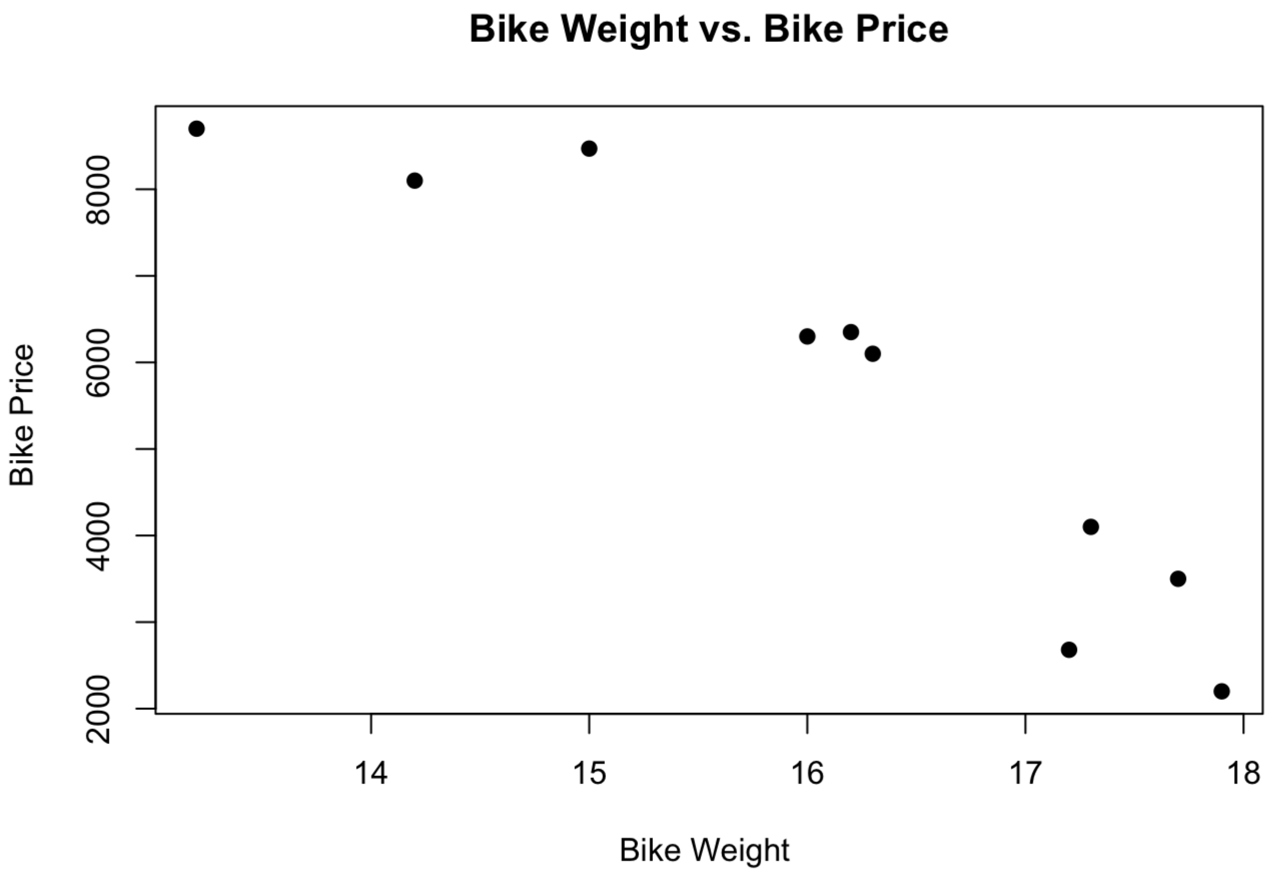
# Regression Analysis

Problem 1: Bicycling World Problem 1. Develop a scatter chart with weight as the independent variable. What does the scatter chart indicate about the relationship between the weight and price of these bicycles?

```
bike <- data.frame(
Model = c("Fierro 7B","HX 5000","Durbin Ultralight","Schmidt","WSilton Advanced",
"bicyclette vélo","Supremo Team","XTC Racer","D’Onofrio Pro","Americana #6"), Price = c(
2200,6350,8470,6300,4100,8700,6100,2680,3500,8100),
Weight = c(17.9,16.2,15,16,17.3,13.2,16.3,17.2,17.7,14.2),
stringsAsFactors = FALSE
)
plot(bike$Weight,bike$Price,main="Bike Weight vs. Bike Price", xlab="Bike Weight ", ylab
="Bike Price ", pch=19)
```



A

negative linear relationship can be obsered between the weight and price of bucycle. Which indicates lighter the bicycle higher is the price of it.

