# YuHW07ST430

Jerry Yu

2023-11-20

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

```
## # A tibble: 212 x 10
##    Asking.Price Year.Made Num.Options Mileage Price.New Loan.Val Avg.Retail
##           <int>     <int>       <int>   <int>     <int>    <int>      <int>
##  1         7500        88           4     100     26578     9200      12825
##  2        10500        88           4      64     26578     9200      12825
##  3        12750        89           4      54     24760    10100      13950
##  4        11400        91           4      29     13825     9725      13050
##  5         2400        85           1      97     17710     2025       3500
##  6         8500        89           4     113     25310     8075      11275
##  7        15500        88           1      50     53000    13325      17825
##  8          500        76           2     134      3935        0        500
##  9         1950        82           2      76      9504      450       1525
## 10         2200        82           2      50     14940      750       2125
## # i 202 more rows
## # i 3 more variables: Make <chr>, Model <chr>, Make.Model <chr>
```

```
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()
avis[2,2]
```

```
## [1] -23.07811
```

# Question 1

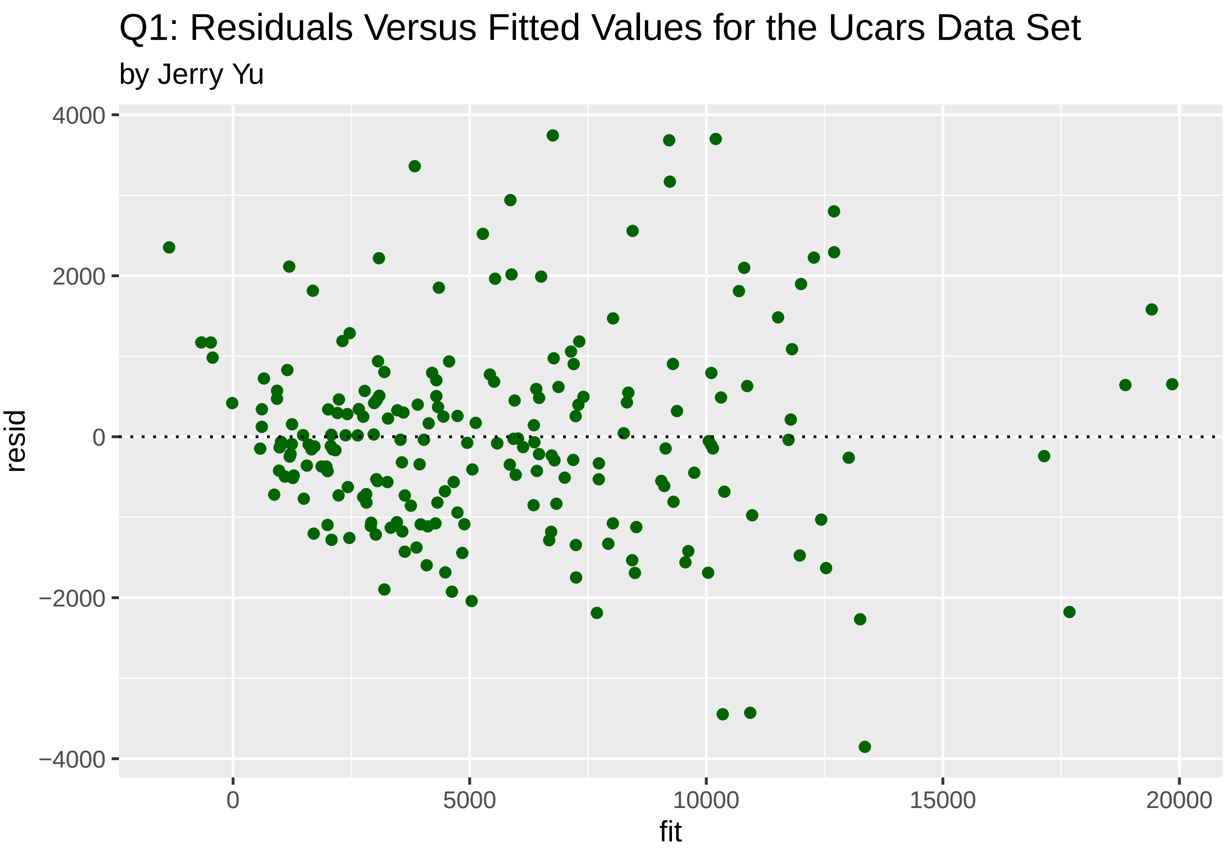
## a) Interpret each of the coefficients in the final model using proper units and a suitable increment (i.e. a 1-unit increase might be too small to consider for some of the terms in your model). Also find and discuss a 95% CI for each predictor variables.

