

# Metrics-Math Bootcamp Day 1

---

Cameron Taylor

August 5, 2021

Stanford GSB

# Overview

1. Intro to Metrics
2. Intro to Stats
3. Intro to Linear Algebra
4. Wrap-up, R, HW

# Metrics

---

# What is Metrics?

- “Econometrics is the application of statistical methods to economic data in order to give empirical content to economic relationships.”
- Economics research broadly consists of three things: theory, data, and econometrics
- Econometrics is the thing that allows you to say something about the theory with the data you have
  - Note that data is an important input! “Garbage-in-garbage-out”
- Common empirical idea/questions: What is causal effect of  $X$  on  $Y$ ? What will be the impact of changing from policy  $A$  to policy  $B$ ?

# Brief General Ideas

- Empirical research generally consists of specifying a model of the data  $X$  and some parameters  $\theta$ ,  $M(X, \theta)$ , and then optimizing some criteria of the model to back out which parameters seem “best” given the data
- OLS will choose parameters  $\theta$  linear in model, and then minimize the sum of the squared deviations between model predictions and data
- MLE and GMM will also do something similar
- Key ideas in each come from statistics and require linear algebra - we will go over both very briefly (MGTECON 603 does this in more depth - a good class that many RF's before took)

# Stats

---

# Random variables

- Many phenomena we can measure with data have "randomness"
- Building block: random variables
- Random variables used to express outcomes with random occurrences
- A random variable  $X$  takes values from some sample space  $\Omega$  into a measurable space, which for now will use  $\mathbb{R}$  (so it is a function)
- $\Omega$  also specifies probability of different events happening
- $Pr(X = x) = Pr(\{\omega \in \Omega | X(\omega) = x\})$
- Examples
  - Number on die role:  $\Omega = \{1, 2, 3, 4, 5, 6\}$ , uniform probability distribution,  $X(\omega) = \omega$ .
  - Coin flip outcome is heads:  $\Omega = \{H, T\}$ ,  $Pr(\omega = H) = Pr(\omega = T) = 1/2$ .  $X(\omega) = 1$  when  $\omega = H$ .

# Distribution and Expectation

- How to describe a random variable?
- A “full” way to do it: CDF  $Pr(X \leq x) = F(x)$  and PDF  $Pr(X = x) = f(x)$ 
  - Why full? Any info we want about the random variable comes from this
- Don't usually take this approach unless doing some descriptive stats because: (1) hard to report and (2) hard to estimate the whole thing
- Instead, usually report other summaries of data
- A particularly useful one: **expectation**  $E(X) = \int xf(x)dx$  (“average”); can generalize to functions  $E(g(X)) = \int g(x)f(x)dx$
- Also care about “spread”- variance  $V(X) = E(X^2) - E(X)^2$
- There are many nice properties about expectations that one can derive: the best one is probably linearity, can look up online

# Joint Dist

- While summarizing a single variable interesting, economics more concerned with relationship between multiple variables
- Consider joint distribution of  $(Y, X)$ , two random variables or a random vector
- These also have a joint dist which can be described by  $f(x, y) \approx \Pr(X = x \text{ and } Y = y)$
- Similarly though, want different ways to summarize

# Covariance, Correlation and Conditional Expectation

- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$  a way to express how  $X$  and  $Y$  move together
- $\text{Cor}(X, Y)$  normalizes the covariance to be in  $[-1, 1]$
- A particular object of interest is  $E[Y|X]$ , the conditional expectation function (CEF). This is a random variable since  $X$  is a random variable
- $E[Y|X = x]$  can be seen as “best guess of  $Y$  given  $X = x$ ” (See Mostly Harmless Thm 3.1.2)
- Note that we can always write  $Y = E[Y|X] + \epsilon$  where  $E[\epsilon|X] = 0$  (Why? HW!)
- A useful tool: Law of Iterated Expectations:  $E(Y) = E(E(Y|X))$

# Independence of Random Variables

- Intuitive: Two random variables are independent if knowing the value of one variable does not help you predict what the value of the other variable will be
- Technically: We can write the joint density as  $f(x, y) = f_X(x)f_Y(y)$ , product of densities
- Independence is a common assumption we make about how the data is generated  $\rightarrow$  makes estimation a lot easier
- Exercise: does covariance of 0 imply independence between two random variables?

