Weekly Assignment 5

Sean Leggett - BDA201 - Winter 2020
March 8, 2020

1. Is the data correlated? What does the correlation mean? What is the pearson coefficient of correlation? Calculate by implementing the formula in r and then use the r function "cor" to verify.
2. What is the p-value for the pearson correlation coefficient? What is meant by calculated p-value?
3. Calculate the regression line for this data, what is the slope and y-intercept?
4. Show the data plotted along with the regression line in brown color.
5. What is residual for the observation (4, 7), which observation has the largest residual?
6. What percentage of variation is explained by the regression line? What percent of variation is due to random and unexplained factors?

```
## generate dataframe
muggings <- data.frame("Officers" = c(10, 15, 16, 1, 4, 6, 18, 12, 14, 7),
                       "Muggings" = c(5, 2, 1, 9, 7, 8, 1, 5, 3, 6)
)
muggings
```

```
##    Officers Muggings
## 1        10        5
## 2        15        2
## 3        16        1
## 4         1        9
## 5         4        7
## 6         6        8
## 7        18        1
## 8        12        5
## 9        14        3
## 10        7        6
```

Question 1 - Correlated?
Let's look at some values and plots…

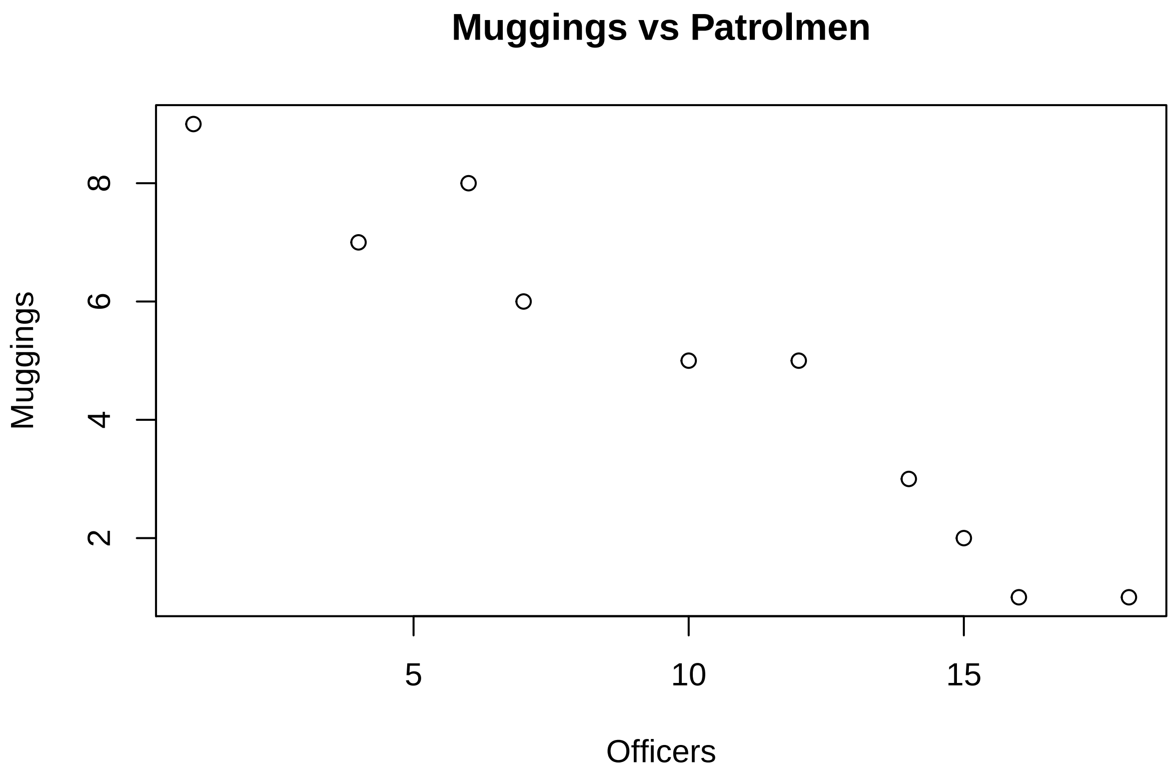
```
## manual calculation of Pearson coefficient of correlation
x <- muggings$Muggings
y <- muggings$Officers
X <- x - mean(x)
Y <- y - mean(y)
r <- sum(X * Y) / sqrt(sum(X*X) * sum(Y*Y))
r
```

```
## [1] -0.9690786
## using cor() function
cor(muggings$Muggings, muggings$Officers)
```

```
## [1] -0.9690786
```

A scatter plot:

```
mugplot <- plot(muggings$Officers, muggings$Muggings,
                main = "Muggings vs Patrolmen",
                xlab = "Officers",
                ylab = "Muggings")
```



```
mugplot
```

```
## NULL
```

There would appear to be a strong negative correlation. More officers = less muggings.

