# YuHW07ST430

Jerry Yu

2023-11-20

```
ucars <- read.csv("Datasets/Used_Cars.txt") %>% as_tibble()
```

## Question 1

### a) Interpret each of the coefficients in the final model using proper units and a suitable increment (i.e. a 1-unit increase might be too small to consider for some of the terms in your model). Also find and discuss a 95% CI for each predictor variables.

```
attach(ucars)
ucarsfm <- lm(Asking.Price~Mileage + Price.New + Avg.Retail)
summary(ucarsfm)
```

```
##
## Call:
## lm(formula = Asking.Price ~ Mileage + Price.New + Avg.Retail)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3852.4  -734.4   -85.0   552.9  3743.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2617.16978  311.93128   8.390 7.39e-15 ***
## Mileage      -29.04579    3.02708  -9.595  < 2e-16 ***
## Price.New      0.05743    0.02131   2.695  0.00761 **
## Avg.Retail     0.75560    0.03620  20.873  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1244 on 208 degrees of freedom
## Multiple R-squared:  0.9115, Adjusted R-squared:  0.9103
## F-statistic: 714.4 on 3 and 208 DF,  p-value: < 2.2e-16
```

```
avis <- confint(ucarsfm, level = 0.95) %>% as.data.frame()
```

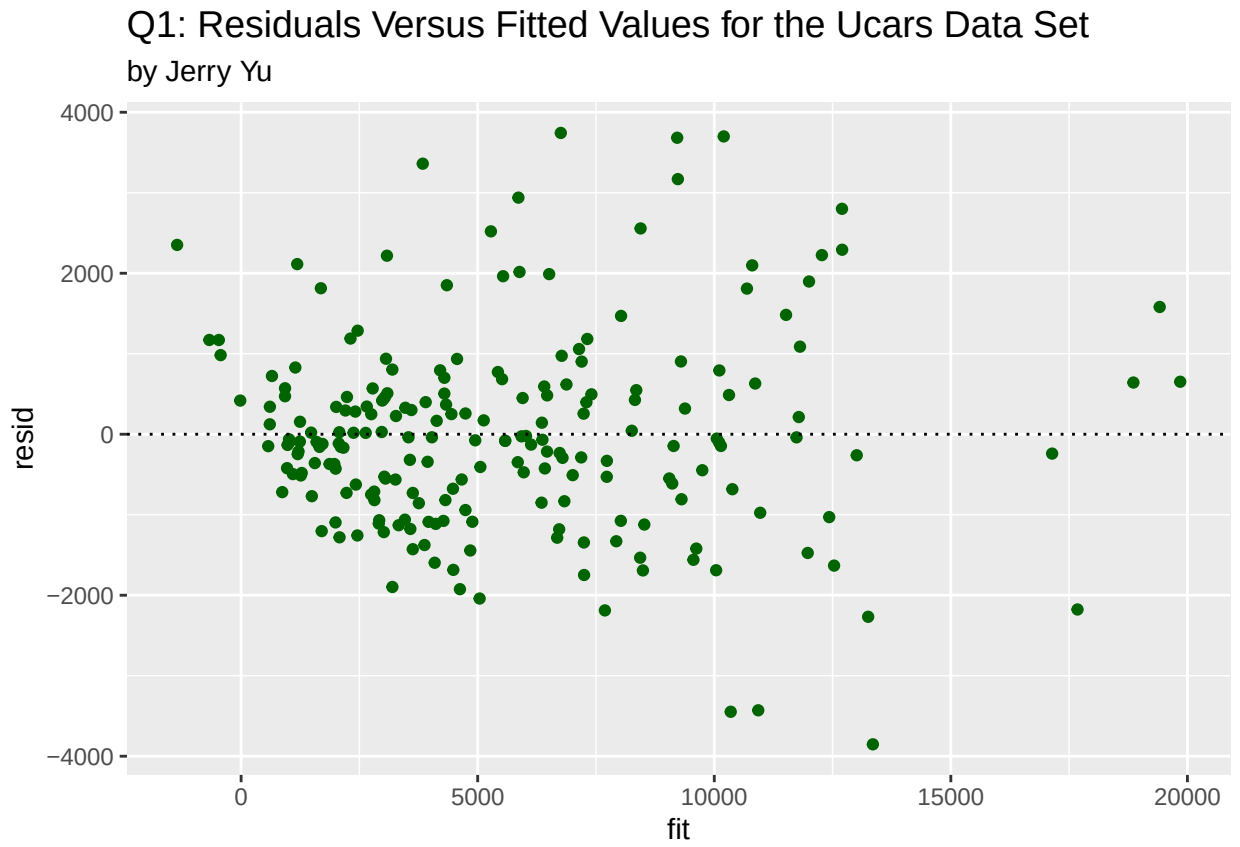
- At a given new price of a car and a given remaining loan value, an increase in 1000 miles driven on the car results in a $29.0457946 decrease in the asking price of a car.

- The 95% CI for Mileage is (-35.0134753,-23.078114). This means that we are 95% confident that the true mean slope of the Mileage variable is within the confidence interval.
- At a given Mileage and a given Average Retail Price, an increase of 1000 dollars of the new price of the cars results in a $57.4289987 increase in the asking price of a car.
- The 95% CI for Price when New is (0.0154229,0.0994351). This means that we are 95% confident that the true mean slope of the Car Price When New variable is within the confidence interval.
- At a given new price of a car and a given mileage, an increase in 100 dollars in the average retail price of the new car results in a $75.5603992 increase in the asking price of a car.
- The 95% CI for Average Retail is (0.6842373,0.8269706). This means that we are 95% confident that the true mean slope of the Average New Retail Price variable is within the confidence interval.

## b) Examine residual plots and a normal quantile plot of and comment on the adequacy of your model.

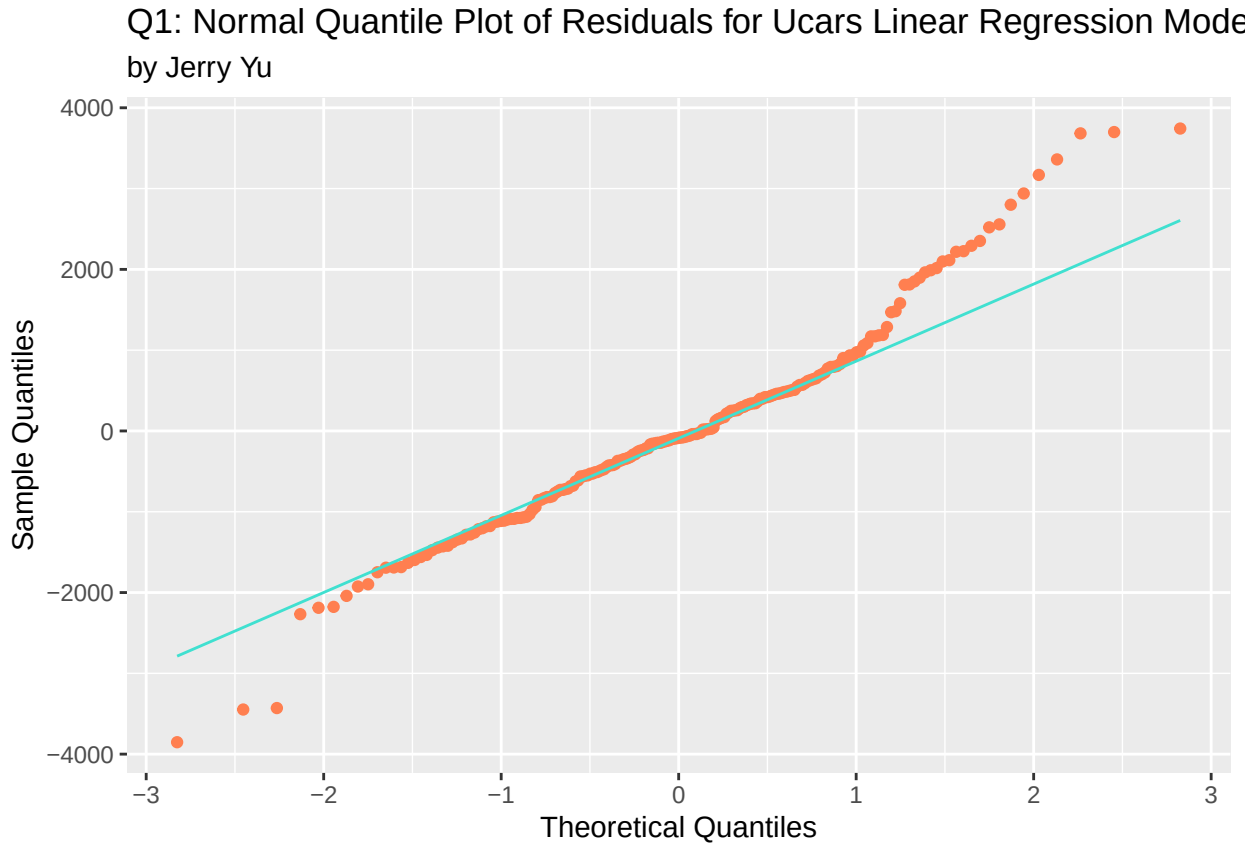
```
ucarsfmr <- tibble(
  "fit" = ucarsfm$fitted.values,
  "resid" = ucarsfm$residuals
)

ggplot(ucarsfmr,aes(x=fit,y=resid))+
  geom_jitter(color="darkgreen")+
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Q1: Residuals Versus Fitted Values for the Ucars Data Set"),
         subtitle = "by Jerry Yu")+
  theme(plot.title = element_text(size = 14))
```




```
ucarsfmrs <- add_column(ucarsfmr, "rstandardized"=rstandard(ucarsfm))

ggplot(data = ucarsfmrs, aes(sample = resid))+
  geom_qq( color="coral")+
  geom_qq_line( color="turquoise")+
  labs(
    title = paste("Q1: Normal Quantile Plot of Residuals for Ucars Linear Regression Model"),
    subtitle = "by Jerry Yu"
  ) +
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")
```




Our Model does not seem adequate. There seems to be a slight fan shaped distribution in the patterns of the residuals on the residuals in the residual plot. This indicates potential non constant error variation which is a violation of the assumption of constant error needed for linear regression. Additionally, the extreme residuals of the residual plot does not look like the theoritical residuals (the amplitude is higher).

