Assignment 3 - Diagnostics

your name goes here

Due: Wednesday, February 7, 2018, noon, to Sakai

Summary

The tasks in this homework assignment focus on residual plots, transformations, and interpreting coefficients after transforming variables.

Assignment

1. Distinguish between
   (a) Residual and semistudentized residual
   (b) $E(\epsilon_i) = 0$ and $\overline{e} = 0$
   (c) error term (ϵ_i) and residual (e_i)
2. Prepare a prototype residual plot (by hand is fine – you can take a picture of your plot and include the image in your assignment) for each of the following causes:
   (a) error variance decreases with X
   (b) true regression function is U shaped, but a linear regression is fitted
3. Consider the following data describing the time spent at sources of pollen (in seconds) and the proportions of pollen removed by bumblebee queens and honeybee workers pollinating a species of lily. (Data from "Evolutionary Options for Maximizing Pollen Dispersal of Animal-pollinated Plants.") Consider only the `QUEEN` bees, and regress the amount of pollen removed (`REMOVED`) on the length of time the bee was at the flower (`DURATION`).

```r
pollen <- read.csv("http://pages.pomona.edu/~jsh04747/courses/math158/pollen.csv",
                   header=TRUE)

pollen <- pollen %>%
  filter(BEE=="QUEEN")
```

   (a) What problems are evident in the residual plot?
   (b) Do log transformations of Y or X help any?
   (c) Try fitting the regression only for those times less than 31 seconds (i.e., excluding the two longest times). Does this fit better?
   (d) Is it acceptable to run a linear regression on the subset of variables suggested above? Explain.
4. Consider the folowing data on life expectancy for a sample of countries, along with the number of TVs per person and doctors per person in thsoe countries.

```
lifeexp <- read.csv("http://pages.pomona.edu/~jsh04747/courses/math158/Doctors_and_Life_Expectancy.
                    header=TRUE)
names(lifeexp) <- c("Country", "LE", "TV", "doctors")
```

(a) Plot life expectancy vs number of TVs. What problems do you see in fitting the normal errors model?

(b) Plot standardized residuals (`.std.resid` from `augment`) vs fitted (`.fitted` from `augment`) values, what problem do you see? Are they they same problems you saw previously?

(c) What transformation(s) seem appropriate for the data? Explain.

(d) Transform the data, refit the model, and replot the standardized residuals. Does the model seem more appropriate now? Explain.

5. Consider the following data on penguin heart rate as a function of duration of dive (in minutes).

```
Penguins <- read_csv("http://www.amherst.edu/~nhorton/sdm4/data/Penguins.csv")
```

(a) Plot Y vs. X. What problems do you see in fitting the normal errors model?

(b) Plot standardized residuals vs fitted values, what problem do you see? Are they they same problems you saw previously?

(c) What transformation(s) seem appropriate for the data? Explain.

(d) Transform the data, refit the model, and replot the standardized residuals. Does the model seem more appropriate now? Explain.

6. A student fit a linear regression function for a class assignment. The student plotted the residuals e_i against the Y_i and found a positive relationship. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relationship. How could this difference arise? Which is the more meaningful plot?

7. Consider a regression model describing the relationship between the area of an island and the number of animal and plant species living on the island:

$$\ln(\widehat{species})|X = 1.94 + 0.25 \ln(area)$$

(a) What can you say about the species count comparing two islands, one of which is half the area of the other?

(b) For what factor of area sizes would we expect the median number of species to double?