Covariance Matrices and Precision Matrices

The matrix of central second moments of a random vector $\boldsymbol{b}$ is $\mathrm{Var}(\boldsymbol{b})$. This is called the **covariance matrix**, **variance matrix**, or **variance-covariance matrix**,

$$\mathrm{Var}(\boldsymbol{b}) \equiv \mathrm{E}\Big(\big(\boldsymbol{b} - \mathrm{E}(\boldsymbol{b})\big)\big(\boldsymbol{b} - \mathrm{E}(\boldsymbol{b})\big)^\top\Big). \tag{1}$$

When $\mathrm{E}(\boldsymbol{b}) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{b}) = \mathrm{E}(\boldsymbol{b}\boldsymbol{b}^\top)$.

The i^{th} diagonal element of $\mathrm{Var}(\boldsymbol{b})$ is $\mathrm{Var}(b_i) = \mathrm{E}\big(b_i - \mathrm{E}(b_i)\big)^2$.

The ij^{th} off-diagonal element of $\mathrm{Var}(\boldsymbol{b})$ is

$$\mathrm{Cov}(b_i, b_j) \equiv \mathrm{E}\Big(\big(b_i - \mathrm{E}(b_i)\big)\big(b_j - \mathrm{E}(b_j)\big)\Big). \tag{2}$$

If $i = j$, $\mathrm{Cov}(b_i, b_j) = \mathrm{Var}(b_i)$.

Since $\mathrm{Cov}(b_i, b_j) = \mathrm{Cov}(b_j, b_i)$, $\mathrm{Var}(\boldsymbol{b})$ must be a symmetric matrix.

If b_i and b_j are statistically independent, then $\mathrm{Cov}(b_i, b_j) = 0$. The converse is not true, however.

The inverse of a covariance matrix is a **precision matrix**.

The **correlation** between b_i and b_j is

$$\rho(b_i, b_j) \equiv \frac{\mathrm{Cov}(b_i, b_j)}{\big(\mathrm{Var}(b_i)\,\mathrm{Var}(b_j)\big)^{1/2}}, \quad \text{with } -1 \leq \rho(b_i, b_j) \leq 1. \tag{3}$$

Correlations can be arranged into a symmetric **correlation matrix** with all elements on the principal diagonal equal to 1.

$\mathrm{Var}(\boldsymbol{b})$ must be **positive semidefinite** as well as symmetric.

In most cases, covariance matrices and correlation matrices are actually **positive definite**.

A $k \times k$ symmetric matrix $\boldsymbol{A}$ is positive definite if, for all nonzero k-vectors $\boldsymbol{x}$, the matrix product $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$ is positive.

$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ is called a **quadratic form**.

A quadratic form always involves a k-vector, in this case $\boldsymbol{x}$, and a $k \times k$ matrix, in this case $\boldsymbol{A}$.

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{k}\sum_{j=1}^{k} x_i x_j A_{ij}. \tag{4}$$

If this can be zero but not negative, then $\boldsymbol{A}$ is said to be positive semidefinite.

Any matrix of the form $\boldsymbol{B}^\top \boldsymbol{B}$ is positive semidefinite. $\boldsymbol{B}^\top \boldsymbol{B}$ is symmetric and, for any nonzero $\boldsymbol{x}$,

$$\boldsymbol{x}^\top \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{x} = (\boldsymbol{B}\boldsymbol{x})^\top (\boldsymbol{B}\boldsymbol{x}) = \|\boldsymbol{B}\boldsymbol{x}\|^2 \geq 0. \tag{5}$$

This is positive unless $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$. In that case, since $\boldsymbol{x} \neq \boldsymbol{0}$, the columns of $\boldsymbol{B}$ are linearly dependent, and $\boldsymbol{B}$ does not have **full column rank**.

- $\boldsymbol{B}$ can have full rank but not full column rank if $\boldsymbol{B}$ has fewer rows than columns. If so, maximum possible rank is number of rows.
- When $\boldsymbol{B}$ does have full column rank, $\boldsymbol{B}^\top\boldsymbol{B}$ is positive definite.
- If $\boldsymbol{A}$ is positive definite, then any matrix $\boldsymbol{B}^\top\boldsymbol{A}\boldsymbol{B}$ is positive definite if $\boldsymbol{B}$ has full column rank, positive semidefinite otherwise.
- A positive definite matrix cannot be singular, because, if $\boldsymbol{A}$ is singular, there must exist a nonzero $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. But then $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = 0$ as well, which means that $\boldsymbol{A}$ is not positive definite.
- Diagonal elements of a positive definite matrix must all be positive. Suppose that A_{22} were negative. Then, if we chose $\boldsymbol{x}$ to be the vector $\boldsymbol{e}_2$, the quadratic form would just be $\boldsymbol{e}_2^\top\boldsymbol{A}\boldsymbol{e}_2 = A_{22} < 0$.
- For a positive semidefinite matrix, diagonal elements may be 0.
- The off-diagonal elements of $\boldsymbol{A}$ may be of either sign.
- The inverse of a positive definite matrix always exists, and it is positive definite.

