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:

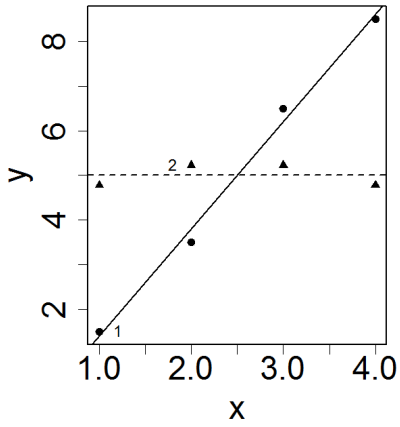
Immediately after the student completes the exam scan it to a pdf file and have student upload to Webassign.

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? (8 points)



   (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.

   Suppose we are referring to a simple linear regression model, a straight line fit to a single scatterplot. One of the assumptions of the model is that the points have the same amount of spread about the regression line at different values of the x variable, the predictor variable. If we create a scatterplot and notice that points are not evenly spread about the line for all values of x, it is possible that a variance stabilizing transformation (the goal of which being to stabilize this variance, or spread of the points about the line) can be used to make those spreads the same across all levels of x.

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.

   We want to weight the model so that the variance of the errors is constant. We fit the model $\sqrt{w_i}y_i = \beta_0\sqrt{w_i} + \beta_1\sqrt{w_i}x_i + \sqrt{w_i}e_i$, so that $\text{Var}(\sqrt{w_i}e_i)$ is constant. We see:

$$\text{Var}(\sqrt{w_i}e_i) = w_i\text{Var}(e_i) = w_i x_i^2\sigma^2$$

   So we can see if we set $w_i = 1/x_i^2$, the variance of our new weighted error term $w_i e_i$ will have constant variance.
   There are two ways to approach the problem from this point; both give the same result. First, we could re-define our vector $\mathbf{Y}$ and our matrix $\mathbf{X}$ to include the weights:

$$\mathbf{Y}_{new} = \begin{bmatrix} y_1/x_1 \\ y_2/x_2 \\ \vdots \\ y_n/x_n \end{bmatrix}, \qquad \mathbf{X}_{new} = \begin{bmatrix} 1/x_1 & 1 \\ 1/x_2 & 1 \\ \vdots & \vdots \\ 1/x_n & 1 \end{bmatrix}$$

   Then $\hat{\boldsymbol{\beta}} = (\mathbf{X}'_{\text{new}}\mathbf{X}_{\text{new}})^{-1}\mathbf{X}'_{\text{new}}\mathbf{Y}_{\text{new}}$.

   Alternatively, we could define a matrix of weights along with the usual vector $\mathbf{Y}$ and design matrix $\mathbf{X}$:

$$\mathbf{W} = \begin{bmatrix} x_1^{-2} & 0 & \ldots & 0 \\ 0 & x_2^{-2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & x_n^{-2} \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

   Then $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$.

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).
No, the straight line model (1) does not seem to fit the observed data well. First, it appears that the mean function relating Market Value to Square Feet is not a straight line. Second, it appears from the plot of the square root of the absolute value of the residuals that the residuals do not have constant variance, also indicating that our model is not valid. Finally, we have some bad leverage points indicated on the Cook's Distance plot and on the plot of Standardized Residuals by Leverage that have too much influence on where the regression line is located.
(I'm less concerned about normality for such a large sample size unless we want to fit prediction intervals. The CLT says that $\hat{\beta}_1$ is approximately normally distributed for such a large sample size, even though the residuals are not normally distributed.)

(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 interval would be too long at that point. The prediction interval should capture approximately 95% of the Market Values of houses at any given Square Feet, but we know that the variability of Market Value is much smaller than average at Square Feet = 1000. A prediction interval using the above model assumes constant variance in Market Value across all values of Square Feet, which is not what is actually happening. (Main idea: non-constant variance.)

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.

If the square footage of a house increases by 1%, our model predicts that its market value will increase by 1.25%. (The percentage interpretation comes from the fact that we're doing a log transformation. Be sure to say something like, "is predicted to" or "on average" or "according to our model" to emphasize that there is variability in actual data.)

(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.

