STATISTICS 608 Linear Models -EXAM I
February 18, 2013

Student's Name: _____

Student's Email Address: _____

INSTRUCTIONS FOR STUDENTS:

1. There are **13** pages including this cover page.
2. You have exactly 50 minutes to complete the exam.
3. There may be more than one correct answer; choose the best answer.
4. You will not be penalized for submitting too much detail in your answers, but you may be penalized for not providing enough detail.
5. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions next week.
6. You may use one 8.5" X 11" sheet of notes and a calculator.
7. At the end of the exam, leave your sheet of notes with your proctor along with the exam.

I attest that I spent no more than 50 minutes to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature: _____

INSTRUCTIONS FOR PROCTOR:

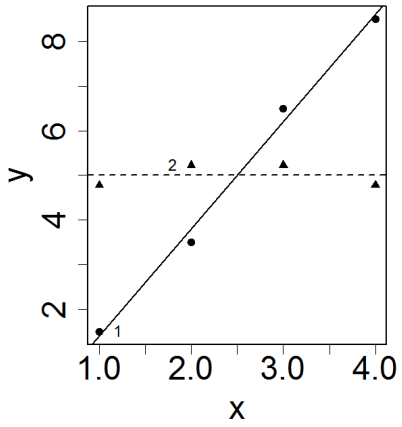
Immediately after the student completes the exam scan it to a pdf file and have student upload to Webassign.

1. I certify that the time at which the student started the exam was _____ and the time at which the student completed the exam was _____.
2. I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
3. I certify that the exam was scanned in to a pdf and uploaded to Webassign in my presence.
4. I certify that the student has left the exam and sheet of notes with me, to be returned to the student no less than one week after the exam or shredded.

Proctor's Signature: _____

Part I: Multiple choice

1. Two linear models were fit to two data sets, shown on the plot below (drawn to scale). Model 1, the solid line fit to the circle points, and Model 2, the dashed line fit to the triangular points, had the same RSS, and were fit to data sets with the same mean for x, 2.5, and the same mean for y, 5. Which of the following statements is true?



   (a) SSReg for model 1 $>$ SSReg for model 2

   (b) SSReg for model 1 $<$ SSReg for model 2

   (c) SSReg for model 1 $=$ SSReg for model 2

   (d) The answer cannot be determined for this data set.

2. Suppose we are interested in the usual simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$ ($i = 1, \ldots, n$), where the errors e_i are independent of each other, and are normally distributed with mean 0 and variance σ^2. Which of the following statements for the model is true about the residuals $\hat{e}_i$ calculated as the difference between the observed and fitted values $y_i - \hat{y}_i$?

   (a) The residuals have the same variance.

   (b) The residuals are correlated.

   (c) The mean of the i^{th} residual $\hat{e}_i$ is $\beta_0 + \beta_1 x_i$.

   (d) A standardized residual r_i larger than 2 in absolute value indicates a point with high leverage in data sets with $n < 100$.

Part II: Short Answer

3. Explain as if to someone with very little statistical experience why a variance stabilizing transformation might be needed in some cases.

4. Suppose that $\text{Var}(e_i|x_i) = x_i^2\sigma^2$ for the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ ($i = 1, 2, \ldots, n$). Clearly define and write down an estimate for the parameters β_0 and β_1. There is no need to finish any algebra that you may set up; simply write down equations and define terms not already defined.

Part III: Long Answer

5. I need help buying a house! The market value (for tax purposes) of 22,208 houses in Houston have been plotted along with square footage (area).[1] I want to model the market value of a house as a function of square footage. The first model I fit to the data was:

$$\text{Market Value} = \beta_0 + \beta_1 \text{Square Feet} + e \tag{1}$$

At the end of this exam are some output from fitting model (1) as well as some plots.

(a) Does the straight line regression model (1) seem to fit the observed data well? If not, list any weaknesses apparent in model (1).

(b) Suppose we used model (1) to calculate a prediction interval for Market Value for a house with Square Feet = 1000. Would the interval be too short, too long, or about right? Give a reason to support your answer.

The second model fitted to the data was

$$\log(\text{Market Value}) = \beta_0 + \beta_1 \log(\text{Square Feet}) + e \tag{2}$$

Output from model (2) also appears at the end of the exam.

[1] Texas laws apparently keep the actual sale price a secret.

(c) Interpret the slope from model (2) in context.

(d) Is model (2) an improvement over model (1) in terms of predicting Market Value? If so, describe all the ways in which it is an improvement.

(e) For a house with Square Feet = 2500, the R output for the confidence interval for Market Value using Model 2 (without transforming the endpoints) was (12.032, 12.038). What is the confidence interval for Market Value in dollars? Interpret the interval in context.

6. Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i \, (i = 1, 2, \ldots, n)$, where e_i is a randomly distributed error term; that is, $E[Y_i|X = x_i] = \beta_0 + \beta_1 x_i$. Assume the usual properties of the errors:

   i. The errors $e_1, e_2, \ldots, e_n$ are independent of one another.

   ii. The errors $e_1, e_2, \ldots, e_n$ have a common variance σ^2.

   iii. The errors are normally distributed with a mean of 0 and a variance of σ^2.

   Since the regression model is conditional on X, the values of the predictor variable are known fixed constants.

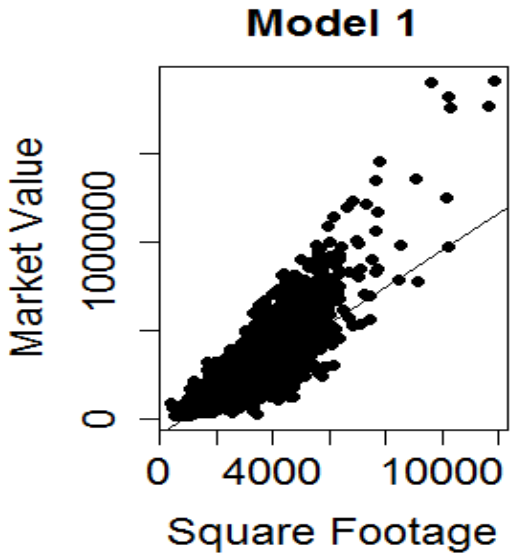
   (a) Show that the least squares estimates of the parameters conditional on X are unbiased. (Hint: matrix notation may make this easier.)

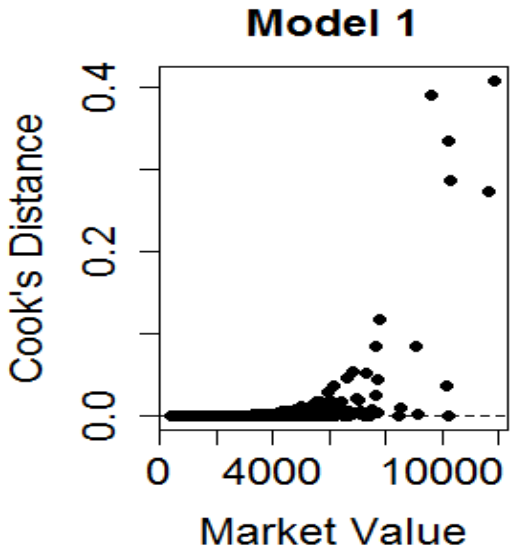
   (b) Show that, where $\boldsymbol{\beta}$ is the vector of parameters and $\mathbf{X}$ is the design matrix for our model, $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

   (c) Show that $\text{Var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$, where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

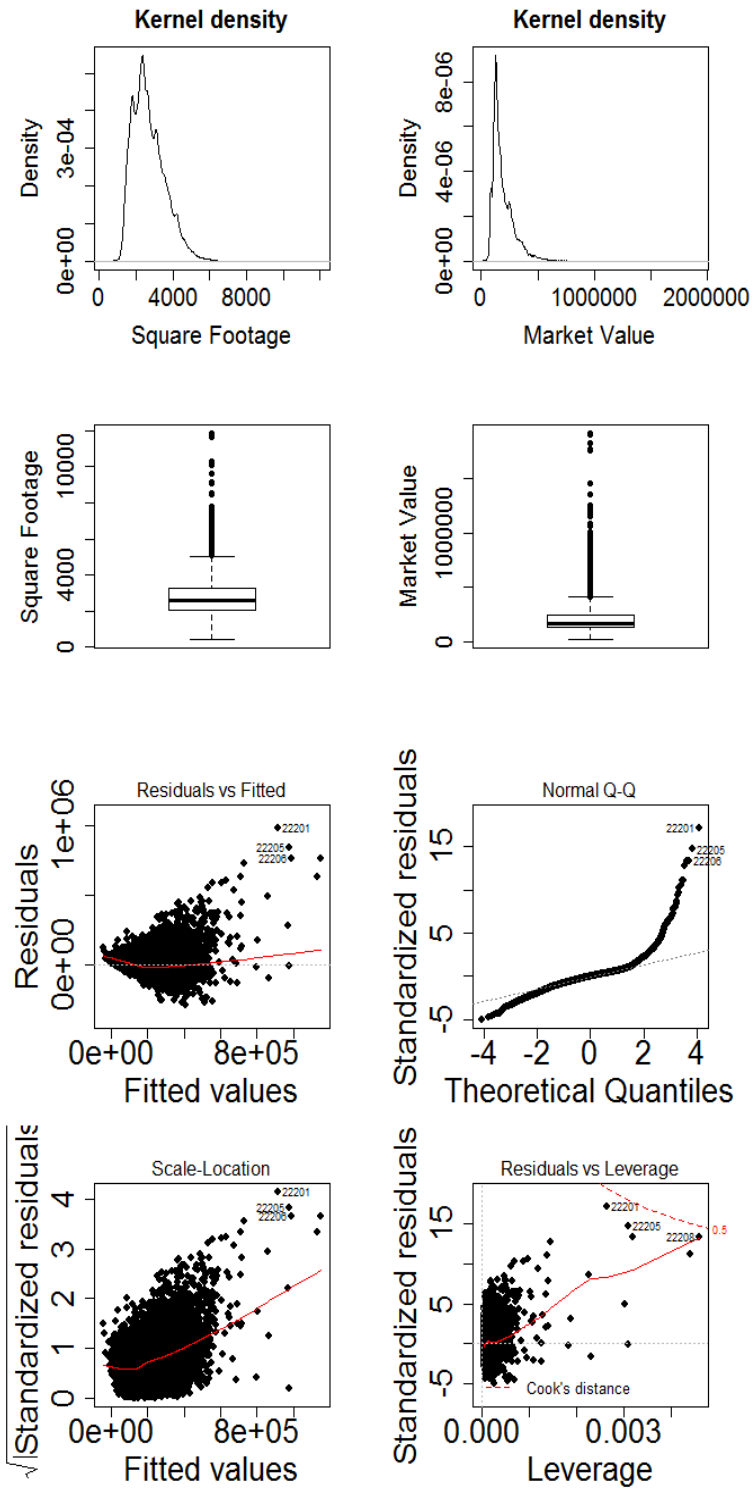
Model 1:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.689e+04  1.219e+03  -71.25   <2e-16 ***
houses$Sq.Ft  1.045e+02  4.256e-01  245.49   <2e-16 ***