### 3 ) Conduct Breusch-Pagan Test for the constancy of the error variance. Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

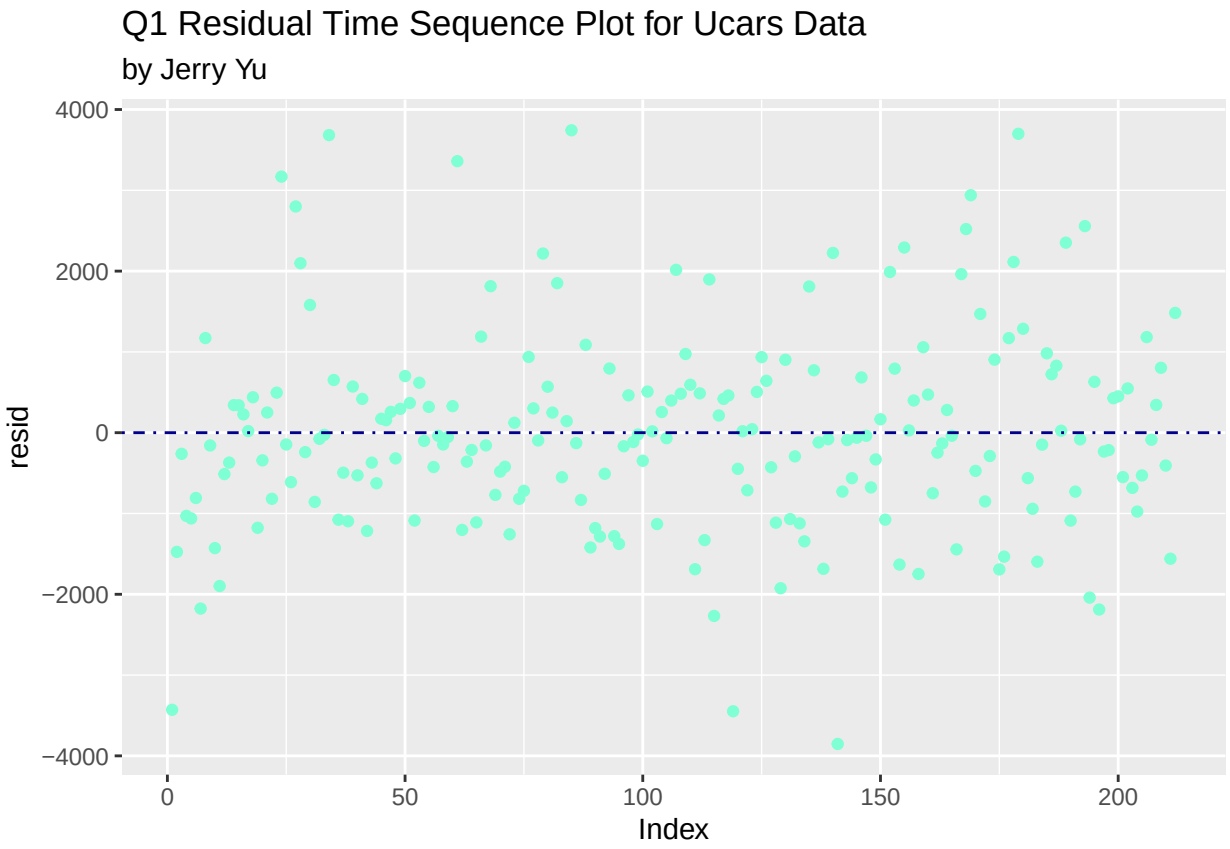
```
ucarsfmbp <- bptest(ucarsfm,studentize = FALSE)
ucarsfmbp
```

```
##
##  Breusch-Pagan test
##
## data:  ucarsfm
## BP = 32.748, df = 3, p-value = 3.641e-07
```

- H0: Equal Variance Among Errors (Homoscedasticity)
- HA: Unequal Variance Among Errors (Heteroscedasticity)
- Statistic: 32.7476717
- Df: 3
- P Value: $3.6405368 \times 10^{-7}$
- Conclusion: Reject H0

### d) Index Plot to test for Independence of errors and write your comments.

```
ggplot(ucarsfmr, aes(x = 1:length(resid), y = resid)) +
  geom_point(color = "aquamarine") +
  labs(x = "Index",
       title = "Q1 Residual Time Sequence Plot for Ucars Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "dotdash")
```




The spread of the errors does not seem to have a pattern. Thus there is no evidence to support the claim that the errors are not independent.

### e) Conduct Durbin-Watson Test. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value.

```
ucarsfmw <-durbinWatsonTest(ucarsfm)
ucarsfmw
```

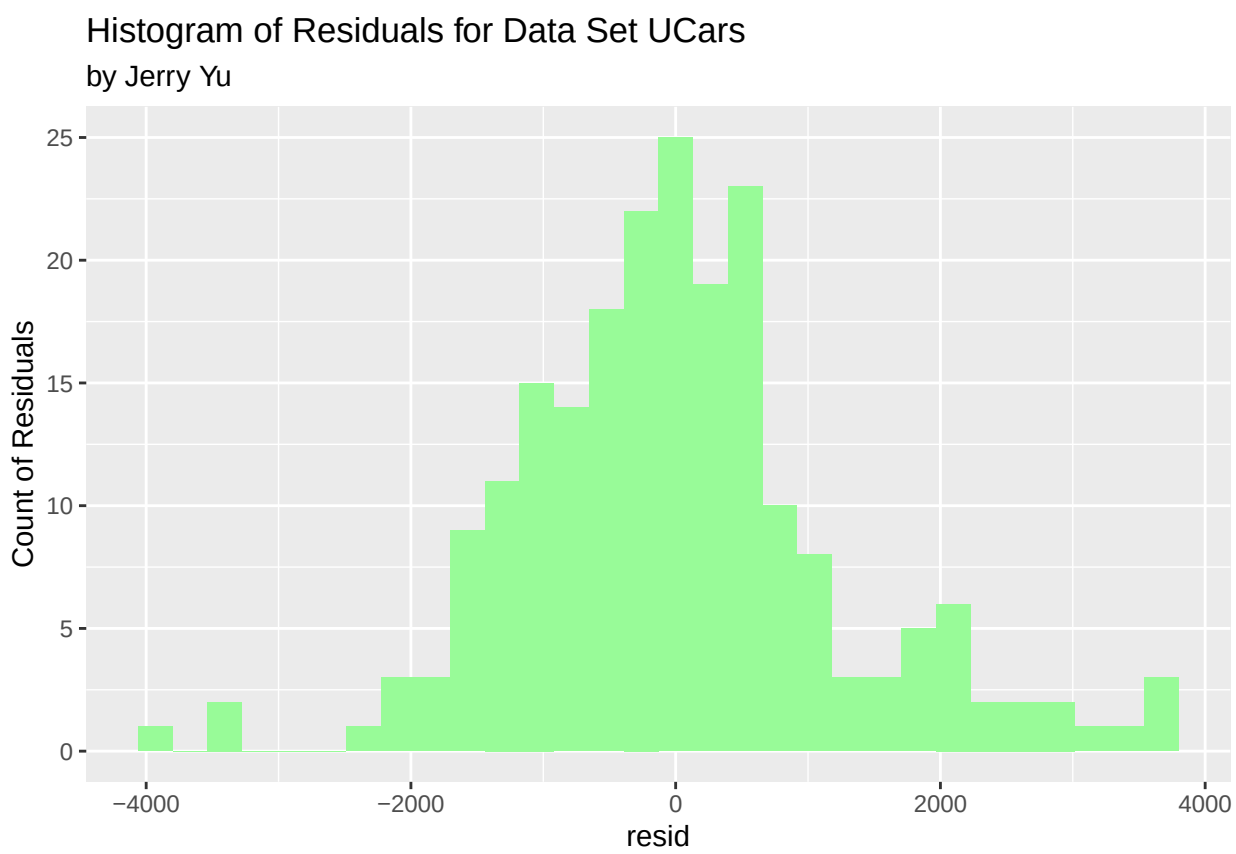
```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01638336      1.923891   0.496
##  Alternative hypothesis: rho != 0
```

- H0: Errors are independent. (autocorrelation = 0)
- HA: Errors are not independent. (autocorrelation $\neq$ 0)
- Statistic: 1.9238914
- P: 0.496
- Conclusion" Reject H0

