Ordinary Least Squares (OLS)

This is your bread and butter. Eat it, drink it, breathe it. Not because it is the answer to everything, but because most economists think it is. Understanding OLS deeply will allow you to 1) engage with the majority of regression-based work (econ or not), and 2) spot design flaws in reports, analysis, or papers.

We will start covering it in more detail in the next section, but for now here is a general

overview.

OLS is a minimization algorithm that finds estimates for a parameter of interesting by minimizing a specific loss function called Sum of Squared Residuals (SSR).

- The residual is the difference between a predicted / fitted value ($\hat{y}_i$) and an observed value (y_i).
- The SSR is the metric we want to optimize
  - Many other loss functions

(or performance metrics) exist. OLS uses SSR.

- Analytically, OLS is an unconstrained optimization problem. Hence we only need to solve for 1) $f'(x) = 0$, and 2) $f''(x) \geq 0$
- Algorithmically, OLS is an iterative process that starts with a random line and checks a bunch of options. Each time adjusting in the direction that reduces SSR

In other words, given the population regression $E[Y|X] = X\beta + \varepsilon = \beta_0 + \beta_1 X_i + \varepsilon_i$, we first want to find the residual

(fitted error)

$$\hat{\varepsilon}_i = y - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$\hat{\varepsilon} = Y - X\hat{\beta}$$

We'll denote β as the true unknown vector of parameters and $\hat{\beta}$ our sample estimates. Then, construct the SSR

$$\sum \hat{\varepsilon}_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\hat{\varepsilon}^T \hat{\varepsilon} = (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

To find the optimal values of $\hat{\beta}$ we have to solve the minimization problem with SSR objective function.

Brief optimization review

The optimization problem consists of max or min an objective function (SSR, MSE, ...) If there are no constraints on the values of the variables to optimize, we call it unconstrained max/min problem. It can be written as

$$\min_{x \in X} f(x) \quad \text{or} \quad \max_{x \in X} f(x)$$

Two conditions must be checked:

1st Order condition

The derivative (scalar) or gradient (vector) at the global max/min must be 0.

$$\nabla f(x^*)$$

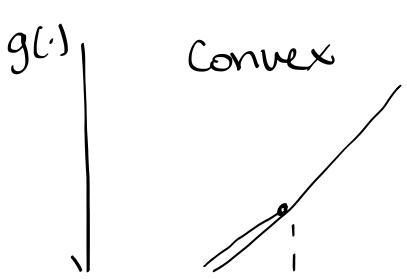
$$\frac{d f(x^*)}{dx} = 0$$

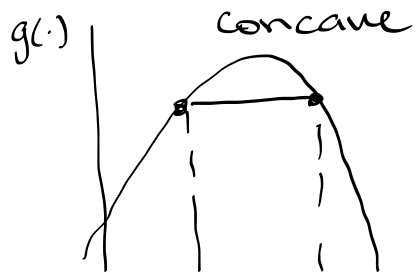
$$\nabla f(x^*) = \begin{bmatrix} \frac{\partial F(x^*)}{\partial x_1} \\ \vdots \\ \frac{\partial F(x^*)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0}$$

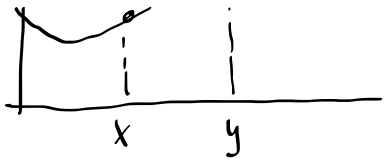
A point x^* satisfying this condition its called a critical or stationary point.

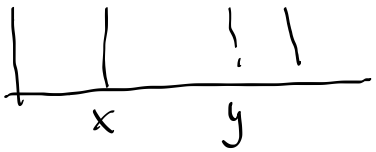
2nd order condition

The 2nd derivative gives us the shape of the derivative (concave, convex)









In minimization we want a convex function because it has a unique min.
In maximization we want concavity

(scalar)

$$\frac{\partial^2 f(x^*)}{\partial x^2} > 0 \quad \text{(convex)}$$

$$\frac{\partial^2 f(x^*)}{\partial x^2} < 0 \quad \text{(concave)}$$

(Vector)

Introduce the Hessian matrix $H(x)$

A square matrix of second-order partial derivatives.

