Homework 3

Kris Hanus, Laura Glathar, Arkya Rakshit, Jace Crist, Brandon Dlugosz
University of Nebraska at Omaha

BSAD 8700 - Business Analytics
Due: February 2, 2015

ANSWER FOR 8:

(a)
```
head(Auto,3)
```

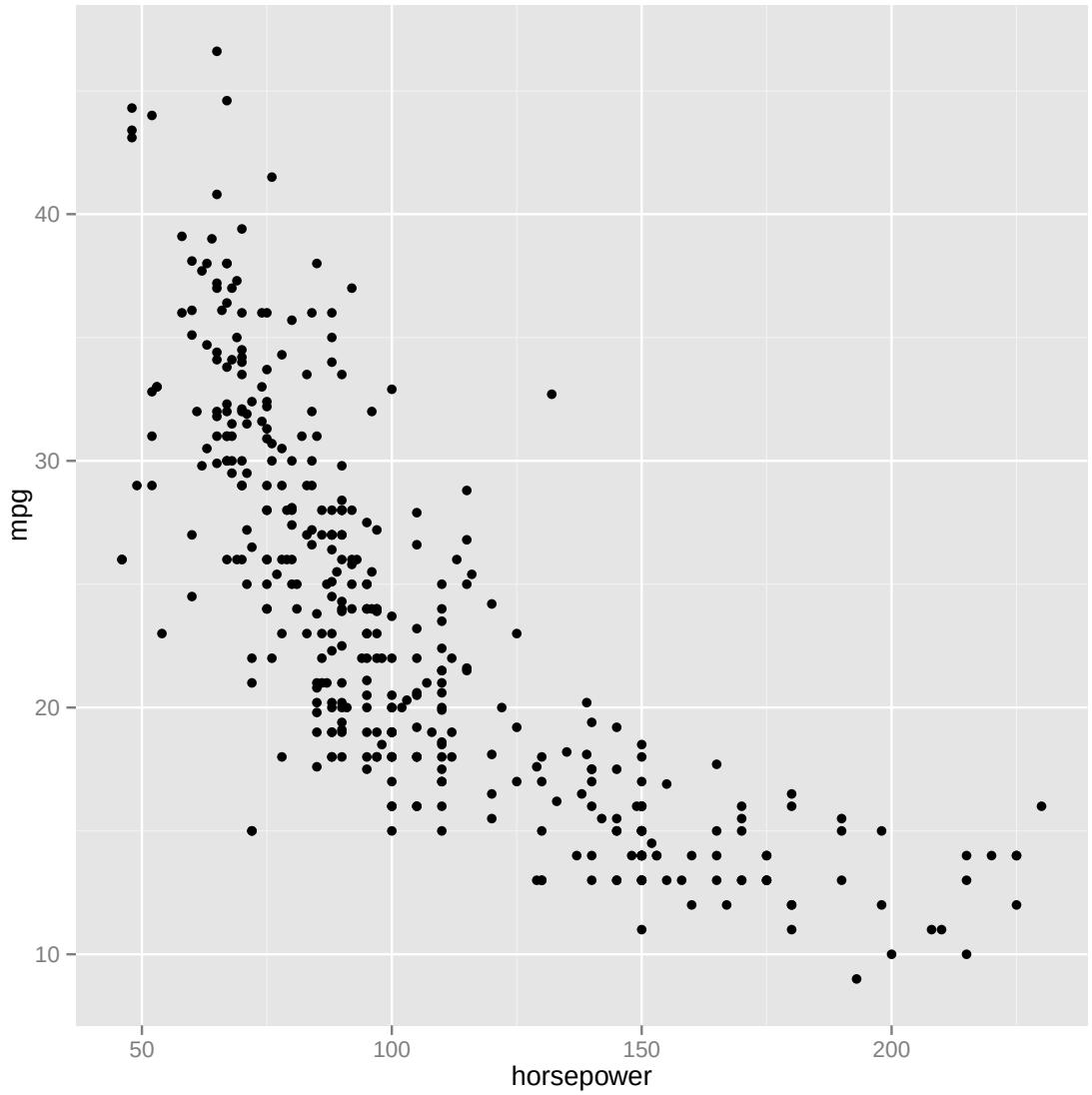
```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
```

(i)
```
predict1<-lm(mpg~horsepower, data=Auto)
summary(predict1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

There is a very strong relationship between the predictor and the response variable. However, according to the hypothesis we would suggest that there might be a transformation that would allow us to create a better prediction. This can be seen from plotting that the Beta values may not be under the assumption of linearity.