- At a given Mileage and a given remaining loan value, an increase of 1000 dollars of the new price of the cars 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 new price of a car and a given remaining loan value, an increase in 1000 miles driven on the car results in a $57.4289987 decrease 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 decrease in the asking price of a car.
- The 95% CI for Average Retail is (-35.0134753,-23.078114). 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.

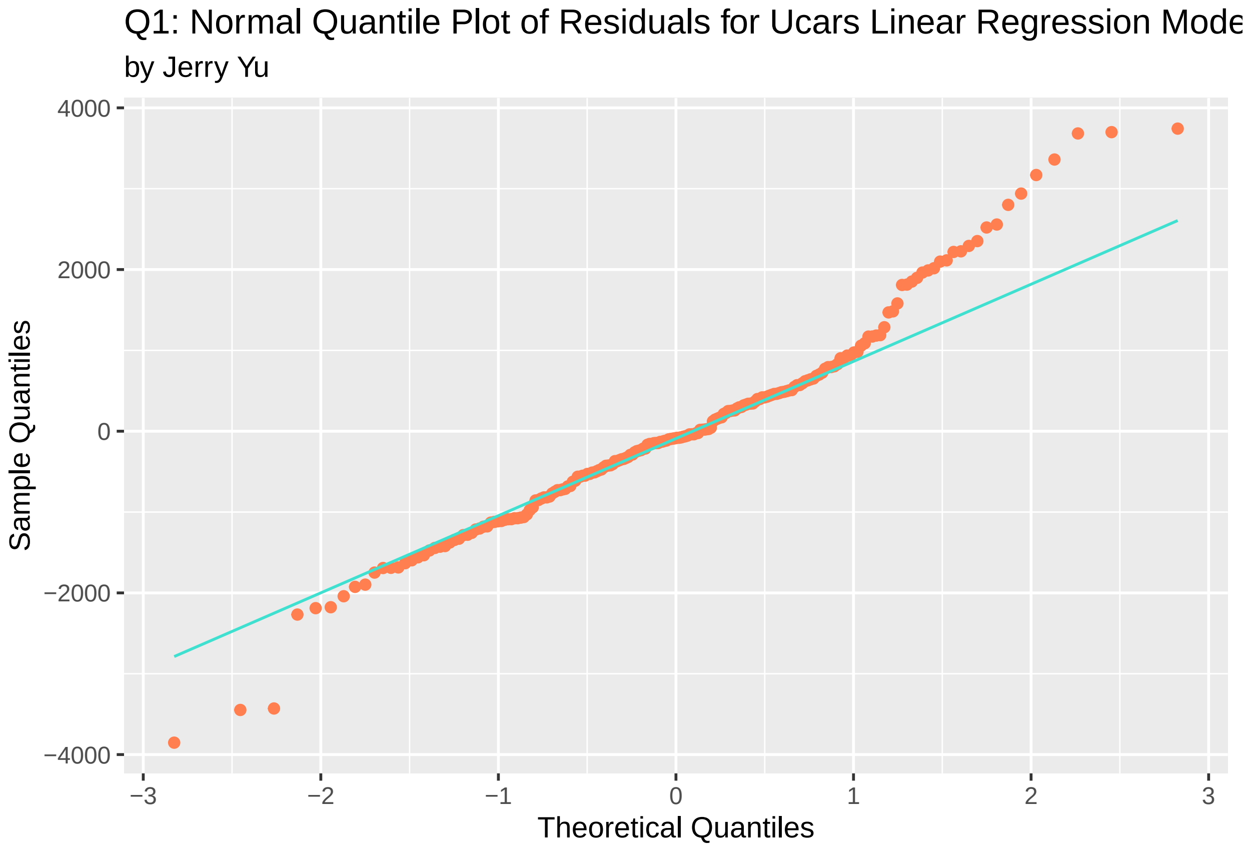
```
ucarsfmr <- tibble(
  "fit" = ucarsfm$fitted.values,
  "resid" = ucarsfm$residuals
)

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
```

