

# 1

## Berkeley Guidance Study

The dataset is located in `BGSgirls2.txt`. It contains one line of data for each of 70 girls with the following variables:

- ID: Girl identification number
- WT2: Weight (kg) at 2 years
- HT2: Height (cm) at 2 years
- WT9: Weight (kg) at 9 years
- HT9: Height (cm) at 9 years
- LG9: Leg circumference (cm) at 9 years
- ST9: Strength (kg) at 9 years
- WT18: Weight (kg) at 18 years
- HT18: Height (cm) at 18 years
- LG18: Leg circumference (cm) at 18 years
- ST18: Strength (kg) at 18 years
- BMI: Body Mass Index at 18 years
- SOMA: Somatotype (SOMA), on a scale from 1 (very thin) to 7 (very obese)

USE SAS TO COMPLETE THE FOLLOWING EXERCISES:

### (a)

Fit a multiple regression model:

$$BMI_i = \beta_0 + \beta_1 WT2_i + \beta_2 HT2_i + \beta_3 WT9_i + \beta_4 HT9_i + \beta_5 ST9_i + \epsilon_i$$

for  $i=1, \dots, 70$ .

And use the following diagnostics to assess model assumptions. (Do not submit the output; just examine the results and briefly describe the insight provided by each).

#### i.

Normal Q-Q plot of residuals and the related Shapiro-Wilk test.

The QQ plot closely aligns with the reference line within the first theoretical quantile, but there are deviations past the first (positive/negative) quantile.

The Shapiro-Wilk test provides a small p-value ( $<0.0001$ ) such that we would have evidence to reject the null hypothesis that the residuals are normally distributed.

Overall, we have reason to suspect our normality assumption is being violated.

ii.

Plot of the residuals versus the estimates of the conditional means for BMI

We generally observe a random spread of residual values across fitted values. However, there are two negative residuals around predicted BMI 25+, such that we'd consider these points to either be candidates for removal (due to being outliers) or that we may in fact be violating our assumption of form of the model.

This notwithstanding, we generally have reason to believe our form of the model and constant variance assumptions are not being violated.

iii.

Individual plots of the residuals versus each of the five explanatory variables

Plots of residuals versus each of the five explanatory variables are generally consistent with the depictions present from the residual v. fitted values graph, insomuch as we may have a few problematic points to address but generally do not have reason to suspect our assumptions are being violated.

**(b)**

Given that an outlier should be detected from part (a), refit the model and recheck the diagnostics listed in (a) to assess whether model assumptions are violated or not. (HINT: You can filter observations from the dataset using the where statement inside the reg procedure in SAS.)

The QQ plot looks better, insomuch as it more closely tracks with the reference line, and this is consistent with a larger Shapiro-Wilk test statistics, such that we would not have evidence to reject the null hypothesis that the residuals are normally distributed. As such we have reason not to suspect the normality assumption is being violated as it was in part (a).

Furthermore, the residual plots consistently (across the x-axis of fitted values as well as individually across the explanatory variables) appear to be randomly spread, such that our constant variance and form of the model assumptions are likely not being violated either.

However, it is worth noting that we still appear to have a potential outlier in our diagnostic plots.

(c)

For the 69 observations (without the outlier that was detected from part (a)), use a backward selection procedure to search for a model using  $\alpha_{stay} = 0.05$ . For this question, just consider the five variables mentioned in part (a): WT2, HT2, WT9, HT9, ST9. For your final model, report the estimated coefficients and their standard errors.

•

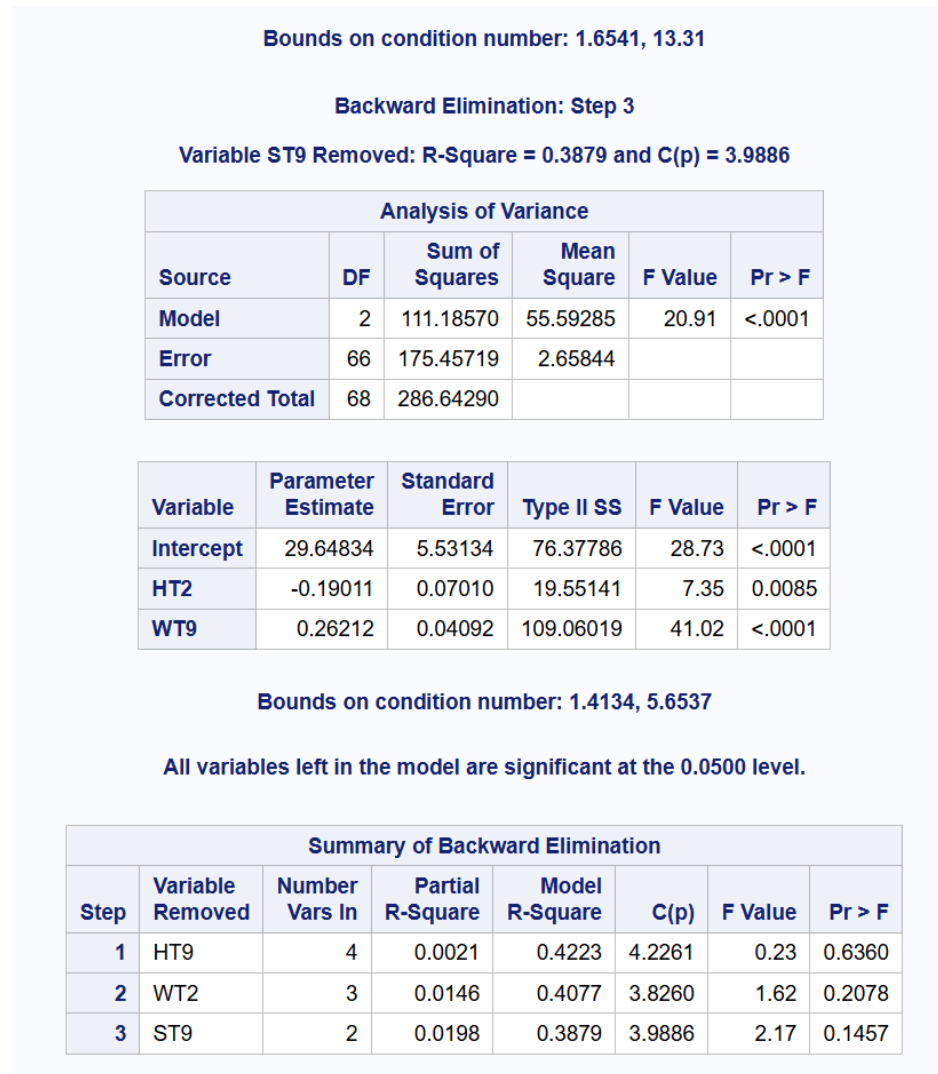


Figure 1: CocoMelon

In the final model from the method described, we have:

Variable: Intercept Coefficient: 29.648 Standard Error: 5.531

Variable: HT2 Coefficient: -0.190 Standard Error: 0.070

Variable: WT9 Coefficient: 0.262 Standard Error: 0.0409

(d)

For the 69 observations (without the outlier that was detected from part (a)), check all possible models that could be constructed using at most the five variables WT2, HT2, WT9, HT9, ST9 and then give the best one that you recommend. Justify your choice.

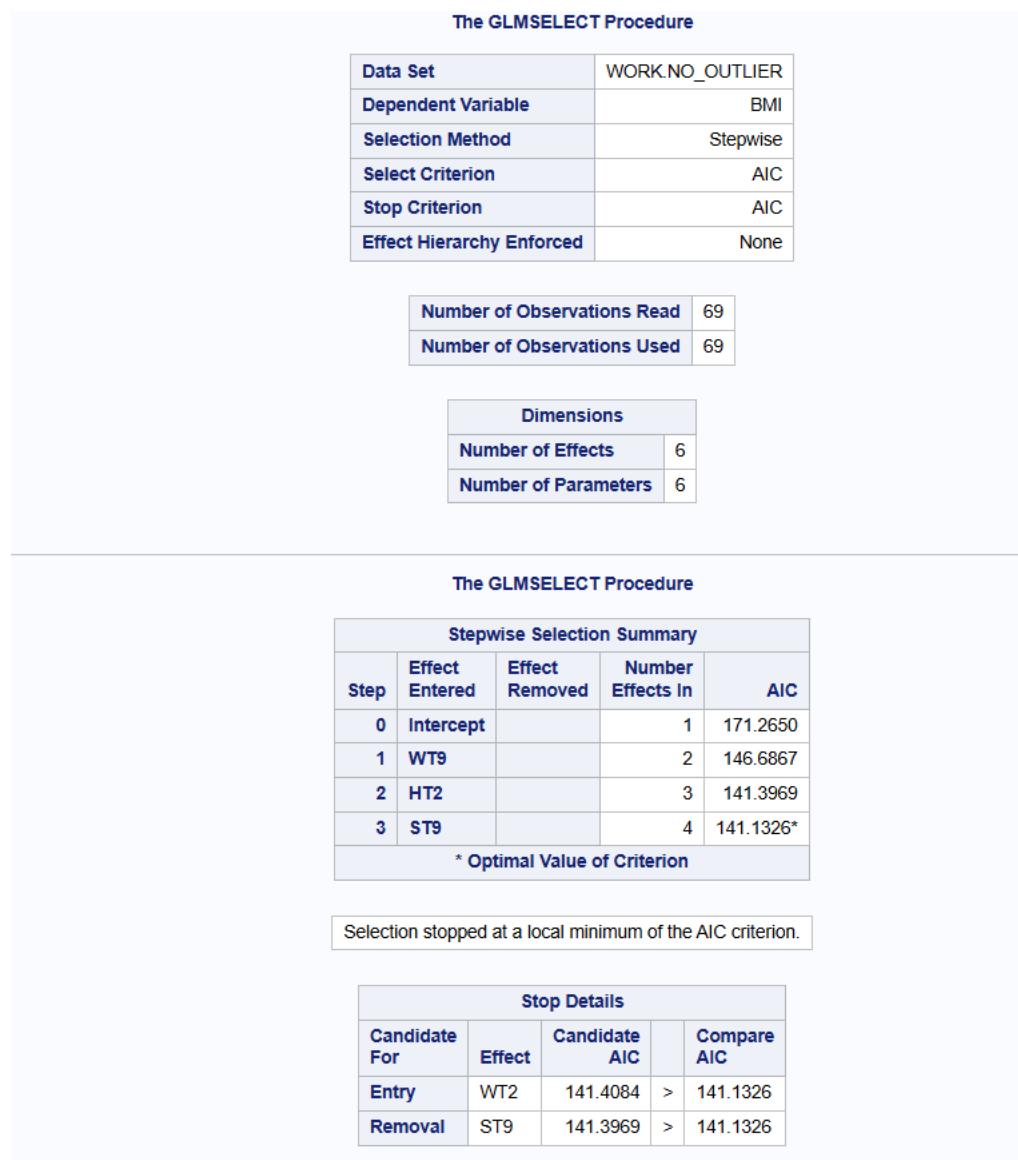


Figure 2: CocoMelon

After much back and forth, I ultimately decided on recommending the following model, which includes an intercept term: HT2, WT9, ST9 explanatory variables/predictors

Descriptively, we have some reason to believe that multicollinearity would be minimized compared to models with similar variables, i.e. that use WT2 and WT9 or HT2 and HT9. Furthermore, we want to avoid overfitting, such that we have a preference for a simpler model (Occam's), which means a preference for less explanatory variables.

Bearing that in mind, there were 3 statistics used as criteria used for evaluating models: AIC, Adjusted R-Squared, and Mallow's  $C_p$ .

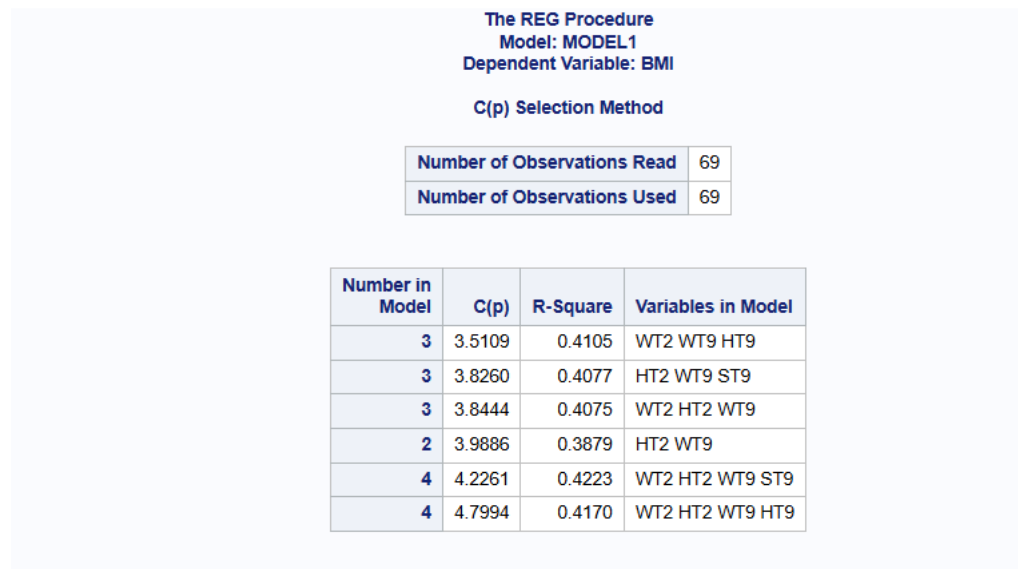


Figure 3: CocoMelon

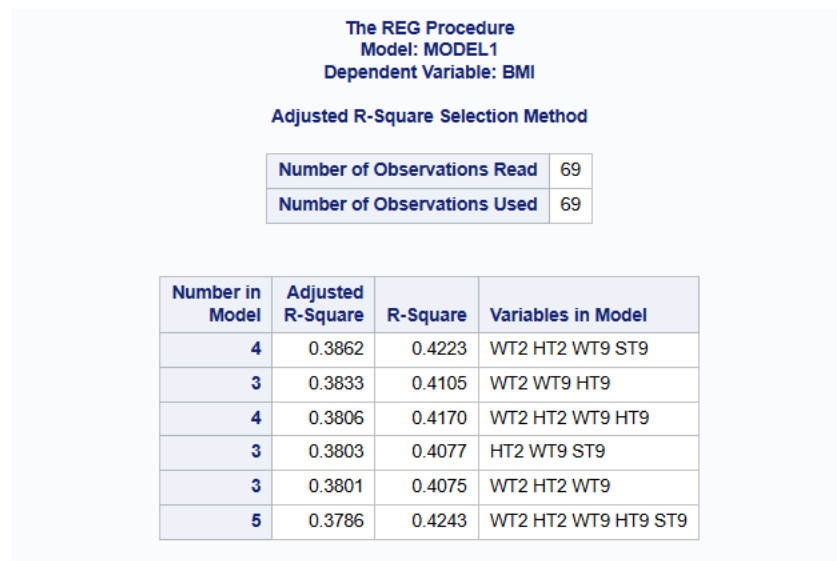


Figure 4: CocoMelon

The model I recommend has the best AIC value (minimized), and it also has a “pretty good” Adjusted R-Squared and Mallows’  $C_p$ .

For Adjusted R-Squared, it has the second best for models with 3 explanatory variables, and 4th overall across all models. Marginally, the difference in Adjusted R-Squared is in the hundredths of decimal places. Also, when comparing values of Mallows (looking at  $k+1$  closest to  $C_p$  value), we see that the model recommended has the second best  $C_p$  value, beaten only by WT2, HT2, and WT9. However, the difference between these two is less than 0.02, so the difference is marginal. And the model we recommend doesn’t “reuse” the same variable WT, such that we potentially avoid multicollinearity. To that end...

(e)

Are there concerns about multicollinearity for the explanatory variables of the model you picked in part (d)?

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	29.06585	5.49727	5.29	<.0001	0
HT2	1	-0.17691	0.07006	-2.52	0.0140	1.43695
WT9	1	0.28677	0.04389	6.53	<.0001	1.65413
ST9	1	-0.02217	0.01505	-1.47	0.1457	1.34562

Figure 5: CocoMelon

VIF values are not especially large (less than cutoff for “moderate” of value of 3 or 5), so minimal issue with multicollinearity based on VIF criteria.

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	HT2	WT9	ST9
1	3.94548	1.00000	0.00007973	0.00006454	0.00122	0.00267
2	0.03694	10.33442	0.00645	0.00390	0.00000907	0.77201
3	0.01701	15.22865	0.00855	0.00316	0.84117	0.21533
4	0.00056290	83.72067	0.98492	0.99288	0.15761	0.00999

Figure 6: CocoMelon

However, when looking at the Condition Index for the Eigenvalues, we do observe a rather high value (larger than 30), with the 83.72 corresponding to a significant proportion of variation for Intercept, HT2, and WT9, or an indication of potential extreme multicollinearity between HT2 and WT9. Given this, we look to the correlations to confirm.

Pearson Correlation Coefficients, N = 69 Prob >  r  under H0: Rho=0			
	HT2	WT9	ST9
HT2	1.00000	0.54083 <.0001	0.36096 0.0023
WT9	0.54083 <.0001	1.00000	0.49445 <.0001
ST9	0.36096 0.0023	0.49445 <.0001	1.00000

Figure 7: CocoMelon

We additionally double check our observation of potential multicollinearity by looking at the correlation between our explanatory variables. To that end: The above is further corroborated by the correlation coefficient between HT2 and WT9 being greater than 0.5 (0.54083); however, the correlation is still less than our “typical significance threshold” of  $|r| > 0.7$ , so this is a good thing (and motivation for choosing this over another hypothetical model like, say, WT2, WT9, and HT9.)

So overall, yes we do have some concerns about multicollinearity for the model we chose in part (d), though not especially extreme.

## 2

### Ames Housing (+25)

A dataset (introduced in the previous homework assignment) was collected from home sales in Ames, Iowa between 2006 and 2010. The variables collected are:

- Year Built: The year the house was built
- Basement Area (in sq. ft): The amount of area in the house below ground level
- Living Area (in sq. ft): The living area in the home (includes Basement Area)
- Total Room: The number of rooms in the house
- Garage Cars: The number of cars that can be placed in the garage
- Year Sold: The year the home was sold
- Sale Price: The sale price of the home (the response variable)
- Garage Size: S = Small (Garage Cars = 0,1) or L = Large (Garage Cars = 2+)
- Age (in yrs.): Age of house = Year Sold - Year Built

Use SAS to complete the following exercises:

The data from 999 sales can be found in the file housing train.csv and for the remaining 1,924 sales in the file housing eval.csv in our course's shared folder in SAS Studio. You will determine a final multiple linear regression model for predicting sale price from the explanatory variables: Basement Area, Living Area, Total Room, Garage Size, and Age.

### (a)

Fit the full model using all 5 explanatory variables listed above to the training data (housing train.csv).

And so it was fit.

#### i.

Find and interpret the  $R^2$  value for the full model.

77.12% of variability in Sales price can be explained using the multiple linear regression using Basement Area, Living Area, Total Room, Garage Size, and Age as explanatory variables (and including an intercept term).

#### ii.

Interpret the value of the estimated regression coefficient corresponding to the Garage Size variable for the full model.

Increasing Garage Size to "Large" is associated with an increased Sales Price of \$15,833 compared to a Garage Size of "Small", all else (all other explanatory variables) being equal.



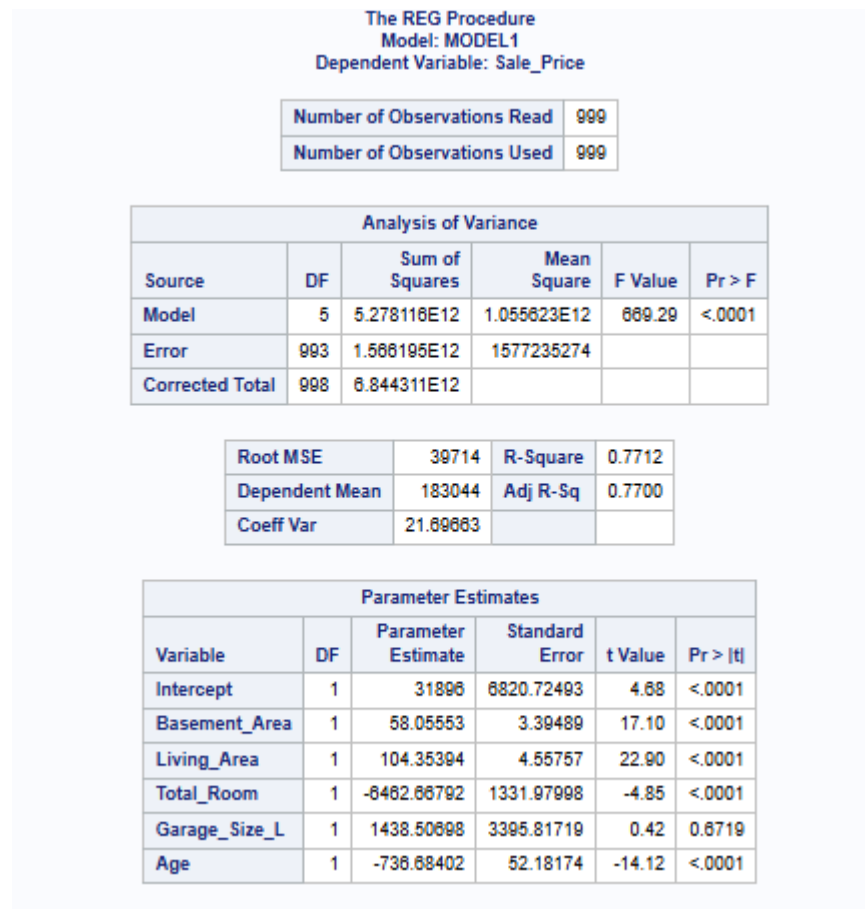


Figure 8: CocoMelon

(b)

Use forward selection to fit a reduced model to the training data using some subset of the 5 explanatory variables listed above. Provide an equation for the estimated MLR model.

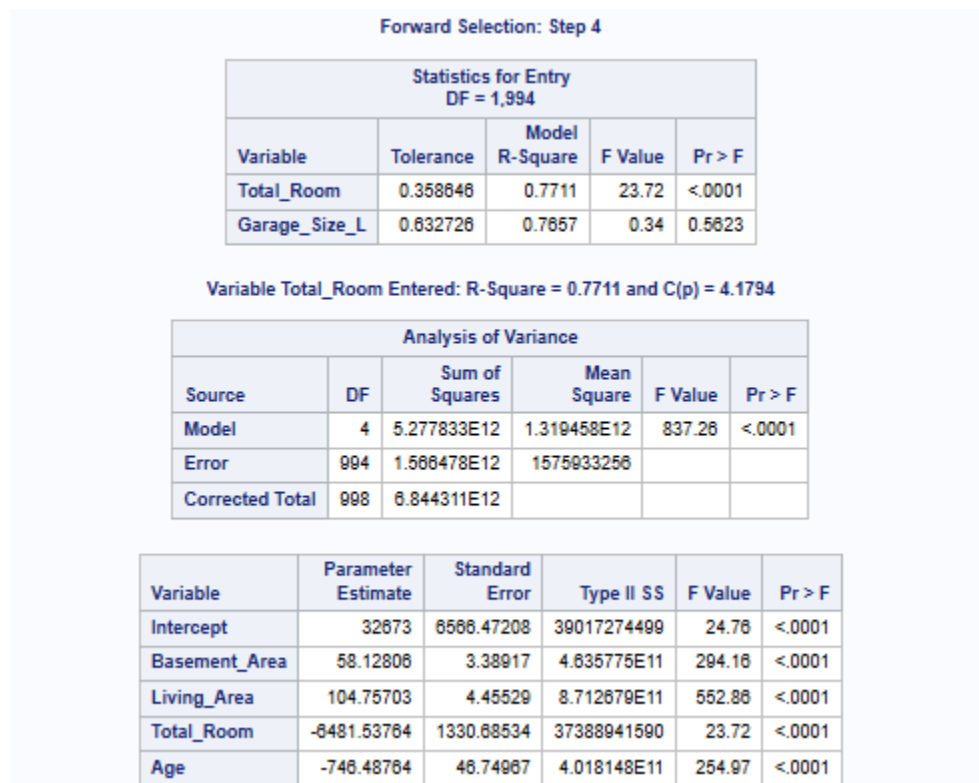


Figure 9: CocoMelon

For the preceding problems, I used the model with “Garage Size” removed, as it didn’t get added for Step 5 of forward selection. The formula corresponding to the above output and the following equation:

$$\widehat{\text{Sale Price}} = 32873 + 58.128 * \text{Basement Area} + 104.75703 * \text{Living Area} - 8481.53784 * \text{Total Room} - 746.48764 * \text{Age}$$

(c)

How does the adjusted  $R^2$  value for the reduced model compare to the full model?

Root MSE	39698	R-Square	0.7711
Dependent Mean	183044	Adj R-Sq	0.7702
Coeff Var	21.68767		

Figure 10: CocoMelon

Full: 0.7700 Reduced: 0.7702 Difference: 0.0002 (2e-04) Note: For Adjusted  $R^2$

The difference of 0.0002 (2e-04) corresponds to a difference of 0.02% between the two models. So the reduced model has a very marginally larger adjusted  $R^2$  value.

(d)

Using the reduced model, check for:

For reference:

Training Data observations:

$$n = 999$$

$$k = 4$$

i.

outliers

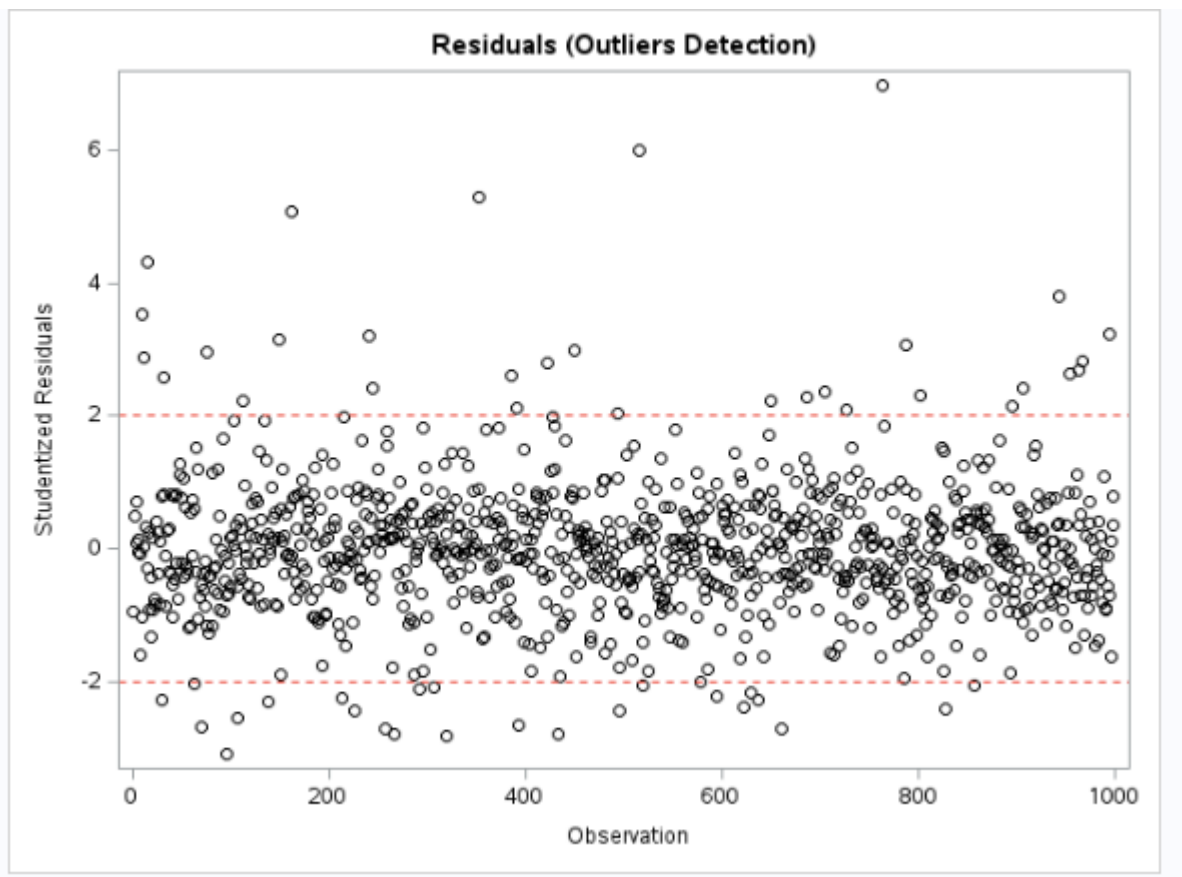


Figure 11: CocoMelon

We do observe there being some potential outliers in the training data, where an outlier is a studentized residual value greater in magnitude to 2, i.e.  $|r| > 2$  where  $r$  is a residual.

ii.

high leverage points

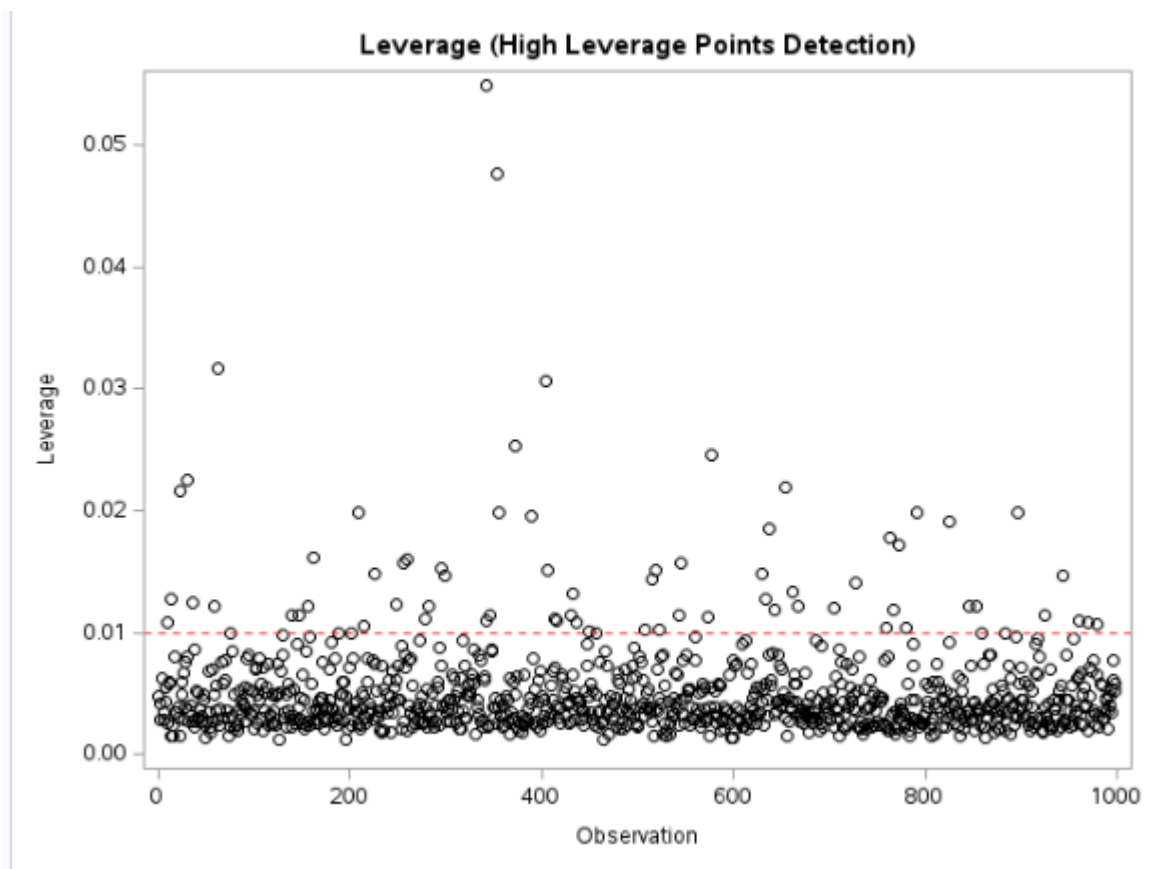


Figure 12: CocoMelon

Leverage threshold:  $2(k + 1)/n = 2(4 + 1)/999 = 0.01001001$

We do observe there being some leverage points in the training data, where our leverage threshold value is calculated to be  $\approx 0.01$ .

iii.

potential influence points

Influence threshold: Cook's D  $D > 2\sqrt{(2/n)} = 2\sqrt{(2/999)} = 0.08948747$

DDFITs:  $2\sqrt{(k + 1/n)} = 2\sqrt{(4 + 1/999)} = 0.1414921$

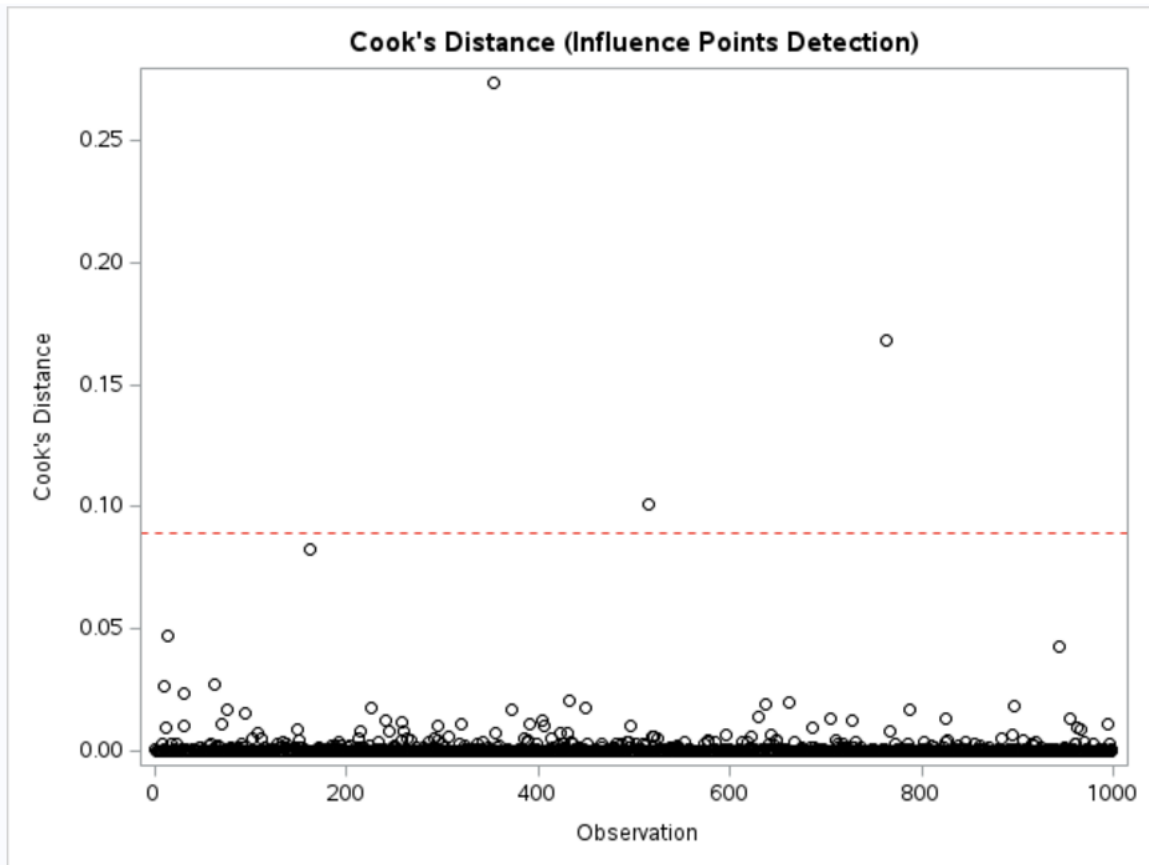


Figure 13: CocoMelon

We do observe there being some potential influence points in the training data, particularly when evaluating based on the Cook's Distance method and a threshold value  $\approx 0.08949$ , as illustrated above. (I also did DDFITS and gave a calculation of the threshold above, but results are consistent with there being some influential points.)

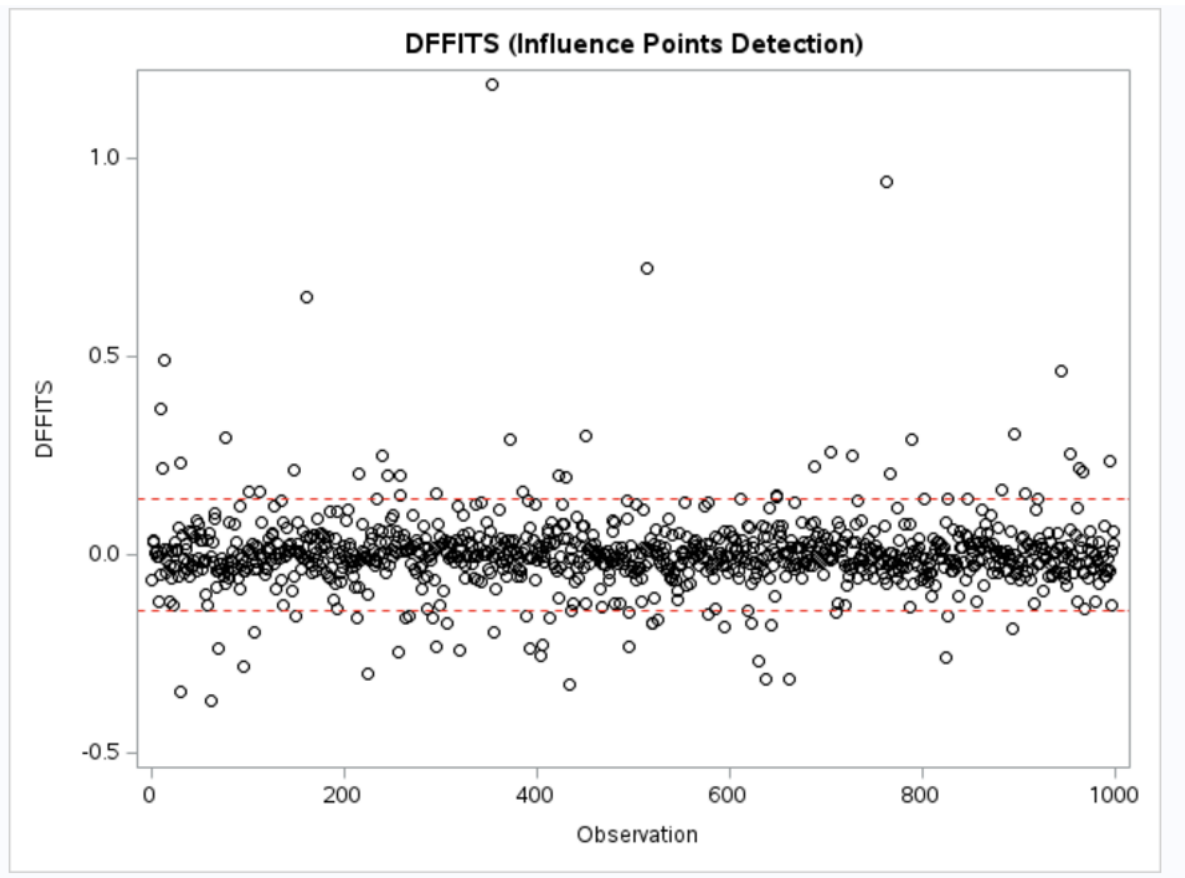


Figure 14: CocoMelon

(e)

Fit the reduced model from part (b) to the evaluation data (housing eval.csv). Compare the mean squared error from fitting the model to the testing data to the mean squared error from fitting the model to the evaluation data. What does this imply?

Eval

The REG Procedure  
Model: MODEL1  
Dependent Variable: Sale\_Price

Number of Observations Read	1924
Number of Observations Used	1924

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9.037262E12	2.259316E12	1553.51	<.0001
Error	1919	2.790858E12	1454329388		
Corrected Total	1923	1.182812E13			

Figure 15: CocoMelon

Train	
The REG Procedure	
Model: MODEL1	
Dependent Variable: Sale_Price	
Number of Observations Read	999
Number of Observations Used	999

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5.277833E12	1.319458E12	837.26	<.0001
Error	994	1.566478E12	1575933256		
Corrected Total	998	6.844311E12			

Figure 16: CocoMelon

$$\text{MSE}_{\text{train}} = 1.5777 \times 10^9$$

$$\text{MSE}_{\text{eval}} = 1.454329 \times 10^9$$

$$\text{Difference in MSE} = \text{MSE}_{\text{train}} - \text{MSE}_{\text{eval}} \approx 1.23371 \times 10^8$$

Importantly, we see that  $\text{MSE}_{\text{eval}} < \text{MSE}_{\text{train}}$ , though the difference is not especially large! This is when looking specifically at the relative difference in MSE, and not the absolute difference in MSE. The relative difference is:



$$\text{Relative Difference} = \frac{\text{Difference in MSE}}{\text{MSE}_{\text{eval}}} * 100 = \frac{1.23371 \times 10^8}{1.454329 \times 10^9} \times 100 \approx 8.48\%$$

Or a roughly 8.48 percent decrease in MSE when going to the eval dataset.

This tells us that: When the MSE (Train) is slightly smaller than the MSE (Eval), meaning the model we chose performs marginally better on the evaluation dataset than the training dataset, and corresponding to a decrease of 8.48%! Generally this marginal relative difference indicates that our model is not overfitting, and that it is generalizing well from the training to the evaluation dataset. It is performing “ok”! (We want MSE to be lower for the eval compared to training, otherwise we’d suspect overfitting based on the training dataset).

### 3

The dataset for this exercise is called diamonds and it is available directly in the ggplot2 package in R. The data set contains prices (response variable – in US dollars) of over 50,000 diamonds, which we will try to explain using the quantitative size measurements:

- carat – weight,
- x – length in mm,
- y – width in mm,
- z – depth in mm,
- depth – total depth percentage =  $z / \text{mean}(x, y)$ ,
- table – width of top of diamond relative to widest point)

And categorical quality (cut, color, and clarity) of the diamonds. The R code used to create the figures below is provided in the diamonds Hmwk 11.R file posted in Canvas.

(a)

Summarize your findings from examining the pairwise scatterplots (on the next page) and correlation matrix (shown below).

	carat	depth	table	price	x	y	z
carat	1.00000000	0.02822431	0.1816175	0.9215913	0.97509423	0.95172220	0.95338738
depth	0.02822431	1.00000000	-0.2957785	-0.0106474	-0.02528925	-0.02934067	0.09492388
table	0.18161755	-0.29577852	1.00000000	0.1271339	0.19534428	0.18376015	0.15092869
price	0.92159130	-0.01064740	0.1271339	1.00000000	0.88443516	0.86542090	0.86124944
x	0.97509423	-0.02528925	0.1953443	0.8844352	1.00000000	0.97470148	0.97077180
y	0.95172220	-0.02934067	0.1837601	0.8654209	0.97470148	1.00000000	0.95200572
z	0.95338738	0.09492388	0.1509287	0.8612494	0.97077180	0.95200572	1.00000000

Figure 17: CocoMelon

Many of the variables are highly correlated with each other; this holds for combinations of variables with the response variable (price) as well as between explanatory variables. Using the “significance” threshold of  $|r| > 0.7$ , we have the following “significant” correlations: Carat and price, carat and x, carat and y, carat and z, price and x, price and y, price and z, x and y, x and z, y and z. Of note is that all of the “significant” correlations are positive. Overall, we do corroborate these findings when looking at the pairwise scatter plots, insomuch as positive correlations generally show a positive linear relationship when looking at the respective graph. On the flip side, we see “small” correlations ( $|r| < 0.20$ , such as x and depth), exhibit a rather large “blob” of points, or for other pairs a general overall coverage of the plot area. This is understandable, as a small-in-magnitude correlation coefficient is evidence that there is not a significant linear relationship between the two variables.

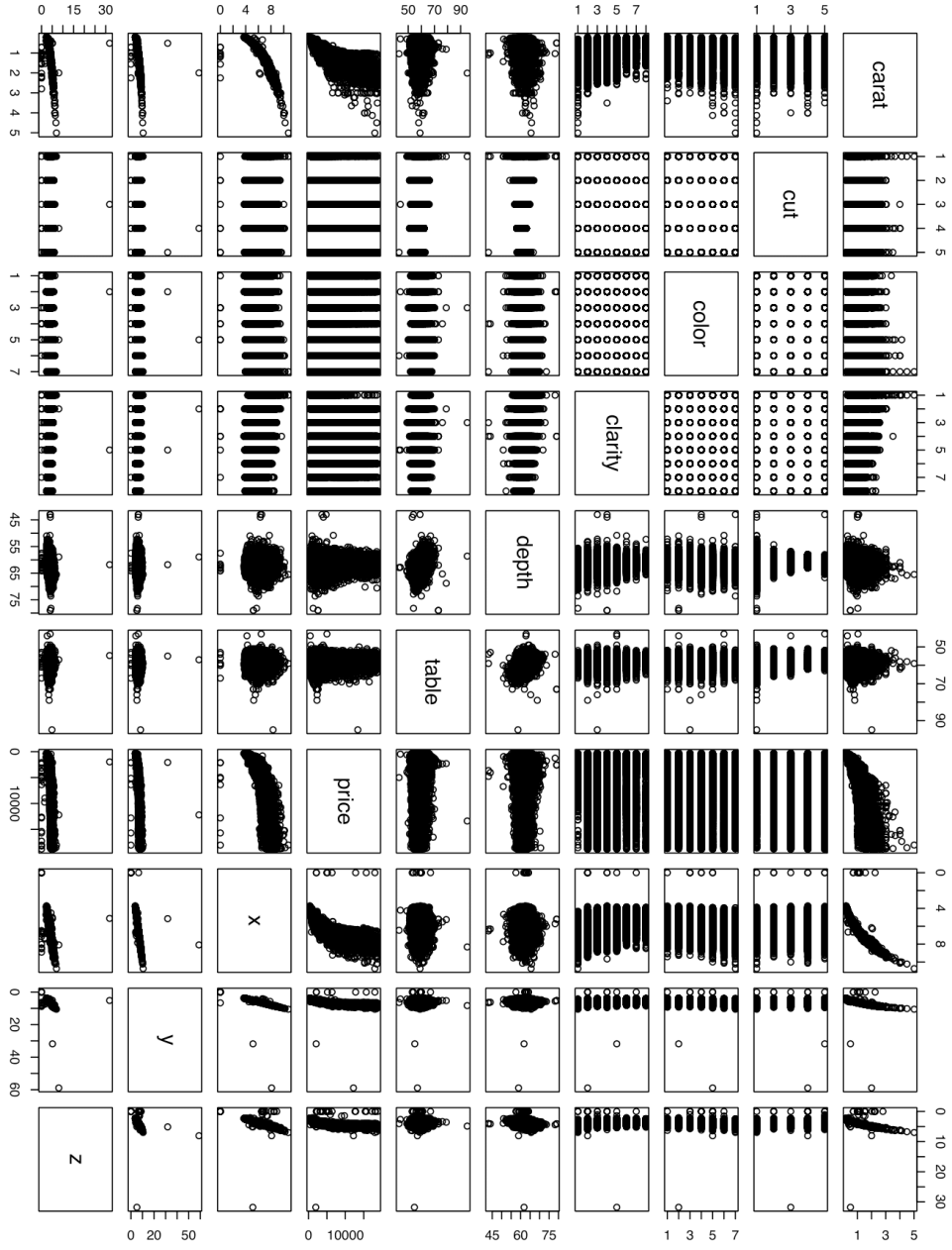


Figure 18: CocoMelon

(b)

Discuss whether the VIFs, shown in the plot below, indicate any explanatory variables exhibiting moderate or extreme multicollinearity.

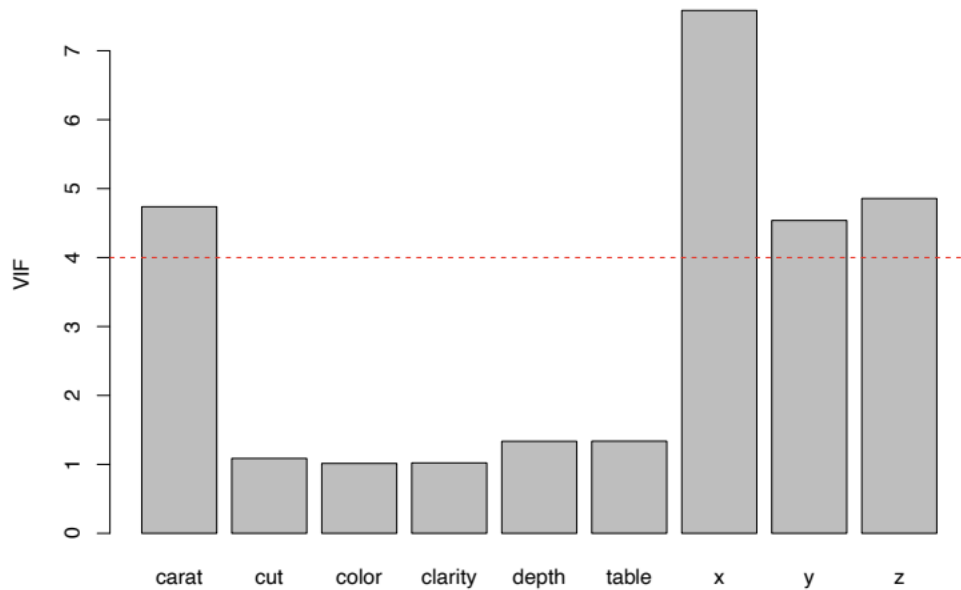


Figure 19: CocoMelon

Carat, x, y, and z all have  $VIF > 4$ , indicating moderate multicollinearity.

The output does not explicitly state if any of the VIF values are greater than 10, as values are cut off at the 7. So for “extreme multicollinearity” corresponding to a VIF greater than 10, we cannot determine explicitly if the “x” explanatory variable exhibits extreme multicollinearity (which is the only explanatory variable that *may* meet that criteria).

(c)

Summarize the backward elimination method of model selection by providing:

```
Call:
lm(formula = price ~ carat + cut + color + clarity + depth +
    table + x + z, data = diamonds)

Residuals:
    Min       1Q   Median       3Q      Max
-21378.8  -592.5   -183.5    376.3  10694.1

Coefficients:
(Intercept)    2198.886    407.163    5.401 6.67e-08 ***
carat       11257.752    48.602 231.630 < 2e-16 ***
cut2         580.325     33.572   17.286 < 2e-16 ***
cut3         727.431     32.214   22.581 < 2e-16 ***
cut4         762.287     32.226   23.654 < 2e-16 ***
cut5         833.352     33.396   24.954 < 2e-16 ***
color2       -209.100     17.893  -11.686 < 2e-16 ***
color3       -272.837     18.093  -15.080 < 2e-16 ***
color4       -482.035     17.716  -27.209 < 2e-16 ***
color5       -980.247     18.836  -52.042 < 2e-16 ***
color6      -1466.257     21.162  -69.287 < 2e-16 ***
color7      -2369.412     26.131  -90.675 < 2e-16 ***
clarity2      2702.855     43.815   61.688 < 2e-16 ***
clarity3      3665.735     43.631   84.018 < 2e-16 ***
clarity4      4267.476     43.850   97.319 < 2e-16 ***
clarity5      4578.702     44.541  102.796 < 2e-16 ***
clarity6      4951.100     45.851  107.983 < 2e-16 ***
clarity7      5008.029     47.156  106.201 < 2e-16 ***
clarity8      5345.420     51.020  104.772 < 2e-16 ***
depth        -64.003      4.517  -14.168 < 2e-16 ***
table        -26.501      2.911   -9.103 < 2e-16 ***
x          -1000.354     28.795  -34.740 < 2e-16 ***
z           -47.925     33.194   -1.444    0.149

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1130 on 53917 degrees of freedom
Multiple R-squared:  0.9198,    Adjusted R-squared:  0.9198
F-statistic: 2.81e+04 on 22 and 53917 DF,  p-value: < 2.2e-16
```

Start: AIC=758426.5  
price ~ carat + cut + color + clarity + depth + table + x + y + z

	Df	Sum of Sq	RSS	AIC
- y	1	3.1549e+05	6.8857e+10	758425
<none>			6.8857e+10	758426
- z	1	2.8609e+06	6.8860e+10	758427
- table	1	1.0558e+08	6.8962e+10	758507
- depth	1	2.5286e+08	6.9110e+10	758622
- cut	4	8.6357e+08	6.9720e+10	759091
- x	1	1.1996e+09	7.0056e+10	759356
- color	6	1.7082e+10	8.5939e+10	770368
- clarity	7	3.5703e+10	1.0456e+11	780945
- carat	1	6.8440e+10	1.3730e+11	795649

Step: AIC=758424.7  
price ~ carat + cut + color + clarity + depth + table + x + z

	Df	Sum of Sq	RSS	AIC
<none>			6.8857e+10	758425
- z	1	2.6622e+06	6.8860e+10	758425
- table	1	1.0584e+08	6.8963e+10	758506
- depth	1	2.5637e+08	6.9114e+10	758623
- cut	4	8.6409e+08	6.9721e+10	759089
- x	1	1.5413e+09	7.0398e+10	759617
- color	6	1.7082e+10	8.5940e+10	770366
- clarity	7	3.5708e+10	1.0457e+11	780946
- carat	1	6.8520e+10	1.3738e+11	795679

Figure 20: CocoMelon

i.

an ordered list of which variable was removed from the model at each step;

Step 1: variable “y” was removed.

There are no more explanatory variables eliminated.

ii.

a list of which variables remained in the final model;

Explanatory variables that remained: Carat, cut, color, clarity, depth, and table.

iii.

a summary of the partial regression coefficients effects tests for the final model.

All the final model partial regression coefficients meet statistical significance to reject null hypothesis at the  $\alpha = 0.05$  level except for “z”. There is only one partial regression coefficients in the final model that does not meet statistical significance to reject null hypothesis at the  $\alpha = 0.05$  level, “z”.

Of note, all explanatory variables except for “z” meet the significance threshold of “overwhelming” or “strong” evidence (in favor of rejecting the null hypothesis).

The statistical significance test is to determine whether there is evidence to reject the null hypothesis that the estimated beta coefficient is equal to zero (statistical significance referring to being statistically significant from zero).

(d)

Summarize the forward selection method of model selection by providing:

```

Start: AIC=894477.9
price ~ 1

+ carat      Df Sum of Sq      RSS      AIC
+ x          1  6.7152e+11  1.8695e+11  812259
+ y          1  6.4296e+11  2.1552e+11  819929
+ z          1  6.3677e+11  2.2170e+11  821454
+ color      6  2.6849e+10  8.3162e+11  892776
+ clarity    7  2.3308e+10  8.3517e+11  893007
+ table      1  1.3876e+10  8.4460e+11  893601
+ cut        4  1.1042e+10  8.4743e+11  893788
+ depth      1  9.7323e+07  8.5838e+11  894474
<none>                      8.5847e+11  894478

Step: AIC=792389.4
price ~ carat

+ clarity    Df Sum of Sq      RSS      AIC
+ color      6  1.2561e+10  1.1678e+11  786891
+ cut        4  6.1332e+09  1.2321e+11  789777
+ x          1  3.5206e+09  1.2583e+11  790903
+ z          1  2.8493e+09  1.2650e+11  791190
+ table      1  1.4377e+09  1.2791e+11  791789
+ y          1  1.2425e+09  1.2810e+11  791871
+ depth      1  1.1546e+09  1.2819e+11  791908
<none>                      1.2935e+11  792389

Step: AIC=772998.5
price ~ carat + clarity

+ color      Df Sum of Sq      RSS      AIC
+ x          1  1.8542e+09  8.8410e+10  771881
+ cut        4  1.7808e+09  8.8483e+10  771932
+ z          1  1.4814e+09  8.8783e+10  772108
+ y          1  7.4127e+08  8.9523e+10  772556
+ table      1  3.7751e+08  8.9886e+10  772774
+ depth      1  3.5822e+08  8.9906e+10  772786
<none>                      9.0264e+10  772998

Step: AIC=762193.4
price ~ carat + clarity + color

+ x          1  2733710969  7.1128e+10  760161
+ z          1  1842294631  7.2020e+10  760833
+ cut        4  1699187372  7.2163e+10  760946
+ y          1  1145039064  7.2717e+10  761353
+ table      1  409645878  7.3452e+10  761895
+ depth      1  174658715  7.3687e+10  762068
<none>                      7.3862e+10  762193

Step: AIC=760161.1
price ~ carat + clarity + color + x

+ cut        4  1918248123  6.9210e+10  758694
+ depth      1  722282102  7.0406e+10  759613
+ table      1  273738191  7.0855e+10  759955
+ z          1  199547343  7.0929e+10  760012
+ y          1  5354253    7.1123e+10  760159
<none>                      7.1128e+10  760161

Step: AIC=758694.4
price ~ carat + clarity + color + x + cut

+ depth      1  244682865  6.8965e+10  758505
+ z          1  726666922  6.9137e+10  758640
+ table      1  9935285    6.9200e+10  758689
<none>                      6.9210e+10  758694
+ y          1  982101    6.9209e+10  758696

Step: AIC=758505.4
price ~ carat + clarity + color + x + cut + depth

+ table      Df Sum of Sq      RSS      AIC
+ z          1  105497218  6.8860e+10  758425
<none>                      6.8965e+10  758505
+ z          1  2323719  6.8963e+10  758506
+ y          1  298553    6.8965e+10  758507

Step: AIC=758424.8
price ~ carat + clarity + color + x + cut + depth + table

+ z          1  2662170  6.8857e+10  758425
<none>                      6.8860e+10  758425
+ y          1  116788  6.8860e+10  758427

Step: AIC=758424.7
price ~ carat + clarity + color + x + cut + depth + table + z

+ y          1  315487  6.8857e+10  758426

```

Figure 21: CocoMelon

i.

an ordered list of which variable was added to the model at each step;

First Step: carat, Second Step: clarity, Third Step: color, Fourth Step: x, Fifth Step: cut, Sixth Step: depth, Seventh Step: table, Eighth Step: z

ii.

a list of which variables never entered the final model;

The explanatory variable “y” never entered the final model.

iii.

a summary of the partial regression coefficients effects tests for the final model.

