IST 687 – Homework 8 Making Predictions

Dan Burke

9/18/2021

Assignment Due: 8/23/2021

Submitted: 9/19/2021

1. Read in data from the following URL:

http://college.cengage.com/mathematics/brase/understandable_statistics/7e/stud ents/datasets/mlr/excel/mlr01.xls

If you view this in a spreadsheet, you will find that four columns of a small dataset.
The first column shows the number of fawn in a given spring (fawn are baby
Antelope).

The second column shows the population of adult antelope, the third
shows the annual precipitation that year, and finally, the last column shows how bad the winter was during that year.

2. You have the option of saving the file save this file to your computer and read it into R, or reading the data directly from the web into a data frame.

I decided to download and read from the same directory as this file is located.

```
#install.packages("readxl")
library("readxl")
```

```
## Warning: package 'readxl' was built under R version 4.1.1
```

```
library(ggplot2)

rawdata <- read_excel("C:\\Users\\danbu\\Desktop\\IntroToDataScience\\mlr01.xls")

antelopeData <- rawdata
colnames(antelopeData) <- c("springFawn", "adultPop", "annualPrecip", "winterbadness")
```

3. You should inspect the data using the str() command to make sure that all of the cases have been read in (n=8 years of observations) and that there are four variables.

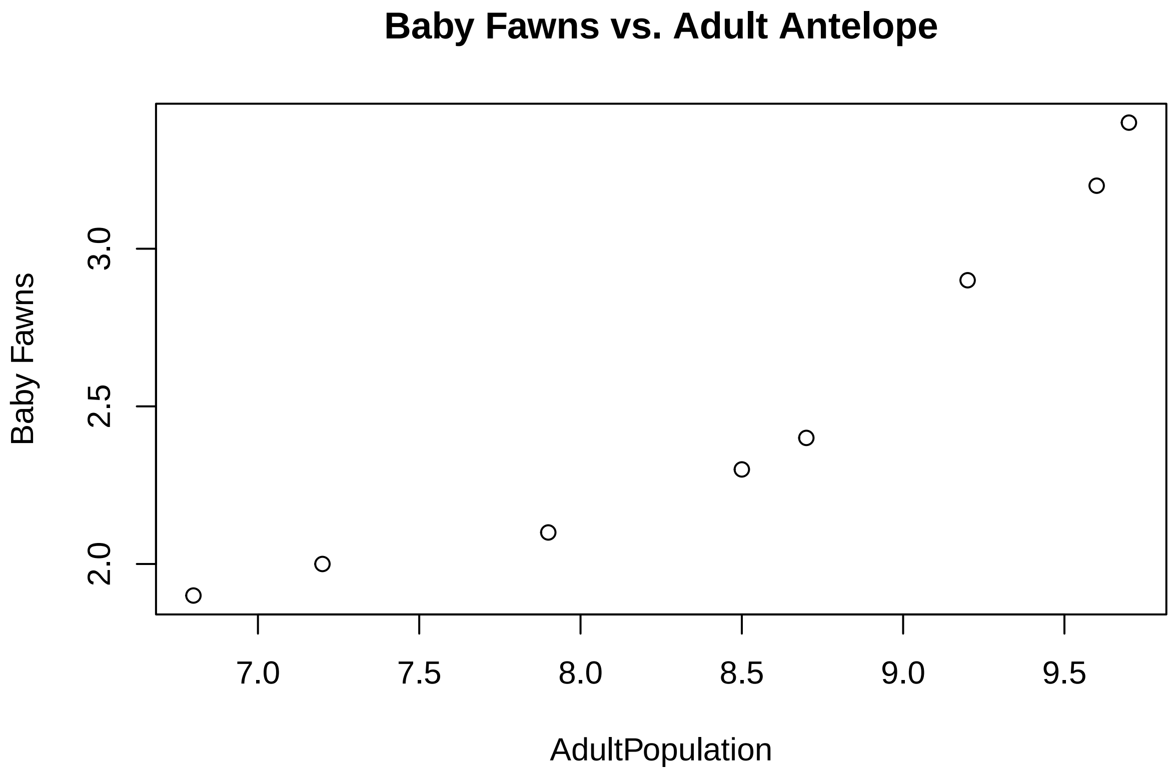
```
str(antelopeData)
```

```
## tibble [8 x 4] (S3: tbl_df/tbl/data.frame)
##  $ springFawn   : num [1:8] 2.9 2.4 2 2.3 3.2 ...
##  $ adultPop     : num [1:8] 9.2 8.7 7.2 8.5 9.6 ...
##  $ annualPrecip : num [1:8] 13.2 11.5 10.8 12.3 12.6 ...
##  $ winterbadness: num [1:8] 2 3 4 2 3 5 1 3
```

