# HW1

*Joaquin Rodriguez*

*9/26/2017*

## 2.4 Table B.3 presents data on the gasoline mileage performance of 32 different automobiles.

### a. Fit a simple linear regression model relating gasoline mileage y (miles per gallon) to engine displacement x1 (cubic inches).

```
library(MPV)
```

```
##
## Attaching package: 'MPV'

## The following object is masked from 'package:datasets':
##
##     stackloss
```

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## The following object is masked from 'package:MPV':
##
##     cement
```

```
auto <- table.b3

names(auto) <- c("Miles/gallon", "Displacement (cubic in)", "Horsepower (ft-lb)", "Torque (ft-lb)", "Co


fit1 <- lm(`Miles/gallon` ~ `Displacement (cubic in)`, data = auto)
```

### b. Construct the analysis-of-variance table and test for significance of regression.

```
sum.fit1 <- summary(fit1)
sum.fit1
```

```
##
## Call:
## lm(formula = `Miles/gallon` ~ `Displacement (cubic in)`, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7923 -1.9752  0.0044  1.7677  6.8171
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               33.722677   1.443903   23.36  < 2e-16 ***
## `Displacement (cubic in)` -0.047360   0.004695  -10.09 3.74e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.065 on 30 degrees of freedom
## Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
## F-statistic: 101.7 on 1 and 30 DF,  p-value: 3.743e-11
```

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: Miles/gallon
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## `Displacement (cubic in)`  1 955.72  955.72  101.74 3.743e-11 ***
## Residuals                 30 281.82    9.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### c. What percent of the total variability in gasoline mileage is accounted for by the linear relationship with engine displacement?

The Adjusted R-squared for the linear model is 0.765. Therefore, the engine displacement explains up to `paste(round(sum.fit1$adj.r.squared, 3), "%", sep = "")` of the variability of in the gasoline mileage data.

### d. Find a 95% CI on the mean gasoline mileage if the engine displacement is 275 in.3

```
newdata2.4 <- data.frame(275)
names(newdata2.4) <- c("Displacement (cubic in)")
predict.lm(fit1, newdata2.4, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 20.69879 19.58807 21.80952
```

### e. Suppose that we wish to predict the gasoline mileage obtained from a car with a 275-in.3 engine. Give a point estimate of mileage. Find a 95% prediction interval on the mileage.

```
predict.lm(fit1, newdata2.4, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 20.69879 14.34147 27.05611
```

### f. Compare the two intervals obtained in parts d and e. Explain the difference between them. Which one is wider, and why?

The prediction interval for the for the mean response provides the confidence interval for the response if we were to use a different sample from the one used to fit the regression. On the other hand, the confidence interval for the prediction estimates the CI for new observations, therefore considering the true error. As a consequence, the prediction interval will be wider as we have to consider the true error associated to the model. In fact, in the prediction formula we use an extra shock of root MSE to increase the width of the prediction.

## 2.6 Table B.4 presents data for 27 houses sold in Erie, Pennsylvania.

### a. Fit a simple linear regression model relating selling price of the house to the current taxes (x1).

```
property <- table.b4

names(property) <- c("sale price of the house (in thousands of dollars)","taxes (in thousands of dollar

fit2 <- lm(`sale price of the house (in thousands of dollars)` ~ `taxes (in thousands of dollars)`, data
sum.fit2 <- summary(fit2)
```

### b. Test for significance of regression.

```
summary(fit2)
```

```
## 
## Call:
## lm(formula = `sale price of the house (in thousands of dollars)` ~ 
##     `taxes (in thousands of dollars)`, data = property)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.8343 -2.3157 -0.3669  1.9787  6.3168 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                        13.3202     2.5717   5.179 3.42e-05 ***
## `taxes (in thousands of dollars)`   3.3244     0.3903   8.518 2.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 2.961 on 22 degrees of freedom
## Multiple R-squared:  0.7673, Adjusted R-squared:  0.7568
## F-statistic: 72.56 on 1 and 22 DF,  p-value: 2.051e-08
```

### c. What percent of the total variability in selling price is explained by this model?

The Adjusted R-squared for the linear model is 0.757. Therefore, the taxes explains up to 75.7% of the variability of the sale price of the house.

### d. Find a 95% CI on B1.

