Homework 4
MSAN 601

due 11.45pm September 27, 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 `R`Markdown or LaTeXto generate the `pdf` deliverable to be uploaded to Canvas. If using `R`Markdown, 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

Use `Transact` from `alr4` package.

The data consists of a sample of branches of a large Australian Bank. Each branch makes transactions of two types, and for each of the branches, we have recoded the number `t1` of type 1 transactions and the number `t2` of type 2 transactions. The response is `time`, the total minutes of time used by the bank.

Define `a=(t1+t2)/2` to be the average transaction time, and `d=t1-t2`, and fit the following four regression functions

$$Y = \beta_{01} + \beta_{11} t_1 + \beta_{21} t_2 \tag{1}$$
$$Y = \beta_{02} + \beta_{32} a + \beta_{42} d \tag{2}$$
$$Y = \beta_{03} + \beta_{23} t_2 + \beta_{43} d \tag{3}$$
$$Y = \beta_{04} + \beta_{14} t_1 + \beta_{24} t_2 + \beta_{34} a + \beta_{44} d \tag{4}$$

(a) In (4), some of the coefficient estimates are labeled as `NA`. Epxlain what this means and why this happens.

(b) What aspects of the fitted regression are the same? What are different?

(c) Why is the estimate for `t2` different in (1) and (3)?

Question 2

Use `UN11` from `alr4` package.

(a) In the SLR of log(`fertility`) $\sim$ `pctUrban`, provide an interpretation of the estimated coefficient for `pctUrban`.

(b) Verify that in the regression of log(`fertility`) $\sim$ log(`ppgdp`) + `lifeExpF` a 25% increase in `ppgdp` is associated with a 1.4% decrease in expected fertility.

Question 3

Use `cakes` from `alr4` package. The response variable is `Y`.

(a) Identify the best possible linear regression function using variables `X1` and `X2` (you can use as many regressors as you like). Show/explain modeling iterations.

(b) Repeat (a), but now include the dummy variable `block` (code the dummy variable 0/1). Show/explain modeling iterations. Is the coefficient of `block` significant?

Question 4

Use `fuel2001` from `alr4` package. `FuelC` is the response. State the null and alternative hypotheses for the overall F-test, perform the test and summarize the results. State your conclusion.

Question 5

Use `cakes` from `alr4` package.

Fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$ and compute the following hypothesis tests:

(a) $H_0 : \beta_5 = 0$

(b) $H_0 : \beta_2 = 0$

(c) $H_0 : \beta_1 = \beta_2 = \beta_5 = 0$

Show all work and state your conclusions.

Question 6

Use `landrent` from `alr4` package.

This data was collected to study the variation in rent paid in 1977 for agricultural land planted to alfalfa. The variables are average rent per acre `Y` planted for alfalfa, average rent paid `X1` for all tillable land, density of dairy cows `X2` (number per square mile), proportion `X3` of farmland used as pasture, and `X4=1` if liming in required to grow alfalfa and 0 otherwise.

The unit of analysis is a county in Minnesota; the 67 counties with appreciable rented farmland are included. Alfalfa is a high-protein crop that is suitable feed for dairy cows. It is thought that rent for land planted to alfalfa relative to rent for other agricultural purposes would be higher in areas with high density of dairy cows and rents would be lower in countries where liming is required, since that would mean additional expense. Use all the techniques at your disposal to explore the data with regard to understanding rent structure. Show relevant work and output. Summarize your results and state your conclusion.

Question 7

In `R`, code a backward stepwise regression algorithm. Choose two different elimination/addition criterion and explain why you chose these criterion over others. Compare the results when employing each criteria on the dataset `Rateprof` from `alr4` package (include *ONLY* `integer` and `numeric` regressors).

Although you may use the `lm` function, you are **not** permitted to use a 'canned' stepwise regression function/algorithm coded up by another author.

Show results for both criterion and explain any difference.

Question 8

Presume you are given a data set with a response variable Y and $p-1$ predictor variables X. In `R`, code an algorithm that will generate the same **textual** output as the `residualsPlots` function in the `car` package. Recall that the `residualsPlots` function tests the null hypothesis H_0 : no lack of fit, where the curvature of the residuals is tested by including a squared term for each predictor variable sequentially, refitting the model, and evaluating the statistical significance of the coefficient of the squared predictor term, resulting in t^* and p-values (see course notes for full details). Recreate the functionality, but now test the residuals by sequentially including a predictor variable in the model raised to the power $\delta \in \{1.2, 1.5, 1.7\}$. Run t-tests for all $p-1$ predictors and all three values of δ. Print results to the console.

Name your file `hw4q8.R`. I will `source()` this file, which should read in a `csv` called `hw4q8.csv` which contains Y in column 1 followed by $p-1$ regressors. Assume `header = F`. Upload your file to Canvas.

Question 9

In `R`, write a function that generates **all** possible combinations of partial coefficients of determination. The algorithm should print the output to the console **neatly** in the following manner

(1) all $R^2_{YX_j}$

(2) all $R^2_{YX_j|X_k}$

(3) all $R^2_{YX_j|X_kX_\ell}$

(4) etc.

Name your file `hw4q9.R`. I will `source()` this file, which should read in a `csv` called `hw4q9.csv` which contains Y in column 1 followed by $p-1$ regressors. Assume `header = F`. Upload your file to Canvas.

Question 10

Online, I observed the following discussion:

> "The Kolgomorov-Smirnov test, Shapiro test, etc.... all reject the hypothesis that a distribution is normal. Yet when I plot the normal quantiles and and histogram, the data is clearly normal. Maybe because the power of the tests are high? The sample size is around 650. So shouldn't at least one of these tests fail to reject the null hypothesis?"

To which a user responded:

> "Normality testing is a waste of time and your example illustrates why. With small samples, the normality test has low power, so decisions about what statistical models to use need to be based on a priori knowledge. In these cases failure to reject the null doesn't prove that

the null is even approximately true at the population level.

When you have large samples, normality tests become ridiculously powerful, but they don't tell you anything you didn't already know. No real quantity is exactly normally distributed. The normal distribution is just a mathematical abstraction that's a good enough approximation in a lot of cases. The simplest proof of this is that there is no real quantity (at least none that I can think of) that could take any real number as its value. For example, there are only so many molecules in the universe. There are only so many dollars in the money supply. The speed of light is finite. Computers can only store numbers of a finite size, so even if something did have a support of all real numbers, you wouldn't be able to measure it.

The point is that you already knew your data wasn't exactly normally distributed but the normality tests tell you nothing about how non-normal the data is. They give you absolutely no hint as to whether your data is approximately normally distributed such that statistical inference methods that assume normality would give correct answers. Ironically, common tests (e.g. the T-test and ANOVA) that assume normality are more robust to non-normality at large sample sizes."

Using the response as a guide, either confirm or reject this user's response to the question, using empirical (simulated) example(s) and directly speaking to the issue of 'power' of the normality test.