ES/STT 7140: Homework 4

Brandon M. Greenwell

Wedesday, March 7, 2018 (due Wednesday, March 14, 2018)

Question 1 (model building) The data set `birds.csv` contains measures on breeding pairs of land bird species collected from 16 islands around Britain over the course of several decades reported in Pimm et al. (1988). For each species (`SPECIES`), the data set contains an average time of extinction (`TIME`) on those islands where it appeared; the average number of nesting pairs (`PAIRS`)—the average over all islands where the birds appeared, of the nesting pairs per year; the size of the species (`SIZE`)—categorized as large (`L`) or small (`S`); and migratory status (`STATUS`)—a factor with levels `R` (resident) and `M` (migrant). It is expected that species with larger numbers of nesting pairs will tend to remain longer before becoming extinct. Pimm et. al. were interested in whether, after accounting for the number of nesting pairs, size or migratory status has any effect (i.e., an ANCOVA). Furthermore, they were also interested in whether the effect of bird size differs depending on the number of nesting pairs. If any species have unusually small or large extinction times compared to other species with similar values of the predictor variables, it would be useful to point them out. Develop a regression model to predict the time to extinction using the number of nesting pairs and the two categorical variables as predictors. You are free to build any linear model, but please explain how you reached your final model (e.g., residual diagnostics, hypothesis testing, etc.). The following code chunk can be used to read in the data. Be sure to include any relevant plots in your write-up. How well does your model seem to fit the data? Does there appear to be any significant interaction(s)?

```
url <- paste0("https://raw.githubusercontent.com/bgreenwell/",
              "eesR/master/R/Data/birds.csv")
birds <- read.csv(url, header = TRUE)

# Print first few observations
head(birds)

# Scatterplot matrix
pairs(birds[, -1L])  # remove SPECIES
```

Question 2 (multicollinearity) The `bodyfat` data frame, defined below, contains a portion of the data for a study of the relation of amount of body fat to several possible predictors, based on a sample of $n = 20$ healthy females 25–34 years old. The possible predictors are tricep skinfold thickness (`triceps`), thigh circumference (`thigh`), and midarm circumference (`midarm`). The amount of body fat (`bodyfat`) for each of the 20 subjects was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictors could provide reliable estimates of the amount of body fat since the measurements needed for the predictors are easy to obtain.

(a) Install the R packages `car` and `corrplot` for this problem. This can be accomplished by running the following two lines of code:

```
# Install required packages
install.packages(c("car", "corrplot"))
```

(b) Starting with the code chunk below, fit an additive model by regressing `bodyfat` on `triceps`, `thigh` and `midarm`. DO any of the predictors seem statistically significant at the $\alpha = 0.05$ level?

```
# Load required packages
library(car)       # for vif() function
library(corrplot)  # for corrplot() function

# Boyfat data
```

```r
bodyfat <- data.frame(
  triceps = c(19.5, 24.7, 30.7, 29.8, 19.1, 25.6, 31.4, 27.9, 22.1, 25.5, 31.1,
              30.4, 18.7, 19.7, 14.6, 29.5, 27.7, 30.2, 22.7, 25.2),
  thigh   = c(43.1, 49.8, 51.9, 54.3, 42.2, 53.9, 58.5, 52.1, 49.9, 53.5, 56.6,
              56.7, 46.5, 44.2, 42.7, 54.4, 55.3, 58.6, 48.2, 51),
  midarm  = c(29.1, 28.2, 37, 31.1, 30.9, 23.7, 27.6, 30.6, 23.2, 24.8, 30,
              28.3, 23, 28.6, 21.3, 30.1, 25.7, 24.6, 27.1, 27.5),
  bodyfat = c(11.9, 22.8, 18.7, 20.1, 12.9, 21.7, 27.1, 25.4, 21.3, 19.3, 25.4,
              27.2, 11.7, 17.8, 12.8, 23.9, 22.6, 25.4, 14.8, 21.1)
)
```

(c) Use the following code to compute the correlation matrix for the data and plot the results. Does there seem to be high correlation between the response and the predictors? If so, how does that explain the results from the previous regression model?

```r
# Correlation matrix
(cor_mat <- cor(bodyfat))

# Visualizing the correlation matrix
corrplot(cor_mat)
```

(d) Use the `vif()` function from the `car` package on your fitted model to obtain the variance inflation factors (VIFs) and interpret the results. Does multicollinearity appear to be a problem?

(e) Obtain the VIF for `midarm` by hand by fitting a separate model regressing `midarm` on the other two predictors (show your work). What seems strange about this regression model when compared to the previously obtain correlation matrix and scatterplot matrix? Explain what you think this means.