Exercises 3 · Nonlinear models · Predictable and Unpredictable Variation

Due Monday, February 9, 2015

1) Transformations

Consider the data on mammalian sleeping patterns, available in the R package `faraway`.[1] After installing the `faraway` package, you can load this data with the following commands:

```
library(faraway)
data(mammalsleep, package="faraway")
```

[1] You can also see a short walkthrough of this data set in the mammalsleep.R script, available from the class website, although this script is not necessary for the present question.

Here we are interested in the "body" and "brain" variables, which are the body weight (in kilos) and the brain weight (in grams) of each animal. Once you've loaded the data, address the following questions.

(A) In light of this data, how does the brain weight of a mammal depend upon its body weight? Briefly explain how you decided what the functional form of the x–y relationship should be. Include any plots or summary statistics you judge relevant in making your argument. A complete answer must specify the functional form; estimates for the specific numbers (i.e. coefficients) that characterize this function; and some quantitative description of the residual uncertainty that cannot be predicted by the model.

(B) Which mammals have the largest and smallest brains, in an absolute sense? Which ones have the largest and smallest brains, adjusting for body size?

(C) Suppose there were an animal not in the data set, but with a typical body weight of 100 kilos. Report a 95% prediction interval for this animal's brain weight, expressed on the original scale (grams).

(2) Polynomial regression and prediction intervals

For this question, you will return to the "utilities.csv" data set from class. Recall that each row has information about a monthly utility bill for a house in Minnesota. The variable "gasbill" is the gas bill for that month, measured in dollars. The "temp" variable depicts the average temperature, in degrees Fahrenheit, for the billing period.

Re-fit first-order through fourth-order polynomial regression models for the gas bill (Y) versus temperature (X). That is, fit the models

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i \\
y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \\
y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i \\
y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + e_i
\end{aligned}
$$

Pick your favorite model of these four, and briefly explain why you chose it trade-off. Use both the linear model and your favorite polynomial model to generate naïve prediction intervals for the gas bill during a month in which the average temperature is 50 degrees Fahrenheit. Make sure you state the empirical coverage level of the intervals. Comment on the differences between the intervals.

3) Should we aggregate or not?

For this question, you will need the "TenMileRace" data set from the `mosaic` package in R, which you will load using the command `data(TenMileRace)` after having loaded the `mosaic` package at the beginning of your R session. Quoting the data set description: "The Cherry Blossom 10 Mile Run is a road race held in Washington, D.C. in April each year. The name comes from the famous cherry trees that are in bloom in April in Washington.... This data frame contains the results from the 2005 race." If you type the command `help(TenMileRace)`, you will get a description of each variable in the data set.

(A) Use a linear model to quantify the relationship between a runner's net finishing time (in seconds) and his or her age in years. what seems to be the effect of one additional year of age on finishing time? What is R^2 for this regression?

(B) Now fit two separate linear models for finishing time versus age: one for men alone, and one for women alone. Within each subset, what seems to be the effect of one additional year of age on finishing time, and what is R^2? Is this consistent with what you found in Part A? Describe what you think is going on here. Note: you can create a new data set from a subset of the original one using the subset command. For example:

```
women = subset(TenMileRace, sex=="F")
```

Notice the quotation marks and the double-equals sign, which is how we test for whether a variable takes a specific value.