Linear Regression Part 2

Carey Kopeikin and Matt Bardoe

4/13/2020 rev 1/25/2021

Linear Regression Day 2

What you will learn:

- To review scatterplots
- To review simple linear regressions
- What a residual is and why it is important.
- How to use the summary function
- What R-squared means

Warmup

We will start by using the data set USArrests.

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

The variables are:

Variable Name	Type	Description
Murder	numeric	Murder arrests (per 100,000)
Assault	numeric	Assault arrests (per 100,000)
UrbanPop	numeric	Percent urban population
Rape	numeric	Rape arrests (per 100,000)

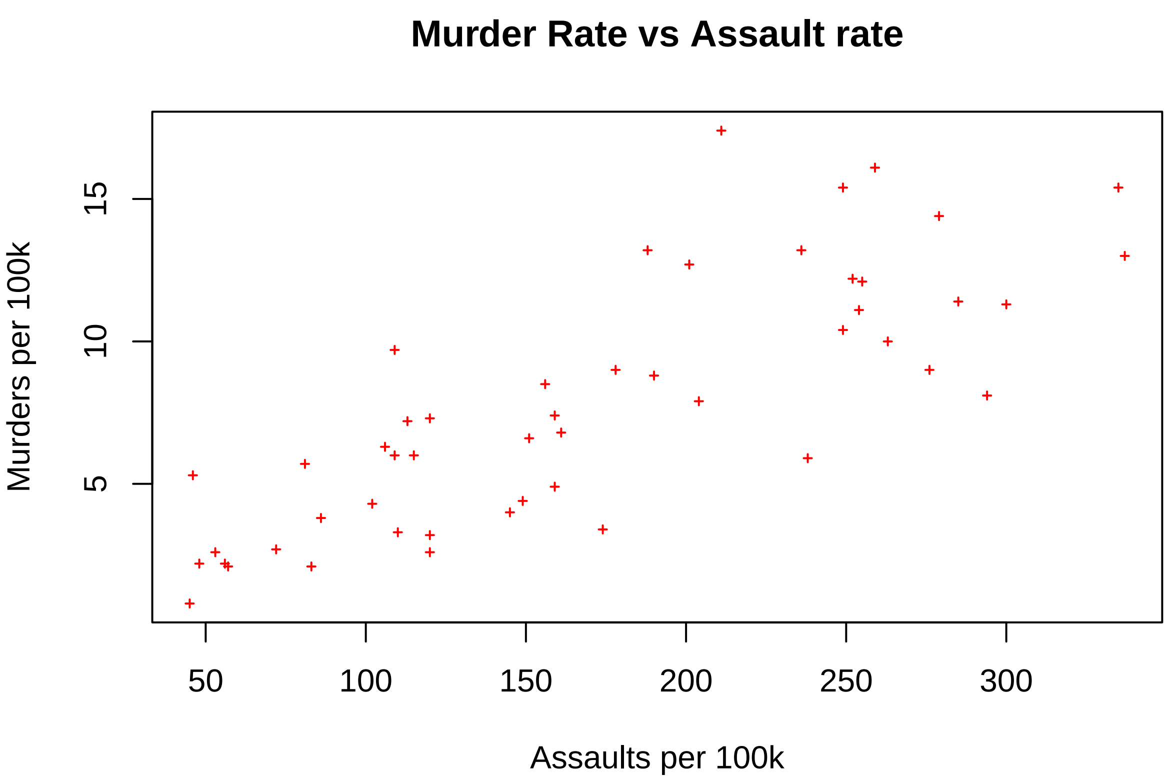
```r
data(USArrests)
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Our goal is to attempt to use the assault rate to predict the murder rate.

Create a scatterplot with labels making sure to put the explanatory and response variables on the correct axis.

```r
plot( USArrests$Murder ~ USArrests$Assault,
     main = "Murder Rate vs Assault rate",
     xlab = "Assaults per 100k",
     ylab = "Murders per 100k",
     col = "red",
     pch = 3,
     cex = .4
      )
```



Describe the shape, direction, and strength of the association:

The shape is linear, the direction is positive, the strength is strong (r ~ .8).

Is it appropriate to use correlation to talk about the relationship between these variables? Explain why or why not.

YES!!! Because the shape is linear.

Find the correlation between assault and murder.

```r
cor(USArrests$Murder, USArrests$Assault)
```

```
## [1] 0.8018733
```

What does this tell you about the strength?

It is a strong linear relationship.

Create a linear model and print out the results. Call this model: linearMod.murder.assault

```r
linearMod.murder.assault=lm(Murder~Assault, data=USArrests)
summary(linearMod.murder.assault)
```