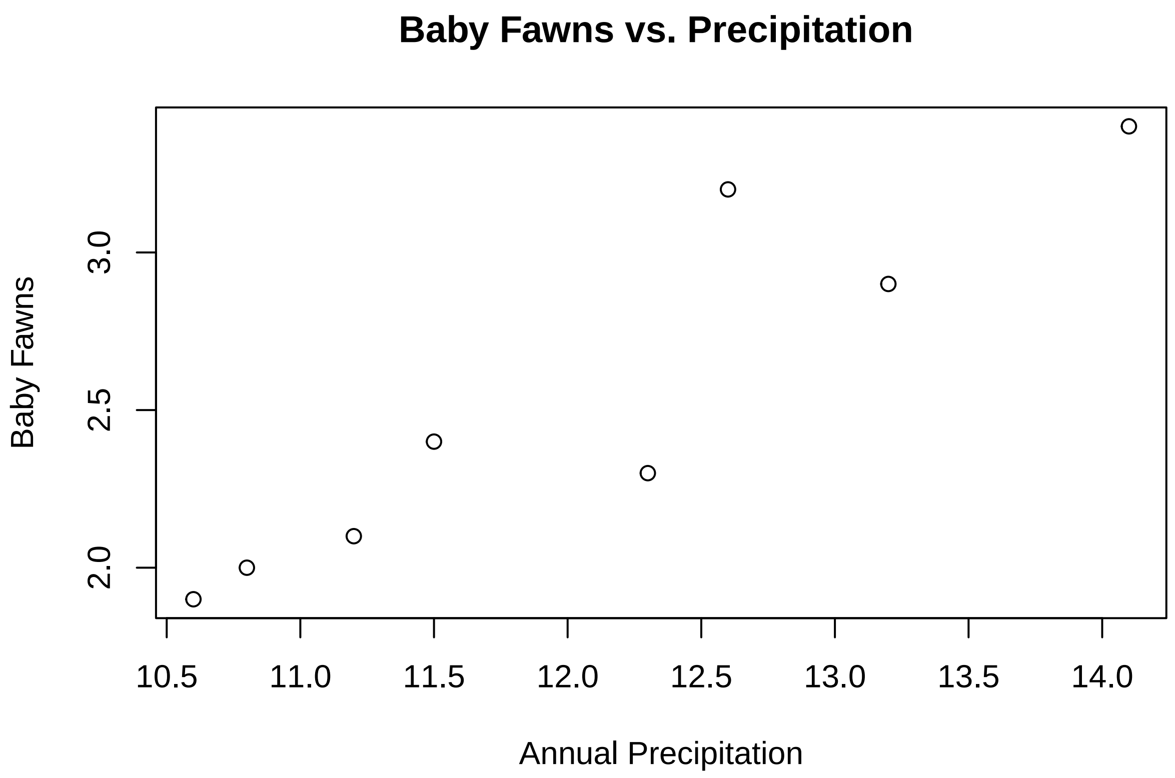
4. Create bivariate plots of number of baby fawns versus adult antelope population, the precipitation that year, and the severity of the winter.

Your code should produce three separate plots. Make sure the Y-axis and X-axis are labeled. Keeping in mind that the number of fawns is the outcome (or dependent) variable, which axis should it go on in your plots?

