

BMI 713: Computational Statistics for Biomedical Sciences

Assignment 8

Nov. 17, 2010 (due Nov. 24 2010)

1. One important assumption in linear regression is that the error term e (see Lecture 9) follows a normal distribution. To check whether this assumption is valid, one can perform the linear regression first and look at the residuals of the linear regression.
 - (a) As in Question 1 of Assignment 7, perform linear regression analysis on the data `father.son` using father's height as predictor and son's height as response.
 - (b) Get the residuals of the linear regression and plot the QQ plot of the residuals. Do these residuals look like following normal distribution?
 - (c) Other than the QQ plot, Shapiro's test can be used to test the normality of the data. Perform Shapiro's test on the residuals and report the p-value. What conclusion can you draw from the test? (Hint: use the function `shapiro.test` in R to perform Shapiro's test.)
2. We use simulation to compare p-value, FWER and FDR.
 - (a) Generate 1000 values from the normal distribution $\mathcal{N}(2, 1)$, 9000 values from the standard normal distribution $\mathcal{N}(0, 1)$ and record these 10,000 values in a vector `x`.
 - (b) For each of the 10,000 observation, perform a z-test to see if its expected mean is 0 or not (null hypothesis: the expected mean is 0; alternative hypothesis: the expected mean is larger than 0).
 - (c) At the significance level $\alpha = 0.05$, how many true positives, false positives, and false negatives do you get? What is the type I error rate, type II error rate, and FDR?
 - (d) Use the Bonferroni method to control for the FWER to be less than 0.05. Report the statistics as in (c) and compare them.
 - (e) Use the Benjamini-Hochberg method to control for the FDR to be less than 0.05. Repeat the same analysis as in (d).