Antelope Population Linear Regression

Tesslyn Knapp

2021-10-04 09:55:43

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:
  - 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.
- 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.

Learning Goals for this activity:

A. Develop skills for manipulating and transforming data that contains missing values.
B. Understand the application of multiple linear regression to simple situations of predicting one numeric variable from one or more other numeric variables.
C. Practice plotting skills.
D. Build debugging skills.
E. Increase familiarity with bringing external data sets into R.
F. Increase familiarity with sources of advice and ideas on R source code.

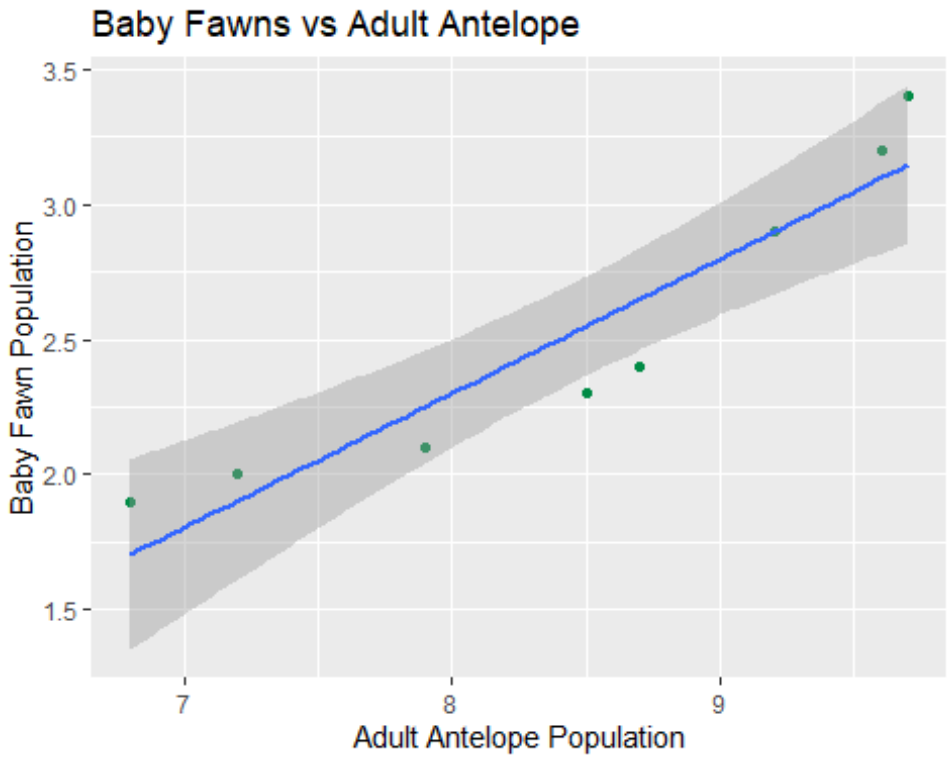
```
# Read dataset into R
fawn.df <- read_excel("C:\\Users\\Tesslyn
Knapp\\Documents\\Syracuse\\Q2\\IST_687\\Week_8\\antelope.xls")
```

- 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.

```
# Inspect dataset using str() command
str(fawn.df) # Looks like we have 8 observations of 4 variables

## tibble[,4] [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

# Rename columns
colnames(fawn.df) <- c("fawns", "adults", "precipitation", "winter")
```