### f) Give a Histogram of the residuals and write your comment.

```
ggplot(data = ucarsfmrs, aes(x = resid)) +
  geom_histogram(fill = "palegreen") +
  labs(
    title = paste("Histogram of Residuals for Data Set UCars"),
    subtitle = "by Jerry Yu"
  ) +
  ylab("Count of Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




There seems to be a right skew of the data, with there being more extreme values on the high side of the residuals, around 4000. This indicates that there are likely outliers.

**g) Conduct a Shapiro-Wilk Test on the residuals. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value. Give the p-value for this test and explain what this means in terms of our model assumptions.**

```
shap1 <- shapiro.test(ucarsfmrs$resid)
shap1
```

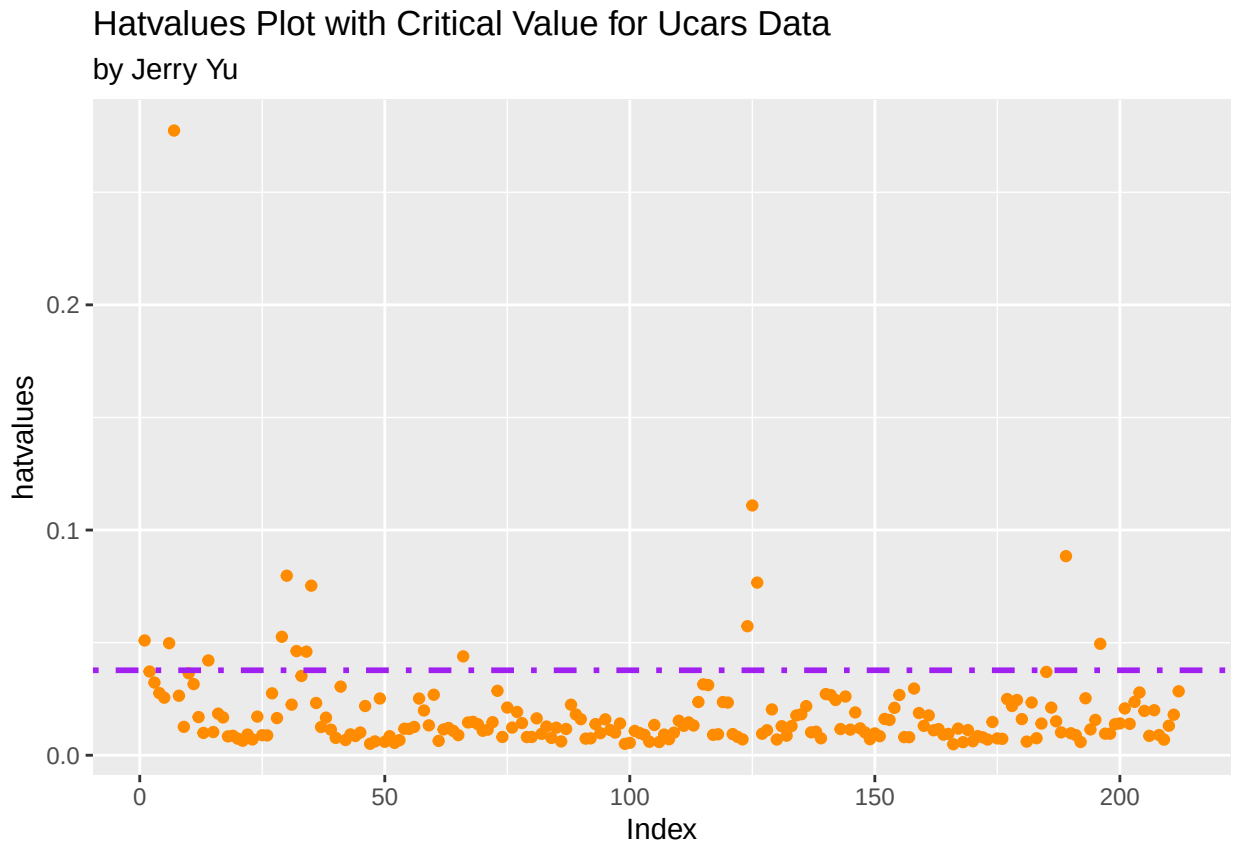
```
##
##  Shapiro-Wilk normality test
##
## data:  ucarsfmrs$resid
## W = 0.96918, p-value = 0.0001376
```

- H0:The random error in our model is normally distributed.
- HA: The random error in our model is not normally distributed.
- Statistic: 0.9691829
- P: $1.3764327 \times 10^{-4}$
- Conclusion: As $1.3764327 \times 10^{-4} < 0.05$, at $\alpha = 0.05$ we conclude that there is evidence to support that the distribution of random error (residuals) in our model is not normal.

**h) Check for large leverage points and identify the row numbers.**

```
# find critical value
crit <- 2*summary(ucarsfm)$coeff %>% nrow() / nrow(ucars)
# derive row numbers and hat values
ucarsfml <- add_column(ucars,"hatvalues"=hatvalues(ucarsfm),
                       "rownum" = rownames(ucars))
# plot
ggplot(ucarsfml, aes(x = 1:length(hatvalues), y = hatvalues)) +
  geom_point(color = "darkorange") +
  labs(x = "Hat Values",
       title = "Hatvalues Plot with Critical Value for Ucars Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = crit,
             color = "purple",
             linetype = "dotdash",
             size = 1) +
  xlab("Index")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

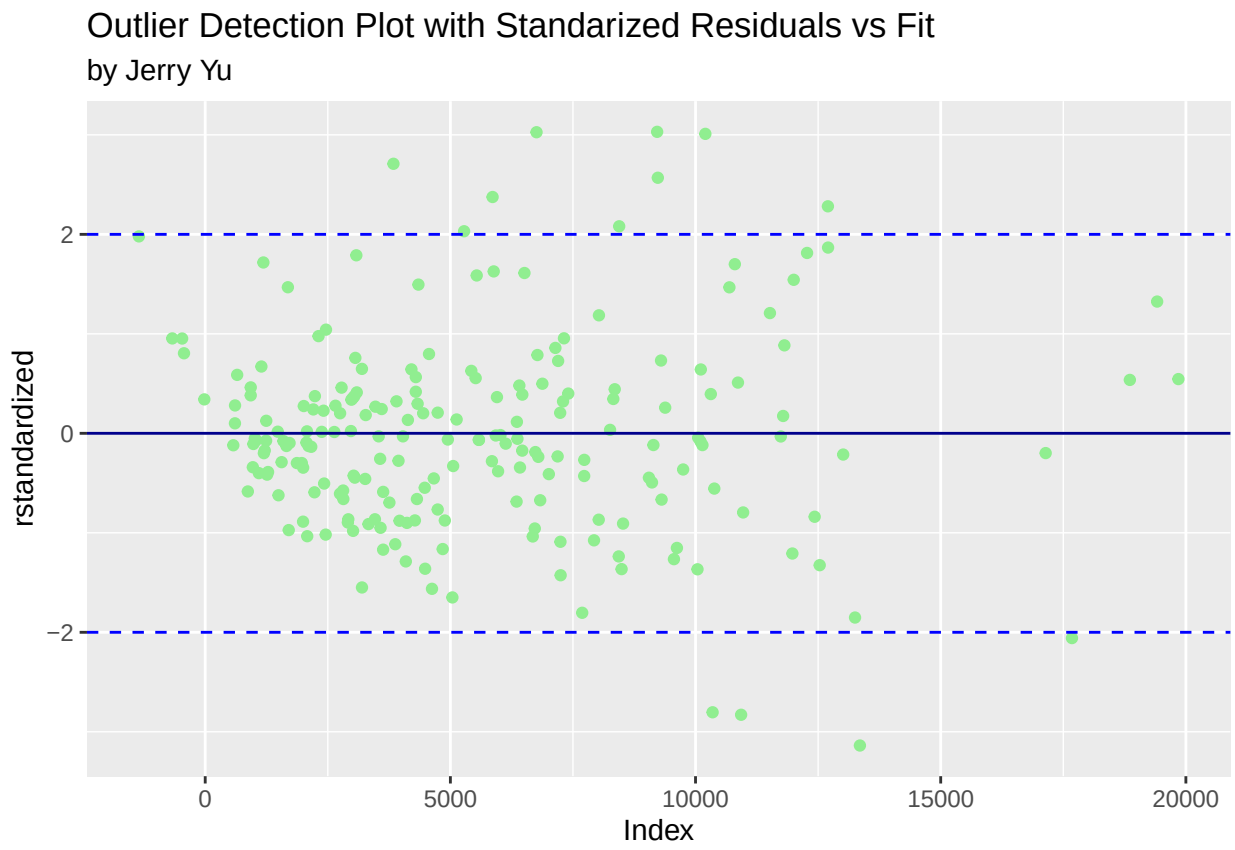



```
# print row numbers for large leverage points
ucarsfml %>% subset(hatvalues> crit) %>% select(c("rownum"))
```

```
## # A tibble: 15 x 1
##    rownum
##    <chr>
##  1 1
##  2 6
##  3 7
##  4 14
##  5 29
##  6 30
##  7 32
##  8 34
##  9 35
## 10 66
## 11 124
## 12 125
## 13 126
## 14 189
## 15 196
```

### i) Check for outliers and identify the row numbers

```
# derive row numbers
ucarsfmro <- add_column(ucars,"rstandardized"=rstandard(ucarsfm),
                        "rownum" = rownames(ucars))
# plot
ggplot(ucarsfmrs, aes(x = fit, y = rstandardized)) +
  geom_point(color = "lightgreen") +
  labs(x = "Index",
       title = "Outlier Detection Plot with Standarized Residuals vs Fit",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "solid") +
    geom_hline(yintercept = -2,
             color = "blue",
             linetype = "dashed")+
  geom_hline(yintercept = 2,
             color = "blue",
             linetype = "dashed")
```



