IST687 - Making Predictions

John Fields

5/19/2019

Assignment

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 data set 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:

Steps 1,2,3

1. Read in data from the following URL: http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/excel/mlr01.xls This URL will enable you to download the dataset into excel.

The more general web site 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 that 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.

2. You have the option of saving the file save this file to your computer and read it into R, or reading the data directly from the web into a data frame.
3. You should inspect the data using the str() command to make sure that all of the cases have been read in (n=8 years of observations) and that there are four variables.

```r
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
```

```
antelope.url <- "http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/data
antelope <- read.xls(antelope.url)
str(antelope)
```

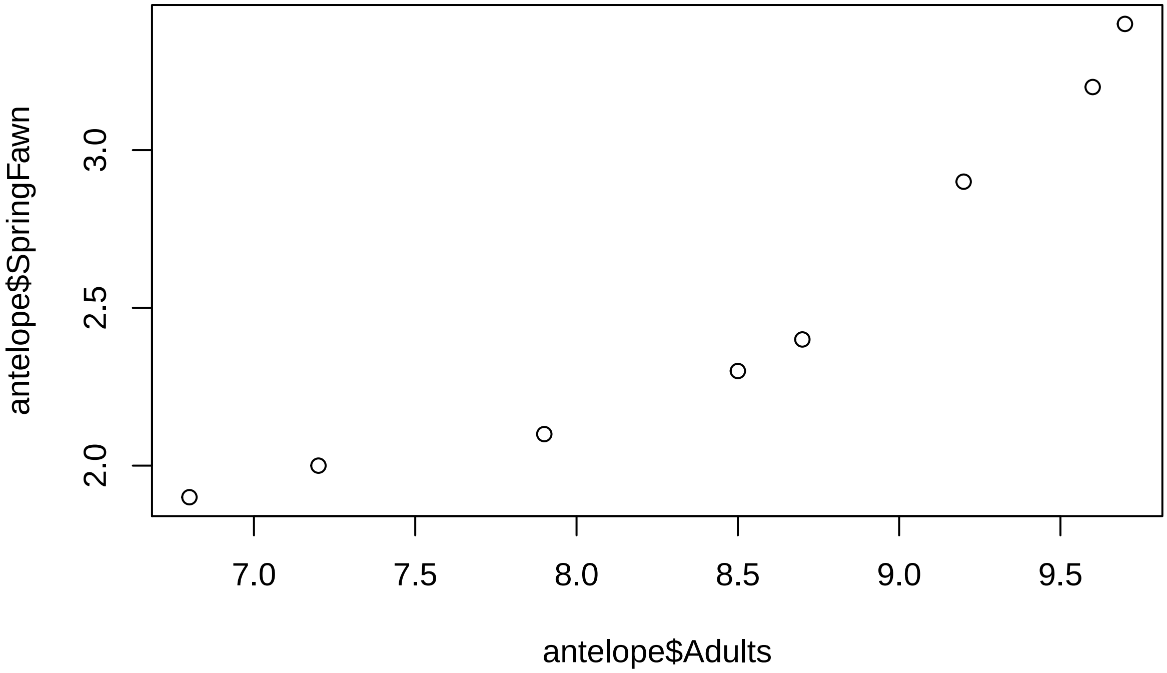
```
## 'data.frame':    8 obs. of  4 variables:
##  $ X1: num  2.9 2.4 2 2.3 3.2 ...
##  $ X2: num  9.2 8.7 7.2 8.5 9.6 ...
##  $ X3: num  13.2 11.5 10.8 12.3 12.6 ...
##  $ X4: int  2 3 4 2 3 5 1 3
```

```
#rename the columns
colnames(antelope)<- c("SpringFawn","Adults","Precipitation","WinterSeverity")
View(antelope)
```

