

Advanced Quantitative Methods: OLS in Matrix Form

Thomas Gschwend | Lisa-Marie Müller | Domantas Undzėnas

Week 2 - 18 February 2026

Introduction

What should you take home from this class today?

- Matrices are your friends
- You will learn to get OLS estimates without actually running a regression command in R.
- We clarify what robust standard errors are and what not.

OLS Model

Some Definitions and Notation

- Suppose we have the following multiple regression model with $k + 1$ parameters (but k independent variables) and $i = 1, \dots, n$ observations,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i$$

- For every observation i the relationship between values of the dependent variable y_i and the corresponding values on the covariates $x_{i1}, x_{i2}, \dots, x_{ik}$ can be written as:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_k x_{1k} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_k x_{2k} + \epsilon_2$$

$$y_3 = \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \dots + \beta_k x_{3k} + \epsilon_3$$

$$\vdots = \vdots$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i$$

$$\vdots = \vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} + \epsilon_n$$

Some Definitions and Notation

- The system of n equations can be elegantly condensed as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}_{[n \times 1]} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ik} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}_{[n \times (k+1)]} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_k \end{pmatrix}_{[(k+1) \times 1]} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}_{[n \times 1]}$$

- This can be rewritten as

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

- The model has a systematic component ($X\beta$) and a stochastic component (ϵ)
- We would like to obtain estimates of the population parameters (β), which we denote as $\hat{\beta}$

Derivation of $\hat{\beta}_{OLS}$

- We derive our first estimator $\hat{\beta}_{OLS}$, i.e., we need to find a $\hat{\beta}$ that minimizes the *sum of squared residuals* (what we used to write as $\sum_{i=1}^n e_i^2$ in scalar notation).
- In matrix notation the vector (dimension?) of the residuals, e , is given as

$$e = y - X\hat{\beta}$$

- Thus, the sum of squared residuals (RSS) is $e'e$ (dimension?)

$$(e_1, e_2, \dots, e_n) \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = e_1 \cdot e_1 + e_2 \cdot e_2 + \dots + e_n \cdot e_n = \sum_{i=1}^n e_i^2$$

Derivation of $\hat{\beta}_{OLS}$

- Using our new matrix notation, we can write the sum of squared residuals as:

$$\begin{aligned}e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}\end{aligned}$$

- In order to minimize the above equation, we need to take the derivative with respect to $\hat{\beta}$ (*first order condition*). Thus,

$$\frac{\partial e'e}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0 \quad (\text{dimension?})$$

Derivation of $\hat{\beta}_{OLS}$

- Ok, lets work with $\hat{\beta}$ that fulfills $\frac{\partial e'e}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0$
- To check wether this is in fact a minimum, we need to take the derivative of the above with respect to $\hat{\beta}$ again (*second order condition*). This gives us

$$\frac{\partial^2 e'e}{\partial \beta \partial \beta'} = 2X'X > 0 \text{ if } \text{rank}(X) = k+1$$

If $\text{rank}(X) = k+1$, i.e., the k independent variables are not linear dependent, then $X'X$ is positive definite ($a'(X'X)a > 0$ for all $n \times 1$ vectors $a \neq 0$).

- Thus, if we find such a $\hat{\beta}$, then it minimizes the sum of squared residuals.

Derivation of $\hat{\beta}_{OLS}$

- How to find such a $\hat{\beta}$? Start with the *first order condition*

$$\frac{\partial e'e}{\partial \beta} = -2X'y + 2X'X\hat{\beta} = 0$$

- Note that $(X'X)$ is symmetric. Rearranging terms, we get the so-called *normal equations* (why plural?):

$$(X'X)\hat{\beta} = X'y$$

- Recall that $(X'X)$ and $X'y$ is known, while $\hat{\beta}$ is unknown. Assuming $(X'X)^{-1}$ exists, pre-multiplying both sides by the inverse yields

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

- Note that we have not had to make any assumption so far (in addition to the existence of $(X'X)^{-1}$).
- Since the OLS estimators $\hat{\beta} = \hat{\beta}_{OLS}$ are a linear combination of an existing random variable (y), they themselves are random variables.

The Gauss-Markov Assumptions

In order to derive the expected value and the variance of $\hat{\beta}_{OLS}$ we need some assumptions

1. $y = X\beta + \epsilon$, i.e., the relationship between X and y is *linear in the parameters*.
2. X is of full rank, i.e. there is *no perfect collinearity* among the covariates.
3. $E[\epsilon|X] = 0$, i.e. the *disturbances* average out to 0 for any value of X (*zero conditional mean*), thus $E(y) = X\beta$ (i.e. on average we get the mean right).
4. $Var(\epsilon|X) = E[\epsilon\epsilon'|X] = \sigma^2 I_n$, i.e. the variance of the disturbances must be constant (*homoskedasticity*) and cannot be correlated across observations (*no serial- or autocorrelation*).
5. $\epsilon|X \sim N(0, \sigma^2 I_n)$ (actually not needed for Gauss-Markov Theorem but for hypothesis testing).