## Concepts → Data

- So far everything has been abstract. How do we use these concepts to think about data?
- We make assumptions about how the data is generated
- We usually think about the data we collect as a **sample** (or random sample)
- Standard assumption: the data  $(x_1, \dots, x_n)$  we collect consists of independent random variables each with the same distribution, described by  $f(x)$
- Our goal: what are the properties of the data using these assumptions?
- We usually estimate objects of interest about our data using statistics or estimators: any function of our data (ex: mean of data)
- Note that because our data are random variables, statistics and estimators are random variables!

# Statistics and Estimator Example

- We have a random sample of heights from Stanford students  $(x_1, \dots, x_n)$  which we assume come from  $N(\mu, \sigma^2)$  distribution
- How can we estimate mean height of Stanford students?
- One trick: analogue principal
  - Sample analog of population:  $\hat{E}(X) = \bar{x} = \frac{1}{n} \sum_i x_i$  (this is like a numerical integral)
- Can show that the expectation of the sample mean is  $\mu$  in this case
- A natural question is how close this will be to the true value, particularly as a function of the number of students we sample
  - To look at this, can look at the variance of the sample mean
- Another natural question: if we add more and more data, will we eventually converge to the correct value?
  - This is called asymptotics, key to metrics theory

# Key Statistical Concepts for Metrics: Small Sample

- Bias
  - An estimator is unbiased if its expectation equals the population object
  - Key tools: Properties of expectations, law of iterated expectations
  - Ex: We will see that OLS is unbiased under some assumptions
- Variance
  - An estimator's variance measures how much it changes between different samples
  - Key tools: Properties of expectation, variance
- Note: Estimators/statistics will be random because they are functions of data which are random!
- Note: An important idea in stats/econometrics is the bias/variance trade-off. More present in non-parametric methods where need to “tune”

# Key Statistical Concepts for Metrics: Asymptotics

- Asymptotics: What happens with lots of data?
- Consistency
  - An estimator is consistent if it “converges in probability” to the true object (can make as close in probability space with more data)
  - Key tool: Law of Large Numbers
  - Idea: The average is consistent in probability under general conditions
- Asymptotic Distribution
  - Since estimators random, will have an asymptotic distribution that they approach as we use more data
  - This is useful because then we can do “asymptotic inference” with estimators
  - Key tool: Central Limit Theorem
  - Idea: The sum of independent random variables converges to a normal distribution

# Linear Algebra

---

# Linear Algebra

- Linear Algebra is an area of mathematics studying linear mappings between vector spaces: we will focus on simple vectors and matrices with real value entries
- A solid understanding of matrices and vectors is essential for studying econometrics, since we often represent our data and model parameters as matrices and vectors
  - Example: write OLS model as

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

where  $y$ ,  $\beta$ ,  $\epsilon$  vectors,  $X$  is matrix

- We will focus on important properties of vectors and matrices for metrics

# Matrices

- A *matrix* is a two dimensional grid

$$\mathcal{M} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

- Each *row* is read left to right: the first row of the above matrix is  $\begin{bmatrix} a & b & c \end{bmatrix}$
- Each *column* is read top to bottom: the first column of the above matrix is  $\begin{bmatrix} a \\ d \end{bmatrix}$
- We call this matrix  $2 \times 3$  (# rows  $\times$  # cols)

# Matrix Transpose

- The transpose of a matrix  $\mathcal{M}$  is denoted  $\mathcal{M}'$  - you rotate the entries

$$\mathcal{M} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

$$\mathcal{M}' = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

- The transpose of an  $N \times K$  matrix is thus a  $K \times N$  matrix.

# Matrix Multiplication

- Let  $A$  be an  $N \times K$  matrix, and let  $B$  be a  $K \times M$  matrix. The *matrix product*  $AB$  is a  $N \times M$  matrix with  $[AB]_{ij} = \sum_l A_{il} B_{lj}$
- For this to be well defined  $A$  must have the same number of columns as  $B$  has rows.

$$(N \times K) \times (K \times M) = (N \times M)$$

$$(N \times K) \times (J \times M) = \text{Not possible!}$$

- Importantly for transposes:  $(AB)' = B'A'$

# Identity Matrix

- A *square* matrix is an  $N \times N$  matrix, i.e., a matrix with the same number of rows and columns
- A *diagonal matrix* is a square matrix whose nondiagonal entries equal 0.
- The  $N$  dimensional *Identity matrix*, often denoted  $I_N$  or just  $I$ , is a square matrix whose diagonal entries are all 1:

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Can verify that for any  $M \times N$  matrix  $A$ ,  $AI_N = A$  and  $I_N A' = A'$ , hence identity

# Matrix Inverse

- Some square matrices are *invertible*. An  $N \times N$  matrix  $\mathcal{M}$  is invertible if and only if there exists an  $N \times N$  matrix  $\mathcal{M}^{-1}$  such that

$$\mathcal{M}^{-1}\mathcal{M} = \mathcal{M}\mathcal{M}^{-1} = I_N$$

- And we call  $\mathcal{M}^{-1}$  the inverse of  $\mathcal{M}$ . It is not hard to prove that if an inverse exists, it is unique.
- There is a deep result that almost all square matrices are invertible, more on this later. The inverse is not well defined for nonsquare matrices.