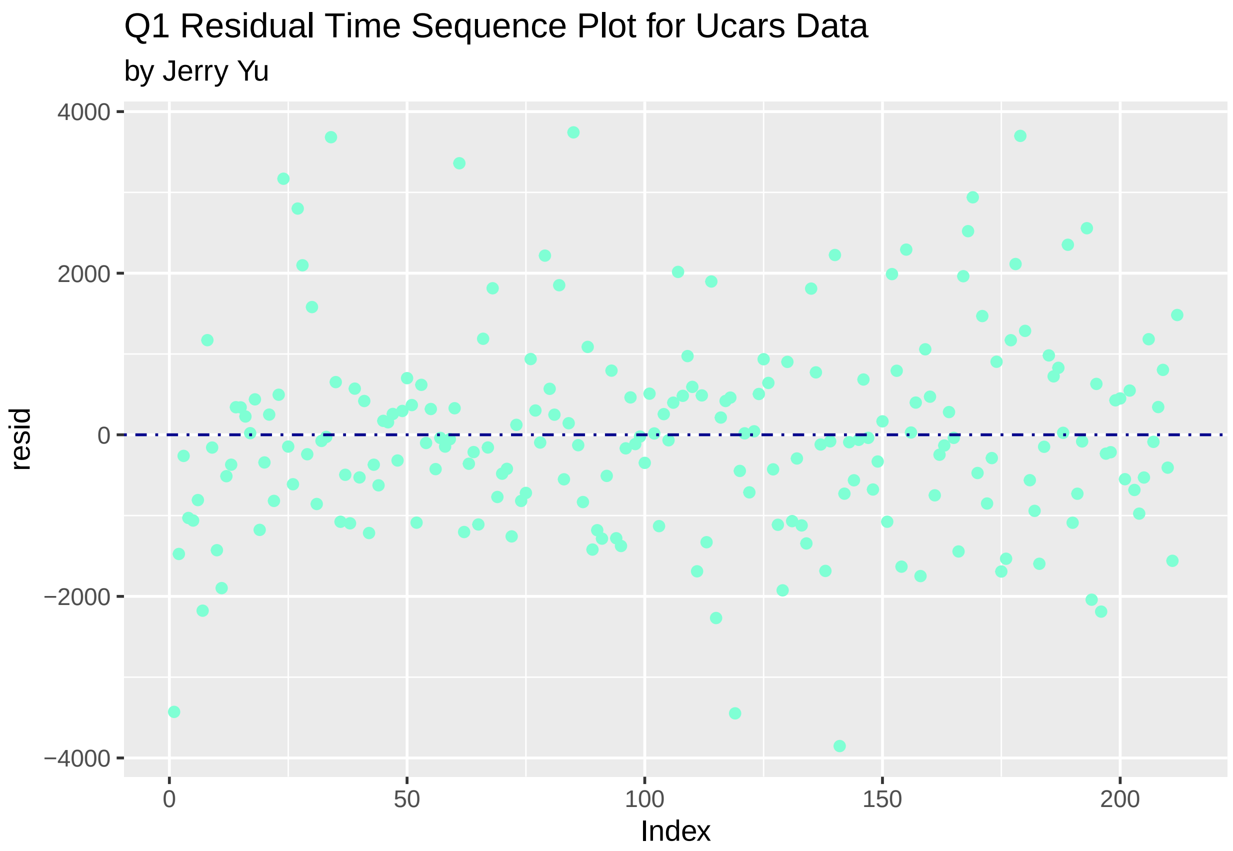
```
ucarsfmbp[[4]]
```

```
##           BP
## 3.640537e-07
```