```
##
## Call:
## lm(formula = Murder ~ Assault, data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8528 -1.7456 -0.3979  1.3044  7.9256
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.631683   0.854776   0.739    0.464
## Assault     0.041909   0.004507   9.298  2.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.629 on 48 degrees of freedom
## Multiple R-squared:  0.643,  Adjusted R-squared:  0.6356
## F-statistic: 86.45 on 1 and 48 DF,  p-value: 2.596e-12
```

Find the equation of the line of best fit and explain in context what the slope and y-intercept tell us.

$\widehat{Murder} = .042 Assault + 0.632$

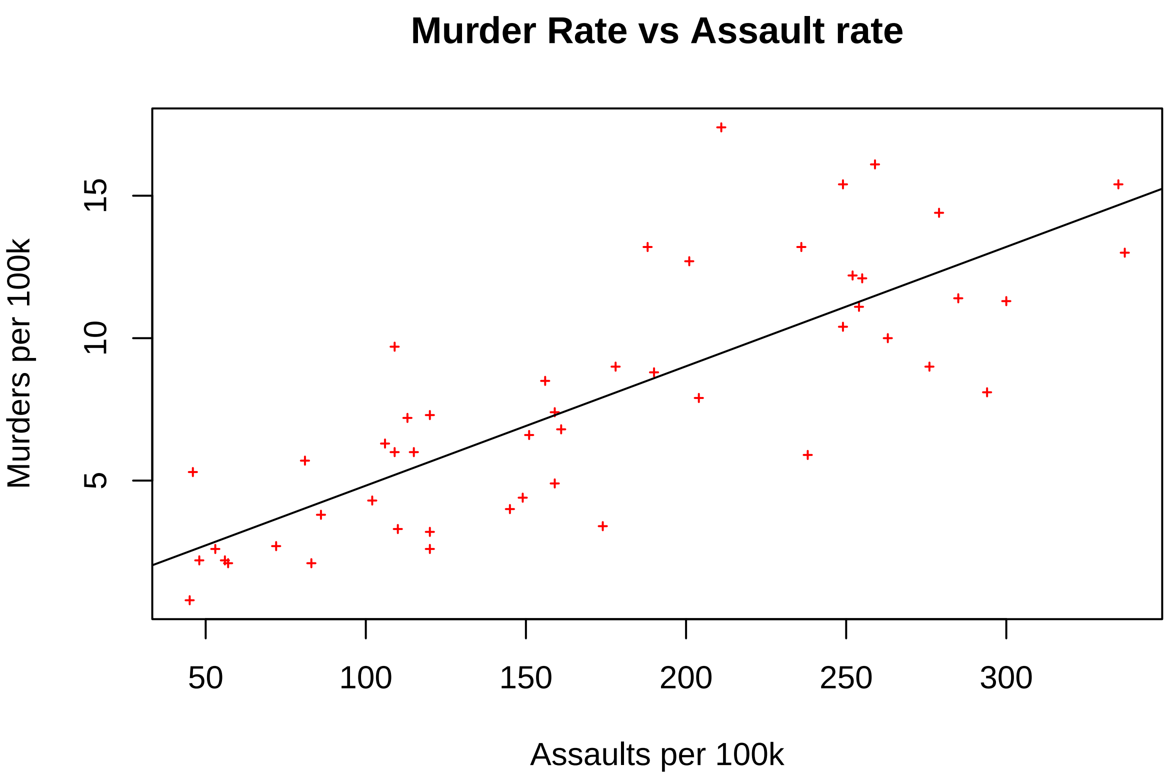
For every increase by 1 in the Assault Rate the Murder Rate increases by 0.042.

The intercept means that if the Assault Rate was zero, then the Murder Rate would be 0.632.

Make a scatterplot that includes the line of best fit.

```r
plot( USArrests$Murder ~ USArrests$Assault,
     main = "Murder Rate vs Assault rate",
     xlab = "Assaults per 100k",
     ylab = "Murders per 100k",
     col = "red",
     pch = 3,
     cex = .4
      )

abline(linearMod.murder.assault)
```



Predict the murder rate of states with 100 assaults per 100K people, 230 assaults per 100K people, and 40 assaults per 100K people

```
0.042 * 100 + 0.632
```

```
## [1] 4.832
```

```
0.042 * 230 + 0.632
```

```
## [1] 10.292
```

```
0.042 * 40 + 0.632
```

```
## [1] 2.312
```

```
# Or even easier ways
0.042 * c(100, 230, 40) + 0.632
```

```
## [1]  4.832 10.292  2.312
```

```
# Or using the linear model
assaults=data.frame(Assault=c(100,230,40))
predict(linearMod.murder.assault, assaults)
```

```
##         1         2         3
##  4.822545 10.270667  2.308028
```

Residuals

So far we have been able to create a linear model and use it to make predictions. One key question is how good is that model? In order to determine that we can look at how wrong our predictions were. If we use the predict function without supplying a data frame of values, predict will output a prediction for each of the values in the original data frame.

```
USArrests$Predictions <- predict(linearMod.murder.assault)

USArrests$Predictions
```

```
##  [1] 10.522119 11.653652 12.952819  8.594322 12.198464  9.181043  5.241632
##  [8] 10.605936 14.671073  9.474403  2.559480  5.660718 11.066931  5.367358
## [15]  2.978566  5.451175  5.199723 11.066931  4.110099 13.204271  6.876068
## [22] 11.318383  3.649104 11.486017  8.091418  5.199723  4.906363 11.192657
## [29]  3.020474  7.295155 12.575642 11.276474 14.754890  2.517571  5.660718
## [36]  6.959886  7.295155  5.073997  7.923784 12.324190  4.235825  8.510505
## [43]  9.055317  5.660718  2.643297  7.169429  6.708434  4.026282  2.852840
## [50]  7.378972
```

Since we named the prediction using the name of the data frame and $ the predictions now appear in the data frame

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape Predictions
## Alabama      13.2     236       58 21.2   10.522119
## Alaska       10.0     263       48 44.5   11.653652
## Arizona       8.1     294       80 31.0   12.952819
## Arkansas      8.8     190       50 19.5    8.594322
## California    9.0     276       91 40.6   12.198464
## Colorado      7.9     204       78 38.7    9.181043
```

