Efficiency of the OLS Estimator

For scalar parameters, one estimator is more **efficient** than another if the precision of the former is greater than that of the latter.

Let $\ddot{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ be unbiased estimators of a k-vector of parameters $\boldsymbol{\beta}$, with covariance matrices $\text{Var}(\ddot{\boldsymbol{\beta}})$ and $\text{Var}(\hat{\boldsymbol{\beta}})$, respectively.

- $\hat{\boldsymbol{\beta}}$ is said to be more efficient than $\ddot{\boldsymbol{\beta}}$ if and only if $\text{Var}(\hat{\boldsymbol{\beta}})^{-1} - \text{Var}(\ddot{\boldsymbol{\beta}})^{-1}$ is a non-zero positive semidefinite matrix.
- If $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive definite matrices, then $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite if and only if $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$ is positive semidefinite.
- Thus $\hat{\boldsymbol{\beta}}$ is more efficient than $\ddot{\boldsymbol{\beta}}$ if and only if $\text{Var}(\ddot{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}})$ is a non-zero positive semidefinite matrix.
- If $\hat{\boldsymbol{\beta}}$ is more efficient than $\ddot{\boldsymbol{\beta}}$, every element of $\boldsymbol{\beta}$, and every linear combination of them, is estimated at least as efficiently by using $\hat{\boldsymbol{\beta}}$ as by using $\ddot{\boldsymbol{\beta}}$.

Consider $\gamma = \boldsymbol{w}^\top \boldsymbol{\beta}$.

$$\begin{aligned} \operatorname{Var}(\ddot{\gamma}) - \operatorname{Var}(\hat{\gamma}) &= \boldsymbol{w}^\top \operatorname{Var}(\ddot{\boldsymbol{\beta}})\boldsymbol{w} - \boldsymbol{w}^\top \operatorname{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{w} \\ &= \boldsymbol{w}^\top \big(\operatorname{Var}(\ddot{\boldsymbol{\beta}}) - \operatorname{Var}(\hat{\boldsymbol{\beta}})\big)\boldsymbol{w}. \end{aligned} \tag{1}$$

This must be either positive or zero if $\operatorname{Var}(\ddot{\boldsymbol{\beta}}) - \operatorname{Var}(\hat{\boldsymbol{\beta}})$ is positive semidefinite. Thus $\operatorname{Var}(\hat{\gamma}) \leq \operatorname{Var}(\ddot{\gamma})$ when $\hat{\boldsymbol{\beta}}$ is more efficient than $\ddot{\boldsymbol{\beta}}$.

Let $\hat{\boldsymbol{\beta}}$ denote the OLS estimator and $\ddot{\boldsymbol{\beta}}$ some other linear estimator.

An estimator is **linear** if we can write it as a linear function of $\boldsymbol{y}$. $\hat{\boldsymbol{\beta}}$ is linear, because it is equal to the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ times the vector $\boldsymbol{y}$.

For any linear estimator that is not the OLS estimator,

$$\ddot{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{C}\boldsymbol{y}, \tag{2}$$

where $\boldsymbol{A}$ and $\boldsymbol{C}$ are $k \times N$ matrices that depend on $\boldsymbol{X}$.

To obtain the second equality in (2), define

$$\boldsymbol{C} \equiv \boldsymbol{A} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top. \tag{3}$$

Alternative linear estimators include **instrumental variables** and **generalized least squares**.

The **Gauss-Markov Theorem** says that $\hat{\boldsymbol{\beta}}$ is the **best linear unbiased estimator**, or **BLUE**. We are comparing it with every other unbiased estimator $\ddot{\boldsymbol{\beta}}$ of the form (2).

If $\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top \mid \boldsymbol{X}) = \sigma^2 \mathbf{I}$, then $\mathrm{Var}(\ddot{\boldsymbol{\beta}}) - \mathrm{Var}(\hat{\boldsymbol{\beta}})$ is a positive semidefinite matrix.

Substituting for $\boldsymbol{y}$ in (2), we find that

$$\ddot{\boldsymbol{\beta}} = \boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{A}\boldsymbol{u}. \tag{4}$$

Since we want $\ddot{\boldsymbol{\beta}}$ to be unbiased, the expectation of the rightmost expression in (4), conditional on $\boldsymbol{X}$, must be $\boldsymbol{\beta}_0$.