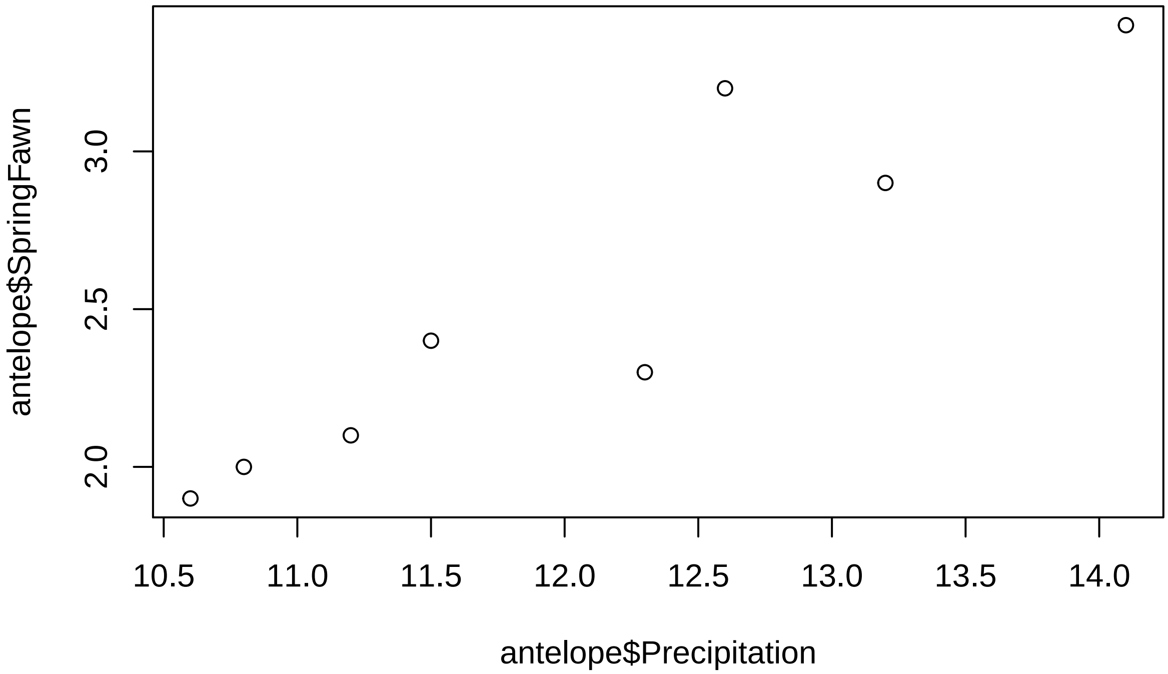
Step 4

4. Create bivariate plots of number of baby fawns versus adult antelope population, the precipitation that year, and the 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?

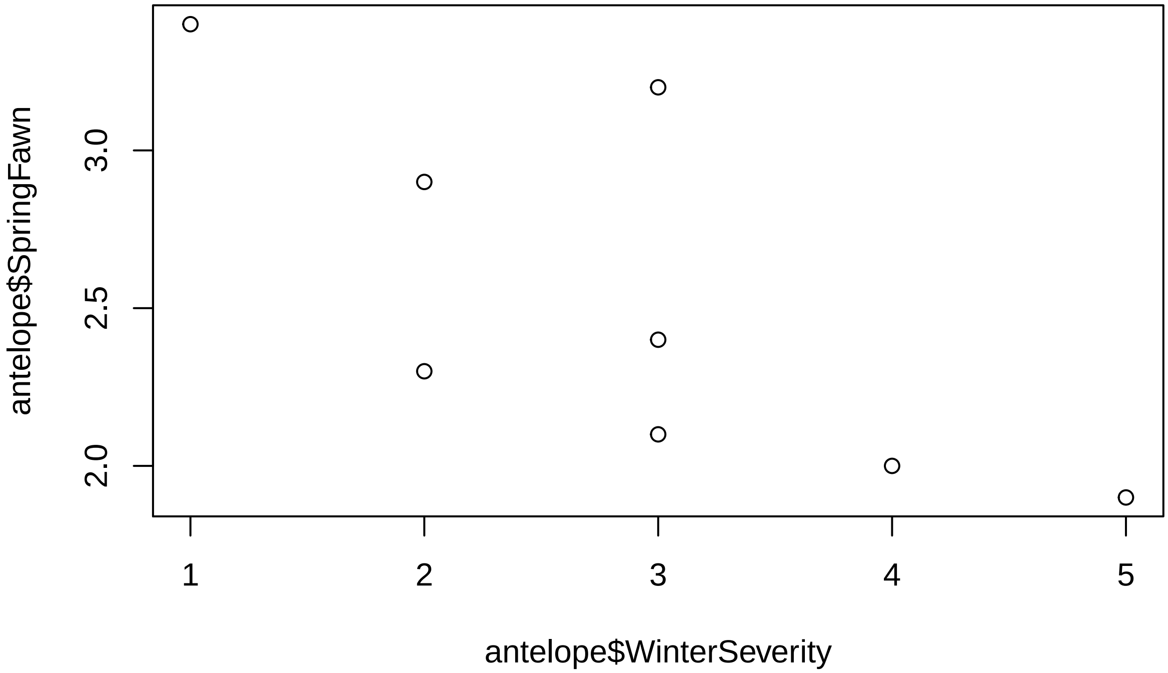
```
plot(antelope$Adults,antelope$SpringFawn)
```



```
plot(antelope$Precipitation,antelope$SpringFawn)
```



```
plot(antelope$WinterSeverity,antelope$SpringFawn)
```



