ST552 Final

Winter 2015

Answer the questions in the spaces provided on this exam.

Name: ______________________________

- You have 110 minutes to complete the exam.
- There are 3 questions. Answer all of the questions.
- Please
  - do not look at the exam until I tell you and
  - stop writing when I announce that the exam is over.
- There is one page of statistical tables at the end of the exam. You may remove the page of tables if you desire.

Question	Points	Score
1	15	
2	15	
3	15	
Total:	45	

1. An experiment was conducted to explore the relationship between the *lifetime* (measured in days) and sexual activity of fruitflys.

   125 fruit flys were divided randomly into 5 treatment groups, each of 25 flys. Each treatment was designed to simulate a different level of sexual activity, with levels: *none*, *one*, *low*, *many* and *high*.

   The *thorax length* of each male was also measured (in mm) as this was known to affect lifetime.

   One observation in the *many* group was lost.

   The following model was fit:

$$\log\left(\text{Lifetime}_i\right) = \beta_0 + \beta_1 \text{Thorax Length}_i + \beta_2 \text{one}_i + \beta_3 \text{low}_i + \beta_4 \text{many}_i + \beta_5 \text{high}_i + \epsilon_i$$

   where *one*, *low*, *many*, and *high* are indicator variables for the respective treatment groups, resulting in the following estimates and standard errors:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{pmatrix} = \begin{pmatrix} 1.84 \\ 2.72 \\ 0.05 \\ -0.12 \\ 0.09 \\ -0.42 \end{pmatrix}, \quad \text{SE}\left(\hat{\beta}\right) = \begin{pmatrix} 0.2 \\ 0.23 \\ 0.05 \\ 0.05 \\ 0.06 \\ 0.06 \end{pmatrix}, \quad \hat{\sigma} = 0.19$$

   (a) Construct a 95% confidence interval for the parameter, β_1. (2)

(b) Interpret the point estimate for β_1, in the context of the study on the **original scale** of lifetime. (`exp(2.72) = 15.18`) (3)

(c) Are any additional assumptions (beyond the usual regression assumptions) required for your interpretation above? (2)

(d) What additional information is required to construct a confidence interval on $\beta_2 - \beta_3$? (2)

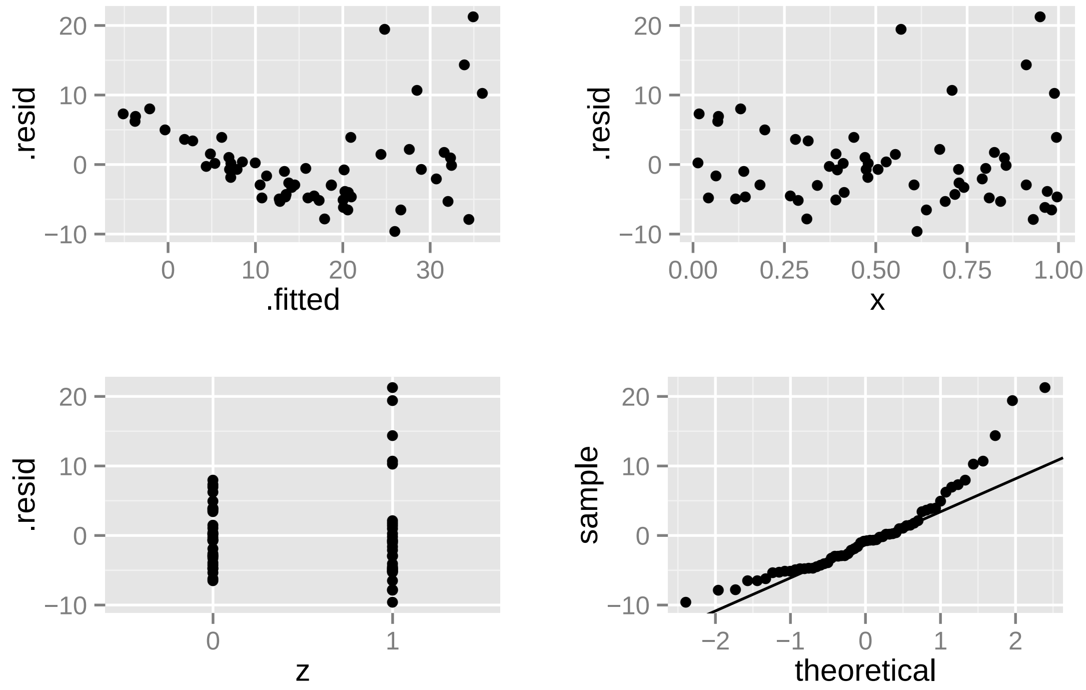
(e) A more complicated model that treats *Thorax Length* as a categorical variable and includes interactions between the treatment groups and thorax length is also fit, with a resulting residual sum of squares (RSS) of 2.59 on 77 degrees of freedom.

