Guo - HWK 1

Andrew Guo

1/29/2022

Problem 1

For each part, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer. ### (a) The sample size n is extremely large, and the number of predictors p is small

It appears that the the inflexible method may be more effective at providing a model due to the smaller number of predictors. But with such a large sample size and with the linearity of the sample distribution unknown, it is unlikely that the inflexible model will be able to accurately predict the true value of f.

Thus, the flexible method, due to the large sample size reducing the amount of reducible error for our model and due to flexible models in general being more apt at accounting for nonlinearity, may therefore be preferred as the tendency for this method of overfitting the data is low, and flexible models in general are more effective at predicting the true value of f. Additionally, the smaller number of predictors does not limit the total nubmer of beta hats that are being estimated. Quadratic, cubic, and higher order terms as well as interaction terms can be made in numerous amounts despite a small number of predictors.

(b) The number of predictors p is extremely large, and the number of observations n is small.

Both options here won't meet their best effectiveness due to the small sample size, but because the sample size is small, a flexible method will have a very high tendency to overfit the data. Because this may result in a considerable variation in the estimation of the true value of f, an inflexible method would be preferred.

(c) The relationship between the predictors and response is highly non-linear.

Inflexible models are less effective than flexible models at interpreting nonlinearity, thus a flexible method would be preferred.

(d) The variance of the error terms, o2 = Var(E), is extremely high.

Because the variance is extremely high, using a flexible model has a higher tendency to overfit the data. Thus, it would be safer and more effective to use an inflexible model to estimate the true value of f.

Problem 2

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Response variable (CEO Salary) is quantitative, thus we are using regression. We are interested in inference; we are not looking to predict the salary, but rather explaining it.

n = 500, p = 4

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Response variable (Outcome of success) is qualitative, thus we are using classification. We are interested in finding out how successful it will be based upon using previously launched products, so we are interested in prediction.

n = 20, p = 14

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Response variable (% change in exchange rate) is quantitative, so we are interested in regression. We are interested in seeing how the weekly % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market determine the % change in the USD/Euro exchange rate, thus we are interested in inference. n = 52, p = 4

Problem 3

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification?

(a) Under what circumstances might a more flexible approach be preferred to a less flexible approach?

A situation where there is a clear nonlinear relation between the response and the predictors, there is a huge sample size and there are many parameters would make a flexible approach much more preferable than an inflexible approach. ### (b) When might a less flexible approach be preferred? When there is a clear linear relationship between the response and the predictors and when there are few parameters and the sample size is not too large, we can estimate the true value of f more closely as compared to the alternatives.

Problem 4

(a) Read in the data set using read.csv(). Then make sure that the missing values have been removed from the data using na.omit(). Use the attach() function on the data frame so that you may refer directly to the column names.

```r
Auto <- read.csv('Auto.csv')
attach(Auto)
Auto = na.omit(Auto)
```

(b) Run the summary() function on your data frame. Which of the predictors are quantitative, and which are qualitative?

```r
summary(Auto)
```

```
##       mpg          cylinders      displacement    horsepower       
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Length:397        
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.0   Class :character  
##  Median :23.00   Median :4.000   Median :146.0   Mode  :character  
##  Mean   :23.52   Mean   :5.458   Mean   :193.5                     
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0                     
##  Max.   :46.60   Max.   :8.000   Max.   :455.0  
##      weight      acceleration        year           origin     
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000  
##  1st Qu.:2223   1st Qu.:13.80   1st Qu.:73.00   1st Qu.:1.000  
##  Median :2800   Median :15.50   Median :76.00   Median :1.000  
##  Mean   :2970   Mean   :15.56   Mean   :75.99   Mean   :1.574  
##  3rd Qu.:3609   3rd Qu.:17.10   3rd Qu.:79.00   3rd Qu.:2.000  
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000  
##      name          
##  Length:397        
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

(c) Use your answer from part (b) to report the min, max, and mean for each quantitative predictor. Summarize the results in a table:

Predictor	Minimum	Maximum	Mean
mpg	9.00	46.60	23.52
year	70.00	82.00	75.99
cylinders	3.000	8.000	5.458
displacement	68.0	455.0	193.5
acceleration	8.00	24.80	15.56
weight	1613	5140	2970

(d) Now remove the 10th through 85th observations. Report the min, max, and mean for each quantitative predictor in the subset of the data that remains. Summarize the results in a table:

