Department of Statistics, The Chinese University of Hong Kong
STAT 5102 Regression in Practice (Term 1, 2018–19)

Assignment 3 · due on 3 December 2018 (Mo)

1. Consider the data in `fevdata.txt`: The first line contains the variable names which are `age`, `fev`, `ht`, `sex`, and `smoke`. In this assignment, we will consider models that use `age` (age of children, measured in years), `ht` (height of children in inches), and `fev` (forced expiratory volume, a measure of lung capacity, measured in litres).

   (a) Fit a linear model to predict FEV from age. You may use the R function `boxcox` to find a simple power transformation (-1 = reciprocal, 0 = log, 0.5 = square root) close to the Box-Cox maximum likelihood estimate. Which simple transformation seems best?

   (b) Fit a linear model with the best simple transformed response predicted by age and examine the residual plot of the fit. Has the transformation improved adherence to the constant variance assumption? Is this linear model acceptable? Briefly explain why or why not.

   (c) Repeat parts (a) and (b) using log(`age`) as the explanatory variable. Which set of transformations of `fev` and age seems to be best to match linear model assumptions?

   (d) Add a second explanatory variable, `ht`, or a transformation of it, to the model found in (c). Write an equation to express the model. Find 95% confidence intervals for each model parameter (one intercept and two slopes) in the (possibly) transformed scale. As best as you can, interpret these confidence intervals in the scale of the initial measurements. (Note: There is no single set of transformations that is unambiguously best. Use your judgment.)

2. The multicollinearity problem can be illustrated with an example using the data shown in the dataset `HAMILTON.txt`. The values of x_1, x_2 and y in the table at right represent appraised land value, appraised improvements value and sale price, respectively, of a randomly selected residential property. (All measurements are in thousands of dollars.)

   (a) Calculate the coefficient of correlation between y and x_1. Is there evidence of a linear relation between sale price and appraised land value? Repeat the same for (y, x_2).

   (b) Based on the results in (a), do you think the model

   $$E(y \mid X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

   will be useful for predicting sale price?

   (c) Fit the model shown in (b) and conduct a test of model adequacy. In particular, note the value of R^2. Does the result agree with your answer in (c)?

   (d) Calculate the coefficient of correlation between x_1 and x_2. What does the result imply? Compute also the VIF's for these two covariates.

3. The text website has a data file (created from data at `www.basketball-reference.com` showing, for each game in 2010-2011 season of the National Basketball Association in which Rajon Rondo of the Boston Celtics played, $x =$ the number of assists he recorded and $y =$ whether the Celtics won (1=yes). Using SAS, or other software,

   (a) show that the logistic model fitted to these data gives

   $$\text{logit}\{\widehat{\Pr}(Y = 1 \mid x)\} = -2.235 + 0.294x;$$

   (b) show that $\widehat{\Pr}(Y = 1 \mid x)$ increases from 0.21 to 0.99 over the observed range of x from 3 to 24 and

   (c) construct a significance test an confidence interval about the effect in the conceptual population that these games represent.

4. Derive the maximum likelihood estimate of p, denoted by $\hat{p}_{MLE}$ from a random sample of Bernoulli random variables $(X_1, \ldots, X_n)$, where $X_i \sim \text{Bern}(p)$, i.i.d..