HW 8

Marley Akonnor

9/6/2021

Step 1: Load the Data Access read.xls function from the gdata package

```
library(gdata)
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
##
## Attaching package: 'gdata'
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
```

Read the data

```
antelope_pop <- read.xls("/Users/m/Documents/M.S Syracuse Data Science/Courses/IST 687 - Intro to Data S
```

Rename the columns

```
new_column_names <- c("Fawn_Population", "Antelope_Population", "Annual_Precipitation", "Winter_Severity
```

Assign the column names to the dataset

```
colnames(antelope_pop) <- new_column_names
```

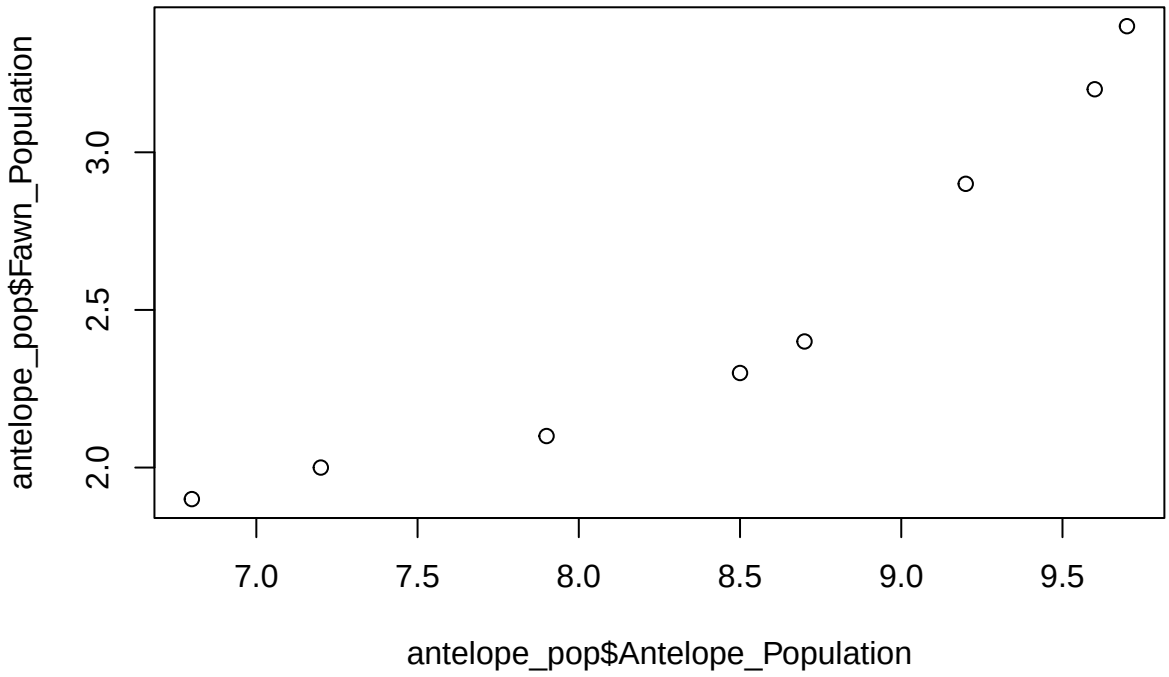
Inspect the dataset

```
str(antelope_pop)
```

```
## 'data.frame':    8 obs. of  4 variables:
##  $ Fawn_Population     : num  2.9 2.4 2 2.3 3.2 ...
##  $ Antelope_Population : num  9.2 8.7 7.2 8.5 9.6 ...
##  $ Annual_Precipitation: num  13.2 11.5 10.8 12.3 12.6 ...
##  $ Winter_Severity     : int  2 3 4 2 3 5 1 3
```

