Homework 3
MSAN 601

due 11.45pm September 19, 2016

n.b. All deliverables are required to be typed and all graphs and statistical output generated in R. Deliverables with *any* handwritten elements will not be accepted and will receive a grade of zero. You are required to use either RMarkdown or LaTeXto generate the pdf deliverable to be uploaded to Canvas. If using RMarkdown, also upload the Rmd file; if using LaTeX, also upload the R file. Show **all** calculations unless otherwise instructed. Also, be sure to upload the two additional .R files for the final question.

Question 1

In general, for a multiple regression model, what are the units of R^2?

Question 2

A researcher estimates the following two econometric models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

where X_3 is an irrelevant variable that does not help generate the Y at all. Will the value of R^2 be higher for the second model than the first? What about adjusted R_a^2? Explain succinctly.

Question 3

Using the anscombe data set available in the default set of R datasets, regress Y_i on X_i for $i = 1, 2, 3, 4$, i.e., generate four separate SLR models. **For each** of the four data sets, run and report summary statistics, generate a scatter plot and run a SLR model, reporting the regression function R^2, R_a^2, and the significance of b_1 (p-value). Intelligently discuss what you observe about each data set and the data sets as a whole.

Question 4

Download the data set extraColumnsOfRandomData.csv, a different version of the body fat percentage data from class, which includes body fat percentage (Y), hip circumference (X_1), age (X_2) and 25 additional columns of purely fictitious independent variables that have been generated at random. Regress the dependent variable against the first i independent variables (reading columns from left to right) as i goes from 1 to 27. For each regression model you execute, note or record the R^2 and R_a^2. Then, create a scatter plot that shows the following: the total number of independent variables on the x axis, and both R^2 and R_a^2 on the y axis. Do not do this manually. Create a R script that will run all 27 regression models, store the R^2 and R_a^2 in an object, **and** generate the graph.

Question 5

Download the data set related to a monthly electric bill called `electricBillData.csv`. While technically, some sort of time series analysis, or at least the use of a lagged variable, would be appropriate in this situation (since the observations are clearly dependent on time), we can still try to analyze the data using a fully cross-sectional approach. Complete the following steps after reading the documentation for this data set.

(a) Try to develop a regression model that predicts the monthly bill (in dollars) against the following variables in the data set: average temperature, heating degree days, cooling degree days, number of family members, whether or not there is a new electric meter, whether or not the home's first heat pump has been replaced, and whether or not the home's second heat pump has been replaced. Notice that the last three variables are dummy variables, but they have already been congured in the data set for you as ones and zeros.

(b) Interpret R^2 for the model.

(c) Did the introduction of the new electric meter and the new heat pumps (the first and the second) make a diffence in the house's bill? Explain your reasoning carefully and in a statistically sound way.

(d) Which variable is more important for explaining the final size of the bill—the number of family members present in the house during the month, or the total number of heating degree days? Explain your reasoning. Does your finding concern you? (It should, and we will address it in the next part of the question.)

(e) Build, and report, a correlation matrix between the following three variables: average temperature, heating degree days, and cooling degree days. What does the matrix suggest to you about the wisdom of having both average temperature and heating degree days in the model?

(f) Improve your model by dropping the variable average temperature and re-execute the regression. Notice that the estimated coefficients associated with the heating degree days and cooling degree days variables now make better sense (because we have removed a primary source of multicollinearity within the model). Read the data documentation again and notice the location of the two professors who provided us with the data. Do a little Internet research. Why might cooling degree days not be a statistically signicant contributor to the model after attending to this multicollinearity problem?

Question 6

Download `SENIC_data.csv`. Fit a second-order regression model regressing the number of nurses (Y) on available facilities and services (X).

1. Fit the second-order regression model. Plot the residuals against the fitted values. How well does the second-order model appear to fit the data?
2. Obtain R^2 for the second-order regression model. Also obtain the coefficient of simple determination for the first-order regression model. Has the addition of the quadratic term in the regression model substantially increased the coefficient of determination?
3. What fraction of the error not explained by the first-power term is explained by the quadratic term when added?
4. Test whether the quadratic term can be dropped from the model using $\alpha = 0.01$. State the alternatives, decision rule and conclusion.

Question 7

A student who used a regression model that included indicator variables was upset when receiving only the following output on the multiple regression printout: `XTRANSPOSE_X_SINGULAR`. What is a likely source of the difficulty?

Question 8

Download `SENIC_data.csv`. Regress length of stay (Y) on age (X_1), routine culturing ratio (X_2), average daily census (X_3), available facilities (X_4) and region (X_4,X_5,X_6).

1. Fit a first order regression model. Let $X_5 = 1$ if NE and 0 otherwise, $X_6 = 1$ if NC and 0 otherwise, and $X_7 = 1$ if S and 0 otherwise.
2. Test whether the routing culturing ratio can be dropped from the model; use a level of significance of 0.05. State the alternatives, decision rule and conclusion.