2. Use the data to develop an estimated regression equation that could be used to estimate the price for a bicycle, given its weight. What is the estimated regression model?

```
x_sq = (bike$Weight)^2
y_sq = (bike$Price)^2
xy = bike$Weight*bike$Price
S_yy = 10*sum(y_sq) - sum(bike$Price)^2
S_xx = 10*sum(x_sq) - sum(bike$Weight)^2
S_xy = 10*sum(xy) - sum(bike$Price)*sum(bike$Weight)
b_1 = S_xy/S_xx
b_0 = (sum(bike$Price)-b_1*sum(bike$Weight))/10
S_yy
```

```
## [1] 521208000
```

```
S_xx
```

```
## [1] 217.4
```

```
b_1
```

```
## [1] -1439.006
```

```
b_0
```

```
## [1] 28818
```

slope of regression equation is b1 = Sxy / Sxx= –1439.01 and intercept of the equation will be b0= 28818. So the regression equation will be y = 28818.00 – 1439.01x

3. Test whether each of the regression parameters and is equal to zero at a 0.05 level of significance. What are the correct interpretations of the estimated regression parameters? Are these interpretations reasonable?

```
bike_model = lm(bike$Price~bike$Weight,bike)
summary(bike_model)
```

```
##
## Call:
## lm(formula = bike$Price ~ bike$Weight, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1387.1  -715.9   164.6   679.9  1237.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28818.0     3267.3   8.820 2.15e-05 ***
## bike$Weight  -1439.0      202.1  -7.121 9.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 942.3 on 8 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8467
## F-statistic:  50.7 on 1 and 8 DF,  p-value: 9.994e-05
```

Since p-value is less than 0.05 so intercept is significant to the model.

4. How much of the variation in the prices of the bicycles in the sample does the regression model you estimated in part b explain? Since from regression output R-square is 0.864 so 86.4% of the variation in the prices of the bicycles is accounted for by the weight of bicycles.
5. The manufacturers of the D’Onofrio Pro plan to introduce the 15-pound D’Onofrio Elite bicycle later this year. Use the regression model you estimated in part a to predict the price of the D’Ononfrio Elite.

```
-15*1439.01+ 28818.00
```

```
## [1] 7232.85
```

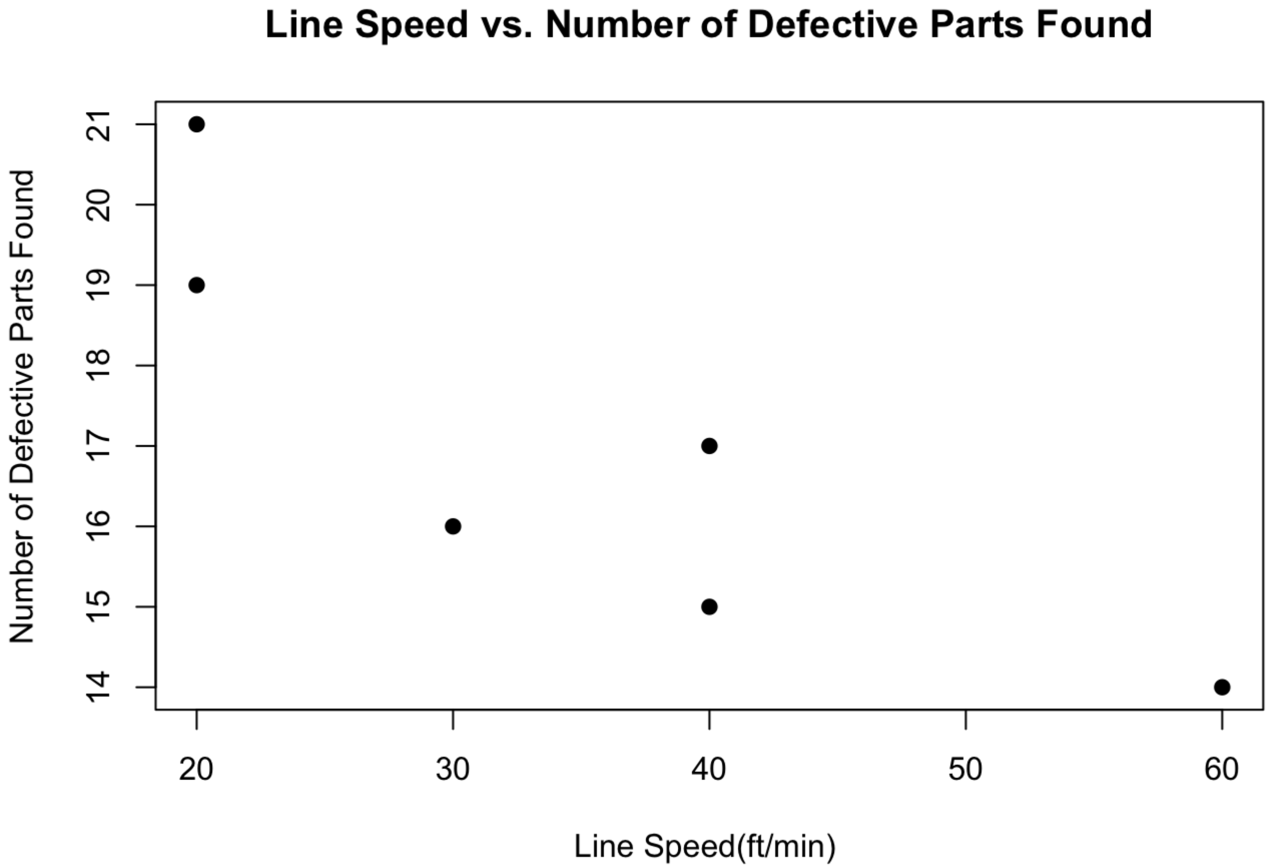
For X= 15 estimated y value is y = 28818.00 – 1439.01 ∗ 15 = 7232.85 hence, required predicted price is $7232.85.

Problem 2: Assembly Line Problem

1. Develop the estimated regression equation that relates line speed to the number of defective parts found.

```
man <- data.frame(line_speed = c(20,20,40,30,60,40), numberofdefectivepart = c(21,19,15,
16,14,17), stringsAsFactors = FALSE)

plot(man$line_speed,man$numberofdefectivepart,main="Line Speed vs. Number of Defective P
arts Found",
   xlab="Line Speed(ft/min) ", ylab="Number of Defective Parts Found", pch=19)
```



Positive correlation between line speed and number of defective parts found

2.Use the data to develop an estimated regression equation that could be used to predict the number of defective parts found, given the line speed. What is the estimated regression model?

```
man_model = lm(man$numberofdefectivepart~man$line_speed,man)
summary(man_model)
```

```
##
## Call:
## lm(formula = man$numberofdefectivepart ~ man$line_speed, data = man)
##
## Residuals:
##       1       2       3       4       5       6
##  1.7826 -0.2174 -1.2609 -1.7391  0.6957  0.7391
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    22.17391    1.65275  13.416 0.000179 ***
## man$line_speed -0.14783    0.04391  -3.367 0.028135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.489 on 4 degrees of freedom
## Multiple R-squared:  0.7391, Adjusted R-squared:  0.6739
## F-statistic: 11.33 on 1 and 4 DF,  p-value: 0.02813
```

y=b0+b1∗xb0 =y−intercept=22.17391b1 =regressioncoefficient of y on x or slopeoftheregression= -0.14783

3. Did the estimated regression equation provide a good fit to the data? The coefficient of determination is r2 = 0.739, which shows that only 73.9% of regression can be explained. So the data is relatively a good fit.

4. Develop a 95% confidence interval to predict the mean number of defective parts for a line speed of 50 feet per minutes. For x=50 feet per minute, predicted number of defective parts will be $ y = 22.1729 - 0.1478 *50 = 14.7829$ Here degree of freedom for t will be df=n-2=4 so for 95% confidence interval critical value of t will be 2.7764.

```
14.7829-2.7764*1.489*sqrt(1/6+((50-35)^2)/1150)
```

```
## [1] 12.29449
```

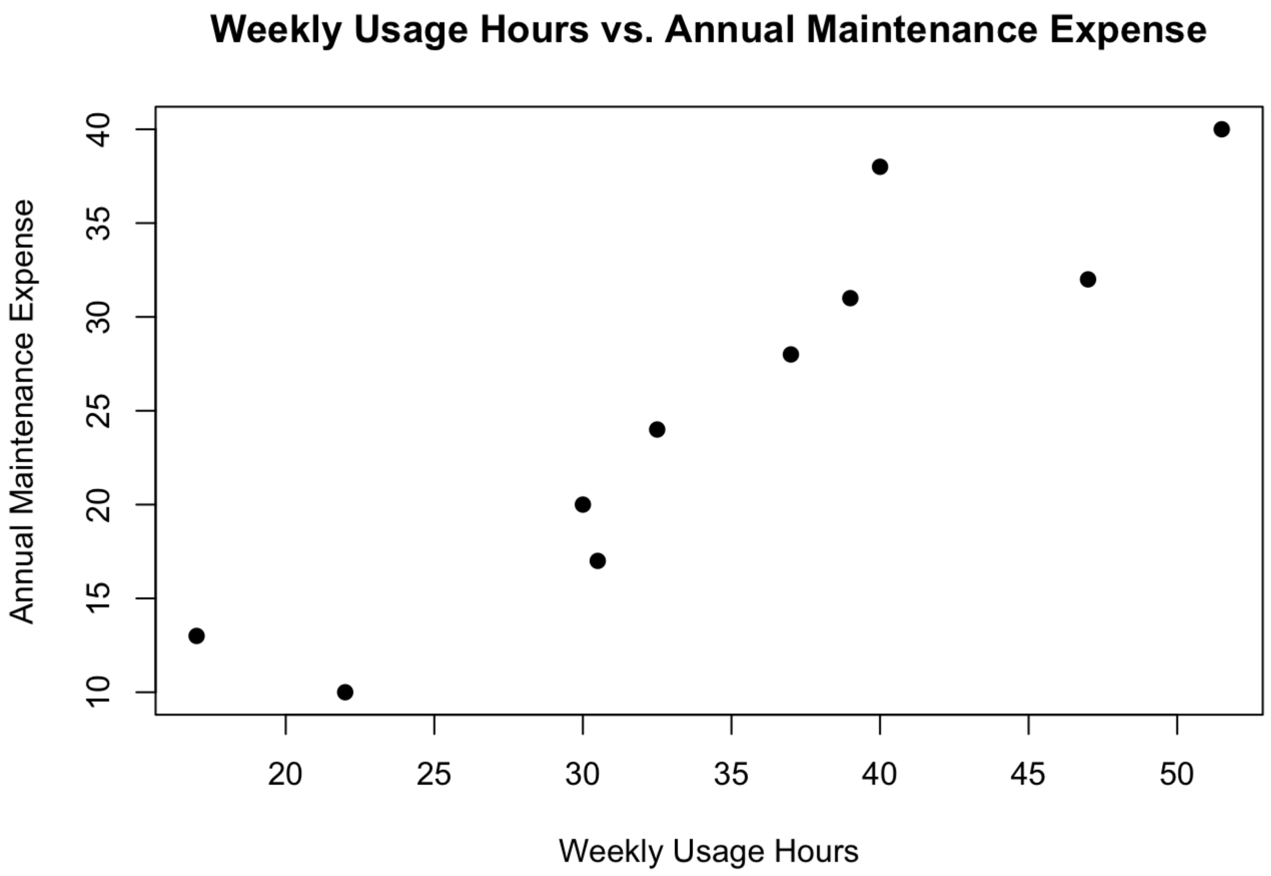
```
14.7829+2.7764*1.489*sqrt(1/6+((50-35)^2)/1150)
```

```
## [1] 17.27131
```

So required confidence interval is (12.2945, 17.2713).

