Alonso\_Week 8 Homework Assignment

Martin Alonso

9/4/2018

IST687 Introduction to Data Science: Week 8 Homework

Making Predictions.

# Part 1: The Code

#1. Load required packages and load the data set.   
require(RCurl)  
require(gdata)  
require(readxl)  
require(ggplot2)  
  
#Note: reading this URL and loading it resulted in an error. I opted for downloading the data set and loading it from my folder using read\_excel from the readxl library.   
url <- 'http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls'  
  
#2. Load the data directly from excel.   
df <- read\_excel('mlr01.xls')  
  
#3. Initial data set exploration - reviewing structure, summary, and dimensions  
str(df)  
summary(df)  
dim(df)  
  
#The data set has 8 rows and 4 columns. The columns need to be renamed.   
#The columns represent, respectively: number of fawn in a given spring, adult antelope population, annual precipitation, how bad the winter was.   
  
names <- c('fawnPerSpring', 'adultPop', 'annualPrecipitation', 'winterSeverity')  
colnames(df) <- names  
  
#It would also be wise to add a column for Year: which will go from 1 to 8.  
df$year <- seq(1, 8, by = 1)  
  
#4. Data visualization.   
#Fawn per Spring vs Adult Antelope Population  
ggplot(df, aes(adultPop, fawnPerSpring)) + geom\_point() +   
 labs(x = 'Adult Antelope Population', y = 'Avg Fawn per Spring') +   
 ggtitle('Adult Antelop Population vs Avg Fawn per Spring')  
  
#Annual Precipitation  
ggplot(df, aes(year, annualPrecipitation)) + geom\_line() +  
 labs(x = 'Year', y = 'Annual Precipitation') +   
 ggtitle('Annual Precipitation', subtitle = 'Years Unknown')  
  
#Winter Severity   
ggplot(df, aes(year, winterSeverity)) + geom\_line() +   
 labs(x = 'Year', y = 'Winter Severity') +  
 ggtitle('Winter Severity', subtitle = 'Years Unknown')  
  
#5. Creating, running, and reviewing linear models.   
#Will first verify if severity of winter can predict number of fawn the following spring.  
fawnVsWinter <- lm(fawnPerSpring ~ winterSeverity, data = df)  
summary(fawnVsWinter)  
  
#Winter severity has a p-value of 0.036 - significant at the 0.05 level. However, Adjusted R-squared states that the severity of the previous winter explains 47 percent of the variance of the number of fawn.   
  
#How about checking the severity of the winter and size of the adult population?   
fawnVsWinterAdult <- lm(fawnPerSpring ~ winterSeverity + adultPop, data = df)  
summary(fawnVsWinterAdult)  
  
#When we add the adult population, the Adjusted R-Squared explains 84.4 percent of the variance in the number of fawn. However, neither of the explaining variables are significant - severity of winter has a p-value of 0.596 and adult population has a p-value of 0.011.   
  
#Let's add the remaining column - annual precipitation - to see how it affects the model.   
fawnVsAll <- lm(fawnPerSpring ~ . - year, data = df)  
summary(fawnVsAll)  
  
#By adding annual precipitation, we are able to explain 95.5 percent of the variance in the number of fawn. Additionally, we find that all variables are significant at a p-value of 0.05; which gives us a very good model.   
  
#Let's plot the model to see how the variables relate  
ggplot(df, aes(x = adultPop, y = fawnPerSpring)) +  
 geom\_point(aes(size = annualPrecipitation, color = as.factor(winterSeverity))) +  
 geom\_smooth(method = 'lm', aes(x = adultPop, y = fawnPerSpring)) +  
 labs(x = 'Adult Antelope Population', y = 'Avg Fawn Per Spring',   
 color = 'Winter Severity', size = 'Annual Precipitation') +  
 ggtitle('Avg Fawn Per Spring vs Adult Pop, Winter Severity,\n and Annual Precipitation')

# Part 2: Running the Code

#1. Load required packages and load the data set.   
require(RCurl)

## Loading required package: RCurl

## Loading required package: bitops

require(gdata)

## Loading required package: gdata

## gdata: Unable to locate valid perl interpreter  
## gdata:   
## gdata: read.xls() will be unable to read Excel XLS and XLSX files  
## gdata: unless the 'perl=' argument is used to specify the location  
## gdata: of a valid perl intrpreter.  
## gdata:   
## gdata: (To avoid display of this message in the future, please  
## gdata: ensure perl is installed and available on the executable  
## gdata: search path.)

