Bing-Je_WU_HW8

Bing-Je Wu

6/1/2019

Step 1. Load the data set

Download data set. 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. Finally, the last column shows how bad the winter was during that year.

```
library(gdata)
mydf <- read.xls("mlr01.xls")
mydf
```

```
##    X1  X2   X3 X4
## 1 2.9 9.2 13.2  2
## 2 2.4 8.7 11.5  3
## 3 2.0 7.2 10.8  4
## 4 2.3 8.5 12.3  2
## 5 3.2 9.6 12.6  3
## 6 1.9 6.8 10.6  5
## 7 3.4 9.7 14.1  1
## 8 2.1 7.9 11.2  3
```

Step 2.

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(mydf)
```

```
## 'data.frame':    8 obs. of  4 variables:
##  $ X1: num  2.9 2.4 2 2.3 3.2 ...
##  $ X2: num  9.2 8.7 7.2 8.5 9.6 ...
##  $ X3: num  13.2 11.5 10.8 12.3 12.6 ...
##  $ X4: int  2 3 4 2 3 5 1 3
```

```
antelopeDF <- data.frame(mydf$X1, mydf$X2, mydf$X3, as.factor(mydf$X4))
str(antelopeDF)
```

```
## 'data.frame':    8 obs. of  4 variables:
##  $ mydf.X1           : num  2.9 2.4 2 2.3 3.2 ...
##  $ mydf.X2           : num  9.2 8.7 7.2 8.5 9.6 ...
##  $ mydf.X3           : num  13.2 11.5 10.8 12.3 12.6 ...
##  $ as.factor.mydf.X4.: Factor w/ 5 levels "1","2","3","4",..: 2 3 4 2 3 5 1 3
```

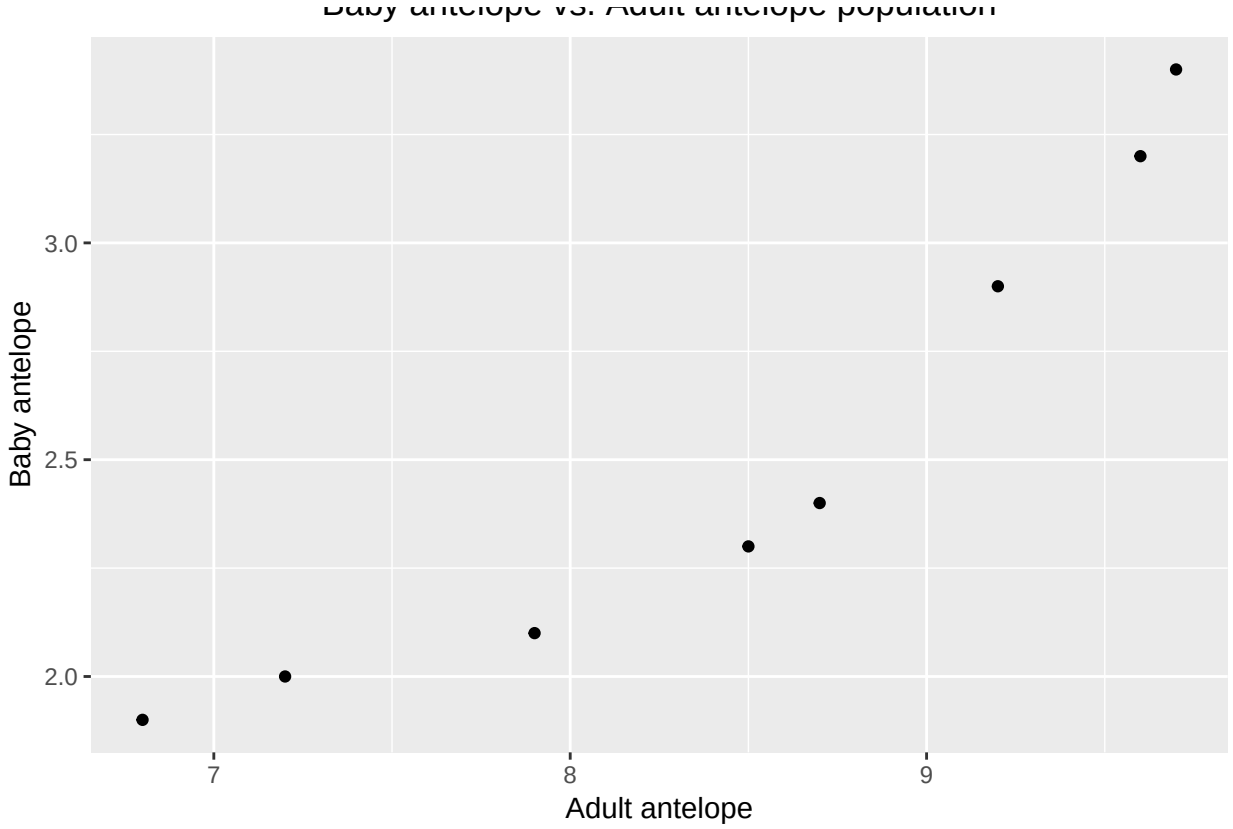
```
colnames(antelopeDF) <- c("baby","adult","precipitation","winter")
```

Step 3.

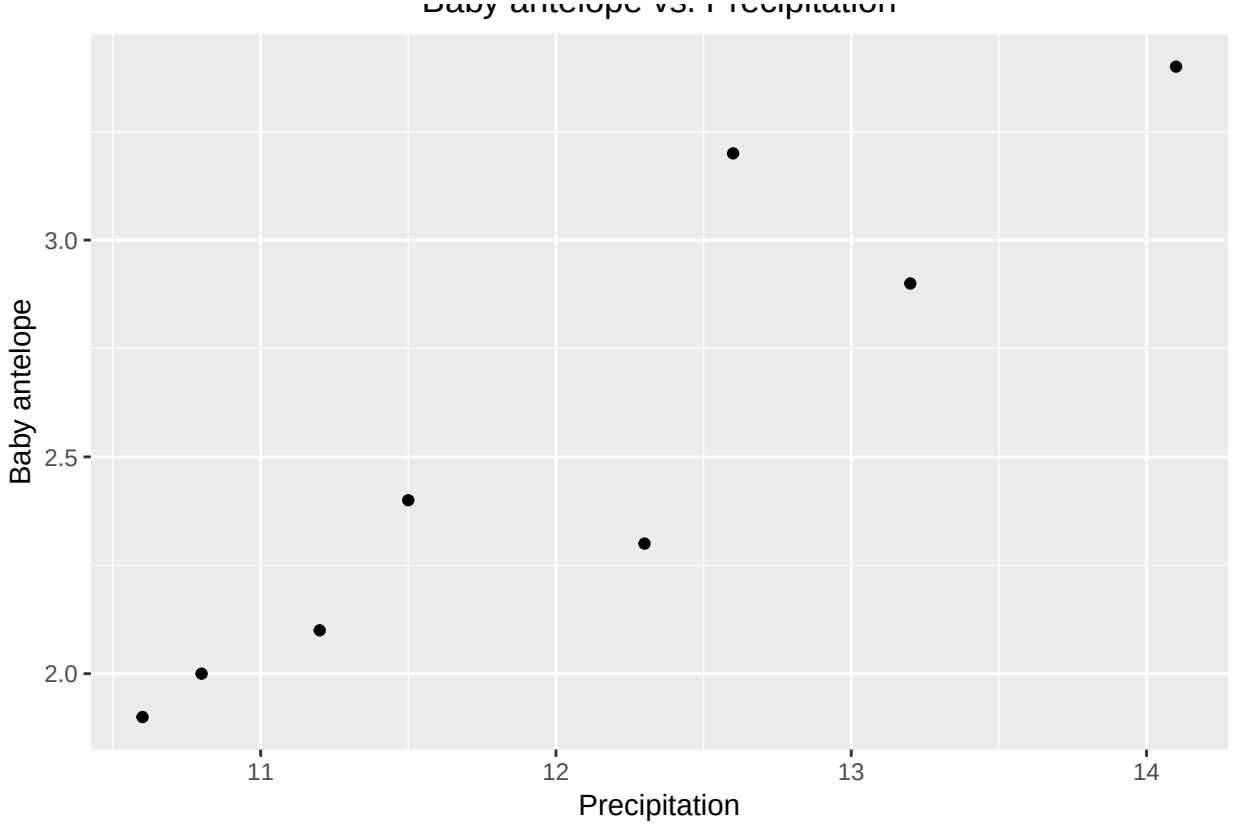
Create bivariate plots of number of baby fawns versus adult antelope population, the precipitation that year, and the severity of the winter. Your code should produce three separate plots. Make sure the Y-axis and

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?

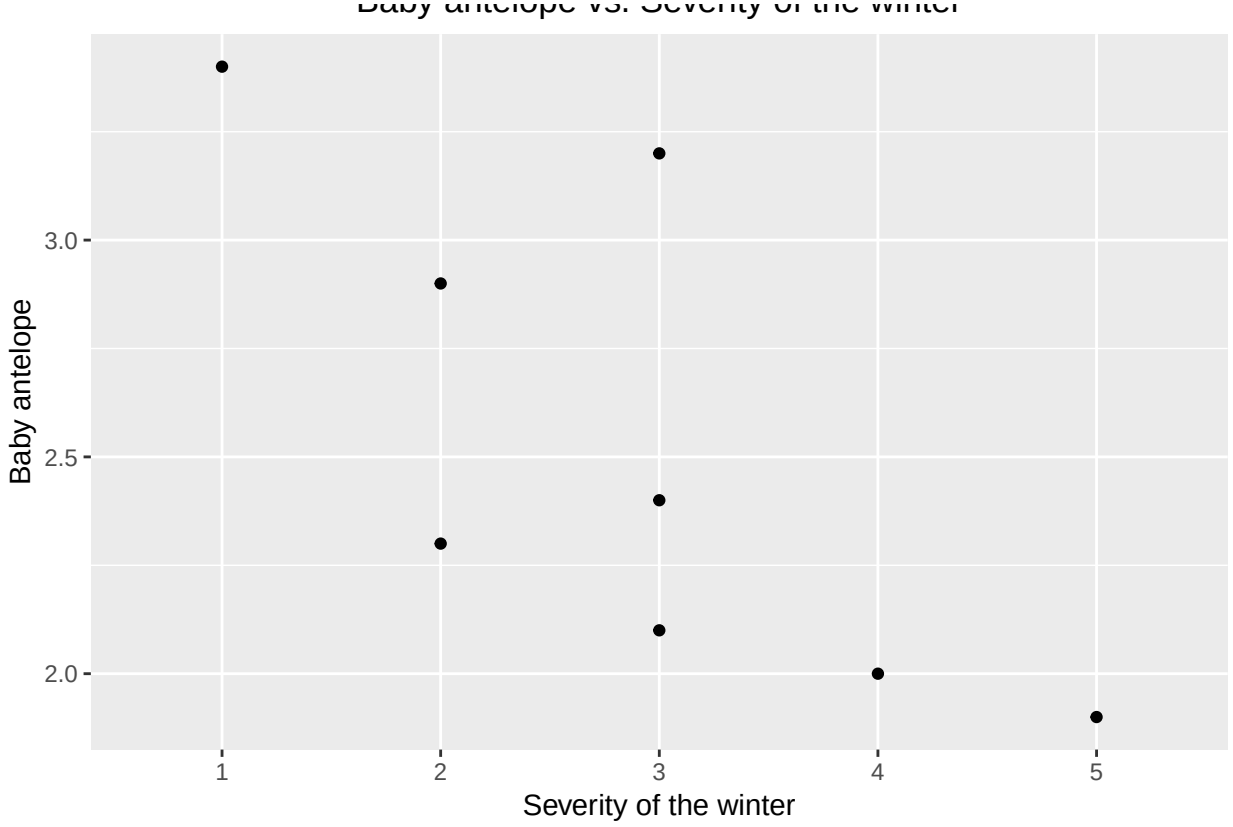
```
library(ggplot2)
 # Baby antelope vs. adult antelope
ggplot(antelopeDF, aes(x=adult, y=baby)) +geom_point() + xlab("Adult antelope") +
  ylab("Baby antelope") + ggtitle("Baby antelope vs. Adult antelope population") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Baby antelope vs. precipitation
ggplot(antelopeDF, aes(x=precipitation, y=baby)) +geom_point() + xlab("Precipitation") +
  ylab("Baby antelope") + ggtitle("Baby antelope vs. Precipitation") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
 # Baby antelope vs. Severity   of the winter
ggplot(antelopeDF, aes(x=winter, y=baby)) +geom_point() +
  xlab("Severity of the winter") + ylab("Baby antelope") +
  ggtitle("Baby antelope vs. Severity of the winter") +
  theme(plot.title = element_text(hjust = 0.5))
```



Step 4.

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. In the second model, predict the number of fawns from two variables (one should be the severity of the winter). In the third model predict the number of fawns from the three other variables. Which model works best? 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?

Fisrt model