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

### 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 erros are not independent.

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

```
dwtest(ucarsfm)
```

```
##
##  Durbin-Watson test
##
```

```
## data:  ucarsfm
## DW = 1.9239, p-value = 0.2633
## alternative hypothesis: true autocorrelation is greater than 0
```

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

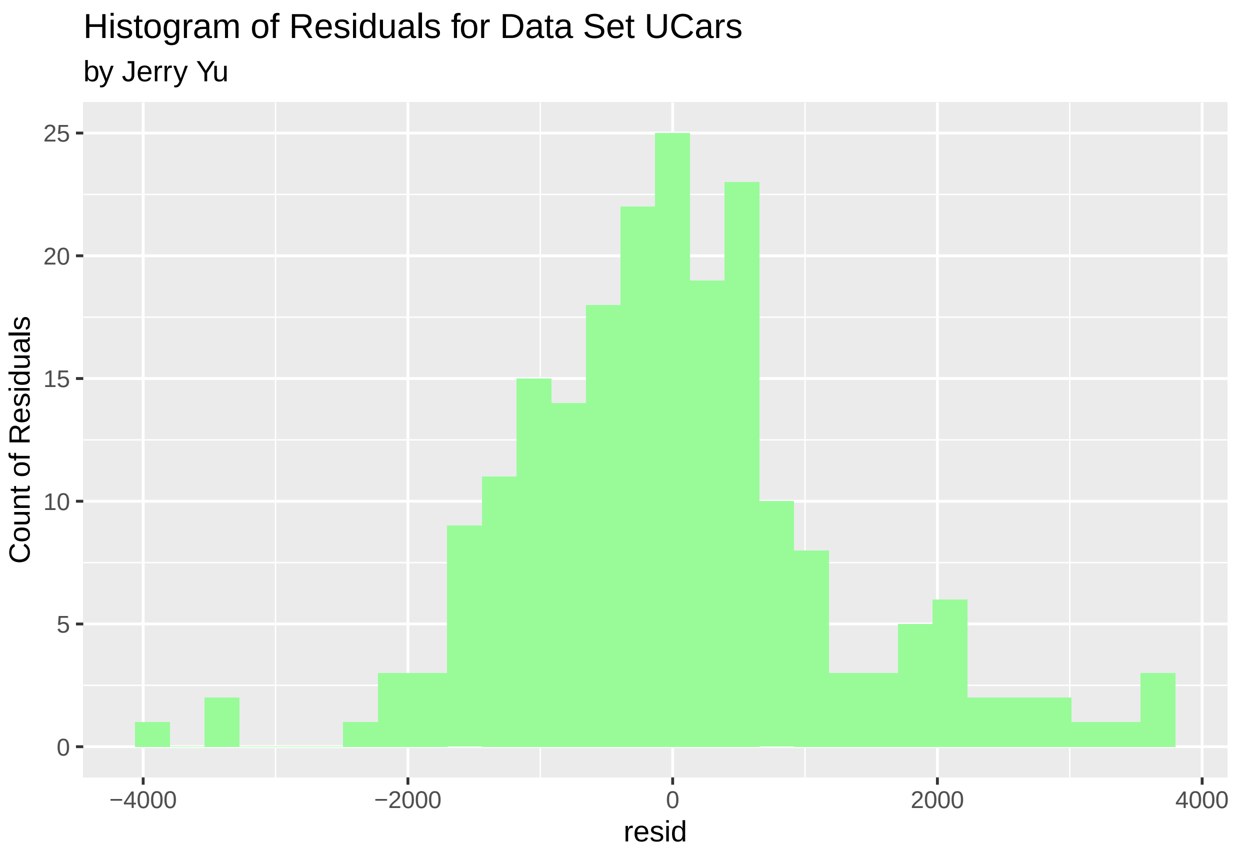
```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01638336      1.923891   0.508
##  Alternative hypothesis: rho != 0
```