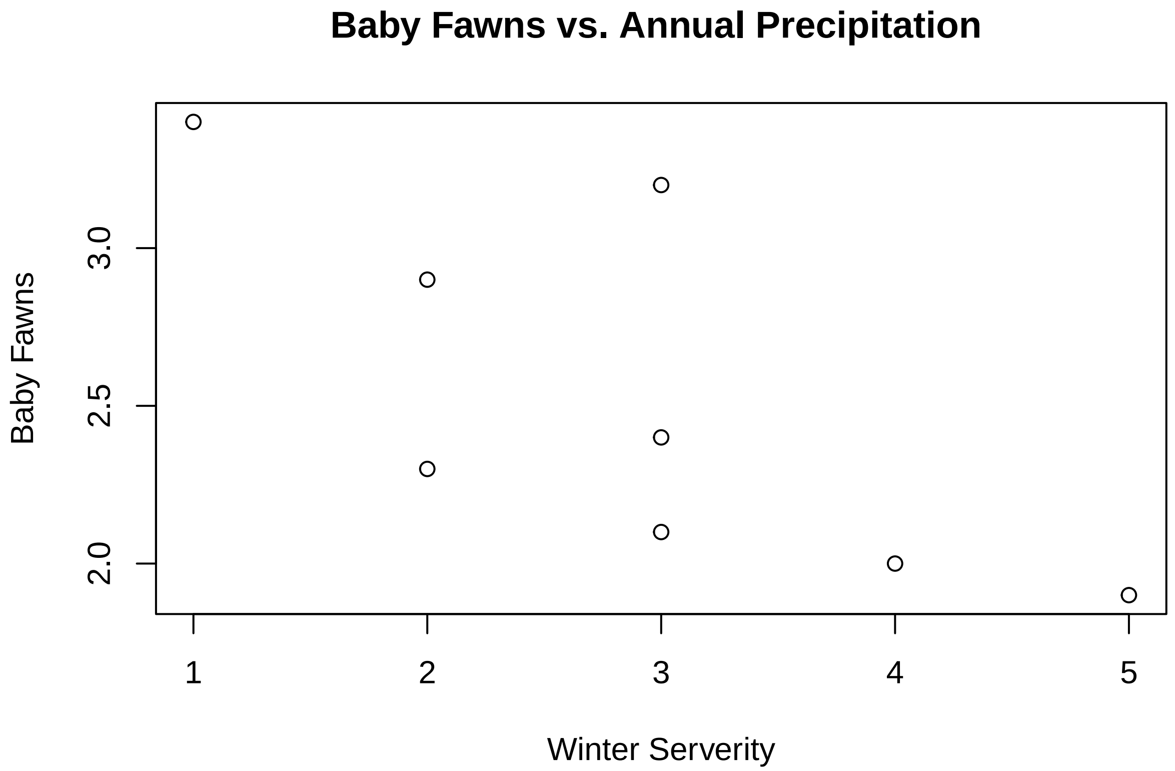
```
#Baby Fawns vs. Adult Antelope
plot(antelopeData$adultPop, antelopeData$springFawn, main = "Baby Fawns vs. Adult Antelope", xlab = "Adu
```



```
#Baby Fawns vs. Precipitation
plot(antelopeData$annualPrecip,antelopeData$springFawn,  main = "Baby Fawns vs. Precipitation ", xlab =
```



```
#Baby Fawns vs. Winter Severity
plot(antelopeData$winterbadness,antelopeData$springFawn, main = "Baby Fawns vs. Annual Precipitation ",
```



5. Next, create three regression models of increasing complexity using lm().

In the first model, predict the number of fawns from the severity of the winter.