```
model_1 <- lm(formula = baby~winter,data = antelopeDF)
summary(model_1)
```

```
##
## Call:
## lm(formula = baby ~ winter, data = antelopeDF)
##
## Residuals:
##          1          2          3          4          5          6
##  3.000e-01 -1.667e-01  1.249e-16 -3.000e-01  6.333e-01 -9.714e-17
##          7          8
## -1.527e-16 -4.667e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4000     0.5249   6.477  0.00747 **
```

```
## winter2      -0.8000     0.6429  -1.244  0.30173
## winter3      -0.8333     0.6061  -1.375  0.26288
## winter4      -1.4000     0.7424  -1.886  0.15579
## winter5      -1.5000     0.7424  -2.021  0.13658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5249 on 3 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.1521
## F-statistic: 1.314 on 4 and 3 DF,  p-value: 0.4282
```

```
# There is no predictor is statistically significant.
```

Second model

```
model_2 <- lm(formula = baby~winter + adult,data = antelopeDF)
summary(model_2)
```

```
##
## Call:
## lm(formula = baby ~ winter + adult, data = antelopeDF)
##
## Residuals:
##          1          2          3          4          5          6
##  6.138e-02 -1.439e-01  6.245e-17 -6.138e-02  4.246e-02 -6.939e-18
##          7          8
## -9.021e-17  1.015e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.2132     1.0688  -3.006   0.0951 .
## winter2      -0.2205     0.1972  -1.118   0.3798
## winter3      -0.1743     0.1951  -0.893   0.4659
## winter4       0.3044     0.3390   0.898   0.4639
## winter5       0.4771     0.3751   1.272   0.3312
## adult         0.6818     0.1092   6.243   0.0247 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.142 on 2 degrees of freedom
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9379
## F-statistic: 22.15 on 5 and 2 DF,  p-value: 0.04376
```

```
# The number of adult antelopes is statistically significant.
# The adjust R-square 0.9379. It means 94% of y values can be explained by X variables.
# This model works best among three models.
```

Third model

```
model_3 <- lm(formula = baby~winter + adult + precipitation, data = antelopeDF)
summary(model_3)
```

```
##
## Call:
## lm(formula = baby ~ winter + adult + precipitation, data = antelopeDF)
##
## Residuals:
```

```
##          1          2          3          4          5          6
## -1.399e-02 -1.261e-02 -1.735e-18  1.399e-02  1.169e-02 -3.469e-18
##          7          8
##  3.469e-18  9.171e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.20532    0.44067 -14.081   0.0451 *
## winter2        0.10518    0.05626   1.870   0.3127
## winter3        0.63878    0.11296   5.655   0.1114
## winter4        0.89031    0.09934   8.962   0.0707 .
## winter5        0.97481    0.09534  10.224   0.0621 .
## adult          0.18389    0.06859   2.681   0.2273
## precipitation  0.55472    0.07305   7.594   0.0834 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02623 on 1 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9979
## F-statistic: 551.1 on 6 and 1 DF,  p-value: 0.0326
```

```
# There is no predictor is statistically significant.
```

The most parsimonious model

```
model_4 <- lm(formula = baby~adult, data=antelopeDF)
summary(model_4)
```

```
##
## Call:
## lm(formula = baby ~ adult, data = antelopeDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24988 -0.17586  0.04938  0.12611  0.25309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.67914    0.63422  -2.648 0.038152 *
## adult        0.49753    0.07453   6.676 0.000547 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2121 on 6 degrees of freedom
## Multiple R-squared:  0.8813, Adjusted R-squared:  0.8616
## F-statistic: 44.56 on 1 and 6 DF,  p-value: 0.0005471
```

```
# Baay versus addult model will be the most parsimonious model.
# adult variable is statistically significant. R-squared is 0.8813.
# P-value is 0.0005 . The equation is also statistically significant.
```