Notes for 751-752
Sections 10-11

Martin Lindquist*

September 25, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

10 Properties of Least Squares Estimates

In this section we continue working with the linear model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. However, we now add a couple of additonal assumptions. Let us begin by assuming that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Equivalently, this can be expressed as as $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}$. Under this assumption the linear model expresses how the mean of the random vector $\mathbf{y}$ changes as a function of the explanatory variables $\mathbf{X}$. It also assumes that the observations y_i and y_j are uncorrelated for $i \neq j$.

It is important to note that the least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ was derived without making these assumptions. Therfore, even if $E(\mathbf{y}) \neq \mathbf{X}\boldsymbol{\beta}$ does not hold, the linear model can still be fit to the data. However, the resulting estimate may have poor properties. In contrast, we will show that under the assumptions above, the estimates $\boldsymbol{\beta}$ have some very good properties. We begin by noting that the least-squares estimator is a random vector, and thus we can compute its expected value and variance.

Theorem: 10.1 If $\mathbf{X}$ is of full rank, then

(a) The least squares estimate is unbiased, i.e., $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

(b) The variance-covariance matrix of the least squares estimate is $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Example: 10.2 To illustrate, consider the case of simple linear regression. We seek to compute the variance-covariance matrix. Using the the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ we obtain:

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\begin{bmatrix} \sum_i x_i^2/n & -\bar{\mathbf{x}} \\ -\bar{\mathbf{x}} & 1 \end{bmatrix}.\end{aligned}$$

Thus, we have that

$$\mathrm{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2/n}{\sum_i (x_i - \bar{x})^2},$$

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2},$$

and

$$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2\bar{\mathbf{x}}}{\sum_i (x_i - \bar{x})^2}$$

Studying $\mathrm{var}(\hat{\beta}_1)$, we note that there are three aspects of the scatter plot affect the variance of the regression slope: (i) the spread around the regression line; (ii) the spread of the $\mathbf{x}$ values; and (iii) the sample size n. Here less scatter around the line indicates the slope will be more consistent from sample to sample, a large variance of $\mathbf{x}$ provides a more stable regression, and having a larger sample size provides more consistent estimates.

Theorem: 10.3 (**Gauss-Markov Theorem**). If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, then the least squares estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimators (BLUE).

Proof: Consider an alternative linear estimator $\mathbf{b} = \mathbf{A}\mathbf{y}$ of $\boldsymbol{\beta}$. As it is a linear estimator we can express it as follows: $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}$ where $\mathbf{D}$ is a non-zero matix. For $\mathbf{A}\mathbf{y}$ to be an unbiased estimator of $\boldsymbol{\beta}$, it must hold that $E(\mathbf{A}\mathbf{y}) = \boldsymbol{\beta}$. This can be expressed as:

$$\begin{aligned} E(\mathbf{b}) &= \mathbf{A}E(\mathbf{y}) \\ &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})\mathbf{X}'\boldsymbol{\beta} \\ &= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})E(\mathbf{y}) \\ &= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

This provides a condition for $\mathbf{b}$ to be an unbiased estimator: $\mathbf{D}\mathbf{X} = \mathbf{0}$.

Now,

$$\begin{aligned} \mathrm{var}(\mathbf{b}) &= \sigma^2\mathbf{A}\mathbf{A}' \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}]' \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}' + \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}'] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{D}\mathbf{D}' \\ &= \mathrm{var}(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{D}\mathbf{D}' \end{aligned}$$

Since $\mathbf{D}\mathbf{D}'$ is positive definite, the variance of $\mathrm{var}(\mathbf{b})$ exceeds that of $\mathrm{var}(\hat{\boldsymbol{\beta}})$.

Note that here 'best' implies minimum variance, and 'linear' that the estimators are linear functions of $\mathbf{y}$ Remarkably, the results holds for any distribution of $\mathbf{y}$. It is important to consider unbiased estimators, since we could always minimize the variance by defining an estimator to be constant (hence variance 0). If one removes the restriction of unbiasedness, then minimum variance cannot be the definition of 'best'. Often one then looks to mean squared error, the squared bias plus the variance, instead.

We can extend these results to linear contrasts of $\boldsymbol{\beta}$ to say that $\mathbf{q}'\hat{\boldsymbol{\beta}}$ is the *best* estimator of $\mathbf{q}'\boldsymbol{\beta}$ in the sense of minimizing the variance among linear (in $\mathbf{Y}$) unbiased estimators.