```r
newAuto <- Auto[-c(10:85), ]
summary(newAuto)
```

```
##       mpg          cylinders      displacement  horsepower        
##  Min.   :11.00   Min.   :3.000   Min.   : 68   Length:321        
##  1st Qu.:18.00   1st Qu.:4.000   1st Qu.:100   Class :character  
##  Median :23.90   Median :4.000   Median :145   Mode  :character  
##  Mean   :24.44   Mean   :5.371   Mean   :187                     
##  3rd Qu.:30.70   3rd Qu.:6.000   3rd Qu.:250                     
##  Max.   :46.60   Max.   :8.000   Max.   :455                     
##      weight      acceleration        year           origin     
##  Min.   :1649   Min.   : 8.50   Min.   :70.00   Min.   :1.000  
##  1st Qu.:2215   1st Qu.:14.00   1st Qu.:75.00   1st Qu.:1.000  
##  Median :2795   Median :15.50   Median :77.00   Median :1.000  
##  Mean   :2934   Mean   :15.72   Mean   :77.15   Mean   :1.598  
##  3rd Qu.:3504   3rd Qu.:17.30   3rd Qu.:80.00   3rd Qu.:2.000  
##  Max.   :4997   Max.   :24.80   Max.   :82.00   Max.   :3.000  
##      name          
##  Length:321        
##  Class :character  
##  Mode  :character  
##                    
```