Problem 3: Jensen Tire & Auto Problem Jensen Tire & Auto is deciding whether to purchase a maintenance contract for its new computer wheel alignment and balancing machine. Managers feel that maintenance expense should be related to usage, and they collected the following information on weekly usage (hours) and annual maintenance expense (in hundreds of dollars). 1. Develop a scatter chart with weekly usage hours as the independent variable. WHat does the scatter chart indicate about the relationship between weekly usage and annual maintenance expense?

```
jt <- data.frame(Weekly = c(13,10,20,28,32,17,24,31,40,38), expense = c(17,22,30,37,47,3
0.5,32.5,39,51.5,40), stringsAsFactors = FALSE)
plot(jt$expense,jt$Weekly,main="Weekly Usage Hours vs. Annual Maintenance Expense",xlab=
"Weekly Usage Hours", ylab="Annual Maintenance Expense", pch=19)
```



Positive correlation between Weekly usage Hours and Annual Maintenance Expense. 2. Use the data to develop an estimated regression equation that could be used to predict the annual maintenance expense for a given number of hours of weekly usage. What is the estimated regression model?

```
jensen_model = lm(jt$expense~jt$Weekly,jt)
summary(jensen_model)
```

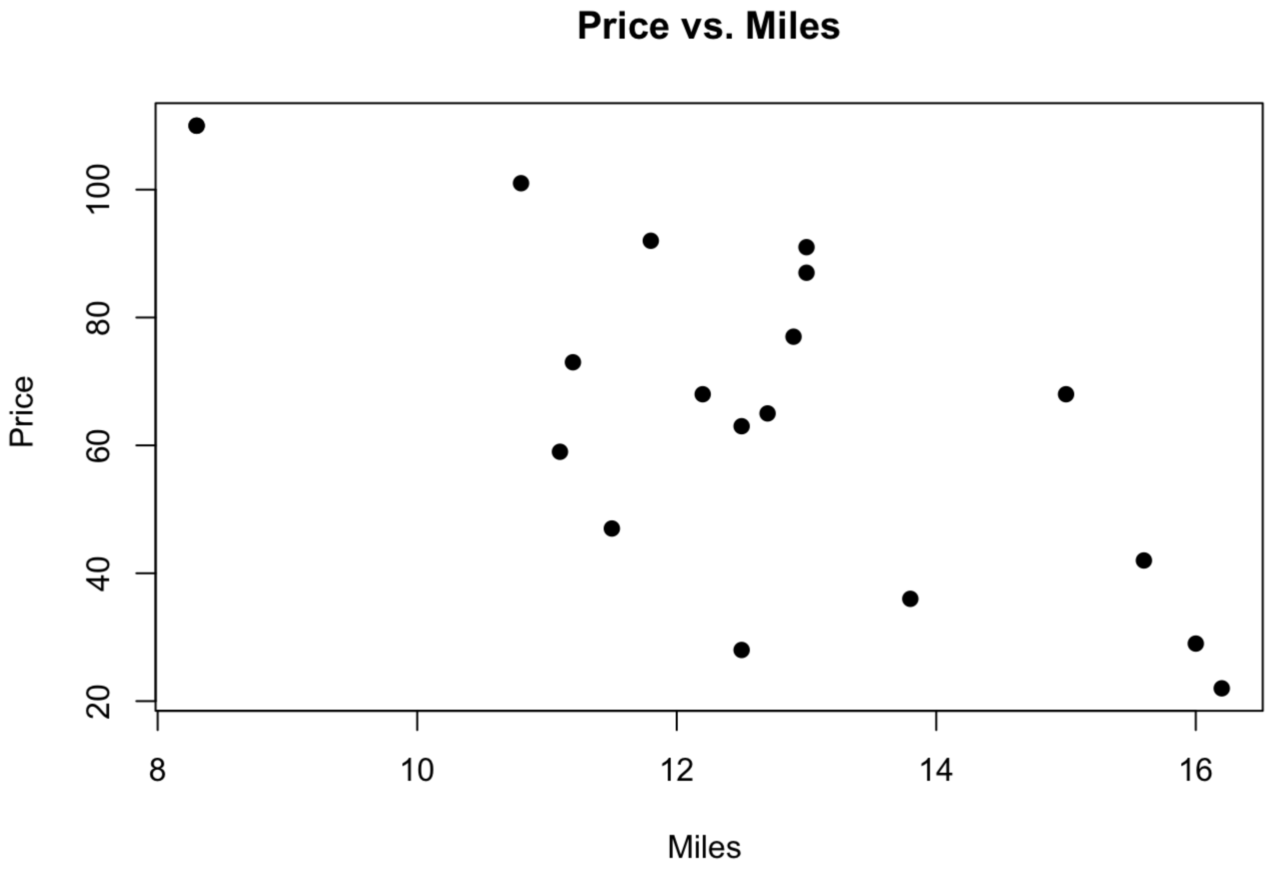
```
##
## Call:
## lm(formula = jt$expense ~ jt$Weekly, data = jt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7587 -1.0411  0.0895  2.6102  5.9619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5280     3.7449   2.811 0.022797 *
## jt$Weekly     0.9534     0.1382   6.901 0.000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.25 on 8 degrees of freedom
## Multiple R-squared:  0.8562, Adjusted R-squared:  0.8382
## F-statistic: 47.62 on 1 and 8 DF,  p-value: 0.0001244
```

$y = b0 + b1 * x = 10.5280 + 0.9534 * x$ This is the estimated regression equation for this problem.