The identity matrix, $\mathbf{I}$, is a positive definite matrix, since

$$\boldsymbol{x}^\top \mathbf{I}\, \boldsymbol{x} = \sum_{i=1}^{k} x_i^2. \tag{6}$$

If the $k \times k$ matrix $\boldsymbol{A}$ is symmetric and positive definite, then there always exists a full-rank $k \times k$ matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}^\top \boldsymbol{B}$.

Think of $\boldsymbol{B}$ as a **square root matrix**. For any matrix $\boldsymbol{A}$, the corresponding matrix $\boldsymbol{B}$ is not unique.

In particular, $\boldsymbol{B}$ can be chosen to be symmetric, but it can also be chosen to be upper or lower triangular.

We can compute a triangular $\boldsymbol{B}$ using the **Cholesky decomposition**.

Since a triangular matrix is extremely easy to invert, we can use the Cholesky decomposition to find $\boldsymbol{A}^{-1} = \boldsymbol{B}^{-1}(\boldsymbol{B}^\top)^{-1}$.

Finding a symmetric $\boldsymbol{B}$ is a lot more work. It involves finding the eigenvectors of $\boldsymbol{A}$.

The OLS Covariance Matrix

The covariance matrix of $\hat{\boldsymbol{\beta}}$ depends on the covariance matrix of the disturbances (error terms).

If the disturbances are IID, the covariance matrix of $\boldsymbol{u}$ is the **scalar matrix** $\sigma^2\mathbf{I}$:

$$\text{Var}(\boldsymbol{u}) = \text{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2\mathbf{I}. \tag{7}$$

Note that (7) does not require the disturbances to be independent, or even to have the same distribution.

But they must all have the same variance, and the covariance of each pair of disturbances must be zero.

When every u_i has the same variance, then they are **homoskedastic**.

When the $\text{Var}(u_i)$ differ, then they are **heteroskedastic**.

In general, we denote the $N \times N$ error covariance matrix by $\boldsymbol{\Omega}$.

For time-series data, when $\mathbf{\Omega}$ has nonzero off-diagonal elements, the disturbances are said to be **autocorrelated** or **serially correlated**.

If the observations of a sample characterize different locations in space, they may display **spatial autocorrelation**.

Another possibility is that the disturbances are **clustered**, correlated within each of G clusters but uncorrelated across them.

When observations are sorted by cluster, $\mathbf{\Omega}$ is block-diagonal, with G diagonal blocks that correspond to the G clusters:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_1 & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \mathbf{\Omega}_2 & \dots & \mathbf{O} \\ \vdots & \vdots & & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{\Omega}_G \end{bmatrix}. \tag{8}$$

Here $\mathbf{\Omega}_g$ is the $N_g \times N_g$ covariance matrix for the observations belonging to the g^{th} cluster.

Disturbances that are autocorrelated or clustered may or may not also be heteroskedastic. The usual approach to cluster-robust inference also allows for heteroskedasticity of unknown form.

When the DGP belongs to the model we estimate,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}. \tag{9}$$

When $\hat{\boldsymbol{\beta}}$ is unbiased, $\text{Var}(\hat{\boldsymbol{\beta}})$ is the expectation of the $k \times k$ matrix

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{10}$$

If we take this expectation, conditional on $\boldsymbol{X}$, we find that

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \text{E}(\boldsymbol{u}\boldsymbol{u}^\top) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{11}$$

$$= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{12}$$

This form of covariance matrix is called a **sandwich covariance matrix**, because $\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}$ is sandwiched between the two instances of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

The diagonal elements of $\text{Var}(\hat{\boldsymbol{\beta}})$ are particularly interesting. The square root of the k^{th} diagonal element is the standard error of $\hat{\beta}_k$.

In practice, we almost never know $\boldsymbol{\Omega}$, so we have to figure out how to estimate the matrix $\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}$.

That is what methods for **heteroskedasticity-consistent** and **cluster-robust** covariance matrix estimation do.