```
##
##
```

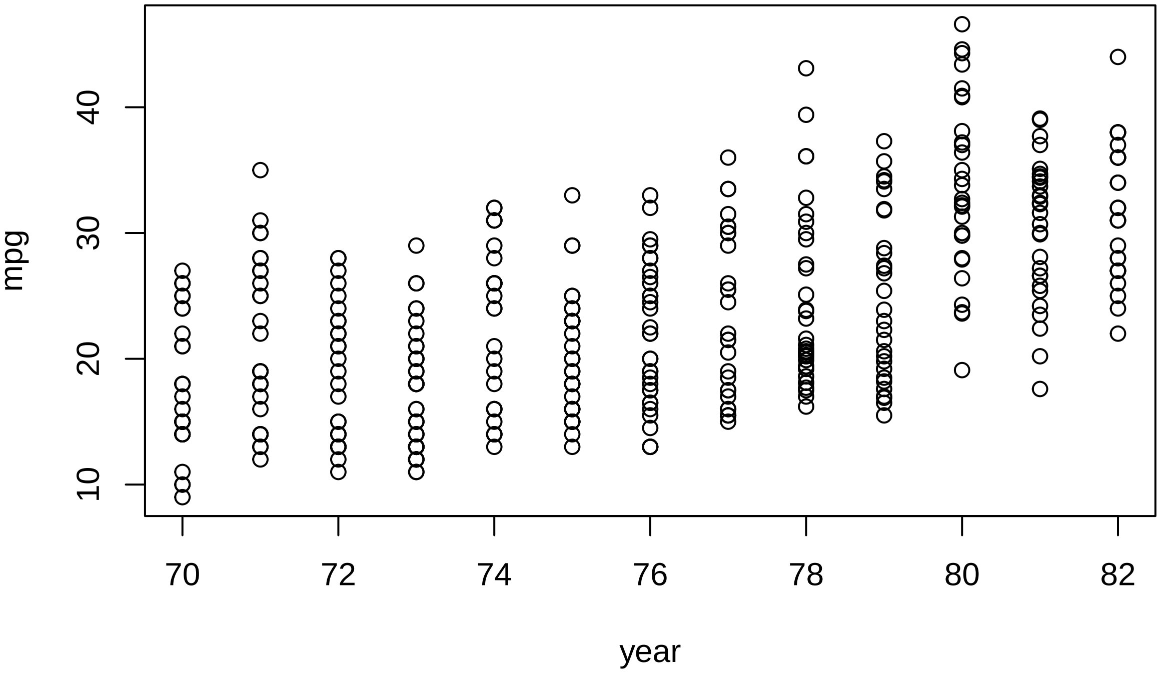
Predictor	Minimum	Maximum	Mean
mpg	11.00	46.60	24.44
year	70	82.00	77.15
cylinders	3.000	8.000	5.371
displacement	68	455	187
acceleration	8.50	15.72	24.80
weight	1649	4997	2934

(e)) Using the full data set, investigate the predictors graphically, using scatter plots. The pairs() and/or plot() functions would be helpful. Create some plots highlighting the relationships among the predictors. Comment on your findings.

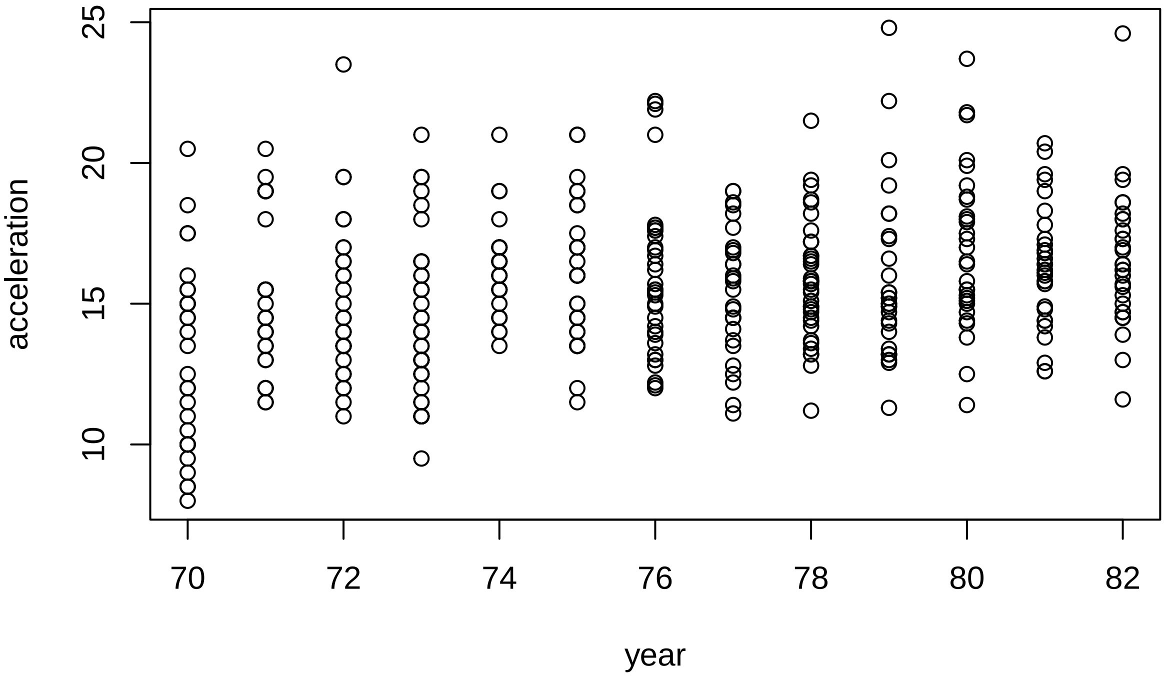
```
attach(Auto)
```