```
confint(fit2)
```

```
##                                      2.5 %    97.5 %
## (Intercept)                       7.986755 18.653604
## `taxes (in thousands of dollars)` 2.514988  4.133754
```

### e. Find a 95% CI on the mean selling price of a house for which the current taxes are $750.

```
newdata <- data.frame(750)
names(newdata) <-  c("taxes (in thousands of dollars)")
predict(fit2, newdata = newdata, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 2506.599 1904.744 3108.453
```

## 3.5 Consider the gasoline mileage data in Table B.3.

### a. Fit a multiple linear regression model relatmg gasoline mileage y (miles per gallon) to engine displacement x1 and the number of carburetor barrels x6.

```
fit3 <- lm(`Miles/gallon` ~ `Displacement (cubic in)` + `Carburetor (barrels)`, data = auto)
```

### b. Construct the analysis-of-variance table and test for significance of regression

```
summary(fit3)
```

```
##
## Call:
## lm(formula = `Miles/gallon` ~ `Displacement (cubic in)` + `Carburetor (barrels)`,
##     data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0623 -1.6687 -0.3628  1.6221  6.2305
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)               32.884551   1.535408  21.417  < 2e-16 ***
## `Displacement (cubic in)` -0.053148   0.006137  -8.660 1.55e-09 ***
## `Carburetor (barrels)`     0.959223   0.670277   1.431    0.163    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.013 on 29 degrees of freedom
## Multiple R-squared:  0.7873, Adjusted R-squared:  0.7726 
## F-statistic: 53.67 on 2 and 29 DF,  p-value: 1.79e-10
```

```
anova(fit3)
```

```
## Analysis of Variance Table
##
## Response: Miles/gallon
##                           Df Sum Sq Mean Sq F value    Pr(>F)    
## `Displacement (cubic in)`  1 955.72  955.72 105.290 3.666e-11 ***
## `Carburetor (barrels)`     1  18.59   18.59   2.048    0.1631    
## Residuals                 29 263.23    9.08                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### c. Calculate R^2 and R^2 for this model. Compare this to the R^2 and the R^2Adj for the simple linear regression model relating mileage to engine displacement in Problem 2.4

```
# full model
fit3.summ <- summary(fit3)
fit3.summ$r.squared
```

```
## [1] 0.7872928
```

```
fit3.summ$adj.r.squared
```

```
## [1] 0.7726233
```

```
# reduced model
fit1.summ <- summary(fit1)
fit1.summ$r.squared
```

```
## [1] 0.7722712
```

```
fit1.summ$adj.r.squared
```

```
## [1] 0.7646803
```

The R^2 and R^2adj for the full model are respectively: 0.79 and 0.77. Whereas, the R^2 and R^2adj for the reduced model are respectively: 0.77 and 0.76.
As we can observe the R^2 and R^2adj for both models are nearly the same; the R^2 for the reduced model is slightly lower compared to the full one. The difference is around 1% and therefore the Carburetor variable does not significantly increase the portion of the variance explained by the model.

### d. Find a 95% CI for B1.

```
confint(fit3)
```

```
##                                 2.5 %      97.5 %
## (Intercept)               29.74428901 36.02481266
## `Displacement (cubic in)` -0.06569892 -0.04059641
## `Carburetor (barrels)`    -0.41164739  2.33009349
```

### e. Compute the t statistics for testing H0: B1 = 0 and H0: B6 = 0. What conclusions can you draw?

```
summary(fit3)
```

```
##
## Call:
## lm(formula = `Miles/gallon` ~ `Displacement (cubic in)` + `Carburetor (barrels)`,
##     data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0623 -1.6687 -0.3628  1.6221  6.2305
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               32.884551   1.535408  21.417  < 2e-16 ***
## `Displacement (cubic in)` -0.053148   0.006137  -8.660 1.55e-09 ***
## `Carburetor (barrels)`     0.959223   0.670277   1.431    0.163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.013 on 29 degrees of freedom
## Multiple R-squared:  0.7873, Adjusted R-squared:  0.7726
## F-statistic: 53.67 on 2 and 29 DF,  p-value: 1.79e-10
```

The Displacement predictor has a really low p-value, therefore we can reject the Null Hypothesis meaning that this beta is significantly different from zero.
The Carburetor predictor has a high p-value, therefore we fail to reject the Null Hypothesis meaning that this beta is not significantly diffrent from zero. Therefore, the Carburetor predictor does not seem to add any significant information compared to the reduced model.

