CH 5440 Multivariate Data Analysis

Assignment 1

Due Date: 29/1/18

1. (a) Let $x_1, x_2, \cdots, x_N$ and $y_1, y_2, \cdots, y_N$ be a set of N measurements of two variables x and y which are linearly related. We are interested in determining the linear regression parameter m where $y = mx + c$. Assume that the measurements of x and y contain errors, with standard deviations σ_δ and σ_ε, respectively. If the ratio of the error variances $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\delta^2}$ is known, derive the weighted TLS (WTLS) estimates of m and c in terms of $s_{xx}, s_{yy}, \bar{x}, \bar{y}, \lambda$. Prove that the standard OLS estimates and inverse OLS estimates for m and c are obtained in the limit as λ tends to ∞ and 0, respectively. (b) How will the solution for m change if it is already known that the constant c is known to be 0?

   Note: The WTLS regression problem when the error variances are known is the solution of the following minimization problem. Multiply the objective function by σ_ε^2 and replace the ratio of the error variances by λ. Differentiate the objective function with respect to the decision variables and solve resulting set of nonlinear algebraic equations for obtaining the parameters m and c.

$$\underset{m,c,\hat{x}_i}{Min} \sum_{i=1}^{N} (y_i - m\hat{x}_i - c)^2 / \sigma_\varepsilon^2 + (x_i - \hat{x}_i)^2 / \sigma_\delta^2$$

   (b) Obtain the solution for the estimates of x and y for each case (OLS, IOLS, TLS) in terms of the regression parameters and measurements

2. The level of phytic acid in urine samples was determined by a catalytic fluorimetric (CF) method and the results were compared with those obtained using an established extraction photometric (EP) technique. The results, in mg/L, are the means of triplicate measurements, as shown in Table 2. Assume that the errors corrupting the measurements of both methods have same variance.

   (a) Is the new method (CF) a good substitute for the established method (EP) for measuring the level of phytic acid in urine? Justify your conclusion.
   (b) Estimate the level of phytic acid in urine if EP measurement is 2.31 mg/l and CF measurement is 2.20 mg/l.

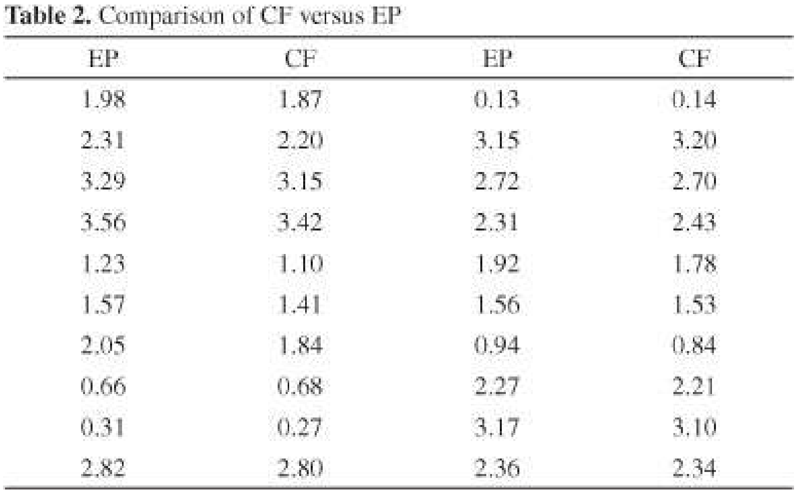

Table 2. Comparison of CF versus EP

EP	CF	EP	CF
1.98	1.87	0.13	0.14
2.31	2.20	3.15	3.20
3.29	3.15	2.72	2.70
3.56	3.42	2.31	2.43
1.23	1.10	1.92	1.78
1.57	1.41	1.56	1.53
2.05	1.84	0.94	0.84
0.66	0.68	2.27	2.21
0.31	0.27	3.17	3.10
2.82	2.80	2.36	2.34

3. Anscombe (1973) has provided four synthetic data sets consisting of two variables x and y (data in file anscombe.xls). Find the best fit linear model for the four data sets using standard OLS. What do you observe? For which of the four data sets do you think that a linear model is appropriate and why?