Homework 2

Seth Marceno

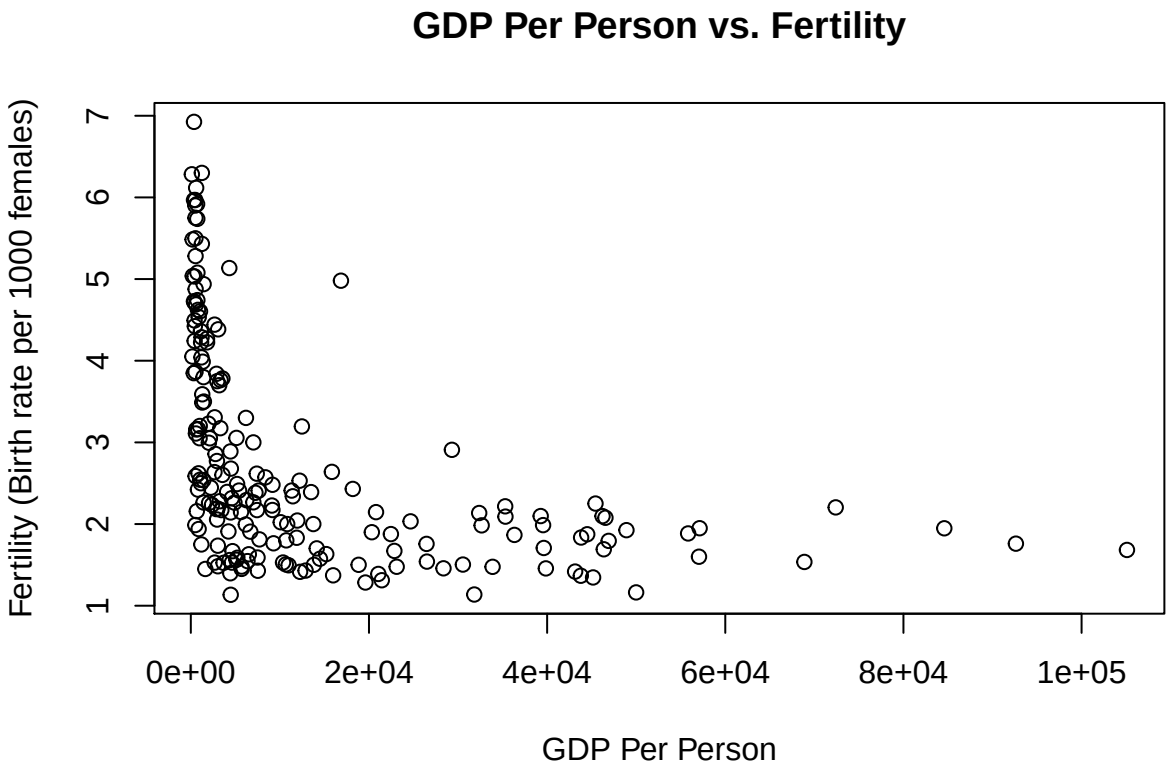
April 18, 2019

Question 1

Part (a)

The predictor is GDP per person and the response is Fertility. ##Part (b) Looking at the scatterplot made below, the trend of GDP per person vs Fertility looks like it is NOT linear.