If we subtract the prediction from the actual value we can see how far off each of our predictions was.

```
USArrests$How.far.off <- USArrests$Murder - USArrests$Predictions
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape Predictions How.far.off
## Alabama      13.2     236       58 21.2   10.522119    2.677881
## Alaska       10.0     263       48 44.5   11.653652   -1.653652
## Arizona       8.1     294       80 31.0   12.952819   -4.852819
## Arkansas      8.8     190       50 19.5    8.594322    0.205678
## California    9.0     276       91 40.6   12.198464   -3.198464
## Colorado      7.9     204       78 38.7    9.181043   -1.281043
```

The term we use in Statistics to describe these numbers is *residuals*.

If we use the `resid` function, then we can find the residuals quickly.

```r
USArrests$Residuals <- resid(linearMod.murder.assault)

head(USArrests)
```

```
##            Murder Assault UrbanPop Rape Predictions How.far.off Residuals
## Alabama      13.2     236       58 21.2   10.522119    2.677881  2.677881
## Alaska       10.0     263       48 44.5   11.653652   -1.653652 -1.653652
## Arizona       8.1     294       80 31.0   12.952819   -4.852819 -4.852819
## Arkansas      8.8     190       50 19.5    8.594322    0.205678  0.205678
## California    9.0     276       91 40.6   12.198464   -3.198464 -3.198464
## Colorado      7.9     204       78 38.7    9.181043   -1.281043 -1.281043
```

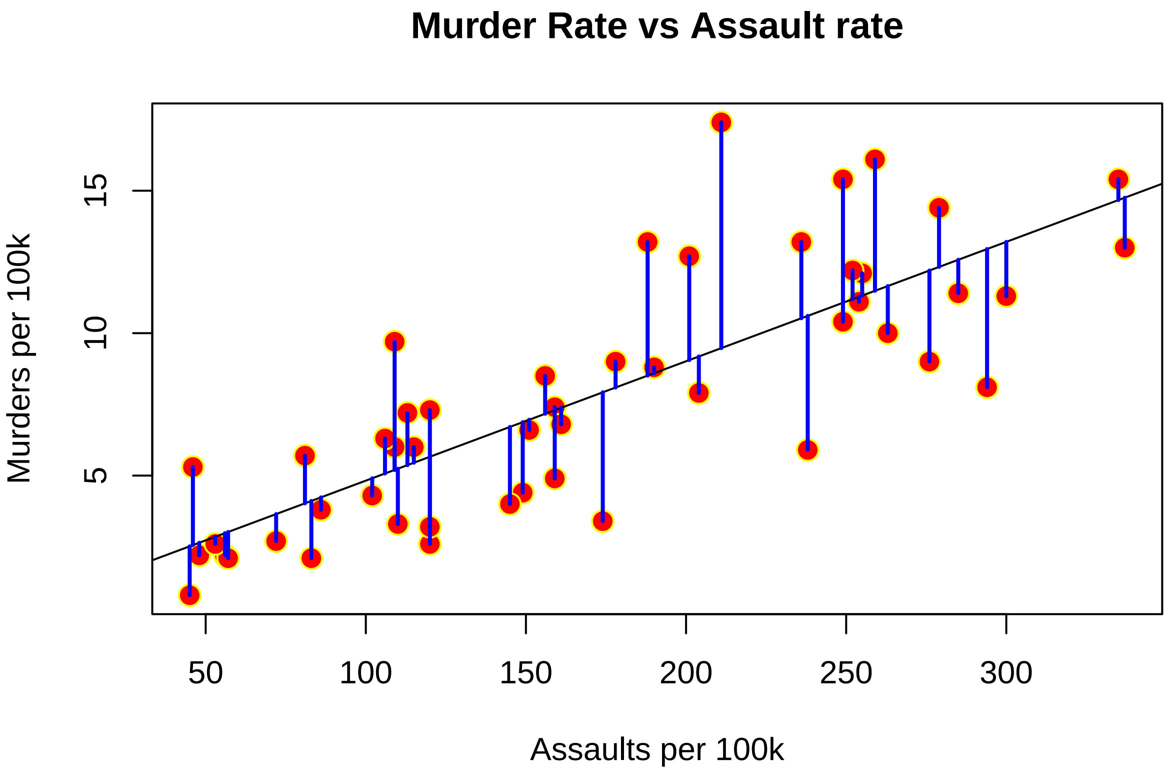
Notice that the Residuals calculated by R using the resid function are exactly the same as those calculated by subtracting the predictions from the actual values.

