# Linear Regression

Nuno Carvalho

April 10, 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:

$$y = X\beta + \epsilon$$

Where:

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

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

$$E[\epsilon|X] = 0$$

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

$$Cov[\epsilon|X] = \Omega$$

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

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

And so it follows that:

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

## 1.2 Least squares estimate of $\beta$

We denote a candidate estimate for the $\beta$ vector as $b$ and define its residual vector as $y - Xb$. The goal of GLS is to find the estimate of $\beta$ 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:

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

We can plug in $\Sigma = \Omega$, $x = y$, and $\mu = Xb$. The $b$ 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:

$$\hat{\beta} = \underset{b}{\operatorname{argmin}}\ (y - Xb)^\top \Omega^{-1}(y - Xb)$$

Through matrix algebra, this is equivalent to:

$$\begin{aligned}
\hat{\beta} &= \underset{b}{\operatorname{argmin}}\ (y - Xb)^\top \Omega^{-1}(y - Xb) \\
&= \underset{b}{\operatorname{argmin}}\ y^\top \Omega^{-1} y - y^\top \Omega^{-1} Xb - (Xb)^\top \Omega^{-1} y + (Xb)^\top \Omega^{-1} Xb \\
&= \underset{b}{\operatorname{argmin}}\ y^\top \Omega^{-1} y + (Xb)^\top \Omega^{-1}(Xb) - 2(Xb)^\top \Omega^{-1} y
\end{aligned}$$

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

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

Which yields:

$$\hat{\beta} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y$$

### 1.2.1 Specific case of Ordinary Least Squares

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

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

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

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

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

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

$$\begin{aligned}
Var[\hat{\beta}] &= Var[(X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y] \\
&= Var[Ay] \\
&= AVar[y]A^\top \\
&= A\Omega A^\top \\
&= (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} \Omega ((X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1})^\top \\
&= (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} \Omega \Omega^{-1} X (X^\top \Omega^{-1} X)^{-1} \\
&= (X^\top \Omega^{-1} X)^{-1} (X^\top \Omega^{-1} X)(X^\top \Omega^{-1} X)^{-1}
\end{aligned}$$

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

$$Var[\hat{\beta}] = (X^\top \Omega^{-1} X)^{-1}$$