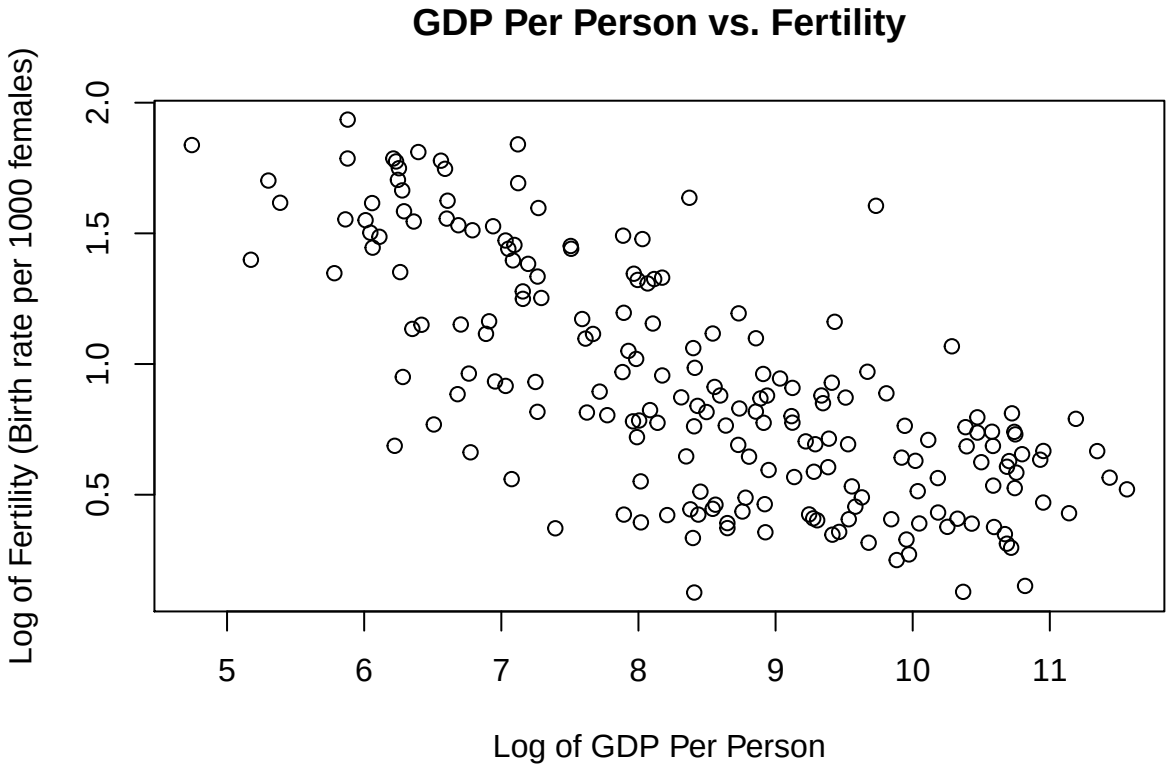
```
#Reading in the dataset UN11 into my Global Envirnment
data(UN11)
#Assigning values to my predictor and response variables, respectively
fertility <- UN11$fertility
ppgdp <- UN11$ppgdp
#Creating and titling my scatterplot
plot( ppgdp, fertility, ylab = 'Fertility (Birth rate per 1000 females)',
      xlab = 'GDP Per Person')
title('GDP Per Person vs. Fertility')
```



Part (c)

Taking the natural log of both variables gives us a scatterplot where a simple linear regression model seems plausible.

```r
plot(log(ppgdp), log(fertility), ylab = 'Log of Fertility (Birth rate per 1000 females)',
     xlab = 'Log of GDP Per Person')
title('GDP Per Person vs. Fertility')
```