Theorem: 10.4 If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}$, the best linear unbiased estimators of $\mathbf{q}'\beta$ is $\mathbf{q}'\hat{\beta}$, where $\hat{\beta}$ is the least-squares estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Proof: Consider estimating $\mathbf{q}'\boldsymbol{\beta}$. Clearly, $\mathbf{q}'\hat{\boldsymbol{\beta}}$ is both unbiased and linear in $\mathbf{Y}$. Also note that $\mathrm{Var}(\mathbf{q}'\hat{\boldsymbol{\beta}}) = \mathbf{q}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{q}\sigma^2$. Let $\mathbf{k}'\mathbf{Y}$ be another linear unbiased estimator, so that $E[\mathbf{k}'\mathbf{Y}] = \mathbf{q}'\boldsymbol{\beta}$. But, $E[\mathbf{k}'\mathbf{Y}] = \mathbf{k}'\mathbf{X}\boldsymbol{\beta}$. It follows that since $\mathbf{q}'\boldsymbol{\beta} = \mathbf{k}'\mathbf{X}\boldsymbol{\beta}$ must hold for all possible $\boldsymbol{\beta}$, we have that $\mathbf{k}'\mathbf{X} = \mathbf{q}'$. Finally note that

$$\mathrm{Cov}(\mathbf{q}'\hat{\boldsymbol{\beta}}, \mathbf{k}'\mathbf{Y}) = \mathbf{q}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{k}'\sigma^2.$$

Since $\mathbf{k}'\mathbf{X} = \mathbf{q}'$, we have that

$$\mathrm{Cov}(\mathbf{q}'\hat{\boldsymbol{\beta}}, \mathbf{k}'\mathbf{Y}) = \mathrm{Var}(\mathbf{q}'\boldsymbol{\beta}).$$

Now we can execute the proof easily.

$$\begin{aligned}
\mathrm{Var}(\mathbf{q}'\hat{\boldsymbol{\beta}} - \mathbf{k}'\mathbf{Y}) &= \mathrm{Var}(\mathbf{q}'\hat{\boldsymbol{\beta}}) + \mathrm{Var}(\mathbf{k}'\mathbf{Y}) - 2\mathrm{Cov}(\mathbf{q}'\hat{\boldsymbol{\beta}}, \mathbf{k}'\mathbf{Y}) \\
&= \mathrm{Var}(\mathbf{k}'\mathbf{Y}) - \mathrm{Var}(\mathbf{q}'\hat{\boldsymbol{\beta}}) \\
&\geq 0.
\end{aligned}$$

Here the final inequality arises as variances have to be non-negative. Then we have that $\mathrm{Var}(\mathbf{k}'\mathbf{Y}) \geq \mathrm{Var}(\mathbf{q}'\hat{\boldsymbol{\beta}})$ proving the result.

Notice, normality was not required at any point in the proof, only restrictions on the first two moments. In following sections, we'll see the consequences of assuming normality.

10.1 Estimating σ^2

We can devise an unbiased estimator for σ^2 based on the least-squares estimator $\hat{\beta}$. Let us define

$$s^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n - r).$$

This is a generalization of the sample variance.

Theorem: 10.5 s^2 is an unbiased estimate of σ^2.

Theorem: 10.6 An unbiased estimate of $\mathrm{var}(\hat{\beta})$ is given by

$$\hat{\mathrm{var}}(\hat{\beta}) = s^2(\mathbf{X}'\mathbf{X})^{-1}.$$

10.2 Model misspecification

Any linear model is only as good as the specified design matrix. Incorrect specification can lead to bias and model misfit, resulting in power loss and an inflated false positive rate. Problems can arise if either irrelevant explanatory variables are included, or relevant variables are omitted. Here we study biases due to model misspecification.

Assume, for example, that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_1\boldsymbol{\beta}_2 + \epsilon.$$

Suppose we erroneously fit the model:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \epsilon.$$

This implies we are under-fitting the model. Then the least-squares estimator is given by $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$. Computing the expectation, we see that

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_1) &= E((\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}) \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'E(\mathbf{y}) \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \end{aligned}$$

In addition,

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_1) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

The estimate of $\boldsymbol{\beta}$ is biased. However, note that the bias disappears if $\boldsymbol{\beta}_2 = 0$ or $\mathbf{X}_1'\mathbf{X}_2 = 0$.

The estimate of σ^2 is also biased, with

$$E(s^2) = \sigma^2 + \frac{1}{n-p}\boldsymbol{\beta}_2'\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2\boldsymbol{\beta}_2.$$

If in contrast, irrelevant variables are included (over-fitting), the parameters will remain unbiased. However, the variance-covariance matrix of $\boldsymbol{\beta}_1$ will be inflated effecting subsequent inference.

To see this, assume that the correctly specified model is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \epsilon.$$

However, suppose we instead use the model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon \\ &= \mathbf{X}\boldsymbol{\beta} + \epsilon \end{aligned}$$

where $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2]'$. Now, one can show:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_1) &= \boldsymbol{\beta}_1 \\ E(s^2) &= \sigma^2 \\ \text{var}(\hat{\boldsymbol{\beta}}_1) &= \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} \\ + &\quad \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{X}_2)[\mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1')^{-1} \end{aligned}$$

11 Generalized Least Squares

What happens if we relax the assumption that $\text{cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$?

