

CH5440: MULTIVARIATE DATA ANALYSIS
QUIZ 11/03/18 2:00-2:50 PM

1. The age (x) in years and systolic blood pressure (y) in mm Hg of 30 patients were measured and the following statistics were obtained

$$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} 45 \\ 143 \end{bmatrix}; s_{xx} = 226; s_{yy} = 493; s_{xy} = 220$$

The variances and cross covariance are computed about the mean and dividing by total number of samples. Assume that a linear relation between blood pressure (BP) and age exists, i.e. $y = ax + b$.

b. Estimate the parameters a and b using

- (a) OLS method assuming age measurements are free of errors: $a = 0.9709$ $b = 98.7147$
- (b) Inverse OLS method assuming BP measurements are free of errors: $a = 2.2453$ $b = 41.194$
- (c) Total least squares assuming both BP and age measurements contain errors and the error variances are equal: $a = 1.7778$ $b = 62.2945$
- (d) OLS method assuming age measurements are free of errors and intercept b is specified as 120: $a = 0.5575$

2. Fisher iris data contains measurements of petal length (pl), petal width (pw), sepal length (sl), sepal width (sw) in cms of 150 iris flowers. The sample mean and covariance matrix of the data is given below

$$\bar{z} = [5.8433 \ 3.0573 \ 3.7580 \ 1.1993]; S_z = \begin{bmatrix} 0.6857 & -0.0424 & 1.2743 & 0.5163 \\ -0.0424 & 0.1900 & -0.3297 & -0.1216 \\ 1.2743 & -0.3297 & 3.1163 & 1.2956 \\ 0.5163 & -0.1216 & 1.2956 & 0.5810 \end{bmatrix}$$

The sample covariance matrix has been computed about the mean and dividing by the total number of samples. The largest three eigenvalues of the covariance matrix are 4.2282, 0.2427, and 0.0782

- (a) The smallest eigenvalue of the covariance matrix is 0.0239
- (b) The orthonormal eigenvector corresponding to the smallest eigenvalue is $v_4 = [-0.3155 \ 0.3201 \ 0.4798 \ -0.7535]^T$
- (c) The number of PCs to be retained if 99% of the variance has to be captured by the retained PCs is 3
- (d) If only three of the four measured variables are linearly independent, and petal length is chosen as the dependent variable, then the regression equation relating the dependent and independent variables using total least squares is given by

$$pl = 0.2339pw - 2.5705sl - 0.9914sw + 5.5023$$

- (e) The denoised estimates corresponding to measurements of $pl = 5.2$ cms, $pw = 3.5$ cms, $sl = 1.4$ cms and $sw = 0.2$ cms are respectively, 6.1569cm, 3.2762cm, 3.6684cm, 1.1487cm