STA 321 Professor Gabrosek Fall 2020 Homework #3

Directions: This homework is the culminating experience for Learning Module 3: Regression Diagnostics. 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_______________

Submitting Your Assignment: The following steps describe how to access the homework assignment and then submit it via BB.

1. Access the Word file homework assignment saved to BB. If you are reading this, then you have done Step 1 successfully!

2. Complete your homework. You may write your answers by hand or you may type your answers into this document. Whichever you choose, be sure to do the following:

    - Write in dark ink or dark pencil if you are writing answers by hand.
    - Show your work on any required calculations.
    - Use the snipping tool to embed any R output you are asked to produce directly into your homework paper. If you choose to write out answers by hand, then print the necessary R output and place it in the appropriate spot on the homework. Please do not place all R output at the end.

3. Scan your homework into a single PDF. I do not want separate PDFs for each page! The Math and Stats (Tutoring) Center has two videos that show how to use a smartphone to do this:

    iPhone – A short video named, “Scan With iPhone” that describes how to scan your homework papers with an iPhone is saved to BB under the Technology Menu item.

    Android – A short video named, “Scan With Android” that describes how to scan your homework papers with an Android is saved to BB under the Technology Menu item.

4. Upload your PDF to BB – A short video named “Uploading PDF to BB” that describes how to upload your homework PDF is saved to BB under the Technology Menu item.

Scoring:

- Any student who chooses to turn in the assignment on or before the due date given in class will be eligible to score 100 points on the assignment.
- Any student who chooses to turn in the assignment by midnight on Tuesday will be eligible to score 80 points on the assignment.

- Any student who chooses to turn in the assignment after midnight on Tuesday will have the assignment reviewed and commented on but the assignment will not be eligible to score any points.

There are six pages to this homework assignment including the cover page.

Software Investigation

Software investigation questions require you to use the statistical software R to analyze a data set and fit regression models.

- Begin by accessing the Virtual Lab Environment. See the document named, "Accessing the Virtual Lab" that is saved to BB – Technology.
- Have available the document named, "STA 321 R How-To Sheet" that is saved to BB – Technology.
- Once you have logged into the Virtual Lab environment, follow the instructions under Utilities – Opening R to open RStudio.
- Now follow the instructions under Utilities – Opening R File to open the R program named **hw3starter.rmd** that will read the dataset named **usedcars.csv** into an R data frame named **UsedCars**. You will add code to this program and produce output. You will want to have the snipping tool ready so that you can snip R output that you produce and paste it into your Word file for HW 3.

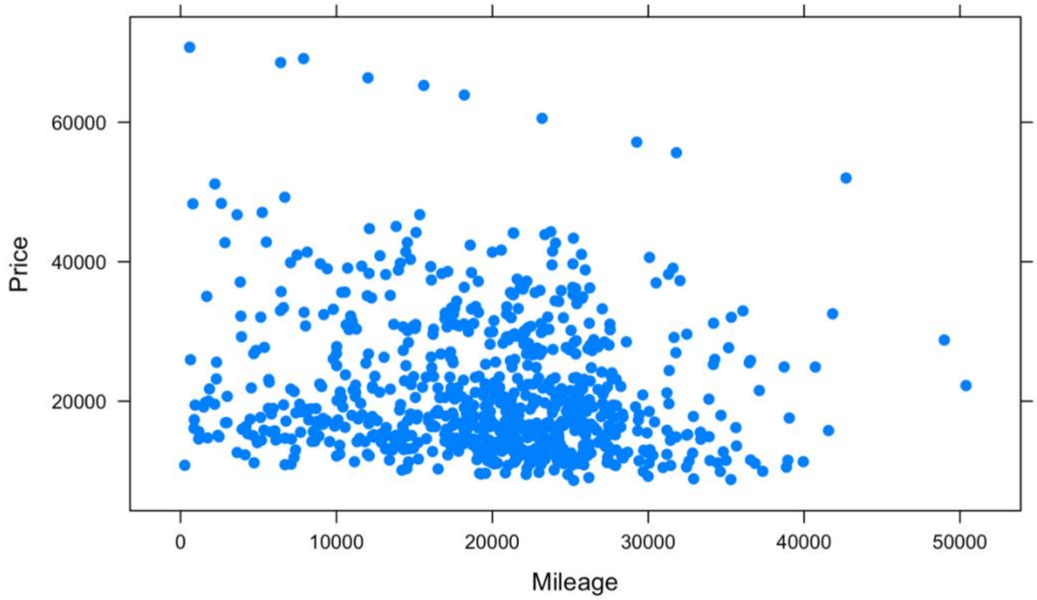
In this software investigation, we will use a data set collected from Kelly Blue Book to assess the value of 804 used 2005 GM cars. At the time of data collection, all cars were less than one year old when priced and considered to be in excellent condition.

