Metrics-Math Bootcamp Day 2

Cameron Taylor

August 13, 2019

Stanford GSB

Overview

1. Intro to OLS
2. OLS derivation
3. Assumptions and Stat Properties
4. Endogeneity
5. OLS in R

- Most of you will have seen these concepts and ideas before, but hopefully can push to deeper level of understanding and get a bit more comfortable with technical tools and econometric intuitions

Intro to OLS

Why OLS?

- OLS is the bread and butter of applied economics
- I think a regression is run in almost every applied economics paper with some data that I have seen
- We will see why: it is so easy to implement and interpret and quite powerful
- But, I think it's ease of implementation makes it deceptively subtle, and so econometric and economic intuition inform what we learn from it

Why OLS? And a little bit of What

- Recall the **CEF**: $E[Y_i|X_i]$ where X_i is (potentially a vector of) random variable(s) and Y_i is our dependent variable of interest
  - This is a nice summary of the relationship between these variables: how the average of Y moves with X
- **OLS approximates the CEF linearly** by $X_i b$ where b is some vector of parameters, and does it very well
- A Thm from Mostly Harmless (Thm 3.1.6):

$$\beta_{OLS} = \text{argmin}_b E\left\{\left(E[Y_i|X_i] - X_i b\right)^2\right\}$$

- There are other senses in which OLS is good (Gauss-Markov) that I won't mention

OLS Derivation

OLS Derivation: Setup

- OLS is derived by minimizing the mean squared error in a linear model
- Suppose we have a sample (Y_i, X_i) with N data points i.i.d.
- At the population level:

$$\min_b E((Y_i - X_i'b)^2)$$

- At sample level - use analogy principle:

$$\min_b \frac{1}{N} \sum_i (Y_i - X_i b)^2$$

OLS Derivation: Algebra

- These are vectors - vector and matrix calculus is not very common, but essentially take FOC as if unidimensional
- Population level:

$$E[X_i'(Y_i - X_i b)] = 0 \Rightarrow b = E[X_i'X_i]^{-1}E[X_i'Y_i]$$

- Sample level: for sample, first easier to remove $1/N$ and then write the objective function as

$$(y - Xb)'(y - Xb)$$

where we stack everything and then the FOC is

$$X'Xb - X'y = 0 \Rightarrow b = (X'X)^{-1}X'y$$

- The dimensions may be confusing: just make sure to follow the matrix multiplication rules and you should be fine!

OLS Estimator

- Thus we have our OLS estimator

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

- What assumptions already needed for this to work? Clearly need $X'X$ invertible!
- $X'X$ only invertible if X has full column rank - if $X'X$ not invertible then columns linearly dependent, so cannot distinguish between effects in the linear model

OLS Interpretation

- To interpet: use Frisch Waugh Lovell Theorem
- Suppose that model is $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$
- Then the interpretation of β_1 is the effect of x_1 on y where both x_1 and y are residualized with respect to x_2
- Formally: let $M_2 = I - x_2(x_2'x_2)^{-1}x_2'$. This is the "residual maker" for x_2 (Why?)
- Then $\hat{\beta}_2 = ((M_2 x_1)'(M_2 x_1))^{-1}((M_2 x_1)'(M_2 y))$, the OLS esitmate in the model
- Can simplify using properties of M_2 you did (or might have done) in HW:

$$\hat{\beta}_2 = (x_1' M_2 x_1)^{-1} x_1' M_2 y$$

Assumptions and Statistical Properties

Assumptions for OLS

- There are many assumptions one can make to give OLS estimators nice properties
- For example, when writing $y = X\beta + \epsilon$ assuming that ϵ is an independent normal gives the OLS estimator great properties
- I will try to look at the assumptions that correspond to the results most often invoked in the research I have seen (ex: normality not often assumed)

