Homework 3 (Solution)

Problem 1. (a) This problem assumes equal variance-covariance matrix for underlying populations and the observations are from multivariate normal distributions.

The error rates for population 1, 2, 3 are 0, 0.04, 0.02 respectively. Use the complete data set twice, once to estimate both the parameters and the second time to the resubstitution error rates, which may cause the error rate estimates not reliable. The samples size for each population is 50, resubstitution method may not be efficient.

(b) Randomly select 30 observations, 10 from each species. When using the remaining data to from a discriminant rule and calculate the resubstitution error rates are 0, 0.05, 0.025 for population 1, 2, 3 respectively. When estimate the error rate based on the test data set, the error rates are all 0’s for the three species. This method gives less bias estimates of error rates.

(c) Without assuming equal covariance matrices, the error rates estimated by the resubstitution methods are the same as the results in part (a). The error rates estimated by the cross-validation are slightly higher than the results from part (a) and the estimates by resubstitution method. Since resubstitution estimates are optimistically biased, we might want to use cross-validation estimates of the error rates and assume unequal covariance matrices.

Problem 2. (a) Skip.

(b) Take $\log(y_i) = \log(\theta_1) + \theta_2\log(x_{1i}) + \theta_3\log(x_{2i}) + \text{error}$

The estimates for $\log(\theta_1)$, $\theta_2$, $\theta_3$ are 2.26811, -0.16947 and 0.95020 respectively.

So we can use the following starting values:

$\theta_1 = 9.6611$, $\theta_2 = -0.16947$, $\theta_3 = 0.95020$.

(c) The single updates of the estimates are:

$\theta^{(1)} = (13.2884, -0.4903, 0.8486)'$ and

$f(\theta^{(1)}) = (13.2884, 31.0467, 72.5366, 9.4597, 51.6372, 7.7543, 18.1168, 42.3227)'$

(d) The single updates of the estimates by Newton-Raphson method are:
\( \theta^{(1)} = (12.4393, -0.4903, 0.9056)' \) and
\[ f(\theta^{(1)}) = (12.4393, 30.7676, 76.1010, 8.8563, 54.1809, 7.2601, 17.9573, 44.4159)' \]
The results are slightly different from what we got in part (c).

(e) The 95% CI's for \( \theta_1, \theta_2, \theta_3 \) are [6.1082, 20.4935], [-0.7483, -0.2027] and [0.5848, 1.1525].
The 95% prediction interval when \( x=(2,2) \) is [38.2235, 70.4761].