Linear Regression

Nuno Carvalho

April 16, 2025

1 Generalized Least Squares (GLS)

Largely adapted from the respective Wikipedia article.

1.1 Model definition

In Generalized Least Squares (GLS), we define an outcome's mean to be a linear function of a set of predictors:

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

Where:

- $\mathbf{y}$: outcome vector. $N \times 1$
  - N: number of samples
- $\mathbf{X}$: design matrix. $N \times K$ matrix
  - K: number of predictors
- $\boldsymbol{\beta}$: coefficients vector. $K \times 1$ vector
- $\boldsymbol{\epsilon}$: error term. $N \times 1$ vector

Because of our conditional mean definition earlier, the error term, $\boldsymbol{\epsilon}$, has a mean of zero for a given set of predictor values ($\mathbf{X}$):

$$\mathrm{E}[\boldsymbol{\epsilon}|\mathbf{X}] = \mathbf{0}$$

We also assume that the variance of the error term given $\mathbf{X}$ is described by an invertible $N \times N$ covariance matrix, $\boldsymbol{\Omega}$:

$$\mathrm{Cov}[\boldsymbol{\epsilon}|\mathbf{X}] = \boldsymbol{\Omega}$$

We further assume that the error term follows a multivariate normal distribution with mean 0 and covariance $\boldsymbol{\Omega}$:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$$

And so it follows that:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega})$$

1.2 Least squares estimate of $\boldsymbol{\beta}$

We denote a candidate estimate for the $\boldsymbol{\beta}$ vector as $\mathbf{b}$ and define its residual vector as $\mathbf{y} - \mathbf{X}\mathbf{b}$. The goal of GLS is to find the estimate of $\boldsymbol{\beta}$ that maximizes the likelihood of the data given the above model, which we can calculate using the probability density function of a multivariate normal distribution:

$$\frac{1}{\sqrt{(2\pi)^K|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

We can plug in $\boldsymbol{\Sigma} = \boldsymbol{\Omega}$, $\mathbf{x} = \mathbf{y}$, and $\boldsymbol{\mu} = \mathbf{X}\mathbf{b}$. The $\mathbf{b}$ that maximizes this likelihood function will be the same that minimizes the inside of the exponent (without the negative), which is also the squared Mahalanobis length of the residual vector:

$$\hat{\beta} = \arg\min_b \ (\mathbf{y}-\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})$$

Through matrix algebra, this is equivalent to:

$$\begin{aligned}
\hat{\beta} &= \arg\min_b \ (\mathbf{y}-\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}) \\
&= \arg\min_b \ \mathbf{y}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y} - \mathbf{y}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X}\mathbf{b} - (\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y} + (\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X}\mathbf{b} \\
&= \arg\min_b \ \mathbf{y}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y} + (\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}(\mathbf{X}\mathbf{b}) - 2(\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y}
\end{aligned}$$

We can use calculus to solve for the $\mathbf{b}$ that minimizes this expression by taking the partial derivative with respect to $\mathbf{b}$ and solving for 0:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial\mathbf{b}}[\mathbf{y}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y} + (\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}(\mathbf{X}\mathbf{b}) - 2(\mathbf{X}\mathbf{b})^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y}] \\
&= 2X^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X}\hat{\beta} - 2X^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y}
\end{aligned}$$

Which yields:

$$\hat{\beta} = (\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y}$$

1.2.1 Specific case of Ordinary Least Squares

Note that in Ordinary Least Squares (OLS), the covariance matrix is an identity matrix, $\boldsymbol{\Omega} = I$. That is, the residuals are uncorrelated with each other. This simplifies the above equation to:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y} \\
&= (\mathbf{X}^\intercal I\mathbf{X})^{-1}\mathbf{X}^\intercal I^{-1}\mathbf{y} \\
&= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}
\end{aligned}$$

1.3 Variance of $\hat{\boldsymbol{\beta}}$

To get the variance of the $\hat{\boldsymbol{\beta}}$ estimate, we only need to focus on the variance of $\mathbf{y}$, as all other terms are not random variables. $\text{Var}[\mathbf{y}] = \boldsymbol{\Omega}$ since $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}\boldsymbol{\beta}$, as shown below:

$$\begin{aligned}
\text{Var}[\mathbf{y}] &= \text{Var}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\
&= \text{Var}[\mathbf{X}\boldsymbol{\beta}] + \text{Var}[\boldsymbol{\epsilon}] \\
&= \mathbf{0} + \text{E}[\text{Var}[\boldsymbol{\epsilon}|\mathbf{X}]] + \text{Var}[\text{E}[\boldsymbol{\epsilon}|\mathbf{X}]] \\
&= \text{E}[\boldsymbol{\Omega}] + \text{Var}[\mathbf{0}] \\
&= \boldsymbol{\Omega}
\end{aligned}$$

Let's define $bA = (\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}$, such that $\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$. Furthermore, note that since $\boldsymbol{\Omega}$ is a covariance matrix, it (and its inverse) is symmetric: $\boldsymbol{\Omega} = \boldsymbol{\Omega}^\intercal$. The same symmetry property applies to $\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X}$. We can then solve:

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}] &= \text{Var}[(\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{y}] \\
&= \text{Var}[\mathbf{A}\mathbf{y}] \\
&= \mathbf{A}\text{Var}[\mathbf{y}]\mathbf{A}^\intercal \\
&= \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\intercal \\
&= (\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}((\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1})^\intercal \\
&= (\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\mathbf{X}(\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})(\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}
\end{aligned}$$

We can define $\mathbf{B} = \mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X}$, allowing us to simplify the above equation, $\mathbf{B}^{-1}\mathbf{B}\mathbf{B}^{-1} = \mathbf{B}^{-1}$, yielding:

$$\text{Var}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^\intercal\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$