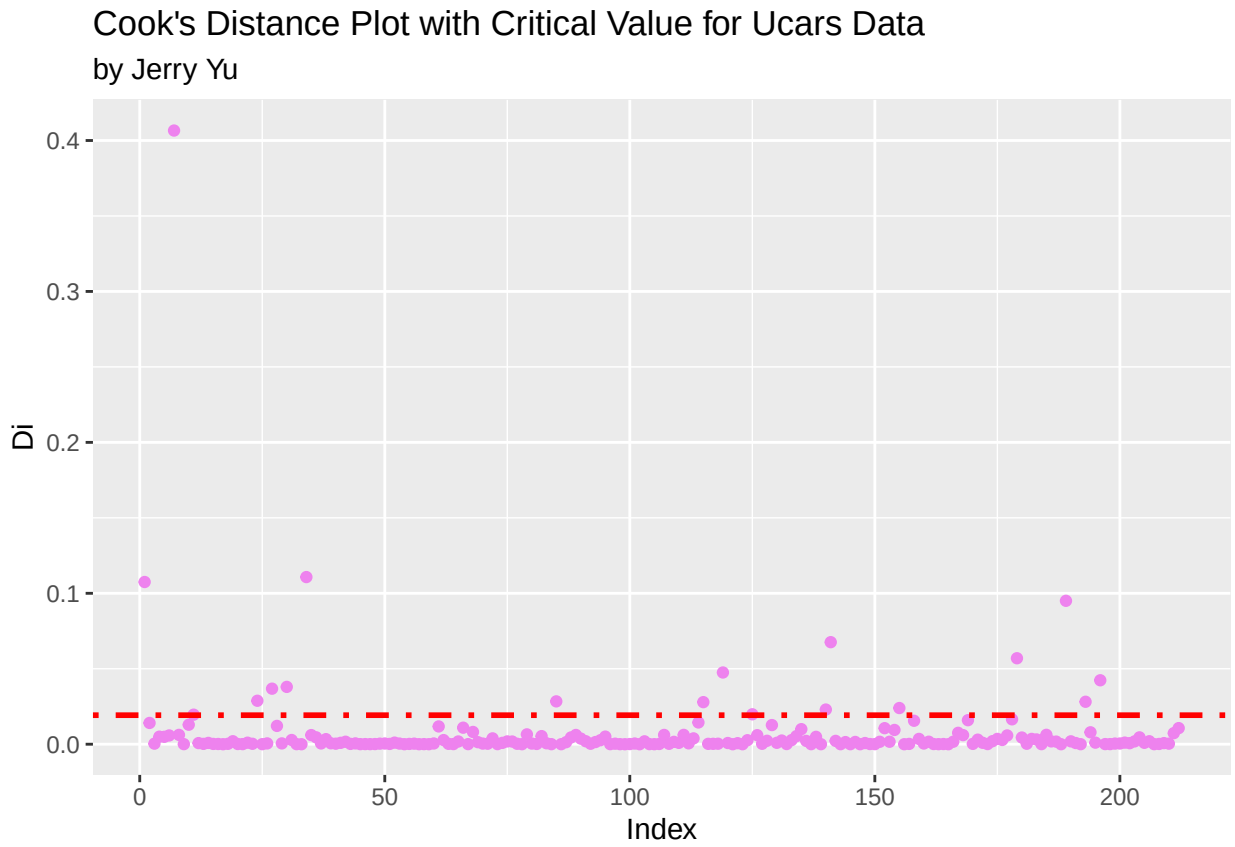

```
# print row numbers for large leverage points
ucarsfmro %>% subset(abs(rstandardized)> 2) %>% select(c("rownum"))
```

```
## # A tibble: 13 x 1
```

```
##    rownum
##    <chr>
##  1 1
##  2 7
##  3 24
##  4 27
##  5 34
##  6 61
##  7 85
##  8 119
##  9 141
## 10 168
## 11 169
## 12 179
## 13 193
```

### j) Check for influential points and identify the row numbers

```
# find cutoff value
cutoff <- with(ucarsfm,4/df.residual)
# derive row numbers and hat values
ucarsfmc <- add_column(ucars,"Di"=cooks.distance(ucarsfm),
                       "rownum" = rownames(ucars))
# plot
ggplot(ucarsfmc, aes(x = 1:length(Di), y = Di)) +
  geom_point(color = "violet") +
  labs(x = "Hat Values",
       title = "Cook's Distance Plot with Critical Value for Ucars Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = cutoff,
             color = "red",
             linetype = "dotdash",
             size = 1) +
  xlab("Index")
```




```
# print row numbers for large leverage points
ucarsfmc %>% subset(Di > cutoff) %>% select(c("rownum"))
```

```
## # A tibble: 18 x 1
##    rownum
##    <chr>
##  1 1
##  2 7
##  3 11
##  4 24
##  5 27
##  6 30
##  7 34
##  8 85
##  9 115
## 10 119
## 11 125
## 12 140
## 13 141
## 14 155
## 15 179
## 16 189
## 17 193
## 18 196
```

### k) Compute Variance inflation factors (VIF) and comment on the degree of collinearity.

```
vif(ucarsfm)
```

```
##    Mileage  Price.New Avg.Retail
##   1.567170   2.165178   2.960574
```

```
mean(vif(ucarsfm))
```

```
## [1] 2.230974
```

As none of the individual VIF values exceed 10, but the mean VIF exceeds 1, there is evidence of multicollinearity in our model, but no variable stands out as being strongly multicollinear with the other variables.

### l) Use your model to estimate the asking price for a car that was $10,000 when new, has 25,000 miles on it, and the average retail is $11,000.

```
samp <- tibble("Mileage"=25,"Price.New" = 10000,"Avg.Retail"=11000)

predict(ucarsfm,samp)
```

```
##        1
## 10776.96
```

Predicted Asking price is $1.0776959 \times 10^4$.

## Question 2

```
cpur <- model.matrix(~.,read.csv("Datasets/Car_Purchase.txt"))[,2:8] %>% as_tibble()
```

### a. Find the correlation matrix and comment on the use of the correlation (r) as a measure of linear association between the response (Y) and the individual X_j's.

```
cor(cpur)
```

```
##                 GenderM     Income        Age   MaritalS   Num.Kids
## GenderM      1.00000000  0.4249836  0.2885128 -0.3434898  0.3013376
## Income       0.42498357  1.0000000  0.7460186 -0.5598847  0.6841437
## Age          0.28851282  0.7460186  1.0000000 -0.6363763  0.7201584
## MaritalS    -0.34348982 -0.5598847 -0.6363763  1.0000000 -0.7861681
## Num.Kids     0.30133761  0.6841437  0.7201584 -0.7861681  1.0000000
```

```
## College.DegYes -0.07059781  0.3516438  0.1621570 -0.2242843  0.1982743
## Price.Paid      0.43560735  0.8751362  0.6154472 -0.4781330  0.5599724
##                College.DegYes Price.Paid
## GenderM           -0.07059781  0.4356073
## Income             0.35164377  0.8751362
## Age                0.16215702  0.6154472
## MaritalS          -0.22428433 -0.4781330
## Num.Kids           0.19827431  0.5599724
## College.DegYes     1.00000000  0.4595652
## Price.Paid         0.45956525  1.0000000
```

We see high positive correlation between Price Paid and Income 0.8751362, and a moderate positive correlation with Age 0.6154472. The dummy variables for the categorical variables all appear to have relatively weak correlations.

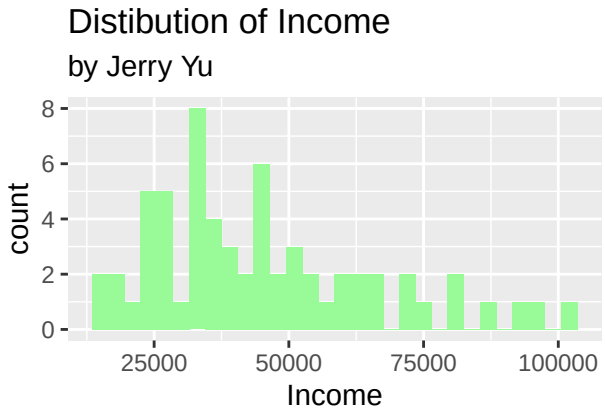
### b. Examine a scatterplot matrix of Price.Paid (Y) and the numeric predictors.Comment on anything interesting you find by examining this plot. You should comment/address the following:

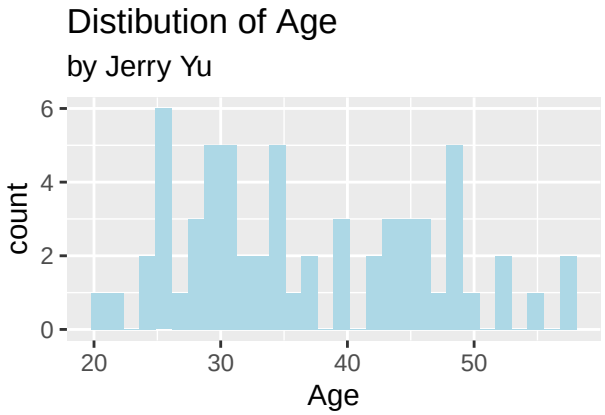
- marginal/univariate distributions of Y and Xj's.
- relationships between Y vs.Xj's
- relationships between Xj's
- any unusual cases

