Introduction

We spent half of our lectures learning the foundations and properties of OLS. The second half has been on understanding the various ways in which OLS may not be optimal, how to identify the source of the error, and how to fix it if possible.

Let's chat about it to recap. These are good "interview" type questions to be comfortable with:

1) How would you define OLS?
2) Is there any difference between linear regression and OLS?
3) What is the objective function OLS optimizes for?

4) What do you think is the most intuitive way to understand SSR?

5) What are the main benefits of OLS?

6) What are the main limitations?

7) Think about the 5 GMAs, Are these about moments or a particular distr.? 8) Once we have an OLS estimate, what is the technical interpretation of any prediction made with the linear model and the OLS-estimated parameters?

The most important feature, imo, of OLS framework is that we need to make assumptions about the

functional form of the DGP (i.e., linear $y = X\beta + \varepsilon$) and the moments (i.e., mean and variance) of the errors (i.e., $E[\varepsilon|X] = 0$, $E[\varepsilon\varepsilon^T|X] = \sigma^2 I$).

Often, this is very restrictive. Most problems, at least the most interesting ones, are not linear. Imposing OLS into a non-linear problem is like trying to force a cube down a circle hole. You might get it through, but it'll require a lot of effort. Think of linearizing

variables, fixing SE, and any data "massaging" technique you can think of.

OLS is powerful, but not almighty.

It is up to you, the modeler, to decide

if it's appropriate or not.

Luckily, there are plenty of alternatives. Today we will study an alternative estimation procedure called Maximum Likelihood Estimation (MLE). In a few weeks, we'll cover more advanced models commonly used in Data Science or ML but that have gained a lot of attention in the econometrics community.

Maximum Likelihood Estimation

In an OLS setting, we considered the data as a random process and the parameters as some fixed unknown values. The linearity condition clearly restricts the type of DGP we could

accurately capture.

In contrast, MLE considers the data

as fixed and the parameters as a random unknown vector. In other words, MLE asks "whats is the vector of parameters that makes the given data most likely to be the true DGP?"

Note there are no restrictions on how

these parameters might look like nor on how the DGP should look like.

The assumption we must make instead

is one about the model (i.e., dependent variable y. Eg, $y = X\beta + \varepsilon$) and/or the error term (e.g., $\varepsilon \sim N(0,1)$).

No need to force values for specific moments of the model or the errors, but choosing an appropriate probability distrib. might be challenging.

Consider the following analogy.

Say I have a bucket of candy. There are only four types: Red, Pink, Yellow, and
Orange. We have been tasked to count the proportions of each candy type in the bucket. As we start to do so, we hear the school's alarm go off. We forgot about a test evacuation!

We leave the classroom and leave the
candy behind. It's one of those really

hot summer days. When we come back, the candy is melted! We are left with a gooey mesh of melted candy.

Good thing I'm not teaching art...

We see that the mesh is mostly orange, with some shades and strikes in red, yellow, and pink.

The mesh is our data, and the true proportion of colors is our parameter vector. Can we figure out the original

composition of candy type in the bucket? How would you approach this problem?

Here is where MLE comes in. We observe the data and make some assumption about the parameters. For the sake of the example we can say that the color distribution in the melted mesh is a sample from a multinomial distribution, so want to estimate the parameters of this underlying distribution. (i.e., the probabilities/frequencies of each color).

We construct a likelihood function

$$L(\theta \mid \text{mesh}) = f_{\text{mesh}}(\text{mesh} \mid \theta)$$

where $f_{mesh}(mesh \mid \theta)$ is the joint PDF of each variable (i.e., color) conditional on parameters θ.

Maximizing this likelihood function wrt θ will yield the proportions of color most likely to have produced the observed mesh, given our assumption of the original distribution of colors.

$$\hat{\theta} = \arg\max_{\theta} L(\theta \mid mesh)$$

In a sense, we are trying to reverse engineer the problem. Instead of assuming the parameters are fixed and hence independent of the data, MLE takes a more humble, and flexible, approach by taking the data as the ground truth and figuring

out how to best explain the way the outcome (i.e., mesh | came to be.