The variables in the dataset are:

- Price: suggested retail price of the used 2005 GM car in excellent condition.
- Mileage: number of miles the car has been driven
- Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet
- Model: specific models for each car manufacturer such as Ion, Vibe, Cavalier
- Trim: specific type of car model such as SE Sedan 4D, Quad Coupe 2D
- Type: body type such as sedan, coupe, etc.
- Cylinder: number of cylinders in the engine
- Liter: a more specific measure of engine size
- Doors: number of doors
- Cruise: indicator variable representing whether the car has cruise control (1 = cruise)
- Sound: indicator variable representing whether the car has upgraded speakers (1 = upgraded)
- Leather: indicator variable representing whether the car has leather seats (1 = leather)

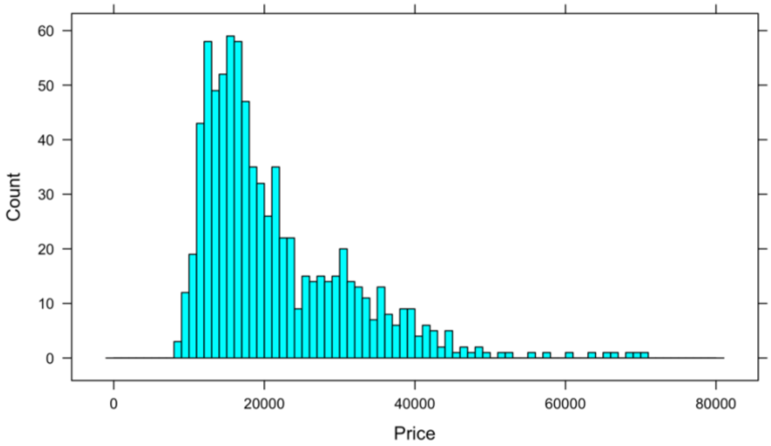
We will focus our attention on Y = Price and X = Mileage.

1. Let's begin with a scatterplot of Y = Price and X = Mileage. Use the snipping tool to copy the output produced and paste it here. **See LM 1 – Scatterplot 321 R How-to.**



2. Now some univariate EDA of the response variable Y = Price. Make a histogram of the response variable Price. Use start = 0, end = 80000, and jump = 1000. **See LM 1 – Histogram 321 R How-to.**