i. What is the value of the F-statistic from an Extra Sum of Squares F-test comparing this model to the one above? (4)

ii. What is the special name for this F-test, and what would we conclude? (2)

2. (a) The following residual plots come from a regression of the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i \quad i = 1, \ldots, 60$$



i. Name the assumption that appears to be violated. (2)

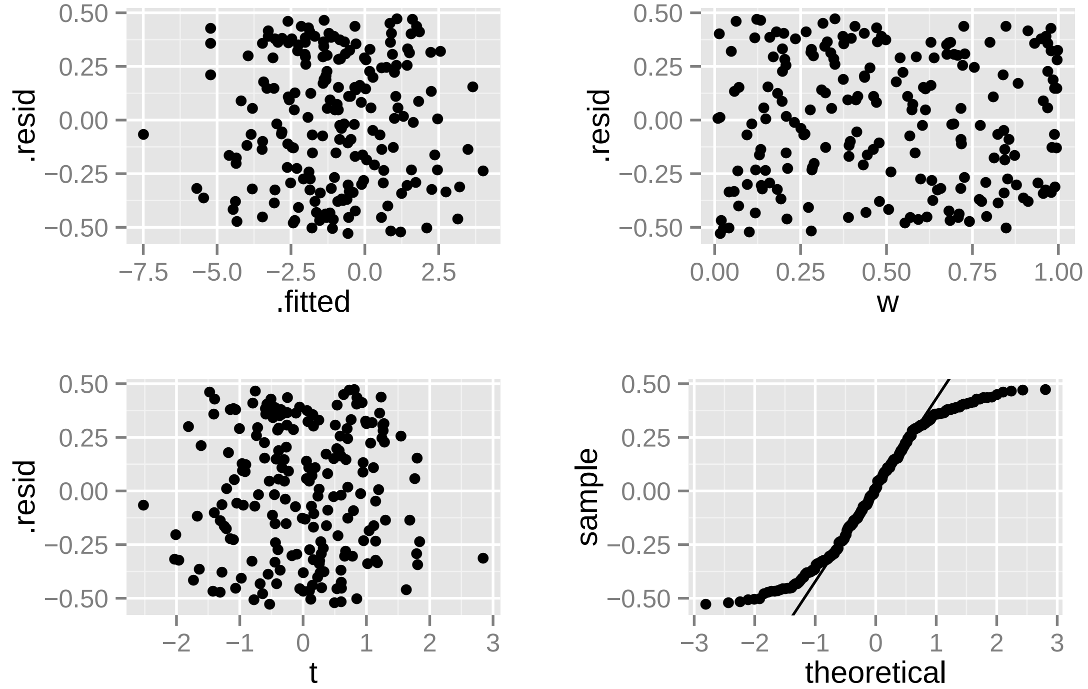
ii. Describe the evidence you see in the plots for the violation. (1)

iii. What are the consequences of proceeding with inference ignoring the violation? (1)

iv. How would you suggest proceeding? (1)

(b) The following residual plots come from a regression of the form:

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 t_i + \epsilon_i \quad i = 1, \dots, 200$$



i. Name the assumption that appears to be violated. (2)

ii. Describe the evidence you see in the plots for the violation. (1)

iii. What are the consequences of proceeding with inference ignoring the violation? (1)

iv. How would you suggest proceeding? (1)

(c) A client has run diagnostics on a regression analysis and identified a single observation with very high leverage, but she admits she doesn't know what leverage is or how to proceed.

i. What does *high leverage* mean? (1)

ii. Sketch a scatterplot that includes a point that has **high leverage** but is **not influential**. (Make sure you label your axes, clearly identify the point of interest, and label any fitted lines you add) (1)

iii. Sketch a scatterplot that includes a point that has **high leverage** and **is influential**.(Make sure you label your axes, clearly identify the point of interest, and label any fitted lines you add) (1)

iv. How would you advise your client to proceed? (2)

3. (a) i. Describe what is meant by **multicollinearity**. (2)

ii. What are the consequences of multicollinearity? (2)

Question 3 continues on the next page...

(b) i. In one sentence, describe what is meant by **variable selection** in the context of multiple linear regression. (1)

ii. Give a reason why variable selection might be recommended. (2)

iii. Give a reason why variable selection might be avoided. (2)

Question 3 continues on the next page...

(c) Some extensions to multiple linear regression include: (6)

- Robust regression
- Generalized least squares
- Regularized regression
- Logistic regression
- Non-linear regression

Pick **two** methods from the above list and describe how they differ from the usual case of linear regression.