled. One approach is to assign numbers from 1 to 16 to each possible outcome. Another approach is to create new dummy variables that are equal one for a specific outcome and zero else. Discuss these two approaches and suggest which approach is more appropriate here. Motivate your decision and keep in mind the nominal, ordinal, and cardinal scales from Statistics. Run a simple regression of `SalePrices` on `MSSubClass` taken as it is (without creating dummies or 1,...,16 variable). Check how does your software handle this type of a variable. (In R your data must be stored as a data frame.)

2. Consider the following numeric explanatory variables `LotFrontage, LotArea, YearBuilt, BsmtFinSF1, BsmtUnfSF`. Plot the scatter plot of `SalePrices` vs. these variables. Analyse every plot and discuss the potential problems/transformations/extentions of the model to take the patterns into account. Remember, that you can either transform variables or create new variables to capture non-standard patterns.

3. Fit a simple linear regression without making the transformations discussed above. Pick up one of the numeric regressors, for example, `LotArea`. 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 these components. Assess the goodness-of-fit of the model. Explain in your own words the difference between R^2 and adjusted R^2.

4. If you wish to argue that `MSSubClass` is insignificant and use the model with dummies than you have to check the simultaneous insignificance of all the dummies that stem from the factor variable `MSSubClass`. Run a test for a general linear hypothesis and conclude about the significance of `MSSubClass`. Write down the matrix and the vector needed in the hypothesis.

5. Give economic interpretation for the parameters of `LotArea`, first dummy for `MSSubClass` (it is a tricky question!), and `BedroomAbvGr`. (Ignore the potential insignificance.) How would the interpretations change if you model log of `SalePrices`?

6. Compute the 95% confidence interval for the parameter of `LotArea` 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.

7. Most of the variables appear to be insignificant and we should find the smallest model that still has a good explanatory power. Choose this model using stepwise model selection (either based on the tests for R^2 or using R^2_{adj}/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.

8. Sometimes data contains outliers which induces bias in the parameter estimates. Check for outliers using plots of the residuals, 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)?

9. Frequently data is missing. Pick up 5 rows in the data set and delete the value for `LotArea`. 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. Note that for imputation you shall use the complete data set and not only the set of variables after feature selection.

10. We consider now the model with `SalePrices` and the model with log(`SalePrices`). 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.

11. We wish to assess the predictive ability of the estimated regression. Consider the full model and the model after features selection. 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 two models.

Problem 6: further issues

(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?

Problem 7: nonlinear regressions

1. A nonlinear regression offers a flexible technique for modelling complex relationships. We wish to explain the `SalePrices` by `YearBuilt`. Consider only the first 500 observations for this purpose.

   (a) Make a bivariate scatter plot and estimate an appropriate linear (!) model. Add the regression line to the plot.

   (b) Estimate now an appropriate nonlinear regression which might fit the data better. Add the regression curve to the plot and compare (quantitatively) the fit with the fit of the linear model.

   (c) Explain in your own words, why all the classical tests and inferences are not directly applicable to the NLS estimators.

   (d) What kind of problems might arise if we decide to fit a non-linear regression using all explanatory variables?