Stat 8311, Fall 2006: Overparameterized two-sample

Here are computations in R for the two sample problem, with m observations per group. We take the parameterized model to be

$$Y = X\beta = \begin{pmatrix} J & J & 0 \\ J & 0 & J \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

where J is a vector of m ones and 0 is a vector of m zeroes. We do the computations with $m = 1$ to find $(X'X)^+$, the Moore Penrose G-inverse of $X'X$):

```
> (xtx <- matrix(c(2, 1, 1, 1, 1, 0, 1, 0, 1), ncol = 3))

     [,1] [,2] [,3]
[1,]    2    1    1
[2,]    1    1    0
[3,]    1    0    1

> (s <- svd(xtx))

$d
[1] 3.000000e+00 1.000000e+00 1.453489e-16

$u
           [,1]          [,2]       [,3]
[1,] -0.8164966  5.724873e-17  0.5773503
[2,] -0.4082483  7.071068e-01 -0.5773503
[3,] -0.4082483 -7.071068e-01 -0.5773503

$v
           [,1]          [,2]       [,3]
[1,] -0.8164966  3.567025e-17  0.5773503
[2,] -0.4082483  7.071068e-01 -0.5773503
[3,] -0.4082483 -7.071068e-01 -0.5773503

> round(s$u %*% diag(s$d) %*% t(s$v))

     [,1] [,2] [,3]
[1,]    2    1    1
[2,]    1    1    0
[3,]    1    0    1

> s1 <- c(1/3, 1, 0)
> round(s$u %*% diag(s1) %*% t(s$v))

     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    1    0
[3,]    0    0    1
```

For the general case of $m > 1$, the eigenvalues of $X'X$ are multiplied by m, so the last expression for $(X'X)^+$ has non-zero diagonals divided by m. Since $X'y = (y_{++}, y_{1+}, y_{2+})'$, a least squares estimate is given by

$$\beta_0 = (X'X)^+X'y = \begin{pmatrix} 0 \\ \bar{y}_{1+} \\ \bar{y}_{2+} \end{pmatrix}$$

The set of *all* least squares estimates are of the form

$$\beta_0 + \gamma \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} \gamma \\ \bar{y}_{1+} - \gamma \\ \bar{y}_{2+} - \gamma \end{pmatrix}$$

where the vector $(1, -1, -1)' \in N(X)$ is a basis for $N(X)$. For example:

1. $\gamma = 0$ corresponds to the *cell means* parameterization.
2. $\gamma = \bar{y}_{++}$ corresponds to the "effects" parameterization with $\beta_1 + \beta_2 = 0$.
3. $\gamma = \bar{y}_{1+}$ corresponds to the "drop first level" parameterization used by R.
4. $\gamma = \bar{y}_{2+}$ corresponds to the "drop last level" parameterization used by SAS.
5. $\gamma = -\bar{y}_{2+}$ gives
$$\hat{\beta} = \begin{pmatrix} -\bar{y}_{2+} \\ \bar{y}_{1+} - \bar{y}_{2+} \\ 0 \end{pmatrix}$$

Estimability

A function of $c'\beta$ is estimable if and only if $c = X'\lambda$, for some vector λ, or if c is a linear combination of the rows of X. For the two sample problem, a basis for the row space is clearly given by $((1, 1, 0)', (1, 0, 1)')$ and so the estimable functions are $(\lambda_1 + \lambda_2, \lambda_1, \lambda_2)$ for any real numbers λ_1 and λ_2. In particular:

1. $\beta_0 + \beta_1$ is estimable, for $\lambda_1 = 1$, $\lambda_2 = 0$.
2. $\beta_2 - \beta_1$ is estimable for $\lambda_1 = -\lambda_2 = 1$.
3. None of the elements of β are themselves estimable.