Project: Multiple Regression Analysis

Proposal due: Friday, November 17, 2017 at 4:00 pm – Submit on Course website / "Assignments"
Project due: Friday, December 8, 2017 at 4:00 pm – Submit on Course website / "Assignments"

Goal: To study a data set using multiple linear regression. You will practice using techniques in Chapter 6 and provide an intelligent discussion and interpretation of your findings.

1) Proposal: Submit one page with: (1) your data source (including any URLs so I can look at your data); (2) state the question you're trying to answer with your analysis; (3) Give a description of each variable (x_i's and y's) and explain why y is predicted by the x_i's.

2) Project: Submit a report (pdf file), with steps clearly numbered, following the structure below. Your report should be self-contained, and include any necessary plots. You should also submit all the code you used (a MATLAB .m file), as well as your data (a .txt or .xls file is appropriate for this.) Your code should run on its own in Matlab.

Evaluation: You will be graded on the following:

- Correct performance of the steps below
- Discussion and interpretation of findings.

You will not be graded for the scientific merit of your data set, but rather on the quality of your analysis. It is fine if none of the regressor variables are significant, provided you analyze your data appropriately.

Who: You should work in pairs and submit one assignment.

Data: You will need to find a data set containing **at least 3 regressor variables (x's) that are related to one dependent variable (y)**, and **at least n = 20 data points**. The data should fit a regression model where $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + \ldots = y$. You can either make the measurements yourself, or use somebody else's measurements; if the latter make sure you properly cite your source. Every pair of students should use a different data set. The data should be continuous (cover a range of numerical values), NOT categorical variables (labels).

What: Your report should include the following numbered sections:

1. **Model and variables.** State the question you're trying to answer with your analysis. Give a description of each variable (x_i's and y's) and your data source (including URLs). State your initial model form. Give a plot for **each** x_i vs. y; note that any linear trend seen in these plots should be consistent with the following numerical analysis, and should be mentioned in your discussion. Also include a plot of each x_i against the other x_j's to check for colinearity (Consider the MATLAB function `corrplot`). Look for outliers and consider their effect in the following regressions; note that removal of outliers must be justified in the context of the data (e.g. erroneous data due to data obtained by a different method than other points, or inclusion of a member of a different population that the rest of the data.) Include these and the following plots in your report in the section where they are generated.
2. **MLR parameter estimation and confidence intervals.** Perform a multiple linear regression analysis on your data to obtain estimates of the regression parameters ($\beta_0, \beta_1, \beta_2, \ldots$) as well as an estimate of σ^2, the variance of the error term. Consider the MATLAB function `fitlm`. Compute confidence intervals for all of the regression parameters, and state the confidence level you are using. Describe the significance of the parameter estimates.
3. **Test for significance of regression.** Perform an F-test to test for the significance of regression.
4. **Hypothesis tests for specific parameters.** Test at least two hypotheses for specific non-zero values of the parameters. Use values that are of interest for your data, or if nothing seems appropriate, use the values from the previous step testing $\beta = 0$.

5. Final model building. Decide which subset of variables, if any, best describe your data set, based on the above results. Compute the new model fit using an F-test and R^2 and adjusted R^2, and discuss the results in step 7.

6. Analysis of residuals. Investigate whether the random component of the model follows the assumptions of a normally distributed, zero mean, constant variance random variable. Consider variation with the independent and dependent variables, as well as variation with factors in the data collection (e.g. time).

(7.) Additional analysis - optional, extra credit. Perform further analysis not described above. Stepwise model building, residual analysis, or transformation of data are suggestions, but other choices are fine.

8. Discussion. Discuss your findings: answer the questions you posed, analyze the assumptions for the model, discuss their appropriateness in your application, comment further. This is often 2-3 paragraphs.

SUBMIT THREE FILES via the course web site "assignments" tab :

(1) Report (pdf file)
(2) Matlab m-file
(3) data file.

You can include your data in the Matlab m-file and submit two files; if so, please note this in your report. I will execute your Matlab code, so please be sure the version you submit runs correctly:
to ensure that the data reads in correctly, please clear your workspace (type `clear`***) before testing.***

Continuous vs. categorical variables