Week 8: Lab - Linear Modeling (Making Predictions)

Alex Koener

3/2/2021

# Instructions

The textbook’s chapter on linear models (“Line Up, Please”) introduces **linear predictive modeling** using the workhorse tool known as **multiple regression**. The term “multiple regression” has an odd history, dating back to an early scientific observation of a phenomenon called **“regression to the mean”**. These days, multiple regression is just an interesting name for using a simple linear modeling technique to measuring the connection between one or more **predictor variables** and an **outcome variable**.

In this exercise, we are going to use an open dataset to explore antelope population.

This is the first exercise of the semester where there is no sample R code to help you along. Because you have had so much practice with R by now, you can create and/or find all of the code you need to accomplish these steps.

# Add your library below.  
library(tidyverse)  
library(readxl)

# Step 1 - Define “Model”

Write a definition of a model, based on how the author uses it in this chapter.

A predictive model is created from a statistical analysis process in R. These models analyze data that the user supplies and then calculate a set of numerical coefficients that help with a prediction.

# Step 2 - Review the data

You can find the data from Cengage’s website. This URL will enable you to download the dataset into excel:

* <http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/excel/mlr01.xls>

The more general website can be found at:

* <http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html>

If you view this in a spreadsheet, you will find 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.

# No code necessary; Just review the data.

# Step 3 - Read in the data

You have the option of saving the file to your computer and reading it into R, or you can read the data directly from the web into a dataframe.

# Write your code below.  
df <- as.data.frame(read\_excel("data.xls"))  
colnames(df) <- c("fawnPop", "antelopePop", "rainAmount", "weatherLevel")

# Step 4 - Inspect the data

You should inspect the data using str() to make sure that 1) all the cases have been read in (n=8 years of observations) and 2) that there are four variables.

# Write your code below.  
str(df)

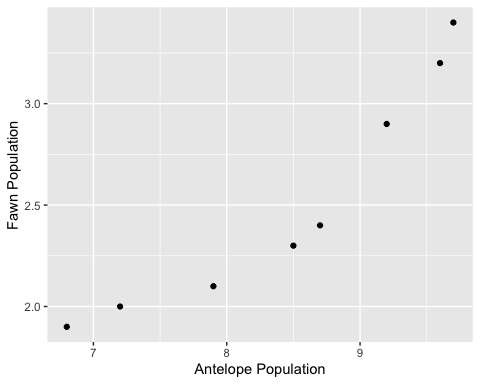
## 'data.frame': 8 obs. of 4 variables:  
## $ fawnPop : num 2.9 2.4 2 2.3 3.2 ...  
## $ antelopePop : num 9.2 8.7 7.2 8.5 9.6 ...  
## $ rainAmount : num 13.2 11.5 10.8 12.3 12.6 ...  
## $ weatherLevel: num 2 3 4 2 3 5 1 3

# Step 5 - Create bivariate plots

Create bivariate plots of the number of baby fawns versus adult antelope population, precipitation that year, and 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? You can also create scatter plots where size and colors reflect the two variables you didn’t use (remember the visualization homework/lab. If you create these plots, you can earn extra 1 point).

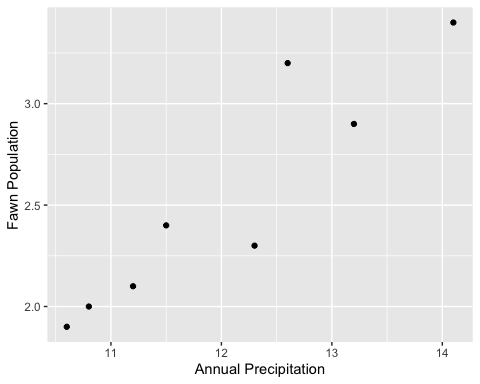
## Step 5.1 - Fawn Count by Adult Population

# Write your code below.  
basePlot <- ggplot(df, aes(y=fawnPop)) +  
 labs(y="Fawn Population")  
  
basePlot +   
 geom\_point(aes(x=antelopePop)) +  
 labs(x="Antelope Population")



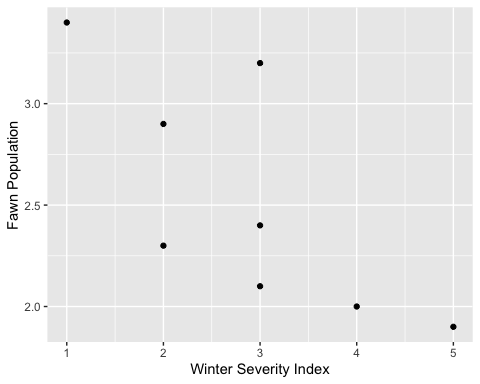
## Step 5.2 - Fawn Count by Annual Precipitation