This is how to create a visual of the residuals. They are the blue lines extending from the data points to the regression line. The length of the line is the value of the residual.

```r
#Scatter plot
plot( USArrests$Murder ~ USArrests$Assault,
     main = "Murder Rate vs Assault rate",
     xlab = "Assaults per 100k",
     ylab = "Murders per 100k",
     col = "Yellow",
     bg = "red",
     pch = 21,
     cex = 1.5
      )

#Regression Line
abline(linearMod.murder.assault)

#Residuals
segments(USArrests$Assault, USArrests$Murder, # (x1, y1 )
         USArrests$Assault, USArrests$Predictions, #(x2, y2)
         col="blue",
         lwd = 2)
```



When the prediction was larger than the actual value the point will be below the line meaning we overestimated in our prediction this results in a negative residual.

When the prediction was smaller than the actual value the point will be above the line meaning we underestimated in our prediction this results in a positive residual.

Ideally if the model is good at predicting the response variable, the residuals should be small.

Residual Plots

Looking at Residual Plots can also help us tell if we should not have been using a linear model after all. If a model tends to overestimate at low values of x and underestimate at high values our model may not be linear. If the model tends to be accurate at low values of x but poor at high values of x it may not be appropriate.

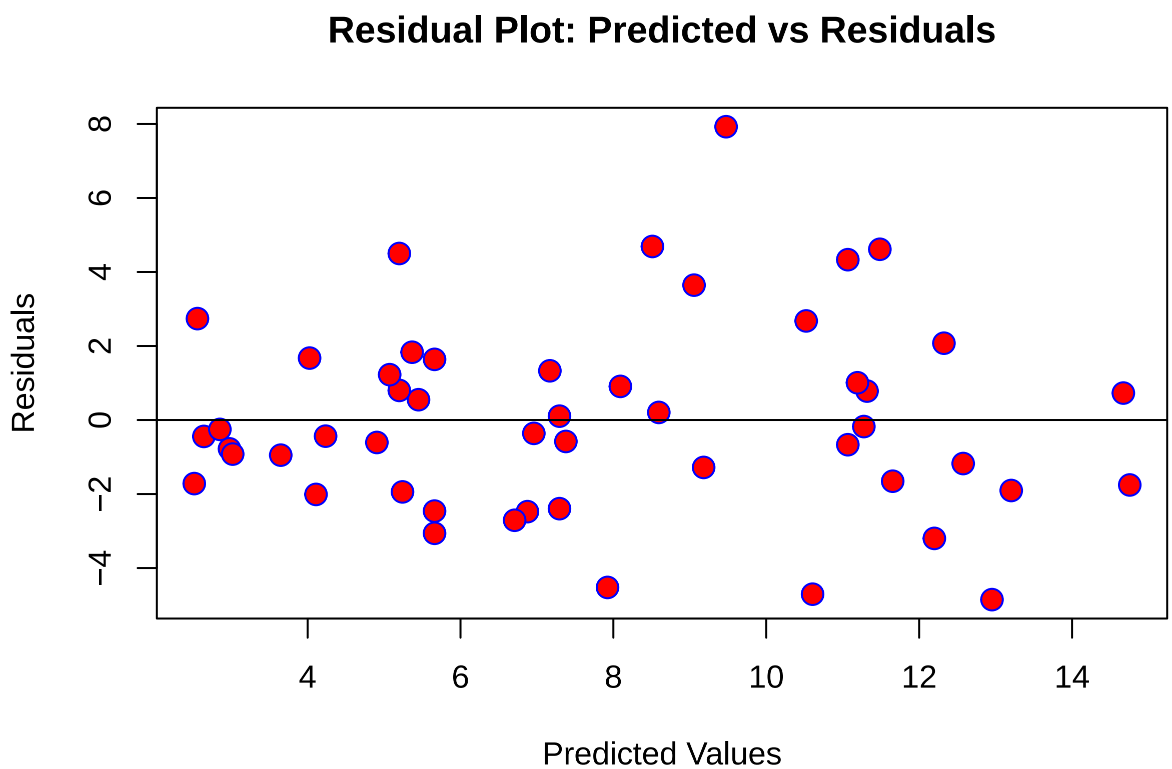
To check a model, we can make a scatterplot with the predicted values as the x variable and the residuals as the y variable. Ideally the shape will be cloudlike showing no patterns at all.

It is helpful to draw a line at $y = 0$ so that we can see which residuals are positive and which are negative.

```
#Residual Plot

plot( USArrests$Residuals ~ USArrests$Predictions,
     main = "Residual Plot: Predicted vs Residuals",
     xlab = "Predicted Values",
     ylab = "Residuals",
     col = "blue",
     bg = "red",
     pch = 21,
     cex = 1.5
      )
```

