Stat 302 Midterm 2

Thursday, March 17, 2016

Student Name _____

Student Number _____

You have exactly 50 minutes to complete this exam.

This test has 6 pages including this one.

Only non-programmable calculators are allowed for electronics.

That means no graphing calculators and no phones.

DO NOT OPEN THIS PACKAGE OR TURN THE PAGE UNTIL INSTRUCTED TO DO SO. OPENING THE EXAM EARLY OR CONTINUING TO WRITE AFTER TIME IS CALLED WILL RESULT IN A SCORE OF ZERO FOR THE EXAMINATION.

Protips:

- Show your work whenever appropriate. It shows understanding, and that's what's being tested.

- Use the backs of pages if space is an issue.

- If you get stuck on a part, don't abandon the question. Often later parts can be answered without earlier ones.

- Try not to panic, it rarely helps.

Good luck!

Question	1	2	3	4	5	Total
Out of	10	5	6	7	4	32

Problem 1, Total / 10 (1 each)

Fill in the letter for each term. Each term is used exactly once. Write clearly.

1. Akaike Information Criterion (AIC) _____

2. Dummy variable _____

3. Imputation _____

4. Interaction _____

5. Leave-One-Out _____

6. Overfitted _____

7. Quantile-Quantile _____

8. Stepwise Method _____

9. Training Set _____

10. Variance Inflation Factor _____

A) Used to find a good model according to some criterion. Works by repeatedly 'dropping' or 'adding' one term to the current model.

B) Used to specify whether an observation is in a specific category, rather than some baseline.

C) Term for a model that applies well to the data set given, but poorly to new observations.

D) The part of the data used to make a model in a cross-validation.

E) Plot used to explore any deviations from normality (or other specified distribution) in a collection of values.

F) A measure to compare statistical models. Considers both model fit and complexity.

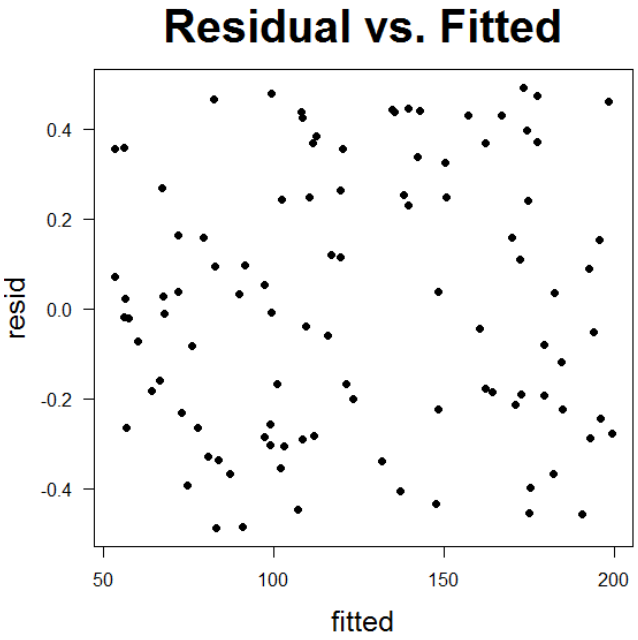
G) A special case of K-fold cross validation using training sets of N-1 observations.

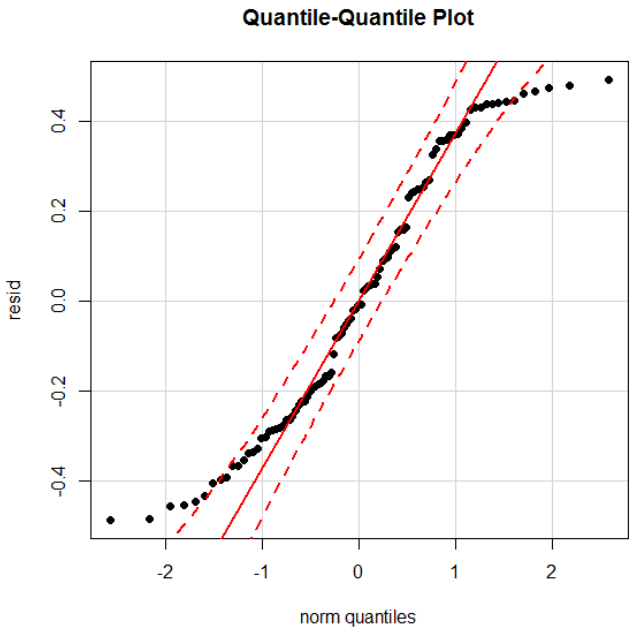
H) A regression term made of two (or more) variables, multiplied together.

I) A measure of co-linearity of a regression term.

J) General term for methods of replacing missing data.

Problem 2, Total / 5





Consider the above diagnostic plots for a regression model.

A) (3 pts) What potential problems can you see from these plots?

B) (2 pts) Would a Shapiro-Test be more likely to have a small (< 0.05) or large (> 0.05) p-value for this dataset? Explain.

Problem 3, Total / 6

Models X, Y, and Z below are used to describe the same variables. There is no missing data.

X: Blood Pressure ~ Sex + Age + Age^2 + Weight

Y: Blood Pressure ~ Sex + Age + Age^2 + (Age:Weight)

Z: Blood Pressure ~ Sex + Age + Age^2 + Weight + Vegetarian

A) (2 pts) Which, if any, of these models would the stepwise method never select? Why?

B) (2 pts) Which model, X or Z, has a higher R-squared? How are you sure?

C) (2 pts) If model X was selected by the stepwise method using AIC, could model Z be selected using BIC? Why? Assume N > 10.

Problem 4, Total / 7

Translink wants to predict the delay for the 135 line from info in the last hour, it is using…

- (hastings) Traffic on Hastings St., the main road of the 135, in 1000's of cars,
- (usage) The number of passengers, in 100's, using the bus
- (event) a dummy variable indicating there is a special event happening.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.24       1.33     -3.182  0.00265 **
hastings     0.13       0.02      5.748 7.37e-07 ***
usage        0.64       0.67      0.958  0.34318
I(usage^2)   0.19       0.08      2.369  0.02217 *
event       -0.32       0.23     -1.353  0.18275
```

A) (2 pts) What does the intercept mean, in this case?

B) (2 pts) Predict the bus usage when, in the last hour 14000 Cars used Hastings, 650 people used the bus, and there is **not** a special event happening. Show your work by including the regression equation.

C) (1 pts) Consider two days with the same amount of bus users and car traffic, but one day has a special event. Is there evidence that the bus delay will be different between the days?

D) (2 pts) For unknown reasons, the drivers sometimes don't record usage on Saturdays, so there is missing data. Is this missing data ignorable?

Problem 5, Total / 4

Consider the following model of adult levels of cholesterol, a naturally occurring substance in the body that can be harmful at high levels.

$$\text{cholesterol} \sim \beta_0 + \beta_1(\text{sex} = \text{M}) + \beta_2(\text{age}) + \beta_3(\text{body mass index}) + \text{error}$$

A) (2 pts) What does the dummy variable 'sex = M' mean, in this case? (Use β_1 in place of a number, if it helps)

B) (2 pts) If the effect of age was different for females that for males. Describe a term that should be added to this model to reflect the difference in effects?