Homework 1
MSAN 601

due 11.45pm September 05, 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.

Question 1

What assumptions are made about the error terms in the normal error regression model? Use **both** English and mathematical notation to explain each assumption. Why are these assumptions made?

Question 2

Prove the equivalence of the F-Test and the T-Test on β_1 for an SLR model.

Question 3

Are hypotheses tested concerning the actual values of the coefficients, e.g., β_1, or their estimated values, e.g., b_1? Why?

Question 4

The OLS estimators we derived in class are considered to be the Best Linear Unbiased Estimators (BLUE). Is it a necessary condition to assume normality of the error terms to obtain BLUE OLS estimators? If so, why? If not, when/why do we assume normality of the error terms?

Question 5

You compute a coefficient of determination for a regression model an obtain an $R^2 = 0.832$. What does the strength of the coefficient of determination say about the causal relationship between the explanatory and response variables?

Question 6

You compute a coefficient of determination for a regression model, regressing crime rate per capita (Y) on the size of municipal police force (X), obtaining an $R^2 = 0.6533$. What can you say about the relationship between Y and X?

Question 7

You run a regression on 6690 observations, and obtain an $SSR = 24,332$ and an $SSTO = 36,234$. Compute MSE.

Question 8

Download copierMaintenanceData.csv from Canvas. **Column 1 are the Y_i and Column 2 are the X_i.** X_i are the number of photocopiers serviced by a service company at a given location, and Y_i are the total number of minutes spent by the service person.

Use `R` to compute the answers to the following questions. **Do not use 'black-box' methods, e.g.,** `lm`. copierMaintenanceData.csvAssume $\alpha = 5\%$ for all relevant questions.

1. Obtain the OLS estimates.
2. Write out the fitted regression equation.
3. What is the numerical value of $\sum e_i^2$?
4. Does b_0 provide any relevant information in this context?
5. Obtain a point estimate of the mean service time when 5 copiers are serviced.
6. Test the hypothesis that $H_0 : \beta_1 = 0$ using a t test. Show all work.
7. Test the hypothesis that $H_0 : \beta_1 \leq 1$. Show all work.
8. Obtain a confidence interval of the expected mean service time when 5 copiers are serviced.
9. Test the hypothesis that $H_0 : \beta_1 = 0$ using an F test. Show all work. Prove (computationally) that $(t^*)^2 = F^*$.
10. Come up with your own estimates for b_0 and b_1 (pick any values for them, so long as they are not equal to the OLS values above) and write out your personal fitted regression equation. Using the fitted regression equation, generate the $\hat{Y}_i$, and subsequently compute the $\sum e_i^2$? What is the value? Repeat this exercise until you are convinced that the OLS estimates are the best.

Question 9

Using the data in Table 1, construct a linear regression model by calculating parameters manually to predict the prostrate cancer death rate from per day dietary consumption in a country and answer the following questions.

1. What are the OLS estimates?
2. Write out the fitted regression equation.
3. Does b_0 provide any relevant information in this context?
4. Interpret the fitted slope coefficient.
5. Construct a 95% confidence interval for b_0 and b_1.

Question 10

'In the context of SLR model, whether we regress X on Y or Y on X, the parameter interpretation does not change.' Verify and explain if this statement is true using the `faithful` dataset from base `R`.

Country	Dietary fat (g/day)	Death rate (per 100,000)
Philippines	29	1.3
Mexico	57	4.5
Colombia	47	5.4
Yugoslavia	72	5.6
Panama	58	7.8
Romania	67	8.8
Czechoslovakia	96	9.1
Spain	97	10.1
Finland	112	11.7
United Kingdom	143	12.4
Canada	142	13.4
France	137	14.4
Australia	129	15.1
United States	147	16.3
Sweden	132	18.4

Table 1: Fat consumption and prostate cancer deaths

Question 11

What do you observe about the F-testa in the summary of the two models in Question 10? Explain.

Question 12

Consider the following data:

1. `X = c(1:100) and Y = x ** 2`
2. `X = c(1:100) and Y = 2 * x`

In both, we can establish a relation between Y and X. What do you expect as the R^2 value when we regress Y on X? Verify using `R` and explain.

Question 13

From the discussion of SLR so far, how do you believe outliers will effect the regression line?