M256/656 Open-ended Multiple Regression project to examine influences on “classic” car *performance* (horsepower OR acceleration) .

This is to be an INDIVIDUAL project. You may use your book and notes but you are NOT to work with others on this project.

This project utilizes an Excel data file, labeled: cars\_regr\_f12 which, for 392 model/year CLASSIC cars of the 1970’s and early 80’s gives their:

Model, Year, (the model/year’s) Horsepower, Acceleration, MPG, Number of Cylinders, Engine Displacement, Weight and Origin (American, European, or Japanese).

I took the liberty of adding to this data set a couple of “new variables” which are “functions” of the above variables

(namely, Weight Squared and dummy random variables for American and for Japanese), you will probably want to create up some other “new variables”.

Each row in the excel file, (besides the header), represents an observation.

The two most obvious *variables to predict* (i.e. potential “y” variables) are: Horsepower and Acceleration.

You choose the y variable (choose either Horsepower OR Acceleration) that you’d like to analyze.

For your “y” variable, you should use several runs of Excel (multiple) regression with the goal being to produce “as good of a model as possible” to predict y (horsepower or acceleration) using (a subset of) the predicting variables: year, mpg, number of cylinders, engine displacement, origin, and/or functions of them. Note: if your y is horsepower do NOT use acceleration as a predicting variable; conversely, if your y is acceleration, do NOT use horsepower as a predicting variable.

Towards that end, you’d like for your final model to be as “simple” (short) as possible but still have very good predictability with well-behaved “residuals” (estimated error terms), see more discussion about this below.

When you have two predictors that are highly correlated, say with |r| > 0.8 (which you could investigate with Data - data analysis - Correlation) , you might want to decide to leave one of the two variables out of the regression. Related ideas concerning potential f multicollinearity are discussed at the end of Chpt 12.

To make a multiple regression run (on the excel data: cars\_f12), with the file open go: Data – data analysis – Regression. Say you choose *acceleration* for y, then for Input Y range, you type in c1:c393 (NOTE: do NOT just click the top of the C column, because excel will not process it correctly, even though logically it should, and will give you an error message; this was the mistake I was making on Monday) Now if for the predicting “X” columns, you want to use mpg, cylinders, enginedisp, AND weight, then for Input X range, you type in d1:g393. Then you check that there are labels, you give a unique name to the New Worksheet Ply, check all boxes relating to residual output and plots, except for the line plots, so that you can examine them.

In making a multiple regression run, if pvalues for a term are considerably high (say > 0.2), you might want to drop that terms, one by one, from the next regression; related to the backward version of the forward stepwise regression discussed at end of Chpt 12.

Other things to take into account in evaluating the “goodness” of a model are: see if all of the standardized residuals are NOT extreme (i.e. NONE are above +3.0 nor below -3.0), see if the Adjusted R-square in the regression is close to the highest you can find, see if MSE (in excel, it is labeled: Residual MS), and its square root (“s”, typical amount off of prediction) are “low”.

You also want to do residual analysis, as discussed at end of Chpt 12, where you examine if your outputted plots of the residuals vs a predictor look like random scatters about a horizontal line, see if the residuals have close to constant variance in the plots, and that the normal probability plot on residuals resembles a straight line.

So, you are to look at several different potential regression models (i.e. set of predicting variables), to search for a “best” model. Suppose the next model you looked at had: mpg, weight, and the mpg\*weight interaction. You’d first need to rearrange the columns of the data set, so that the columns labeled mpg and weight are beside each other (in excel, the Input X variables for a regression must be a contiguous reference). You can do this by going to the Home tab, where, on the right, there are Insert and Delete icons. You could insert a (blank) Column after the mpg column, then copy the weight column to that (blank column) position,

Then delete the old weight column. Since you also want the interaction term, you insert another blank column after the (new) weight column, label this column as, say, mpgweightint and form the interaction term by multiplying the row respective entries in the mpg and weight columns, together. Note that you only need to enter the formula for the first blank row (here, row 2), then you can highlight this cell, and double-click on the dot in the lower right-hand corner in order to have excel to enter the remaining row’s formulas. Be sure to give a (new) unique name to the worksheet and examine the output and residuals from this analysis. Based on analyses to this point, decide which next analysis you’d like to run next, etc.

What and how to submit for your completed project?

By Wednesday December 5, address to: walsdm12@wfu.edu the following under one email, with the subject of:

M256\_mult\_reg:

1. (an attached) Microsoft Word portion of the project should consist of about a 10-20 line introduction. Then, a one to three page discussion, where you discuss your findings, being sure to justify your process of constructing your final model. Followed by a less than one page conclusion. Also,
2. (an attached) copy of your excel files, including all excel worksheets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |