

Directions: Type or clearly handwrite your solutions to each of the following exercises. Partial credit cannot be given unless all work is shown. You may work in groups provided that each person takes responsibility for understanding and writing out the solutions. Additionally, you must give proper credit to your collaborators by providing their names on the line below (if you worked alone, write “No Collaborators”):

1. [+25]: Let’s reconsider the full dataset for the Berkeley Guidance Study from the previous homework assignment. The data are posted in the file `BGSgirls2.txt`. There is one line of data for each of 70 girls with the variables appearing in the following order:

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

The goal of this exercise is to determine how well the measurements at ages 2 and 9 can predict BMI at age 18.

Use SAS to complete the following exercises:

- (a) Fit the multiple regression model

$$\text{BMI}_i = \beta_0 + \beta_1 \text{WT2}_i + \beta_2 \text{HT2}_i + \beta_3 \text{WT9}_i + \beta_4 \text{HT9}_i + \beta_5 \text{ST9}_i + \epsilon_i, \quad \text{for } i = 1, \dots, 70.$$

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

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

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

- (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.)
- (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 $\alpha_{\text{stay}} = 0.05$. 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.

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

2. [+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`).
- Find and interpret the R^2 value for the *full* model.

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

(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.

(c) How does the adjusted R^2 value for the *reduced* model compare to the *full* model?

(d) Using the *reduced* model, check for:

i. outliers,

ii. high leverage points,

iii. potential influence points.

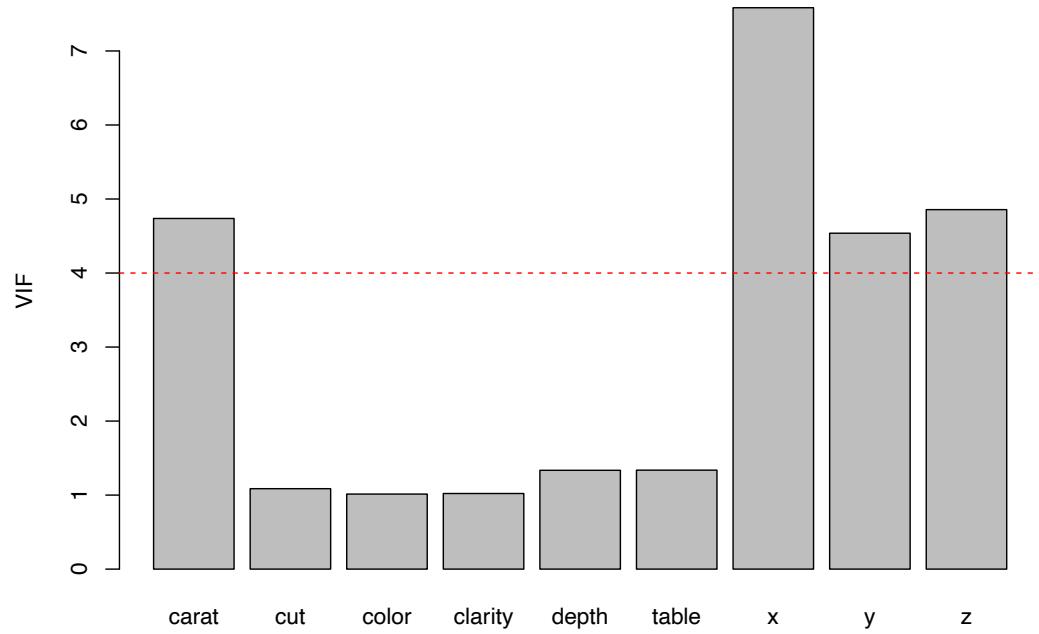
- (e) Fit the *reduced* model from part (b) to the evaluation data (`housing_eval.csv`). Compare the mean squared error from fitting the model to the testing data to the mean squared error from fitting the model to the evaluation data. What does this imply?

3. [+50]: 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_Hmwk11.R` file posted in Canvas.

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

	<code>carat</code>	<code>depth</code>	<code>table</code>	<code>price</code>	<code>x</code>	<code>y</code>	<code>z</code>
<code>carat</code>	1.0000000	0.02822431	0.1816175	0.9215913	0.97509423	0.95172220	0.95338738
<code>depth</code>	0.02822431	1.0000000	-0.2957785	-0.0106474	-0.02528925	-0.02934067	0.09492388
<code>table</code>	0.18161755	-0.29577852	1.0000000	0.1271339	0.19534428	0.18376015	0.15092869
<code>price</code>	0.92159130	-0.01064740	0.1271339	1.0000000	0.88443516	0.86542090	0.86124944
<code>x</code>	0.97509423	-0.02528925	0.1953443	0.8844352	1.0000000	0.97470148	0.97077180
<code>y</code>	0.95172220	-0.02934067	0.1837601	0.8654209	0.97470148	1.0000000	0.95200572
<code>z</code>	0.95338738	0.09492388	0.1509287	0.8612494	0.97077180	0.95200572	1.0000000

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



Backward Elimination Model Selection Output from R:

```

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

```

- (c) Summarize the backward elimination method of model selection by providing:
- an ordered list of which variable was removed from the model at each step;

- a list of which variables remained in the final model;

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

Forward Selection Model Selection Output from R:

```

Start: AIC=894477.9
price ~ 1

Step: AIC=762193.4
      Df Sum of Sq    RSS   AIC
+ carat   1 7.2913e+11 1.2935e+11 792389
+ 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
<none>
+ 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

      Df Sum of Sq    RSS   AIC
+ clarity 7 3.9082e+10 9.0264e+10 772998
+ 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

      Df Sum of Sq    RSS   AIC
+ color   6 1.6402e+10 7.3862e+10 762193
+ 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=760161.1
price ~ carat + clarity + color + x

      Df Sum of Sq    RSS   AIC
+ 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=758424.8
price ~ carat + clarity + color + x + cut + depth + table

      Df Sum of Sq    RSS   AIC
+ 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

      Df Sum of Sq    RSS   AIC
<none>   6.8857e+10 758425
+ y       1 315487 6.8857e+10 758426

```

- (d) Summarize the forward selection method of model selection by providing:
- i. an ordered list of which variable was added to the model at each step;
 - ii. a list of which variables never entered the final model;
 - iii. a summary of the partial regression coefficients effects tests for the final model.
(See output on the next page ...)

Forward Selection Final Model Output from R:

```

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

```

All-Possible-Subsets Model Selection Output from R:

1 subsets of each size up to 8
 Selection Algorithm: exhaustive

	carat	cut2	cut3	cut4	cut5	color2	color3	color4	color5	color6	color7	clarity2	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

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

i. Which model would you choose based on the adjusted R^2 values?

ii. Which model would you choose based on the Mallow's C_p criteria?

iii. Which model would you choose based on the BIC values?