```
## The following objects are masked from Auto (pos = 3):
##
##     acceleration, cylinders, displacement, horsepower, mpg, name,
##     origin, weight, year
```

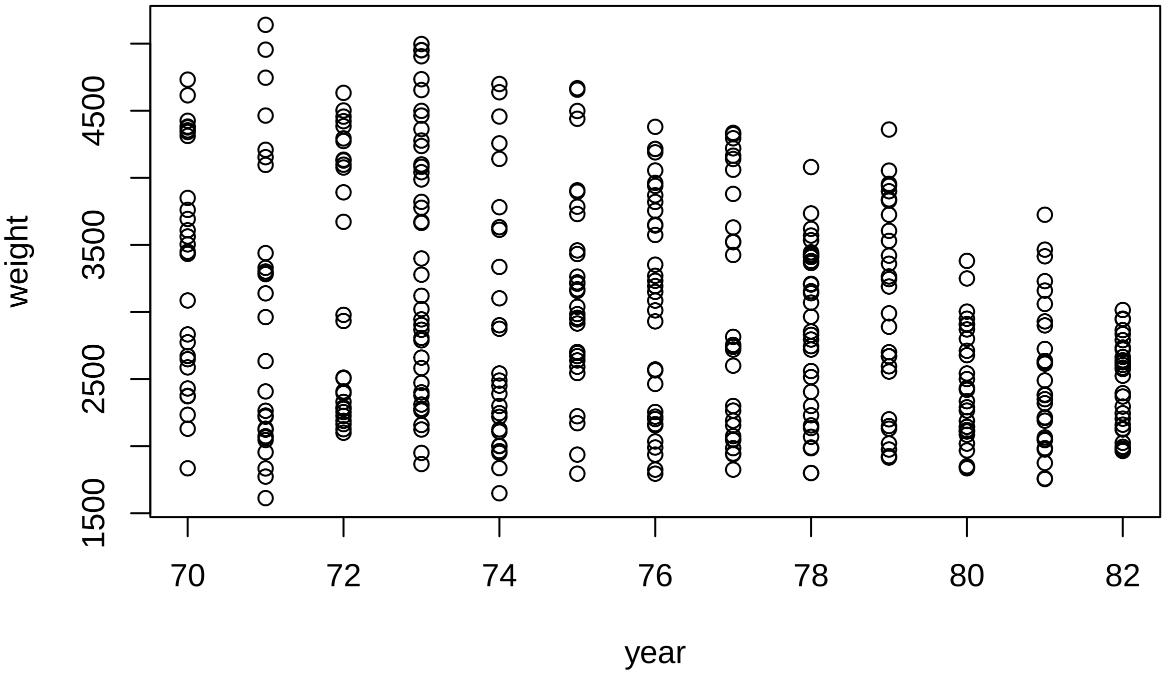
```
plot(year, mpg)
```



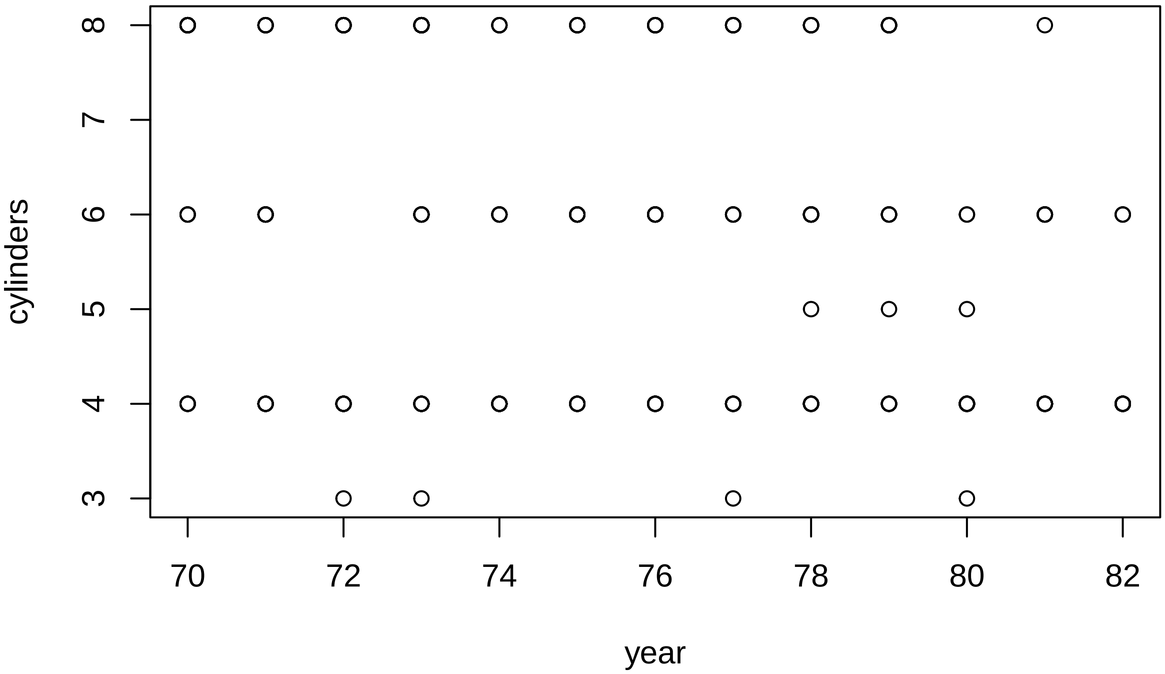
```
plot(year, acceleration)
```



