**Unexecuted Code**

## Homework 8 - 23 Aug 2018

#------------------------------------

#Step 1 - Read in the data

#------------------------------------

library(gdata)

url = 'http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls'

df <- read.xls(url)

col.names <- c("numFawn", "popAnt", "precip", "winter")

colnames(df) <- col.names

#Step 3 - Check the structure of the data

#-----------------------------------

str(df)

#eight observations of four variables

#Columns are numeric

#Step 4: Plot baby fawn vs independent variables

#-----------------------------------

library(ggplot2)

#Fawn vs. Adults

gPop <- ggplot(df, aes(popAnt, numFawn))

gPop <- gPop + geom\_point()

gPop <- gPop + ylab("Number of Fawn") + xlab("Adult Antelope Population")

gPop <- gPop + ggtitle("Adult Population vs. Fawn Born")

gPop <- gPop + theme(plot.title = element\_text(hjust = 0.5))

gPop

#Fawn vs. Rain

gRain <- ggplot(df, aes(precip, numFawn))

gRain <- gRain + geom\_point()

gRain <- gRain + ylab("Number of Fawn") + xlab("Annual Precipitation")

gRain <- gRain + ggtitle("Annual Precipitation vs. Fawn Born")

gRain <- gRain + theme(plot.title = element\_text(hjust = 0.5))

gRain

#Fawn vs Winter

gWinter <- ggplot(df, aes(winter, numFawn))

gWinter <- gWinter + geom\_point()

gWinter <- gWinter + ylab("Number of Fawn") + xlab("Harshness of Winter")

gWinter <- gWinter + ggtitle("Harshness of Winter vs. Fawn Born")

gWinter <- gWinter + theme(plot.title = element\_text(hjust = 0.5))

gWinter

#Step 5: Regression Models

#winter vs fawn model

model1 <- lm(numFawn ~ winter, data = df)

#winter and adult population vs fawn

model2 <- lm(numFawn ~ winter + popAnt, data = df)

#winter, adult population, and precipitation vs fawn

model3 <- lm(numFawn ~ winter + precip + popAnt, data = df)

#Looking at the model summaries

#Model 1

summary(model1)

#adjusted R2 of 0.47

#Model 2

summary(model2)

#adjusted R2 of 0.84

#Model 3

summary(model3)

#adjusted R2 of 0.96

#Which model works best?

#Model 3 is the best model. All variables are significant

#and it has the highest R2 value of 0.955. Meaning that

#the change in fawn born is 95.5% due to the variability

#in the winter, adult population, and precipitation

#Which predictors are statistically significant in each of the models?

#With an alpha of 0.05, the winter variable is significant. In model 2,

#only the population of adult antelope is significant. In model 3, all

#of the variables (num antelope, precipitation, and winter) are statistically

#significant

#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?

#To create the best linear model with the fewest parameters, I would use

#Precipitation as the only parameter. The adult population had teh lowest p-value,

# but it looked like it had a non-linear (exponential) correlation, and therefore

#would not be best for a linear model without some sort of data transformation

**Console log w/plot**

**Executed code**

> ## Homework 8 - 23 Aug 2018

> #------------------------------------

> #Step 1 - Read in the data

> #------------------------------------

>

> library(gdata)

> url = 'http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls'

> df <- read.xls(url)

trying URL 'http://college.cengage.com/mathematics/brase/understandable\_statistics/7e/students/datasets/mlr/excel/mlr01.xls'

Content type 'application/vnd.ms-excel' length 5632 bytes

==================================================

downloaded 5632 bytes

> col.names <- c("numFawn", "popAnt", "precip", "winter")

> colnames(df) <- col.names

>

> #Step 3 - Check the structure of the data

> #-----------------------------------

>

> str(df)

'data.frame': 8 obs. of 4 variables:

$ numFawn: num 2.9 2.4 2 2.3 3.2 ...

$ popAnt : num 9.2 8.7 7.2 8.5 9.6 ...

$ precip : num 13.2 11.5 10.8 12.3 12.6 ...

$ winter : int 2 3 4 2 3 5 1 3

> #eight observations of four variables

> #Columns are numeric

>

> #Step 4: Plot baby fawn vs independent variables

> #-----------------------------------

> library(ggplot2)

> #Fawn vs. Adults

> gPop <- ggplot(df, aes(popAnt, numFawn))

> gPop <- gPop + geom\_point()

> gPop <- gPop + ylab("Number of Fawn") + xlab("Adult Antelope Population")

> gPop <- gPop + ggtitle("Adult Population vs. Fawn Born")

> gPop <- gPop + theme(plot.title = element\_text(hjust = 0.5))

> gPop



>

> #Fawn vs. Rain

> gRain <- ggplot(df, aes(precip, numFawn))

> gRain <- gRain + geom\_point()

> gRain <- gRain + ylab("Number of Fawn") + xlab("Annual Precipitation")

> gRain <- gRain + ggtitle("Annual Precipitation vs. Fawn Born")

> gRain <- gRain + theme(plot.title = element\_text(hjust = 0.5))

> gRain



>

> #Fawn vs Winter

> gWinter <- ggplot(df, aes(winter, numFawn))

> gWinter <- gWinter + geom\_point()

> gWinter <- gWinter + ylab("Number of Fawn") + xlab("Harshness of Winter")

> gWinter <- gWinter + ggtitle("Harshness of Winter vs. Fawn Born")

> gWinter <- gWinter + theme(plot.title = element\_text(hjust = 0.5))

> gWinter



>

> #Step 5: Regression Models

>

> #winter vs fawn model

> model1 <- lm(numFawn ~ winter, data = df)

>

> #winter and adult population vs fawn

> model2 <- lm(numFawn ~ winter + popAnt, data = df)

>

> #winter, adult population, and precipitation vs fawn

> model3 <- lm(numFawn ~ winter + precip + popAnt, data = df)

>

> #Looking at the model summaries

> #Model 1

> summary(model1)

Call:

lm(formula = numFawn ~ winter, 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 \*\*\*

winter -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

> #adjusted R2 of 0.47

>

> #Model 2

> summary(model2)

Call:

lm(formula = numFawn ~ winter + popAnt, 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

winter 0.07058 0.12461 0.566 0.5956

popAnt 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

> #adjusted R2 of 0.84

>

> #Model 3

> summary(model3)

Call:

lm(formula = numFawn ~ winter + precip + popAnt, 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 \*\*

winter 0.26295 0.08514 3.089 0.0366 \*

precip 0.40150 0.10990 3.653 0.0217 \*

popAnt 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

> #adjusted R2 of 0.96

>

> #Which model works best?

>

> #Model 3 is the best model. All variables are significant

> #and it has the highest R2 value of 0.955. Meaning that

> #the change in fawn born is 95.5% due to the variability

> #in the winter, adult population, and precipitation

>

> #Which predictors are statistically significant in each of the models?

>

> #With an alpha of 0.05, the winter variable is significant. In model 2,

> #only the population of adult antelope is significant. In model 3, all

> #of the variables (num antelope, precipitation, and winter) are statistically

> #significant

>

> #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?

>

> #To create the best linear model with the fewest parameters, I would use

> #Precipitation as the only parameter. The adult population had teh lowest p-value,

> # but it looked like it had a non-linear (exponential) correlation, and therefore

> #would not be best for a linear model without some sort of data transformation