Analytics 512 Homework 2

Arif Ali

02/09/2016

Exercise 3

Part A

The equation is set up as the following:

$\hat{y} = 50 + GPA * \beta_1 + IQ * \beta_2 + Gender * \beta_3 + (GPA * IQ) * \beta_4 + (GPA * Gender) * \beta_5$

By putting in the beta values, we get:

$\hat{y} = 50 + GPA * 20 + IQ * 0.07 + Gender * 35 + (GPA * IQ) * 0.01 + (GPA * Gender) * -10$

From the updated $\hat{y}$, we know that i and ii are wrong because depending on the value of the GPA, females could make more.

Part B

In [1]:
```
50+20*4+110*0.07+1*35+110*4*0.01+4*1*-10
```

Out[1]: 137.1

Part C

This isn't true, LASSO regression incorporates variable selection by adding a coefficient of zero for predictors to are not statistically significiant. The p-value needs to be computed for each of the predictors first.

Exercise 4

Part A

Based on the equations, I would expect that cubic regression model would have a lower RSS compared to the simple linear regression.

Exercise 8

In [2]:
```
library(ISLR)
data(Auto)
```

Part A

In [6]:
```
horsepower.lm = lm(mpg~horsepower, data = Auto)
summary(horsepower.lm)
confint(horsepower.lm, level = 0.95)
predict(horsepower.lm, interval = "confidence")[Auto$horsepower==98,]
```

Out[6]:
```
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
```

Out[6]:

	2.5 %	**97.5 %**
(Intercept)	38.52521	41.34651
horsepower	-0.1705170	-0.1451725

Out[6]:

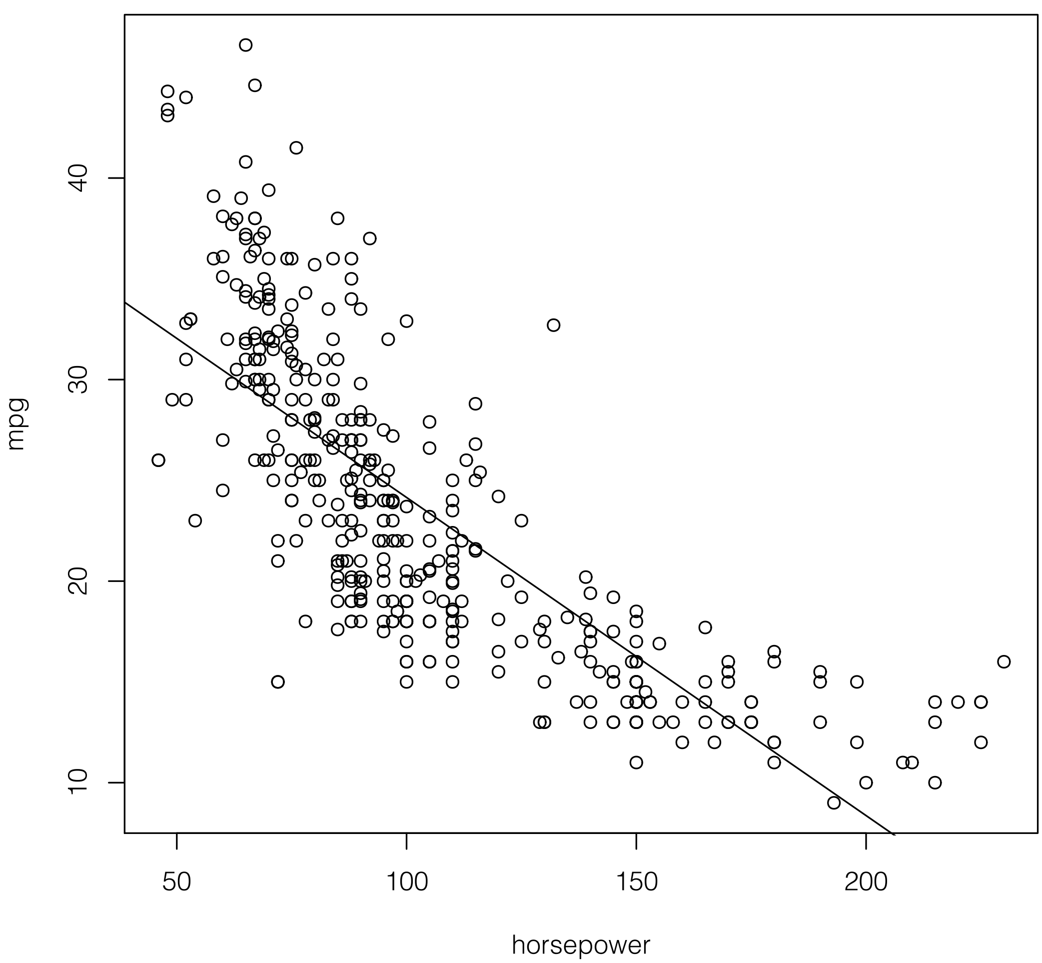
	fit	**lwr**	**upr**
180	24.46708	23.97308	24.96108
229	24.46708	23.97308	24.96108

