STA 321 Professor Gabrosek Fall 2020 Homework #2

Directions: This homework is the culminating experience for Learning Module 2: Multiple Linear Regression. This homework is to be completed individually. While you may discuss problems with each other, the work you turn in must be your own. You will turn in this cover page with your assignment. Note: Any problem that begins CH Problem is a problem from the textbook.

By signing your name below, you verify that the work done on this homework is your own.

Print Name _Monica Klosin___

Software Investigation

In this software investigation, we will use a data set for a random sample of 299 home sales in Ames, Iowa from 2006 to 2010. The population consists of 2930 home sales during that time frame. We will look at several models to predict the sales price of a home.

The variables we will investigate are:

- X1 = PID: Parcel identification number
- X9 = Total Bsmt SF: Total square feet of basement area
- X12 = Gr Liv Area: Above grade (ground) living area square feet
- X15 = Full Bath: Full bathrooms above grade
- X16 = Half Bath: Half baths above grade
- X18 = TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- X20 = Garage Area: Size of garage in square feet
- Y = SalePrice: Sale price $$

1. Find numerical summaries of the response variable Y = SalePrice. **See LM 1 – Basic Numerical Summaries 321 R How-to.**

   (a) Use the snipping tool to copy the output produced and paste it here.

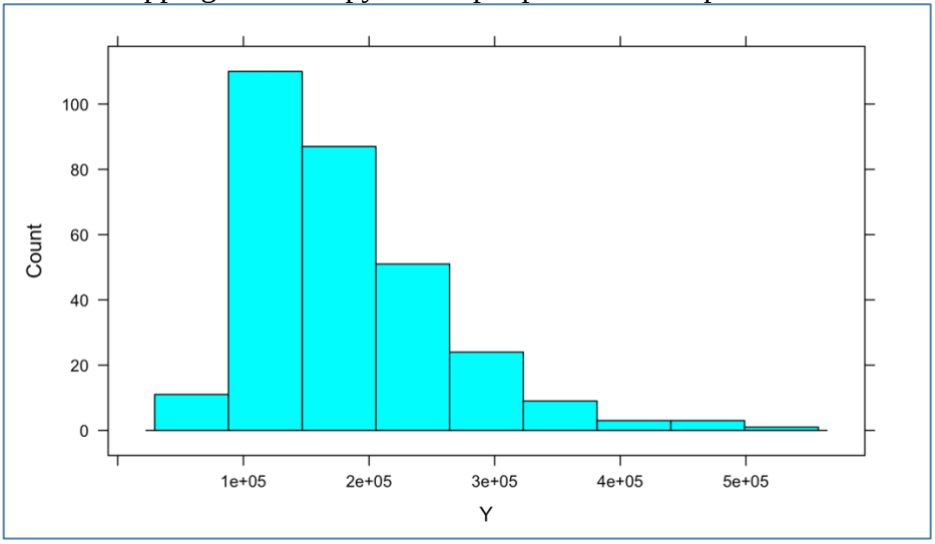
min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
68400	131875	160000	222500	538000	183806	76879.28	299	0

   (b) Interpret the median.

The median sale price of houses was $160,000, this means that out of 299 sales of homes between 2006 and 2010, half of the homes sold under $160,000 and half of the homes sold above $160,000.

2. Make a histogram of the response variable Y = SalePrice. Make logical choices for the start, end, and jump values. **See LM 1 – Histogram 321 R How-to.**

   (a) Use the snipping tool to copy the output produced and paste it here.



   (b) Describe the distribution of the response variable SalePrice. Mention peaks and shape.

The histogram is bell shaped skewed right with peak near 1e+05 to 1e+06. There are no obvious outliers.

Your answer to Question 2(b) suggests that we should transform the response variable. We will do that. In the R program I have written the code to create the variable logY = log(Y). For the remainder of this homework use logY as the response variable.

3. Using proper notation write out the model that includes all of the X variables listed on this homework except X1. For the regression coefficients, use the same subscript numbers as the X variables.

$$\hat{Y} = \beta_0 + \beta_9 X_9 + \beta_{12} X_{12} + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{18} X_{18} + \beta_{20} X_{20} + \varepsilon$$

4. Fit the regression model saving the model output to an R Object named FullModel. **See LM 2 – Fitting the MLR Model on 321 R How-to.**