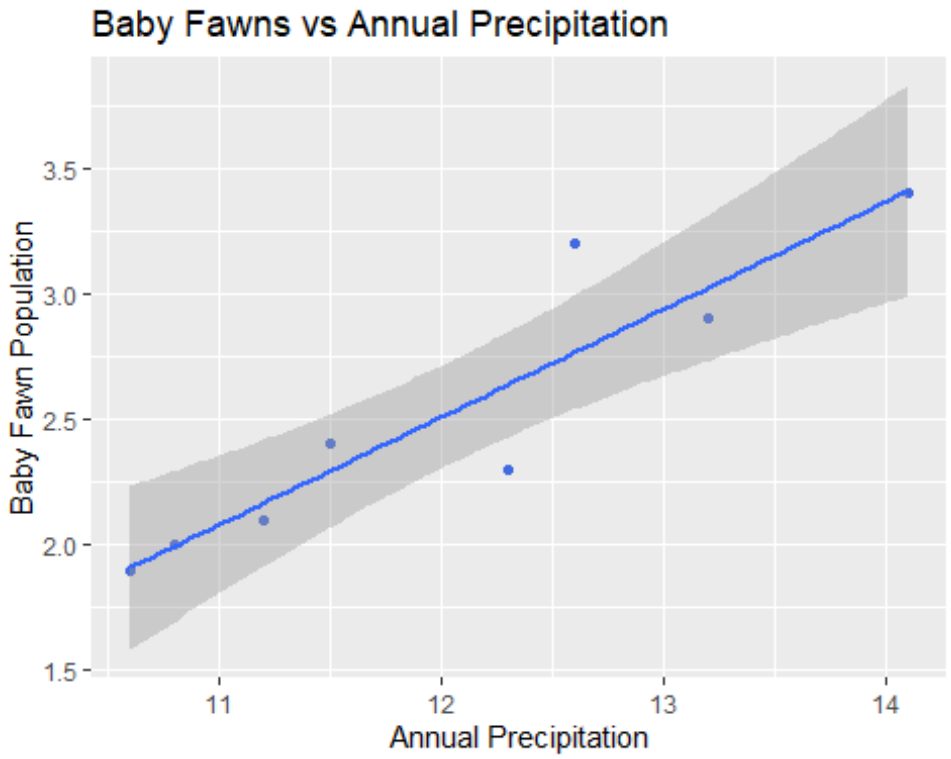
- Create bivariate plots of number of baby fawns versus adult antelope population, the precipitation that year, and the severity of the winter. Your code should produce three separate plots. Make sure the Y-axis and X-axis are labeled. Keeping in mind that the number of fawns is the outcome (or dependent) variable, which axis should it go on in your plots?

```
# Fawns are the dependent variable, so they should go on the Y-axis
# Baby Fawns vs Adult Antelope
ggplot(fawn.df, aes(x=adults, y=fawns)) + geom_point(color="springgreen4") +
geom_smooth(method = "lm") + ggtitle("Baby Fawns vs Adult Antelope") +
xlab("Adult Antelope Population") + ylab("Baby Fawn Population")

## `geom_smooth()` using formula 'y ~ x'
```

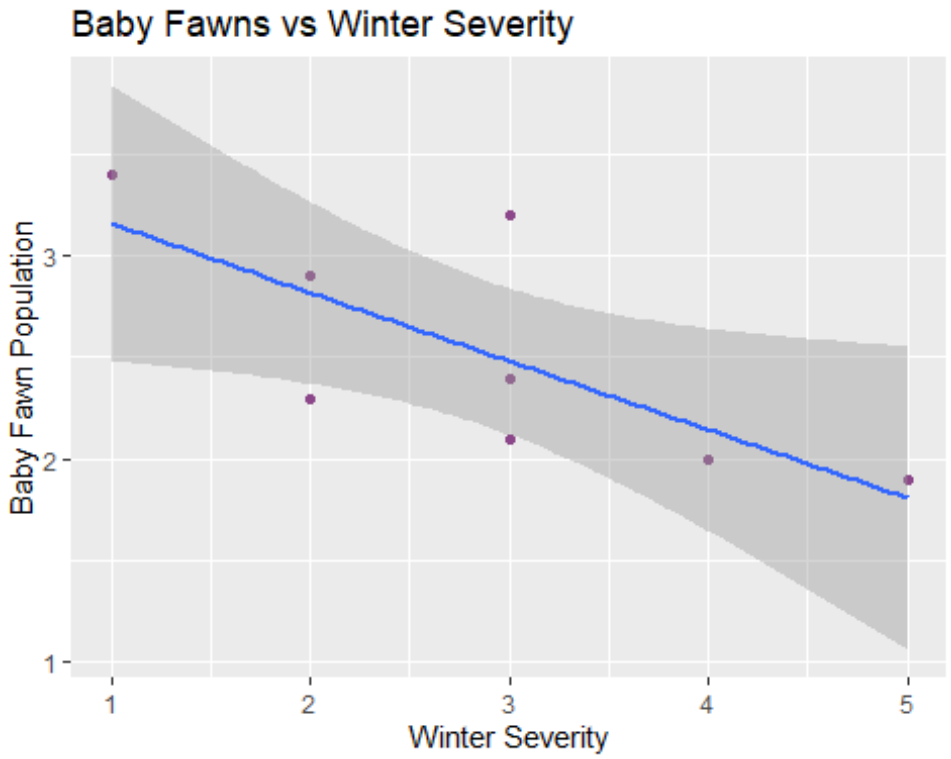



```
# Baby Fawns vs Annual Precipitation
ggplot(fawn.df, aes(x=precipitation, y=fawns)) +
geom_point(color="royalblue") + geom_smooth(method = "lm") + ggtitle("Baby
Fawns vs Annual Precipitation") + xlab("Annual Precipitation") + ylab("Baby
Fawn Population")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Baby Fawns vs Winter Severity
ggplot(fawn.df, aes(x=winter, y=fawns)) + geom_point(color="orchid4") +
geom_smooth(method = "lm") + ggtitle("Baby Fawns vs Winter Severity") +
xlab("Winter Severity") + ylab("Baby Fawn Population")

## `geom_smooth()` using formula 'y ~ x'
```