### f. Find a 95% CI on the mean gasoline mileage when x1 = 275 in.3 and x6 = 2 barrels.

```
newdata3.5 <- data.frame(275, 2)
names(newdata3.5) <- c("Displacement (cubic in)", "Carburetor (barrels)")
predict.lm(fit3, newdata3.5, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 20.18739 18.87221 21.50257
```

### g. Find a 95% prediction interval for a new observation on gasoline mileage when x1 = 275 in.3 and x6 = 2 barrels.

```
predict.lm(fit3, newdata3.5, interval = "prediction", level = 0.95)
```

```
##        fit     lwr      upr
## 1 20.18739 13.8867 26.48808
```

### 3.6 In problem 2.4 you were asked to compute a 95% CI on mean gasoline prediction interval on mileage when the engine displacement x1 = 275 in.^3 Compare the lengths of these intervals to the lengths of the confidence and prediction intervals from Problem 3.5 above. Does this tell you anything about the benefits of adding x6 to the model?

```
# reduced model
a <- predict.lm(fit1, newdata2.4, interval = "prediction", level = 0.95) %>% print()
```

```
##        fit      lwr      upr
## 1 20.69879 14.34147 27.05611
```

```
a[1,3]- a[1,2]
```

```
## [1] 12.71464
```

```
# full model
a <- predict.lm(fit3, newdata3.5, interval = "prediction", level = 0.95) %>% print()
```

```
##        fit     lwr      upr
## 1 20.18739 13.8867 26.48808
```