Since $\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{X}) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{A}\boldsymbol{u} \mid \boldsymbol{x}) = \mathbf{0}$, and, by the law of iterated expectations, $\mathrm{E}(\boldsymbol{A}\boldsymbol{u}) = \mathbf{0}$.

So the key condition that $\boldsymbol{A}$, and thus $\boldsymbol{C}$, must satisfy for $\ddot{\boldsymbol{\beta}}$ to be unbiased is that $\mathrm{E}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}_0 \mid \boldsymbol{X}) = \boldsymbol{\beta}_0$.

This is the case for all $\boldsymbol{\beta}_0$ if and only if $\boldsymbol{A}\boldsymbol{X} = \mathbf{I}$.

Equivalently, it holds for all $\boldsymbol{\beta}_0$ whenever $\boldsymbol{C}\boldsymbol{X} = \mathbf{O}$, since

$$\boldsymbol{C}\boldsymbol{X} = \boldsymbol{A}\boldsymbol{X} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{A}\boldsymbol{X} - \mathbf{I}. \tag{5}$$

The condition that $\boldsymbol{C}\boldsymbol{X} = \mathbf{O}$ implies that $\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\boldsymbol{u}$. Since $\boldsymbol{C}\boldsymbol{y} = \ddot{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$, this makes it clear that $\ddot{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$ has conditional expectation zero.

The unbiasedness condition on $\ddot{\boldsymbol{\beta}}$ also implies that the covariances of $\ddot{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}$ are all zero.

In other words, $\ddot{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \boldsymbol{v}$, where $\boldsymbol{v}$ is uncorrelated with $\hat{\boldsymbol{\beta}}$.

Since the covarances of $\ddot{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}$ are zero,

$$\begin{aligned} \mathrm{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\ddot{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^\top\big) &= \mathrm{E}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{C}^\top\big) && (6)\\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\sigma_0^2\mathbf{I}\boldsymbol{C}^\top && (7)\\ &= \sigma_0^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{C}^\top = \mathbf{O}. && (8) \end{aligned}$$

Once again, we see that $\boldsymbol{C}\boldsymbol{X} = \mathbf{O}$, so that $\mathrm{E}(\boldsymbol{C}\boldsymbol{y}) = \mathrm{E}(\boldsymbol{C}\boldsymbol{u}) = \mathbf{0}$.

To complete the proof, note that

$$\begin{aligned} \mathrm{Var}(\ddot{\boldsymbol{\beta}}) &= \mathrm{Var}\left(\hat{\boldsymbol{\beta}} + (\ddot{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\right) \\ &= \mathrm{Var}\left(\hat{\boldsymbol{\beta}} + \boldsymbol{C}\boldsymbol{y}\right) && (9)\\ &= \mathrm{Var}(\hat{\boldsymbol{\beta}}) + \mathrm{Var}(\boldsymbol{C}\boldsymbol{y}). \end{aligned}$$

The difference between $\mathrm{Var}(\ddot{\boldsymbol{\beta}})$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is $\mathrm{Var}(\boldsymbol{C}\boldsymbol{y})$, which must be a positive semidefinite matrix.

Any unbiased linear estimator $\ddot{\boldsymbol{\beta}}$ equals $\hat{\boldsymbol{\beta}}$ plus $\boldsymbol{Cy}$, which has expectation zero and is uncorrelated with $\hat{\boldsymbol{\beta}}$:

$$\ddot{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \boldsymbol{Cy} = \hat{\boldsymbol{\beta}} + \boldsymbol{v}. \tag{10}$$

- The Gauss-Markov theorem requires the disturbances to be independent and homoskedastic, but they do not have to be normally distributed.
- It applies only to a correctly specified model with exogenous regressors and disturbances with a scalar covariance matrix.
- It does *not* say that $\hat{\boldsymbol{\beta}}$ is more efficient than every imaginable estimator. Nonlinear and/or biased estimators may well perform better than OLS.

Something very similar to (10) holds asymptotically for estimators that are asymptotically efficient and asymptotically unbiased.

Residuals and Disturbances

Least-squares residuals $\hat{\boldsymbol{u}} \equiv \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ have both numerical and statistical properties.