- 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. Which model works best? Which of the predictors are statistically significant in each model? 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?

```
# Create a regression model to predict the number of fawns from the severity
of winter
summary(lm(formula = fawns ~ winter, data = fawn.df))

##
## Call:
## lm(formula = fawns ~ winter, data = fawn.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
```

- R-squared: 0.5459
- Winter severity is statistically significant (< 0.05): p-value = 0.036

```
# Create a regression model to predict the number of fawns from the severity
of winter and precipitation
summary(lm(formula = fawns ~ winter + precipitation, data = fawn.df))

##
## Call:
## lm(formula = fawns ~ winter + precipitation, data = fawn.df)
##
## 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
```

- Adjusted R-squared: 0.86
- Winter severity is not statistically significant: p-value = 0.188
- Annual precipitation is statistically significant: p-value = 0.008

```
# Create a regression model to predict the number of fawns from the severity
of winter, precipitation, and adult antelope
summary(lm(formula = fawns ~ winter + precipitation + adults, data =
fawn.df))

##
## Call:
## lm(formula = fawns ~ winter + precipitation + adults, data = fawn.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 *
## precipitation  0.40150    0.10990   3.653   0.0217 *
## adults         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 R-squared: 0.955
- Winter severity is statistically significant: p-value = 0.0366
- Annual precipitation is statistically significant: p-value = 0.0217
- Adult antelope population is statistically significant: p-value = 0.0273

The third model works best because we achieve an Adjusted R-squared of 0.955, which is the closest to 1.0 out of all of the models.

```
summary(lm(formula = fawns ~ precipitation, data = fawn.df))

##
## Call:
## lm(formula = fawns ~ precipitation, data = fawn.df)
##
## 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

summary(lm(formula = fawns ~ adults, data = fawn.df))

##
## Call:
## lm(formula = fawns ~ adults, data = fawn.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24988 -0.17586  0.04938  0.12611  0.25309
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.67914    0.63422  -2.648 0.038152 *
## adults       0.49753    0.07453   6.676 0.000547 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2121 on 6 degrees of freedom
## Multiple R-squared:  0.8813, Adjusted R-squared:  0.8616
## F-statistic: 44.56 on 1 and 6 DF,  p-value: 0.0005471
```

In order to create the most parsimonious model, it would contain two variables: precipitation and adults because these variables are more statistically significant on their own (p-values of 0.001 and 0.0005, respectively) and generate a higher R-squared value individually (0.8536 and 0.8813, respectively).

```
summary(lm(formula = fawns ~ precipitation + adults, data = fawn.df))

##
## Call:
## lm(formula = fawns ~ precipitation + adults, data = fawn.df)
##
## Residuals:
##       1        2        3        4        5        6        7        8
## -0.07265 -0.09701  0.08698 -0.29029  0.22233  0.14526  0.10497 -0.09960
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.3155     0.7595  -3.049   0.0285 *
## precipitation   0.1916     0.1421   1.348   0.2355
## adults          0.2999     0.1624   1.847   0.1241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.199 on 5 degrees of freedom
## Multiple R-squared:  0.913,  Adjusted R-squared:  0.8782
## F-statistic: 26.23 on 2 and 5 DF,  p-value: 0.002234
```

According to this model, we generate an adjusted R-squared of 0.8782 and a p-value < 0.05 of 0.002, which means this is a strong model. However, both of the variables of annual precipitation and adult antelope population size have p-values greater than 0.05, meaning that we cannot confirm statistical significance that they are strong predictors of baby fawn population size.