Consistent with the prior model,

All the final model partial regression coefficients meet statistical significance to reject null hypothesis at the  $\alpha = 0.05$  level except for “z”. There is only one partial regression coefficients in the final model that does not meet statistical significance to reject null hypothesis at the  $\alpha = 0.05$  level, “z”.

```

Call:
lm(formula = price ~ carat + clarity + color + x + cut + depth +
    table + z, data = diamonds)

Residuals:
    Min       1Q   Median       3Q      Max
-21378.8  -592.5   -183.5    376.3  10694.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2198.886    407.163   5.401 6.67e-08 ***
carat       11257.752    48.602  231.630 < 2e-16 ***
clarity2     2702.855    43.815   61.688 < 2e-16 ***
clarity3     3665.735    43.631   84.018 < 2e-16 ***
clarity4     4267.476    43.850   97.319 < 2e-16 ***
clarity5     4578.702    44.541  102.796 < 2e-16 ***
clarity6     4951.100    45.851  107.983 < 2e-16 ***
clarity7     5008.029    47.156  106.201 < 2e-16 ***
clarity8     5345.420    51.020  104.772 < 2e-16 ***
color2       -209.100    17.893  -11.686 < 2e-16 ***
color3       -272.837    18.093  -15.080 < 2e-16 ***
color4       -482.035    17.716  -27.209 < 2e-16 ***
color5       -980.247    18.836  -52.042 < 2e-16 ***
color6      -1466.257    21.162  -69.287 < 2e-16 ***
color7      -2369.412    26.131  -90.675 < 2e-16 ***
x           -1000.354    28.795  -34.740 < 2e-16 ***
cut2          580.325    33.572   17.286 < 2e-16 ***
cut3          727.431    32.214   22.581 < 2e-16 ***
cut4          762.287    32.226   23.654 < 2e-16 ***
cut5          833.352    33.396   24.954 < 2e-16 ***
depth        -64.003     4.517  -14.168 < 2e-16 ***
table        -26.501     2.911   -9.103 < 2e-16 ***
z           -47.925    33.194   -1.444    0.149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1130 on 53917 degrees of freedom
Multiple R-squared:  0.9198,    Adjusted R-squared:  0.9198
F-statistic: 2.81e+04 on 22 and 53917 DF,  p-value: < 2.2e-16

```

Figure 22: CocoMelon



Consistent with part (d), of note, all explanatory variables except for “z” meet the significance threshold of “overwhelming” or “strong” evidence (in favor of rejecting the null hypothesis).

The statistical significance test is to determine whether there is evidence to reject the null hypothesis that the estimated beta coefficient is equal to zero (statistical significance referring to being statistically significant from zero).

(e)

Summarize the all-possible-subsets method of model selection by providing:

1 subsets of each size up to 8													
Selection Algorithm: exhaustive													
		carat	cut2	cut3	cut4	cut5	color2	color3	color4	color5	color6	color7	clarity2
1	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
2	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"
3	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	"*"
4	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	"*"
5	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	"*"	"*"	"*"
6	( 1 )	"*"	" "	" "	" "	" "	" "	" "	"*"	"*"	"*"	"*"	"*"
7	( 1 )	"*"	" "	" "	" "	" "	" "	" "	"*"	"*"	"*"	"*"	"*"
8	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	"*"	"*"	" "
		clarity3	clarity4	clarity5	clarity6	clarity7	clarity8	depth	table	x	y	z	
1	( 1 )	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
2	( 1 )	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
3	( 1 )	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
4	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
5	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
6	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
7	( 1 )	"*"	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "
8	( 1 )	"*"	"*"	"*"	"*"	"*"	"*"	" "	" "	" "	" "	" "	" "
Model		1	2	3	4	5	6	7	8				
adj $R^2$		0.8493	0.8643	0.8728	0.8806	0.8855	0.8890	0.8927	0.8971				
$C_p$		47344	37249	31567	26311	23020	20657	18212	15269				
$BIC$		-102069	-107722	-111183	-114596	-116846	-118518	-120307	-122544				

Figure 23: CocoMelon

i.

Which model would you choose based on the adjusted R2 values?

Model 8, using the explanatory variables of carat, color 7, and clarity 3 through 8 (possibly with an intercept term as well).

ii.

Which model would you choose based on the Mallows' Cp criteria?

Model 8, using the explanatory variables of carat, color 7, and clarity 3 through 8 (possibly with an intercept term as well).

iii.

Which model would you choose based on the BIC values?

Model 8, using the explanatory variables of carat, color, and clarity (possibly with an intercept term as well).

(f)

Interpret the values of the estimated regression coefficients for the final model selected:

i.

one of the values corresponding to the categorical variable of your choice;

The mean price of diamonds that are cut 2 are priced \$580.325 more than diamonds that are cut 1, holding all other variables constant (given the other variables in the model).

ii.

one of the values corresponding to the quantitative variable of your choice.