## gdata: Unable to load perl libaries needed by read.xls()  
## gdata: to support 'XLX' (Excel 97-2004) files.

##

## gdata: Unable to load perl libaries needed by read.xls()  
## gdata: to support 'XLSX' (Excel 2007+) files.

##

## gdata: Run the function 'installXLSXsupport()'  
## gdata: to automatically download and install the perl  
## gdata: libaries needed to support Excel XLS and XLSX formats.

##   
## 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

require(readxl)

## Loading required package: readxl

require(ggplot2)

## Loading required package: ggplot2

#Note: reading this URL and loading it resulted in an error. I opted for downloading the data set and loading it from my folder using read\_excel from the readxl library.   
url <- 'http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls'  
  
#2. Load the data directly from excel.   
df <- read\_excel('mlr01.xls')  
  
#3. Initial data set exploration - reviewing structure, summary, and dimensions  
str(df)

## Classes 'tbl\_df', 'tbl' and '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: num 2 3 4 2 3 5 1 3

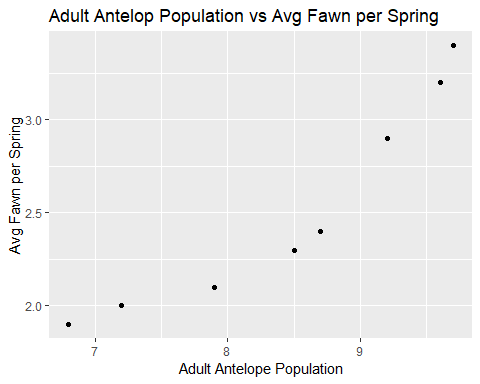
summary(df)

## X1 X2 X3 X4   
## Min. :1.900 Min. :6.800 Min. :10.60 Min. :1.000   
## 1st Qu.:2.075 1st Qu.:7.725 1st Qu.:11.10 1st Qu.:2.000   
## Median :2.350 Median :8.600 Median :11.90 Median :3.000   
## Mean :2.525 Mean :8.450 Mean :12.04 Mean :2.875   
## 3rd Qu.:2.975 3rd Qu.:9.300 3rd Qu.:12.75 3rd Qu.:3.250   
## Max. :3.400 Max. :9.700 Max. :14.10 Max. :5.000

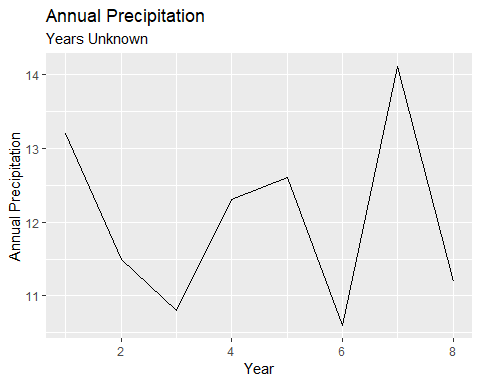
dim(df)

## [1] 8 4

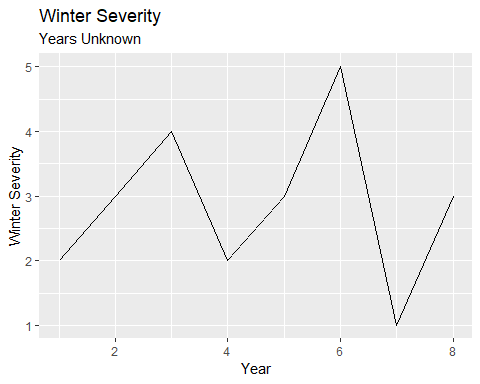
#The data set has 8 rows and 4 columns. The columns need to be renamed.   
#The columns represent, respectively: number of fawn in a given spring, adult antelope population, annual precipitation, how bad the winter was.   
  