Residual standard error: 57620 on 22206 degrees of freedom
Multiple R-squared: 0.7307,     Adjusted R-squared: 0.7307
F-statistic: 6.027e+04 on 1 and 22206 DF,  p-value: < 2.2e-16

Analysis of Variance Table
                Df     Sum Sq    Mean Sq F value    Pr(>F)
houses$Sq.Ft     1 2.0006e+14 2.0006e+14   60265 < 2.2e-16 ***
Residuals    22206 7.3717e+13 3.3197e+09
```



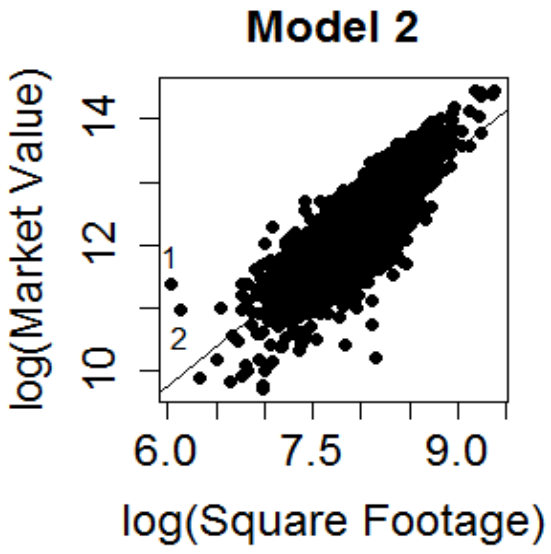


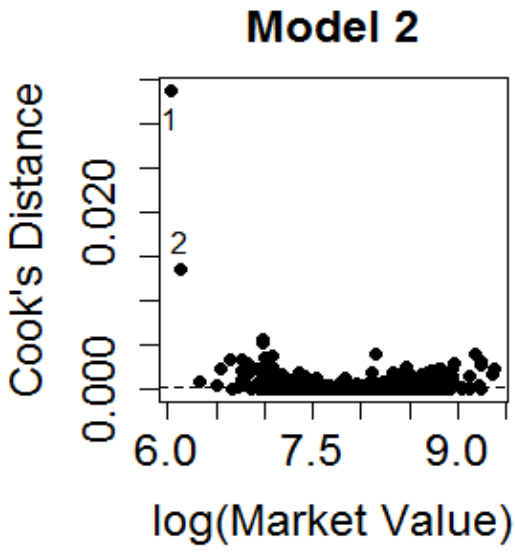
Model 1:


Kernel density
Density
0e+00
3e-04
0
4000
8000
Square Footage
Kernel density
Density
0e+00
4e-06
8e-06
0
1000000
2000000
Market Value
Square Footage
0
4000
10000
Market Value
0
1000000
Residuals vs Fitted
Residuals
0e+00
1e+06
22201
22205
22206
0e+00
8e+05
Fitted values
Normal Q-Q
Standardized residuals
-5
5
15
22201
22205
22206
-4
-2
0
2
4
Theoretical Quantiles
Scale-Location
Standardized residuals
0
1
2
3
4
22201
22205
22206
0e+00
8e+05
Fitted values
Residuals vs Leverage
Standardized residuals
-5
5
15
22201
22205
22206
0.5
Cook's distance
0.000
0.003
Leverage

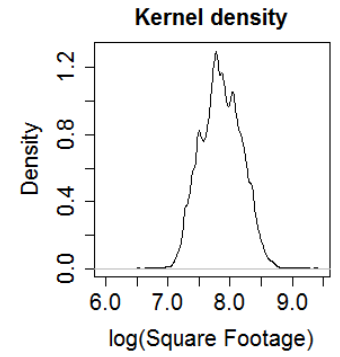
Model 2:

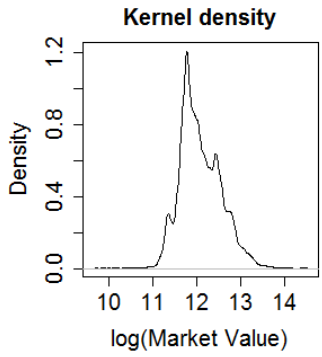
```
x2<-log(houses$Sq.Ft)
y2<-log(houses$Market)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.222704   0.037547    59.2   <2e-16 ***
x2          1.254116   0.004776   262.6   <2e-16 ***