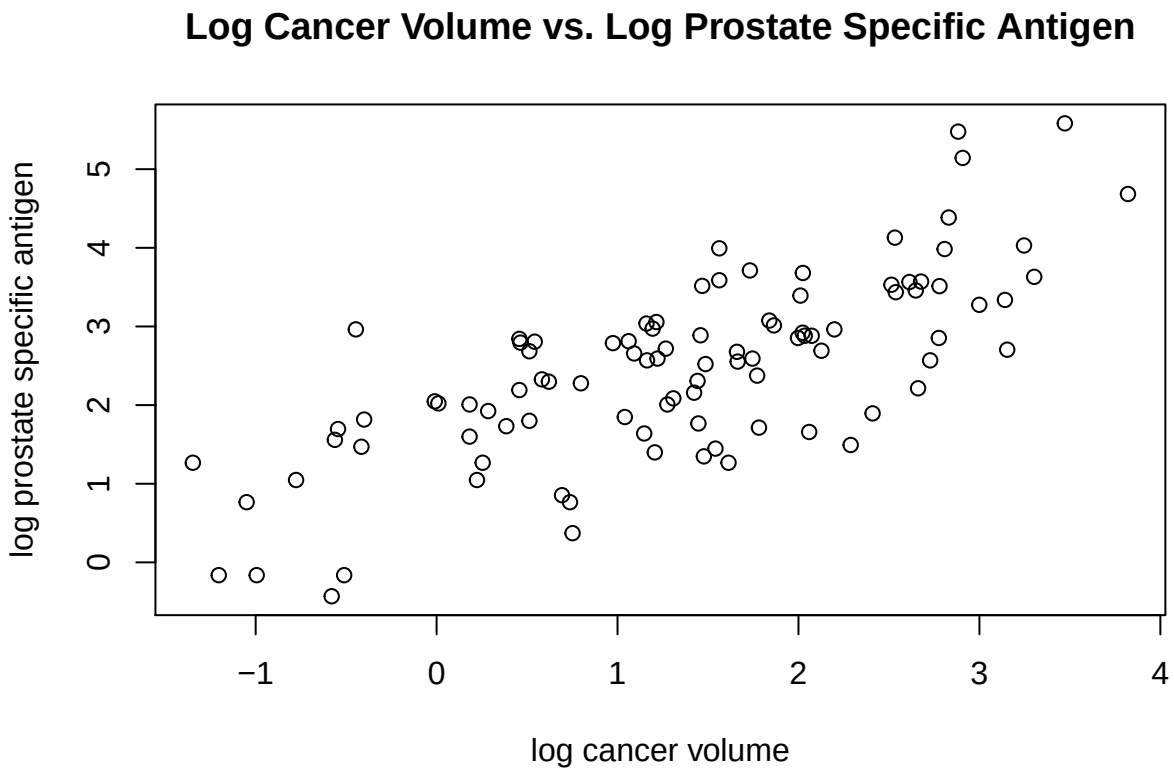


Question 2

Part (a)

Based off our resutls of the scatter plot below, a simple linear regression model seems reasonable.

```r
#Reads in the dataset prostate into my global environment
data(prostate)
#Assigning values to my predictor and response variables, respectively
lpsa <- prostate$lpsa
lcavol <- prostate$lcavol
#Creates and titles my scatterplot
plot(lcavol, lpsa, xlab = 'log cancer volume', ylab = 'log prostate specific antigen')
title('Log Cancer Volume vs. Log Prostate Specific Antigen')
```



Part (b)

Here I compute the values of Xbar, Ybar, Sxx, Syy, Sxy, B0hat, and B1hat without using the built in R function lm(). Then I draw the fitted line on my plot from part (a)

```
Xbar <- 1/length(lcavol) * sum(lcavol)
Xbar
```

```
## [1] 1.35001
```

```
Ybar <- 1/length(lpsa) * sum(lpsa)
Ybar
```

```
## [1] 2.478387
```

```
Sxx <- sum(lcavol^2)-length(lcavol)*Xbar^2
Sxx
```

```
## [1] 133.359
```

```
Syy <- sum(lpsa^2)-length(lpsa)*Ybar^2
Syy
```

```
## [1] 127.9176
```

```
Sxy <- sum(lcavol*lpsa) - length(lpsa)*Xbar*Ybar
Sxy
```

```
## [1] 95.92784
```