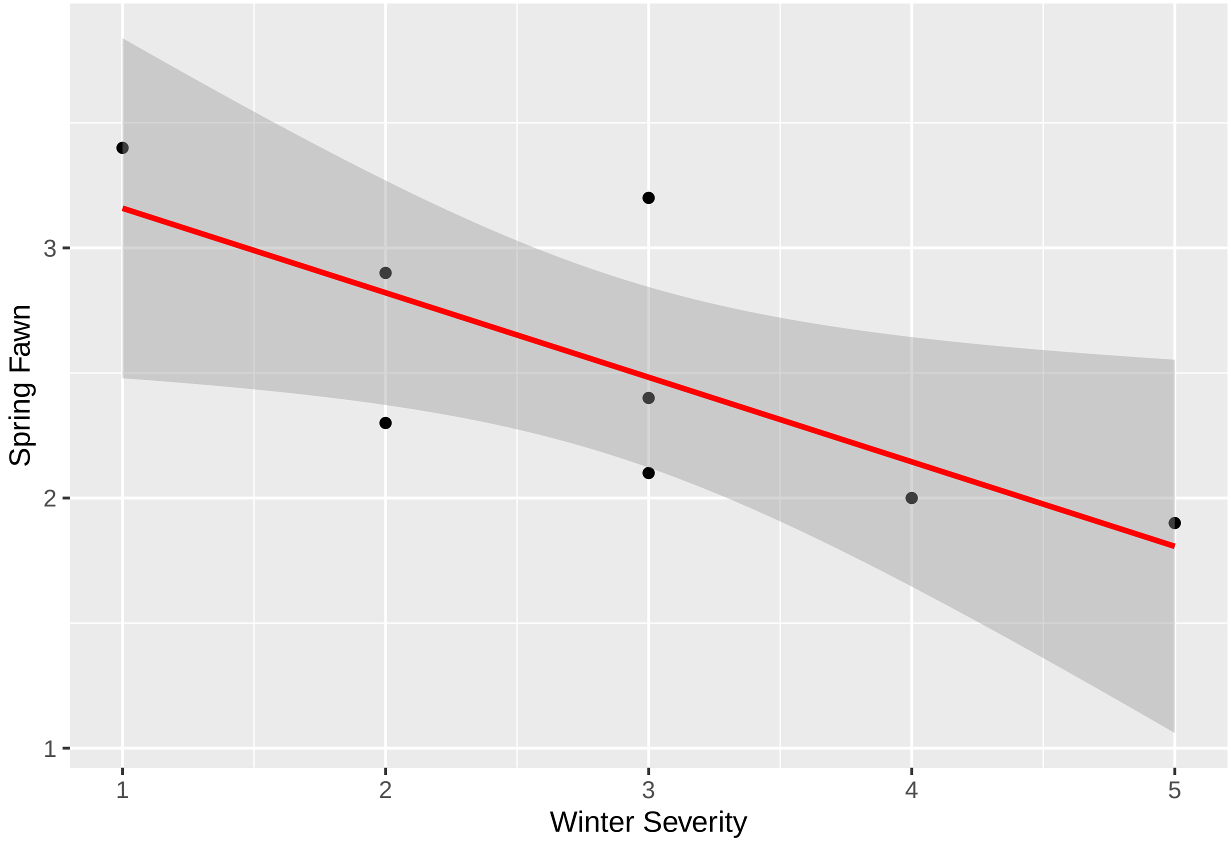
```
modelFawnsWinter <- lm(formula = springFawn ~ winterbadness, data=antelopeData)
summary(modelFawnsWinter)
```

```
##
## Call:
## lm(formula = springFawn ~ winterbadness, data = antelopeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52069 -0.20431 -0.00172  0.13017  0.71724
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.4966     0.3904   8.957 0.000108 ***
## winterbadness  -0.3379     0.1258  -2.686 0.036263 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

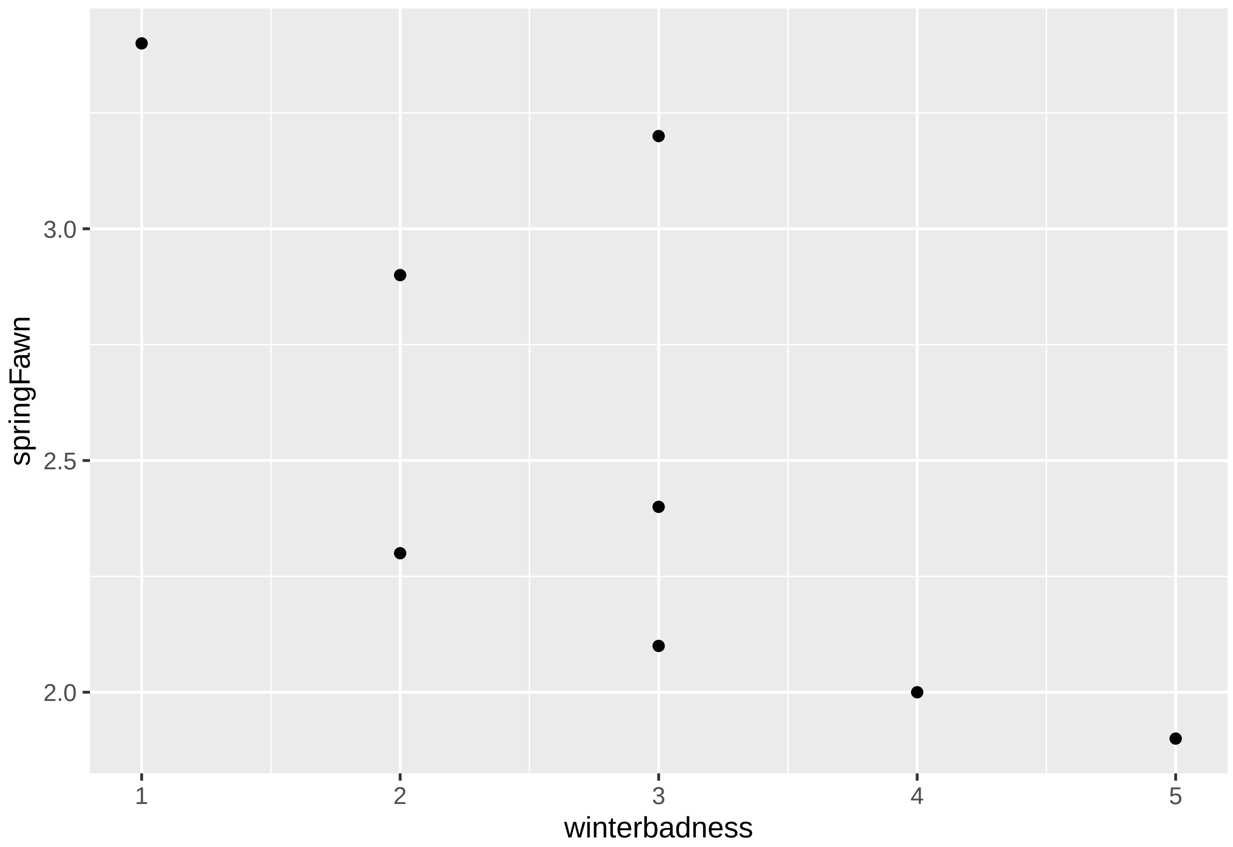
```
##
## Residual standard error: 0.415 on 6 degrees of freedom
## Multiple R-squared:  0.5459, Adjusted R-squared:  0.4702
## F-statistic: 7.213 on 1 and 6 DF,  p-value: 0.03626
```