Yes, it is a vast improvement. First, the relationship between log(Square Footage) and log(Market Value) appears to be linear, unlike in model (1). Second, the residuals appear to have much more constant variance. Third, there are fewer points that appear to be bad leverage points; points 1 and 2 should still be investigated, however.
(The model still isn't perfect, though. The plot of the residuals vs. the fitted values seems to indicate some curvature, rather than a simple shotgun shape. In further investigating points 1 and 2, I found that they were mobile homes on large plots of land. I should definitely include land size as another predictor in my model, and possibly also use an indicator for whether a home is a mobile home.)

(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.

Remember to use the correction factor with your back-transformation: $(exp(12.032 + 0.1/2), exp(12.038 + 0.1/2) = (\$176,663, \$177,726)$.
I am 95% confident that the **average** market value for homes with 2500 square feet is between $176,663 and $177,726.

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.)

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{X}\boldsymbol{\beta} + \mathbf{e}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{0}) \\ &= \boldsymbol{\beta} \end{aligned}$$

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

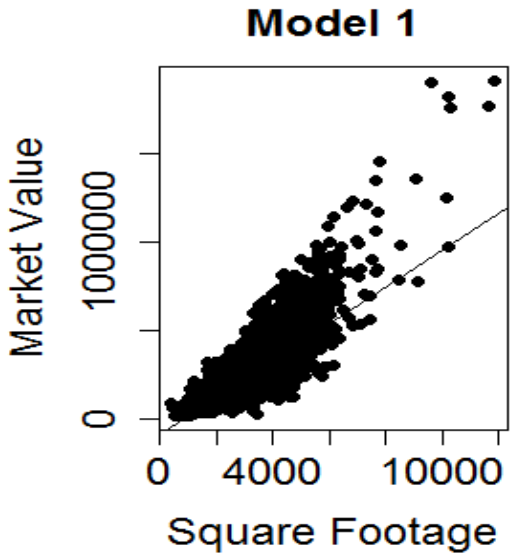
$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \text{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{Y}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{e}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\text{Variance of a constant} = 0) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\text{Independent, constant variance errors} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

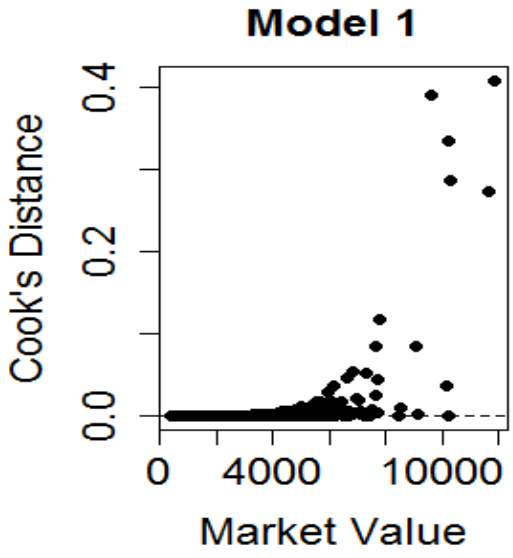
   (c) Show that $\text{Var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$, where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

$$\begin{aligned} \text{Var}(\hat{\mathbf{Y}}) &= \text{Var}(\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{X}\text{Var}(\hat{\boldsymbol{\beta}})\mathbf{X}' \\ &= \mathbf{X}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \sigma^2\mathbf{H} \end{aligned}$$