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



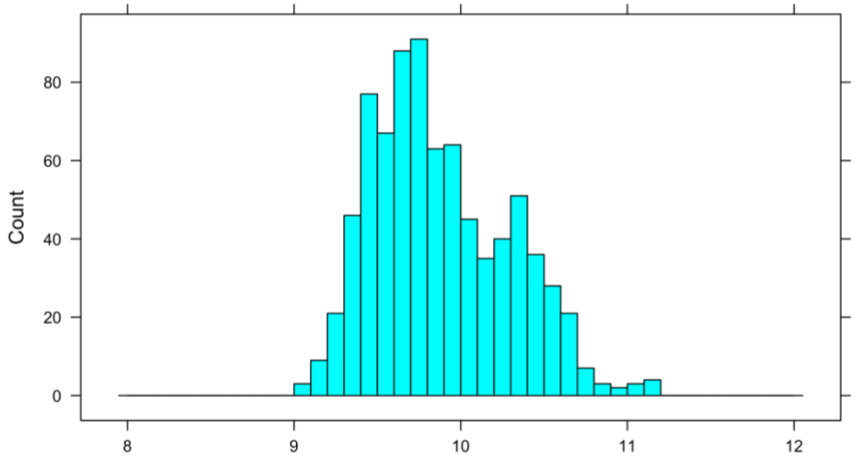
   (b) Describe the distribution of the response variable Price. Mention peaks and shape.

   The histogram is unimodal and bell-shaped skewed right with a peak near slightly less than $20,000. There are a few outliers skewing around $70,000, causing the data to be skewed more right.

Your answer to Question 2(b) suggests that we should transform the response variable. We will do that. In the R program under "Transform Y" write the code to create a variable named logPrice that is the logarithm of the Price.

3. Make a histogram of the response variable logPrice. Use start = 8, end = 12, and jump = 0.1. **See LM 1 – Histogram 321 R How-to.**

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



   (c) Describe the distribution of the response variable logPrice. Mention peaks and shape.

   The histogram is bimodal and bell shaped slightly skewed right. Peak around 9.7, and a slightly smaller peak is seen around 10.3.

Taking the log helps "normalize" the response. Is it perfect – No, but we are going to use logPrice as our response in the rest of the homework. We will investigate the model below.

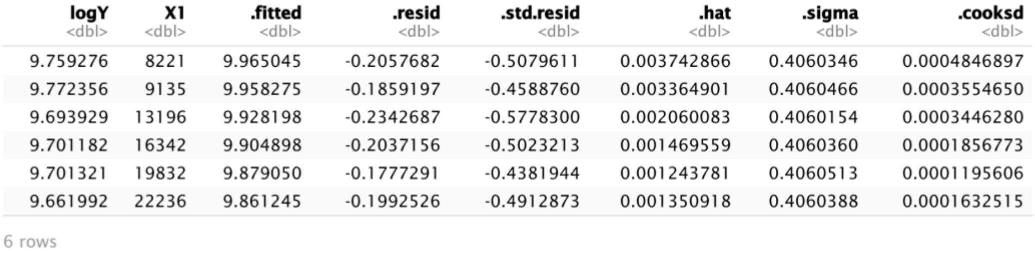
Model 1: logY = logPrice, X1 = Mileage

4. Using proper notation write out Model 1. Be sure to indicate that the response is logY.

$$logY = \beta_0 + \beta_1 X_1 + \varepsilon$$

5. Fit Model 1 saving the model output to an R Object named Model1. **See LM 1 – Linear Regression on 321 R How-to.** Then, save the residuals to an RObject named Model1_metrics. **See LM 2 – Saving Residuals on 321 R How-to.**

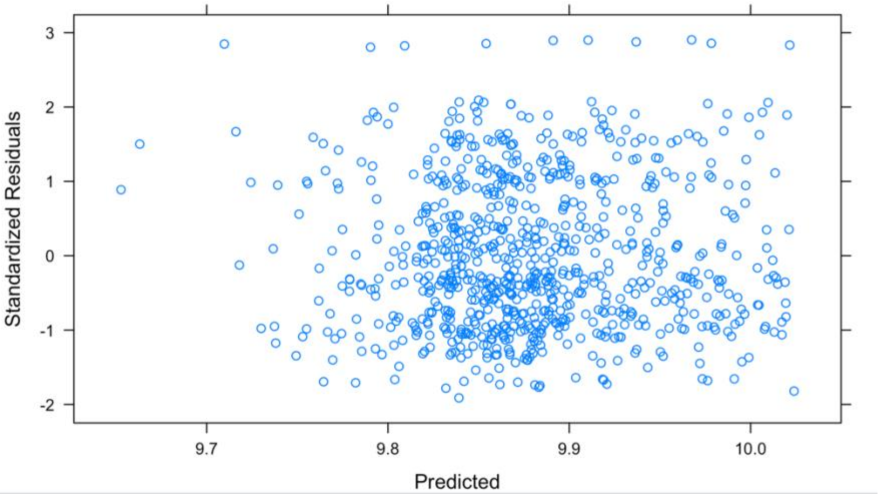
   (a) Use the following code to print the first six rows of residuals: head(Model1.metrics). Use the snipping tool to copy and paste the output here.