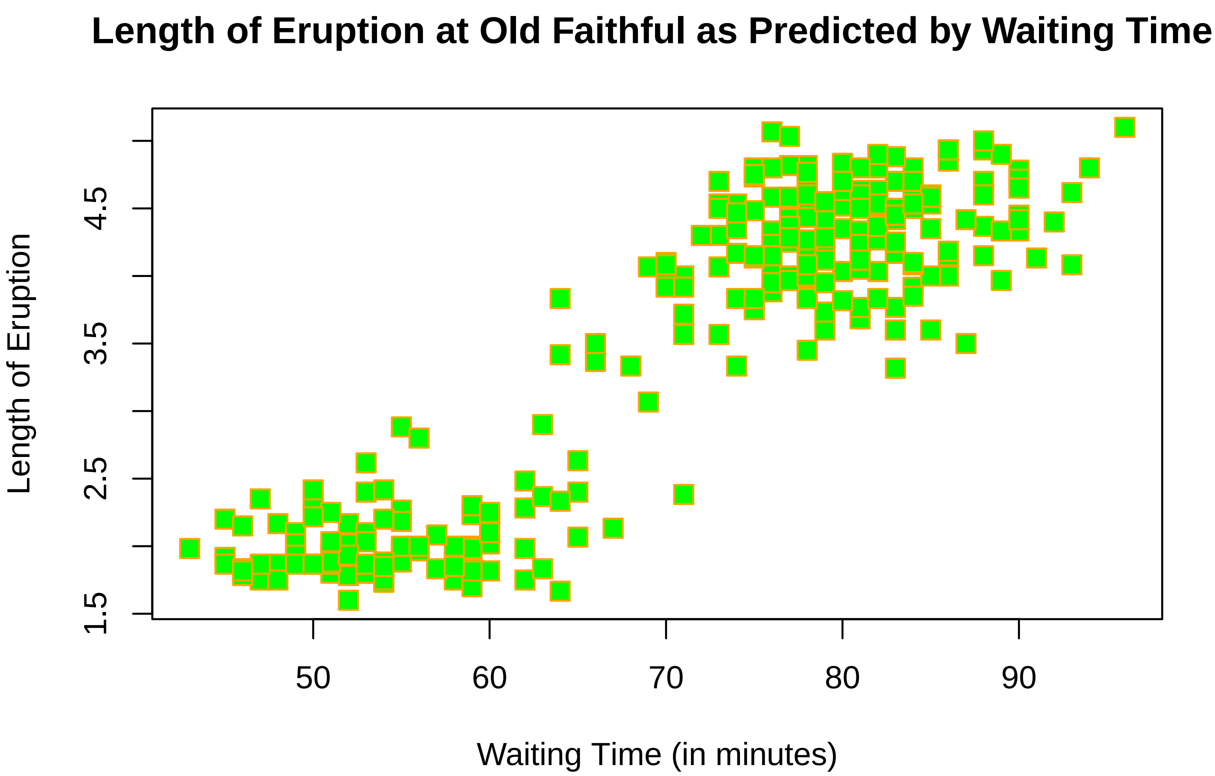
```r
#Line at y = 0
abline(0, 0)
```



That worked out very well! This is strong evidence that using a linear model was appropriate.

Here is an example of a graph that originally looks like a linear model might be appropriate

```r
data(faithful)
plot(faithful$eruptions ~ faithful$waiting,
     main = "Length of Eruption at Old Faithful as Predicted by Waiting Time ",
     xlab = "Waiting Time (in minutes)",
     ylab = "Length of Eruption",
     col =  "orange",
     bg = "green",
     pch = 22,
     cex = 1.5
      )
```



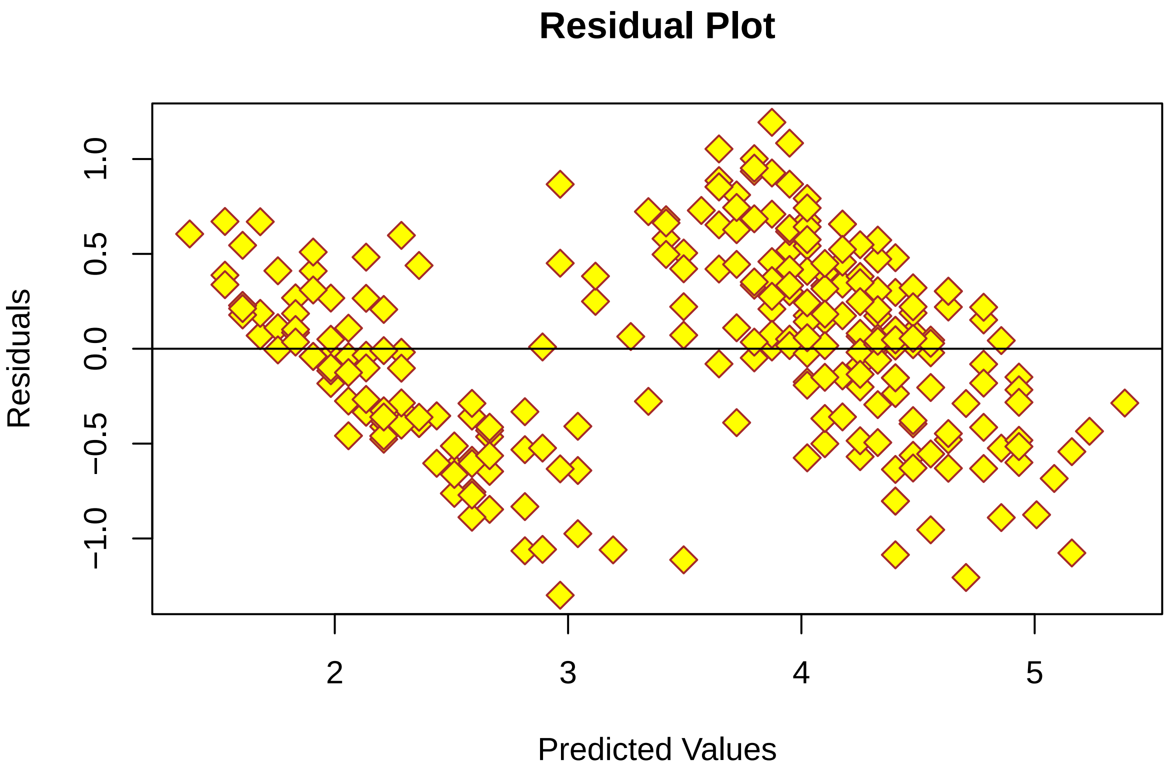
But when we check the residuals there is a strange pattern.

```
linearMod.eruptions.waiting <- lm(eruptions ~ waiting, data = faithful)

faithful$Predictions <- predict(linearMod.eruptions.waiting)
faithful$Residuals <- resid(linearMod.eruptions.waiting)


plot(faithful$Residuals ~ faithful$Predictions,
     main = "Residual Plot",
     xlab = "Predicted Values",
     ylab = "Residuals",
     col = "brown",
     bg = "yellow",
     pch = 23,
     cex = 1.5
      )

abline(0,0)
```