Residual standard error: 0.2306 on 22206 degrees of freedom
Multiple R-squared: 0.7564,     Adjusted R-squared: 0.7564
F-statistic: 6.895e+04 on 1 and 22206 DF,  p-value: < 2.2e-16

Analysis of Variance Table
             Df Sum Sq Mean Sq F value    Pr(>F)
x2            1 3665.8  3665.8   68946 < 2.2e-16 ***
Residuals 22206 1180.7     0.1
```

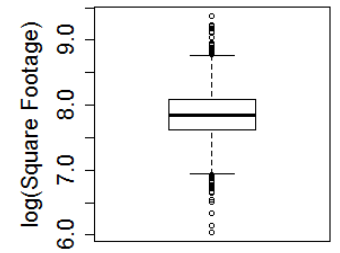


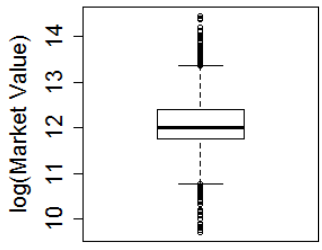


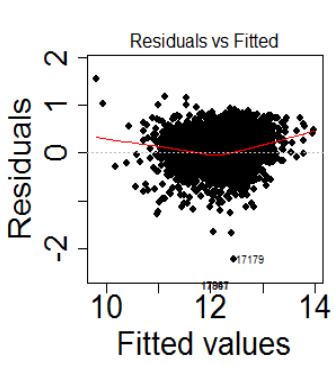
Model 2:

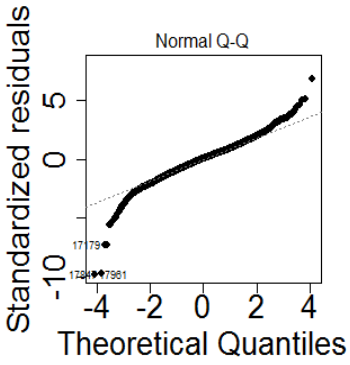

Kernel density
Density
log(Square Footage)

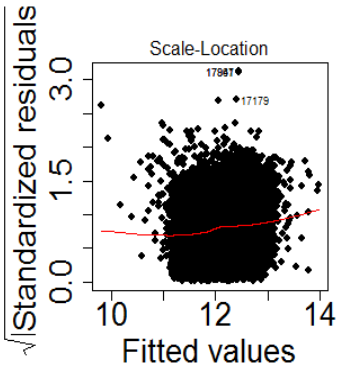

Kernel density
Density
log(Market Value)

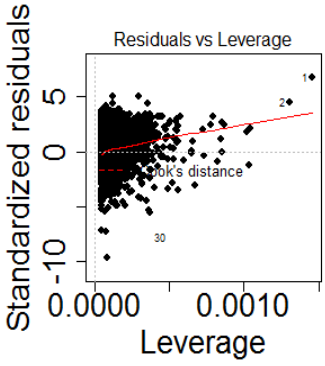

log(Square Footage)


log(Market Value)


Residuals vs Fitted
Residuals
Fitted values
17179


Normal Q-Q
Standardized residuals
Theoretical Quantiles
17179


Scale-Location
√|Standardized residuals|
Fitted values
17179


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1
2
30