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

$H(x)_{i,j} > 0$ (convex) "Positive Definite"

$H(x)_{i,j} < 0$ (concave) "Negative Definite"

The OLS minimization problem is an

unconstrained optimization one.

(scalar)

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

$$\min_{\beta_0, \beta_1} S(\beta_0, \beta_1) = \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_K}, \quad K: 0, 1$$

For the int. β_0

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = \sum (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$-2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum y_i - n\beta_0 - \beta_1 \sum x_i = 0$$

$$n\bar{y} - \beta_1 n\bar{x} = n\beta_0$$

$$\hat{\beta}_0^{ols} = \bar{y} - \beta_1 \bar{x}$$

For slope β_1,

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum (x_i)(y_i - \hat{\beta}_0 - \beta_1 x_i) = 0$$

$$\sum x_i y_i - \hat{\beta}_0 \sum x_i - \beta_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum x_i - \beta_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - \bar{y} \sum x_i + \beta_1 \bar{x} \sum x_i - \beta_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - n \bar{y} \bar{x} + \beta_1 n \bar{x}^2 - \beta_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - n \bar{y} \bar{x} = \beta_1 \left(\sum x_i^2 - n \bar{x}^2 \right)$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{y} \bar{x}}{\sum x_i^2 - n \bar{x}^2}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}$$

Exc.

$$\sum x_i y_i - n\bar{y}\bar{x} = \sum x_i y_i - n\bar{x}\bar{y} + n\bar{x}\bar{y} - n\bar{x}\bar{y}$$

$$\sum x_i y_i - \bar{y}\sum x_i + \sum \bar{x}\bar{y} - \bar{x}\sum y_i$$

$$\sum \left(x_i y_i - \bar{y} x_i + \bar{x}\bar{y} - \bar{x} y_i \right)$$

$$\sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right) = Cov(x, y)$$

$$\sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2 + n\bar{x}^2 - n\bar{x}^2$$

$$= \sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2$$

$$= \sum \left(x_i^2 - 2\bar{x} x_i + \bar{x}^2 \right)$$

$$= \sum \left(x_i - \bar{x} \right)^2$$

So, the OLS estimates are

$$\hat{\beta}_0 = \bar{y} - \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}\bar{x}$$

$$\hat{\beta}_1 = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}$$

(Matrix form) $Y = X\beta + \varepsilon \Rightarrow \varepsilon = Y - X\beta$

$\varepsilon^T\varepsilon = \sum \varepsilon_i \varepsilon_i = \sum \varepsilon_i^2 = (Y - X\beta)^T (Y - X\beta)$

↳ Dot product * Recall the Transpose distributes

$$\frac{\partial \varepsilon^T\varepsilon}{\partial \beta} = (Y^T - (X\beta)^T)(Y - X\beta) = 0$$

(Matrix ↓, ↳ Vector)

$$= Y^TY - Y^TX\beta - (X\beta)^TY + (X\beta)^T(X\beta)$$

$$= 0 - Y^TX - X^TY + \beta^TX^TX\beta$$

$$= -2X^TY + 2X^TX\beta = 0$$

$$-2X^TY = -2X^TX\beta$$

$$\boxed{(X^TX)^{-1}X^TY = \hat{\beta}_{OLS}}$$

Assuming the inverse $(X^TX)^{-1}$ exists, which it does as long as there is no perfect multicollinearity. $\text{Det}(X^TX) \neq 0$, $\text{Rank}(X^TX) = K$.