i/ii: Based on the F-statistic and the p-value, there is a strong relationship between the predictor (horsepower) and the response variable (mpg)

iii: The Coefficient is negative which indicates a negative relationship between the predictor and response

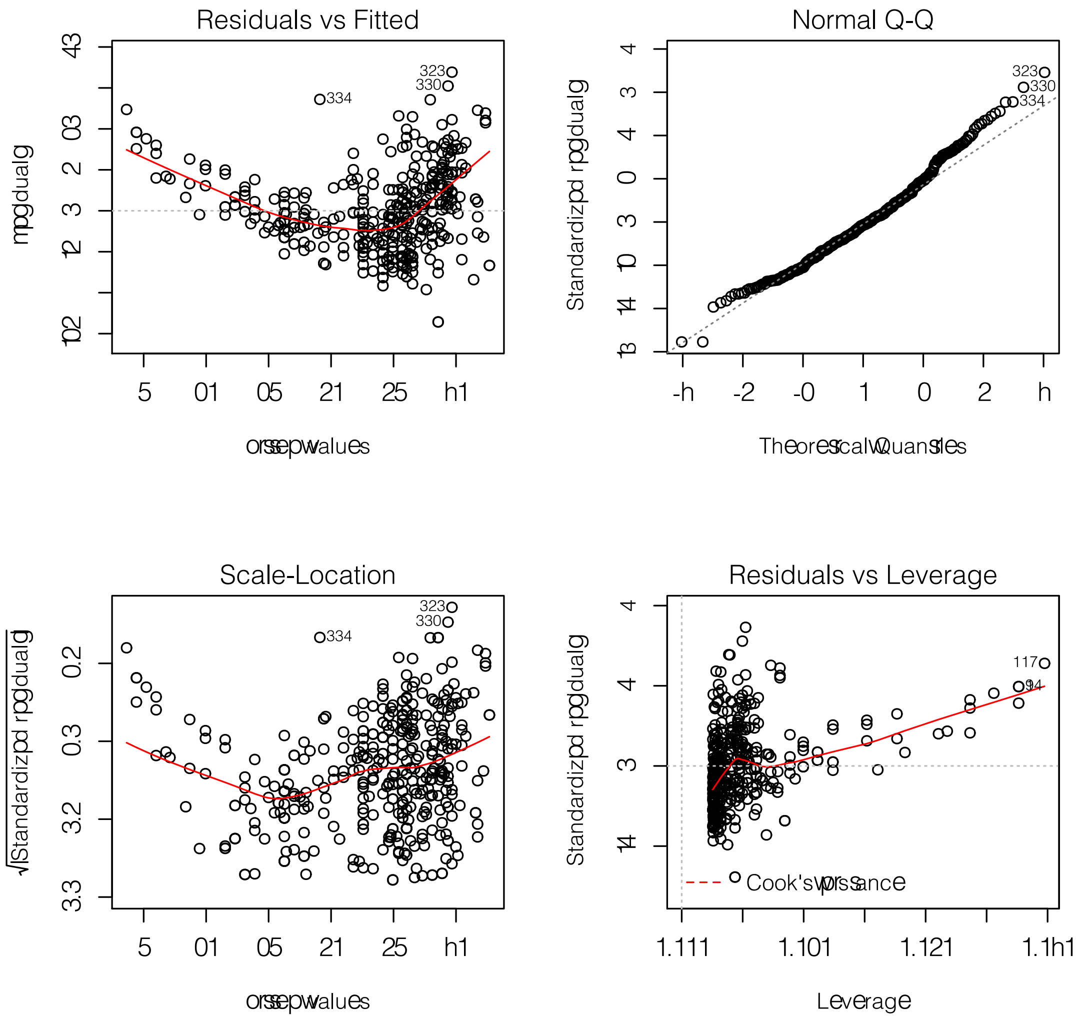
Part B

In [35]:
```
plot(mpg~horsepower, data = Auto)
abline(horsepower.lm)
```



Part C

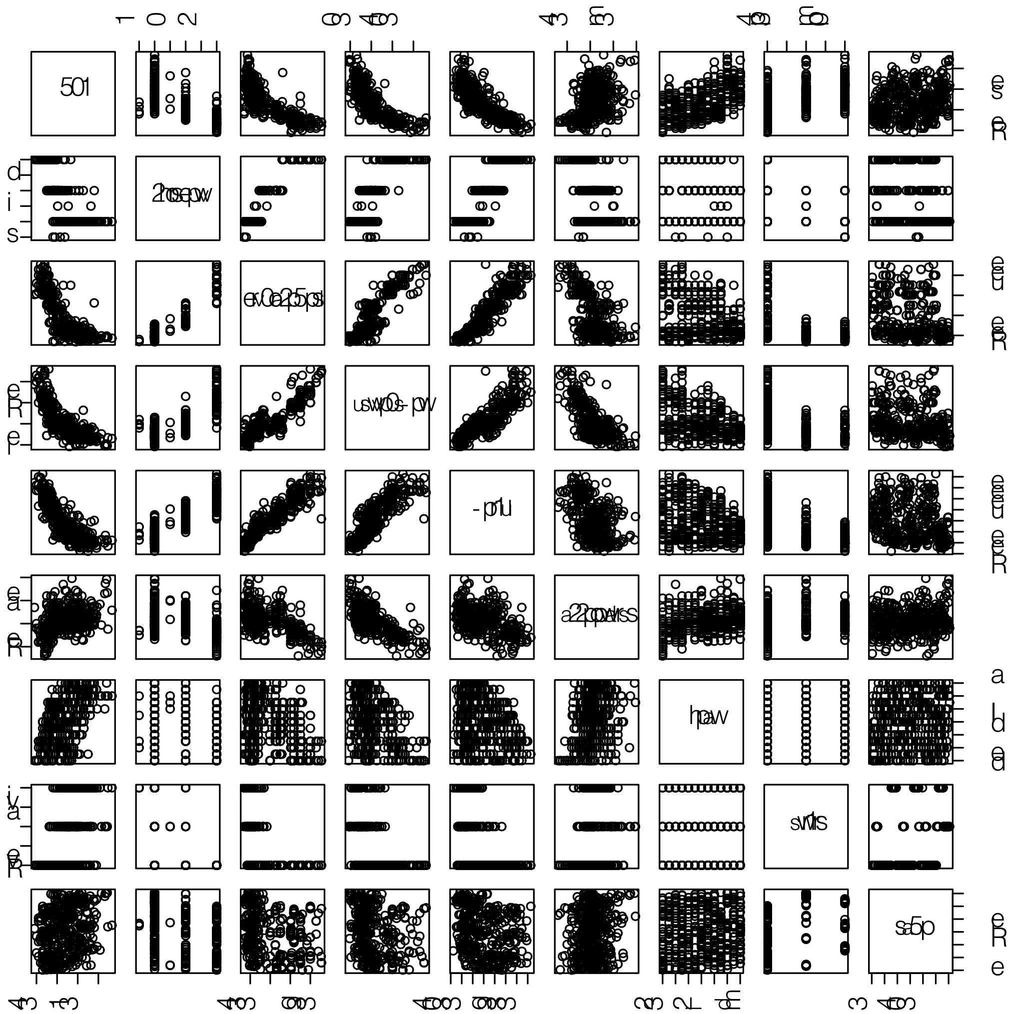
```
In [36]: par(mfrow = c(2,2))
         plot(horsepower.lm)
```



Exercise 9

Part A

In [3]:
```
pairs(Auto)
```



Part B

In [4]:
```
cor(Auto[,-ncol(Auto)])
```

Out[4]:

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

Part C

In [7]:
```
auto.lm = lm(mpg~.,data=Auto[,-ncol(Auto)])
summary(auto.lm)
```

Out[7]:
```
Call:
lm(formula = mpg ~ ., data = Auto[, -ncol(Auto)])

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
```