R Output for the Final 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

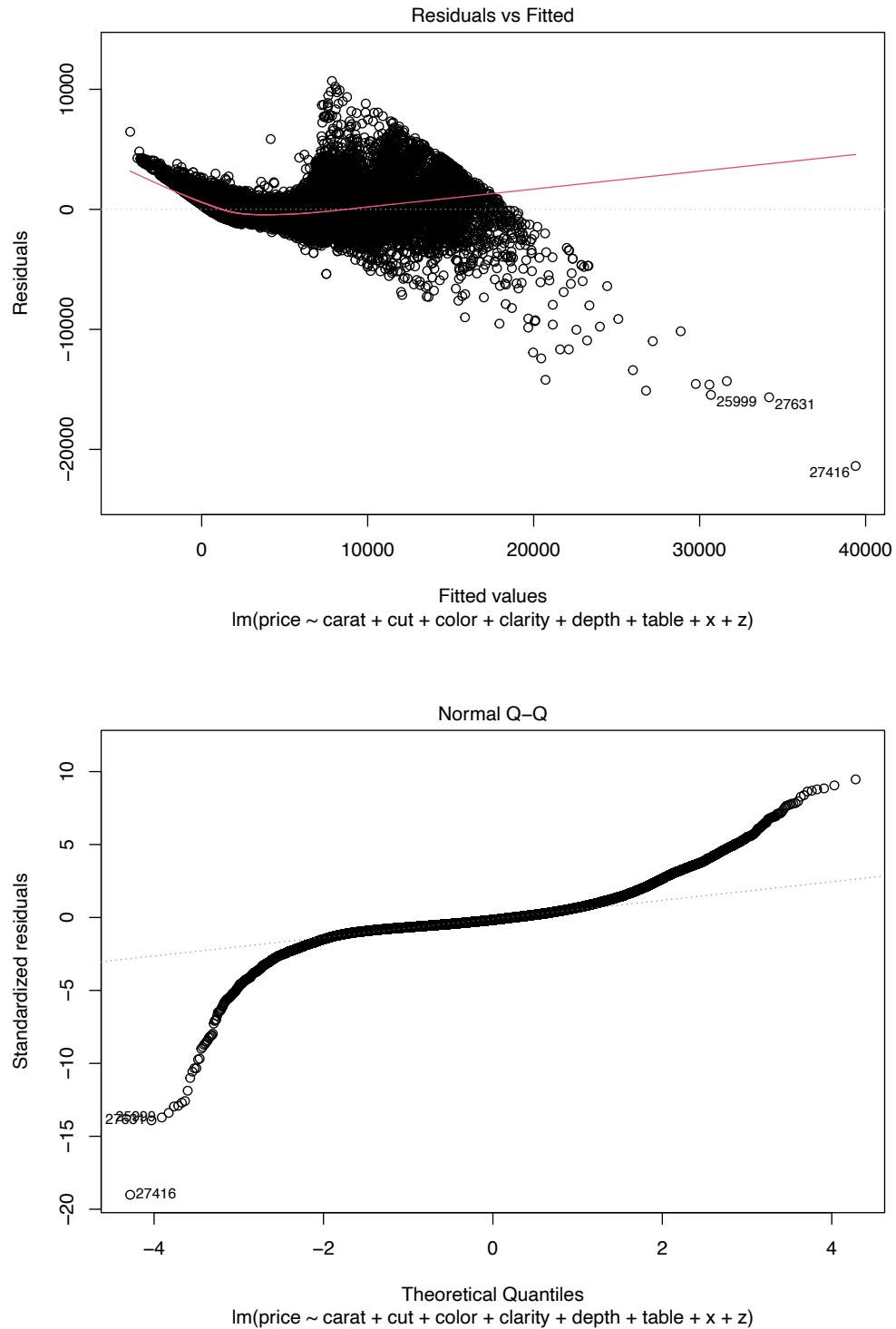
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
carat	1	7.2913e+11	7.2913e+11	5.7093e+05	<2e-16 ***
cut	4	6.1332e+09	1.5333e+09	1.2006e+03	<2e-16 ***
color	6	1.2598e+10	2.0997e+09	1.6441e+03	<2e-16 ***
clarity	7	3.8452e+10	5.4931e+09	4.3012e+03	<2e-16 ***
depth	1	4.9405e+06	4.9405e+06	3.8686e+00	0.0492 *
table	1	9.2727e+07	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

- (f) Interpret the values of the estimated regression coefficients for the final model selected:

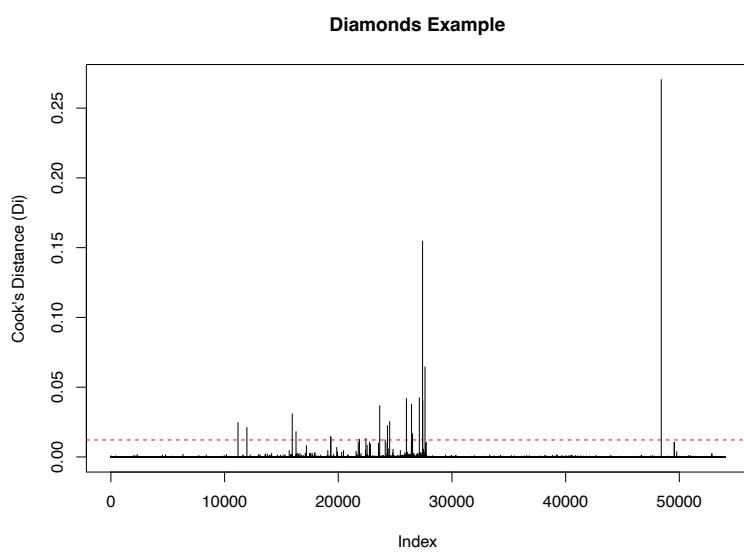
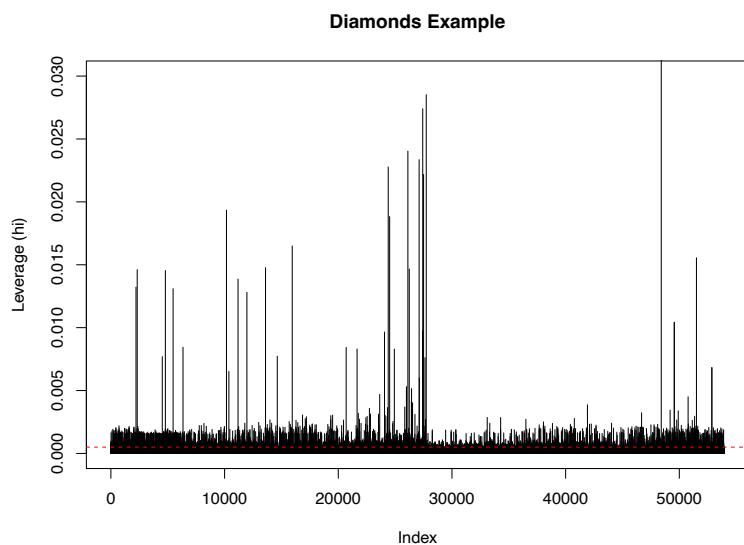
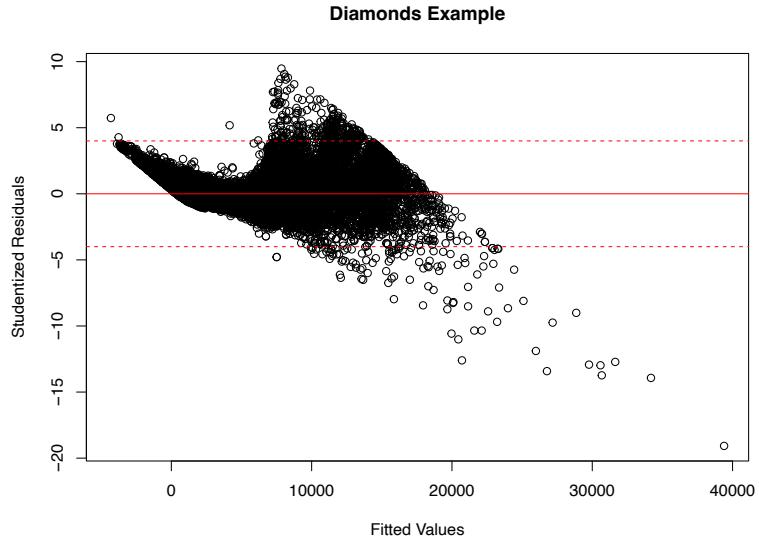
 - one of the values corresponding to the categorical variable of your choice;
 - one of the values corresponding to the quantitative variable of your choice.

Residual Plots for the Final Model:



- (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.

Case Diagnostic Plots for the Final Model:



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

Total: 100 points **# correct:** _____ **%:** _____