Although this assumption has no effect on the actual ordinary least-squares (OLS) estimate, it does affect the properties of the estimator and any subsequent inference. To illustrate, assume $\text{cov}(\mathbf{Y}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is assumed to be known. Note, in practice, we will also have to estimate $\mathbf{V}$. In this setting, $\boldsymbol{\beta}$ is still unbiased. However, the variance-covariance matrix is

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

In addition, there is no longer any guarantee that the estimator is the BLUE of $\boldsymbol{\beta}$.

Here we introduce the method of generalized least squares (GLS) to improve upon estimation efficiency for the case when $\text{cov}(\mathbf{Y}) \neq \sigma^2\mathbf{I}$

Example: 11.1 (Clustered data). Let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_K \end{pmatrix},$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is a vector of responses on the ith cluster (patient, household, school, etc). Assuming clusters are independent,

$$\text{cov}(\mathbf{Y}) = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{V}_K \end{pmatrix},$$

where we might assume a common variance σ^2 and common pairwise correlation ρ within a cluster, i.e. an exchangeable correlation structure:

$$cov(\mathbf{Y}_i) = \sigma^2\mathbf{V}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}_{n_i \times n_i}$$

The solution to the problem when $\mathrm{cov}(\mathbf{Y}) \neq \sigma^2\mathbf{I}$, is to transform the model to a new set of observations that satisfy the constant variance assumption, and therafter use the ordinary least squares to estimate the parameters. Since $\sigma^2\mathbf{V}$ is a variance-covariance matrix, $\mathbf{V}$ is a symmetric non-singular matrix, and we can write: $\mathbf{V} = \mathbf{KK}$, where $\mathbf{K}$ is called the squared root of $\mathbf{V}$. Using this matrix, let $\tilde{\mathbf{y}} = \mathbf{K}^{-1}\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{K}^{-1}\mathbf{X}$, and $\tilde{\boldsymbol{\varepsilon}} = \mathbf{K}^{-1}\boldsymbol{\varepsilon}$. Then, it holds that

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}},$$

where $E(\tilde{\boldsymbol{\varepsilon}}) = \mathbf{0}$ and $\mathrm{var}(\tilde{\boldsymbol{\varepsilon}}) = \sigma^2\mathbf{I}$. We are now back to the assumptions of ordinary least squares. The least squares function can be given by

$$\begin{aligned} f(\boldsymbol{\beta}) &= ||\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}||^2 \\ &= (\mathbf{K}^{-1}\mathbf{y} - \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta})'(\mathbf{K}^{-1}\mathbf{y} - \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{K}^{-1}\mathbf{K}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

To minimize $f(\boldsymbol{\beta})$, begin by taking the derivative with respect to $\boldsymbol{\beta}$ and set the results equal to 0. This gives the normal equations:

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Thus, least squares applied to the transformed $\mathbf{Y}$ yields

$$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},$$

the Generalized Least Squares (GLS) estimate.

Theorem: 11.2 Properties of $\boldsymbol{\beta}^*$:

(a) $E[\boldsymbol{\beta}^*] = \boldsymbol{\beta}$,

(b) $\mathrm{cov}(\boldsymbol{\beta}^*) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$,

Let $\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ be the generalized least squares (GLS) estimate, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ be the ordinary least squares (OLS) estimate.

Theorem: 11.3 (Optimality of GLS estimates). If $E[Y] = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{cov}(\mathbf{Y}) = \sigma^2\mathbf{V}$, then for any constant vector $\mathbf{a}$, $\mathbf{a}'\boldsymbol{\beta}^*$ is the BLUE of $\mathbf{a}'\boldsymbol{\beta}$.

Example: 11.4 (Weighted least squares). Let $Y_1, \ldots, Y_n$ be independent, $E[Y_i] = \beta x_i$, and $\text{var}(Y_i) = \sigma^2 w_i^{-1}$. The GLS estimate of β is

$$\boldsymbol{\beta}^* = \frac{\sum_{i=1}^n w_i x_i Y_i}{\sum_{i=1}^n w_i x_i^2}.$$

The OLS estimate is

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

The variances are

$$\text{var}(\boldsymbol{\beta}^*) = \frac{\sigma^2}{\sum_{i=1}^n w_i x_i^2} \qquad \text{and} \qquad \text{var}(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2 \sum_{i=1}^n \frac{x_i^2}{w_i}}{(\sum_{i=1}^n x_i^2)^2}.$$

Theorem: 11.5 The GLS estimate and the OLS estimate are equal only when either one of the following conditions holds:

1. $\mathcal{R}(\mathbf{V}^{-1}\mathbf{X}) = \mathcal{R}(\mathbf{X})$.
2. $\mathcal{R}(\mathbf{V}\mathbf{X}) = \mathcal{R}(\mathbf{X})$.