```
ggplot(Auto, aes(horsepower, mpg))+geom_point()
```



(ii)
```
fvalue<-qf(.95,390,1)
fvalue
```

```
## [1] 253.9873
```

The Adjusted R-Squared value would suggest that a linear line would fit the data fairly well, but as the F-statistic shows and p-value supports that the fit is significant. However, as can be shown by graph or residuals this is a curvi-linear plot. We would need either a transformation or an adjusted curvi-linear independent value.

(iii)
```
cor(Auto[1:8])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
```

```
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

There is a negative relationship

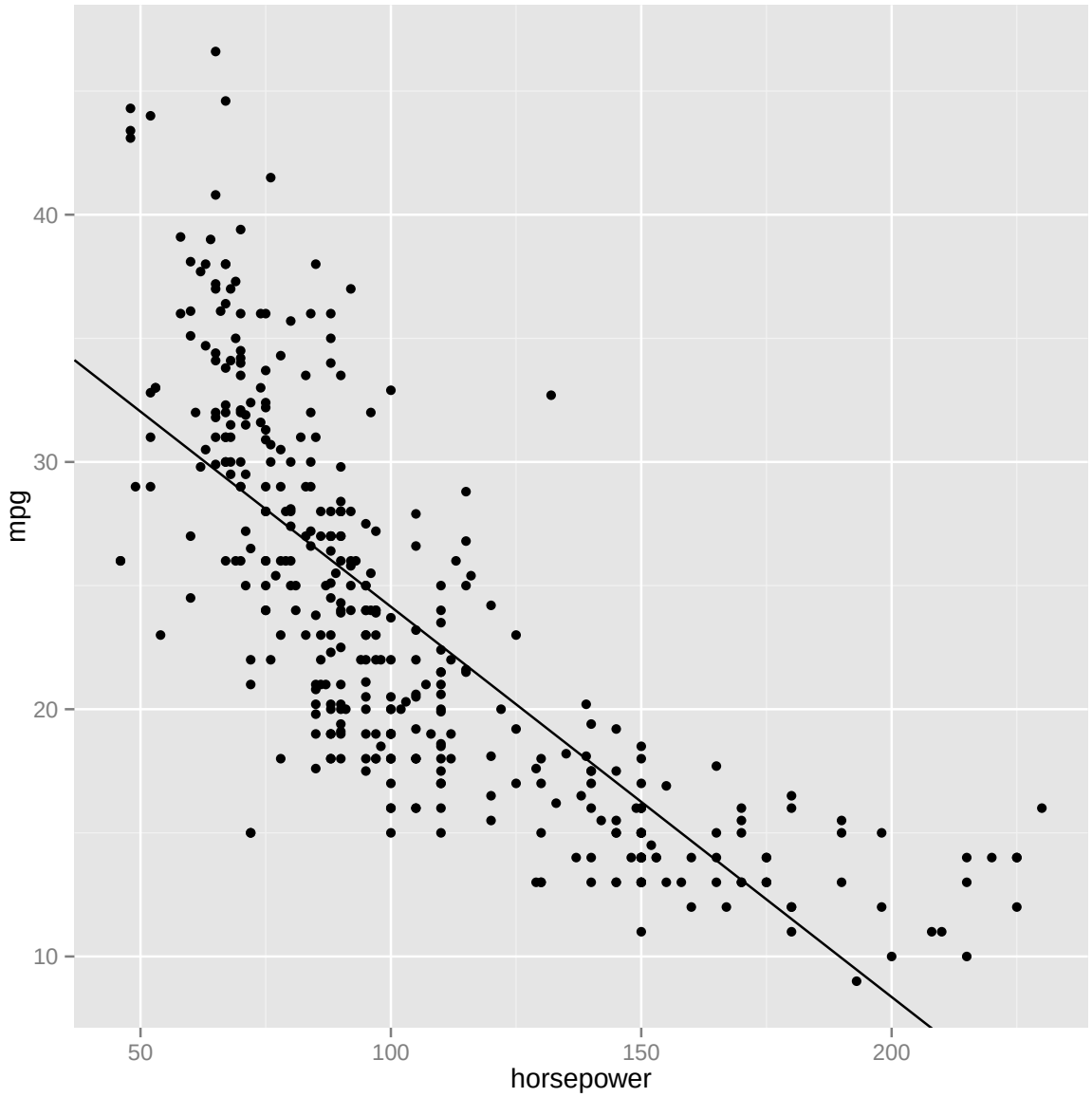
(iv)
```
NewDat<-data.frame(horsepower=98)
predict(predict1, NewDat,interval = "confidence", level=0.95, se.fit=FALSE)
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```
predict(predict1, NewDat,interval = "prediction", level=0.95, se.fit=FALSE)
```

