Introduction to Computational Statistics INSH 5301

Final exam

04/13/2020

For this exam you need to use the datasets **Housing** and **HealthInsurance**. Housing can be obtained from the package Ecdat using the command data(Housing), and, HealthInsurance from the package **AER** using the command data(HealthInsurance). Before starting your work, please verify that you can load the data correctly and use the ? option in R to check the variable definitions for each dataset. Some of the variables in the dataset are stored as factors, you want to transform them to numeric dummies before proceeding. (See Module 13's Memo 3: dummies)
Some rules regarding the submission of your solution:
• All regression outputs must be displayed using the stargazer package and properly compiled to LaTeX (i.e. using the type = text or type = html options will be penalized).
• Display regression tables with more than 5 regression outputs in separate tables.
• Must submit both PDFand Rmd file. Verify the .pdf before submitting your solution, submitting a wrong or corrupted file implies an automatic grade of zero, the same applies to solutions without .Rmd file.

1.Bivariate Regression

1.Using the Housing dataset, create a scatter plot of sale price of a house (y-axis) and the lot size of the property (x-axis). Use the ggplot function and include a regression line. Using the graph, describe the relation between the two variables.
2. Estimate a bivariate regression of the sale price of a house on the lot size of the property. Interpret the estimated β parameters, the statistical significance and R^2.
3. Is there any reason to believe that the estimated slope parameter in the previous regression is biased? (Explain)

2. Multivariate Regression

4. Using the rest of the variables in the dataset, construct a correlation matrix and use it to check if the assumption of exogeneity is valid in the estimated model in question (2). (Explain)
Hint: See module 13's Memo 3 for dummies and then see Module 13's Memo 1 to get familiar with OV B, explain accordingly.
5. Estimate a set of multivariate models to address the potential issue of OV B, adding at most one additional variable each time.(See Memo 2: Multivariate Models under Real World Example) Display all the estimated models side-by-side (you may need two or more stargazer tables here). Using the multivariate models, do you think there is evidence that the estimated parameter in (2) was biased? which of the estimated models you consider the least bias (from now own, we'll call this model the best model)?
Hint: See Module 13's Memo 1 to get familiar with OV B, and follow MEMO2's Multivariate Models to add one variable each time.
6. Check if the best model suffers from multicollinearity (if it does, don't try to fix it, just explain β what problems it may cause).
Hint:Use vif() in car package to easily calculate VIF and lecture 13's Memo 2 of multicollinearity to explain.

3. Non-linear Functional Forms

7. Take a look at the graph from part (1), do you think there is any reason to believe that the effect of lot size on price is not the same for all the domain of lot size? if yes, is the effect increasing or decreasing?

8. Take a look at the graph from part (1), do you think there is any reason to believe that the effect of lot size on price is not the same for all the domain of lot size? if yes, is the effect increasing or decreasing?
9. Estimate the best model twice: (a) first, adding a quadratic term for lot size, and, (b) second, adding a quadratic and cubic terms. Using the change in lot size as a one standard deviation change from the mean, compare the effect of lot size in the original model, model (a), and, model (b). Can you reject the hypothesis that the relation between lot size and price is linear? quadratic? cubic? (Explain)
10. Using the best model as the nested model, test the hypothesis that the effect of lot size on price is moderated by prefarea.

4. Unsupervised Machine Learning

11. Run a factor analysis or PCA on the Housing dataset, examine the loadings of the factors on the variables. Sort the variables by their loadings, and try to interpret what the first one "mean".
12. Use k-means algorithm and examine the centers of each cluster using only two centroids. How are they similar to and different from the factor loadings of the first factor?

5. Supervised Machine Learning

13. Divide the Housing data into two equally sized samples (one for training and one for testing). The dependent variable is price. Using the training sample, estimate a ridge model using the Housing dataset and find the optimal value of λ.
14. How does the model performs in the testing sample? Compare the results of the ridge model with a linear regression. Which model performs best?
15. Using the HealthInsurance dataset. Divide the data into two equally sized samples (one for training and one for testing). The dependent variable is health. Using the training sample; and a radial kernel and the following two values for cost C = (1e - 05, 1e + 01), estimate a support vector machine model and choose the optimal cost parameter using the function tune.(In this part, feel free to reduce the size of each sample to improve the speed of the calculations.)
16. How does the svm model performs in the testing sample? How does the model compares to a logit in terms of accuracy?