OLS has no chance of operationalizing a problem like this, but it is more robust to distributional mis specification. If we assume the wrong distribution then MLE will be wrong. But if our "melting model", the assumption of how the mesh came to be, is correct then MLE is far better because it uses the full information (i.e., distribution).

The technical details

For a random sample $X = (x_1, x_2, \ldots, x_n)^T$ with joint PDF $f_x(x|\theta)$

conditional on parameter vector θ, the likelihood function is

$$L(\theta|x) = f_x(x|\theta)$$

If the sample is iid, this simplifies to

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

Recall statistical independence means that we can express a joint distribution as the product of the individual distrib.

$$f_x(x|\theta) = f(x_1, x_2, \dots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n)$$

We want to maximize $L(\theta|x)$ wrt θ. Good old calculus is the way. But if we have iid variables we can apply

a log transformation to avoid dealing with all these multiplications

$$\ln L(\theta|x) = \ell(\theta|x) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

Because the ln is a monotonic transf. the same global optima will be preserved.

With this set up, the MLE estimator is the value of θ that maximizes the likelihood of the observed data.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta|x) = \arg\max_{\theta} \ell(\theta|x)$$

To find $\hat{\theta}_{MLE}$ take the partials wrt to each $\theta_i \in \theta$, set them to 0, solve, and check 2nd order conditions.

$$\nabla_\theta \ell(\theta|x) = \begin{bmatrix} \frac{\partial \ell(\theta|x)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\theta|x)}{\partial \theta_n} \end{bmatrix} = \vec{0}$$

Importantly, note that $L(\theta|x)$ is <u>not</u> a PDF. It is a function of θ that takes the observed data as given.

Ex: Normal mean

Suppose $y \sim N(\mu, 1)$ and we observe one observation $y = 3$. Let's find the expected value μ with MLE.

$$L(\mu|y) = f_y(y|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(3-\mu)^2}{2}}$$

↖ PDF of normal dist.

$$\ell(\mu|y) = -\frac{1}{2}\ln(2\pi) - (3-\mu)^2$$

$$\frac{\partial \ell(\mu|y)}{\partial \mu} = 0 - (3-\mu)(-1) = 0$$

$$\boxed{\mu = 3}$$

The best estimate is the single observation we have, makes sense.

Ex: Linear Regression with normal errors

Suppose we have $y = X\beta + \varepsilon$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with n observations

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow y_i | X_i, \beta, \sigma^2 \sim \mathcal{N}(X_i\beta, \sigma^2)$$

$$L(\beta, \sigma^2 | y, X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(y_i - X_i\beta)^2}{2\sigma^2}\right)}$$

$$\ell(\beta, \sigma^2 | y, X) = \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{(y_i - X_i\beta)^2}{2\sigma^2} \right]$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2$$

$$\frac{\partial \ell(\beta, \sigma^2 | y, X)}{\partial \beta} = \frac{1}{\sigma^2} X^T(y - X\beta) = 0$$

$$X^T y - X^T X \beta = 0$$

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

Same as OLS!

$$\frac{\partial \ell(\beta, \sigma^2 \mid y, X)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (y_i - x_i\beta)^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (y_i - x_i\hat{\beta})^2 = \frac{\sum \hat{\varepsilon}_i^2}{n}$$

