Maximum likelihood estimation

Does LS estimate have other optimality properties?

Since we have assumed that $Y \sim N_n(X\beta, \sigma^2 I_n)$ to derive distributional properties of $\hat{\beta}$, let us derive the maximum likelihood estimates of β and σ^2 under this assumption. $\hat{\beta}_{mle}$ and $\hat{\sigma}^2$ are values of β and σ^2 which maximize the likelihood,

$$(2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right).$$

Equivalently, we may maximize the loglikelihood,

$$-\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta).$$

Fix σ^2 and maximize over β, then maximize over σ^2. Now note that maximizing over β (for any fixed σ^2) is equivalent to minimizing $(Y-X\beta)'(Y-X\beta) = ||Y-X\beta||^2$, which yields the same estimate as the least squares. i.e., $\hat{\beta}_{mle} = \hat{\beta}_{ls}$. However, $\hat{\sigma^2} = \text{RSS}/n$, which is not unbiased.

Estimation under linear restrictions or constraints

Consider the following examples.
(i) $y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij}$. Test $H_0 : \tau_1 = \tau_2$. i.e., test whether there is any difference between treatments 1 and 2. Under H_0, $\tau_1 - \tau_2 = 0$, or $A\beta = c$ where $A = a' = (0, 0, \ldots, 0, 1, -1, 0, \ldots, 0)$, $\beta = (\mu, \alpha_1, \ldots, \alpha_I, \tau_1, \tau_2, \ldots)'$.
(ii) $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1}x_{i(p-1)} + \epsilon_i$. Test $H_0 : X_1, \ldots, X_{p-1}$ are not useful.

Recall that, to derive the GLRT, we need to estimate the parameters of the model, both with and without restrictions. While testing linear hypotheses in a linear model, we need to estimate β under the linear constraint $A\beta = c$.

Consider $Y = X\beta + \epsilon$, $X_{n\times p}$ of rank p, first. We will consider the deficient rank case later. Let us see how we can find the least squares estimate of β subject to $H : A\beta = c$, where $A_{q\times p}$ of rank q and c is given. We can use the Lagrange multiplier method of calculus for this as follows.

$$\begin{aligned}&\min_\beta ||Y-X\beta||^2 + \lambda'(A\beta - c)\\ &= \min_\beta \left\{Y'Y - 2\beta'X'Y + \beta'X'X\beta + \lambda'A\beta - \lambda'c\right\},\end{aligned}\tag{1}$$

differentiating which (w.r.t. β) and setting equal to 0, we get,

$$-2X'Y + 2X'X\beta + A'\lambda = 0 \text{ or } X'X\beta = X'Y - \frac{1}{2}A'\lambda_H.$$

Therefore,

$$\hat{\beta}_H = (X'X)^{-1}\left\{X'Y - \frac{1}{2}A'\lambda_H\right\} = \hat{\beta} - \frac{1}{2}(X'X)^{-1}A'\lambda_H \quad (*).$$

Differentiating (1) w.r.t. λ, we get $A\beta - c = 0$. Since

$$\begin{aligned} c = A\hat{\beta}_H &= A\hat{\beta} - \frac{1}{2}A(X'X)^{-1}A'\lambda_H, \\ c - A\hat{\beta} &= -\frac{1}{2}A(X'X)^{-1}A'\lambda_H, \text{ and hence} \\ -\frac{1}{2}\lambda_H &= \left[A(X'X)^{-1}A'\right]^{-1}(c - A\hat{\beta}), \text{ and therefore} \\ \hat{\beta}_H &= \hat{\beta} + (X'X)^{-1}A'\left[A(X'X)^{-1}A'\right]^{-1}(c - A\hat{\beta}). \end{aligned}$$

To establish minimization subject to $A\beta = c$, note that

$$\begin{aligned} ||X(\hat{\beta} - \beta)||^2 &= (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \\ &= (\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta)'X'X(\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta) \\ &= (\hat{\beta} - \hat{\beta}_H)'X'X(\hat{\beta} - \hat{\beta}_H) + (\hat{\beta}_H - \beta)'X'X(\hat{\beta}_H - \beta) \\ &\quad +2(\hat{\beta} - \hat{\beta}_H)'X'X(\hat{\beta}_H - \beta) \\ &= ||X(\hat{\beta} - \hat{\beta}_H)||^2 + ||X(\hat{\beta}_H - \beta)||^2, \end{aligned}$$

since, from (*) above, and subject to $A\beta = c$,

$$\begin{aligned} (\hat{\beta} - \hat{\beta}_H)'X'X(\hat{\beta}_H - \beta) &= \frac{1}{2}\lambda'_H A(X'X)^{-1}X'X(\hat{\beta}_H - \beta) \\ &= \frac{1}{2}\lambda'_H A(\hat{\beta}_H - \beta) = \frac{1}{2}\lambda'_H(A\hat{\beta}_H - A\beta) = 0. \end{aligned}$$

Therefore,

$$\begin{aligned} ||Y - X\beta||^2 &= ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \beta)||^2 \\ &= ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \hat{\beta}_H)||^2 + ||X(\hat{\beta}_H - \beta)||^2 \\ &\geq ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \hat{\beta}_H)||^2, \end{aligned}$$

and is a minimum when $\beta = \hat{\beta}_H$. (Note, $X(\hat{\beta}_H - \beta) = 0$ implies $X'X(\hat{\beta}_H - \beta) = 0$, so $\hat{\beta}_H - \beta = 0$ since columns of X are linearly independent.) Also, from above, we get,

$$||Y - X\hat{\beta}_H||^2 = ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \hat{\beta}_H)||^2.$$

If we let $\hat{Y} = X\hat{\beta}$ and $\hat{Y}_H = X\hat{\beta}_H$, then

$$||Y - \hat{Y}_H||^2 = ||Y - \hat{Y}||^2 + ||\hat{Y} - \hat{Y}_H||^2.$$

Note that this can also be established using projection matrices, and not just for the full column rank case. Let us first establish it for the case $\text{Rank}(X_{n\times p}) = p$ again, and next extend it.