   (a) Use the summary function to get the Coefficients table. Use the snipping tool to copy the all output produced below Coefficients and paste it here.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.104e+01  4.838e-02 228.186  < 2e-16 ***
X9           2.734e-04  2.931e-05   9.327  < 2e-16 ***
X12          3.895e-04  4.310e-05   9.035  < 2e-16 ***
X15          7.650e-02  2.451e-02   3.121  0.00198 **
X16          8.113e-02  2.400e-02   3.381  0.00082 ***
X18         -5.056e-02  1.077e-02  -4.694 4.13e-06 ***
X20          6.683e-04  5.699e-05  11.727  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1698 on 292 degrees of freedom
Multiple R-squared:  0.8033,    Adjusted R-squared:  0.7993
F-statistic: 198.8 on 6 and 292 DF,  p-value: < 2.2e-16
```

   (b) Write out the regression equation.

$\hat{Y}$ = 1.104e+01 + 2.734e-04X_9 + 3.895e-04X_{12} + 7.650e-02X_{15} + 8.113e-02X_{16} - 5.056e-02X_{18} + 6.683e-04X_{20}

   (c) Interpret the regression coefficient for X12.

After adjusting for the other variables in the model, we estimate that the sale prices increases by 3.895e-04 when the living area increases by 1 square foot in a house in Ames, IA.

   (d) Note that the regression coefficient for X18 is negative. Explain why this does not mean that you can state that in a SLR model with Y against X18 that the regression coefficient will be negative.

It is not guaranteed that in a SLR model with Y against X18 that the regression coefficient will be negative. Currently in the Full Model, B_{18} is negative due to the other regression. In a SLR model with Y against X18, there is only one predictor variable – meaning it could have a positive regression coefficient since there are no other predictor variables that can skew its relation to the response variable.

(e) Predict the Sale Price for a home with the following: X9 = 1000 sq. ft., X12 = 1750 sq. ft., X15 = 2, X16 = 1, X18 = 7, and X20 = 500 sq. ft.

$\log\hat{Y}$ = 1.104e+01 + 2.734e-04(1000) + 3.895e-04(1750) + 7.650e-02(2) + 8.113e-02(1)- 5.056e-02(7) + 6.683e-04(500)

$$log\hat{Y} = \$12.209385$$
$$\hat{Y} = \$200{,}663.57$$

We expect the sale price to be $200,663.57 for a house with the above specs in Ames, IA.

(f) Suppose the house actually sold for $197,500. Find AND interpret the value of the residual.

residual = (actual sale price) – (predicted sale price)
residual = 197500 – 200,663.57
residual = -$3,163.57

The sale price of the house is $3,163.57 less than expected.

(g) Interpret the value of R^2.

R^2 = 0.8033. R^2 indicates that 80.33% of the totally variability in the response variable (sale price) is accounted for by the predictor variables (house specs; Parcel ID, Total Bsmt SF, Gr Liv Area, Full Bath, Half Bath, TotRmsAbvGrd, Garage Area). The value of R^2 indicates a strong learner relationship between the sale price and the house specs.

(h) Use the Anova function to get the anova table for the FullModel. Use the snipping tool to copy the output and paste it here.

```
Analysis of Variance Table

Response: logY
           Df  Sum Sq Mean Sq F value    Pr(>F)
X9          1 16.2976 16.2976 565.479 < 2.2e-16 ***
X12         1 11.2292 11.2292 389.622 < 2.2e-16 ***
X15         1  0.5635  0.5635  19.551 1.384e-05 ***
X16         1  1.1138  1.1138  38.645 1.752e-09 ***
X18         1  1.2094  1.2094  41.964 3.939e-10 ***
X20         1  3.9632  3.9632 137.511 < 2.2e-16 ***
Residuals 292  8.4157  0.0288
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(i) Show how the adjusted R^2 value from Question 4(a) was calculated.

SSE = 8.4157
1 – SSE/SST = .8033 ->
0.1967 = SSE/SST ->
0.1967 (SST) = SSE ->
0.1967 (SST) = 8.4157 ->
SST = 42.79

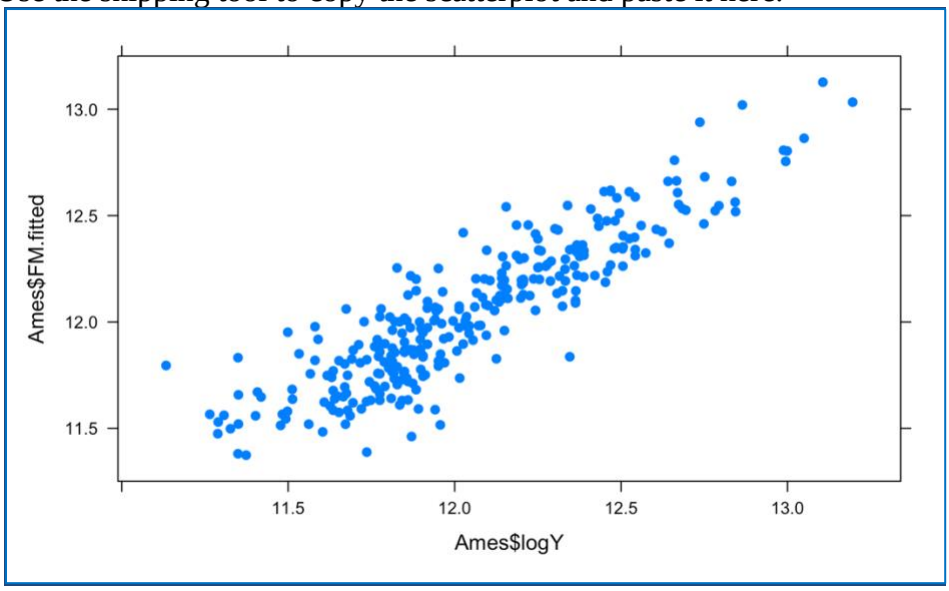
n = 299
p = 6

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{8.4157/(299-6-1)}{42.79/(299-1)} = 1 - \frac{.02882089041}{.014357195743} = 1 - .20072 =$$