```
plot(year, weight)
```



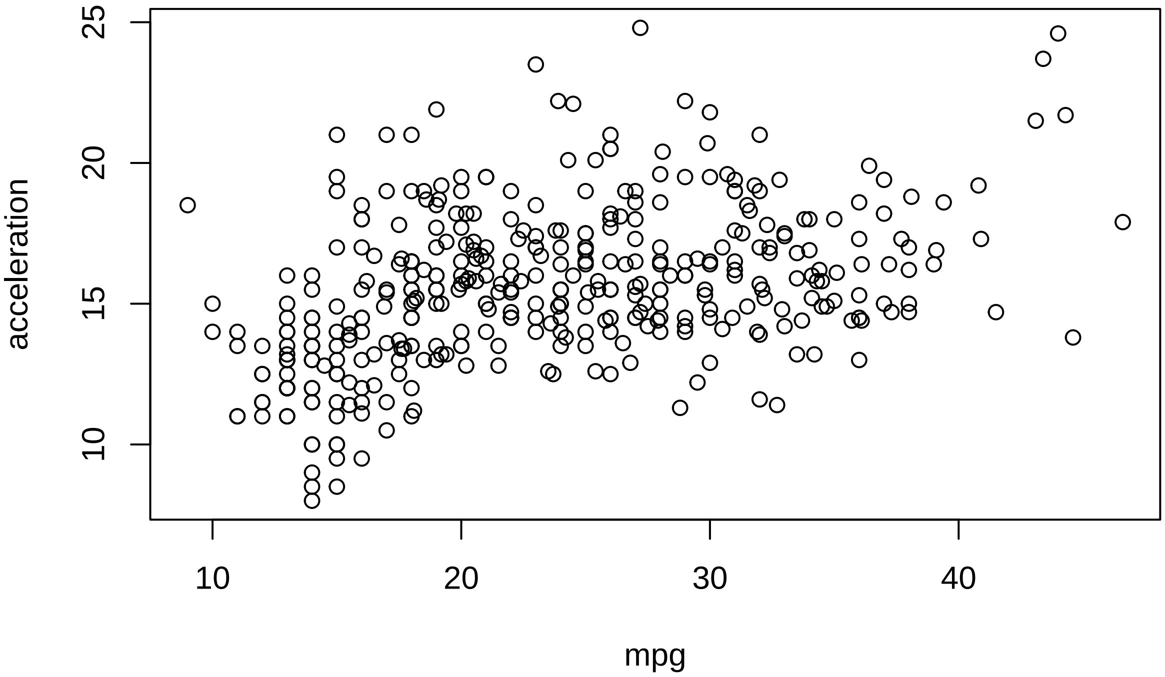
```
plot (year, cylinders)
```



Between 1970 and 1982, we see that the overall variability in the mpg, acceleration and weight of cars has roughsly stayed consistent over time. However, especially in our year vs. mpg graph, we notice that the upper bounds for mpg have increased especially since the 80s, with the lower bound also increasing. This indicates that cars are technologically improving in terms of mpg. We also see