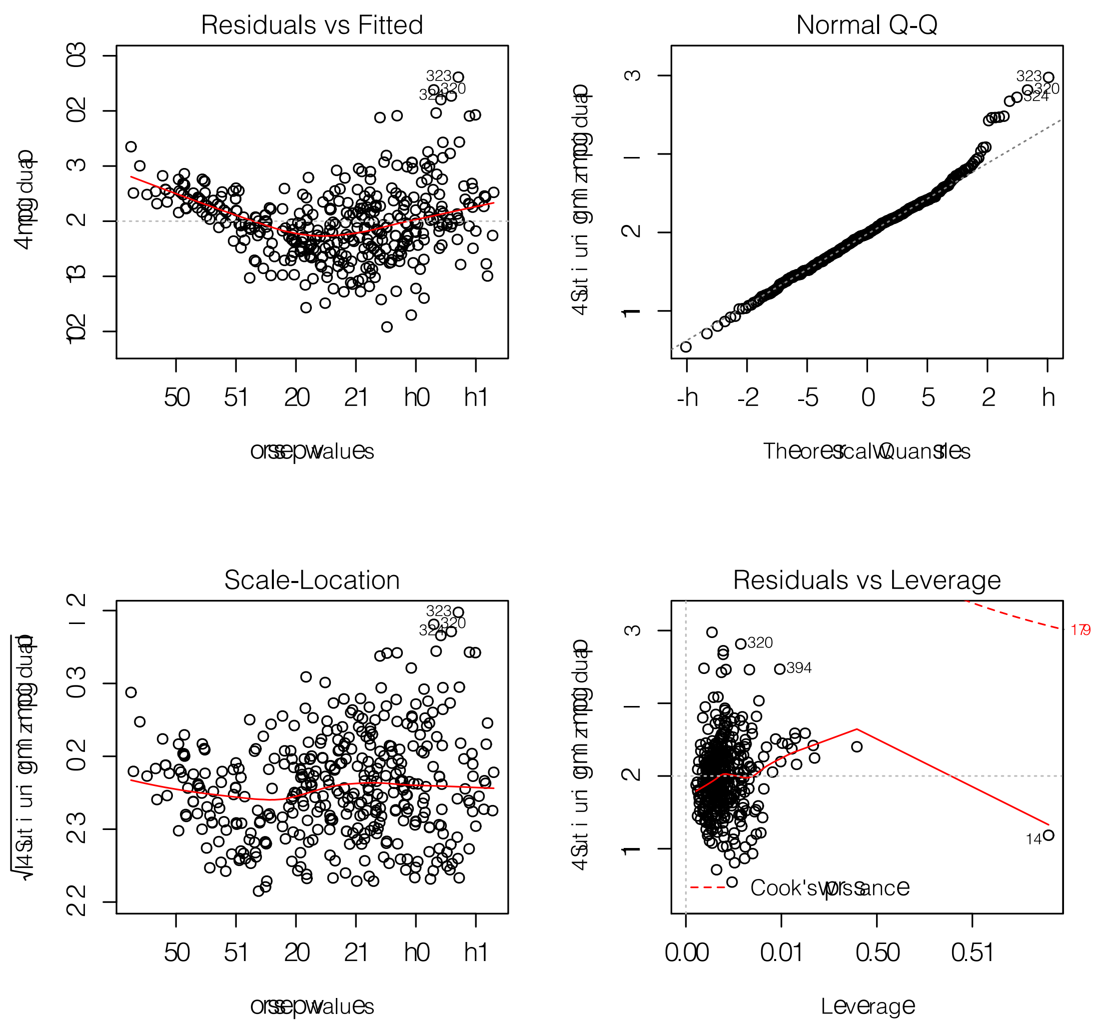
i: The F-statistic is very high and the p-value associated with it is very low, so there is an overall strong relationship between the predictors and response variable (mpg)

ii: The Predictors with regards to displacement, weight, year, and orgin are statistically significant with respect to mpg.

iii: The coefficient is positive, so the newer the car, the better the mpg.

Part D

```
In [8]: par(mfrow = c(2,2))
        plot(auto.lm)
```



Point 14 seems to have some high leverage as opposed to 327 and 394 which which noted are not that far out as 14. From the normal Q-Q plot indicates that the standardized residuals do not follow a normal distribution.

Part E

For part E and F, I got rid of the non-significant predictors.