The Gauss-Markov Theorem

There will be no other linear, unbiased estimator of β coefficients that has smaller sampling variance. Thus, $\hat{\beta}_{OLS}$ will be BLUE (Best Linear Unbiased Estimator).

Proof that $\hat{\beta}$ is an unbiased estimator of β .

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ \hat{\beta} &= \beta + (X'X)^{-1}X'\epsilon\end{aligned}$$

Taking the expectation conditional on X gives

$$\begin{aligned}E[\hat{\beta}|X] &= \beta + (X'X)^{-1}X'E[\epsilon|X] \\ &= \beta + (X'X)^{-1}X'0 \\ &= \beta\end{aligned}$$

The Variance-Covariance Matrix of OLS estimator $\hat{\beta}$

We can derive the variance-covariance matrix of the OLS estimator $\hat{\beta}$ (i.e., the “variance” of a multidimensional estimator) conditional on X as

$$\begin{aligned}\text{Var}(\hat{\beta}|X) &= E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] \\ &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[((X'X)^{-1}X'\epsilon)((X'X)^{-1}X'\epsilon)'] \\ &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2I_nX(X'X)^{-1} \\ &= \sigma^2I_n(X'X)^{-1}(X'X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned}$$

The Variance-Covariance Matrix of OLS estimator $\hat{\beta}$

- We observe $(X'X)^{-1}$ but we cannot observe σ^2 . One can show that

$$\hat{\sigma}^2 (= s^2) = \frac{e'e}{n - (k + 1)}$$

is an unbiased estimator of σ^2 . The positive square root of it ($\hat{\sigma}$ or s) is called the *standard error of the regression* (or *root mean squared error*) and is an estimator of the standard deviation of the regression error term.

- The denominator $n - (k + 1)$ is the difference of the number of observations and the number of estimated parameters and is called *degrees of freedom* (df).
- Note: As n increases $\hat{\sigma}^2$ decreases, while as k increases $\hat{\sigma}^2$ increases

The Variance-Covariance Matrix of OLS estimator $\hat{\beta}$

- Thus, the (symmetric) variance-covariance matrix $Var(\hat{\beta}|X) =$

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \begin{pmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & \cdots & cov(\hat{\beta}_0, \hat{\beta}_k) \\ cov(\hat{\beta}_1, \hat{\beta}_0) & var(\hat{\beta}_1) & \cdots & cov(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \vdots & \vdots \\ cov(\hat{\beta}_k, \hat{\beta}_0) & cov(\hat{\beta}_k, \hat{\beta}_1) & \cdots & var(\hat{\beta}_k) \end{pmatrix}$$

- ...can be estimated through

$$\widehat{Var}(\hat{\beta}|X) = \hat{\sigma}^2(X'X)^{-1} = \frac{e'e}{n - (k + 1)}(X'X)^{-1}$$

- As you can see, the standard errors of $\hat{\beta}$ are given by the square root of the elements along the main diagonal of the above (symmetric) matrix, i.e.

$$se(\hat{\beta}_i) = \sqrt{\widehat{var}(\hat{\beta}_i)}$$

Hypothesis Testing

In order to test hypotheses about $\hat{\beta}_{OLS}$, we need a distributional assumption.

- We assume that $\epsilon | X \sim N(0, \sigma^2 I_n)$, i.e., the disturbances are multivariate normally distributed with mean zero and variance $\sigma^2 I_n$.
- We have shown that $\hat{\beta}_{OLS} = \beta + (X'X)^{-1}X'\epsilon$, i.e., the OLS estimator is a linear function of the disturbances.
- Since a linear transformation of a multivariate normal vector is again multivariate normal, it follows that

$$\hat{\beta}_{OLS} | X \sim N(\beta, \sigma^2 (X'X)^{-1}).$$

This gives us the sampling distribution (of our OLS estimator) needed for hypothesis testing.

- In particular, the variance of $\hat{\beta}_j$ (conditional on X) is

$$\text{Var}(\hat{\beta}_j | X) = \sigma^2 [(X'X)^{-1}]_{jj}.$$

Again, since we do not know σ^2 , we replace it in practice by $\hat{\sigma}^2$.