# Matrix Addition

- Matrix addition works exactly the way you want it to. Let  $A$  and  $B$  be  $N \times K$  matrices:

$$[A + B]_{ij} = A_{ij} + B_{ij}$$

- Matrix addition is only defined if  $A$  and  $B$  have the same dimensions
- Componentwise multiplication or addition by a scalar (number) is also straightforward: If  $A$  is a matrix and  $c$  is a number,  $cA$  the matrix  $A$  with each element multiplied by  $c$ . Informally,  $c + A$  is the matrix whose  $ij$ -th element is  $c + A_{ij}$ .
  - This often comes up in programming languages where the code interpreter will make assumptions about what you want to happen when you add a vector or constant to matrix

# Linear Combinations

- Recall a vector is a  $N \times 1$  or  $1 \times N$  dimensional matrix.
- Let  $V^1, V^2, \dots, V^K$  be a collection of  $K$  column vectors, or  $N \times 1$  dimensional matrices
- We just learned that the sum  $\sum_{i=1}^K V^i$  of the  $K$  vectors will also be a  $N \times 1$  dimensional column vector. We also learned that if we multiply  $V^i$  by a scalar  $c_i$ , then  $c_i V^i$  is an  $N \times 1$  dimensional vector.
- A *linear combination* of the vectors  $V^1, \dots, V^K$  is the sum of the scalar multiples  $\sum_{i=1}^K c_i V^i$  for any arbitrary numbers  $c_1, c_2, \dots, c_K$
- The same concept applies to a collection of  $1 \times N$  dimensional row vectors

# Linear Independence

- A collection of column vectors  $V^1, V^2, \dots, V^K$  is said to be *linearly dependent* if one of the vectors  $V^l$  can be written as a linear combination of the other  $K - 1$  vectors:

$$V^l = \sum_{i \neq l} c_i V^i \text{ for some } \{c_i\}_{i \neq l}$$

- A collection of column vectors  $V^1, V^2, \dots, V^K$  is *linearly independent* if they are NOT linearly dependent.
- For example,  $\left( \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)$  is a linearly dependent collection, because

$$1 * \begin{bmatrix} 1 \\ 1 \end{bmatrix} + (-1) * \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

# Rank

- For any  $N \times K$  matrix  $A$ , the  $K$  columns of  $A$  are a collection of  $N \times 1$  dimensional column vectors.
- The *column rank* of a  $N \times K$  matrix  $A$  is the maximum number of columns that can form a linearly independent collection. For example, the  $2 \times 3$  matrix

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

has a column rank of 2: we just saw that the three vectors are linearly dependent, but

$\left( \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)$  are linearly independent.

# Full Rank and Invertibility

- For any  $N \times K$  matrix  $A$ , we say the matrix  $A$  has *full column rank* if it has column rank  $K$ .
- It turns out that for any  $N \times K$  matrix  $A$ , the  $K \times K$  square matrix  $A'A$  is invertible if and only if  $A$  has full column rank.
- The  $N \times N$  square matrix  $AA'$  is invertible if and only if  $A$  has full row rank.
- In general, an  $N \times N$  matrix  $B$  is invertible if and only if  $B$  has full column rank.

# Wrap-Up

- Hopefully you have an idea of
  - What metrics is
  - How stats is useful for metrics and some basic stats concepts
  - How linear algebra is useful for metrics and some basic understanding of matrices and vectors

## Some Extra Resources for Today's Content

- For intro to metrics: Mostly Harmless has a nice intro, Angrist and Pischke have a nice JEP piece (2010)
  - For an alternate viewpoint: can look at Frank Wolak and Peter Reiss *Structural Econometric Modeling: Rationales and Examples from Industrial Organization* - talks about reduced form vs. structural
  - Also Deaton (2010) on randomized control trials
- For intro stats for metrics (and general tools related to metrics): MGTECON 603 is a good intro, even a bit of overkill
  - Goes over some stuff we did not have time to do from stat perspective: MLE, hypothesis testing, suff stats, more on probability theory, etc.
  - Accompanying book (which I like): Casella and Berger *Statistical Inference*
- For linear algebra: I really like Gilbert Strang *Intro to Linear Algebra*