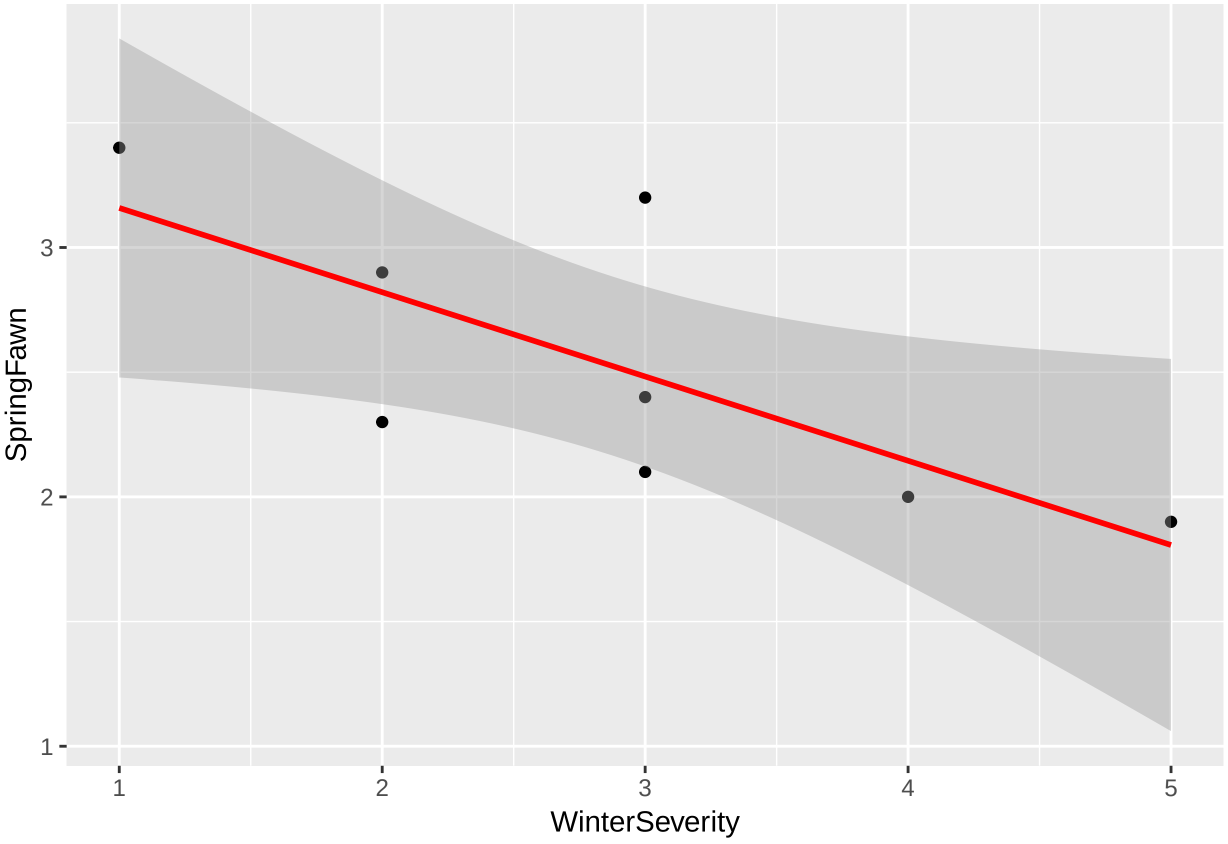
Step 5 Regression Models

5. Next, create three regression models of increasing complexity using lm(). In the first model, predict the number of fawns from the severity of the winter. In the second model, predict the number of fawns from two variables (one should be the severity of the winter). In the third model predict the number of fawns from the three other variables.

```r
#model 1 with winter severity
model1 <- lm(formula=SpringFawn ~ WinterSeverity, data=antelope)
summary(model1)
```

```
##
## Call:
## lm(formula = SpringFawn ~ WinterSeverity, data = antelope)
##
## 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
```

```r
library(ggplot2)
ggplot(antelope,aes(x=WinterSeverity,y=SpringFawn)) + geom_point() + stat_smooth(method="lm",col="red")
```



```r
#model 2 with winter severity and precipitation
model2 <- lm(formula=SpringFawn ~ WinterSeverity + Precipitation, data=antelope)
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = SpringFawn ~ WinterSeverity + Precipitation, data = antelope)
##
## Residuals:
##         1         2         3         4         5         6         7
## -0.165458  0.188313  0.006417 -0.193358  0.289080 -0.193312 -0.010695
##         8
##  0.079013
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.7791     2.2139  -2.610  0.04765 *
## WinterSeverity   0.2269     0.1490   1.522  0.18842
## Precipitation    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
```

```r
#model 3 with adutls, precipitation and winter severity
model3 <- lm(formula=SpringFawn ~ Adults + Precipitation + WinterSeverity, data=antelope)
summary(model3)
```

```
##
## Call:
## lm(formula = SpringFawn ~ Adults + Precipitation + WinterSeverity,
##     data = antelope)
##
## 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 **
## Adults          0.33822    0.09947   3.400   0.0273 *
## Precipitation   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
```