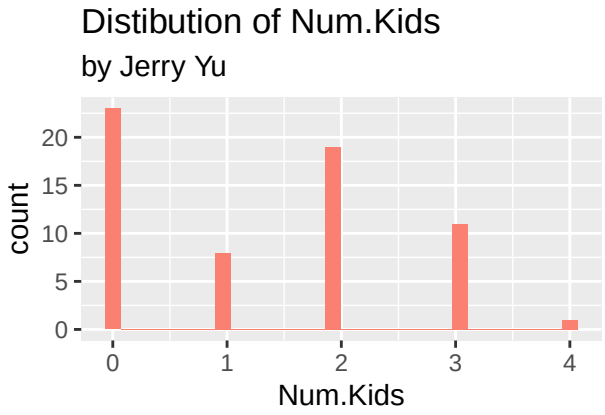
```
cpurn <- cpur %>% select(c(Income,Age,Num.Kids,Price.Paid))
p1 <- ggplot(data = cpurn, aes(x = Income)) +
  geom_histogram(fill = "palegreen") +
  labs(
    title = paste("Distibution of Income"),
    subtitle = "by Jerry Yu"
  )
p2 <- ggplot(data = cpurn, aes(x = Age)) +
  geom_histogram(fill = "lightblue") +
  labs(
    title = paste("Distibution of Age"),
    subtitle = "by Jerry Yu"
  )
p3 <- ggplot(data = cpurn, aes(x = Num.Kids)) +
  geom_histogram(fill = "salmon") +
  labs(
    title = paste("Distibution of Num.Kids"),
    subtitle = "by Jerry Yu"
  )
p4 <- ggplot(data = cpurn, aes(x = Price.Paid)) +
  geom_histogram(fill = "violet") +
  labs(
    title = paste("Distibution of Price.Paid"),
    subtitle = "by Jerry Yu"
  )
grid.arrange(p1,p2,p3,p4)
```

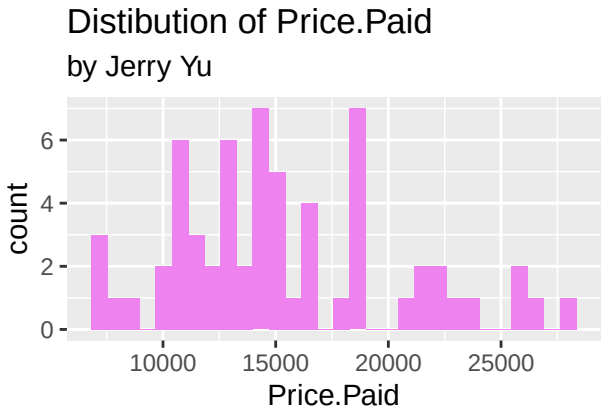
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

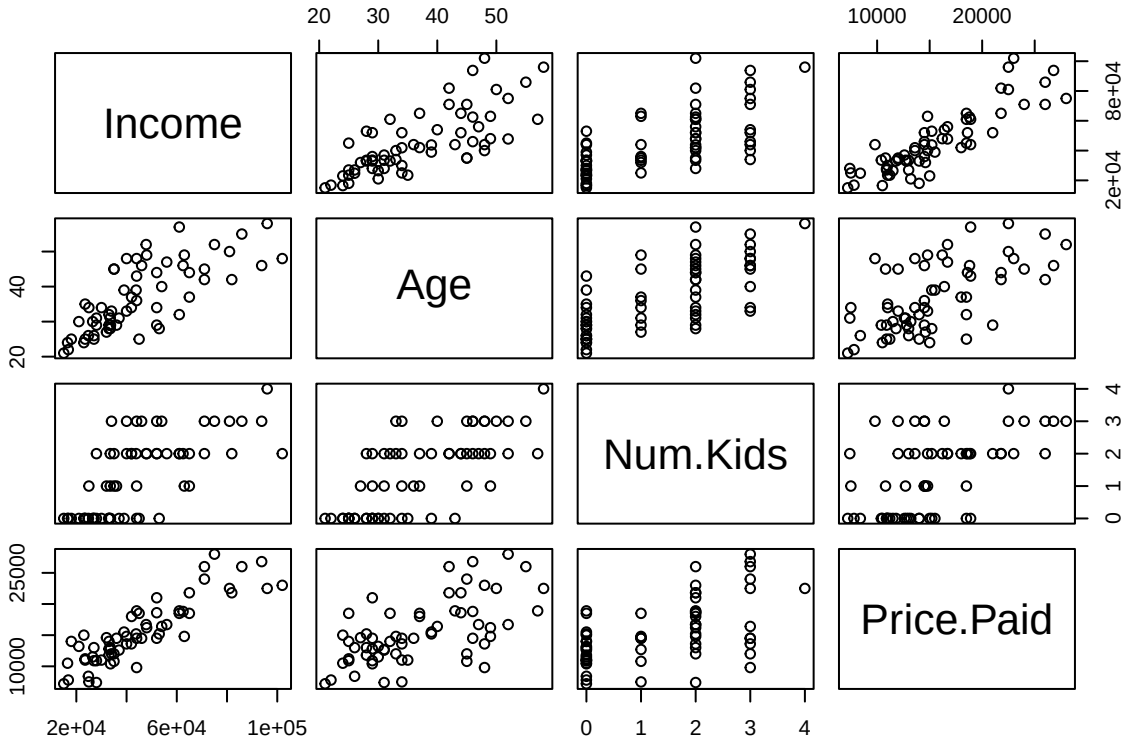



```
pairs(cpurn)
```




Looking At the Marginal/Univariate Scatterplots of the Variables (not required, but I created them to better analyze the data) income has a right skew and Number of Kids has a left skew (as well as being discrete data). Age is a bit more spread out, as is Price Paid, though Price Paid also has a slight right skew.

Moving on to the Scatterplot Matrix, there seem to be varying levels of positive correlation between Price Paid (Y) and all of the predictor variables, the strongest being Income. This matches our correlation matrix. Meanwhile, there also seem to be positive correlations between the different X variables. Basically every graph in the matrix appears to show some level of positive linear correlation. This could be an indicator of multicollinearity. There do not seem to be any unusual cases in the scatterplot matrix or the single distributions.

### c. Write out the full model using all available predictors.

A full first order linear model of Car Purchase would be

Price Paid $= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$

- Where $X_1$ = GenderM, a dummy variable where is Gender is Male GenderM=1
- Where $X_2$ = Annual Income is $
- Where $X_3$ = Age in Years
- Where $X_4$ = MaritalS, a dummy variable where if Marital is Single then MaritalS =1
- Where $X_5$ = Number of Children
- Where $X_6$ = College.DegYes, a dummy variable where if College.Deg is yes then College.DegYes =1

### d. Which predictors/terms in model are statistically significant at the =0.05 level?

```
cpurm <- lm(Price.Paid~ GenderM + Income + Age + MaritalS + Num.Kids + College.DegYes,cpur)
summary(cpurm)
```

```
##
## Call:
## lm(formula = Price.Paid ~ GenderM + Income + Age + MaritalS +
##     Num.Kids + College.DegYes, data = cpur)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5236.2 -1162.3  -467.4  1657.6  6053.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4599.49206 1886.41569   2.438  0.01802 *
## GenderM        1504.23087  770.16293   1.953  0.05590 .
## Income            0.18950    0.02623   7.225 1.61e-09 ***
## Age               1.20892   52.88467   0.023  0.98185
## MaritalS        219.56769 1025.81795   0.214  0.83131
## Num.Kids       -183.68338  475.20535  -0.387  0.70059
## College.DegYes 2118.26644  695.32550   3.046  0.00355 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2327 on 55 degrees of freedom
## Multiple R-squared:  0.8078, Adjusted R-squared:  0.7868
## F-statistic: 38.52 on 6 and 55 DF,  p-value: < 2.2e-16
```

According to the individual T tests, the 2 significant variables at $\alpha = 0.05$ are Income, and College.DegYes.

### e. Conduct a F-test for removing the insignificant terms from the full model.

```
avis2 <- matrix(c(0,1,0,0,0,0,0,
                  0,0,0,1,0,0,0,
                  0,0,0,0,1,0,0,
                  0,0,0,0,0,1,0),
                   nrow=4,
                   byrow = TRUE)
ftest(cpurm,avis2)
```

```
##          F        df1        df2    p-value
##  1.1149923  4.0000000 55.0000000  0.3588412
```

At $\alpha = 0.05$, p $> \alpha$, so we fail to reject the null hypothesis and conclude that there is no statistically significant difference between the full and the reduced model, and thus go with the simpler reduced model.