Since there is a clear pattern in the Residual Plot, a linear model in not appropriate for this data. In order to analyze it we need a different method.

The Summary Function

We can learn more about a regression model by using the summary function:

```
print(linearMod.murder.assault)
```

```
##
## Call:
## lm(formula = Murder ~ Assault, data = USArrests)
##
## Coefficients:
## (Intercept)      Assault
##     0.63168      0.04191
```

```
summary(linearMod.murder.assault)
```

```
##
## Call:
## lm(formula = Murder ~ Assault, data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8528 -1.7456 -0.3979  1.3044  7.9256
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.631683   0.854776   0.739    0.464
## Assault     0.041909   0.004507   9.298  2.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.629 on 48 degrees of freedom
## Multiple R-squared:  0.643,  Adjusted R-squared:  0.6356
## F-statistic: 86.45 on 1 and 48 DF,  p-value: 2.596e-12
```

There is a lot of information that is printed out. The most important values for our analysis are:

- The Estimates which are the same values that you found earlier using the print() function
- The numbers under the column Pr(> t). These are the p-values of each variable. They essentially tell us the likelihood that the association we found could have been because of random variation.
- The stars next to the p-values. These tell us if and at what level the variables are significant. The key to understanding them is in the line that starts with Signif. codes.

From this summary we can see that:

- The slope of the regression line is .0419 and the y-intercept is .6316.
- The extremely low p-value by assault tells me that it is highly unlikely that it is just random variation that is masquerading as association.
- The three stars tell us that the variable is significant at the .001 level.

R-squared

Another key value we can see in the summary is R-squared (labeled Multiple R-squared in the summary table). R-squared is the percentage of the variation in the response variable is explained by the model. This value will always be between 0 and 1. The closer this value is to 1 the better the model is.

Here is an example to see how to understand R-squared:

Say we did not do a regression and therefore had no idea what the relationship between assault and murder was. In order to guess the murder rate for a randomly selected state what would our best guess be?

Since you know nothing else your guess should be the average murder rate. So how would we do if we just guessed the average murder rate for each state?

```r
avg.murder.rate=mean(USArrests$Murder)
avg.murder.rate
```

```
## [1] 7.788
```

Now we can compare the residuals to figure out how much better our regression model is than just guessing the mean. We can't just add them up to compare because due to some being positive and some being negative they will both add to 0.

```r
USArrests$Avg.Error=USArrests$Murder-avg.murder.rate
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape Predictions How.far.off Residuals
## Alabama      13.2     236       58 21.2   10.522119    2.677881  2.677881
## Alaska       10.0     263       48 44.5   11.653652   -1.653652 -1.653652
## Arizona       8.1     294       80 31.0   12.952819   -4.852819 -4.852819
## Arkansas      8.8     190       50 19.5    8.594322    0.205678  0.205678
## California    9.0     276       91 40.6   12.198464   -3.198464 -3.198464
## Colorado      7.9     204       78 38.7    9.181043   -1.281043 -1.281043
##            Avg.Error
## Alabama        5.412
## Alaska         2.212
## Arizona        0.312
## Arkansas       1.012
## California     1.212
## Colorado       0.112
```

```
sum(USArrests$Avg.Error)
```

```
## [1] -1.953993e-14
```

Note the only reason they do not add to exactly zero is due to rounding.

So we do what we have done several times already this year and square them before adding them up together.

```
sum(USArrests$Avg.Error^2)
```

```
## [1] 929.5528
```

Our regression model has a much smaller sum of squared residuals than the just guessing model. It only has

```
sum(USArrests$Residuals^2)
```

```
## [1] 331.8496
```

```
(sum(USArrests$Avg.Error^2)-sum(USArrests$Residuals^2))/sum(USArrests$Avg.Error^2)
```

```
## [1] 0.6430008
```

That means that our model explains

This means that the Assault Rate explain 64.3% of the variation in the Murder Rate

Running the summary function again we see that R-squared is

```
summary(linearMod.murder.assault)
```

```
##
## Call:
## lm(formula = Murder ~ Assault, data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.8528 -1.7456 -0.3979  1.3044  7.9256
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.631683   0.854776   0.739    0.464
## Assault     0.041909   0.004507   9.298  2.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.629 on 48 degrees of freedom
## Multiple R-squared:  0.643,  Adjusted R-squared:  0.6356
## F-statistic: 86.45 on 1 and 48 DF,  p-value: 2.596e-12
```

Your Turn

Now we will look at a data set that contains the monthly totals of car drivers in Great Britain killed or seriously injured Jan 1969 to Dec 1984. We will be looking to see if the price of gasoline (PetrolPrice) has an effect on the number of drivers killed (DriversKilled).