```r
# model 4 fawn predictions with different precipitation levels
model4 <- lm(formula=SpringFawn ~ Precipitation, data=antelope)
summary(model4)
```

```
##
```

```
## Call:
## lm(formula = SpringFawn ~ Precipitation, data = antelope)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33747 -0.08040 -0.00889  0.03023  0.43399
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.63251    0.87591  -3.005  0.02384 *
## Precipitation  0.42845    0.07244   5.915  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2356 on 6 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8292
## F-statistic: 34.99 on 1 and 6 DF,  p-value: 0.001039
```

```
test=data.frame(Precipitation=9)
predict(model4,test,type="response")
```

```
##        1
## 1.223573
```

```
test=data.frame(Precipitation=16)
predict(model4,test,type="response")
```

```
##        1
## 4.222746
```

Step 5 Questions

Which model works best?

The model that includes all variables (adults, precipitation, winter severity) is the best model since it has an adjusted R-squaured of .955. This indicates that the independent variables account for 95.5% of the dependent variable (spring fawns).

Which of the predictors are statistically significant in each model?

Model 1 - winter severity has a P-value of .036 Model 2 - winter severity has a P-value of .189 and precipitation .001. Model 3 - winter severity has a P-value of .037, precipitation .022 and adults .027

All of these are below the typical alpha value of .05 except for winter severity in Model 2. However, precipitation is the most significant (lowest P-value) in the models where it is included.

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?