### f. Interpret each of the parameter estimates in the final model using proper units and increments.

```
cpurf <- lm(Price.Paid ~ Income + College.DegYes,cpur)
summary(cpurf)
```

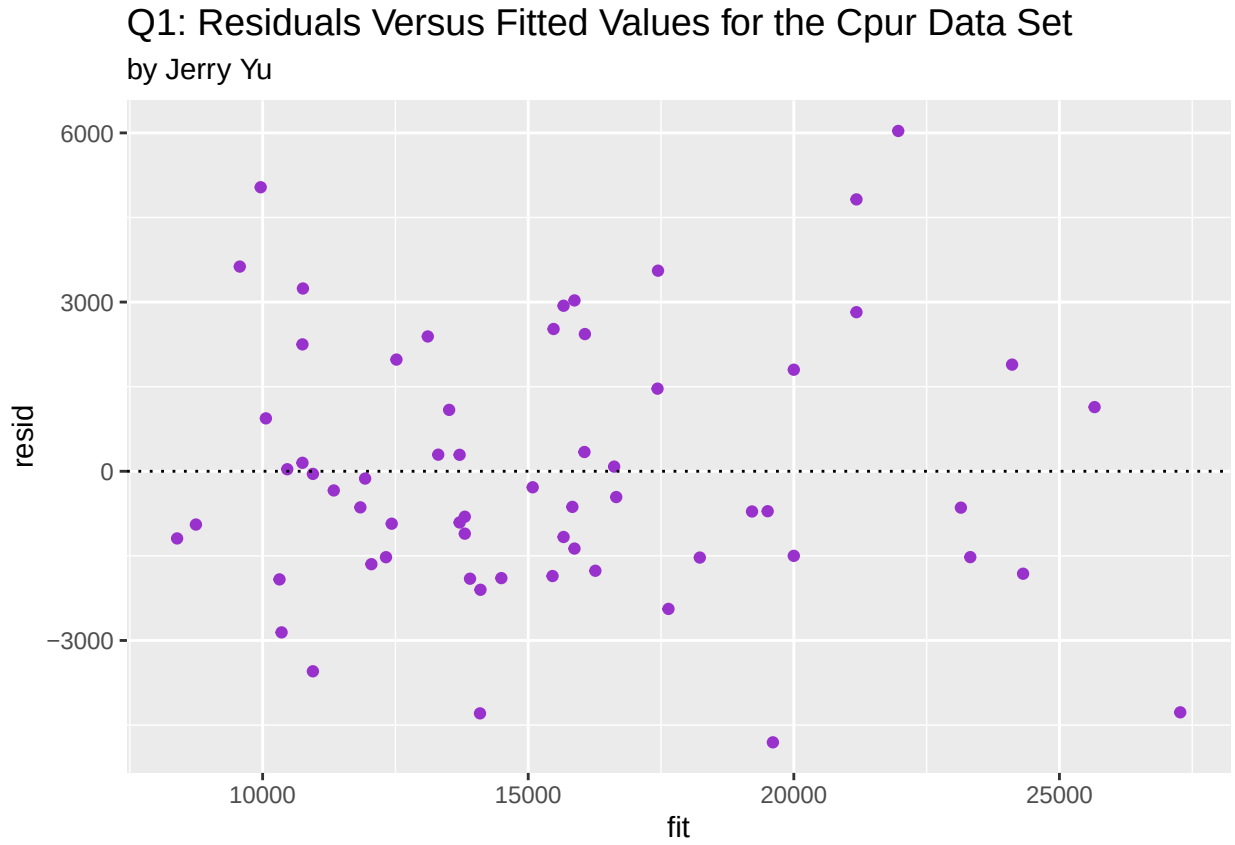
```
##
## Call:
## lm(formula = Price.Paid ~ Income + College.DegYes, data = cpur)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4806  -1522   -544   1717   6034
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.441e+03  7.292e+02   7.462 4.46e-10 ***
## Income         1.966e-01  1.531e-02  12.842  < 2e-16 ***
## College.DegYes 1.778e+03  6.508e+02   2.733  0.00828 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2336 on 59 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7851
## F-statistic: 112.4 on 2 and 59 DF,  p-value: < 2.2e-16
```

- At a given College Degree (either Yes or No), an increase in 1000 dollars in Annual Income, Average Car Purchase Price will increase by 196.6124715 dollars.
- At a Given Annual Income, people who gained a college degree will on average spend 1778.309332 dollars more on a car than people without a college degree.

### g. Examine residual plots and a normal quantile plot and comment on the adequacy of your model.

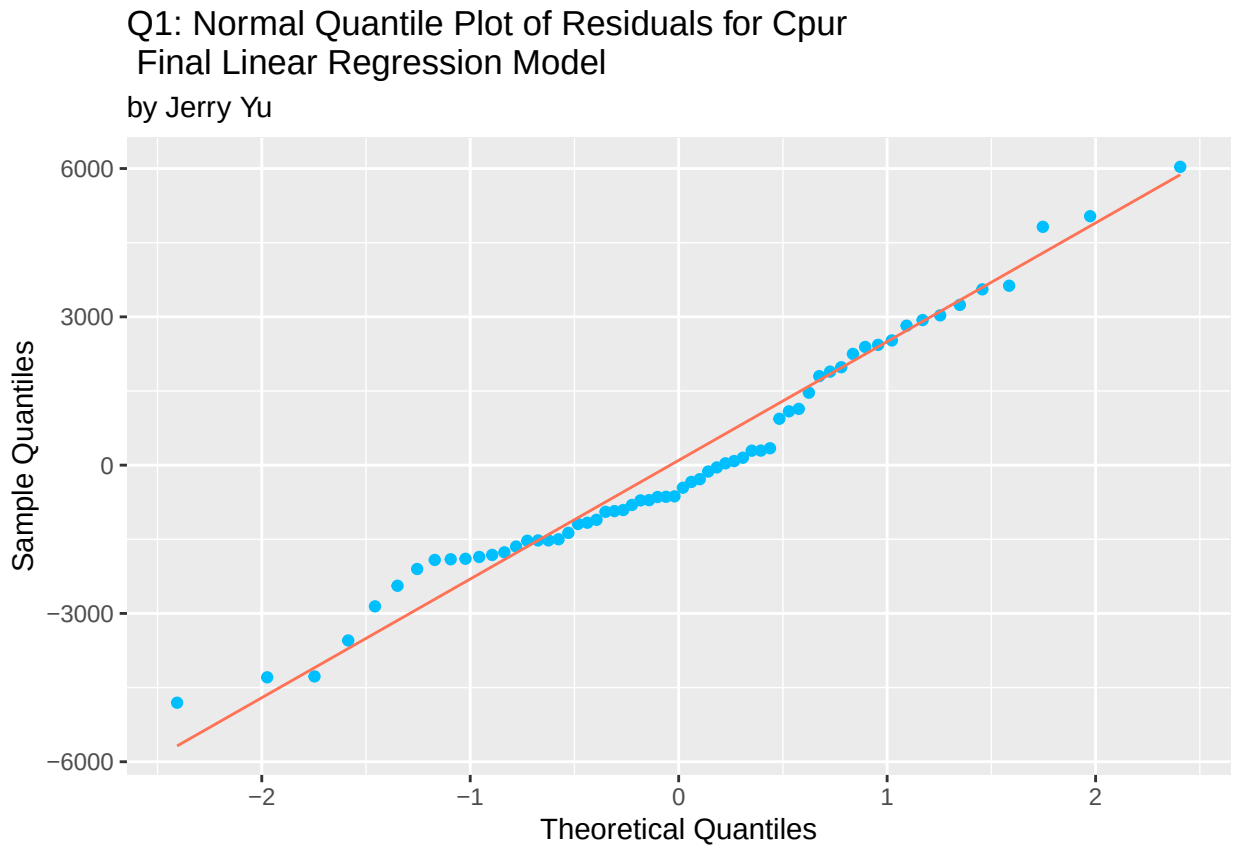
```
cpurfr <- tibble(
  "fit" = cpurf$fitted.values,
  "resid" = cpurf$residuals
)

ggplot(cpurfr,aes(x=fit,y=resid))+
  geom_jitter(color="darkorchid")+
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Q1: Residuals Versus Fitted Values for the Cpur Data Set"),
         subtitle = "by Jerry Yu")+
  theme(plot.title = element_text(size = 14))
```




```
cpurfrs <- add_column(cpurfr, "rstandardized"=rstandard(cpurf))

ggplot(data = cpurfrs, aes(sample = resid))+
  geom_qq( color="deepskyblue")+
  geom_qq_line( color="coral1")+
  labs(
    title = paste("Q1: Normal Quantile Plot of Residuals for Cpur \n Final Linear Regression Model"),
    subtitle = "by Jerry Yu"
  ) +
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")
```




There are no pattern or funnel shapes in the residual plots, so it unlikely that the error variance is linear and Homoscedastic (constant variance). The Normal Quantile Plot also does not have many values that deviate too much from the theoretical quantiles, so it likely that the distribution of the error is normal. Thus, since all of our assumptions are not broken, our model seems adquate at least in regards to linearity, constant variance (homoscedastic), and normality.