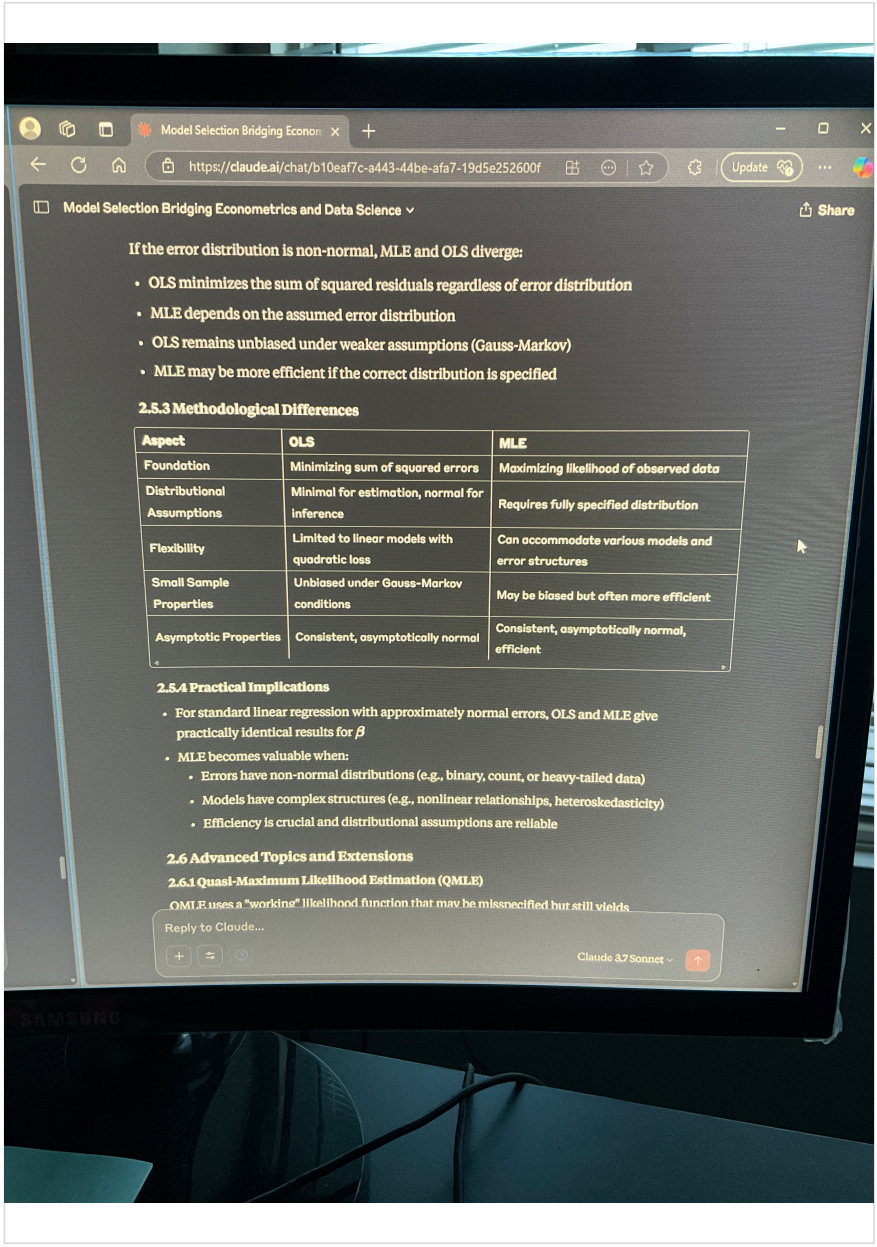
Unlike the OLS estimator for variance, MLE does not correct for # of regressors (K) so it may be biased in small samples

MLE vs OLS

OLS	MLE
$\min \sum \varepsilon_i^2$	$\max L(\beta, \sigma^2 \mid X, y)$
BLUE if errors are spherical	Requires normality or for a specific distribution
Only good for linear	Extends to non-linear

Only good for linear models	Extends to non-linear models
Statistical-based inference	Likelihood-based inference
Not sensitive to distrib. misspecification	Sensitive to dist. mis.
Less computationally extensive	More computationally extensive

Another good comparison by Claude:



Notice we are now stepping into territory of some bias is okay as long as we

get efficiency and flexibility...

Building a model

After reviewing these challenges, your head should be screaming "how do I even build a model then?". Assumptions are often violated in the real world, we have imperfect information about the DGP, and sometimes even theory can be outdated.

Unfortunately, there is no short-cut. Fortunately, that means our jobs will remain in demand even if AI takes over. This is where the art of modeling comes in, and any creative

endeavor requires trial and error.
Data Science has squeezed
a lot out of automated trial and
error. Thanks to cheap compute,
nowadays we can just try and
test as many models as we want.
The focus then falls on selecting
the one the best optimizes your
criteria, rather than focusing on
building
the perfect model from the start.

So for a general guideline I use:

- Survey the literature to find
someone
that modeled the same thing and
got
it peer-reviewed.

- If no attempt suitable to your case exists, then assume you won't have a strike of genius to figure it out.