Numerical: $\hat{\boldsymbol{u}}$ is orthogonal to every vector that lies in $\mathcal{S}(\boldsymbol{X})$.

Consistency of $\hat{\boldsymbol{\beta}}$ implies that $\hat{\boldsymbol{u}} \to \boldsymbol{u}$ as $N \to \infty$.

$$\boldsymbol{M}_X \boldsymbol{y} = \boldsymbol{M}_X \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{M}_X \boldsymbol{u} = \boldsymbol{M}_X \boldsymbol{u} = \hat{\boldsymbol{u}}. \tag{11}$$

Each residual is a linear combination of every one of the disturbances:

$$\hat{u}_i = u_i - \boldsymbol{X}_i (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} \tag{12}$$

$$= u_i - \boldsymbol{X}_i (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sum_{j=1}^{N} \boldsymbol{X}_j{}^\top u_j. \tag{13}$$

Even when the u_i are independent, the $\hat{u}_i$ are not independent.

Now assume that $\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{X}) = \mathbf{0}$, so that $\mathrm{E}(u_i \mid \boldsymbol{X}) = 0$ for all i.

Since, by (13), $\hat{u}_i$ is just a linear combination of all the u_i, $\mathrm{E}(\hat{u}_i \mid \boldsymbol{X}) = 0$. Therefore, $\mathrm{Var}(\hat{u}_i)$ is just $\mathrm{E}(\hat{u}_i^2)$.

We know that $\|\hat{\boldsymbol{u}}\| < \|\boldsymbol{u}(\boldsymbol{\beta})\| = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|$ for any $\boldsymbol{\beta} \neq \hat{\boldsymbol{\beta}}$. In particular, $\|\hat{\boldsymbol{u}}\| \leq \|\boldsymbol{u}(\boldsymbol{\beta}_0)\|$.

This implies that $\mathrm{E}\big(\|\hat{\boldsymbol{u}}\|^2\big) \leq \mathrm{E}\big(\|\boldsymbol{u}\|^2\big)$. If $\mathrm{Var}(u_i) = \sigma_0^2$, then

$$\sum_{i=1}^{N} \mathrm{Var}(\hat{u}_i) = \sum_{i=1}^{N} \mathrm{E}(\hat{u}_i^2) = \mathrm{E}\left(\sum_{i=1}^{N} \hat{u}_i^2 \right) = \mathrm{E}\big(\|\hat{\boldsymbol{u}}\|^2\big) \tag{14}$$

$$\leq \mathrm{E}\big(\|\boldsymbol{u}\|^2\big) = \mathrm{E}\left(\sum_{i=1}^{N} u_i^2 \right) = \sum_{i=1}^{N} \mathrm{E}(u_i^2) = N\sigma_0^2. \tag{15}$$

This suggests that, at least for most observations, $\mathrm{Var}(\hat{u}_i)$ must be less than σ_0^2. In fact, $\mathrm{Var}(\hat{u}_i)$ is less than σ_0^2 for *every* observation.

The covariance matrix of the entire vector $\hat{\boldsymbol{u}}$ is

$$\begin{aligned} \operatorname{Var}(\hat{\boldsymbol{u}}) &= \operatorname{Var}(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}) = \operatorname{E}(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}\boldsymbol{u}^{\top}\boldsymbol{M}_{\boldsymbol{X}}) && (16)\\ &= \boldsymbol{M}_{\boldsymbol{X}}\operatorname{E}(\boldsymbol{u}\boldsymbol{u}^{\top})\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{M}_{\boldsymbol{X}}\operatorname{Var}(\boldsymbol{u})\boldsymbol{M}_{\boldsymbol{X}} && (17)\\ &= \boldsymbol{M}_{\boldsymbol{X}}(\sigma_0^2\mathbf{I})\boldsymbol{M}_{\boldsymbol{X}} = \sigma_0^2\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{X}} = \sigma_0^2\boldsymbol{M}_{\boldsymbol{X}}. && (18) \end{aligned}$$

This uses the facts that $\operatorname{E}(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}) = \mathbf{0}$ and that $\boldsymbol{M}_{\boldsymbol{X}}$ is idempotent.