OLS Assumption Overview

Consider model

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

1. (Full Rank) X has full rank - obvious why we need it
2. (Homoscedasiticty) $V(\epsilon) = \sigma^2 I$ (recall ϵ is a vector)
3. (Exogeneity) $E[\epsilon|X] = 0$
4. (I.i.d.) ϵ_i is i.i.d. across observations

Probing Assumptions

- The full rank assumption can be directly tested
  - It often only fails due to the dummy variable trap: include too many dummies so that constant is spanned by all dummies. To avoid, always leave out a comparison group.
- Homoscedasitity: can plot errors and see if move with X - less worried about this for causal inference, etc. (I see much less heteroscedastic corrections in economics papers). If this is not true, will change standard errors calculation - will go over briefly.
- Exogeneity: this is the big one; much of empirical research is based on plausible "identification strategies" to deal with this. We will talk more in depth about it
- I.i.d.: can assess based on situation (ex: unlikely true in time series, etc.)

Statistical Properties Overview

- Unbiased estimator
- Simple form for variance of estimator; other forms for variance
- Consistency
- The estimator has an asymptotic normal distribution

Unbiasedness

- The OLS estimator is unbiased for the population coefficient if the model is correct
- In other words: $E[\hat{\beta}_{OLS}|X] = E[\hat{\beta}_{OLS}] = \beta$ where β is the true population coefficient
- To prove this: look at formula for the OLS estimator and use the exogeneity condition
- Most useful step: write $y = X\beta + \epsilon$ and then multiply out

Variance and Standard Errors

- Important for inference and std errors to know variance of estimator
- Under homoskedasticity, can compute relatively easily using linear algebra tricks (important to know how trace multiplication works; also recall variance of random vector)

$$\begin{aligned} V(\hat{\beta}_{OLS}|X) &= E[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)|X] \\ &= E((X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

- For in sample, compute $\hat{\sigma}$ using residuals and degrees of freedom correction
- Note that this directly gives standard errors of estimator (square root of diagonal of this matrix)

Alternate Variance Assumptions and Standard Errors

- Two standard ways to approach dealing with non-homoscedasticity
- First, can use the *robust* SE estimator:

$$(X'X)^{-1}(\sum_i e_i^2 x_i x_i')(X'X)^{-1}$$

  where e_i is the i-th residual and x_i is the k vector for the i-th observation
- This allows for arbitrary structure of variance-covariance along the diagonal
- Alternatively can model the heteroscedasticity directly and use *Generalized Least Squares* (example: auto-correlation, etc.)
- My sense is that in applied work robust SE is more common, but some critiques of it (see in Mostly Harmless)

Consistency

- Recall consistency: want the estimator to converge to the population object as our sample size grows
- Under the exogeneity conditions and some technical assumptions on X, we get consistency of the OLS estimator using a Law of Large numbers
- The proof is very similar to the proof of unbiasedness and consists of expressing the estimator as means and applying Law of Large Numbers and Continuous Mapping and Slutsky Theorem

Asymptotic Normal Distribution

- Consider the object
$$\sqrt{n}(\hat{\beta} - \beta)$$
this is the object we perform asymptotic analysis on
- This object is equal to
$$(\frac{1}{n}X'X)^{-1}\frac{1}{\sqrt{n}}X'e$$
- Treat X as constant and by the exogeneity condition, $E[X'e] = 0$ and so using a CLT, get that this will have an asymptotically normal distribution
- Thus OLS estimator has an asymptotic normal distribution!
- Useful for asymptotic inference - usually what we refer to even if n is not huge (but highlights usefulness of large n!)