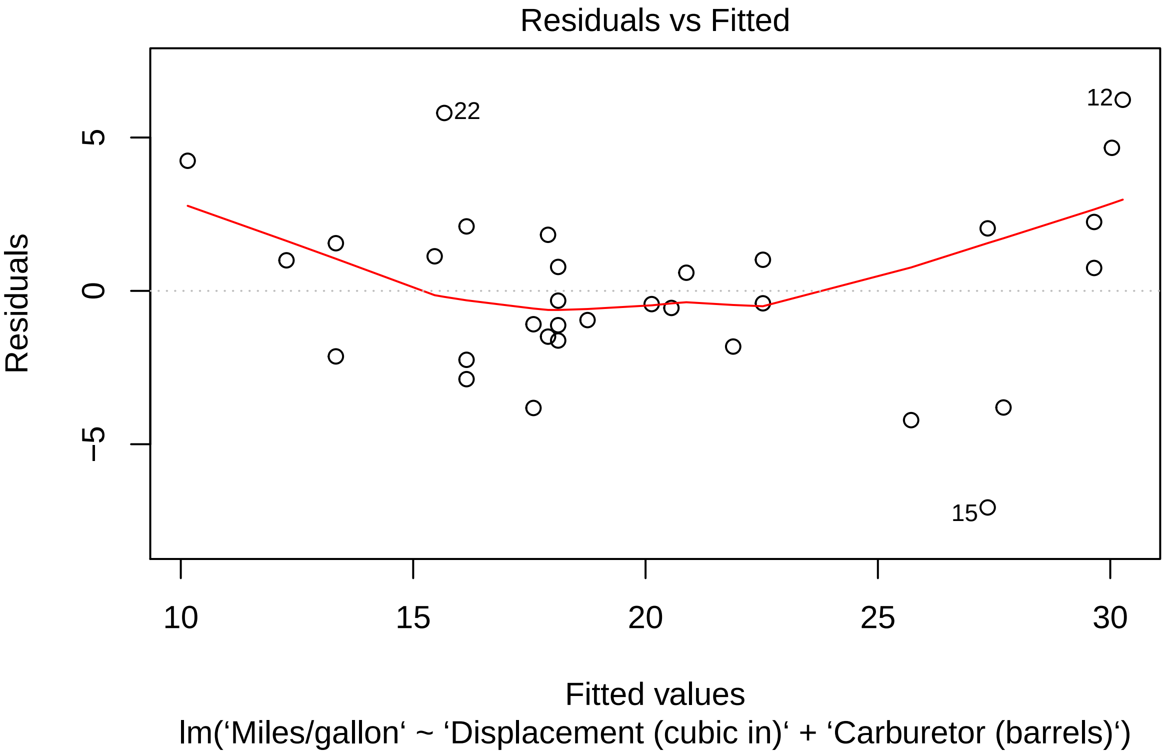
```
a[1,3]- a[1,2]
```

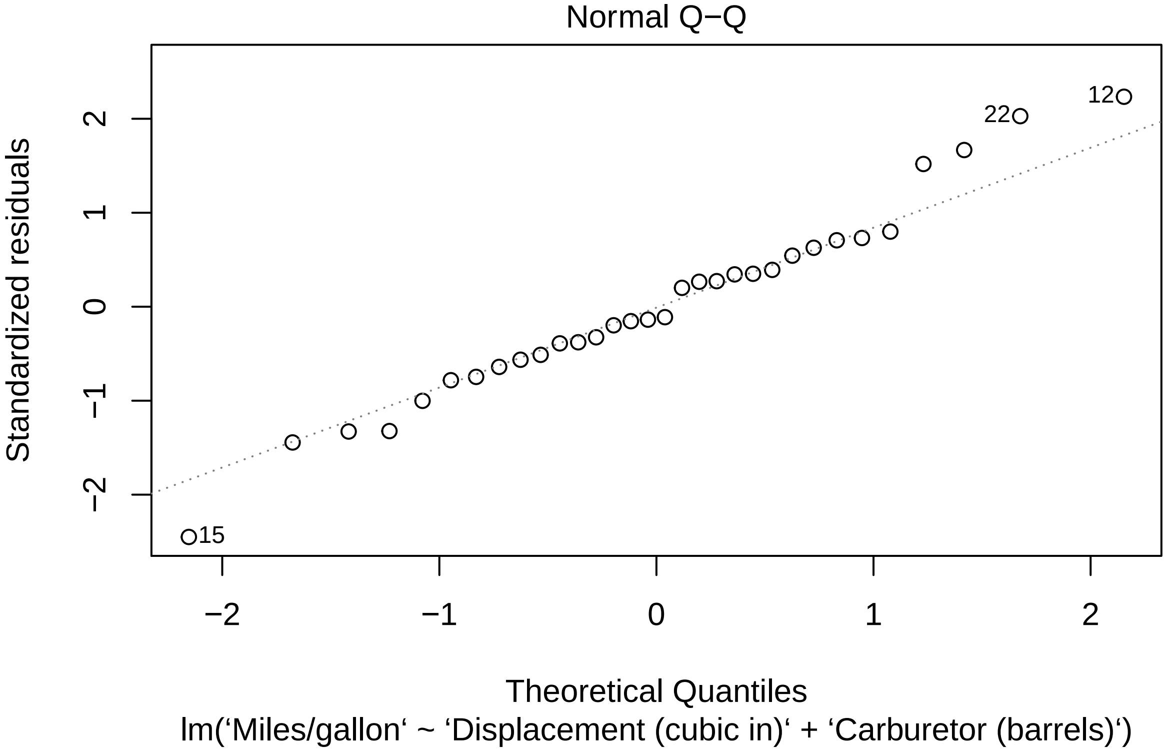
```
## [1] 12.60138
```

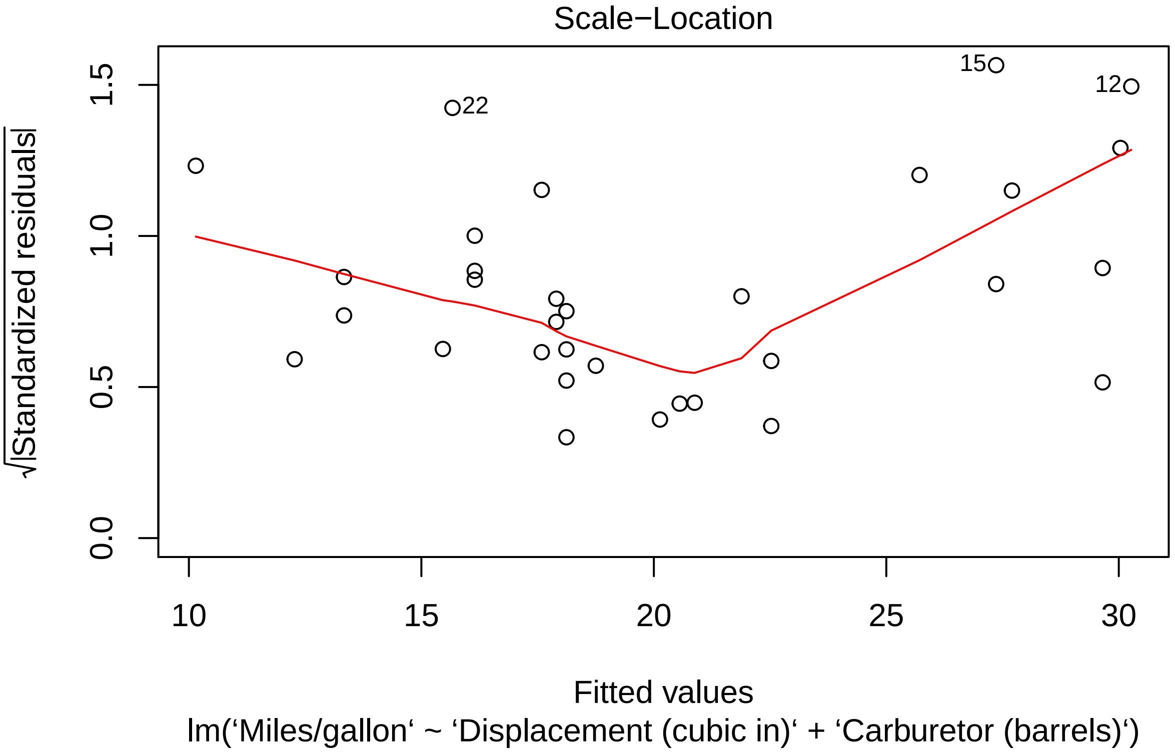
As we can observe interval length for the full model is slightly lower compared to the reduced model. Therefore, the full model is able to predict with a better confidence the mean gasoline