Matrix Calculus concepts used

$$\frac{\partial a^T b}{\partial b} = \frac{\partial a b^T}{\partial b} = a \quad , \quad a, b \in \mathbb{R}^{K \times 1}$$

$$\frac{\partial b^T A b}{\partial b} = 2Ab = 2b^T A \quad , \quad A \in \mathbb{R}^{K \times K}$$

and symmetric, $(A + A^T)b$ if not symmetric

$$\frac{\partial 2\beta^T X^T y}{\partial \beta} = \frac{\partial \text{ vector } (\text{vector})}{\partial \text{ vector}}$$

$$= \frac{\partial 2\beta^T (X^T y)}{\partial \beta} = 2X^T y$$

$$\frac{\partial \beta^T X^T X \beta}{\partial \beta} = \frac{\partial \beta^T A \beta}{\partial \beta} = 2A\beta = 2\beta^T A$$

We can now use our OLS estimate to get the residuals

$$\hat{\varepsilon} = Y - X\hat{\beta}_{OLS}$$

$$\hat{\varepsilon} = Y - X(X^T X)^{-1} X^T Y$$

$$\hat{\varepsilon} = (I - X(X^TX)^{-1}X^T)\,Y$$
$$\hat{\varepsilon} = M\,Y$$

M is called the residual maker. It yields the

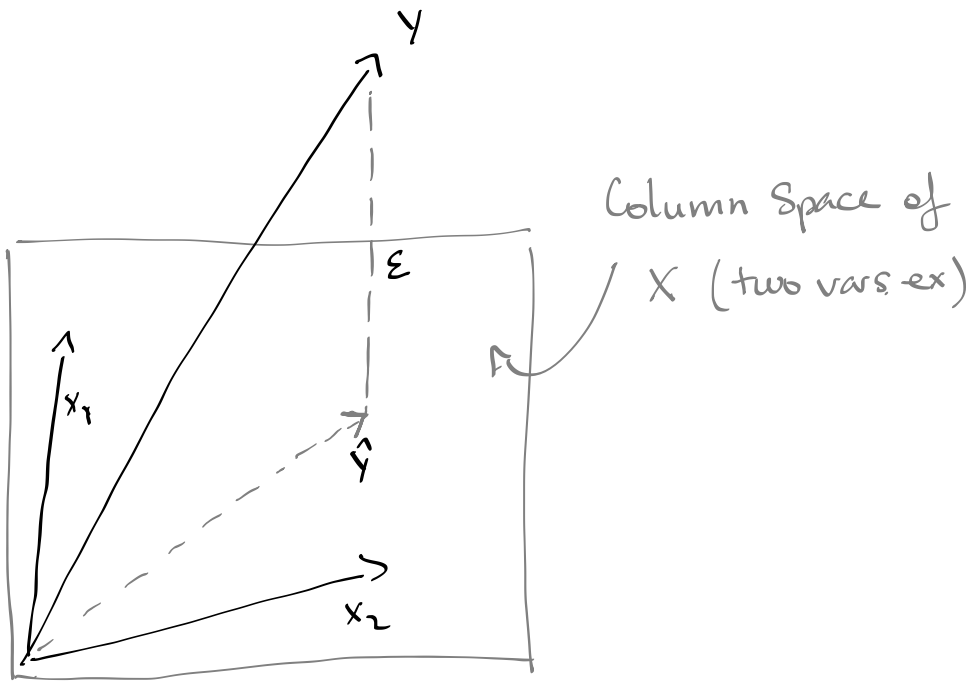
OLS residuals (true - fitted values) for our regression. M is symmetric ($M = M^T$) and idempotent ($M = M^2$)

M is also called "Annihilator matrix"

Note that from $Y = X\beta + \varepsilon$, we can get the fitted or estimated values of Y

$$\hat{Y} = Y - \hat{\varepsilon} = Y - MY = (I - M)Y$$
$$= (I - I - X(X^TX)^{-1}X^T)\,Y$$
$$= X(X^TX)^{-1}X^TY$$
$$= PY$$

P is called the Projection matrix, also symmetric and Idempotent. It gets this name because it "projects" the vector Y into the column space of X



The idea is that Y lives in some other linear space, and the projection gives its value in the linear space our data lives in.