Testing

- The asymptotic results gives a basis for testing hypotheses within the model
- There are many different types of tests we might formulate for the underlying parameter β that we estimate
- I will go over some here
- The basic formula consists of: state the null in terms of the OLS estimator; find the asymptotic distribution and variance-covariance matrix of the form of the null hypothesis, and then see what critical values allow you to reject the hypothesis at different levels (usually 95% level)

Classic t-Test

- The most common test is whether some specific parameter $H_0 : \beta_k = 0$
- The null of this test states that factor/variable k has no statistical impact in this linear model on y
- This is very common so comes out automatically in most regression output (R, Stata, etc.)
- To test: use asymptotic inference

$$\hat{\beta}_k / \text{S.E.}(\hat{\beta}_k)$$

  will have a standarx normal distribution
- Rule of thumb: since use 95% and 2-sided test, z-value is about 2 for something to feel comfortable "rejecting"
- Can generalize to $H_0 : \beta_k = \beta_0$ by having $\hat{\beta}_k - \beta_0$ in numerator

F-Test

- A common test for how "good" you're model is is to test $H_0 : \beta_k = 0, \forall k$
- This says: just guessing a constant is better than using any of the X's
- Also common enough that shows up in most regression output (R, Stata, etc.)
- This is also closely related to the R^2 ("R-squared") of a regression, which measures the fit of the model
- Not always a useful "economic" tool, but useful for basic model assessment

General Linear Test: Wald Test

- In general, any linear restriction of all the coefficients β can be tested using the *Wald* test
- By linear restriction I mean null takes form:

$$R\beta = q$$

  where R is a matrix and q is a vector
- Both the standard t and F test above are special cases of this
- See Greene or any standard econometrics book for more details on this

Non-Linear Tests

- Sometimes we might be interested in non-linear tests: for example $\beta_k^2 = 1$
- In general, the key tool for these tests is the *Delta Method*
- This allows for an asymptotic approximate to a test $c(\beta) = q$ where $c(\cdot)$ is any function and q is some vector
- I won't go over details, but can find details in any standard econometrics book

Endogeneity

Endogeneity Intro

- Think about two things
- First: what are consequences of failure of assumption $E[\epsilon|X] \neq 0$? (or even $E[X'\epsilon] \neq 0$)
- Second: how does one know when $E[\epsilon|X] = 0$, or how can one achieve this?
- If the first fails, then we have inconsistency and mispecified asmptotic distribution - statistical inferences will be "wrong"
- Sometimes we know *how* it is wrong (direction, for example), but often not good enough to know this
- Major problem: cannot test this assumption since ϵ is unobserved (by definition!)
- Will discuss some common scenarios of failure

Omitted Variable Bias

- Consider univariate situation and normalize all variables to have mean 0 for simplicity $y = \beta x + \epsilon$
- Suppose that $\epsilon = \eta w + \nu$ and that $E[xw] \neq 0$ where ν is independent random noise
- Then $y = \beta x + \eta w + \nu$
- Run OLS:

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\sum_i x_i(\beta x_i + \eta w_i + \nu_i)}{\sum_i x_i^2} = \beta + \beta\eta \frac{\sum_i x_i w_i}{\sum_i x_i^2} + \frac{\sum_i x_i \nu_i}{\sum_i x_i^2}$$

- What is the expectation and probability limit of this?

Reverse Causality

- Suppose that want to claim that x has effect on y through $\beta \neq 0$ in the model $y = \beta x + \epsilon$
- However, what if the true data generating process is: $x = \delta y + \nu$ is the correct model - i.e. y changes x and x has no effect on y
- Then $y = \frac{1}{\delta} x - \frac{\nu}{\delta}$ and so $\beta \neq 0$ will be found in the misspecified model!
- So need some way to argue that direction is the direction of importance

Measurement Error

- Suppose that we have an imperfect measure of x, $\tilde{x} = x + \nu$ where ν is mean 0 independent noise (classical measurement error)
- How will this effect the estimate?
- Model is now

$$y = \beta\tilde{x} + \epsilon = \beta\nu + \beta x + \epsilon$$

- The OLS estimator is

$$\hat{\beta} = \frac{\sum_i y_i \tilde{x}_i}{\sum_i \tilde{x}_i^2}$$

- It turns out this will always be *attenuated* - biased towards 0 (do this in HW!)
- For applied people, not as much a worry if looking for an effect

Acheiving Identification

- So how do we make sure the assumption is verified?
- Large part of Mostly Harmless devoted to this
- Idea: Random experiments are ideal, so look for "experiments in the data" - where x, the variable(s) of interest are randomly assigned (or quasi randomly assigned, or conditionally randomly assigned, etc.)
- Example: Suppose we wanted to look at effect of the sex of a child on labor decisions. Is the sex of a child randomly assigned?