From (18), $\operatorname{E}(\hat{u}_i\hat{u}_j) \neq 0$ for $i \neq j$. Residuals are correlated even when disturbances are uncorrelated.

Residuals do not have constant variance, and the variance of every residual must always be smaller than σ_0^2.

A typical diagonal element of $\boldsymbol{M}_{\boldsymbol{X}}$ is $1 - h_i$. Therefore, it follows from (18) that

$$\operatorname{Var}(\hat{u}_i) = \operatorname{E}(\hat{u}_i^2) = (1 - h_i)\sigma_0^2. \tag{19}$$

Since $0 \leq 1 - h_i < 1$, we see that $\operatorname{E}(\hat{u}_i^2) < \sigma_0^2$.

Estimating the Variance of the Disturbances

(19) tells us that high-leverage observations, for which h_i is relatively large, must have residuals with unusually small variances.

The method of moments suggests that we can estimate σ^2 by using the corresponding sample moment.

If we observed the u_i, this sample moment would be

$$\frac{1}{N}\sum_{i=1}^{N} u_i^2. \tag{20}$$

In fact, we only observe the $\hat{u}_i$, so a natural MM estimator is

$$\hat{\sigma}^2 \equiv \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i^2. \tag{21}$$

Because $\mathrm{E}(\hat{u}_i^2) < \sigma_0^2$, by (19), $\hat{\sigma}^2$ must be biased downward.

We know that $\sum_{i=1}^{N} h_i = k$. Therefore,

$$\mathrm{E}(\hat{\sigma}^2) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}(\hat{u}_i^2) = \frac{1}{N}\sum_{i=1}^{N}(1-h_i)\,\sigma_0^2 = \frac{N-k}{N}\,\sigma_0^2. \tag{22}$$

Since $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$ and $\boldsymbol{M}_{\boldsymbol{X}}$ is idempotent, the SSR is just $\boldsymbol{u}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, and

$$\mathrm{E}\big(\boldsymbol{u}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}\big) = \mathrm{E}\big(\mathrm{SSR}(\hat{\boldsymbol{\beta}})\big) = \mathrm{E}\left(\sum_{i=1}^{N}\hat{u}_i^2\right) = (N-k)\,\sigma_0^2. \tag{23}$$

Adding one more regressor has exactly the same effect on the expectation of the SSR as taking away one observation.

The result (23) suggests another MM estimator which is unbiased whenever $\hat{\boldsymbol{\beta}}$ is:

$$s^2 \equiv \frac{1}{N-k}\sum_{i=1}^{N}\hat{u}_i^2. \tag{24}$$

The square root of s^2 is s, the **standard error of the regression** (or **regression standard error**). Even though s^2 provides an unbiased estimate of σ^2, s does not provide an unbiased estimate of σ.

An unbiased estimator for $\text{Var}(\hat{\boldsymbol{\beta}})$ is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{25}$$

This is the usual estimator of $\text{Var}(\hat{\boldsymbol{\beta}})$ under the assumption of IID disturbances.

- Suppose that $\boldsymbol{X}$ is fixed across samples. Then $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is the same for every sample, but $s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ varies.
- However, $s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is always proportional to $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. Like the latter, it is normally $O(N^{-1})$
- Evidently, $s^2 = O_p(1)$, and $s^2/\sigma_0^2 \to 1$ as $N \to \infty$.

Over-specification

A regression model is **over-specified** if some variables that belong to Ω_i, but do not appear in the DGP, are mistakenly included.

Since the DGP remains a special case of the model, there is no actual misspecification, merely a failure to incorporate zero restrictions that are true.

Consider the over-specified linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{26}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ belong to Ω_i. The data are actually generated by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_0^2\mathbf{I}). \tag{27}$$

The DGP (27) is a special case of (26), with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma} = \boldsymbol{0}$, and $\sigma^2 = \sigma_0^2$.

The estimates $\hat{\boldsymbol{\beta}}$ from (26) are the same as those from the FWL regression

$$\boldsymbol{M_Z y} = \boldsymbol{M_Z X \beta} + \text{residuals}, \tag{28}$$

where, as usual, $\boldsymbol{M_Z} = \mathbf{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$.