consumption. This result is a consequence of the fact that the full model has a slightly higher adj R^2 compared to the reduced model. However, we can observe how the difference in interval length is quite small.

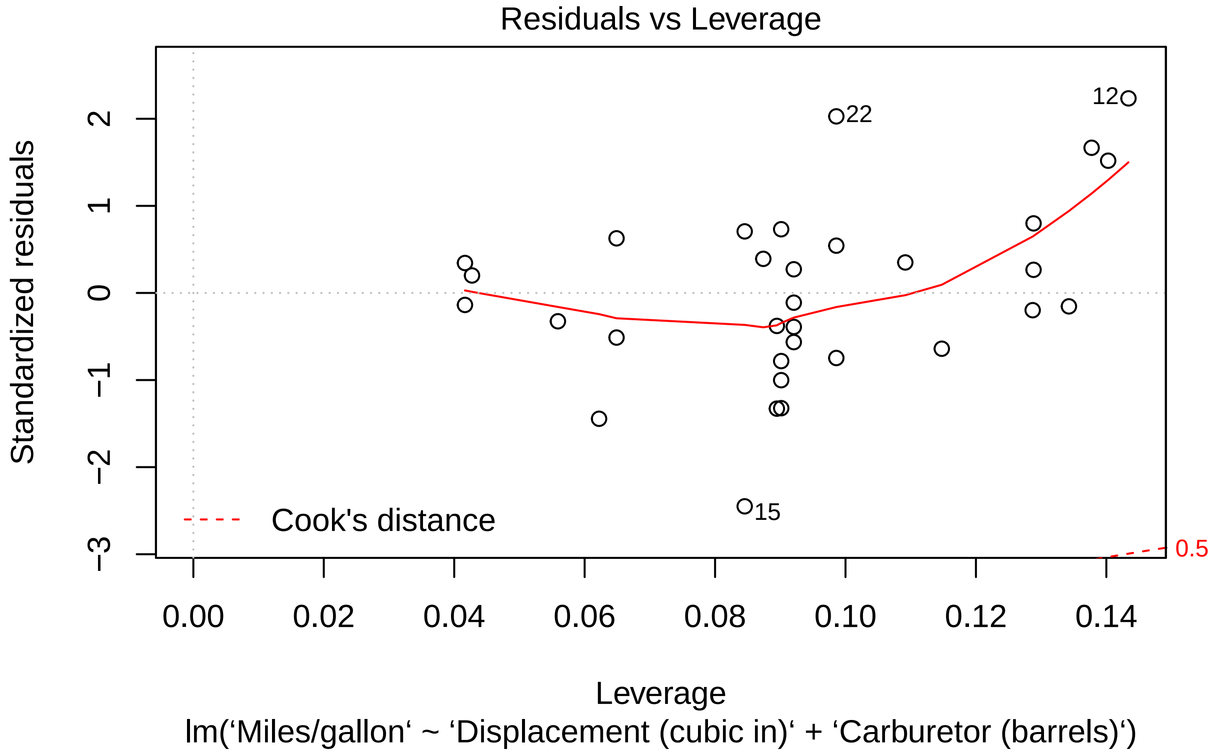
### 4.4 Consider the multiple regression model fit to the gasoline mileage data in Problem 3.5.

```
plot(fit3)
```


