IST 687
Homework 8
Due Date: 11/30

Code requires the following packages to run

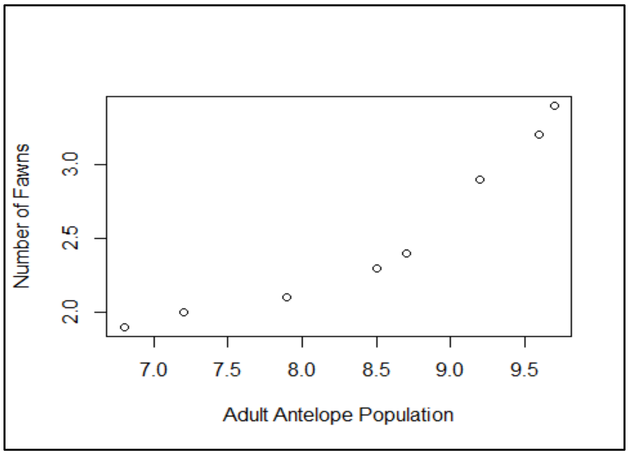
```
library(gdata) #install.packages('gdata')
library(readxl) #install.pacakges('readxl')
```

Read in data from the following URL

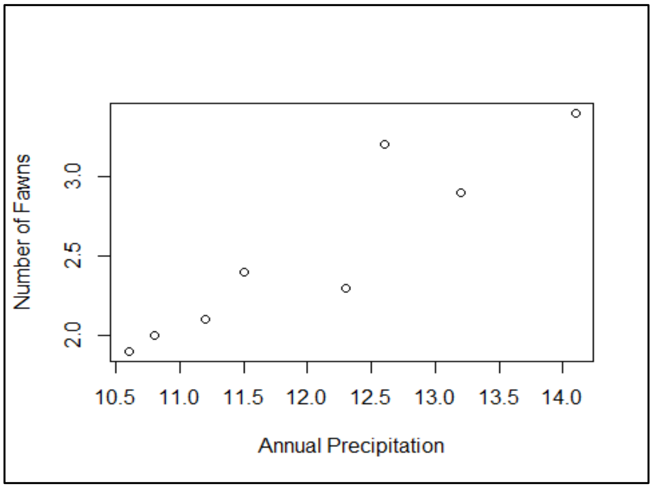
```
#read in dataset (saved local file)
raw_data_url <- "http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/excel/mlr01.xls"
raw_data_temp <- tempfile()
download.file(raw_data_url, raw_data_temp, mode="wb")
raw_data <- read_excel(path=raw_data_temp)
str(raw_data)

## tibble [8 x 4] (S3: tbl_df/tbl/data.frame)
##  $ X1: num [1:8] 2.9 2.4 2 2.3 3.2 ...
##  $ X2: num [1:8] 9.2 8.7 7.2 8.5 9.6 ...
##  $ X3: num [1:8] 13.2 11.5 10.8 12.3 12.6 ...
##  $ X4: num [1:8] 2 3 4 2 3 5 1 3

colnames(raw_data) <- c("BabyFawn", "AdultAntelope", "Precipitation", "Winter")
```

Create bivariate plots

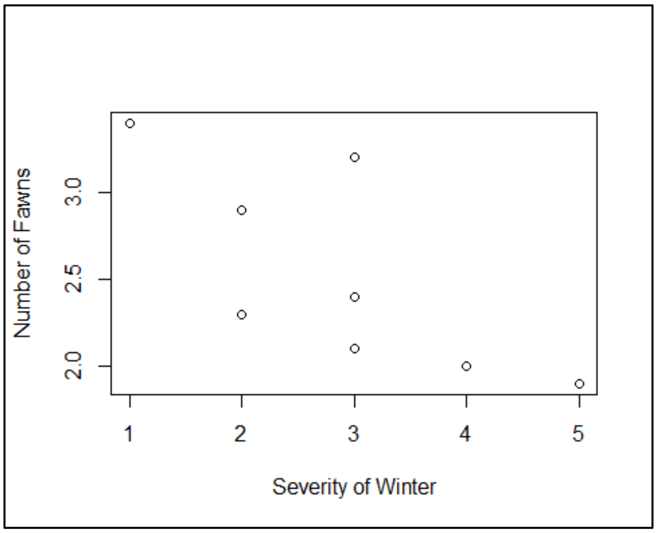
```
#number of baby fawns versus adult antelope population
plot(raw_data$AdultAntelope, raw_data$BabyFawn, xlab="Adult Antelope Population", ylab = "Number of Fawns")
```



```r
#number of baby fawns versus annual precipitation
plot(raw_data$Precipitation, raw_data$BabyFawn, xlab="Annual Precipitation", ylab = "Number of Fawns")
```



```r
#number of baby fawns versus severity of winter
plot(raw_data$Winter, raw_data$BabyFawn, xlab="Severity of Winter", ylab = "Number of Fawns")
```



first regression model

```
model1 <- lm(formula=BabyFawn ~ Winter, data=raw_data)

summary(model1)

##
## Call:
## lm(formula = BabyFawn ~ Winter, data = raw_data)
##
## 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
```

Second regression model

```
model2 <- lm(formula=BabyFawn ~ Winter + Precipitation, data=raw_data)
summary(model2)

##
## Call:
## lm(formula = BabyFawn ~ Winter + Precipitation, data = raw_data)
##
## 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 *
## Winter          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
```

Third regression model

```
model3 <- lm(formula=BabyFawn ~ Winter + Precipitation + AdultAntelope, data=raw_data)
summary(model3)

##
## Call:
## lm(formula = BabyFawn ~ Winter + Precipitation + AdultAntelope,
##     data = raw_data)
##
## 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 *
## Precipitation  0.40150    0.10990   3.653   0.0217 *
## AdultAntelope  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
```

Which model works the best?

The third model works the best at predicting the baby fawns because it has the last amount of residual standard error. It can also be noted that it has the highest R2 value at .9743 which says that there is a very strong correlation between the independent variables (the combination of severity of winter, annual precipitation, and adult antelope population) and the dependent variable (number of baby fawns).

Which of the predictors are statistically significant in each model?

The benchmark for statistical significance is .05 in other words anything with a Pr(>|t|) of less than .05 is considered to be statistically significant. In this case, all of the predictors are statistically significant in each model because they are less than .05.

What would the most parsimonious model contain?

I would say that the second model because it only uses 2 independent variables but produces a result that is only off by about .1 more fawns that the third model that uses 3 independent variables. The second model uses the severity of winter and annual precipitation to predict the number of baby fawns. I also think this could be better because estimating the population of adult antelope is based on statistical inference itself (noway to know exactly the population) whereas the winter and annual precipitation are more objective measures.