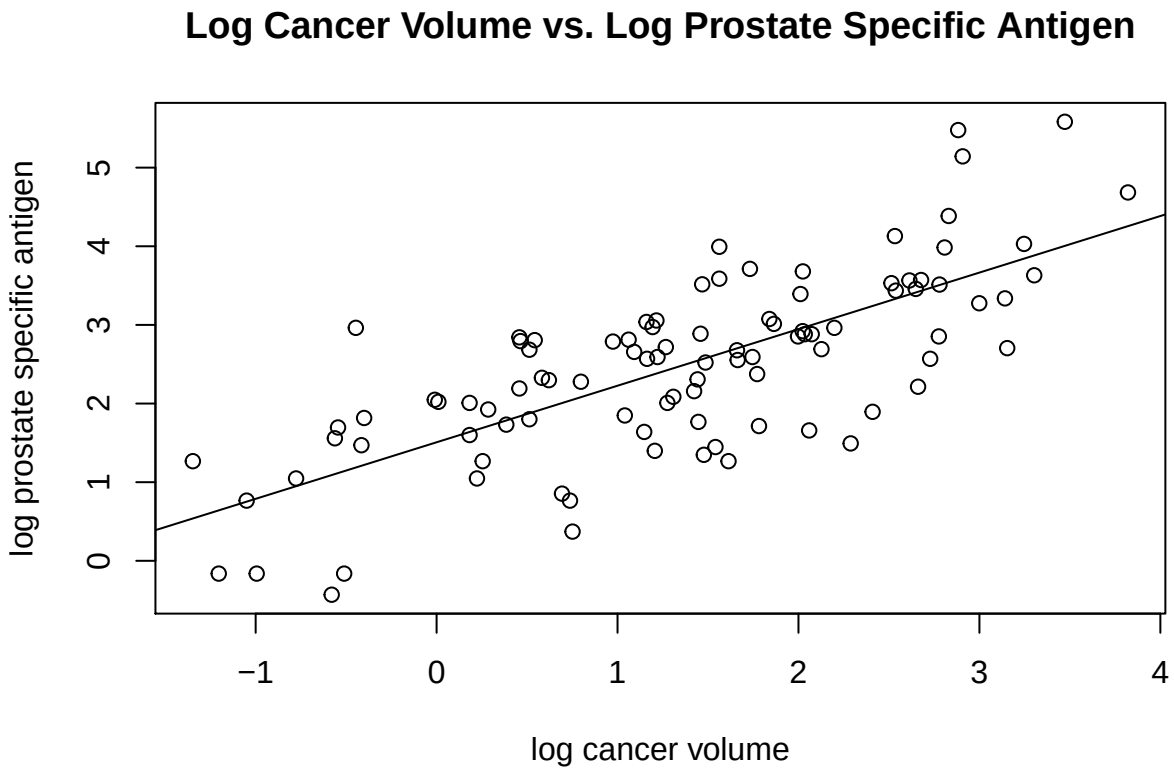
```
B1hat <- Sxy/Sxx
B1hat
```

```
## [1] 0.7193201
```

```
B0hat <- Ybar - B1hat*Xbar
B0hat
```

```
## [1] 1.507298
```

```
plot(lcavol, lpsa, xlab = 'log cancer volume', ylab = 'log prostate specific antigen')
title('Log Cancer Volume vs. Log Prostate Specific Antigen')
abline(a = B0hat, b = B1hat)
```



Part (c)

Here I estimate Sigma Squared Hat

```
Yhat <- B0hat + B1hat*lcavol
SigmaHatSquared <- 1/(length(lpsa)-2) * sum((lpsa - Yhat)^2)
SigmaHatSquared
```

```
## [1] 0.6201553
```

Standard Error of B0hat

```
SE_B0hat  <- sqrt(SigmaHatSquared) * sqrt((1/length(lpsa)) + (Xbar^2/Sxx))
SE_B0hat
```

```
## [1] 0.1219368
```

Standard Error of B1hat

```
SE_B1hat <- sqrt(SigmaHatSquared)/sqrt(Sxx)
SE_B1hat
```

```
## [1] 0.06819288
```

Finding covariance of B1hat and B0hat

```
covB1hat_B0hat <- (-Xbar*SigmaHatSquared)/Sxx
covB1hat_B0hat
```

```
## [1] -0.006277907
```

T-test for B0hat: H_0: B0 = 0 vs. H_1: B0 != 0 with alpha = 0.05

```
test_stat_B0 <- B0hat/SE_B0hat
crit_val_B0 <- qt(p = .975, df = length(lpsa)-2)
P_value_B0 <- 2*pt(abs(test_stat_B0), df = length(lpsa)-2, lower.tail = FALSE)

P_value_B0
```

```
## [1] 1.722234e-21
```

```
test_stat_B0
```

```
## [1] 12.3613
```

```
crit_val_B0
```

```
## [1] 1.985251
```

Since |test_stat_B0| > crit_val_b0 we reject our null hypothesis that B0=0. Similarly since P_value_b0 < alpha = 0.05 we also conclude that we reject the null hypothesis.

T-test for B1hat: H_0: B1 = 0 vs. H_1: B1 != 0 with alpha = 0.05

```
test_stat_B1 <- B1hat/SE_B1hat
crit_val_B1 <- qt(p = .975, df = length(lpsa)-2)
P_value_B1 <- 2*pt(abs(test_stat_B1), df = length(lpsa)-2, lower.tail = FALSE)

P_value_B1
```

```
## [1] 1.118616e-17
```

```
test_stat_B1
```

```
## [1] 10.54832
```

```
crit_val_B1
```

```
## [1] 1.985251
```

Since |test_stat_B1| > crit_val_b1 we reject our null hypthesis that B1=0. Similarly since P_value_b1 < alpha = 0.05 we also conlcude that we reject the null hypothesis. Thus we can say that a simple linear regression model is reasonable.