Thus we see that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{M_Z X})^{-1} \boldsymbol{X}^\top \boldsymbol{M_Z y}. \tag{29}$$

$\hat{\boldsymbol{\beta}}$ must be unbiased if $\boldsymbol{X}$ and $\boldsymbol{Z}$ are exogenous.

If we replace $\boldsymbol{y}$ by $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$, we find that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{M_Z X})^{-1} \boldsymbol{X}^\top \boldsymbol{M_Z u}. \tag{30}$$

The second term on the r.h.s. has conditional mean $\mathbf{0}$, provided we take expectations conditional on $\boldsymbol{Z}$ as well as on $\boldsymbol{X}$.

Imposing the restriction that $\boldsymbol{\gamma} = \mathbf{0}$ yields $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, which is also unbiased because the restriction is true.

We now have two unbiased linear estimators, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$.

When the restriction holds, the matrix $\mathrm{Var}(\hat{\boldsymbol{\beta}}) - \mathrm{Var}(\tilde{\boldsymbol{\beta}})$ must be positive semidefinite. We know that

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) = \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{31}$$

It is easily shown that

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big) && (32)\\
&= (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} && (33)\\
&= \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \mathbf{I} \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} && (34)\\
&= \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1}. && (35)
\end{aligned}$$

When there is only one regressor, $\boldsymbol{x}$, and one parameter, β, it is easy to see that $\tilde{\beta}$ is more efficient than $\hat{\beta}$.

Since $\boldsymbol{M_Z}$ is a projection matrix, $\|\boldsymbol{M_Z x}\|$ must be smaller (or at least, no larger) than $\|\boldsymbol{x}\|$. Thus $\boldsymbol{x}^\top \boldsymbol{M_Z x} \leq \boldsymbol{x}^\top \boldsymbol{x}$, which implies that

$$\sigma_0^2 (\boldsymbol{x}^\top \boldsymbol{M_Z x})^{-1} \geq \sigma_0^2 (\boldsymbol{x}^\top \boldsymbol{x})^{-1}. \tag{36}$$

Notice that $\text{Var}(\hat{\beta})$ is proportional to the inverse of the SSR from a regression of $\boldsymbol{x}$ on the other regressors.

Showing that $\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\tilde{\boldsymbol{\beta}})$ is positive semidefinite is equivalent to showing that $\text{Var}(\tilde{\boldsymbol{\beta}})^{-1} - \text{Var}(\hat{\boldsymbol{\beta}})^{-1}$ is positive semidefinite:

$$\begin{aligned} \boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{X}^\top \boldsymbol{M_Z X} &= \boldsymbol{X}^\top (\mathbf{I} - \boldsymbol{M_Z}) \boldsymbol{X} \\ &= \boldsymbol{X}^\top \boldsymbol{P_Z X} = (\boldsymbol{P_Z X})^\top \boldsymbol{P_Z X}. \end{aligned} \tag{37}$$

This is the transpose of a matrix times itself. Thus $\sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{M_Z X})^{-1} - \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ must be positive semidefinite.

Adding additional variables that do not really belong in a model leads to less accurate estimates unless $\boldsymbol{P_Z X} = \mathbf{O}$, so that $\boldsymbol{M_Z X} = \boldsymbol{X}$.

We used the FWL theorem to show that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}$.

If we were unaware of this theorem, we would need to write

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{Z} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{y} \\ \boldsymbol{Z}^\top \boldsymbol{y} \end{bmatrix}. \tag{38}$$

Then we would have to invert the matrix in (38) and multiply the first row of the inverse by the vector that follows it.

The inverse is (not easily!) seen to be

$$\begin{bmatrix} (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} & -(\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \\ -(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} & (\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{-1} \end{bmatrix}. \tag{39}$$

This leads directly to $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}$.

To obtain $\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$, it is easier to write the lower left-hand submatrix as $-(\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

Under-specification

Now suppose the DGP is really

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_0^2 \mathbf{I}). \tag{40}$$

The estimator $\hat{\boldsymbol{\beta}}$ is now the "correct" one to use.