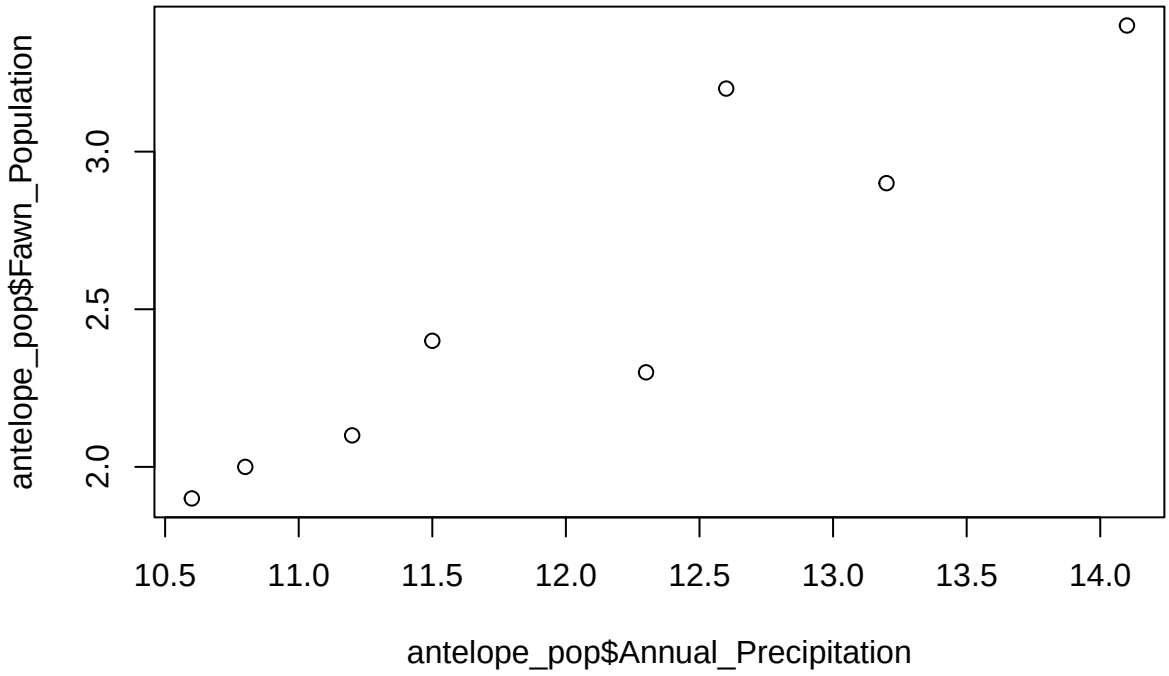
Create bivariate plots of number of baby fawns versus adult antelope population, the precipitation that year, and the severity of the winter. fawn(y) vs antelope(x)

```
plot(antelope_pop$Antelope_Population,antelope_pop$Fawn_Population)
```