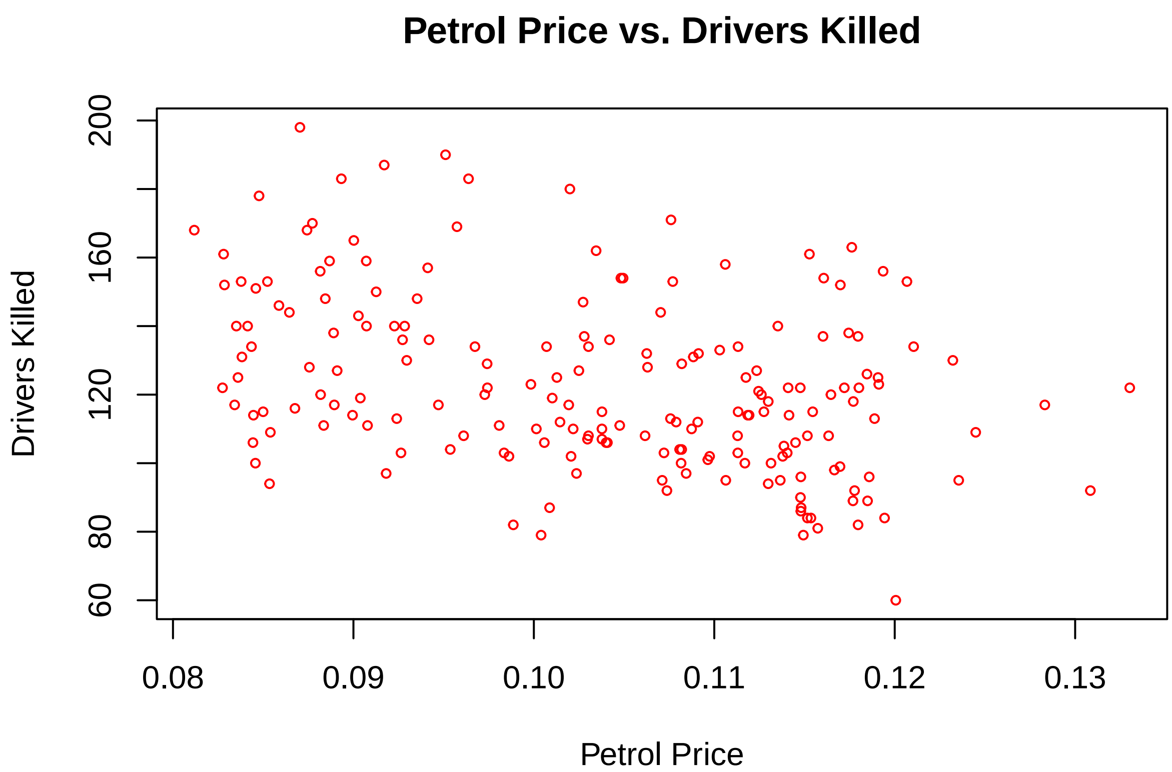
Load up the data set

```
df.seatbelts <- as.data.frame(Seatbelts)
head(df.seatbelts)
```

```
##   DriversKilled drivers front rear   kms PetrolPrice VanKilled law
## 1           107    1687   867  269  9059   0.1029718        12   0
## 2            97    1508   825  265  7685   0.1023630         6   0
## 3           102    1507   806  319  9963   0.1020625        12   0
## 4            87    1385   814  407 10955   0.1008733         8   0
## 5           119    1632   991  454 11823   0.1010197        10   0
## 6           106    1511   945  427 12391   0.1005812        13   0
```

Create a scatter plot of gas prices vs drivers killed. Make sure to put the explanatory and response variables on the correct axis.

```
plot(df.seatbelts$DriversKilled~df.seatbelts$PetrolPrice,
     main="Petrol Price vs. Drivers Killed",
     col="red",
     pch=1,
     xlab="Petrol Price",
     ylab="Drivers Killed",
     cex=.6)
```



Is it appropriate to use a linear model to describe the association between the variables? Why or why not?

It looks like the relationship is linear, so yes it is appropriate.

Create a linear model for the two variables.

```
linear.mod.driving=lm(DriversKilled~PetrolPrice, data=df.seatbelts)
summary(linear.mod.driving)
```

```
##
## Call:
## lm(formula = DriversKilled ~ PetrolPrice, data = df.seatbelts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.558 -16.921  -3.594  13.638  61.830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   206.31      14.55  14.179  < 2e-16 ***
## PetrolPrice  -805.86     139.46  -5.778 3.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.47 on 190 degrees of freedom
## Multiple R-squared:  0.1495, Adjusted R-squared:  0.145
## F-statistic: 33.39 on 1 and 190 DF,  p-value: 3.044e-08
```

Find the predicted values and the residuals and add them to the data frame.