H0: Errors are independent. (autocorrelation = 0) HA: Errors are not independent. (autocorrelation ≠ 0) Statistic: 1.9238914 P: 1.9238914

### 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 likley 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 %>% str()
```

```
## List of 4
##  $ statistic: Named num 0.969
##   ..- attr(*, "names")= chr "W"
##  $ p.value  : num 0.000138
##  $ method   : chr "Shapiro-Wilk normality test"
##  $ data.name: chr "ucarsfmrs$resid"
##  - attr(*, "class")= chr "htest"
```

```
shap1[[2]]
```

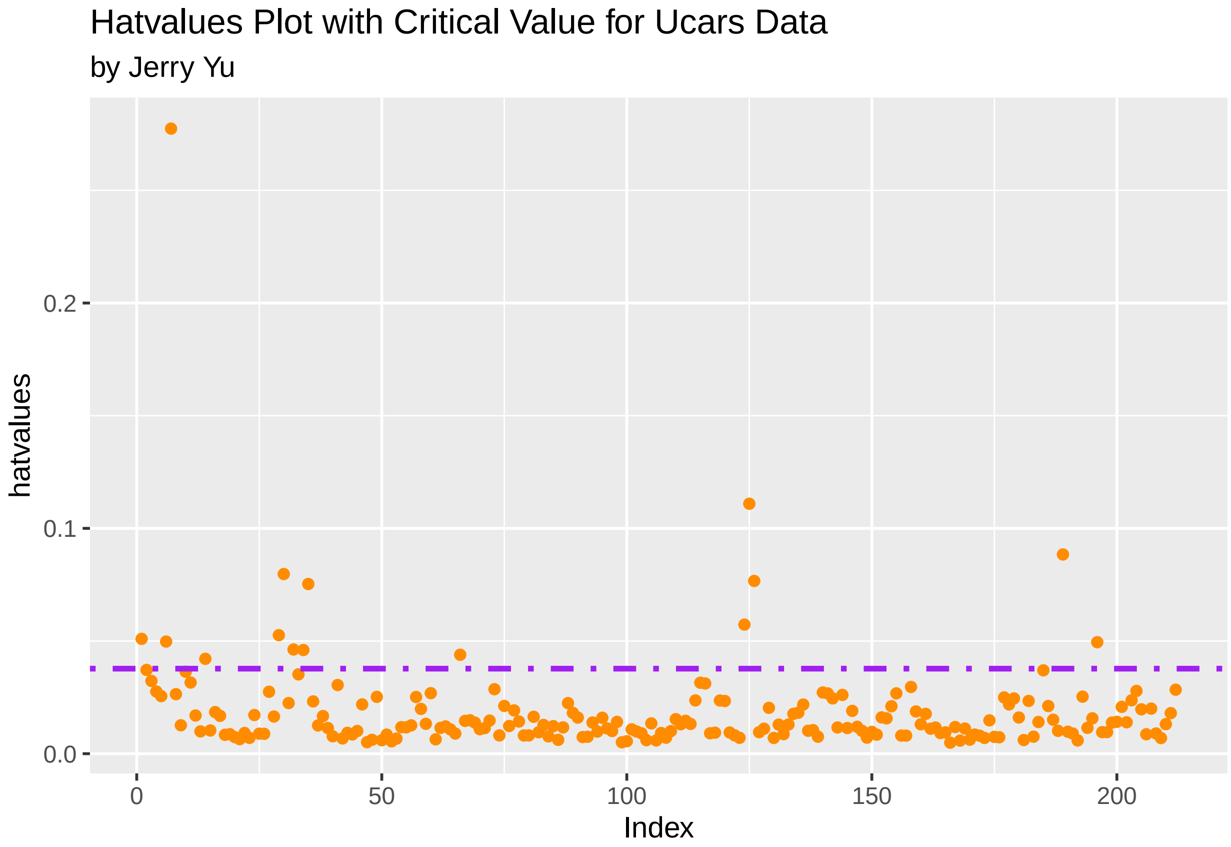
```
## [1] 0.0001376433
```

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.