a similar pattern in our year vs. acceleration graph. However, we also notice that therer is no clear correlation between the year the car was produced and its number of cylinders. Other variables may be at play impacting the number of cylinders in side a car, such as the type of car (race cars for instance may have more cylinders to increase speed and acceleration).

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

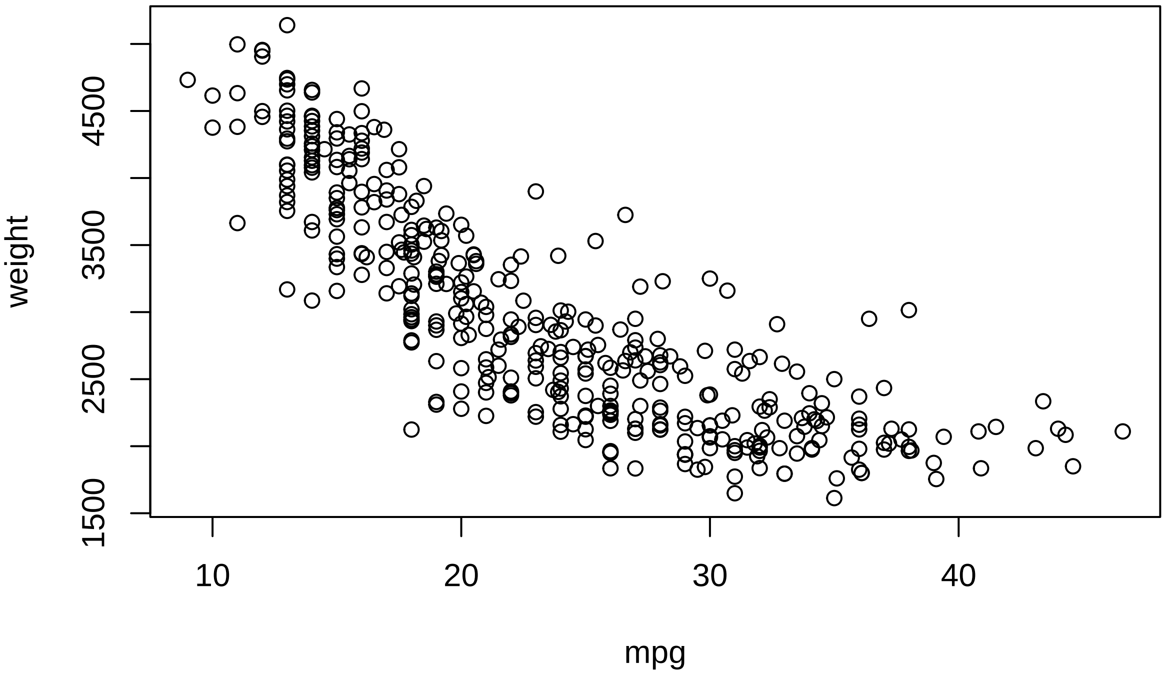
```
plot(mpg, acceleration)
```