logY	X1	.fitted	.resid	.std.resid	.hat	.sigma	.cooksd
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
9.759276	8221	9.965045	-0.2057682	-0.5079611	0.003742866	0.4060346	0.0004846897
9.772356	9135	9.958275	-0.1859197	-0.4588760	0.003364901	0.4060466	0.0003554650
9.693929	13196	9.928198	-0.2342687	-0.5778300	0.002060083	0.4060154	0.0003446280
9.701182	16342	9.904898	-0.2037156	-0.5023213	0.001469559	0.4060360	0.0001856773
9.701321	19832	9.879050	-0.1777291	-0.4381944	0.001243781	0.4060513	0.0001195606
9.661992	22236	9.861245	-0.1992526	-0.4912873	0.001350918	0.4060388	0.0001632515

6 rows

   (b) Make a plot of the standardized residuals vs y-hat. Use snipping tool to cut and paste plot. **See LM 3 – Standardized Residuals -Y-Hat Plot on 321 R How-to**.



   (c) The residual plot in part (b) can be used to check Assumption 1 Form of Model – Linearity. Does this plot suggest that the assumption is met? Answer Yes/No and give a brief explanation as to why.

   No. There is no visible linear relationship between the standardized residuals and the predicted values. This is because we don't see any type of linear trend amongst the variables.

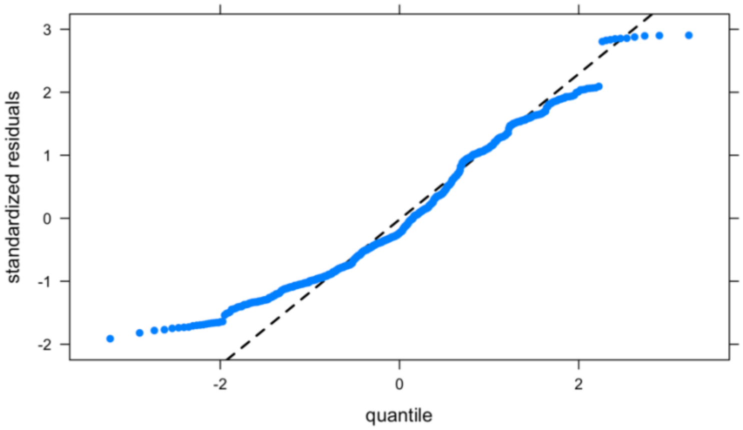
(d) The residual plot in part (b) can be used to check Assumption 3 Errors – Mean Zero. Does this plot suggest that the assumption is met? Answer Yes/No and give a brief explanation as to why.

Assumption 3 (mean zero) states that the errors ε_i have a mean of 0 for all values of the X variables. Based on the plot, this assumption is met, because there are some errors > 0 and some errors < 0 , the average error events to 0. There is no concentrated data points in any region of the y-axis.

(e) The residual plot in part (b) can be used to check Assumption 4 Errors – Constant Variance. Does this plot suggest that the assumption is met? Answer Yes/No and give a brief explanation as to why.

Assumption 4 (constant variance) states that the errors ε_i have the same unknown variance σ^2 for all values of the X variables. Based on the residual plot in part be, this assumption is met, the data isn't concentrated, the spread of data remains constant throughout the graph.

