Sharat\_Sripada\_HW8.R

ssharat

2020-03-07

#  
# Course: IST-687  
# Name: Sharat Sripada  
# Homework #8  
# Due Date: 3/8/2020  
# Date Submitted: 3/7/2020  
# Topic: Making predictions  
  
library("gdata")

## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##   
## 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

# Step-1: Read the xls  
antelopes <- read.xls("/Users/ssharat/Downloads/mlr01\_2\_2\_2\_2\_2\_2.xlsx")  
summary(antelopes)

## 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

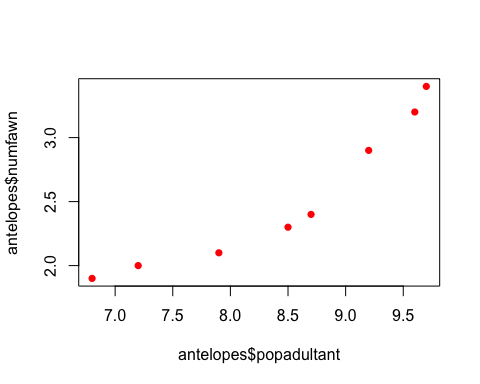
View(antelopes)  
  
# Step-2: Rename the columns:  
# X1 -> numfawn  
# X2 -> popadultant  
# X3 -> anprecip  
# X4 -> wintergrade  
colnames(antelopes) <- c('numfawn', 'popadultant', 'anprecip', 'wintergrade')  
colnames(antelopes)

## [1] "numfawn" "popadultant" "anprecip" "wintergrade"

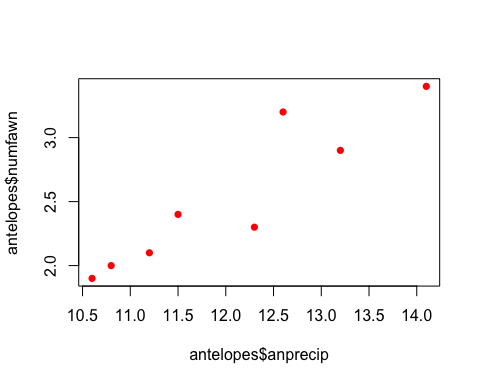
# Step-3: str()  
str(antelopes)

## 'data.frame': 8 obs. of 4 variables:  
## $ numfawn : num 2.9 2.4 2 2.3 3.2 ...  
## $ popadultant: num 9.2 8.7 7.2 8.5 9.6 ...  
## $ anprecip : num 13.2 11.5 10.8 12.3 12.6 ...  
## $ wintergrade: int 2 3 4 2 3 5 1 3

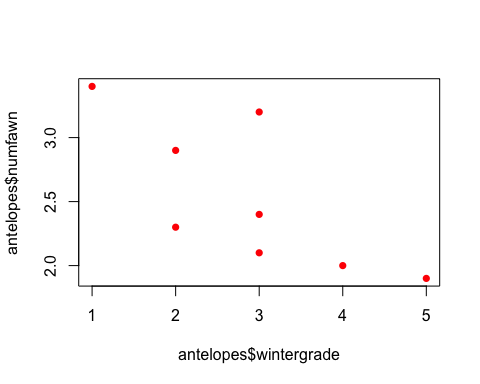
# Step-4: Create bivariate plots  
# baby fawns vs adult antelope population  
# pch: point character - 15: square, 16: circle etc  
# col: color  
plot(antelopes$popadultant, antelopes$numfawn, pch=16, col='red')



# baby fawns versus precipitation  
plot(antelopes$anprecip, antelopes$numfawn, pch=16, col='red')



# baby fawns versus severity of winter  
plot(antelopes$wintergrade, antelopes$numfawn, pch=16, col='red')



# Step-5: Create 3 regression models  
# Model-1: predict the number of fawns from the severity of the winter  
model1 <- lm(formula=numfawn ~ wintergrade, data=antelopes)  
summary(model1)

##   
## Call:  
## lm(formula = numfawn ~ wintergrade, data = antelopes)  
##   
## 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 \*\*\*  
## wintergrade -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

summary(model1)$r.squared

## [1] 0.5458886

# Summary for model1:  
# Co-efficients:  
# wintergrade P-val: 0.0362 < 0.05  
# R-square: 0.5459 (which shows not very strong correlation)  
# numfawn(Y) = -0.3379 \* wintergrade(X) + 3.4966  
test <- data.frame(wintergrade=2)  
predict(model1, test, type="response")

## 1   
## 2.82069

# Prediction: 2.82 (actual-data: 2.9, 2.3)  
  
# Model-2: predict the number of fawns from the severity of the winter and precipitation  
model2 <- lm(formula=numfawn ~ wintergrade + anprecip, data=antelopes)  
summary(model2)

##   
## Call:  
## lm(formula = numfawn ~ wintergrade + anprecip, data = antelopes)  
##   
## 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 \*   
## wintergrade 0.2269 0.1490 1.522 0.18842   
## anprecip 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

summary(model2)$r.squared

## [1] 0.8999734

# Summary for model2:  
# Co-efficients:  
# wintergrade P-val: 0.18843 > 0.1  
# anprecip P-val: 0.00843 < 0.01  
# R-square: 0.9 (which shows strong correlation)  
test <- data.frame(wintergrade=2, anprecip=13.2)  
predict(model2, test, type="response")

## 1   
## 3.065458

# Prediction: 3.06 (actual-data: 2.9)  
  
# Model-3: predict the number of fawns from the severity of the winter, precipitation, adult population  
model3 <- lm(formula=numfawn ~ wintergrade + anprecip + popadultant, data=antelopes)  
summary(model3)

##   
## Call:  
## lm(formula = numfawn ~ wintergrade + anprecip + popadultant,   
## data = antelopes)  
##   
## 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 \*\*  
## wintergrade 0.26295 0.08514 3.089 0.0366 \*   
## anprecip 0.40150 0.10990 3.653 0.0217 \*   
## popadultant 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

summary(model3)$r.squared

## [1] 0.9742884

# Summary for model3:  
# Co-efficients:  
# wintergrade P-val: 0.0466 < 0.05  
# anprecip P-val: 0.0217 < 0.05  
# popadultant P-val: 0.0273 < 0.05  
# R-square: 0.973 (which shows very strong correlation)  
test <- data.frame(wintergrade=2, anprecip=13.2, popadultant=9.2)  
predict(model1, test, type="response")

## 1   
## 2.82069

# Prediction: 2.82 (actual-data: 2.9)  
  
# So, the best model here model-3 - theoretical & prediction is very close.   
  
# Step-5: Parsimonious model using the step() function  
model <- lm(formula=numfawn ~ ., data=antelopes)  
step(model, data=antelopes, direction="backward")

## Start: AIC=-31.35  
## numfawn ~ popadultant + anprecip + wintergrade  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.058494 -31.346  
## - wintergrade 1 0.13950 0.197989 -23.592  
## - popadultant 1 0.16907 0.227561 -22.478  
## - anprecip 1 0.19518 0.253673 -21.609

##   
## Call:  
## lm(formula = numfawn ~ popadultant + anprecip + wintergrade,   
## data = antelopes)  
##   
## Coefficients:  
## (Intercept) popadultant anprecip wintergrade   
## -5.9220 0.3382 0.4015 0.2629

# The step() function showed a single iteration   
# with all three variables - wintergrade, popadultant, anprecip.