```
lm.fit.MA = lm(acceleration~mpg, data = Auto)
summary(lm.fit.MA)
```

```
##
## Call:
## lm(formula = acceleration ~ mpg, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1436 -1.7179 -0.2371  1.4661  8.6976
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.06601    0.39716  30.381   <2e-16 ***
## mpg          0.14840    0.01603   9.259   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.496 on 395 degrees of freedom
## Multiple R-squared:  0.1783, Adjusted R-squared:  0.1763
## F-statistic: 85.73 on 1 and 395 DF,  p-value: < 2.2e-16
```

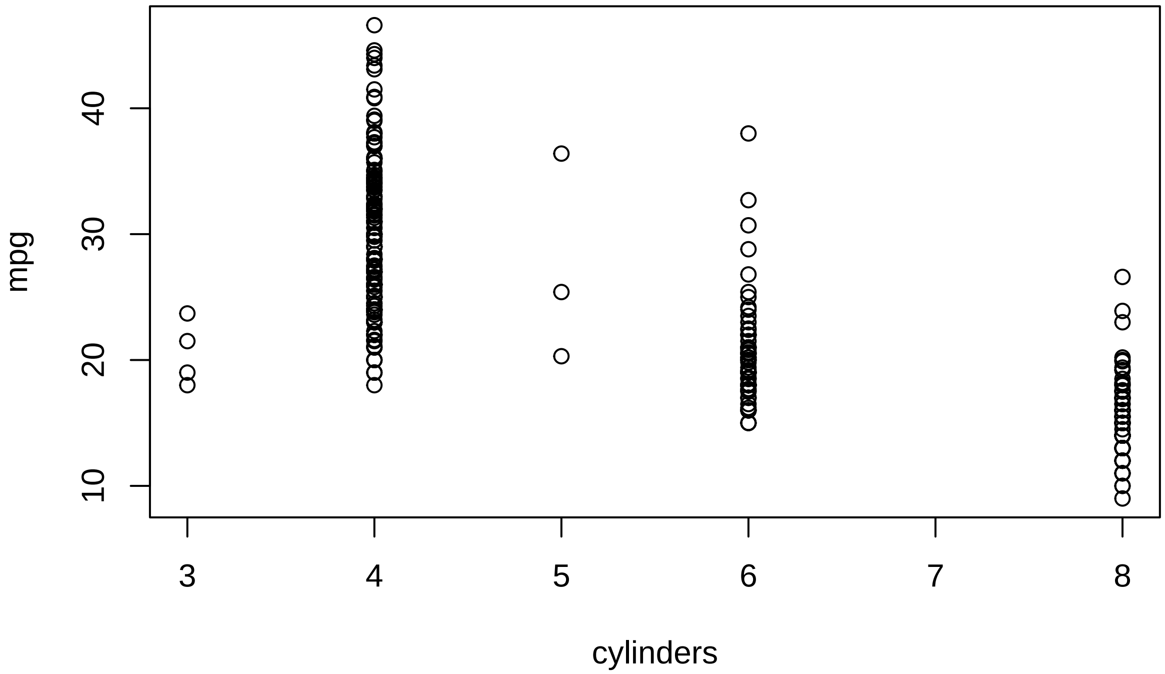
```
plot(mpg, weight)
```