```
g <- ggplot(antelopeData, aes(y=springFawn, x=winterbadness)) + geom_point()
g + stat_smooth(method = lm, "col" = "red") + ylab("Spring Fawn") + xlab("Winter Severity")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
g
```



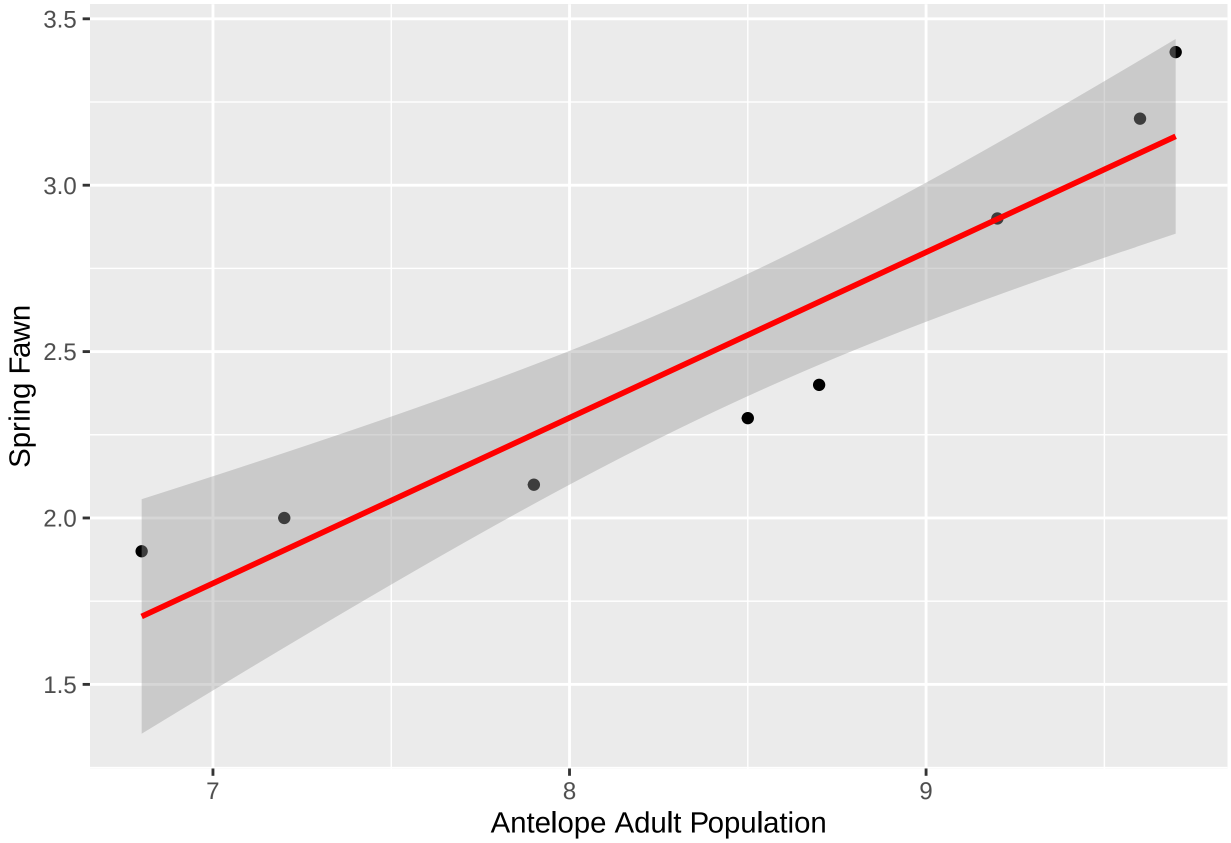
In the second model, predict the number of fawns from two variables (one should be the severity of the winter).