If $\boldsymbol{\Omega} = \sigma_0^2 \mathbf{I}$, so that there is neither heteroskedasticity not autocorrelation, then equation (12) simplifies greatly. It becomes

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \sigma_0^2 \mathbf{I} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{13}$$

$$= \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{14}$$

For the next few lectures, we will assume that (14) holds.

But this simplification is often not valid, and most modern empirical work uses sandwich covariance matrices.

Precision of the Least-Squares Estimates

When (14) holds, the precision matrix is

$$\frac{1}{\sigma_0^2}\boldsymbol{X}^\top\boldsymbol{X}. \tag{15}$$

This is inversely proportional to σ_0^2. The more random variation there is in the disturbances, the less precise are the parameter estimates.

It is illuminating to rewrite (15) as

$$\frac{N}{\sigma_0^2}\left(\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}\right). \tag{16}$$

Under standard assumptions, $N^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ is $O(1)$.

Thus the precision matrix (16) must be $O(N)$. When $\boldsymbol{X}$ is stochastic, it is $O_p(N)$.

If we double the sample size, the precision of $\hat{\boldsymbol{\beta}}$ should roughly double, and the standard errors of the individual $\hat{\beta}_i$ will be, approximately, divided by $\sqrt{2}$.

Consider the model $\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \boldsymbol{u}$. Replacing $\boldsymbol{X}$ with $\boldsymbol{\iota}$, we find that

$$\hat{\beta}_1 = (\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top \boldsymbol{y} = \frac{1}{N} \sum_{i=1}^{N} y_i, \text{ and} \tag{17}$$

$$\text{Var}(\hat{\beta}_1) = \sigma_0^2 (\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1} = \frac{1}{N} \sigma_0^2. \tag{18}$$

The precision of the sample mean is exactly proportional to N, since the variance is proportional to $1/N$.

We have seen that σ_0^2 and N affect the precision of $\hat{\boldsymbol{\beta}}$. The third thing that does so is the matrix $\boldsymbol{X}$.

High-leverage observations affect it much more than average ones.

We can rewrite a linear regression model as

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \tag{19}$$

where $\boldsymbol{X}$ has been partitioned into $\boldsymbol{x}_1$ and $\boldsymbol{X}_2$.

By the FWL Theorem, (19) yields the same estimate of β_1 as the FWL regression

$$\boldsymbol{M}_2\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{x}_1\beta_1 + \text{residuals}, \tag{20}$$

where $\boldsymbol{M}_2 \equiv \mathbf{I} - \boldsymbol{X}_2(\boldsymbol{X}_2{}^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2{}^\top$.

This estimate is $\hat{\beta}_1 = \boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{y}/\boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{x}_1$, and its variance is $\sigma_0^2(\boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{x}_1)^{-1}$. Thus the precision of $\hat{\beta}_1$ is

$$\frac{1}{\sigma_0^2}\,\boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{x}_1. \tag{21}$$

How much information the sample gives us about β_1 is proportional to the squared Euclidean length of the vector $\boldsymbol{M}_2\boldsymbol{x}_1$.

When $\|\boldsymbol{M}_2\boldsymbol{x}_1\|$ is big, because N is large or at least some elements of $\boldsymbol{M}_2\boldsymbol{x}_1$ are large, $\hat{\beta}_1$ is relatively precise.

When $\|\boldsymbol{M}_2\boldsymbol{x}_1\|$ is small, because N is small or all the elements of $\boldsymbol{M}_2\boldsymbol{x}_1$ are small, $\hat{\beta}_1$ is relatively imprecise.

The squared Euclidean length of $\boldsymbol{M}_2\boldsymbol{x}_1$ is just the SSR from

$$\boldsymbol{x}_1 = \boldsymbol{X}_2\boldsymbol{c} + \text{residuals}. \tag{22}$$

Thus the precision of $\hat{\beta}_1$ is proportional to the SSR from regression (22). It depends on $\boldsymbol{X}_2$ just as much as it depends on $\boldsymbol{x}_1$.

The quadratic form $\boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{x}_1$ can be written as $\sum_{i=1}^{N}(\boldsymbol{M}_2\boldsymbol{x}_1)_i^2$. Each term in the sum is a squared residual. For high-leverage observations, these squared residuals are large.

When $\boldsymbol{X}_2$ explains $\boldsymbol{x}_1$ much better than a constant alone, the length of $\boldsymbol{M}_2\boldsymbol{x}_1$ is much less than the length of $\boldsymbol{M}_\iota\boldsymbol{x}_1$.