(f) Make a normal probability plot of the standardized residuals. Use snipping tool to cut and paste plot. **See LM 3 – Normal Probability Plot of Residuals on 321 R How-to**.



(g) The normal probability plot in part (f) can be used to check Assumption 2 Errors – Normality. This plot suggests that the assumption is violated. Write a sentence that tells me what the plot tells you about the residuals. I want more than "They aren't normal." Tell me what they are

The closer the blue line is to the dotted black line the more our residuals are fitting a normal distribution. Assumption 2 checks Normality, seeing if the errors ε_i follow a normal distribution. . Based on the plot, Assumption 2 Normality is not met because the residuals are heavy tailed. The blue line is pretty far from the dotted black line at very low (< -1) and very high (> 1.5) quantiles. This means residuals from our model at high quantiles are not fitting the normal dist. well. Residuals on the far left (-3, -2) part of the graph tell us that we have heavy tails.

There is no time component to the data collection so checking Assumption 5 Errors – Independence is not applicable. We can make the argument that the error in predicting the price of one used car should have no effect on the error of predicting the price of any other used car. That's what the assumption says.

6. The assumptions that we investigated in Question 5 deal with the model and the errors. In Question 6 we focus on the predictors.

(a) Assumption 7 Predictors – No Measurement Error. We don't have a diagnostic to check this, but write a sentence that explains what this assumption means for this example.

The 7th regression assumption states that there is no measurement error, meaning, all the predictor variables (X1,X2,…Xn) are measured without error. This means, for this example, that we assume that the measurement of Millage is measured without error.

Since we only have one predictor in the model, Assumption 8 Predictors – Predictors Independence is not relevant.

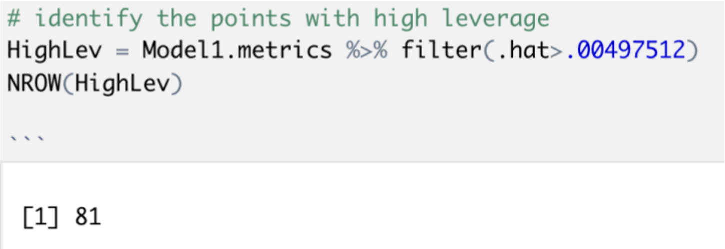
(b) Page 108 of the text states that points are considered to have high leverage if the leverage value exceeds $2(p+1)/n$. What is the cutoff value for leverage for this data set?

p = 1 ; there is one predictor
n = 804; there are 804 cars in the dataset

High Leverage = 2(1+1)/804 = 4/804 = .00497512

.00497512 is the leverage cutoff for this dataset.

(c) Write the code to identify the points with high leverage and save them to an R Object named HighLev. **See LM 3 – High Leverage on 321 R How-to**. How many of the 804 points were identified?



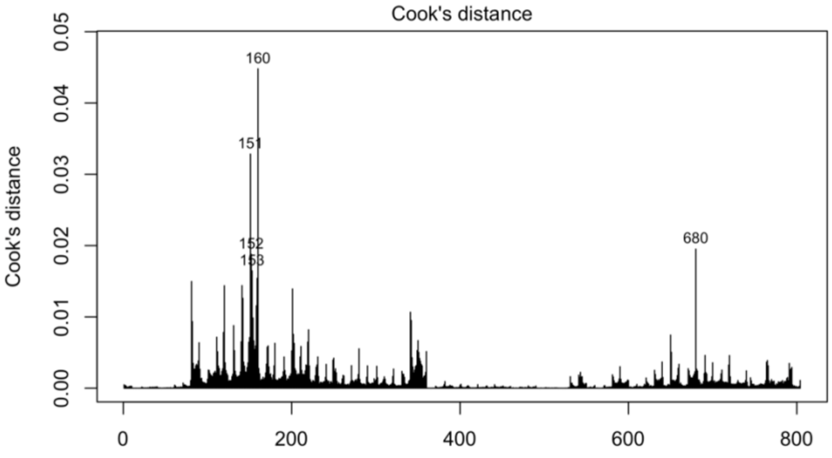
```
# identify the points with high leverage
HighLev = Model1.metrics %>% filter(.hat>.00497512)
NROW(HighLev)

```

```
[1] 81
```

81 points were defined with high leverage.