  Shift your focus to the data you need. Grab as much of it you can. Clean it and construct the most complicated known to earth. Iteratively test simpler versions (using Ocam's razor) until you arrive to a logical solution with good performance. Test, test, test.

The key challenge is to balance model fit and complexity (# of parameter).

I didn't just pull this advice from a magic hat. Specification has been a huge problem for econometricians and modelers since models became a thing. Generally, there have been two approaches I've found (Peter Kennedy's "Guide to Econometrics" is a great source).

Before we get into more details of model selection, let's remind ourselves why specification matters.

Broadly speaking, the specification problem deals with deciding what variables to include in the data matrix X and which functional form to

apply
(linear, log-linear, non-linear, etc.)

There are three problems that we may face:

1) OVB (under-specification)
   - True model: $Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon$
   - Estimate: $Y = X_1 \beta_1 + \varepsilon$
   - Bias: $E[\hat{\beta}_1] = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$
     ↳ Depends on correlation between included (X_1) and omitted (X_2) variables

2) Over-Specification
   - Unbiased but inefficient estimates

- Variances gets inflated
- higher chance of type II error and harder to interpret

3) Form misspecification
   - Assume linearity when problem is inherently non-linear
   - Biased and inconsistent estimates

If you've chosen OLS, then we must assume linearity. Always check if the relationship actually looks linear. Very hard to plot when $k > 2$ but there exist tests to do so numerically. Eg, Ramsey RESET test, Box-Cox transform., partial correlation plots, and probably a bunch more I have no idea about.

Alright, hopefully Linearity is not a horrible choice. We must weigh the trade-offs between under- and over-specification. We've learnt ways to
diagnose and deal with inflated variances. But we have no way to correct for the things we are missing. Because OVB leads to bias, we should prefer a recipe for building models that reduces the probability of OVB. Enter the kitchen sink approach. More formally known as the General-to-Specific approach.

1) Start with a model $y = X\beta + \varepsilon$ that includes all variables (and

their transformations) that you think could be relevant.

2) For each variable $x_j \in X$ test

$$H_0 : \beta_j = 0 \quad , \quad H_1 : \beta_j \neq 0$$

3) Remove the least significant variable
   - First degree of freedom: You have to choose a significance level α.

4) Re-estimate the model

5) Repeat 2-4 until all remaining variables are significant.

The test to run is a sequential F-test

$$F_j = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2 \cdot (X^T X)^{-1}_{jj}} \sim F_{1, n-k}$$

$(X^T X)^{-1}_{jj}$ is the j^{th} diagonal element of $(X^T X)^{-1}$

David Henry, from LSE, is known for promoting this procedure. He outlined five important principles:

a) Congruence: Models should be consistent with available evidence and theory
   - Otherwise we are doing DS!

b) Encompassing: Final model should explain competing model's results.
   - If the model predicts

something
that contrasts economic intuition,
then
be weary.

c) Parameter Constancy: Estimated coefficients
should be stable across samples
- If you change the data or drop a variable and one of the estimates changes drastically, then inspect further before proceeding.

d) Data Coherence: Residuals should be
white noise
- If the residuals seem informative to explain y then

we are prob. missing an important variable

e) Theory Consistency: Models should align with economic theory

The main advantages is mitigating OVB, helping to account for multicollinearity, and it is a systematic procedure.

But it is not without limitations. For one, data might be a problem. Perhaps you don't even know if you are missing an important variable in your kitchen sink. Also, huge models are more expensive to estimate. Although not a big issue nowadays. A third, more important,

limitation is that exclusion decisions might be trivial. T- or F- tests are not some sort of holy grail. You still need to practice agency and follow intuition. Lastly, the final model is path dependent. There is a chance that the first variable you drop conditions which variables are more likely to be insignificant later on.

Today we have automated algorithms to do this. I learnt this while writing these notes. Checkout Automated GETS algorithms by Jurgen A. Doornik

Model Selection

So now we have a way to build

models. But how do we choose amongst competing models?

The idea behind model selection strategies is to define criteria based on quantitative metrics to evaluate and compare models that have already been specified.

When we talk about fit vs complexity, we really are thinking about the bias-variance tradeoff, remember?