For every 1 carat increase in weight, the mean price of diamonds will increase by \$11,257.752 (associated with an increase of), holding all other variables constant (given all other variables in the model).

```
Call:
lm(formula = price ~ carat + cut + color + clarity + depth +
    table + x + z, data = diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-21378.8	-592.5	-183.5	376.3	10694.1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2198.886	407.163	5.401	6.67e-08 ***
carat	11257.752	48.602	231.630	< 2e-16 ***
cut2	580.325	33.572	17.286	< 2e-16 ***
cut3	727.431	32.214	22.581	< 2e-16 ***
cut4	762.287	32.226	23.654	< 2e-16 ***
cut5	833.352	33.396	24.954	< 2e-16 ***
color2	-209.100	17.893	-11.686	< 2e-16 ***
color3	-272.837	18.093	-15.080	< 2e-16 ***
color4	-482.035	17.716	-27.209	< 2e-16 ***
color5	-980.247	18.836	-52.042	< 2e-16 ***
color6	-1466.257	21.162	-69.287	< 2e-16 ***
color7	-2369.412	26.131	-90.675	< 2e-16 ***
clarity2	2702.855	43.815	61.688	< 2e-16 ***
clarity3	3665.735	43.631	84.018	< 2e-16 ***
clarity4	4267.476	43.850	97.319	< 2e-16 ***
clarity5	4578.702	44.541	102.796	< 2e-16 ***
clarity6	4951.100	45.851	107.983	< 2e-16 ***
clarity7	5008.029	47.156	106.201	< 2e-16 ***
clarity8	5345.420	51.020	104.772	< 2e-16 ***
depth	-64.003	4.517	-14.168	< 2e-16 ***
table	-26.501	2.911	-9.103	< 2e-16 ***
x	-1000.354	28.795	-34.740	< 2e-16 ***
z	-47.925	33.194	-1.444	0.149

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1130 on 53917 degrees of freedom  
 Multiple R-squared: 0.9198, Adjusted R-squared: 0.9198  
 F-statistic: 2.81e+04 on 22 and 53917 DF, p-value: < 2.2e-16

Analysis of Variance Table

Response: price

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
carat	1	7.2913e+11	7.2913e+11	5.7093e+05	<2e-16 ***
cut	4	6.1332e+09	1.5333e+09	1.2006e+03	<2e-16 ***
color	6	1.2598e+10	2.0997e+09	1.6441e+03	<2e-16 ***
clarity	7	3.8452e+10	5.4931e+09	4.3012e+03	<2e-16 ***
depth	1	4.9405e+06	4.9405e+06	3.8686e+00	0.0492 *
table	1	9.2727e+07	9.2727e+07	7.2607e+01	<2e-16 ***
x	1	3.2053e+09	3.2053e+09	2.5098e+03	<2e-16 ***
z	1	2.6622e+06	2.6622e+06	2.0846e+00	0.1488
Residuals	53917	6.8857e+10	1.2771e+06		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Figure 24: CocoMelon

(g)

Summarize your findings from examining all the residual plots used to diagnose the MLR model assumptions. Are there any assumptions that aren't met for this analysis? Briefly justify your response.

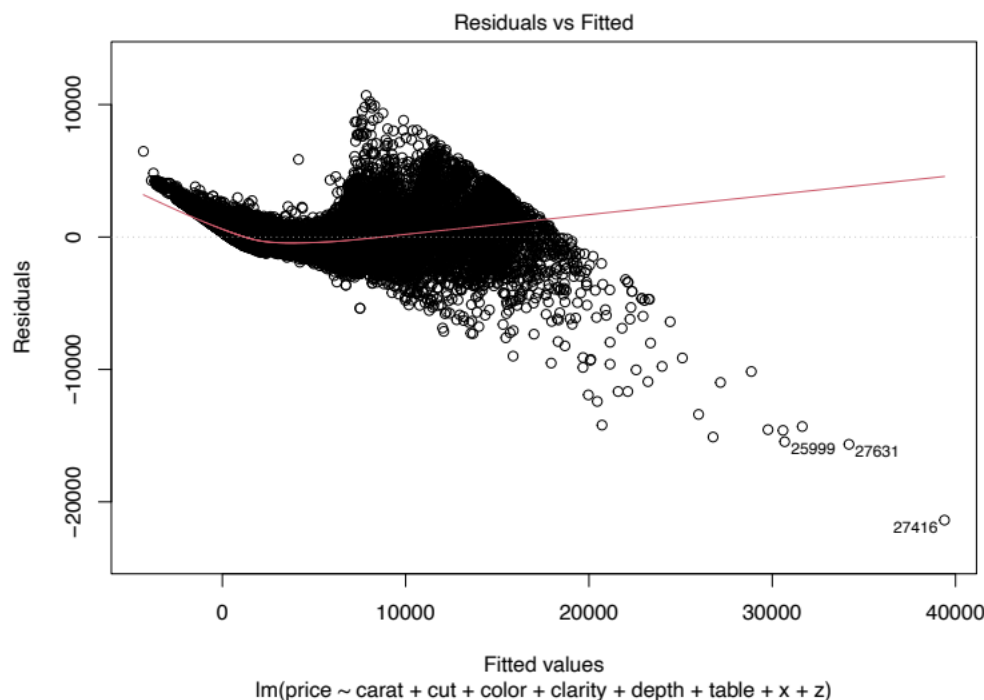


Figure 25: CocoMelon

The key model assumptions we evaluate with the given plots and graphs are: Equal variance, linearity (form of the model), and normality. That being said:

The residual plot has a clearly obvious non-random-scattering trend, suggesting the linearity and equal variance assumptions may be violated. (We observe a trend in the residual plot, particularly indicating a violation of linearity).

Additionally, the QQ is roughly linear as it follows the reference line closely in the middle of the plot. However, there are deviations in the left tail and right of the middle, suggesting the normality assumption is violated.

Taken together, we have reason to be concerned that our key assumptions are being violated, with regards to the residuals.

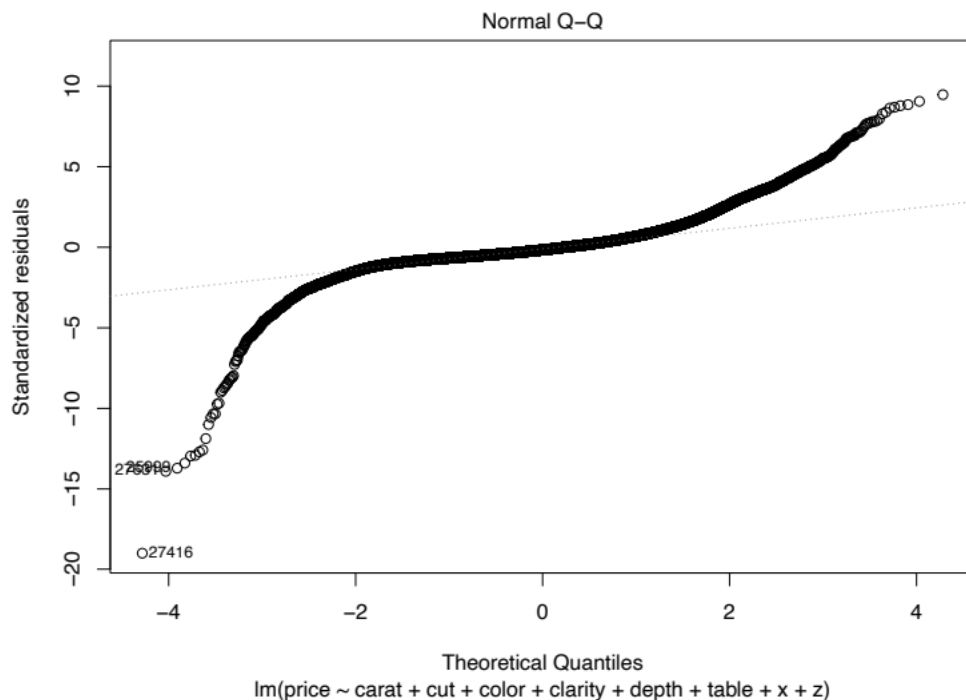


Figure 26: CocoMelon

(h)

Summarize your findings from examining the case diagnostic values/plots. Are there any outliers, leverage points, or influential observations?

Outliers: Looking at the studentized residual plot, there appear to be many outliers (looking for observations where the studentized residuals exceed 2, based on magnitude, i.e.  $|r|$  for  $r$  residuals, as indicated by the dashed red lines).

Leverage: From the leverage plot, there appears to be many high leverage points (leverage values exceeding  $2(8 + 1)/50000 \approx 0.00036$ , as indicated by the red line).

From the cook's D plot, there appear to be several high influence points (influence value exceeding  $2\sqrt{2}/50000 \approx 0.01264911$ , as indicated by the red line).

Overall, we have reason to believe there are outliers, leverage points, and influential points, such that we may recommend considering other models or undergoing a transformation of our data and reevaluating our assumptions again. As-is, we have reason to be suspect.

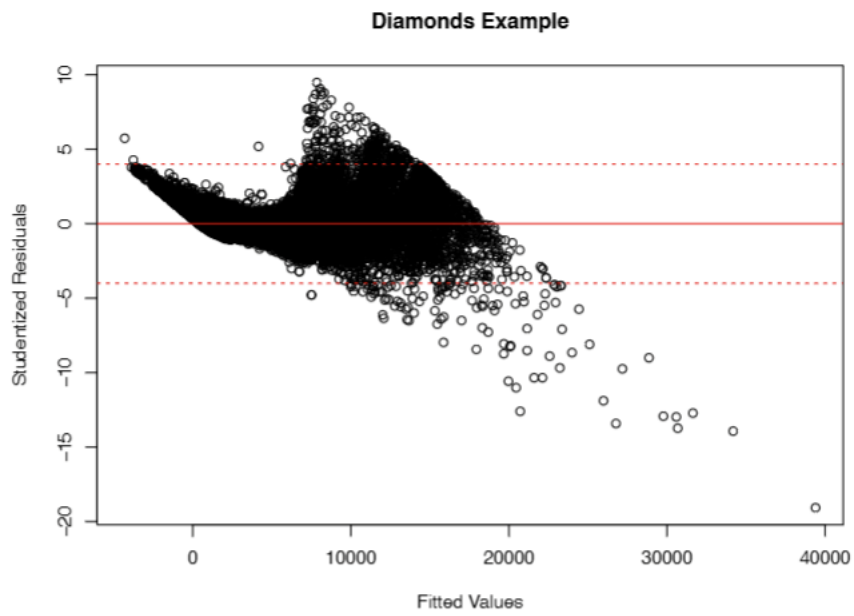


Figure 27: CocoMelon

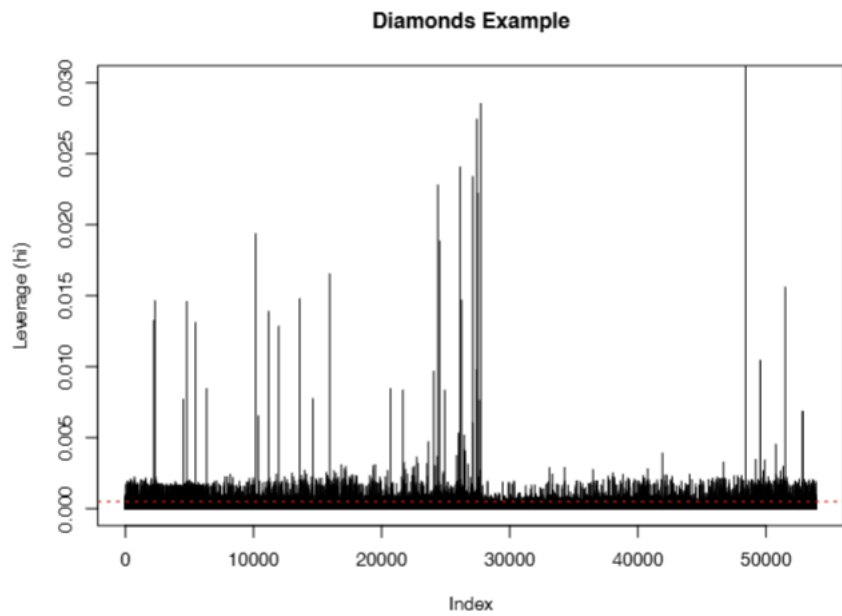


Figure 28: CocoMelon

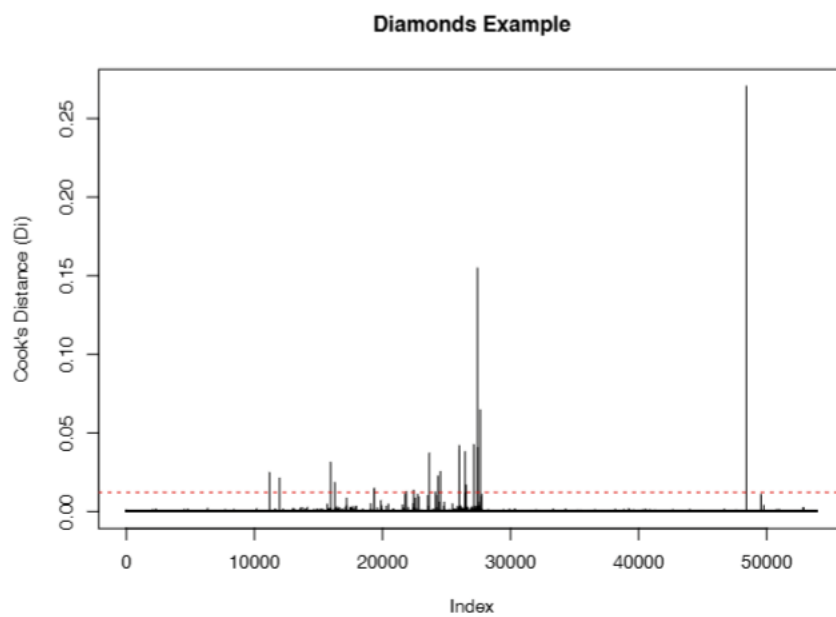


Figure 29: CocoMelon