```
model2FawnsAdultsWinter <- lm(formula = springFawn ~ winterbadness + adultPop, data=antelopeData)
summary(model2FawnsAdultsWinter)
```

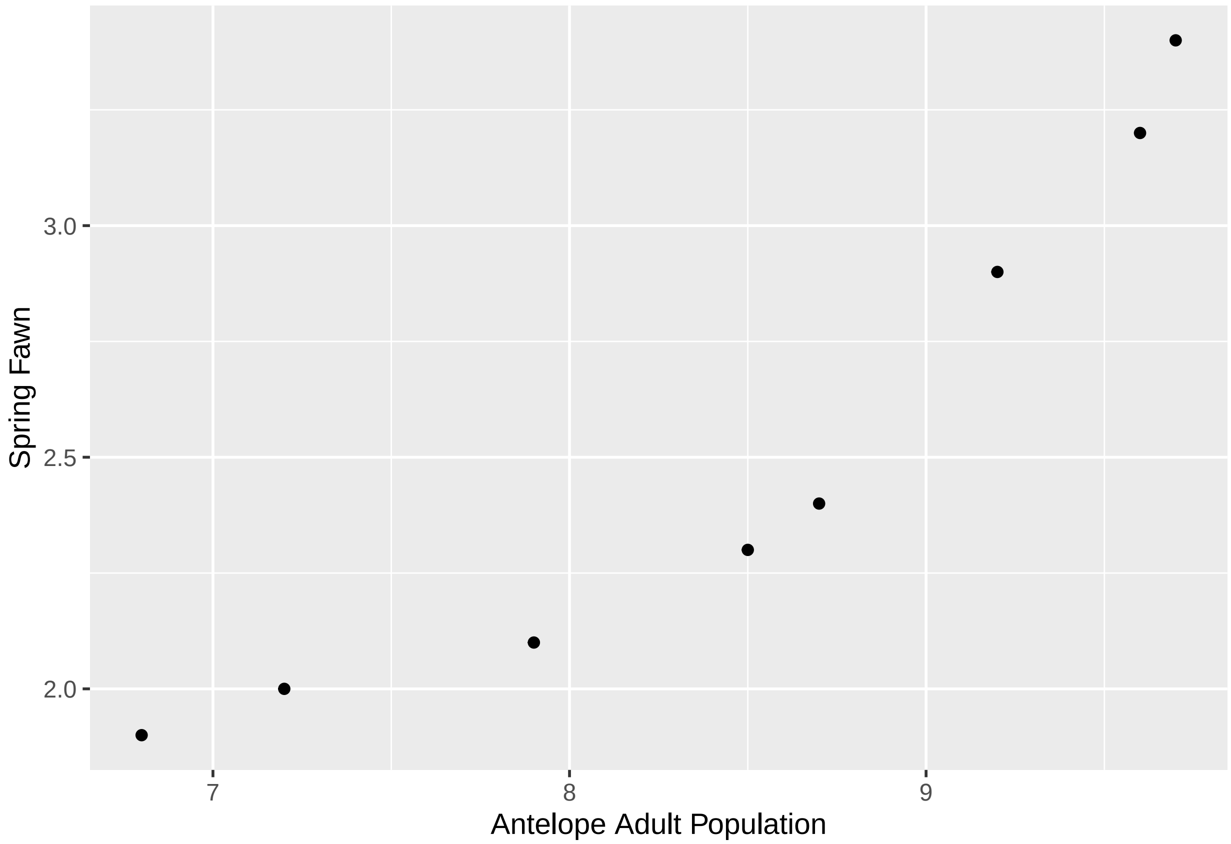
```
##
## Call:
## lm(formula = springFawn ~ winterbadness + adultPop, data = antelopeData)
##
## Residuals:
##        1        2        3        4        5        6        7        8
##  0.01231 -0.27531  0.10301 -0.19154  0.01535  0.15880  0.29992 -0.12256
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.46009    1.53443  -1.603   0.1698
## winterbadness  0.07058    0.12461   0.566   0.5956
## adultPop       0.56594    0.14439   3.920   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2252 on 5 degrees of freedom
## Multiple R-squared:  0.8885, Adjusted R-squared:  0.8439
## F-statistic: 19.92 on 2 and 5 DF,  p-value: 0.004152
```