### h. Conduct Breusch-Pagan Test for the constancy of the error variance. Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

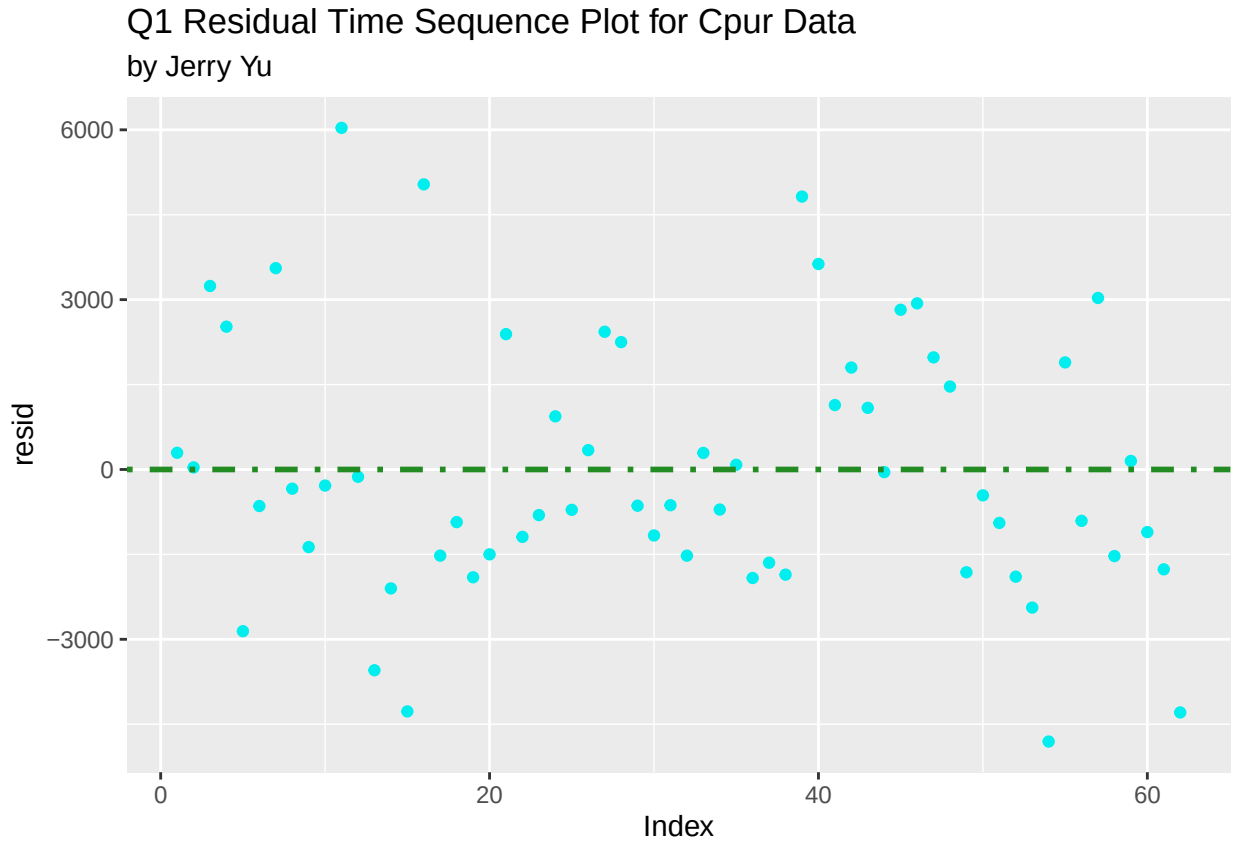
```
cpurfbp <- bptest(cpurf,studentize = FALSE)
cpurfbp
```

```
##
##  Breusch-Pagan test
##
## data:  cpurf
## BP = 3.0431, df = 2, p-value = 0.2184
```

- H0: Equal Variance Among Errors (Homoscedasticity)
- HA: Unequal Variance Among Errors (Heteroscedasticity)
- Statistic: 3.0430623
- Df: 2
- P Value: 0.2183773

**i. Index Plot to test for Independence of errors.**

```
ggplot(cpurfr, aes(x = 1:length(resid), y = resid)) +
  geom_point(color = "cyan2") +
  labs(x = "Index",
       title = "Q1 Residual Time Sequence Plot for Cpur Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "forestgreen",
             linetype = "dotdash",
             size=1)
```




**j. Conduct Durbin-Watson Test. Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.**

```
cpurfw <-durbinWatsonTest(cpurf)
cpurfw
```

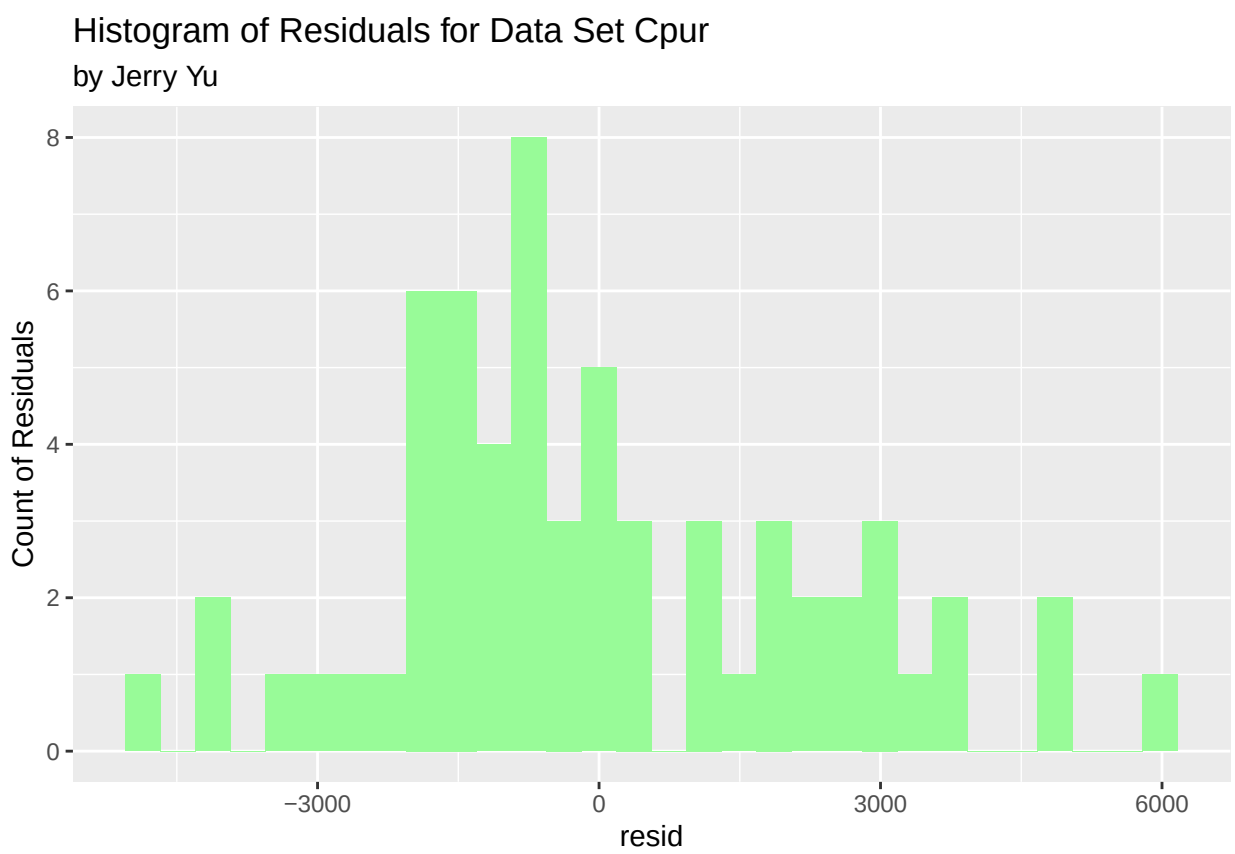
```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1186807      1.705172    0.23
##  Alternative hypothesis: rho != 0
```

- H0: Errors are independent. (autocorrelation = 0)
- HA: Errors are not independent. (autocorrelation ≠ 0)
- Statistic: 1.7051718
- Df: 59
- P: 0.23