```r
# 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.
```

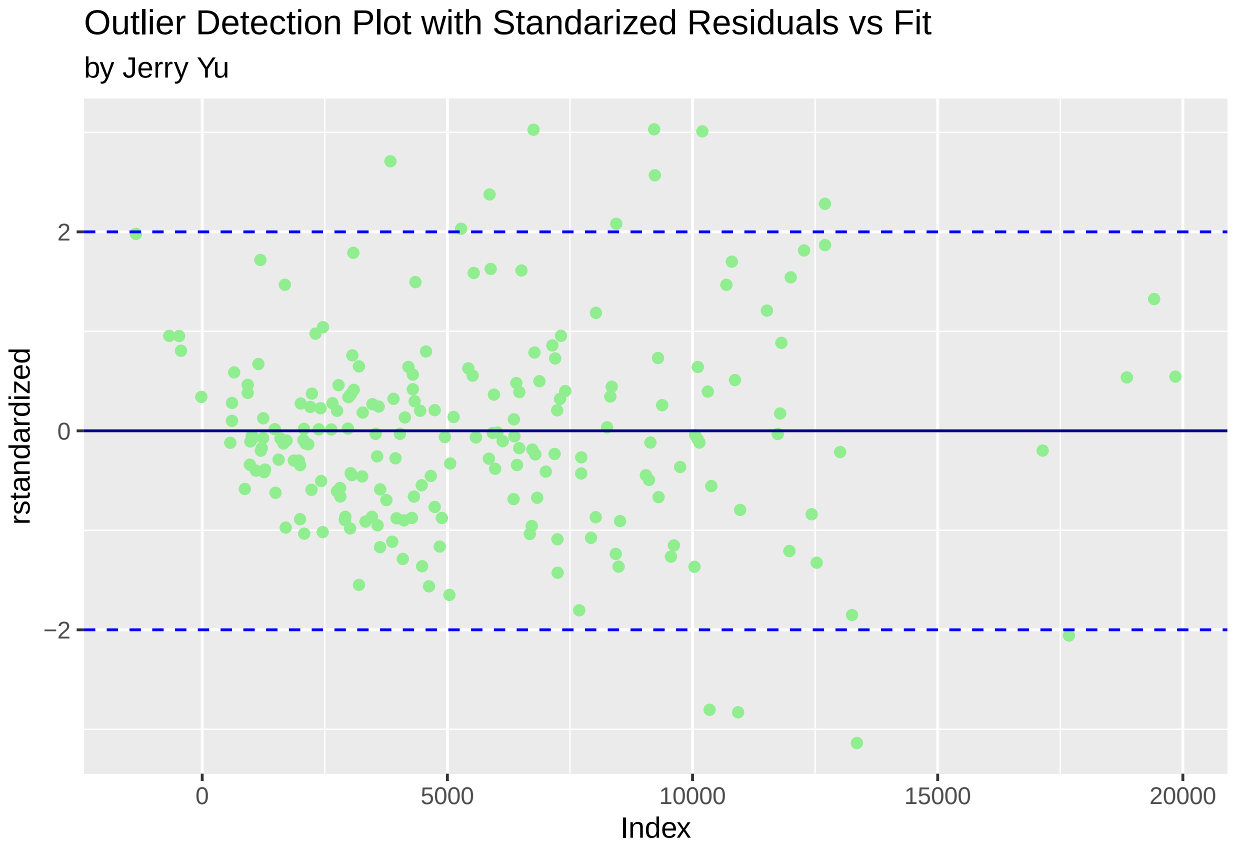