Residuals vs Fitted
Residuals
5
0
−5
22
12
15
10
15
20
25
30
Fitted values
lm(‘Miles/gallon‘ ~ ‘Displacement (cubic in)‘ + ‘Carburetor (barrels)‘)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
12
22
15
−2
−1
0
1
2
Theoretical Quantiles
lm(‘Miles/gallon‘ ~ ‘Displacement (cubic in)‘ + ‘Carburetor (barrels)‘)





**a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?**

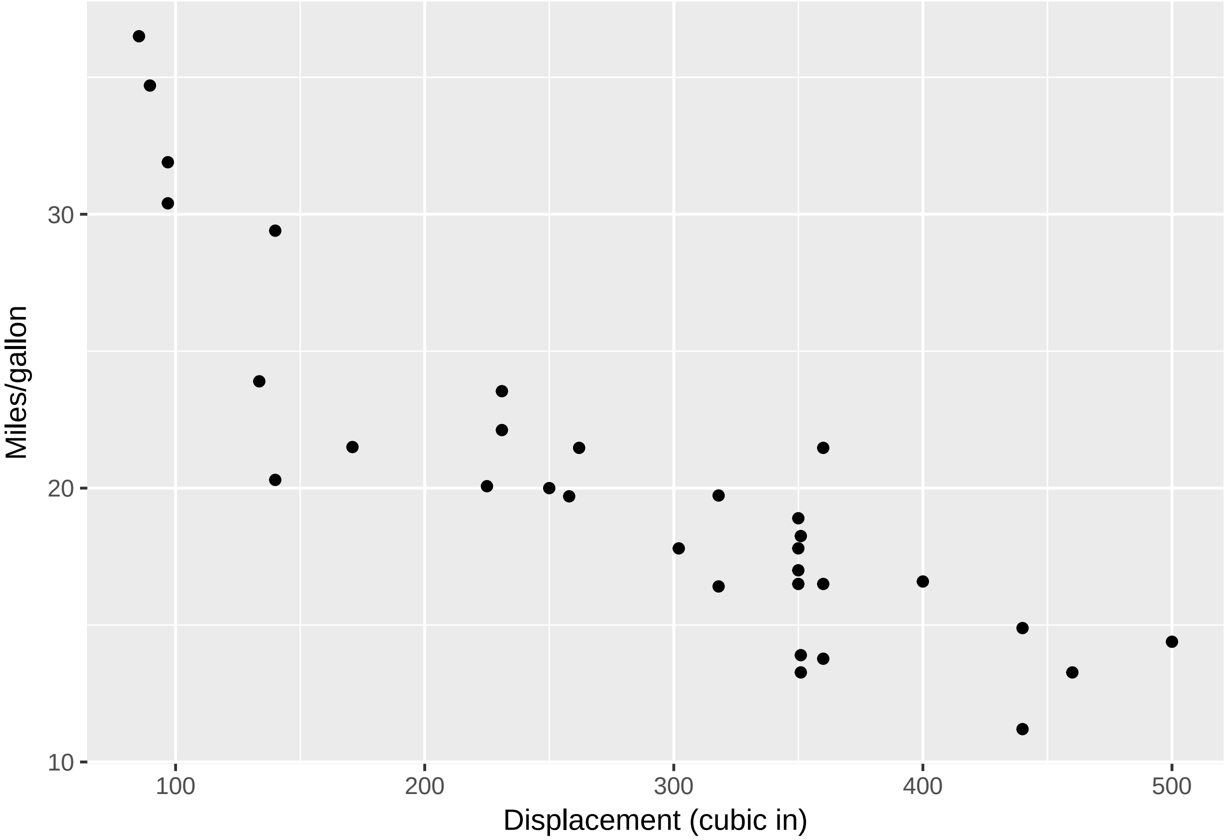
From the Normal Q-Q Plot we can observe how the assumption of normality appears to hold. There is a small departure from normaliy for some observations, however these departures do not seem to be significant enough to undermine the normality assumption.