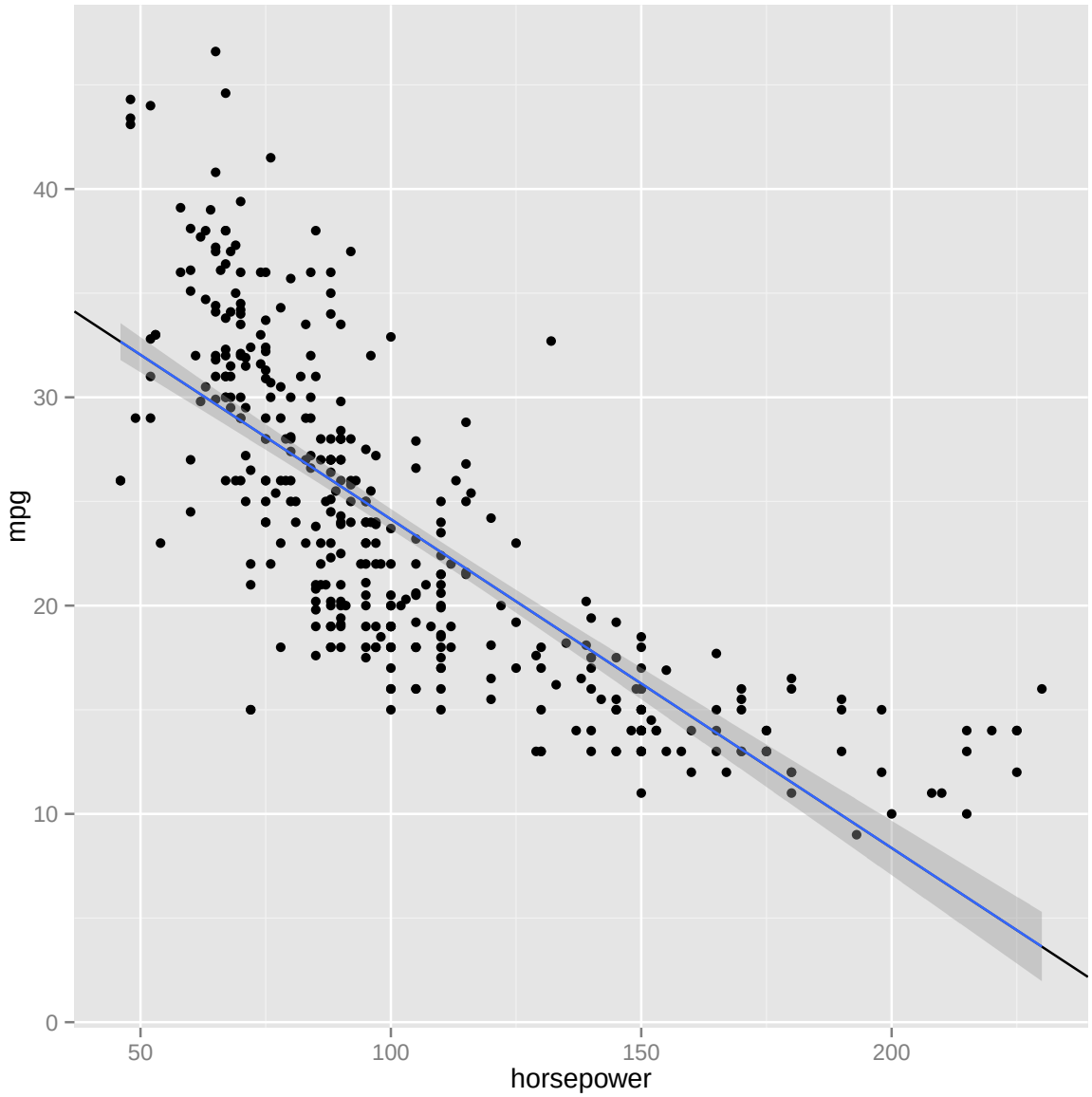
```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

(b)
```
ggplot(Auto, aes(horsepower,mpg))+
  geom_point()+
  geom_abline(intercept=39.93, slope=-0.157845)
```



(c) Extra Credit:

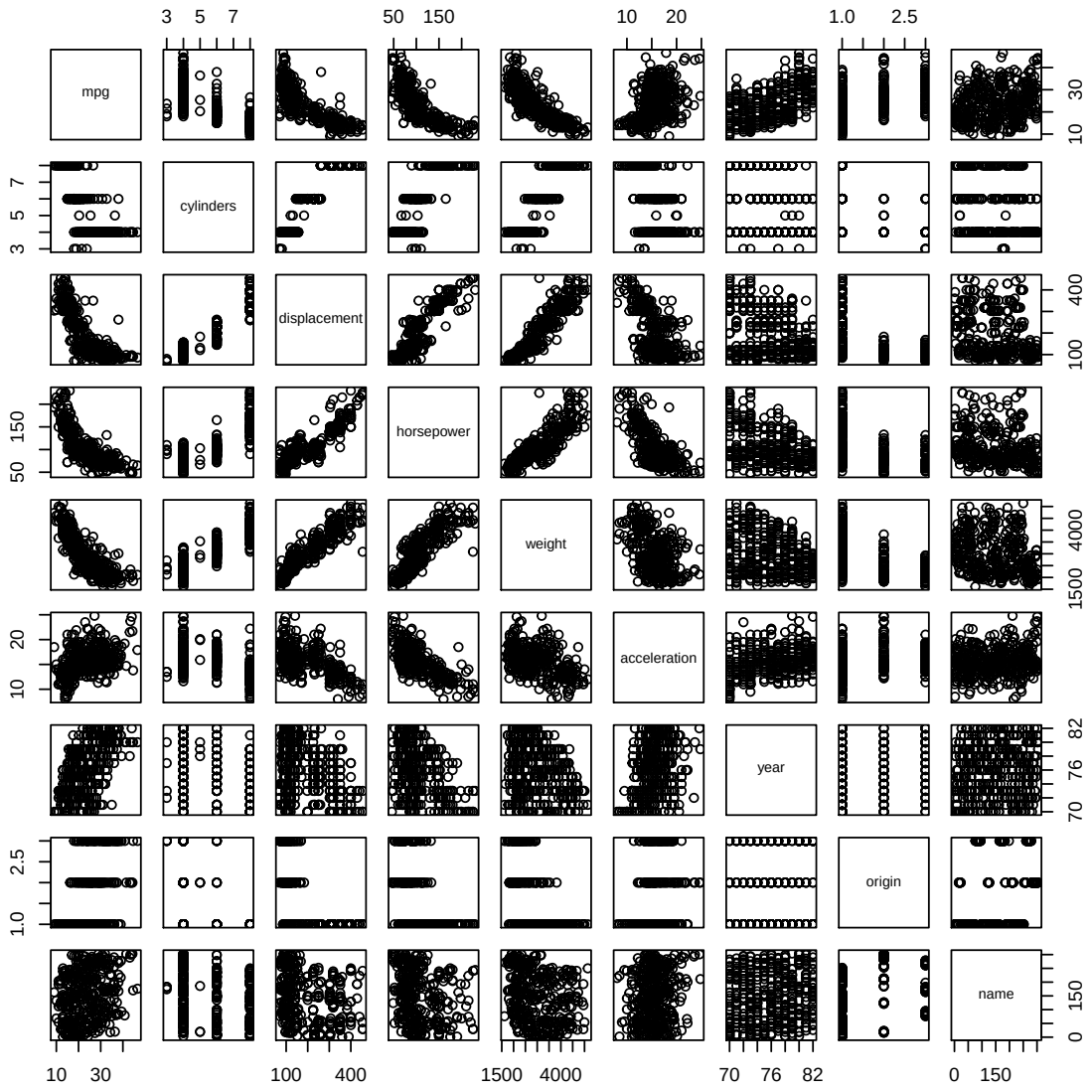
```
ggplot(Auto, aes(horsepower,mpg))+geom_point()+
  geom_abline(intercept=39.93, slope=-0.157845)+
  geom_smooth(method="lm", se=T)
```



The fitted line goes through the mean. It is evident that the residuals of the points (from left to right) would start off positive, would then become negative, and will finish positive. By looking at this trend we would suggest that the data is either logistic or $\frac{1}{x}$ and there might be a better fit model to the data.

ANSWER FOR 9:

(a)
```
pairs(Auto)
```



(b)
```
str(Auto)

