# Homework #8 – Making Predictions

## R Script (Code)

#

# Course: IST687

# Name: Joyce Woznica

# Homework 8 - Making Predictions

# Due Date: 03/05/2019

# Date Submitted:

#

# Package Section

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

install.packages("readxl")

library(readxl)

# 1. Read in the data

# 2. Option to save to computer (had to based on bad URL)

# read in a dataset so that it can be useful.

# use this package to read in a XLS file

fawnData<-read\_excel("C:/Users/Joyce/Desktop/Syracuse/IST687/Submissions/Week8/mlr01\_2\_2.xlsx")

fawnDF <- as.data.frame(fawnData)

# rename the columns to make easier:

# X1 = SpringFawns

# X2 = AdultPop

# X3 = AnnualPercip

# X4 = WinterRating

newColNames <- c("SpringFawns", "AdultPop", "AnnualPrecip", "WinterRating")

colnames(fawnDF)<-newColNames

View(fawnDF)

# 3. Confirm the data

# confirming all numbers, correct # of variables and observations

str(fawnDF)

# 4. Create bivariate plots of baby fawns versus:

# a) Adult Population

# b) Annual Percipitation

# c) Severity of Winter

# Independent variables on X axis, Dependent (SpringFawns) on the Y axis

plot(fawnDF$SpringFawns ~ fawnDF$AdultPop, xlab="Adult Antelope Population", ylab="Number of Spring Fawns")

plot(fawnDF$SpringFawns ~ fawnDF$AnnualPrecip, xlab="Annual Percipitation", ylab="Number of Spring Fawns")

plot(fawnDF$SpringFawns ~ fawnDF$WinterRating, xlab="Severity of Winter", ylab="Number of Spring Fawns")

# 5. Create 3 regression models of increasing complexity using lm().

# a) Predict Number of Fawns from Severity of Winter

# b) Predict Number of Fawns from Severity of Winter (and another variable)

# c) Predict Number of Fawns from all 3 variables

oneVarM<-lm(formula=SpringFawns ~ WinterRating, data=fawnDF)

twoVarM<-lm(formula=SpringFawns ~ WinterRating + AdultPop, data=fawnDF)

threeVarM<-lm(formula=SpringFawns ~., data=fawnDF)

# inspect each model and see which is best

oneVarMSum<-summary(oneVarM)

twoVarMSum<-summary(twoVarM)

threeVarMSum<-summary(threeVarM)

ModResidsV<-residuals(threeVarM)

ModCoeffV<-coefficients(threeVarM)

ModelsRSq<-c(oneVarMSum$adj.r.squared, twoVarMSum$adj.r.squared, threeVarMSum$adj.r.squared)

# Which model works best?

# ANSWER: The model with all three variables i best because the Adjusted (multiple) R-squared is .955

# each independent variable shows a p-value below the 0.05 alpha.

AICModels<-step(threeVarM,data=fawnDF, direction="backward")

# the model would include all variables, as we see using the AIC method and to determine the

# optimal selection of independent variables, so the fewest is actually ALL the variables

# Bonus #5 on second page

# I was surprised to lean that the Annual Precipitation had a bearing on the number of fawns

# but then I considered the fact that the vegetation would be affecting and that would have

# an influence

## Console Log (Executed Code)

> #

> # Course: IST687

> # Name: Joyce Woznica

> # Homework 8 - Making Predictions

> # Due Date: 03/05/2019

> # Date Submitted:

> #

> # Package Section

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

> install.packages("readxl")

Error in install.packages : Updating loaded packages

> install.packages("readxl")

Installing package into ‘C:/Users/Joyce/Documents/R/win-library/3.5’

(as ‘lib’ is unspecified)

Warning in install.packages :

package ‘readxl’ is in use and will not be installed

> library(readxl)

> # 1. Read in the data

> # 2. Option to save to computer (had to based on bad URL)

> # read in a dataset so that it can be useful.