# Stats and LA in R

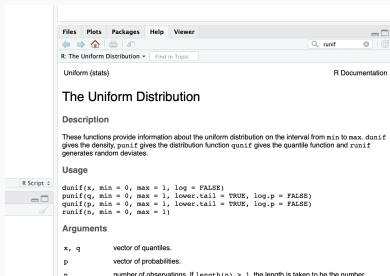
---

# Overview of Stats and LA in R

- Many of you are familiar with R or use something similar (like Python)
- For this bootcamp I don't care what language you use, but I think R is a good one (seems like many RFs used it in the past)
- More important that you get practice applying the concepts on real data than learning specific syntax
- Here: some basic stats and matrix algebra in R

# Intro to R

- R has a nice editor (RStudio) and a great online community
- Some important basics
  - `#` symbol for comments
  - People use arrows to define variables (I don't know why) but you can use equals
  - The “Help” sidebar in RStudio is extremely useful



# Simulating in R

- I think simulating is extremely useful for understanding basic metrics problems and working through examples
- In R usually specify r-DISTRIBUTION NAME to simulate data. First argument is number of draws, and then require parameter specifications.
- R stores draws in a vector...

```
# Simulating data
n_samp <- 10
set.seed(1)
data1 <- rnorm(n=n_samp, mean=0, sd=1)
data2 <- runif(n=n_samp, min=0, max=1)
data1
data2
```

```
> data1
[1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078 -0.8204684  0.4874291  0.7383247
[9]  0.5757814 -0.3053884
> data2
[1] 0.93470523 0.21214252 0.65167377 0.12555510 0.26722067 0.38611409 0.01339033 0.38238796
[9] 0.86969085 0.34034900
```

Figure 2: Simulating

# Vectors in R

- The notation `c()` is used for defining vectors
- Adding vectors and multiplying works very intuitively: for example `a*b` will multiply each entry of the vector if the same length

```
> # Define a vector
> a <- c(1,2,3)
> b <- c(4,5,6)
> # Multiply by constant and vector
> a*2
[1] 2 4 6
> a*b
[1] 4 10 18
```

**Figure 3:** Vectors

# Matrices in R

- There are multiple ways to define matrices
  - Can use the `matrix()` command
  - I prefer using `cbind()` or `rbind()` which combines matrices
- `%*%` is used for matrix multiplication (including multiplying vectors by matrices) (make sure dimensions match! R will tell you)

```
> # Define a matrix
> # Directly
> m1 <- matrix(1, nrow=2, ncol=2)
> # From vectors
> m2 <- cbind(c(1,0), c(0,1))
> m1
      [,1] [,2]
[1,]    1    1
[2,]    1    1
> m2
      [,1] [,2]
[1,]    1    0
[2,]    0    1
```

```
> # Multiply matrices
> m1%m2
      [,1] [,2]
[1,]    1    1
[2,]    1    1
```

- Some useful basic functions
  - `mean(x)` takes average of vector `x`
  - `sd(x)` takes standard deviation of vector `x`
  - `lm()` for running OLS (see Help in R for more details)
  - Many packages in R will give you standard errors and other important stats objects when you estimate (ex: `lm()` for OLS)
- Combining these functions with simulation tools allows you to test your intuition and experiment

# HW

1. Prove the CEF property: we can always write  $Y = E[Y|X] + \epsilon$  where  $E[\epsilon|X] = 0$  and show that  $\epsilon$  is uncorrelated with any function of  $X$ .
2. Calculate  $V(\bar{x})$  where  $\bar{x}$  is the mean of  $N$  i.i.d. draws  $\{x_i\}_{i=1}^N$  where  $V(x_i) = \sigma^2$ . Calculate it when the draws are independent but not identical and  $V(x_i) = \sigma_i^2$  (make sure you know what i.i.d. means!). Bonus: under what conditions will this go to 0 as  $n \rightarrow \infty$ ?
3. Look up idempotent matrices. Prove that  $I - X'(X'X)^{-1}X'$  is idempotent.
4. Plot the true and empirical CEF from simulated data with  $N = 100$  for  $(Y, X)$  where  $Y = X + 3 * X^2 + \epsilon$ ,  $X \sim U\{0, \dots, 10\}$  (uniform distribution over integers 0 to 10) and  $\epsilon \sim N(0, 1)$ . Also try it for  $N = 1000$ . Note: I do NOT want you to run a linear model, but to approximate the CEF using conditional means.