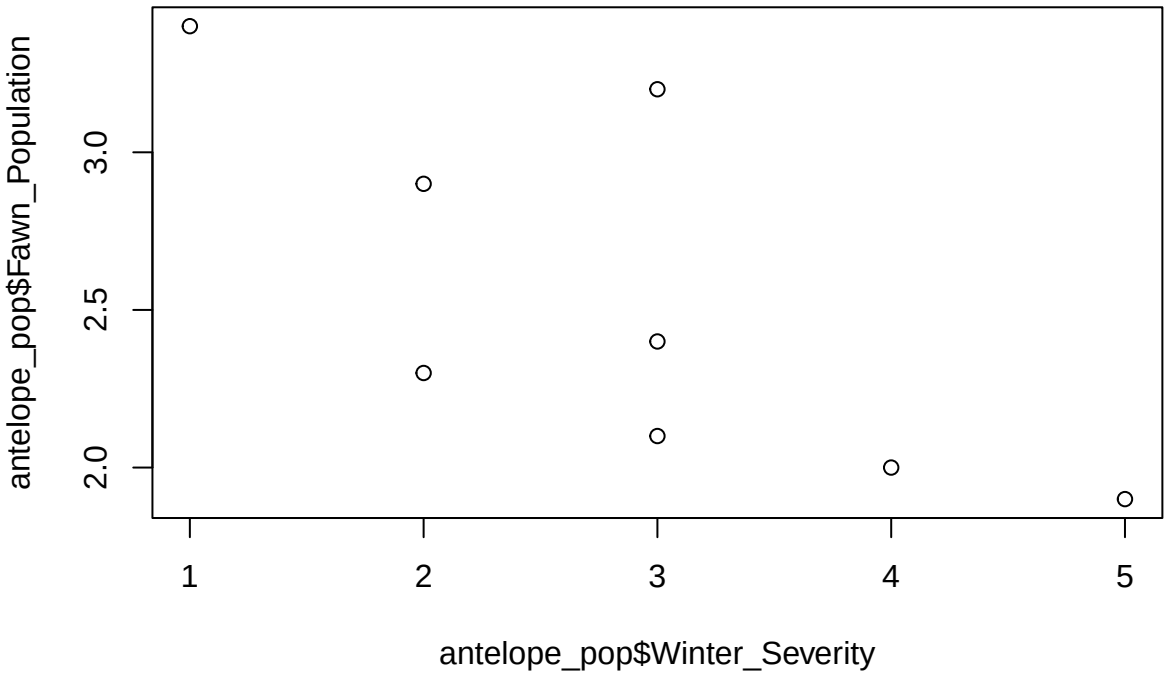
fawn(y) vs precipitation(x)

```
plot(antelope_pop$Annual_Precipitation,antelope_pop$Fawn_Population)
```



fawn(y) vs winter(x)

```
plot(antelope_pop$Winter_Severity,antelope_pop$Fawn_Population)
```



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

```
fawn_wint_model_1 <- lm(formula = Fawn_Population ~ Winter_Severity, data = antelope_pop)
summary(fawn_wint_model_1)
```

```
##
## Call:
## lm(formula = Fawn_Population ~ Winter_Severity, data = antelope_pop)
##
## 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 ***
## Winter_Severity  -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
```

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

```
fawn_precip_wint_model_2 <- lm(formula = Fawn_Population ~ Annual_Precipitation + Winter_Severity, data
summary(fawn_precip_wint_model_2)
```

```
##
```

```
## Call:
## lm(formula = Fawn_Population ~ Annual_Precipitation + Winter_Severity,
##     data = antelope_pop)
##
## Residuals:
##         1         2         3         4         5         6         7         8
## -0.165458  0.188313  0.006417 -0.193358  0.289080 -0.193312 -0.010695  0.079013
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -5.7791     2.2139  -2.610  0.04765 *
## Annual_Precipitation   0.6357     0.1511   4.207  0.00843 **
## Winter_Severity        0.2269     0.1490   1.522  0.18842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2133 on 5 degrees of freedom
## Multiple R-squared:    0.9,  Adjusted R-squared:   0.86
## F-statistic: 22.49 on 2 and 5 DF,  p-value: 0.003164
```

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

```
all_var_model_3 <- lm(formula = Fawn_Population ~ Antelope_Population + Annual_Precipitation + Winter_S
summary(all_var_model_3)
```

```
##
## Call:
## lm(formula = Fawn_Population ~ Antelope_Population + Annual_Precipitation +
##     Winter_Severity, data = antelope_pop)
##
## 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 **
## Antelope_Population   0.33822    0.09947   3.400   0.0273 *
## Annual_Precipitation  0.40150    0.10990   3.653   0.0217 *
## Winter_Severity       0.26295    0.08514   3.089   0.0366 *
## ---
## 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
```

Which model works best? The third model works the best. The multiple R-squared is 97% and the F statistic P-value is the lowest of the 3.

Which of the predictors are statistically significant in each model? The statistically significant predictor in model 1 is: Winter Severity The statistically significant predictor in model 2 is: Annual Precipitation The statistically significant predictors in model 3 are: Winter Severity, Annual Precipitation, and Antelope Population

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? It would contain the Antelope Population and Annual Precipitation.