```
lm.fit.MW = lm(weight~mpg, data = Auto)
summary(lm.fit.MW)
```

```
##
## Call:
## lm(formula = weight ~ mpg, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1343.33  -325.82   -31.33   317.67  1350.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5089.432     74.999   67.86   <2e-16 ***
## mpg          -90.117      3.027  -29.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.3 on 395 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.691
## F-statistic: 886.6 on 1 and 395 DF,  p-value: < 2.2e-16
```

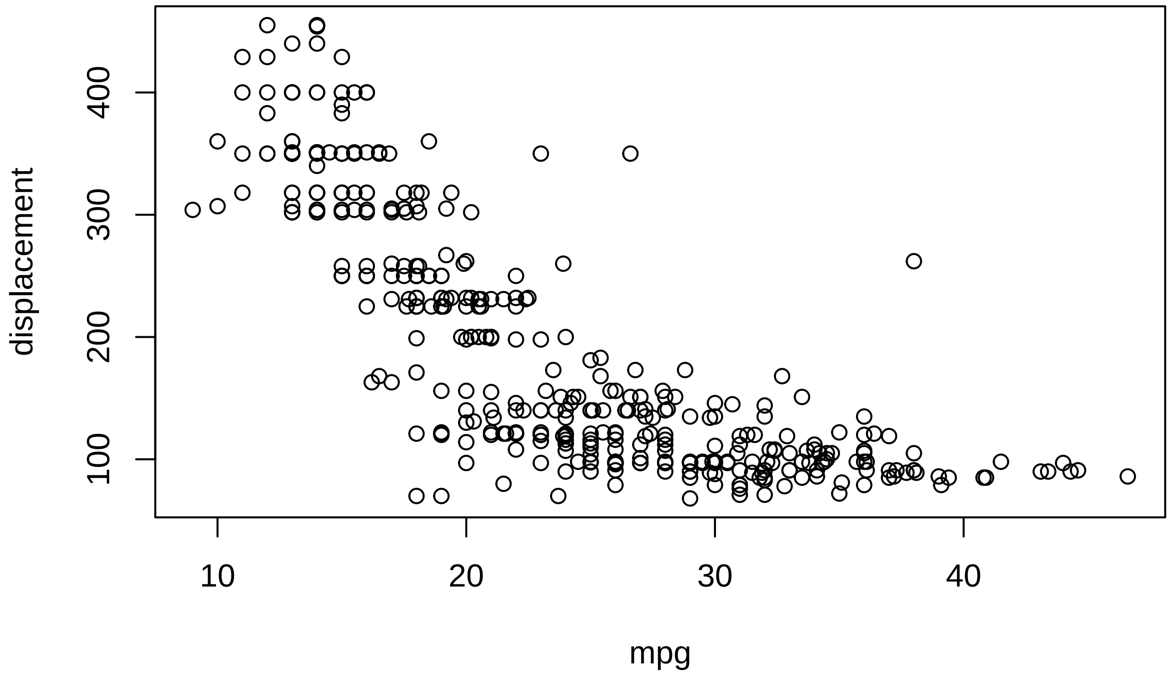
```
plot (cylinders, mpg)
```



```
lm.fit.MC = lm(cylinders~mpg, data = Auto)
summary(lm.fit.MC)
```

```
##
## Call:
## lm(formula = cylinders ~ mpg, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3894 -0.7270 -0.0265  0.7921  3.0621
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.427527   0.170908   55.16   <2e-16 ***
## mpg         -0.168783   0.006897  -24.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 395 degrees of freedom
## Multiple R-squared:  0.6026, Adjusted R-squared:  0.6016
## F-statistic: 598.9 on 1 and 395 DF,  p-value: < 2.2e-16
```

```
plot(mpg, displacement)
```



```r
lm.fit.MD = lm(displacement~mpg, data = Auto)
summary(lm.fit.MD)
```

```
##
## Call:
## lm(formula = displacement ~ mpg, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -182.716  -37.690   -2.716   39.489  223.876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 445.8477     9.8789   45.13   <2e-16 ***
## mpg         -10.7296     0.3987  -26.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.08 on 395 degrees of freedom
## Multiple R-squared:  0.6471, Adjusted R-squared:  0.6462
## F-statistic: 724.4 on 1 and 395 DF,  p-value: < 2.2e-16
```

We notice that there is a rough, general linear relationship between mpg and acceleration; as mpg increases, the acceleration generally decreases as well. For mpg and weight, there is a negative relationship between the two; as mpg increases, the weight decreases. This relationship appears to be nonlinear, as it appears that the weight decreases at a steadily increasing rate as mpg increases. We also see a similar, nonlinear, negative relationship between mpg and displacement; as mpg increases, displacement decreases at an increasing rate. Additionally, for the number of cylinders, we notice that cars with more cylinders tend to have lower and possibly less deviation in mpg. Cars with 4 cylinders contained mpgs varying between roughly 20 and 50 mpg, whereas cars with eight cylinders contained roughly between 10 and 25 mpg. The smaller variability in the mpg as cylinder count increased is not definitively a result of the number of cylinders; this could be explained by there being a smaller number of cars that have more cylinders.

As such, acceleration, weight, cylinders, and displacement can be useful predictors for mpg: cars with lower weight and lower displacememnt tend to have higher mpgs, cars with more cylinders tend to have less mpg, and cars with a higher acceleration tend to have a higher mpg.