In this case, $\boldsymbol{x}_1$ is said to be **collinear** with some of the other regressors.

This should be called **approximate collinearity**, but it is often (wrongly) called **multicollinearity**.

- Collinearity can greatly reduce the precision of OLS estimates.
- Estimates can be imprecise even when N is very large.

When the disturbances are not independent and identically distributed, things get more complicated.

In general, for a sandwich covariance matrix, the precision of $\hat{\beta}_1$ is

$$\frac{(\boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{x}_1)^2}{\boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{\Omega} \boldsymbol{M}_2 \boldsymbol{x}_1}. \tag{23}$$

It is usual to assume that $\boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{x}_1 = O(N)$. But, when there is dependence, it may be that $\boldsymbol{x}_1^\top \boldsymbol{M}_2 \boldsymbol{\Omega} \boldsymbol{M}_2 \boldsymbol{x}_1 = O(N^*)$, with $N^* > N$.

In that case, the precision of $\hat{\beta}_1$ may be $O(N^2/N^*) < O(N)$. Thus information accumulates less rapidly than it does for independent observations.

Linear Functions of Parameter Estimates

Suppose we are interested in the variance of $\hat{\gamma}$, where $\gamma = \boldsymbol{w}^\top \boldsymbol{\beta}$, $\hat{\gamma} = \boldsymbol{w}^\top \hat{\boldsymbol{\beta}}$, and $\boldsymbol{w}$ is a k-vector of known coefficients.

By choosing $\boldsymbol{w}$ appropriately, we can make γ equal to any one of the β_i, or to the sum of the β_i, or to any linear combination of the β_i.

For example, if $\gamma = 3\beta_1 - \beta_4$, $\boldsymbol{w}$ would be a vector with 3 as the first element, -1 as the fourth element, and 0 for all the other elements.

The variance of $\hat{\gamma}$ is just

$$\begin{aligned} \text{Var}(\hat{\gamma}) &= \text{Var}(\boldsymbol{w}^\top \hat{\boldsymbol{\beta}}) = \text{E}\big(\boldsymbol{w}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{w}\big) && (24) \\ &= \boldsymbol{w}^\top \text{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big) \boldsymbol{w} && (25) \\ &= \boldsymbol{w}^\top \text{Var}(\hat{\boldsymbol{\beta}}) \boldsymbol{w}. && (26) \end{aligned}$$

In general, $\text{Var}(\hat{\gamma})$ depends on every element of $\text{Var}(\hat{\boldsymbol{\beta}})$.

The **delta method**, discussed in Chapter 5, generalizes the result (26) to nonlinear functions of $\boldsymbol{\beta}$.

Consider the special case in which $\gamma = 3\beta_1 - \beta_4$. In this case,

$$\begin{aligned} \operatorname{Var}(\hat{\gamma}) &= w_1^2 \operatorname{Var}(\hat{\beta}_1) + w_4^2 \operatorname{Var}(\hat{\beta}_4) + 2w_1 w_4 \operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_4) && (27) \\ &= 9 \operatorname{Var}(\hat{\beta}_1) + \operatorname{Var}(\hat{\beta}_4) - 6 \operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_4). && (28) \end{aligned}$$

The variance of $\hat{\gamma}$ depends on $\operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_4)$ as well as on the variances of $\hat{\beta}_1$ and $\hat{\beta}_4$.

When that covariance is large and positive, $\operatorname{Var}(\hat{\gamma})$ may be small, even when $\operatorname{Var}(\hat{\beta}_1)$ and $\operatorname{Var}(\hat{\beta}_4)$ are both large.

Just looking at standard errors of the $\hat{\beta}_i$ can be extremely misleading when we care about functions of the β_i.

Instead of using (28), we may be able to rearrange the regression so that γ is estimated directly. Or use `lincom` in `Stata`.

The Variance of Forecast Errors

Suppose we have computed $\hat{\boldsymbol{\beta}}$ and wish to predict y_j, for j not in $i = 1, \ldots, N$, using observed regressors $\boldsymbol{X}_j$.

The forecast of y_j is $\boldsymbol{X}_j\hat{\boldsymbol{\beta}}$. The **prediction error**, or **forecast error**, has mean zero, and variance

$$\begin{aligned} \mathrm{E}(y_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}})^2 &= \mathrm{E}(\boldsymbol{X}_j\boldsymbol{\beta}_0 + u_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}})^2 && (29)\\ &= \mathrm{E}(u_j^2) + \mathrm{E}(\boldsymbol{X}_j\boldsymbol{\beta}_0 - \boldsymbol{X}_j\hat{\boldsymbol{\beta}})^2 && (30)\\ &= \sigma_0^2 + \mathrm{Var}(\boldsymbol{X}_j\hat{\boldsymbol{\beta}}). && (31) \end{aligned}$$

- The first equality depends on correct specification.
- The second depends on disturbances being serially uncorrelated, which ensures that $\mathrm{E}(u_j\boldsymbol{X}_j\hat{\boldsymbol{\beta}}) = 0$.
- The third uses the fact that $\hat{\boldsymbol{\beta}}$ is assumed to be unbiased.

Under classical assumptions,

$$\text{Var}(y_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}) = \sigma_0^2 + \sigma_0^2 \boldsymbol{X}_j(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_j^\top. \tag{32}$$

The first term is the variance of u_j. The second term arises because we use $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}_0$.

Thus the variance of the forecast error is greater than the variance of the disturbance u_j.

In contrast, as we will see shortly, the variance of the residual $\hat{u}_i$ is less than the variance of the disturbance u_i.

Suppose the y_i were not generated by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ but by the DGP

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u}, \tag{33}$$

where some columns of $\boldsymbol{Z}$ may belong to $\mathcal{S}(\boldsymbol{X})$. Thus the model we estimate is misspecified.

If the regressors are fixed,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} \tag{34}$$

$$= \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}, \tag{35}$$

where the **pseudo-true** parameter vector $\boldsymbol{\beta}_0$ is $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0$.

The expected squared forecast error based on the false model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ is

$$\begin{aligned}
\mathrm{E}(y_j - \boldsymbol{X}_j \hat{\boldsymbol{\beta}})^2 &= \mathrm{E}(\boldsymbol{Z}_j \boldsymbol{\gamma}_0 + u_j - \boldsymbol{X}_j \hat{\boldsymbol{\beta}})^2 \\
&= \mathrm{E}(u_j^2) + \mathrm{E}(\boldsymbol{Z}_j \boldsymbol{\gamma}_0 - \boldsymbol{X}_j \hat{\boldsymbol{\beta}})^2 && (36) \\
&= \mathrm{E}(u_j^2) + \mathrm{E}(\boldsymbol{Z}_j \boldsymbol{\gamma}_0 - \boldsymbol{X}_j \boldsymbol{\beta}_0 + \boldsymbol{X}_j \boldsymbol{\beta}_0 - \boldsymbol{X}_j \hat{\boldsymbol{\beta}})^2 && (37) \\
&= \sigma_0^2 + \mathrm{E}(\boldsymbol{Z}_j \boldsymbol{\gamma}_0 - \boldsymbol{X}_j \boldsymbol{\beta}_0)^\top (\boldsymbol{Z}_j \boldsymbol{\gamma}_0 - \boldsymbol{X}_j \boldsymbol{\beta}_0) + \mathrm{Var}(\boldsymbol{X}_j \hat{\boldsymbol{\beta}}). && (38)
\end{aligned}$$

The middle term in the last line is essentially a squared bias.

Under classical assumptions, $\text{Var}(\boldsymbol{X}_j\hat{\boldsymbol{\beta}}) = \sigma_0^2 \boldsymbol{X}_j(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_j^\top$.

Thus the expected squared forecast error for the false model is equal to the one for the true model, given in (32), plus a squared bias term.

If $\boldsymbol{X}_j\boldsymbol{\beta}_0$ provides a good approximation to $\boldsymbol{Z}_j\boldsymbol{\gamma}_0$, then the middle term in (38) will be small.

But if it provides a poor approximation, the middle term may be large.

In that case, the expected squared forecast error may be much larger than expression (32) suggests.

We conclude that forecast errors will generally be larger than residuals for three reasons:

- $\text{Var}(u_i) > \text{Var}(\hat{u}_i)$ for all i (to be proved).
- $\text{Var}(\boldsymbol{X}_j\hat{\boldsymbol{\beta}}) > 0$ because $\hat{\boldsymbol{\beta}}$ is random.
- $\text{E}(\boldsymbol{Z}_j\boldsymbol{\gamma}_0 - \boldsymbol{X}_j\boldsymbol{\beta}_0)^\top(\boldsymbol{Z}_j\boldsymbol{\gamma}_0 - \boldsymbol{X}_j\boldsymbol{\beta}_0) > 0$.