BMI 713: Computational Statistics for Biomedical Sciences

Assignment 7

October 28, 2010 (due November 4)

Simple Linear Regression

1. To study the relationship between a father's height and his son's height, Karl Pearson (1857-1936) collected the data of heights from 1078 father-son pairs.

   (a). Get the dataset by the following R commands:

   ```
   install.packages("UsingR")
   library(UsingR)
   data(father.son)
   ```

   Then the data frame `father.son` contains the 1078 observations on 2 variables: *fheight* (father's height in inches, x) and *sheight* (adult son's height in inches, y).

   (b). Draw a scatter plot of son's height versus father's height. Does the relationship appear linear?

   (c). Fit the simple linear regression of son's height on father's height. What are the estimated regression coefficients, a and b, respectively?

   (d). Add the regression line $y = a + bx$ to the plot in (b).

   (e). Calculate Pearson correlation coefficient r between father's height and son's height. Perform a proper test to test the null hypothesis $\rho = 0$, where ρ is the population correlation coefficient.

   (f). What is the 95% confidence interval for the slope coefficient β?

   (g). Calculate the coefficient determination R^2. What does the R^2 statistic mean?

   (h). Draw a residual plot. Are the residuals normally distributed with constant variance?

   (i). What are the estimated means of son's height given that his father's height is 72, 75, 60, and 63 inches, respectively?

   (Notice that sons of tall fathers tended to be tall, but on average not as tall as their fathers. Similarly, sons of short fathers tended to be short, but on average not as short as their fathers. This phenomenon was first described by Sir Francis Galton, as "regression towards mediocrity", where the term *regression* came from. The regression effect – phenomenon of regression toward the mean – appears in any test-retest situation.)

   (j). Given a father's height, we can use simulation method to construct the 100(1-α)% confidence interval for the mean of his son's height. First draw 1000 samples each of size 1078 with replacement from the 1078 pairs of father-son heights, then from each sample fit a linear regression model by the method of least squares, and compute the estimated mean of son's height.

   What are the mean and standard deviation of these 1000 simulated values?

   Sort these 1000 estimated means in ascending order. Denote the 25th largest as h_{25} and the 975th largest as h_{975}, which are our estimates of the 0.025 and 0.975 quantiles of the sampling distribution for the mean of son's height. Then the 100(1-α)% confidence interval for the mean of the son's height is (h_{25}, h_{975}). Compute the 95% confidence interval for the mean of son's height if his father is 72 inches tall.

Contingency Table

2. In an investigation of the association between smoking habit and lung cancer, lung cancer patients and controls were obtained. The patients and controls were matched for age, sex, and community. The data are shown in the table below. (Data from P. Notani and L. D. Sanghvi, "A Retrospective Study of Lung Cancer in Bomby", *Br. J. Cancer* 29(6): 477-482, 1974.)

	Lung Cancer	Controls
Smokers	413	318
Nonsmokers	107	201

   (a). What is the type of this study in terms of study design?

   (b). Calculate the odds of lung cancer for smokers, the odds of lung cancer for nonsmokers, and the ratio of two odds.

   (c). Calculate the 95% confidence interval for the odds ratio.

3. The Salk polio vaccine trials of 1954 included a double-blind experiment in which elementary school children of consenting parents were assigned at random to injection with the Salk vaccine of with a placebo. Both treatment and control groups were set at 200,000 because the target disease, infantile paralysis, was uncommon (but greatly feared). (Data from J. M. Tanur *et al.*, *Statistics: A Guide to the Unknown*, San Francisco: Holden-Day, 1972.)

	Infantile paralysis victim?	
	Yes	No
Placebo	142	199,858
Salk polio vaccine	56	199,944

   (a). Is this a randomized experiment or a cohort study?

   (b). Calculate the proportion of infantile paralysis victims among placebo group, and the proportion of infantile paralysis victims among vaccine group, respectively.

   (c). What is the risk difference? Calculate the 95% confidence interval for the risk difference.

   (d). What is the relative risk? Calculate the 95% confidence interval for the relative risk.