(d) Make a plot of Cooks D by observation number using the HighLev R Object. Write the code so that the observation number for the top five values are identified. Use snipping tool to cut and paste plot. **See LM 3 – Cooks D on 321 R How-to**.



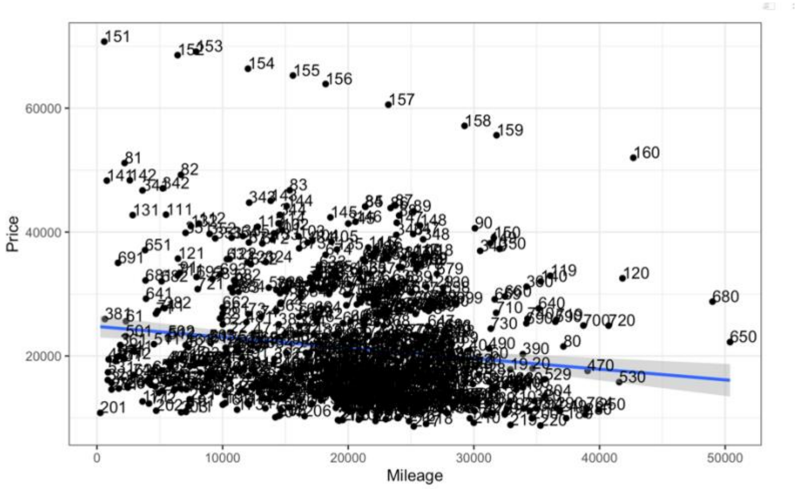
(e) Complete the table below with information from the original dataset and the Model1.metrics on the two observations with the largest Cooks D.

Observation	Price	Mileage	Standardized residual	Hat	Cooks D
160	52001.99	583	2.847434884	0.010930209	0.04480011422456
151	70755.47	42691	2.833693779	0.008112252	0.03283633891952

(f) For each point explain what makes the point have high leverage

More extreme values in millage cause a high leverage. High leverage is based on the variable that will cause the linear regression line the most change, which in this example is the variable Millage.
The observations with high leverage are the ones with either high millage or low millage, but also based on where they are in the graph based on how they can skew the linear regression line.

(g) For each point explain what makes the point have high Cooks D. The scatterplot in Question 1 will help you see these points.



I made another scatterplot to list the observation values of the raw data.

As you see, observation 151 has a very high price, but very low mileage. This makes observation 151 have a very big high leverage since if this observation was taken out of the data, it would skew the slope (the coefficient on mileage) to be less-steep.

Observation 160 has a very high mileage, and a semi-high price. There are no other observations very close to 160, causing this observation to have an impact on the linear model. If observation 160 was taken out, I assume the slope (the coefficient on mileage) would be steeper.

7. The scatterplot in Question 1 suggests that a SLR model with predictor Mileage is not going to fit the data well. Consider the other variables in the dataset and write a sentence or two that explains why a MLR model is likely to be need in this situation.

A MLR model is needed for this situation, since it is difficult to predict the price of a car one just one variable type. In the real world, a cars price is not just based on mileage, it is also based on other things like Model, Type, and Sound. There are many things that make a car its worth in price, so it is difficult to predict a cars price just on one variable (Mileage). There would be better prediction of car price if we used a MLR model, since we could factor in more variables to help predict the price of a car.

Final note on HW 3: I did not include any work with dummy variables or interaction variables on this homework. You will work with them in homework 4a.