In [17]:
```
auto.lm.interaction = lm(mpg~(displacement:weight)+(year:origin),data=Auto[,-ncol(Auto)])
summary(auto.lm.interaction)
```

Out[17]:
```
Call:
lm(formula = mpg ~ (displacement:weight) + (year:origin), data = Auto[,
    -ncol(Auto)])

Residuals:
    Min      1Q  Median      3Q     Max
-13.198  -2.832  -0.279   2.193  16.860

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.679e+01  8.319e-01  32.200  < 2e-16 ***
displacement:weight -9.940e-06  5.398e-07 -18.416  < 2e-16 ***
year:origin          2.690e-02  4.471e-03   6.016 4.14e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.562 on 389 degrees of freedom
Multiple R-squared:  0.6601,    Adjusted R-squared:  0.6584
F-statistic: 377.8 on 2 and 389 DF,  p-value: < 2.2e-16
```

Part F

In [25]:
```
auto.lm.transformation = lm(mpg~log(displacement)+weight+
                             sqrt(year)+I(origin)^2,data=Auto[,-ncol(Auto)])
summary(auto.lm.transformation)
```

Out[25]:
```
Call:
lm(formula = mpg ~ log(displacement) + weight + sqrt(year) +
    I(origin)^2, data = Auto[, -ncol(Auto)])

Residuals:
     Min       1Q   Median       3Q      Max
-10.8260  -1.9314  -0.0845   1.7774  13.2013

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.069e+01  8.953e+00  -6.779 4.54e-11 ***
log(displacement) -2.982e+00  1.006e+00  -2.964  0.00322 **
weight            -4.483e-03  5.712e-04  -7.849 4.17e-14 ***
sqrt(year)         1.280e+01  8.433e-01  15.181  < 2e-16 ***
I(origin)          7.782e-01  2.860e-01   2.721  0.00681 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.322 on 387 degrees of freedom
Multiple R-squared:  0.8207,    Adjusted R-squared:  0.8188
F-statistic: 442.7 on 4 and 387 DF,  p-value: < 2.2e-16
```

Exercise 12

Part A

Part B

In [9]:
```
X = rnorm(100)
Y = rpois(n = 100, lambda = 2)
train = data.frame(X,Y)
```

In [12]:
```
summary(lm(Y~X, data = train))
```

Out[12]:
```
Call:
lm(formula = Y ~ X, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1529 -1.0049 -0.0166  0.9011  5.2616 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.9968     0.1398  14.288   <2e-16 ***
X            -0.1312     0.1289  -1.018    0.311    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.396 on 98 degrees of freedom
Multiple R-squared:  0.01046,	Adjusted R-squared:  0.00036 
F-statistic: 1.036 on 1 and 98 DF,  p-value: 0.3113
```

In [13]:
```
summary(lm(X~Y, data = train))
```

Out[13]:
```
Call:
lm(formula = X ~ Y, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9381 -0.7046 -0.0222  0.7164  2.7078 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.21037    0.19005   1.107    0.271
Y           -0.07969    0.07831  -1.018    0.311

Residual standard error: 1.088 on 98 degrees of freedom
Multiple R-squared:  0.01046,	Adjusted R-squared:  0.00036 
F-statistic: 1.036 on 1 and 98 DF,  p-value: 0.3113
```

Part C

In [2]:
```
X = rnorm(100)
Y = X
train = data.frame(X,Y)
```

In [3]:
```
summary(lm(Y~X, data = train))
```

```
Warning message:
In summary.lm(lm(Y ~ X, data = train)): essentially perfect fit: summary may be unreliable
```

Out[3]:
```
Call:
lm(formula = Y ~ X, data = train)

Residuals:
       Min         1Q     Median         3Q        Max 
-3.636e-16 -2.366e-17  7.170e-18  3.854e-17  1.714e-16 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept) 0.000e+00  7.478e-18 0.00e+00        1    
X           1.000e+00  7.353e-18 1.36e+17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.441e-17 on 98 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 1.85e+34 on 1 and 98 DF,  p-value: < 2.2e-16
```

In [4]:
```
summary(lm(X~Y, data = train))
```

```
Warning message:
In summary.lm(lm(X ~ Y, data = train)): essentially perfect fit: summary may be unreliable
```

Out[4]:
```
Call:
lm(formula = X ~ Y, data = train)

Residuals:
       Min         1Q     Median         3Q        Max 
-3.636e-16 -2.366e-17  7.170e-18  3.854e-17  1.714e-16 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept) 0.000e+00  7.478e-18 0.00e+00        1    
Y           1.000e+00  7.353e-18 1.36e+17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.441e-17 on 98 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 1.85e+34 on 1 and 98 DF,  p-value: < 2.2e-16
```