### f) Give a Histogram of the residuals and write your comment.

```
ggplot(data = cpurfrs, aes(x = resid)) +
  geom_histogram(fill = "palegreen") +
  labs(
    title = paste("Histogram of Residuals for Data Set Cpur"),
    subtitle = "by Jerry Yu"
  ) +
  ylab("Count of Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




There seems to be a right skew of the residuals. There are what could possibly be a high and a low outlier with a residual value around -1600 and 6000, respectively.Overall though, the distirbution of the residuals still seems mostly normal, supporting our Normal Quantile Plot.

**l. Conduct a Shapiro-Wilk Test on the residuals. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value. Give the p-value for this test and explain what this means in terms of our model assumptions.**

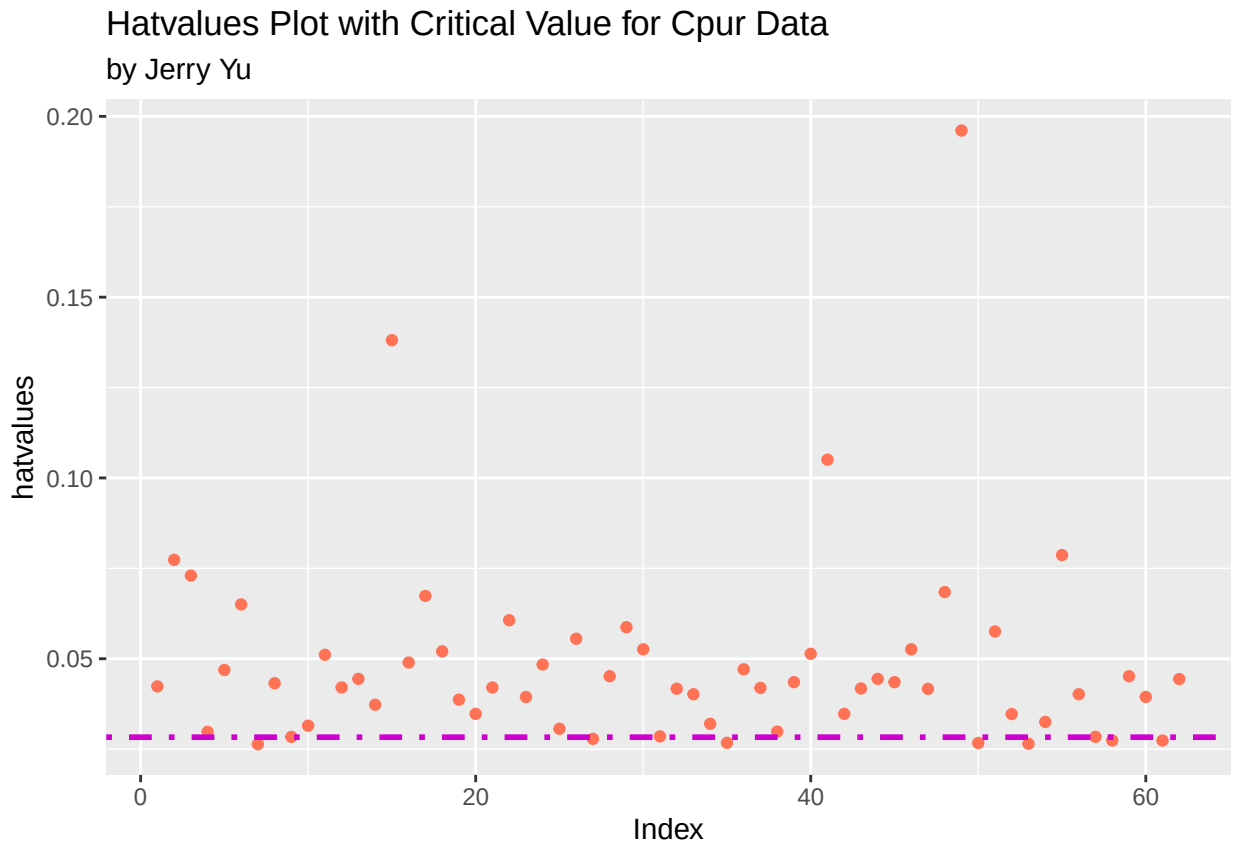
```
shap1 <- shapiro.test(cpurfrs$resid)
shap1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cpurfrs$resid
## W = 0.96903, p-value = 0.1188
```

- H0:The random error in our model is normally distributed.
- HA: The random error in our model is not normally distributed.
- Statistic: 0.9690348
- P: 0.1187514
- Conclusion: As $0.1187514 > 0.05$, at $\alpha = 0.05$ we conclude that there is no evidence to support that the distribution of random error (residuals) in our model is not normal.

**m. Check for large leverage points and identify the row numbers**

```
# find critical value
crit2 <- 2*summary(cpurf)$coeff %>% nrow() / nrow(ucars)
# derive row numbers and hat values
cpurfl <- add_column(cpur,"hatvalues"=hatvalues(cpurf),
                     "rownum" = rownames(cpur))
# plot
ggplot(cpurfl, aes(x = 1:length(hatvalues), y = hatvalues)) +
  geom_point(color = "coral1") +
  labs(x = "Hat Values",
       title = "Hatvalues Plot with Critical Value for Cpur Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = crit2,
             color = "magenta3",
             linetype = "dotdash",
             size = 1) +
  xlab("Index")
```




```
# print row numbers for large leverage points
cpurfl %>% subset(hatvalues> crit2) %>% select(c("rownum"))
```

```
## # A tibble: 55 x 1
##    rownum
##    <chr>
##  1 1
##  2 2
##  3 3
##  4 4
##  5 5
##  6 6
##  7 8
##  8 9
##  9 10
## 10 11
## # i 45 more rows
```

### n. Check for outliers

```
# derive row numbers
cpurfro <- add_column(cpur,"rstandardized"=rstandard(cpurf),
                      "rownum" = rownames(cpur))
```

```
# plot
ggplot(cpurfrs, aes(x = fit, y = rstandardized)) +
  geom_point(color = "aquamarine1") +
  labs(x = "Index",
       title = "Outlier Detection Plot with Standarized Residuals vs Fit",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "salmon1",
             linetype = "solid") +
    geom_hline(yintercept = -2,
             color = "orchid1",
             linetype = "dashed",
             size=1)+
  geom_hline(yintercept = 2,
             color = "orchid1",
             linetype = "dashed",
             size=1)
```

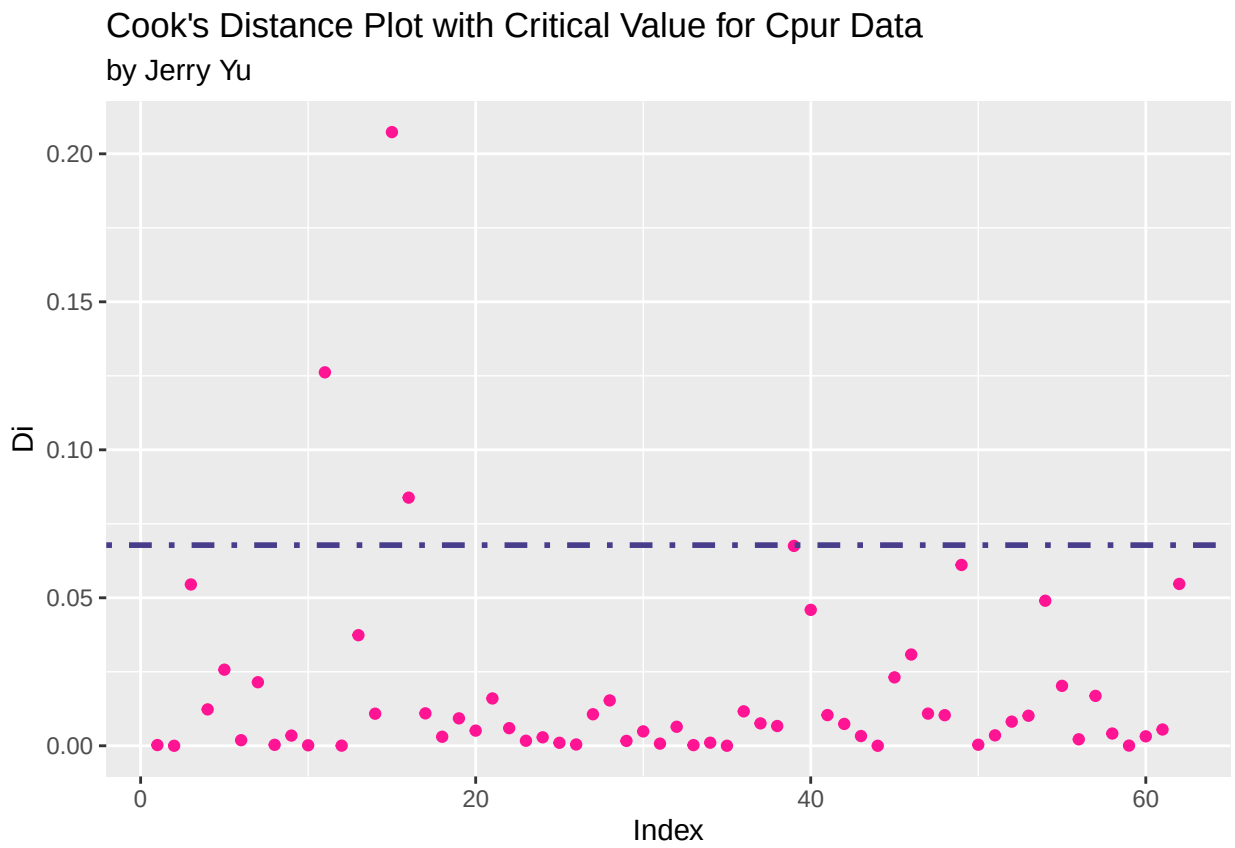

```
# print row numbers for large leverage points
cpurfro %>% subset(abs(rstandardized)> 2) %>% select(c("rownum"))
```

```
## # A tibble: 4 x 1
##   rownum
##   <chr>
## 1 11
```

```
## 2 16
## 3 39
## 4 54
```

### o. Check for influential points

```
# find cutoff value
cutoff2 <- with(cpurf,4/df.residual)
# derive row numbers and hat values
cpurfc <- add_column(cpur,"Di"=cooks.distance(cpurf),
                         "rownum" = rownames(cpur))
# plot
ggplot(cpurfc, aes(x = 1:length(Di), y = Di)) +
  geom_point(color = "deeppink") +
  labs(x = "Hat Values",
       title = "Cook's Distance Plot with Critical Value for Cpur Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = cutoff2,
             color = "darkslateblue",
             linetype = "dotdash",
             size = 1) +
  xlab("Index")
```



```r
# print row numbers for large leverage points
cpurfc %>% subset(Di > cutoff2) %>% select(c("rownum"))
```

```
## # A tibble: 3 x 1
##   rownum
##   <chr>
## 1 11
## 2 15
## 3 16
```

### p. Compute Variance inflation factors (VIF) and comment on the degree of collinearity.

```r
vif(cpurf)
```

```
##         Income College.DegYes
##       1.141101       1.141101
```

```r
mean(vif(cpurf))
```

```
## [1] 1.141101
```

As none of the individual VIF values exceed 10, but the mean VIF exceeds 1, there is some evidence of multicollinearity in our model, but no variable stands out as being strongly multicollinear with the other variables (the VIFs are the same)