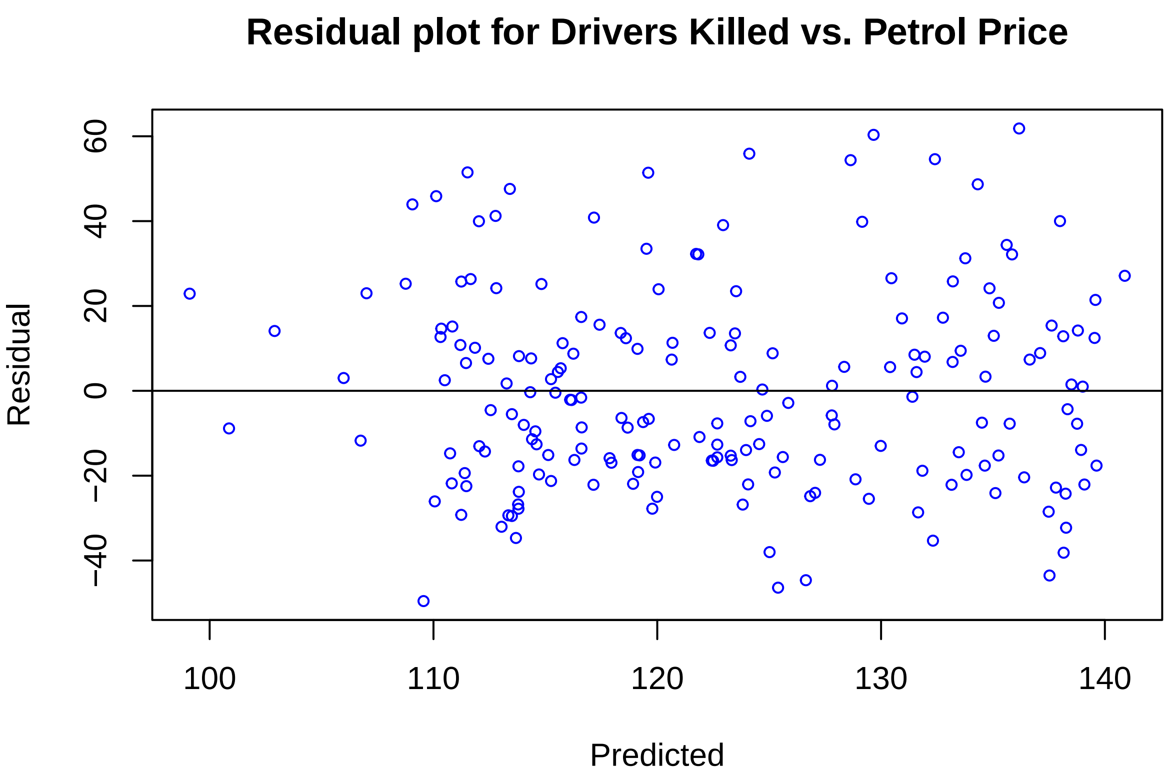
```r
df.seatbelts$predicted=predict(linear.mod.driving)
head(df.seatbelts)
```

```
##   DriversKilled drivers front rear   kms PetrolPrice VanKilled law predicted
## 1           107    1687   867  269  9059   0.1029718        12   0  123.3277
## 2            97    1508   825  265  7685   0.1023630         6   0  123.8183
## 3           102    1507   806  319  9963   0.1020625        12   0  124.0604
## 4            87    1385   814  407 10955   0.1008733         8   0  125.0188
## 5           119    1632   991  454 11823   0.1010197        10   0  124.9008
## 6           106    1511   945  427 12391   0.1005812        13   0  125.2542
```

Make a graph of the predicted values and the residuals.

```r
df.seatbelts$resid=resid(linear.mod.driving)
plot(df.seatbelts$resid~df.seatbelts$predicted,
     main="Residual plot for Drivers Killed vs. Petrol Price",
     xlab="Predicted",
     ylab="Residual",
     col="blue",
     pch=21,
     cex=.7)

abline(0,0)
```



Do you still think a linear model is appropriate? Why?

The residuals are seemingly randomly placed on the residual plot, so it looks good.

Find the summary of the linear model.

```r
summary(linear.mod.driving)
```

```
##
## Call:
## lm(formula = DriversKilled ~ PetrolPrice, data = df.seatbelts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.558 -16.921  -3.594  13.638  61.830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   206.31      14.55  14.179  < 2e-16 ***
## PetrolPrice  -805.86     139.46  -5.778 3.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.47 on 190 degrees of freedom
## Multiple R-squared:  0.1495, Adjusted R-squared:  0.145
## F-statistic: 33.39 on 1 and 190 DF,  p-value: 3.044e-08
```

Write the equation of the model:

$\widehat{DriversKilled} = -805.86 PetrolPrice + 206.31$

Could the association between gas price and driver deaths be due to random variation? Why or why not?

It could not be due to random variation because the summary indicates that the slope is tremendously unlikely to be zero.

How good is the model at explaining the variation in driver deaths?

Not very good. The Petrol Price only explains about 15% of the variation in the number of Drivers Killed. Therefore there are probably other variables that are more important.