GR5205
Introduction to linear regression Dec. 3, 2016

Take-home Final Exam

Please read this page carefully before you start. We expect you to follow all the instructions. Violating any of these instructions may cost you several marks!

- Turn in your solutions in person at 5:00(PM) on Sunday Dec. 4. We will be waiting in the lobby area of the school of social work (first floor). To be fair to the other students, for every **1 minute** of delay we will deduct **1 mark**. We will **not** accept any exams after 5:20 (PM). Therefore, I suggest you account for random events (e.g., subway delays, traffic, etc.) and turn in your solutions in time. Please do **not** drop them off at SSW 904 and do **not** slip them under the doors of our offices. Otherwise, we will deduct **25 marks**.

- Make sure to turn in **all the codes and graphs** you write for the problems, even if it is not mentioned specifically in the question.

- To avoid confusion of the graders, make sure that you write or print your solutions on blank papers. Try to write your solutions as clean as possible. Depending on the neatness of your response sheet, you may get up to 5 extra marks. For those of you who got permission to submit via email, please make sure that your exam is readable after we print it.

- If you have any problem regarding the questions, email the instructor and the TAs. We **ONLY** answer emails that are sent to all of us. If you email a question after 11:59PM on Saturday, there is a good chance that you will not get any response until 8:00 (AM) Sunday. Otherwise, we hope to answer your questions within 30 minutes of the time we receive them.

- You can cite any result from the class lecture notes without proof. However, you have to prove every result that you use and we have not proved in the lecture notes.

- No discussions are allowed (or related topics) during the exam hours (except discussion with me or TAs). You are not even allowed to chat with or send emails to classmates or people who are familiar with linear regression. Any discussion will be a violation of honor code as defined below. You can of course use your computer, the software packages, text books, lecture notes, and the Internet.

"I affirm that I will not plagiarize, use unauthorized materials, or give or receive illegitimate help on assignments, papers, or examinations. I will also uphold equity and honesty in the evaluation of my work and the work of others. I do so to sustain a community built around this Code of Honor." *https://www.college.columbia.edu/ccschonorcode*

Good luck

In this exam, I try to guide you through a real data processing problem. Most of the problems are independent of each other. Hence, if you are stuck in a certain problem then move to other problems. Instead of worrying about your grade, please try to think more deeply why we do each step. In case you think you can come up with better models or you have better methods to visualize and understand the data, use them. Before you even start answering the questions, I suggest you do some visual inspection to get some feeling for the dataset. I suggest you at least do the following:

1. Use scatter plots to study the relationship between the price and bedrooms, bathrooms, sqft_living, etc.

2. Use scatter plots to study the relationship between some of the predictors.

3. In some cases, box-plots help you understand the relationship between variables better. For instance, you may use box-plot to study the relationship between the price and number of bathrooms or price and zipcode.

Visual inspection usually helps you build good simple models that can be used as a baseline for your more advanced investigations. Also, I suggest that you read these two sections of the lecture note before you start the exam.

1. Section 1.2 of Lecture note 6.

2. I have given you an extra dataset as a "test set". Please study Section 4.3 of Lecture note 8 for the definition of test and training sets. You should always use the training set **only** to estimate the regression coefficients, and then use the test to evaluate your estimates. Hence, when you need to do cross validation, please do not do leave-one-out or B-fold cross validation. Instead, use the test set that is given to you.

You are given three files: Training.csv, Test.csv, Readme. Training includes the price of 7089 houses sold in certain zipcodes. In addition to the price you can find several features of these houses. Study the Readme file to understand these features. Similarly Test.csv has the price of 6428 houses. In this problem you are supposed to estimate the regression coefficients on the Training.csv and test them on Test.csv. Since all the houses are sold in 2014 and 2015 we ignore the inflation in our studies. Please submit all the codes and graphs you use to answer the following questions.

(a) (3 points) Suppose that based on your visual inspection only you would like to pick 5 predictors to include in your linear model. Which 5 of the following will you pick: bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, yr_built, lat, long? I suggest you use box-plots on the categorical variables. Justify your choice carefully and give all the graphs based on which you reached your conclusion.

(b) (4 points) Suppose that I want to use the predictor sqft_living for predicting house prices. Hence, if the price of house i is denoted with P_i, then I want to use the model

$$P_i = \beta_0 + \beta_1 \text{sqft\_living}_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Suppose that the only information I have given you is the scatter plot of sqft_living and P. Based on this scatter plot, I claim that my model is not correct. Can you explain why?

(c) (8 points) Suppose that we calculate the average prices of the apartments that have 0 bathroom, 0.5 bathroom, 1 bathroom, etc. Plot these average prices in terms of the number of bathrooms and fit a linear model to this graph.

(d) We study the following two models:

Model 1: $P_i = \beta_0 + \beta_1 \text{sqft\_living}_i + \epsilon_i$

Model 2: $\log(P_i) = \tilde{\beta}_0 + \tilde{\beta}_1 \log(\text{sqft\_living}_i) + \epsilon_i$

i. (4 points) Compare the scatter plots of $(\text{sqft\_living}_i, P_i)$ and $(\log(\text{sqft\_living}_i), \log(P_i))$. What important differences do you notice? Suppose that our goal is to apply least squares to estimate the regression coefficients. Which of the above models will you pick?

ii. (4 points) Fit linear model to estimate all the coefficients in the above two models. Calculate R^2 for both models.

iii. (5 points) Propose a fair strategy to compare the prediction performance of the above two models, and use your strategy to compare Model 1 and Model 2.

(e) (9 points) In the rest of this question we work with $\log(P_i)$. Consider the following predictors: bedrooms, bathrooms, log(sqft_living), log(sqft_lot), floors, waterfront, view, condition, grade, yr_built, lat, long. Use the best subset selection technique to find the best predictor (subset of size 1).

(f) We improve our Model 2 by including more predictors. Our new model is (we call it Model 3 in the rest of the exam)

$$\log(P_i) \approx \beta_0 + \beta_1 \log(\text{sqft\_living}_i) + \beta_2 \text{bedrooms}_i + \beta_3 \text{bathrooms}_i + \beta_4 \text{grade}_i + \beta_5 \text{waterfront}_i.$$

Note that for waterfront predictor we would like to use a dummy variable.

i. (3 points) Explain what you would do if you wanted to use stratification for the predictor waterfront. Compare the bias and variance of a model that uses stratification with Model 3. We do not expect you to do anything quantitative to answer this question.

ii. (4 points) Use scatter plots of $\log(P_i)$ and the predictors we used in Model 3 to see if the way we have used bedrooms, bathrooms, and grade makes sense? Do you think it is possible to include these variables in a different form and improve the performance? Do not do any quantitative analysis to answer this question. Answer it based on your visual inspection only.

iii. (3 points) Estimate the coefficients β_i and report the R^2. Compare the R^2 of Model 2 and Model 3.

iv. (4 points) Intuitively argue which model has higher bias: Model 2 or Model 3? Which model has higher variance, Model 2 or Model 3?

v. (7 points) Propose a method to quantitatively compare the performance of these two models in predicting apartment prices. Which Model makes better predictions? Support your answer with a quantitative analysis.

vi. (6 points) Calculate 90% confidence interval for β_1, β_2, β_3, β_4, and β_5.

vii. (5 points) Use Bonferroni's approach to calculate a joint 95 percent confidence interval for β_1 and β_2.

(g) (4 points) Our next goal is to improve Model 3. Toward this goal we would like to add more predictors to our model. We would like to follow the strategy that I explained in the class: we first calculate the residuals

$$r_i = \log(P_i) - \hat{\beta}_0 - \hat{\beta}_1 \log(\text{sqft\_living}_i) - \hat{\beta}_2 \text{bedrooms}_i - \hat{\beta}_3 \text{bathrooms}_i - \hat{\beta}_4 \text{grade}_i - \hat{\beta}_5 \text{waterfront}_i,$$

where $\hat{\beta}_i$ are the estimates of OLS for Model 3. Then we plot residuals in terms of different predictors. If we see that the plot shows a non-random pattern for a specific predictor, then we conclude that, this predictor may help our prediction. Can you explain why this approach is a good idea?

(h) Exhibit scatter plots of r_i in terms of different predictors. Convince yourself that lat, yr_built, zipcode can still help our predictions.

i. (4 points) Based on visual inspections only, how would you include the predictors lat and yr_built in your model. Write the equations you would use for your model.

ii. (4 points) As the first step we would like to include predictors lat and yr_built in our model. Hence, we construct Model 4 in the following way:

$$\begin{aligned} \log(P_i) \approx\ & \hat{\beta}_0 + \hat{\beta}_1 \log(\text{sqft\_living}_i) + \hat{\beta}_2 \text{bedrooms}_i + \hat{\beta}_3 \text{bathrooms}_i \\ & + \hat{\beta}_4 \text{grade}_i + \hat{\beta}_5 \text{waterfront}_i + \hat{\beta}_6 \text{yr\_built} + \hat{\beta}_7 lat. \end{aligned} \quad (1)$$

Estimate the coefficients $\hat{\beta}_i$ and report the R^2. Has the R^2 improved from Model 3?

iii. (8 points) Which model predicts the prices better: Model 3 or Model 4?

(i) (10 points) Another predictor that can improve the performance of our model is the zipcode. Use boxplot to visualize the residuals we calculated in part (f) in terms of the zipcode. How would you like to include the zip-code in your model. Construct a model in which zipcode is also considered. Calculate R^2 of your model. Compare the prediction error of your Model with that of Model 4. We call this model Model 5.

(j) If you improve the prediction error of Model 5 by 15 percent, you will get 5 extra marks. If you improve it by 25 percent, you will get 10 extra marks. Note that your result must be 25 percent better than our Model 5. Also, note that your models should not use Test.csv. In case we feel that you have indirectly used the test set, we will ask for your code, and we will run it on a new test data and compare your method with our Model 5 on the new dataset.