```
# print row numbers for large leverage points
ucarsfml %>% subset(hatvalues> crit) %>% select(c("rownum"))
```

```
## # 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")
```

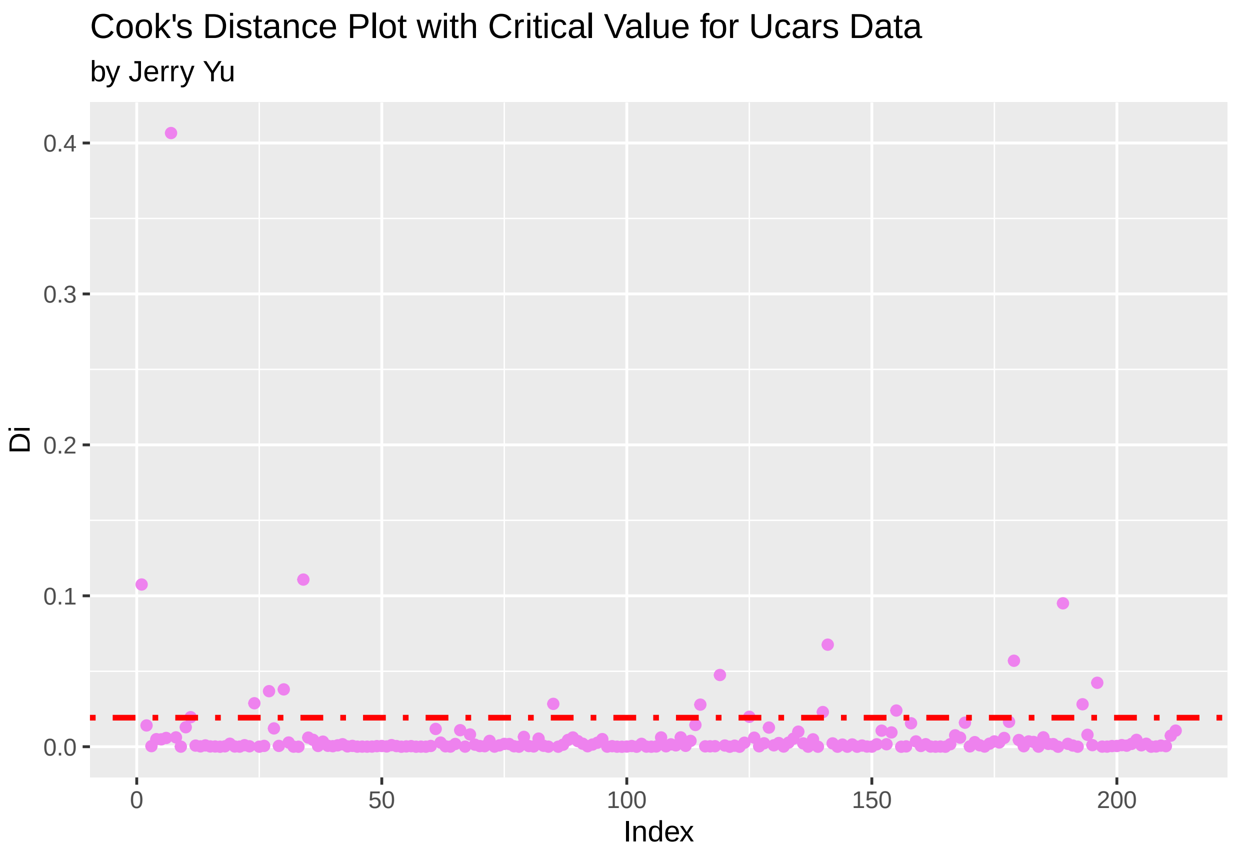


```
# print row numbers for large leverage points
ucarsfmro %>% subset(abs(rstandardized)> 2) %>% select(c("rownum"))
```

```
## # 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

```r
# 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")
```



```r
# 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 $1.0776959 \times 10^4$.