If instead we use $\tilde{\boldsymbol{\beta}}$, there really is misspecification, and the restricted estimator $\tilde{\boldsymbol{\beta}}$ is biased.

$$\begin{aligned} \text{E}(\tilde{\boldsymbol{\beta}}) &= \text{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u})\big) && (41) \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0 + \text{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\big) && (42) \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0. && (43) \end{aligned}$$

The second term in (43) equals zero only when $\boldsymbol{X}^\top \boldsymbol{Z} = \mathbf{O}$ or $\boldsymbol{\gamma}_0 = \boldsymbol{0}$.

In all other cases, $\tilde{\boldsymbol{\beta}}$ is biased. The magnitude of the bias depends on γ_0 and on the $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices.

This bias does not vanish as $N \to \infty$, so $\tilde{\boldsymbol{\beta}}$ is also generally inconsistent.

Since $\tilde{\boldsymbol{\beta}}$ is biased, we cannot use $\text{Var}(\tilde{\boldsymbol{\beta}})$ to evaluate its accuracy. Instead, we can use the **mean squared error matrix**, or **MSE matrix**:

$$\text{MSE}(\tilde{\boldsymbol{\beta}}) \equiv \text{E}\big((\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big). \tag{44}$$

The MSE matrix is equal to $\text{Var}(\tilde{\boldsymbol{\beta}})$ if $\tilde{\boldsymbol{\beta}}$ is unbiased, but not otherwise.

For a scalar parameter β,

$$\text{MSE}(\tilde{\beta}) = \text{Var}(\tilde{\beta}) + \big(\text{E}(\tilde{\beta}) - \beta_0\big)^2. \tag{45}$$

In this case, it is common to report the **root mean squared error**, or **RMSE**, instead of the MSE.

It is easy to see that

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{\gamma}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}. \tag{46}$$

Therefore, $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ times itself transposed is equal to

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{\gamma}_0 \boldsymbol{\gamma}_0{}^\top \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} \boldsymbol{u}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{47}$$

$$+ (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{\gamma}_0 \boldsymbol{u}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} \boldsymbol{\gamma}_0{}^\top \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{48}$$

The second term here has expectation $\sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and the third and fourth terms have expectation zero. Thus

$$\text{MSE}(\tilde{\boldsymbol{\beta}}) = \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{\gamma}_0 \boldsymbol{\gamma}_0{}^\top \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{49}$$

The first term here is $O(N^{-1})$, and the second is $O(1)$.

We would like to compare $\text{MSE}(\tilde{\boldsymbol{\beta}})$ with $\text{MSE}(\hat{\boldsymbol{\beta}}) = \text{Var}(\hat{\boldsymbol{\beta}})$.

If the bias is small, the second term in (49) must be small, and in that case $\tilde{\boldsymbol{\beta}}$ may well have smaller MSE than $\hat{\boldsymbol{\beta}}$.

However, if the bias is large, the second term in (49) is necessarily large, and $\tilde{\boldsymbol{\beta}}$ must have larger MSE than $\hat{\boldsymbol{\beta}}$.

$\tilde{\boldsymbol{\beta}}$ may yield smaller MSE for some parameters and $\hat{\boldsymbol{\beta}}$ for others.

$\widehat{\text{Var}}(\tilde{\boldsymbol{\beta}})$ calculated by a least-squares regression program attempts to estimate the first term in (49), but it ignores the second.

Because s^2 is typically larger than σ_0^2 if some regressors have been incorrectly omitted, this estimate is biased.

- Under-specification causes bias and inconsistency, but over-specification "merely" causes inefficiency. Which problem is more severe?
- It depends on how much information the sample contains. In sufficiently large samples, avoid under-specification at all costs.
- However, in samples of modest size, the gain in efficiency from omitting some variables may be very large relative to the bias that is caused by their omission. There is a **bias-variance tradeoff**.

Measures of Goodness of Fit

The most commonly used (and misused) measure of goodness of fit is the **coefficient of determination**, or $\boldsymbol{R^2}$.

The **uncentered $\boldsymbol{R^2}$**, denoted R_u^2, is the ratio of the explained sum of squares (ESS) of the regression to the total sum of squares (TSS).

As a consequence of Pythagoras' Theorem,

$$\text{TSS} = \|\boldsymbol{y}\|^2 = \|\boldsymbol{P_X y}\|^2 + \|\boldsymbol{M_X y}\|^2 = \text{ESS} + \text{SSR}. \tag{50}$$

Therefore,

$$R_u^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\|\boldsymbol{P_X y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\|\boldsymbol{M_X y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\text{SSR}}{\text{TSS}} = \cos^2\theta, \tag{51}$$

where θ is the angle between $\boldsymbol{y}$ and $\boldsymbol{P_X y}$.