names <- c('fawnPerSpring', 'adultPop', 'annualPrecipitation', 'winterSeverity')  
colnames(df) <- names  
  
#It would also be wise to add a column for Year: which will go from 1 to 8.  
df$year <- seq(1, 8, by = 1)  
  
#4. Data visualization.   
#Fawn per Spring vs Adult Antelope Population  
ggplot(df, aes(adultPop, fawnPerSpring)) + geom\_point() +   
 labs(x = 'Adult Antelope Population', y = 'Avg Fawn per Spring') +   
 ggtitle('Adult Antelop Population vs Avg Fawn per Spring')



#Annual Precipitation  
ggplot(df, aes(year, annualPrecipitation)) + geom\_line() +  
 labs(x = 'Year', y = 'Annual Precipitation') +   
 ggtitle('Annual Precipitation', subtitle = 'Years Unknown')



#Winter Severity   
ggplot(df, aes(year, winterSeverity)) + geom\_line() +   
 labs(x = 'Year', y = 'Winter Severity') +  
 ggtitle('Winter Severity', subtitle = 'Years Unknown')



#5. Creating, running, and reviewing linear models.   
#Will first verify if severity of winter can predict number of fawn the following spring.  
fawnVsWinter <- lm(fawnPerSpring ~ winterSeverity, data = df)  
summary(fawnVsWinter)

##   
## Call:  
## lm(formula = fawnPerSpring ~ winterSeverity, 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 \*\*\*  
## winterSeverity -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

#Winter severity has a p-value of 0.036 - significant at the 0.05 level. However, Adjusted R-squared states that the severity of the previous winter explains 47 percent of the variance of the number of fawn.   
  
#How about checking the severity of the winter and size of the adult population?   
fawnVsWinterAdult <- lm(fawnPerSpring ~ winterSeverity + adultPop, data = df)  
summary(fawnVsWinterAdult)

##   
## Call:  
## lm(formula = fawnPerSpring ~ winterSeverity + adultPop, data = df)  
##   
## 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   
## winterSeverity 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

#When we add the adult population, the Adjusted R-Squared explains 84.4 percent of the variance in the number of fawn. However, neither of the explaining variables are significant - severity of winter has a p-value of 0.596 and adult population has a p-value of 0.011.   
  
#Let's add the remaining column - annual precipitation - to see how it affects the model.   
fawnVsAll <- lm(fawnPerSpring ~ . - year, data = df)  
summary(fawnVsAll)

##   
## Call:  
## lm(formula = fawnPerSpring ~ . - year, 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 \*\*  
## adultPop 0.33822 0.09947 3.400 0.0273 \*   
## annualPrecipitation 0.40150 0.10990 3.653 0.0217 \*   
## winterSeverity 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

#By adding annual precipitation, we are able to explain 95.5 percent of the variance in the number of fawn. Additionally, we find that all variables are significant at a p-value of 0.05; which gives us a very good model.   
  
#Let's plot the model to see how the variables relate  
ggplot(df, aes(x = adultPop, y = fawnPerSpring)) +  
 geom\_point(aes(size = annualPrecipitation, color = as.factor(winterSeverity))) +  
 geom\_smooth(method = 'lm', aes(x = adultPop, y = fawnPerSpring)) +  
 labs(x = 'Adult Antelope Population', y = 'Avg Fawn Per Spring',   
 color = 'Winter Severity', size = 'Annual Precipitation') +  
 ggtitle('Avg Fawn Per Spring vs Adult Pop, Winter Severity,\n and Annual Precipitation')