Problem 4: Toyota Problem The Toyota Camry is one of the best-selling cars in North America. The cost of a previously owned Camry depends upon many factors, including the model year, mileage, and condition. To investigate the relationship between the car's mileage and the sales price for a 2007 model year Camry, the following data show the mileage and sale price for 19 sales (PriceHub website, February 24, 2012).

1. Develop a scatter chart for these data with miles as the independent variable. What does the scatter chart indicate about the relationship between price and the miles?

```
to <- data.frame(
Miles = c(22,29,36,47,63,77,73,87,92,101,110,28,59,68,68,91,42,65,110), Price = c(16.2,1
6,13.8,11.5,12.5,12.9,11.2,
             13,11.8,10.8,8.3,12.5,11.1,15,12.2,13,15.6,12.7,8.3),
   stringsAsFactors = FALSE
)
plot(to$Price,to$Miles,main="Price vs. Miles", xlab="Miles ", ylab="Price ", pch=19)
```



2. Use the data to develop an estimate regression equation that could be used to predict the number of defective parts found, given the line speed. What is the estimated regression model?

```
toy_model = lm(to$Price ~ to$Miles, to)
summary(toy_model)
```

```
##
## Call:
## lm(formula = to$Price ~ to$Miles, data = to)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32408 -1.34194  0.05055  1.12898  2.52687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.46976    0.94876  17.359 2.99e-12 ***
## to$Miles    -0.05877    0.01319  -4.455 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.541 on 17 degrees of freedom
## Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
## F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

R square is 0.53, 53% Intercept is 16.46 slope is -0.05877 y = b0 + b1 ∗ x y = 16.36 + (-0.058) * x is the equation

3. Test whether each of the regression parameters β0 and β1 is equal to zero at a 0.01 level of significance. What are the correct interpretatios of the estimated regression parameters? Are these interpretation resonable?

The slope of the line is -0.05878 which means that for every 1000 miles of increase in the mielage the sales price of the Camry decreases by $58.78. P-value of intercept is 2.99e-12. Since p-value is less than 0.01 so intercept is significant to the model. P value of slope is 0.000348 which is less that 0.01 so slope is significant to the model.

4. How much of the variation in the sample values of price does the model estimated in part b explain?

Since from regression output R-square is 0.5115 so 51.15% of the variation in the prices of the car is accounted for by the miles of car.

5. For the model estimated in part b, calculate the predicted price and residual for each automobile in the data. Identify the two automobiles that were the biggest bargains.

```
predict(toy_model)
```

```
##        1        2        3        4        5        6        7        8
## 15.17673 14.76531 14.35389 13.70738 12.76700 11.94416 12.17926 11.35642
##        9       10       11       12       13       14       15       16
## 11.06255 10.53359 10.00462 14.82408 13.00209 12.47313 12.47313 11.12133
##       17       18       19
## 14.00125 12.64945 10.00462
```

```
residuals(toy_model)
```

```
##           1           2           3           4           5           6 
##  1.02327147  1.23468899 -0.55389349 -2.20738023 -0.26699732  0.95583772 
##           7           8           9          10          11          12 
## -0.97925801  1.64357704  0.73744670  0.26641209 -1.70462253 -2.32408494 
##          13          14          15          16          17          18 
## -1.90209305  2.52687234 -0.27312766  1.87867277  1.59875011  0.05055054 
##          19 
## -1.70462253
```

There are two data points with most largest residuals, which are 12th(-2.32408) and 4th(-2.207380).

6. Suppose that you are considering purchasing a previously owned Camry that has been driven 60000 miles. Use the estimate regression equation developed in part b to predict the price for this car. Is this the price you would offer the seller?

```
16.47-0.0587*60
```

```
## [1] 12.948
```