Econometrics Discussion Section 2

John Green

Spring 2024

Linear regression

We now want to move on to examining how to understand the relationship between two variables.

- Linear regression model
- Necessary assumptions
- R^2, SER, F-test
- Distinguish *prediction* from *causation*

Minimizing error with one variable

- Suppose we have data $X_1, X_2, X_3, \ldots, X_n = \{X_i\}_{i \leq n}$
- What is the best guess for the value of an arbitrary X_j?
  - The best fit will be the mean, $E[X]$
- But wait: which error are we minimizing?

Minimizing error with one variable

- Suppose we have data $X_1, X_2, X_3, \ldots, X_n = \{X_i\}_{i \leq n}$
- What is the best guess for the value of an arbitrary X_j?
    - The best guess is going to be the mean, $E[X]$
    - This is intuitively embedded in the language we use: the "expectation" of X
- But wait: which error are we minimizing?

Minimizing error with one variable

- Suppose we have data $X_1, X_2, X_3, \ldots, X_n = \{X_i\}_{i \leq n}$
- What is the best guess for the value of an arbitrary X_j?
  - The best guess is going to be the mean, $E[X]$
  - This is intuitively embedded in the language we use: the “expectation” of X
- But wait: which error are we minimizing?
  - *Mean squared error*

Minimizing mean squared error

Our problem:

$$\underbrace{\arg\min}_{f} \mathbb{E}\left[\frac{1}{2}(X - f)^2\right]$$

How do we find the f that minimizes this?

Minimizing mean squared error

Our problem:

$$\underbrace{\arg\min}_{f} \mathbb{E}\left[\frac{1}{2}(X - f)^2\right]$$

How do we find the f that minimizes this?

Set the derivative to 0:

$$\mathbb{E}[(X - f)] = 0 \iff f = \mathbb{E}[X]$$

Minimizing error with two variables

- Now what if we have data on two variables:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) = \{(X_i, Y_i)\}_{i \leq n}$$

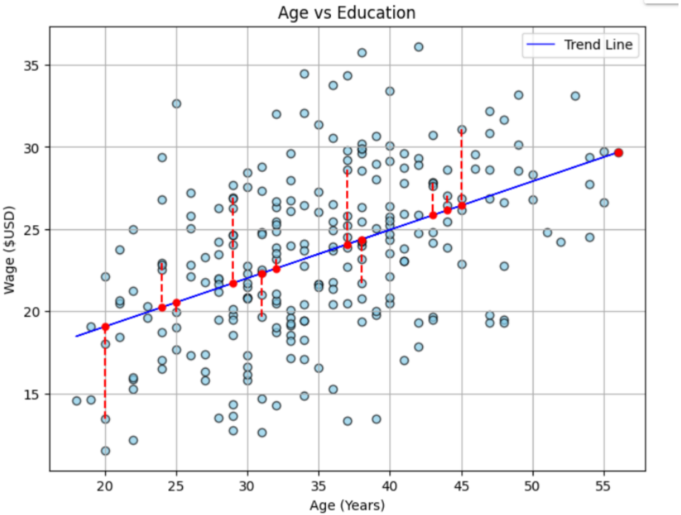
- We want to understand their relationship!
- What is our best guess for an arbitrary Y_j, given X_j?
  - X is age and Y is income; if I tell you that someone is 20 years old, what is your best guess for their income?
  - This will be the *conditional expectation*: $E[Y|X = 20]$

Minimizing error with two variables

- What is our best guess for an arbitrary Y_j, given X_j?
    - X is age and Y is income; if I tell you that someone is 20 years old, what is your best guess for their income?
    - This will be the *conditional expectation*: $E[Y|X = 20]$
- Let's assume a *linear* relationship:

$$Y_i = \beta_0 + \beta_1 X_i$$

- Then our problem is to draw a line through the 2D data which minimizes the errors
    - β_0 is the intercept, β_1 is the slope


Age vs Education
Trend Line
35
30
25
20
15
Wage ($USD)
20
25
30
35
40
45
50
55
Age (Years)

Ordinary least squares

- The line which minimizes the errors is the *ordinary least squares* regression line
- Found by solving

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 - \beta_1 X_i))^2$$

  just like we did before (derived in lecture notes and book)
- In practice, we only have a **sample**, and so we have to estimate these two parameters:
  - Add an error term: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
  - We will denote these estimates as $\hat{\beta}_0$ (intercept) and $\hat{\beta}_1$ (slope)

Assumptions

- What assumption have we made so far?

Assumptions

- What assumption have we made so far?
    - Specified a **linear** relationship between X and Y
    - This is not a small assumption!
- What if we want to think about X as having a causal effect on Y?
    - eg if we were to hold all other factors constant (as in a controlled and randomized experiment) what do we expect an extra year of education to do to income?