> # use this package to read in a XLS file

>

> fawnData<-read\_excel("C:/Users/Joyce/Desktop/Syracuse/IST687/Submissions/Week8/mlr01\_2\_2.xlsx")

> fawnDF <- as.data.frame(fawnData)

> # rename the columns to make easier:

> # X1 = SpringFawns

> # X2 = AdultPop

> # X3 = AnnualPercip

> # X4 = WinterRating

>

> newColNames <- c("SpringFawns", "AdultPop", "AnnualPrecip", "WinterRating")

> colnames(fawnDF)<-newColNames

> View(fawnDF)

> # 3. Confirm the data

> # confirming all numbers, correct # of variables and observations

> str(fawnDF)

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

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

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

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

$ WinterRating: num 2 3 4 2 3 5 1 3

> # 4. Create bivariate plots of baby fawns versus:

> # a) Adult Population

> # b) Annual Percipitation

> # c) Severity of Winter

> # Independent variables on X axis, Dependent (SpringFawns) on the Y axis

> plot(fawnDF$SpringFawns ~ fawnDF$AdultPop, xlab="Adult Antelope Population", ylab="Number of Spring Fawns")

> plot(fawnDF$SpringFawns ~ fawnDF$AnnualPrecip, xlab="Annual Percipitation", ylab="Number of Spring Fawns")

> plot(fawnDF$SpringFawns ~ fawnDF$WinterRating, xlab="Severity of Winter", ylab="Number of Spring Fawns")

> # 5. Create 3 regression models of increasing complexity using lm().

> # a) Predict Number of Fawns from Severity of Winter

> # b) Predict Number of Fawns from Severity of Winter (and another variable)

> # c) Predict Number of Fawns from all 3 variables

> oneVarM<-lm(formula=SpringFawns ~ WinterRating, data=fawnDF)

> twoVarM<-lm(formula=SpringFawns ~ WinterRating + AdultPop, data=fawnDF)

> threeVarM<-lm(formula=SpringFawns ~., data=fawnDF)

> # inspect each model and see which is best

> oneVarMSum<-summary(oneVarM)

> twoVarMSum<-summary(twoVarM)

> threeVarMSum<-summary(threeVarM)

> ModResidsV<-residuals(threeVarM)

> ModCoeffV<-coefficients(threeVarM)

> ModelsRSq<-c(oneVarMSum$adj.r.squared, twoVarMSum$adj.r.squared, threeVarMSum$adj.r.squared)

> oneVarMSum

Call:

lm(formula = SpringFawns ~ WinterRating, data = fawnDF)

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 \*\*\*

WinterRating -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

> twoVarMSum

Call:

lm(formula = SpringFawns ~ WinterRating + AdultPop, data = fawnDF)

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

WinterRating 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

> threeVarMSum

Call:

lm(formula = SpringFawns ~ ., data = fawnDF)

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 \*

AnnualPrecip 0.40150 0.10990 3.653 0.0217 \*

WinterRating 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

> ModelsRSq

[1] 0.4702033 0.8438937 0.9550047

> # Which model works best?

> # ANSWER: The model with all three variables i best because the Adjusted (multiple) R-squared is .955

> # each independent variable shows a p-value below the 0.05 alpha.

> AICModels<-step(threeVarM,data=fawnDF, direction="backward")

Start: AIC=-31.35

SpringFawns ~ AdultPop + AnnualPrecip + WinterRating

Df Sum of Sq RSS AIC

<none> 0.058494 -31.346

- WinterRating 1 0.13950 0.197989 -23.592

- AdultPop 1 0.16907 0.227561 -22.478

- AnnualPrecip 1 0.19518 0.253673 -21.609

> # the model would include all variables, as we see using the AIC method and to determine the

> # optimal selection of independent variables, so the fewest is actually ALL the variables

>

> # Bonus #5 on second page

> # I was surprised to lean that the Annual Precipitation had a bearing on the number of fawns

> # but then I considered the fact that the vegetation would be affecting and that would have

> # an influence

## Visualizations