```
g <- ggplot(antelopeData, aes(y=springFawn, x=adultPop)) + geom_point() + ylab("Spring Fawn") + xlab("A
g + stat_smooth(method = lm, "col" = "red")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
g
```



In the third model predict the number of fawns from the three other variables.

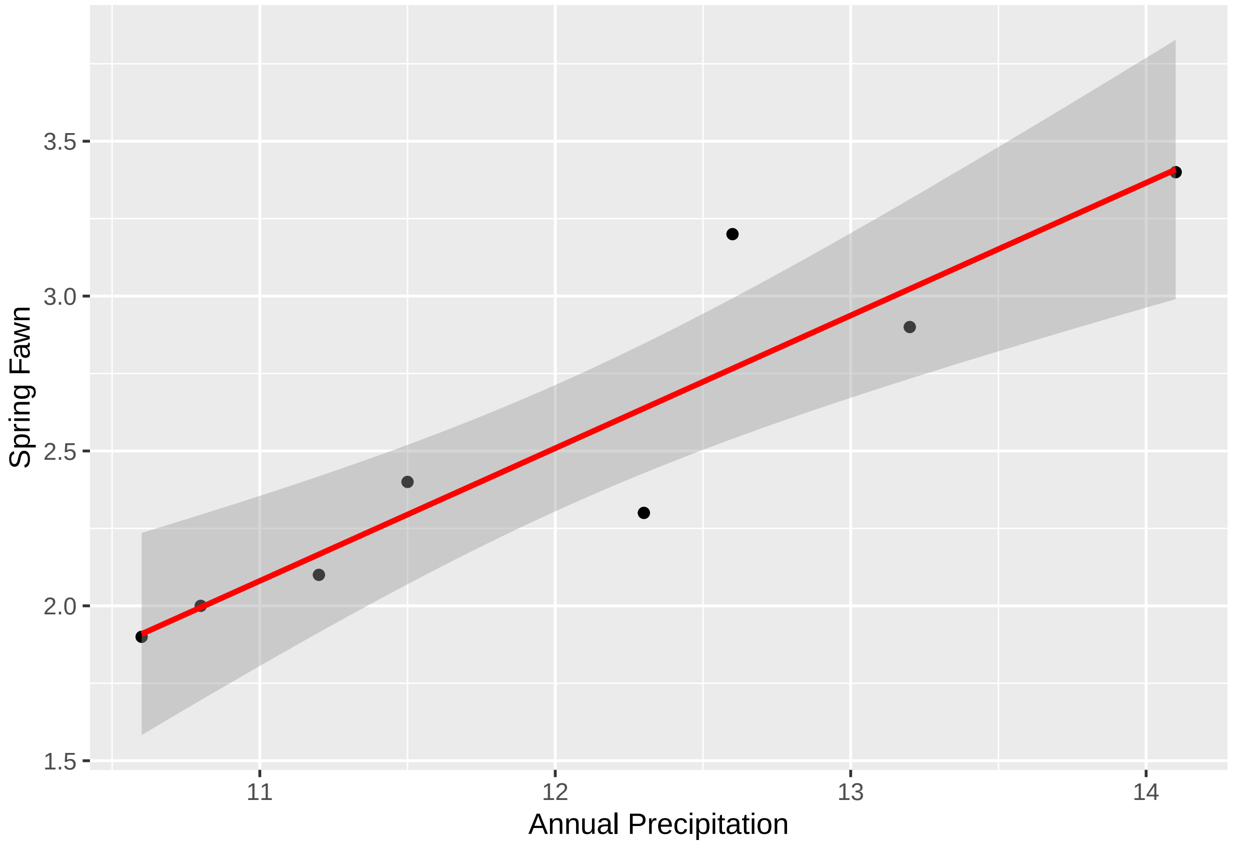
```
model3FawnsAllData <- lm(formula = springFawn ~ winterbadness + adultPop + annualPrecip, data=antelopeDa
thirdModelSum <-summary(model3FawnsAllData)
thirdModelSum
```

```
##
## Call:
## lm(formula = springFawn ~ winterbadness + adultPop + annualPrecip,
##     data = antelopeData)
##
## Residuals:
##        1        2        3        4        5        6        7        8
## -0.11533 -0.02661  0.09882 -0.11723  0.02734 -0.04854  0.11715  0.06441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.92201    1.25562  -4.716   0.0092 **
## winterbadness  0.26295    0.08514   3.089   0.0366 *
## adultPop       0.33822    0.09947   3.400   0.0273 *
## annualPrecip   0.40150    0.10990   3.653   0.0217 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