## 'data.frame': 392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 2

cor(Auto[1:8])
```

```
##                      mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

(c) (i)

```
lm.fit<-lm(mpg~.,data=Auto[1:8])
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = Auto[1:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215,Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
lm.fit<-lm(mpg~.-acceleration-cylinders-horsepower-displacement,data=Auto[1:8])
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - acceleration - cylinders - horsepower -
```

```
##     displacement, data = Auto[1:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9440 -2.0948 -0.0389  1.7255 13.2722
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.805e+01  4.001e+00  -4.510 8.60e-06 ***
## weight      -5.994e-03  2.541e-04 -23.588  < 2e-16 ***
## year         7.571e-01  4.832e-02  15.668  < 2e-16 ***
## origin       1.150e+00  2.591e-01   4.439 1.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.348 on 388 degrees of freedom
## Multiple R-squared:  0.8175,Adjusted R-squared:  0.816
## F-statistic: 579.2 on 3 and 388 DF,  p-value: < 2.2e-16
```

There is some relationship between predictors and the response variable.

(ii) The variables which have the highest correlation would be weight, year, and origin. They are all significant to within 99.9%. There were two other variables (horsepower and displacement) that would have been significant together to a 90%, however; we are looking for correlations above 95% and preventing over-fitting the model they were eliminated.

(iii) The β_{year} is equal to .7571. This is a positive correlation between year and mpg.

(d) Extra Credit

(e) Extra Credit

(f) Extra Credit