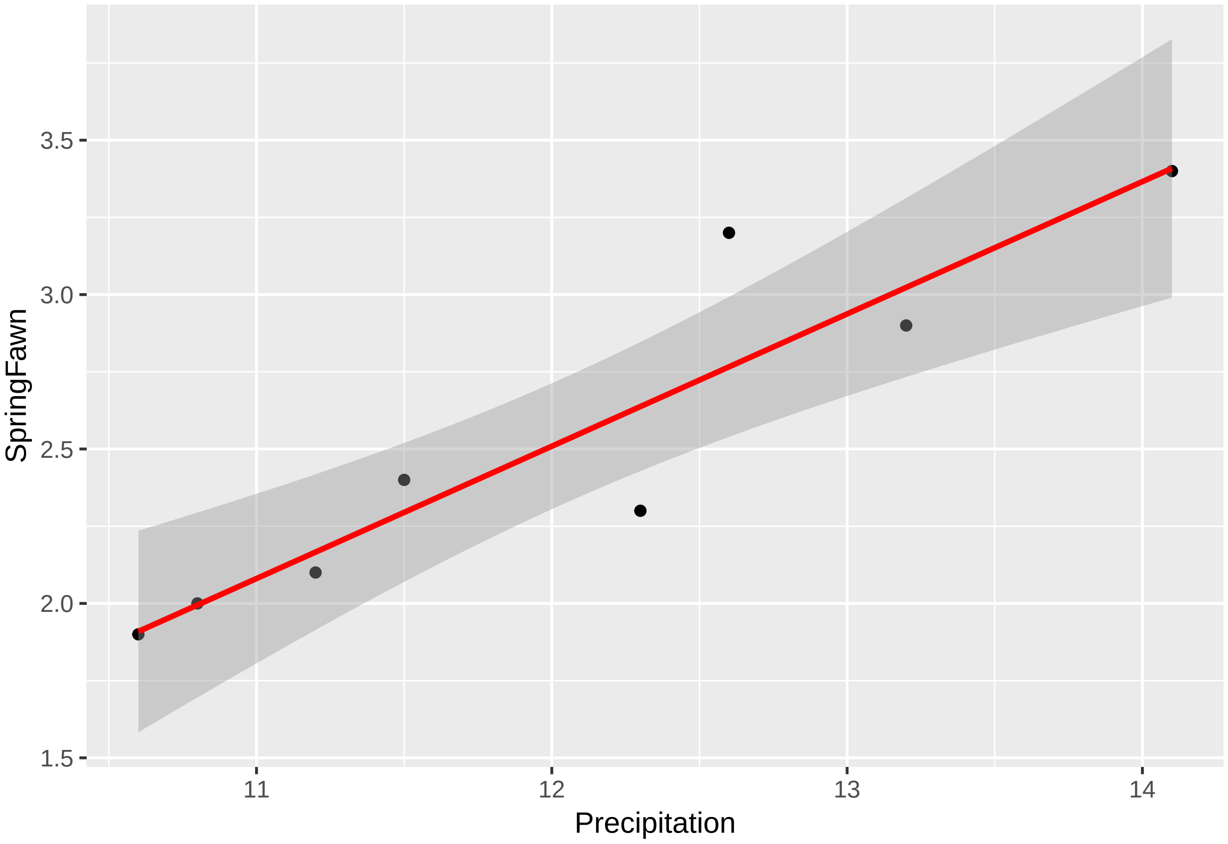
Model 4 (below) which only uses the precipitation variable would be the most parsimonious. Using only this variable, the adjusted R-squared is .829 which indicates this one variable explains 82.9% of the variation in spring fawn births.

Note: the additional models below were developed to look at the impact of precipitation on spring fawns during possible drought and flood conditions.

```
# fawn predictions with different precipitation levels
model4 <- lm(formula=SpringFawn ~ Precipitation, data=antelope)
summary(model4)
```

```
##
## Call:
## lm(formula = SpringFawn ~ Precipitation, data = antelope)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33747 -0.08040 -0.00889  0.03023  0.43399
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.63251    0.87591  -3.005  0.02384 *
## Precipitation  0.42845    0.07244   5.915  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2356 on 6 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8292
## F-statistic: 34.99 on 1 and 6 DF,  p-value: 0.001039
```

```r
ggplot(antelope,aes(x=Precipitation,y=SpringFawn)) + geom_point() + stat_smooth(method="lm",col="red")
```



```r
test=data.frame(Precipitation=9)
predict(model4,test,type="response")
```

```
##        1
## 1.223573
```

```r
test=data.frame(Precipitation=16)
predict(model4,test,type="response")
```

```
##        1
## 4.222746
```