Note the projection is orthogonal. This is because from M and P being symmetric and idempotent we have that these two
are orthogonal as well

$$PM = MP = 0$$

So we can reconstruct or recover the true values Y by adding the fitted values $\hat{Y}$ and the residuals

$$Y = Py + My = \text{projection} + \text{residuals}$$

Small Sample Properties

What are the two properties of an estimator we care about? Unbiasedness & Efficiency

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

$$E[\hat{\beta}_{OLS}] = E[(X^T X)^{-1} X^T Y]$$

$$= E[(X^T X)^{-1} X^T (X\beta - \varepsilon)]$$

$$= E[(X^T X)^{-1} X^T X\beta - (X^T X)^{-1} X^T \varepsilon]$$

$$= E[I\beta - (X^T X)^{-1} X^T \varepsilon]$$

$$= E[\beta] - E[\underbrace{(X^T X)^{-1} X^T \varepsilon}]$$

estimated from data so it's deterministic

$$= \beta - (X^T X)^{-1} X^T E[\varepsilon]$$ ← what assumption do we need to use?

$E[\hat{\beta}_{OLS}] = \boxed{\beta}$ Unbiased if $E[\varepsilon] = 0$

If X is random we basically take a conditional expectation

$$E[(X^TX)^{-1}X^T\varepsilon|X] = (X^TX)^{-1}X^T E[\varepsilon|X]$$

in which case we need $E[\varepsilon|X] = 0$

$$Var(\hat{\beta}_{OLS}) = E[(\hat{\beta}_{OLS_i} - \beta)^2]$$ "Deviations from the true value"

(matrix form)

$$= E\left[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)^T\right]$$

In general, x^Tx is a dot product and xx^T is the variance-covariance matrix.

$$= E\left[\left((X^TX)^{-1}X^TY - \beta\right)\left((X^TX)^{-1}X^TY - \beta\right)^T\right]$$

Recall $\hat{\beta}_{ols} = (X^TX)^{-1}X^Ty$, so $\hat{\beta}_{ols} = (X^TX)^{-1}X^T(X\beta+\varepsilon)$

$$\Rightarrow \hat{\beta}_{ols} = \beta + (X^TX)^{-1}X^T\varepsilon$$

$$= E\left[\left((X^TX)^{-1}X^T\varepsilon + \beta - \beta\right)\left((X^TX)^{-1}X^T\varepsilon + \beta - \beta\right)^T\right]$$

$$= E\left[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}\right]$$

$$= (X^TX)^{-1}X^T E[\varepsilon\varepsilon^T] X (X^TX)^{-1}$$

↳ We need homoskedasticity

$E[\varepsilon\varepsilon^T] = \sigma^2 I$

$$= (X^TX)^{-1}X^T \sigma^2 I X (X^TX)^{-1}$$

$$= \sigma^2 I (X^TX)^{-1}X^T X (X^TX)^{-1}$$

$$= \sigma^2 I (X^TX)^{-1}$$

↳ How to estimate σ^2?

Recall the variance measures deviations from the mean. Here, our "mean" is the model. So we want to know how much do observations vary around their predicted value. This is exactly what the residuals measure

The unbiased estimator for σ^2 (the true variance of the error term) is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{N-K} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{N-K} = (y - X\hat{\beta})^T(y - X\hat{\beta})$$

To show $E[\hat{\sigma}^2] = \sigma^2$ you'll need the residual maker matrix (M), remember its properties to compute $E[\varepsilon^T\varepsilon]$ using the trace of M, and get $\frac{\sigma^2(N-K)}{(N-K)} = \sigma^2$

$$Var\left(\hat{\beta}_{OLS}\right) = \frac{\varepsilon^T\varepsilon}{N-K}\left(X^TX\right)^{-1} = \hat{\sigma}^2\left(X^TX\right)^{-1}$$

Under the assumptions covered last week, the OLS estimator is BLUE.

- Best Linear Unbiased Estimator.

In fact, a few years ago Hansen proved that it is BLUE (best even amongst non-linear estimators), which is not surprising given the linearity assumption.

Everything rests on the assumptions!

Make sure these are clear to you. Even better, write an R function that runs all the checks for a given regression model to save you time in

the future.

Goodness of fit (R^2 and adjusted R^2)
and Basic Stats Review II next week.