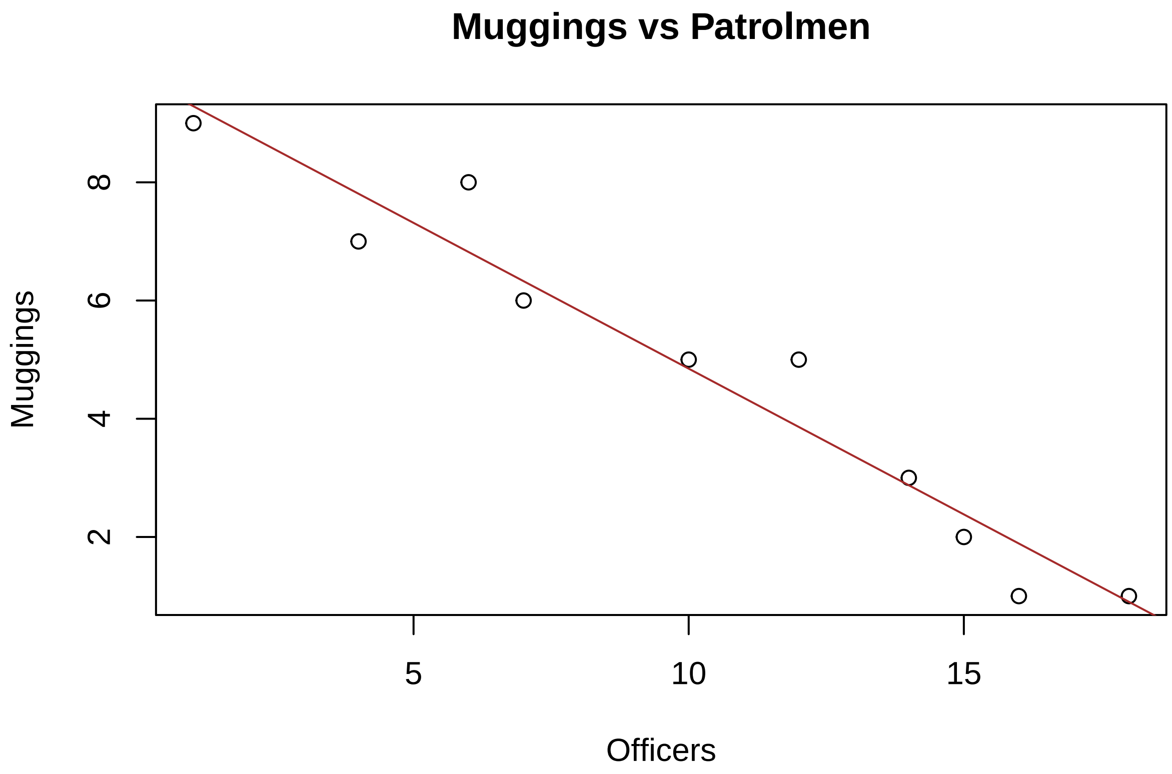
Question 2 - P-Value

We opted for the cor.test() function to yield a variety of critical points including P-Value of 0.000003853 which is statistically significant. The P-Value indicates the probability that future randomly chosen values for these variables will demonstrate the same close relationship we have displayed thus far.

```
cor.test(muggings$Muggings, muggings$Officers, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  muggings$Muggings and muggings$Officers
## t = -11.108, df = 8, p-value = 3.853e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9928873 -0.8707417
## sample estimates:
##        cor
## -0.9690786
```

```
mugmodel <- lm(muggings$Officers ~ muggings$Muggings)
regres1 <- plot(muggings$Officers, muggings$Muggings,
                main = "Muggings vs Patrolmen",
                xlab = "Officers",
                ylab = "Muggings")
abline(lm(muggings$Muggings ~ muggings$Officers), col = "brown")
abline(h = mean(muggings$Officers))
```



```
regres1
```

```
## NULL
```

slope is 9.7798 and Y-Intercept is -.4932

Question 5 - Residuals
We can calculate residulas using R. Code and output below shows the residuals for all 10 data points. The point (4,7) is the 5th entry, or residual -1.9203779. The pair in the 8th position (12,5) demonstrates the highest residual at 2.2712551

```
mugres <- resid(mugmodel)
mugres
```

```
##          1          2          3          4          5          6 
##  0.2712551 -0.4412955 -1.3454791 -1.1120108 -1.9203779  1.9838057 
##          7          8          9         10 
##  0.6545209  2.2712551  0.4628880 -0.8245614
```

Question 6 - RSquared
The RSquared value can be obtained or calculated in a number of ways. The summary() function includes this value and it can be accessed directly using $ if preferred. RSquared can also be easily calculated as R*R given that we know R from Question 1 (0.9690786)

```
r2 <- .9690786^2
r2
```

```
## [1] 0.9391133
```

```
summary(mugmodel)$r.squared
```

```
## [1] 0.9391132
```

```
summary(mugmodel)
```

```
## 
## Call:
## lm(formula = muggings$Officers ~ muggings$Muggings)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.92038 -1.04015 -0.08502  0.60661  2.27126 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        19.2497     0.9311   20.68 3.14e-08 ***
## muggings$Muggings  -1.9042     0.1714  -11.11 3.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.476 on 8 degrees of freedom
## Multiple R-squared:  0.9391, Adjusted R-squared:  0.9315 
## F-statistic: 123.4 on 1 and 8 DF,  p-value: 3.853e-06
```

This means that 94% of variation is explained by the regression line and 6% remains statistically unexplained by the model.