The total error metric is often operationalize as Mean Squared Error (MSE)

$$MSE(\hat{\beta}) = E[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] = Bias(\hat{\beta})^2 + Var(\hat{\beta})$$

Usually, the more complex the model is the lower the bias. But variance increases. An optimal model would balance these two to minimize MSE.

From a prediction perspective, this decomposition makes it very easy to see that regardless of our ultimate objective (explain or predict) this trade-off is something we must deal with.

Suppose we have the model $\vec{Y} = X\beta + \vec{\varepsilon}$ and we get a new observation $\vec{x}_0$. The prediction is $\vec{y}_0 = \vec{x}_0^T \beta + \vec{\varepsilon}_0$

(MSE) — true model (not fitted)

$$E\left[(y_0 - x_0^T \hat{\beta})^2\right] = \sigma^2 + \text{Bias}(x_0^T \hat{\beta})^2 + \text{Var}(x_0^T \hat{\beta})$$

Prove this

$$\text{Var}(\varepsilon) = E[\varepsilon^2] + E[\varepsilon]^2 \quad \to 0 \text{ from exog.}$$

$$\text{Bias}(\hat{\beta}) = \beta - E[\hat{\beta}] \quad \to \text{deterministic}$$

$$\text{Var}(\hat{\beta}) = E[\hat{\beta}^2] + E[\hat{\beta}]^2$$

Proof

$$E\left[(y_0 - x_0^T \hat{\beta})^2\right] = E\left[(x_0^T \beta + \varepsilon_0 - x_0^T \hat{\beta})^2\right] \quad (\varepsilon_0)$$

$$= E\left[(\varepsilon_0 + x_0^T(\beta - \hat{\beta}))^2\right]$$

$$= E\left[\varepsilon_0^2 + 2\varepsilon_0 x_0^T(\beta - \hat{\beta}) + (x_0^T(\beta - \hat{\beta}))^2\right]$$

(ε_0: given; x_0^T: constant)

$$= E[\varepsilon_0^2] + 2\,E[\varepsilon_0]\,x_0^T(\beta-\hat{\beta}) + E[(x_0^T(\beta-\hat{\beta}))^2]$$

↳ exogeneity

$$= \sigma^2 + 0 + ??$$

* this term is tricky.

Let $z_0 = x_0^T(\beta-\hat{\beta})$

$$E[z_0^2] = Var(z_0) + E[z_0]^2$$

$$E[z_0] = E[x_0^T(\beta-\hat{\beta})] = x_0^T\,Bias(\hat{\beta})$$

$$Var(z_0) = Var(x_0^T(\beta-\hat{\beta}))$$

$$= Var(x_0^T\beta - x_0^T\hat{\beta})$$

$$= Var(x_0^T\beta) + Var(-x_0^T\hat{\beta})$$

$x_0^T\beta$ ↳ constant

$-x_0^T\hat{\beta}$ ↳ constant

$$= 0 + Var(-x_0^T\hat{\beta}) = Var(x_0^T\hat{\beta})$$

$$= x_0^T \, \mathrm{Var}(\hat{\beta}) \, x_0$$

$$\Rightarrow E[z_0^2] = x_0^T \, \mathrm{Var}(\hat{\beta}) \, x_0 + \left(x_0^T \, \mathrm{Bias}(\hat{\beta})\right)^2$$

Putting everything together

$$E\left[(y_0 - x_0^T \hat{\beta})^2\right] = \sigma^2 + x_0^T \, \mathrm{Var}(\hat{\beta}) \, x_0 + \left(x_0^T \, \mathrm{Bias}(\hat{\beta})\right)^2$$

irreducible error — Variance — Bias squared

We could also apply a "decomposition" trick and write

$$\left(x_0^T (\beta - \hat{\beta})\right)^2 = \left(x_0^T (\beta - E[\hat{\beta}] + E[\hat{\beta}] - \hat{\beta})\right)^2$$

and then solve. It'll give you the same result.

* Have one group solve it with the decomposition trick and one group without. The idea is that we are decomposing the estimation error $(\beta - \hat{\beta})$ into a "systematic" error (Bias) and a "random" error (variance)

$\beta - E(\hat{\beta})$ is the Bias

$E[\hat{\beta}] - \hat{\beta}$ is the fluctuation of $\hat{\beta}$ around its mean (i.e., variance)