0.79925

5. Make a scatterplot of logY on the vertical axis and Y-hat (which is the predicted log sale price) on the x axis. I have written the code to create a variable named FM.fitted that has the y-hat values from the FullModel. The variable is added on to the Ames dataframe. **See LM 1 – Scatterplot on 321 R How-to.**

   (a) Use the snipping tool to copy the scatterplot and paste it here.



   (b) Describe what you see in the scatterplot.

   We see a positive moderately strong fit for the model.

   (c) Find the value of the linear correlation between the actual logY values and FM.fitted. **See LM 1 – Linear Correlation on 321 R How-to.**

```
# linear correlation between the actual logY values and FM.fitted
cor(logY~FM.fitted, data = Ames)
```

```
[1] 0.8962904
```

   The linear correlation is 0.8962904, which confirms 4d on what we saw that there is a moderately strong fit in the model.

(d) In MLR what do we call the correlation between actual and predicted values of the response? Notice that if you square this value you get R^2 just like in SLR where $(r)^2 = R^2$.

That is the mutiple correlation coefficient.

6. Consider the full model stated in Question 3 and fit in Question 4.

(a) Complete the Overall F-Test. Your solution should include the null and alternative hypotheses (notation only is ok), the value of the test statistic with appropriate degrees of freedom, the p-value, decision, and your conclusion. **Note: You can simply use the output generated in Question 4(a) to answer this.**

H_0: $\beta_0 = \beta_9 = \beta_{12} = \beta_{15} = \beta_{16} = \beta_{18} = \beta_{20} = 0$
H_a: At least one B_x != 0
F-statistic = 198.8; **df** = 292
p-value = 2.2e-16
decision: We reject H_0
conclusion: At least one predictor variable is useful in predicting Y.

(b) Complete the single variable T-Test on X_{20} = garage area . Your solution should include the null and alternative hypotheses (notation only is ok), the value of the test statistic with appropriate degrees of freedom, the p-value, decision, and your conclusion. **Note: You can simply use the output generated in Question 4(a) to answer this.**

H_0: $\beta_{20} = 0$ given the X's in the model
H_a: $\beta_{20} != 0$ given the X's in the model
t = 11.727 ; **df** = 292
p-value: < 0.001 or <2e-16
decision: Reject H_0
conclusion: After adjusting for the other predictor variables in the model, Garage Area is a statistically significant predictor of Sale Price of homes in Ames, IA.

7. Partial F-Test for X_{15}, X_{16}, and X_{18} in a model that has X_9, X_{12}, and X_{20}. Fit the regression model that has only X_9, X_{12}, and X_{20}. Save the model output to an R Object named ReducedModel. Use the anova function to get the SSE for the reduced model**.** **See LM 2 – Fitting the MLR Model on 321 R How-to.**

   (a) Use the snipping tool to cut and paste the anova table for the reduced model.

```
Analysis of Variance Table

Response: logY
           Df  Sum Sq Mean Sq F value    Pr(>F)
X9          1 16.2976 16.2976  504.41 < 2.2e-16 ***
X12         1 11.2292 11.2292  347.55 < 2.2e-16 ***
X20         1  5.7341  5.7341  177.47 < 2.2e-16 ***
Residuals 295  9.5314  0.0323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   (b) Give values for the following: SSE(FM), SSE(RM), p, n, and k. Recall that the anova output for the full model is in Question 4(h).

SSE(FM) = 8.4157
SSE(RM) = 9.5315
p = 229
n = 3
k = 4

   (c) Complete the Partial F-Test for X_{15}, X_{16}, and X_{18} in a model that has X_9, X_{12}, and X_{20}. Your solution should include the null and alternative hypotheses (notation only is ok), the value of the test statistic with appropriate degrees of freedom, the p-value, decision, and your conclusion. **See LM 2 – F Statistic P-Value on 321 R How-to.**

H_0: $\beta_{15} = \beta_{16} = \beta_{18} = 0$ | X_9,X_{12}, X_{18} in model
H_a: $\beta_{15}, \beta_{16}, \beta_{18}$!= 0 | X_9,X_{12}, X_{18} in model
F-statistic = 343.1 ; **df** = 295
p-value: 0
decision: Reject H_0
conclusion: At least one of the predictor variables, X_{15}, X_{16}, and X_{18}, adds useful information to the prediction of Y in the model that includes the 3 other predictors. The full model is better than the reduced model for predicting Y.