OLS in R

Dataframes in R

- To take full advantage of R's regression, could to use dataframes
- These are basically "special matrices" where columns are named as variables
- Once a dataframe is specified, R has clean syntax for running regressions
- Oftentimes when read in data, R will read in as a dataframe (will get experience with this)

Dataframes in R

```
> # Dataframes
> set.seed(1)
> x1 <- rnorm(10)
> x2 <- rnorm(10)
> y <- 1+x1+2*x2+rnorm(10)
> df <- data.frame(y,x1,x2)
> df
             y         x1          x2
1   4.31608590 -0.6264538  1.51178117
2   2.74546610  0.1836433  0.38984324
3  -1.00354479 -0.8356286 -0.62124058
4  -3.82347067  1.5952808 -2.21469989
5   4.19919536  0.3295078  1.12493092
6   0.03353566 -0.8204684 -0.04493361
7   1.29925302  0.4874291 -0.01619026
8   2.15524474  0.7383247  0.94383621
9   2.74007369  0.5757814  0.82122120
10  2.30035582 -0.3053884  0.59390132
```

Figure 1: Dataframes

lm() Function

- lm() stands for linear model
- The basic setup is
  lm(y $\sim$ x1+x2+...+xk, data=dataframe)
- To get the most information out of the regression output define the regression object as something
  reg $<-$ lm(...)
- Then apply the summary() function over the lm() object
- This will apply the tests we have discussed, estimates, etc.

Regression Example

```
> reg <- lm(y~x1+x2, data=df)
> summary(reg)

Call:
lm(formula = y ~ x1 + x2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28920 -0.05421  0.01321  0.06747  0.22564

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.96795    0.05481   17.66 4.60e-07 ***
x1           0.82584    0.07464   11.06 1.09e-05 ***
x2           2.08704    0.05448   38.31 2.15e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1619 on 7 degrees of freedom
Multiple R-squared:  0.9953,    Adjusted R-squared:  0.994
F-statistic: 740.4 on 2 and 7 DF,  p-value: 7.143e-09
```

Figure 2: Regression Output

Wrap-Up

- OLS is super useful for describing data - has a very clean nice interpretation
- Some basic tools to do statistical inference on estimates - OLS has really nice properties
- Combine with economic modeling and conceptual frameworks to interpret estimates in economic way
- Extremely powerful when combined with causal inference and taking endogeneity seriously
- Very easy to implement in R (and other softwares)

References for Content

- Mostly Harmless has excellent material on regression fundamentals and understanding regression (Read Ch 3 is your HW)
- Greene has quite a bit of material on regression, but less intuitive and more useful for the consistency and asymptotic distribution proofs

HW

1. Read Chapter 3 of Mostly Harmless Metrics
2. Do out the matrix algebra on the OLS Derivation: Algebra slide. Why is the FOC sufficient in this case?
3. Prove that the OLS estimator is unbiased when the exogeneity condition holds.
4. Prove that OLS has attenuation bias in the univariate model with classical measurement error
5. Demonstrate the Frisch-Waugh-Lovell Theorem in an example in R using lm()
6. Run a regression on simulated (or real!) data using the lm() function and using the matrix formula for the OLS coefficient $(X'X)^{-1}X'Y$. Bonus: calculate (plain) standard errors only using matrix and vector algebra in R.