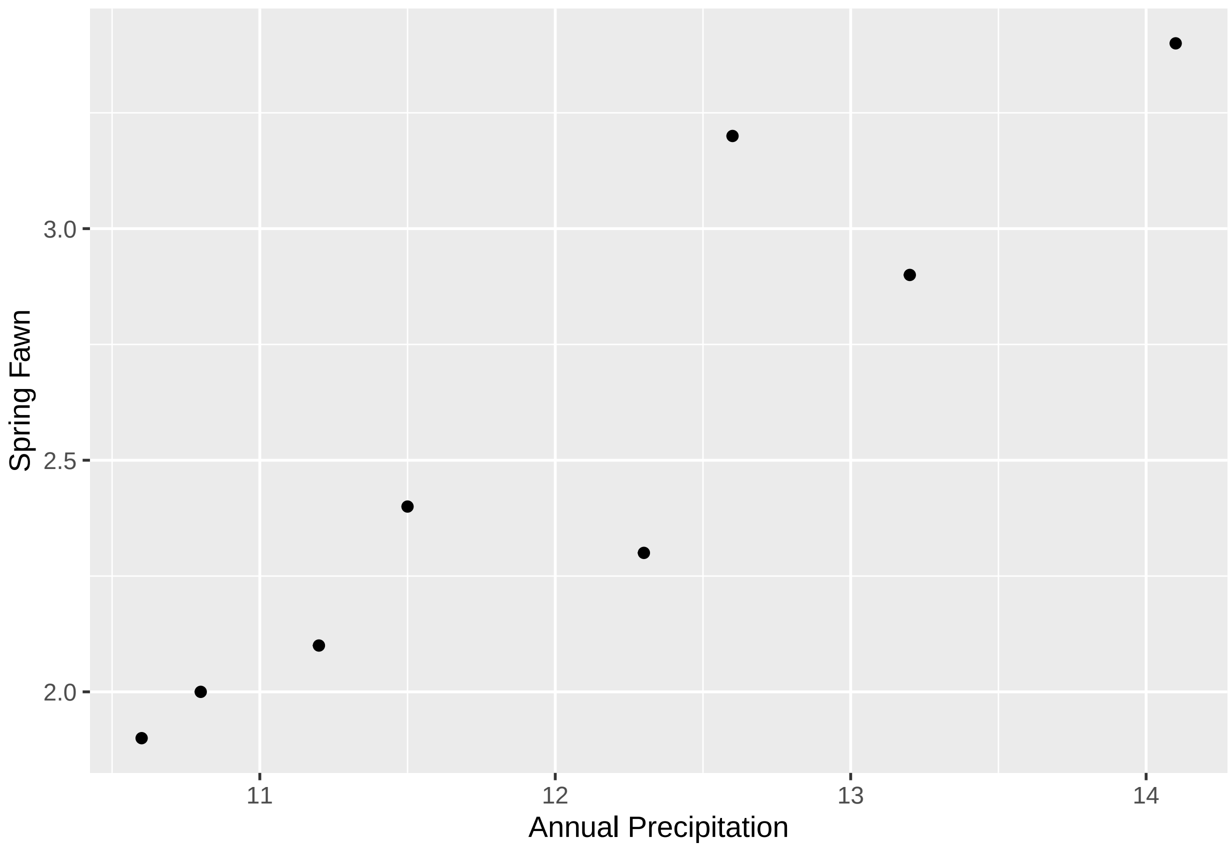
```
## Residual standard error: 0.1209 on 4 degrees of freedom
## Multiple R-squared:  0.9743, Adjusted R-squared:  0.955
## F-statistic: 50.52 on 3 and 4 DF,  p-value: 0.001229
```

```
g <- ggplot(antelopeData, aes(y=springFawn, x=annualPrecip)) + geom_point()+ ylab("Spring Fawn") + xlab
g + stat_smooth(method = lm, "col" = "red")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
g
```



Which model works best?

After developing the three models, the last ("model3FawnsAllData") proved to be the "best" model as it possesses the highest "Adjusted R-squared" value of 0.955. This model is more accurate as it incorporates the most data (adult population, annual precipitation and winter severity). We are also able to see that as we added columns of our data frame to the model, we observed an increasing adjusted R-squared value, this indicates that as we added data we gained a more accurate model, however this does not mean that should we gain new data that new data added to the data frame will increase prediction accuracy. We would expect it will be we cannot be certain.

Which of the predictors are statistically significant in each model? If you wanted to create the most parsimonious model (i.e., the one that did the best job with the fewest predictors), what would it contain?

Below are the P values for each of the coefficients. Here we are able to observe that the "winterbadness" (severity), possesses the highest value. This shows us that if we are to create the most parsimonious model, we would choose the second model which contains the columns "springfawns" and "winterbadness" (winter severity).

```
# Display P-Values
thirdModelSum$coef[,4]
```

```
##   (Intercept) winterbadness      adultPop  annualPrecip
##   0.009196072   0.036626174   0.027272444   0.021707219
```