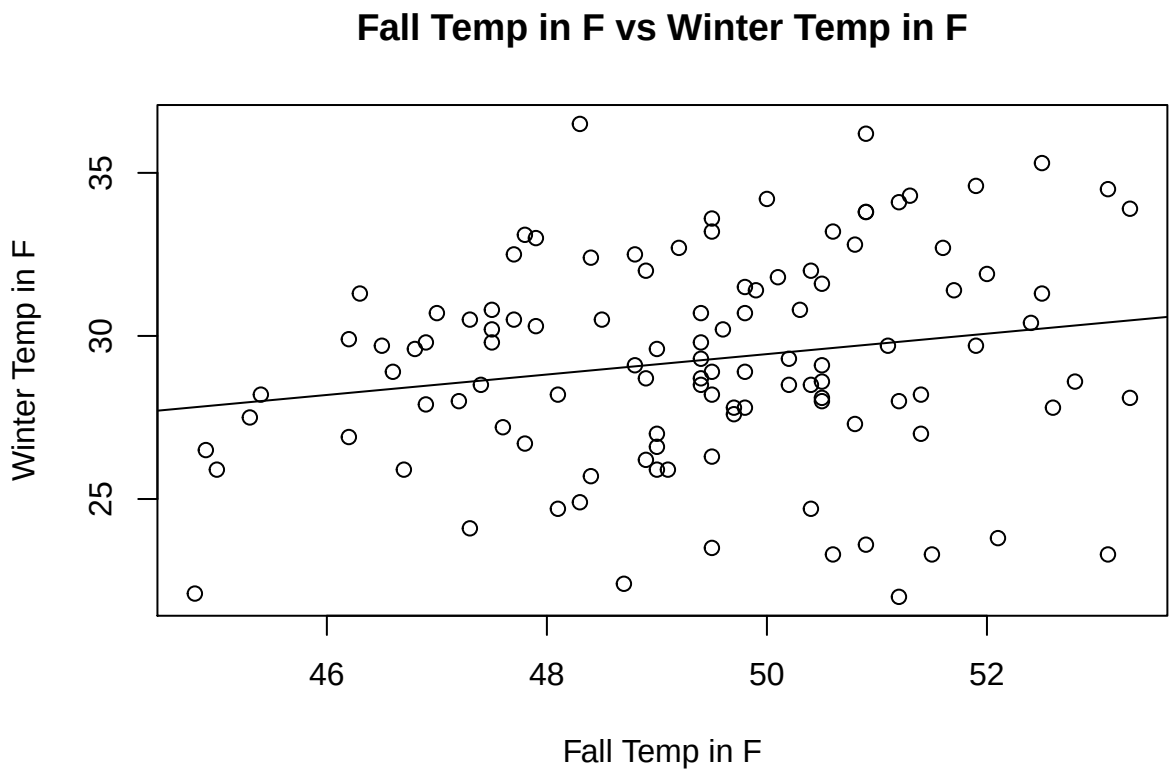
Question 3

Part (a)

Reads in data, creates a linear regression model, and then plots that data.

```
data(ftcollinstemp)
fall <- ftcollinstemp$fall
winter <- ftcollinstemp$winter
```

```r
fit1 <- lm(winter~fall)
plot(fall, winter, xlab = 'Fall Temp in F', ylab = 'Winter Temp in F')
abline(fit1)
title('Fall Temp in F vs Winter Temp in F')
```



Part (b)

Testing H_0 B1 = 0 vs. H_1: B1 != 0 with alpha = 0.01

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = winter ~ fall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8186 -1.7837 -0.0873  2.1300  7.5896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7843     7.5549   1.825   0.0708 .
## fall          0.3132     0.1528   2.049   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 109 degrees of freedom
## Multiple R-squared:  0.0371, Adjusted R-squared:  0.02826
```

```
## F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

Based off of our summary function, we can see that our test statistic for B1 is .3132 and our t value is 2.049. Thus, since |.3132| !> 2.049 we fail to reject the null hypothesis.

Part (c)

We know that the percentage of the variability in winter as explained by fall is equal to R squared. Based off our summary from above we see that R squared is equal to 0.0371. Therefore we can conclude that 3.71% of the variablilty in winter is explained by fall.