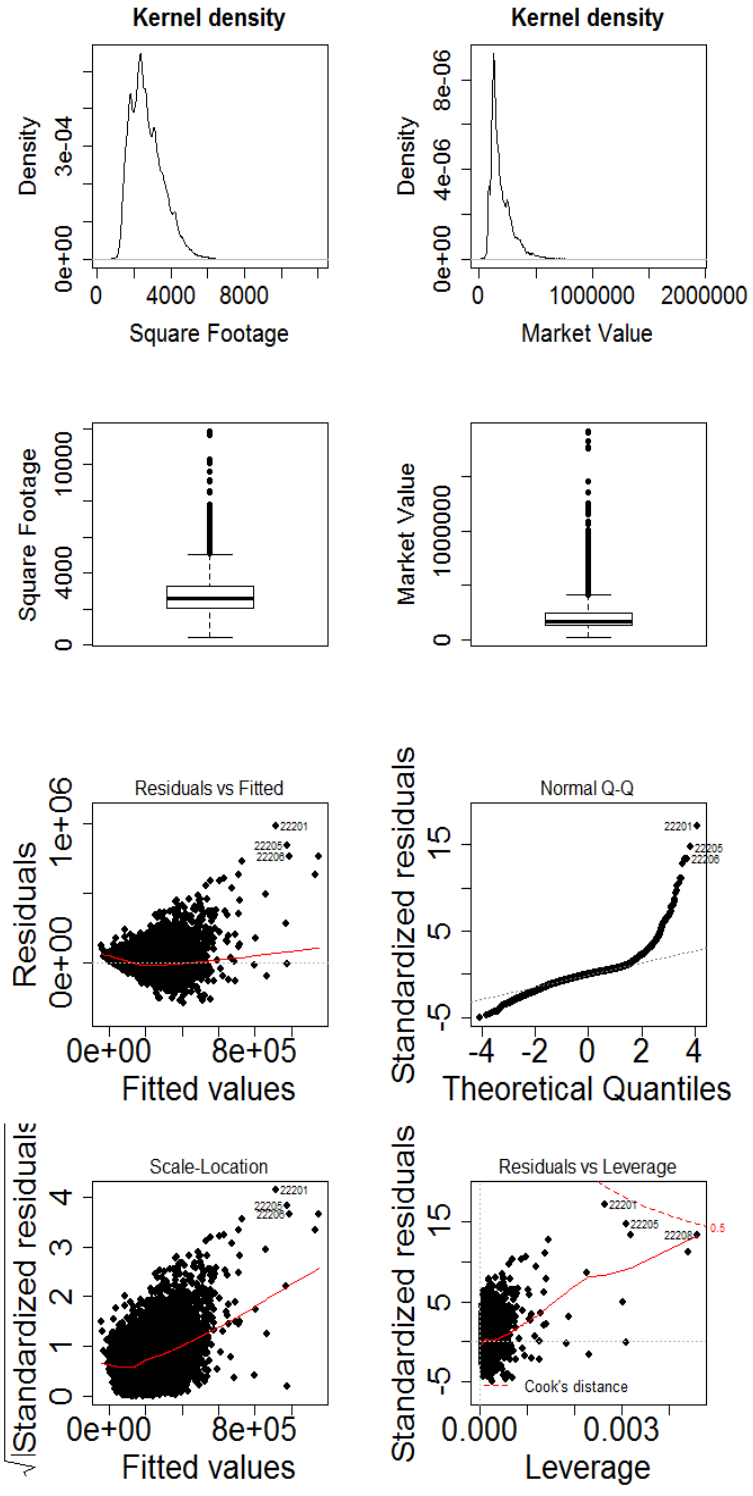
Model 1:

```
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
```



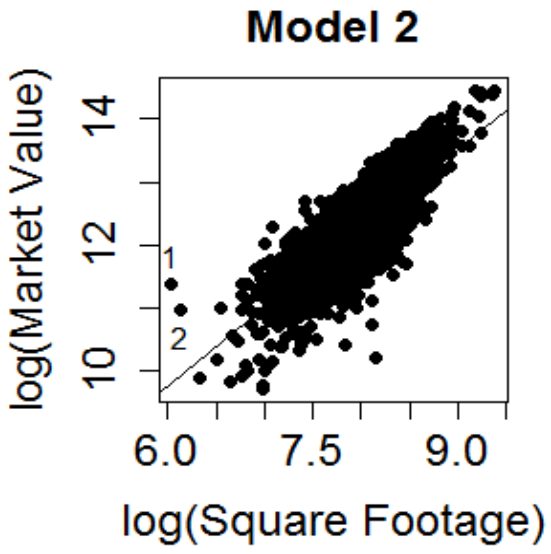


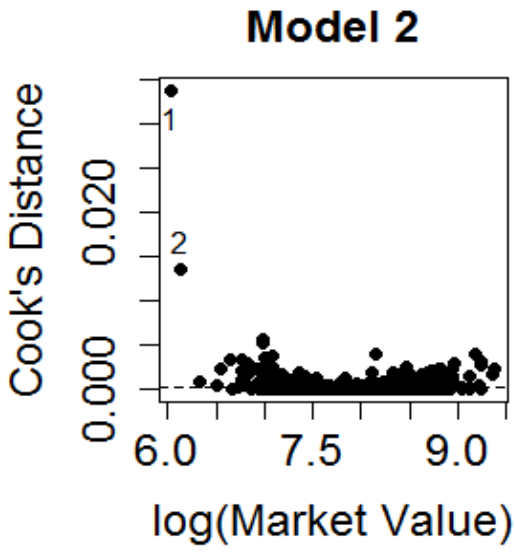
Model 1:


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

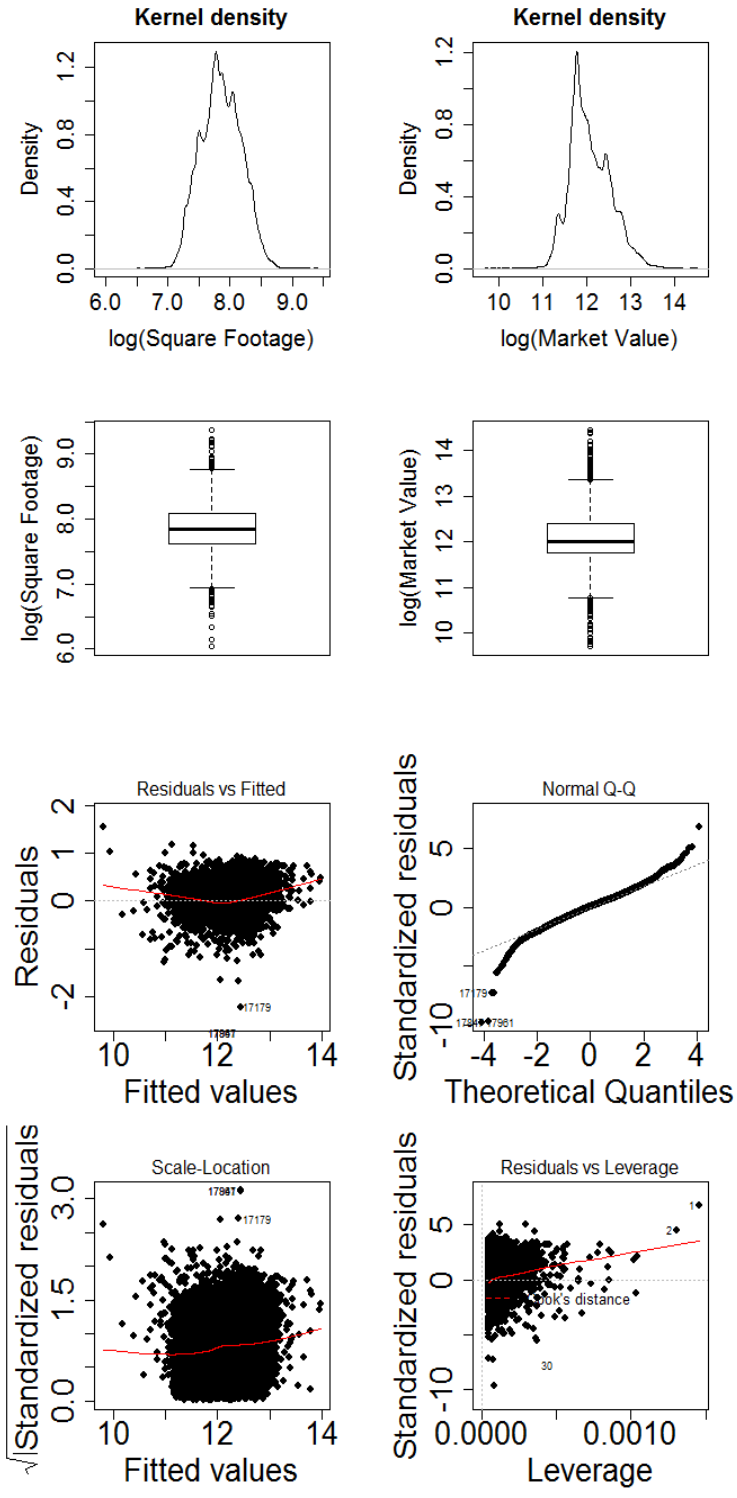
Model 2:

```
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
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage