CSE 544, Fall 2018, Probability and Statistics for Data Science

Assignment 5: Regression Due: 11/28

(6 questions, 70 points total)

I/We understand and agree to the following:

(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.

(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

(write down the name of all collaborating students on the line below)

---

1. Regression Analysis **(Total 10 points)**

Assume Simple Linear Regression on *n* sample points (Y_1, X_1), (Y_2, X_2), …, (Y_n, X_n); that is, $Y = \beta_0 + \beta_1 X + \varepsilon_i$, where $E[\varepsilon_i] = 0$.

(a) Derive the estimates of β when minimizing the sum of squared errors and show that:
$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$ and $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$, where $\bar{X} = (\sum_{i=1}^{n} X_i)/n$ and $\bar{Y} = (\sum_{i=1}^{n} Y_i)/n$. (6 points)

(b) Show that the above estimators are unbiased (Hint: Treat X's as constants) (4 points)

2. Finding the right predictor (Total 10 points)

For this problem, refer to the A5_q2.dat dataset on the website. The first column (X) refers to server CPU usage percentage and the second column (Y) refers to the observed response time for a web application hosted on that server. Feel free to use any programming tools and attach the final plots. Using simple linear regression ($Y = \beta_0 + \beta_1 X$), plot the original data and the regression fit, and also calculate the SSE and MAPE, when using the following predictors:

(a) X (3 points)

(b) 1/(1-(X/100)) (3 points)

(c) A simple variation of (b) which is an almost perfect predictor for Y, which you must find. (4 points)

3. OLS as MLE **(Total 5 points)**

Assume Multiple Linear Regression on *n* sample points, $Y_1, \ldots, Y_n$, with regressor vectors $X_1, \ldots, X_n$. Assume, as in class, that the errors, ε_i, have conditional mean zero and are homoskedastic. Further assume that the errors are Normally distributed conditional on the regressors. Show that the OLS estimate of the β vector is also its MLE.

4. Multicollinearity and NBA (Total 20 points)

For this problem, refer to the "Team Per Game Stats" found on https://www.basketball-reference.com/leagues/NBA_2018.html#all_team-stats-per_game (this is 2018 data); you can obtain data for different years by changing the year in the url. We only care about columns 7, 17, 18, 19, and 25 (FG% - Field Goal percentage, ORB - Offensive Rebounds/game, DRB - Defensive Rebounds/game, TRB - Total Rebounds/game, and PTS - points/game). Use the "Share & more" option to format data. Feel free to use any programming tools for this problem.

(a) Using 2017 data (change year to 2017 in url) as training, find the multiple linear regression fit for PTS as a function of only FG% and TRB (no β_0). Treat each row as a new observation. What can you infer about the contribution of TRB to total points? (4 points)

(b) Repeat part (a) but now also include ORB and DRB as predictors. What can you infer about the contribution of TRB to total points in this case? Explain, in words, why this happens. This is a result of Multicollinearity. (4 points)

(c) Using 2018 as the test data (all 30 rows), and FG% and TRB as the two predictors, compute SSE and MAPE when using the following training data: (i) 2017, (ii) 2016-2017, (iii) 2010-2017. (6 points)

(d) For (c), compute the residuals for (i), (ii), and (iii), and plot them against the predicted PTS.You should have 3 plots. What can you say about the variance of residuals as a function of predicted PTS? Why is this important? (3 points)

(e) For (c), compute the residuals for (i), (ii), and (iii), and plot their distribution. You should have 3 plots. Comment on whether they are Normal. Why does Normality matter? (3 points)

5. Time series analysis (Total 15 points)

For this problem, refer to the A5_q5.dat dataset on the website. This file contains 720 values. Each value represents the average request rate per 10 minutes for a sports website. Thus, the data spans 5 days. Use the first 4 days for training, and the 5th for testing. Report the average error over the 5th day, where error is defined as percentage difference between predicted and actual request rate. That is, use the required number of samples from the first 576 days to predict the 577th sample; compute the error for 577th sample. Now use the required number of samples from the first 577 samples to predict the 578th sample and compute its error. Finally, report the average error across the last 144 samples (577 to 720). Also plot the actual and predicted values for the 5th day in each case.

(a) Using EWMA with $\alpha = 0.5$. (3 points)

(b) Using EWMA with $\alpha = 0.8$. (3 points)

(c) Using Seasonal last observed with season = 144 samples. (2 points)

(d) Using AR(144) (3 points)

(e) Using AR(576) (3 points)

(f) What is your conclusion? (1 point)

6. More on Regression and Time series analysis (Total 10 points)

In this problem, we use the data from Azure trace; refer to A5_q6.dat dataset on the website. The file contains 576 values. Each value represents the number of VMs running in a data center for a 5 minutes interval. Thus, the data spans exactly 2 days.

(a) Split the dataset into 4 equal parts. For each quarter of the data, using simple linear regression (include β_0 term), plot the original data and the regression fit (using the corresponding quarter of data as training), and calculate the SSE in all 4 cases. (5 points)

(b) Split the dataset into 2 equal parts. Use the first half of the data as the training set. Predict the data points for the second half of the data using exponential moving average (α=0.5), auto regression (p=3), and seasonal average (s=288). For each technique report the average errors across all the 288 predictions. Note that you may have to use predicted data for training. From the original data, use only the first 288 values as part of training (you can augment them with predictions for 289th point, 290th point, etc.), and use the final 288 points for computing the error. (5 points)