### b. Construct and interpret a plot of the residuals versus the predicted response.

From that standardized residual vs predicted response plot we can observe how the residuals are generally speaking equally spread along the ranges of the predictors. From the plot we can observe how there are three observations (#12, #15, #22) that have high residuals, thus indicating possible outliers or leverage points.

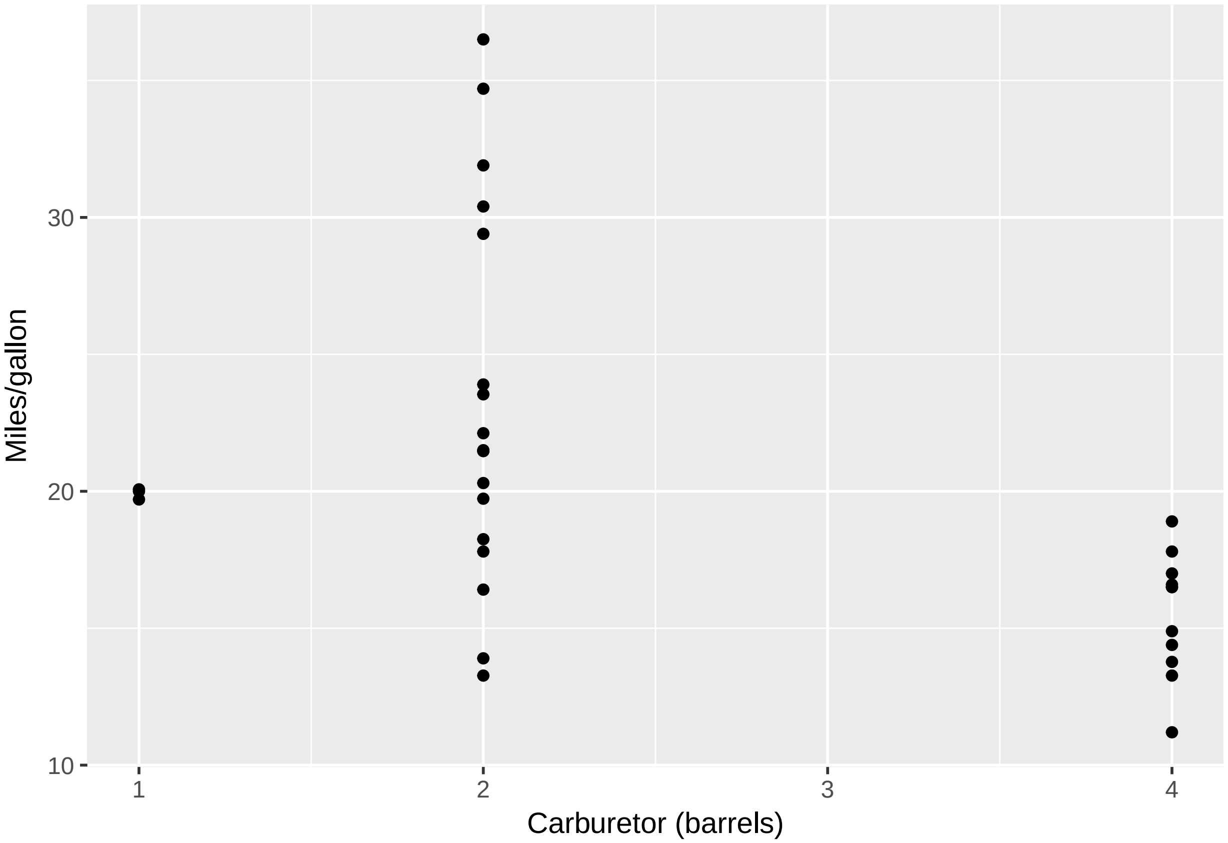
### c. Construct and interpret the partial regression plots for this model.

```
auto %>%
  ggplot(aes(x = `Displacement (cubic in)`, y = `Miles/gallon`)) +
  geom_point()
```