# Write your code below.  
basePlot +   
 geom\_point(aes(x=rainAmount)) +  
 labs(x="Annual Precipitation")



## Step 5.3 - Fawn Count by Winter Severity Index

# Write your code below.  
basePlot +   
 geom\_point(aes(x=weatherLevel)) +  
 labs(x="Winter Severity Index")



# Step 6 - Create regression models

Create three regression models of increasing complexity using lm(), then analyze the results.

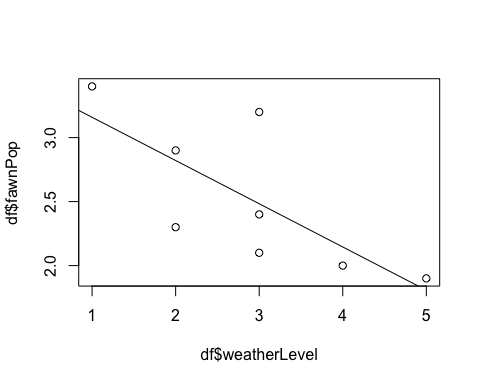
* **Model one**: Fit the model to predict the number of fawns from the severity of the winter.
* **Model two**: Fit the model to predict the number of fawns from two variables (one should be the severity of the winter).
* **Model three**: Fit the model to predict the number of fawns from the three other variables.

## Step 6.1 - Predict Fawn Count by Winter Severity Index

# Write your code below.  
m1 <- lm(formula=fawnPop ~ weatherLevel, df)  
summary(m1)

##   
## Call:  
## lm(formula = fawnPop ~ weatherLevel, data = df)  
##   
## 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 \*\*\*  
## weatherLevel -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

plot(x=df$weatherLevel, y=df$fawnPop) +  
abline(m1)



## integer(0)

test1 <- data.frame(weatherLevel=1)  
predict(m1, test1, type="response")

## 1   
## 3.158621

## Step 6.2 - Predict Fawn Count by Winter Severity Index + your choice of variable

# Write your code below.  
m2 <- lm(formula=fawnPop ~ weatherLevel + rainAmount, df)  
summary(m2)

##   
## Call:  
## lm(formula = fawnPop ~ weatherLevel + rainAmount, data = df)  
##   
## 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 \*   
## weatherLevel 0.2269 0.1490 1.522 0.18842   
## rainAmount 0.6357 0.1511 4.207 0.00843 \*\*  
## ---  
## 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

test2 <- data.frame(weatherLevel=0, rainAmount=0)  
predict(m2, test2, type="response")

## 1   
## -5.779064

## Step 6.3 - Predict Fawn Count by the three other variables

# Write your code below.  
m3 <- lm(formula=fawnPop ~ weatherLevel + rainAmount + antelopePop, df)  
summary(m3)

##   
## Call:  
## lm(formula = fawnPop ~ weatherLevel + rainAmount + antelopePop,   
## data = df)  
##   
## 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 \*\*  
## weatherLevel 0.26295 0.08514 3.089 0.0366 \*   
## rainAmount 0.40150 0.10990 3.653 0.0217 \*   
## antelopePop 0.33822 0.09947 3.400 0.0273 \*   
## ---  
## 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

test3 <- data.frame(weatherLevel=0, rainAmount=0, antelopePop=0)  
predict(m3, test3, type="response")

## 1   
## -5.922011

## Step 6.4 - Analysis

Which regression model works best?

The final regression model that includes 3 variables (antelopes, weather, rain) seems to be our most effective model, the adjusted r-squared value is the highest at (0.955). This number represents that there is ALMOST a certain correlation between the independent and dependant variables within this model.

Which of the predictors are statistically significant in each model?

All the predictors (antelopes, weather, rain) were statistically significant with (alpha = .05). In other words: we’re 95% sure that the results variation is due to each predictor/variable and not random chance. OR: There’s less than a 5% chance that the varibales aren’t effecting the dawn 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?

The most parismonious model would include only the rain amount as the independent variable because it is most likely, based on the p-value, that the rain amount is potentially effecting the fawn population.

# Step 7 - Upload the compiled file

Please only include print outs of data sets using “head” function. I will take points off if you include more than two pages of dataset print outs.