960:463 Fall 2021 Final – Take Home

Due: Dec 20, 2021, 11:59 PM ET

(1) (20 pts) Final Problem 1 Dataset.txt data set on smoking

(a) Plot the scatter diagram showing the country names on the plot

(b) Plot the regression against the scatter diagram (Cigarette (X) and Deaths (Y)

(c) Discuss which are outliers, influence, and high leverage points. Perform Cook's Distance and DFBETA and DFFITS analysis.

(d) Estimate the influence of any outlier observation(s), if any.
Note: Please make sure that your graphs are clearly marked and easily interpretable.

(2) (20 pts) Final Problem 2 Dataset.csv. Do a thorough analysis of the data covering the following

(a) Multiple linear regression fit

(b) Statistical significance of the coefficients,

(c) What the residual and QQ plots convey, outliers, Cook's Distance, etc.

(d) Observations about multi-collinearity

(e) Perform a stepwise and all possible regression analysis.

(3) (15 pts) Final Problem 3 Dataset.csv. (# of Bacteria (y) vs Exposure time in min (x))

(a) Is a straight line model adequate?

(b) Based on residual analysis suggest an appropriate transformation either of the response variable or predictor or both?

(c) Assess the benefits of the transformation in (b).

(4) (15 pts) Final Problem 4 Dataset.txt

(a) Fit a simple linear regression for Y vs X and plot the externally Studentized residuals vs fitted values.

(b) Based on the residual plot, what modification would you suggest to the model to improve the fit? Compare the two fits to demonstrate the improvement.

(c) What might be a way to reduce computational errors in the beta estimates?

(5) (15 pts) Final Problem 5 Dataset.txt

(a) Fit a simple linear regression for Y (Revenue Estimate) vs X (Expenses). Comment on the homoscedasticity assumption.

(b) Do you have reason to conclude that the constant-variance assumption is not reasonable? If so, suggest a weighted least squares approach to address this issue.

(c) Does your weighted residual plot show any improvement? Use the nearest neighbor approach as well the "residual fit" method discussed in class.

(6) (15 pts) Final Problem 6 Dataset.txt

This data was collected to model a disaster scenario during tornados. Response variable y (0 or 1) refers to whether or no the neighborhood homes sustained substantial damage (1) or not (0). The dataset has two predictor variables - D which is a measure of the size of the home, and S is the intensity of the storm.

(a) Fit an appropriate regression model using S and the log2 (ie, to the base 2) of D as predictors and interpret the parameter estimates

(b) Does the data show that the model is adequate? Are the coefficients significant?

(c) Plot the estimated response against log2(D). Provide an interpretation to this plot.