Multiple Linear Regression Estimators

Properties of Least-Squares Estimators

Gauss-Markov Theorem

Let $\hat{\theta}$ be the least-squares estimate of $\theta = \mathbf{X}\beta$, where $\theta \in \Omega = \mathcal{C}(\mathbf{X})$ and $\mathbf{X}$ may not have full rank. Then among the class of unbiased estimates of $\mathbf{c}^T\theta$, $\mathbf{c}^T\hat{\theta}$ is the unique estimate with minimum variance. We say that $\mathbf{c}^T\hat{\theta}$ is the best linear unbiased estimate (BLUE) of $\mathbf{c}^T\theta$.
Proof.
We know that $\hat{\theta} = \mathbf{X}\hat{\beta} = \mathbf{H}Y$, where $\mathbf{H}\theta = \mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta = \theta$. Hence, $\mathbb{E}(\mathbf{c}^T\hat{\theta}) = \mathbf{c}^T\mathbf{H}\theta = \mathbf{c}^T\theta$, for all $\theta \in \Omega$, so that $\mathbf{c}^T\hat{\theta}$ is unbiased estimate of $\mathbf{c}^T\theta$. Then, $\mathbf{c}^T\theta = \mathbf{E}(\mathbf{d}^T\mathbf{Y}) = \mathbf{d}^T\theta$ or $(\mathbf{c}-\mathbf{d})^T\theta = 0$, so that $(\mathbf{c}-\mathbf{d})$ is orthogonal to Ω. Therefore, $\mathbf{H}(\mathbf{c}-\mathbf{d}) = 0$ and $\mathbf{H}c = \mathbf{H}d$.

Now,

$$\begin{aligned}
Var(\mathbf{c}^T\hat{\theta}) &= Var\left((\mathbf{H}\mathbf{c})^T\mathbf{Y}\right) \\
&= Var\left((\mathbf{H}\mathbf{d})^T\mathbf{Y}\right) \\
&= \sigma^2\mathbf{d}^T\mathbf{H}^T\mathbf{H}\mathbf{d} \\
&= \sigma^2\mathbf{d}^T\mathbf{H}^2\mathbf{d} \\
&= \sigma^2\mathbf{d}^T\mathbf{H}\mathbf{d}
\end{aligned}$$

so that

$$\begin{aligned}
Var(\mathbf{d}^T\mathbf{Y}) - Var(\mathbf{c}^T\hat{\theta}) &= Var(\mathbf{d}^T\mathbf{Y}) - Var((\mathbf{H}\mathbf{d})^T\mathbf{Y}) \\
&= \sigma^2(\mathbf{d}^T\mathbf{d} - \mathbf{d}^T\mathbf{H}\mathbf{d}) \\
&= \sigma^2\mathbf{d}^T(\mathbf{I}_n - \mathbf{H})\mathbf{d} \\
&= \sigma^2\mathbf{d}^T(\mathbf{I}_n - \mathbf{H})^T\underbrace{(\mathbf{I}_n - \mathbf{H})\mathbf{d}}_{:=\mathbf{d}_1} \\
&= \sigma^2\mathbf{d}_1^T\mathbf{d}_1 \geq 0
\end{aligned}$$

with equality only if $(\mathbf{I}_n - \mathbf{H})\mathbf{d} = 0$ or $\mathbf{d} = \mathbf{H}\mathbf{d} = \mathbf{H}\mathbf{c}$. Hence, $\mathbf{c}^T\hat{\theta}$ has minimum variance and is unique.

Corollary

If $\mathbf{X}$ has full rank, then $\mathbf{a}^T\hat{\beta}$ is the BLUE of $\mathbf{a}^T\beta$ for every vector $\mathbf{a}$.
Proof.
Now $\theta = \mathbf{X}\beta$ implies that $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\theta$ and $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\theta}$. Hence, setting $\mathbf{c}^T = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ we have that $\mathbf{a}\hat{\beta}(= \mathbf{c}^T\hat{\theta})$ is the BLUE of $\mathbf{a}\beta(= \mathbf{c}^T\theta)$ for every vector $\mathbf{a}$.

Theorem (Unbiased Estimator of σ^2)

If $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\beta$, where $\mathbf{X}$ is an $n \times p$ matrix of rank r $(r \leq p)$, and $Var(\mathbf{Y}) = \sigma^2\mathbf{I}_n$, then

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \hat{\theta})^T(\mathbf{Y} - \hat{\theta})}{n-r} = \frac{RSS}{n-r}$$

is an unbiased estimate of σ^2.

Proof.
Consider the full-rank representation $\theta = \mathbf{X}_1\alpha$, where $\mathbf{X}_1$ is $n \times r$ of rank r. Then,

$$\mathbf{Y} - \hat{\theta} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$. Using the properties of the Hat matrix we have the following:

$$\begin{aligned}(n-r)\hat{\sigma}^2 &= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\ &= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})^2\mathbf{Y} \\ &= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\end{aligned}$$

Since $\mathbf{H}\theta = \theta$, we have

$$\mathbb{E}(\mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}) = \sigma^2 tr(\mathbf{I}_n - \mathbf{H}) + \theta^T(\mathbf{I}_n - \mathbf{H})\theta = \sigma^2(n-r)$$

and hence $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$.

Maximum Likelihood Estimation

Assuming normality, the likelihood function $L(\beta, \sigma^2)$ for the full rank regression model is the probability density of $\mathbf{Y}$, namely,

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\beta||^2\right\}$$

Let $\ell(\beta, \sigma^2) = \log L(\beta, \sigma^2)$. Then ignoring constants, we have

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\beta||^2$$

So, taking derivatives with respect to β and σ^2, we have that

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= -\frac{1}{2\sigma^2}(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^t\mathbf{X}\beta) \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}||\mathbf{y} - \mathbf{X}\beta||^2\end{aligned}$$

Setting $\frac{\partial \ell}{\partial \beta} = 0$, we get the LS estimate of β, which clearly maximizes $\ell(\beta, \sigma^2)$ for any $\sigma^2 > 0$. Hence,

$$L(\beta, \sigma^2) \leq L(\hat{\beta}, \sigma^2)$$

for all $\sigma^2 > 0$ with equality if and only if $\beta = \hat{\beta}$.

We now wish to maximize $L(\hat{\beta}, \sigma^2)$, or equivalently $\ell(\hat{\beta}, \sigma^2)$ with respect to σ^2. Setting $\frac{\partial \ell}{\partial \sigma^2} = 0$, we get a stationary value of $\hat{\sigma}^2 = ||\mathbf{y} - \mathbf{X}\beta||^2/n$. Then,

$$\ell(\hat{\beta}, \hat{\sigma}^2) - (\hat{\beta}, \sigma^2) = -\frac{n}{2}\left(\log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) + 1 - \frac{\hat{\sigma}^2}{\sigma^2}\right) \geq 0$$

since $x \leq e^{x-1}$ and therefore $\log x \leq x - 1$ for $x \geq 0$ (with equality when $x = 1$). Hence,

$$L(\beta, \sigma^2) \leq L(\hat{\beta}, \hat{\sigma}^2), \text{ for all } \sigma^2 > 0$$

with equality if and only if $\beta = \hat{\beta}$ and $\sigma^2 = \hat{\sigma}^2$. Thus, $\hat{\beta}$ and $\hat{\sigma^2}$ are the maximum likelihood estimates of β and σ^2. Also, we can compute the maximum value of the likelihood as

$$L(\hat{\beta}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-n/2}e^{-n/2}.$$