White-Huber Standard Errors

Robust (White-Huber) Standard Errors

- Recall that we derived the variance-covariance matrix of the OLS estimator $\hat{\beta}$ conditional on X as

$$\begin{aligned}\text{Var}(\hat{\beta}|X) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1} \\ &= (X'X)^{-1}(X'\Omega X)(X'X)^{-1}, \text{ where } \Omega = \text{Var}(\epsilon | X).\end{aligned}$$

This also helps us to understand the so-called *Beck and Katz panel-corrected standard errors* in the context of cross-sectional time-series models.

- Note that we can compute $\hat{\beta}$ without making any distributional assumption about the disturbances (because $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$).
- However, to conduct inference and to test hypotheses we need distributional assumptions about $\text{Var}(\epsilon | X)$.
- The assumption of homoskedasticity is strong and often violated. The error variance is often not the same for all observations. Some observations are better to predict than others.

Robust (White-Huber) Standard Errors

If we detect (or suspect) heteroskedastic error variances (σ_i^2 instead of σ^2) there are two basic strategies to deal with this

1. **Model the non-constant variance** to extract *substantive* information from it, e.g., Weighted Least Squares (WLS - using a weight that is proportional to the variance) or *parameterize the variance* (heteroskedastic regression, multi-level models) and estimate it. In order to do this, we need to assume that the variance is correctly specified.
2. **Use a statistical fix** to treat non-constant variance as *nuisance* (e.g., robust standard errors)

$$\begin{aligned}\text{Var}(\hat{\beta}|X) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= (X'X)^{-1}(X'E[\epsilon\epsilon']X)(X'X)^{-1} \\ \widehat{\text{Var}}(\hat{\beta}|X) &= (X'X)^{-1}(X'\hat{\Omega}X)(X'X)^{-1}\end{aligned}$$

Robust (White-Huber) Standard Errors

- Use a **statistical fix** to treat non-constant variance as nuisance (e.g., robust standard errors)

$$\text{Var}(\hat{\beta}|X) = (X'X)^{-1}(X'E[\epsilon\epsilon']X)(X'X)^{-1}$$

$$\widehat{\text{Var}}(\hat{\beta}|X) = (X'X)^{-1}(X'\hat{\Omega}X)(X'X)^{-1}$$

- White (1980) showed that $X'\hat{\Omega}X$ with $\hat{\Omega} = \text{diag}(ee')$, a diagonal matrix with the squared residuals as non-zero elements, is a consistent (but not unbiased) variance-covariance matrix estimator of $X'E[\epsilon\epsilon']X$, also called Heteroskedastic-Consistent (HC) estimator.
- One way to modify HC is to implement a *degrees of freedom correction* similar to the one used to obtain unbiased estimates of σ^2 . This strategy to get White robust standard errors yields the following so-called HC1 variance-covariance matrix estimator $\frac{n}{n-(k+1)}X'\hat{\Omega}X$ for $X'E[\epsilon\epsilon']X$, that is, as with HC, only appropriate in large samples.

Classical vs. Robust Standard Errors

- Point estimates do not require distributional assumptions:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y.$$

- But standard errors depend on the error variance:

$$\text{Var}(\hat{\beta} | X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}, \quad \Omega = \text{Var}(\varepsilon | X).$$

- Classical (homoskedastic) case:* if $\Omega = \sigma^2 I_n$ then

$$\widehat{\text{Var}}(\hat{\beta} | X) = \hat{\sigma}^2 (X'X)^{-1}.$$

- Heteroskedasticity:* if $\Omega \neq \sigma^2 I_n$, classical SEs are wrong. White-Huber SEs replace Ω with an estimator (sandwich form):

$$\widehat{\text{Var}}_{rob}(\hat{\beta} | X) = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}.$$

What are robust (White-Huber) standard errors good for?

- In R one can implement HC1 using `diag(diag(.))` (turns vector into square diagonal matrix) to implement $\hat{\Omega}$ directly or load `library(sandwich)`.
- King and Roberts (2015) argue that if robust and classical standard error estimates differ substantially, this indicates that the model's assumptions about the error structure are incorrect. Such variance misspecification often reflects deeper structural problems (omitted variables, incorrect functional form, ignored model dynamics, ect) that robust standard errors do not fix.
- Robust standard errors only fix the reported standard errors but not other quantities of interest (predicted probabilities, ect.).
- Thus, robust standard errors are some sort of a weak *specification test* only. If standard errors diverge this indicates model misspecification. If not, we do *not* know!
- Take a look at: [King, Gary, and Margaret E. Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix". *Political Analysis* 23, 159–179.](#)