We can observe how the relationship between Consumption vs Displacement approximates a simple linear trend. It is likely that the relationship would be better approximated by a second order polinomial.

```
auto %>%
  ggplot(aes(x = `Carburetor (barrels)`, y = `Miles/gallon`)) +
  geom_point()
```



As we can observe from the plot the relationship between Consumption and Carburetor do not seem to be correctly approximated by a linear regression. In fact, for each level of Carburetor we have significant variation. This is coherent with the fact that the Carburetor variable does not result to be significantly different from zero.

**Compute the studentized residuals and the R-student residuals for this model. What information is conveyed by these scaled residuals?**

```
#studentized residuals
residuals(fit3) / fit3.summ$sigma * sqrt (1 - influence(fit3)$hat)
```

```
##          1          2          3          4          5          6
##  0.2467640 -0.3541423 -0.1725258  0.6654468 -0.5669752 -0.6732633
##          7          8          9         10         11         12
## -0.1318844  0.1921501  1.4375977  0.2314219 -0.5122755  1.9139963
##         13         14         15         16         17         18
## -1.3547238 -0.1333150 -2.2428180 -0.3071533  1.3056208  0.4895626
##         19         20         21         22         23         24
## -0.1011291 -0.4788926  0.3295161  1.8277972  0.3575470  0.6961169
##         25         26         27         28         29         30
##  0.6471153  0.3121734 -1.2038682  0.5867122 -0.7117755 -0.9112353
##         31         32
## -1.2093191 -0.3446788
```

```
#R-student residual
rstudent(fit3)
```

```
##          1          2          3          4          5          6 
##  0.2674019 -0.3842810 -0.1946983  0.7253733 -0.6338410 -0.7410735 
##          7          8          9         10         11         12 
## -0.1352667  0.1973731  1.7229000  0.2613455 -0.5574801  2.4130447 
##         13         14         15         16         17         18 
## -1.4734791 -0.1513731 -2.7032887 -0.3202682  1.5553718  0.5363974 
##         19         20         21         22         23         24 
## -0.1094707 -0.5055101  0.3385474  2.1507428  0.3859964  0.7939593 
##         25         26         27         28         29         30 
##  0.7006450  0.3450656 -1.3412473  0.6207433 -0.7769267 -1.0015660 
##         31         32 
## -1.3466593 -0.3728890
```

The studentized residuals provide a proper standardization of the errors as it considers also the hat values of the observations. In fact, the studentized residual follow closely a Z standard normal.
The R-student provides a studentized residuals where the estemiate of sigma squared is estimated leaving out the current observation. Therefore, the R-student provides a externally studentized residuals.
From these residuals we can identify those observations that the model does not produce an adequate estimate.
These measures are useful in order to identigy outliers and influential observations, but also deviations of the data from the underlying model assumed.

### 4.14 Problems 2.4 and 3.5 asked you to fit two different models to the gasoline mileage data in Table B.3. Calculate the PRESS statistic for these two models. Based on this statistic, which model is most likely to provide better predictions of new data?

```
# PRESS 2.4
sum( (resid(fit1) / (1- influence(fit1)$hat)) ^2 )
```

```
## [1] 337.206
```

```
# PRESS 3.5
sum( (resid(fit3) / (1- influence(fit3)$hat)) ^2 )
```

```
## [1] 328.7654
```

The PRESS conveys better predictions of new data for lower levels. Therefore, the model that is more likely predict better predictions of new data is the model we fit in point 3.5.