- For any angle θ, $-1 \leq \cos\theta \leq 1$. Consequently, $0 \leq R_u^2 \leq 1$.
- If $\theta = 0$, $\boldsymbol{y}$ and $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ would coincide, the residual vector $\hat{\boldsymbol{u}}$ would vanish, and we would have a **perfect fit**, with $R_u^2 = 1$.
- If $R_u^2 = 0$, the fitted value vector would vanish, and $\boldsymbol{y}$ would coincide with $\hat{\boldsymbol{u}}$.
- R_u^2 depends on the data only through residuals and fitted values. It is invariant to nonsingular linear transformations of $\boldsymbol{X}$.
- Because it is a ratio, R_u^2 is invariant to changes in the scale of $\boldsymbol{y}$.

The **centered R^2**, denoted R_c^2, is much more commonly encountered than R_u^2. All variables are centered, that is, expressed as deviations from their means, before ESS and TSS are calculated.

By adding a large enough constant to all the y_i, we could make R_u^2 become arbitrarily close to 1, since the SSR would stay the same and the TSS would increase without limit.

But R_c^2 is invariant to changes in the mean of the regressand.

- Both versions of R^2 are valid only if a regression model is estimated by least squares. Only then is TSS = ESS + SSR.
- The centered version is not valid if the regressors do not include a constant term (or equivalent), i.e. if $\boldsymbol{\iota}$ does not belong to $\mathcal{S}(\boldsymbol{X})$.
- Both R_u^2 and R_c^2 increase whenever more regressors are added.

Consider restricted and unrestricted models with the same dependent variable. They have the same TSS, so the regression with the larger ESS (smaller SSR) must also have the larger R^2.

The ESS from the unrestricted model is $\|\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}}\,\boldsymbol{y}\|^2$, and the ESS from the restricted model is $\|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}\|^2$. The difference between them is

$$\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{X}})\,\boldsymbol{y}. \tag{52}$$

Since the matrix $\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{X}}$ is an orthogonal projection matrix, (52) must be non-negative.

Why is $\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{X}}$ an orthogonal projection matrix? Because $\mathcal{S}(\boldsymbol{X}) \subset \mathcal{S}(\boldsymbol{X},\boldsymbol{Z})$. This implies that $\boldsymbol{P}_{\boldsymbol{X}}$ projects on to a subspace of the image of $\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}}$.

We conclude that the ESS, and hence the R^2, from the unrestricted model can be no less than those from the restricted model.

The R^2 can be modified so that adding additional regressors does not necessarily increase its value.

If $\boldsymbol{\iota} \in \mathcal{S}(\boldsymbol{X})$, the centered R^2 can be written as

$$R_c^2 = 1 - \frac{\sum_{i=1}^N \hat{u}_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \tag{53}$$

The SSR has expectation $(N-k)\sigma_0^2$ under standard assumptions. The denominator is $N-1$ times an unbiased estimator of the variance of y_i about its true mean. As such, it has expectation $(N-1)\operatorname{Var}(y)$.

Thus the second term of (53) can be thought of as the ratio of two biased estimators.

If we replace these by unbiased ones, we obtain the **adjusted R^2**,

$$\bar{R}^2 \equiv 1 - \frac{\frac{1}{N-k}\sum_{i=1}^{N}\hat{u}_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{(N-1)\boldsymbol{y}^\top \boldsymbol{M}_X \boldsymbol{y}}{(N-k)\boldsymbol{y}^\top \boldsymbol{M}_\iota \boldsymbol{y}}. \tag{54}$$

$\bar{R}^2$ and R_c^2 are generally very similar, except when $(N-k)/(N-1) << 1$.

One nice feature of R_u^2 and R_c^2 is that they are constrained to lie between 0 and 1.

In contrast, $\bar{R}^2$ can actually be negative when $(N-1)/(N-k)$ is greater than TSS/SSR.

- Never compare any form of R^2 for models that are estimated using different datasets!
- Models with high R^2 can be complete nonsense.