Assumptions

For β_1 to have a *causal* interpretation, several assumptions are necessary:

1. $E[\epsilon|X = x] = 0$
   - So that $\hat{\beta}_1$ is unbiased estimator
2. $\{(X_i, Y_i)\}_{i \leq n}$ are i.i.d.
   - Will give us sampling distributions for coefficient estimates
3. Large outliers are rare
   - OLS is sensitive to outliers, and our estimate $\hat{\beta}$ might be meaningless otherwise

Exogeneity

- $E[\epsilon|X = x] = 0$ implies that X is *exogeneous*
  - Uncorrelated with the error term
- When might this fail?

Omitted variable bias

- $E[\epsilon|X = x] = 0$ implies that X is *exogeneous*
    - Uncorrelated with the error term
- When might this fail?
- If there is an omitted variable that is correlated with X and Y, then X is not exogeneous
- Suppose we regress years of education on wages:

$$W_i = \beta_0 + \beta_1 E_i + \epsilon_i$$

- What might an omitted variable be?

Omitted variable bias

- Suppose we regress years of education on wages:

$$W_i = \beta_0 + \beta_1 E_i + \epsilon_i$$

- What might an omitted variable be?
- Suppose parental income is correlated with both years of education and wages, so the true model is:

$$W_i = \beta_0 + \beta_1 E_i + Z_i + u_i$$

- So when we estimate the first model, $\epsilon_i = Z_i + u_i$ and exogeneity will fail

Intuition behind exogeneity

- ϵ_i contains all of the things that influence Y_i that are not explicitly included in our model
- So, $E[\epsilon|X = x] = 0$ means that none of those factors are correlated with X
- ϵ_i will **always** contain extra terms we have not included, since there are always unobserved variables
- The problem only emerges when such variables are correlated with the included X

Regression output

- R^2 measures the variance in Y that is explained by X
  - $R^2 = \frac{ESS}{TSS}$
- Standard error of regression (SER): spread of the residual ϵ
- Root mean squared error (RMSE): similar, just calculated with n instead of $n - 2$ in the denominator

Sampling distribution of $\hat{\beta}$

- Just like $\bar{Y}$, $\hat{\beta}$ is a random variable and has a sampling distribution! (Why?)
- So, if we want to be able to say something about the relationship between an X and a Y using $\hat{\beta}$, we need to know something about its distribution
  - Will allow us to test hypotheses, eg that $\beta_1 = 0$
  - Will let us construct confidence intervals and indicate uncertainty
- Extra assumption on top of the OLS assumptions: that the relationship between X and Y is linear (how might we relax this?)

Hypothesis testing for $\hat{\beta}$

- Will work in a very similar way to the hypothesis testing we've done for the sample mean
  - Variance of $\hat{\beta}$ is decreasing in the sample size and in the variance of X
- We will construct a t-statistic, $t = \frac{\hat{\beta}-\beta}{SE(\hat{\beta})}$
- So we can test if $\beta = 0$ (ie if X has no relationship with Y), or if $\beta < 0$ (ie X has a negative relationship with Y)
- Can CIs as well

Binary regression

- Sometimes we have a binary regressor: for example, participation in some program
  - Effect of taking a drug
- OLS works in the same way but interpretation is a little different:
  - $Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$
  - β_0 is mean with no treatment
  - $\beta_0 + \beta_1$ is the mean with treatment
  - So β_1 is the average treatment effect

Heteroskedasticity

- Homoskedasticity: ϵ has constant variance (doesn't depend on X)
- If we assume Homoskedasticity, we can say some stronger things about the OLS
  - Gauss-Markov theorem: OLS is the best linear unbiased estimator ie has smallest variance
  - Simpler formula for the variance of $\hat{\beta}$
- What if we assume Homoskedasticity but it's not true?
  - SE will be too small (probably) meaning that we are *overconfident* in our inference
- If we assume normal errors, we can say some even stronger things
- Make strong assumptions, get strong results! These are usually difficult to justify.

Consistency and bias

- There are many features we would like our estimators to have: being *consistent* and *unbiased* are (usually) two of them
- Let's say we have an estimator $\hat{\theta}(x_i)$ for some statistic θ
  - $\bar{x}$ is estimator for μ_x, $\hat{\beta}^{OLS}$ is estimator for β
  - In other words, we look at some data x_i, use it to come up with a "best guess" for the θ
- Consistent: as our sample size grows, the estimator converges to the true value of the parameter
  - $\hat{\theta}(x_i) \xrightarrow{p} \theta$
- Unbiased: in expectation, our estimator is equal to the true value of the parameter
  - $\mathbb{E}[\hat{\theta}(x_i) - \theta] = 0$

Showing consistency and unbiasedness

- Bias is usually straightforward to show: take the expectation of the estimator, try to show it is equal to the object of interest
- Consistency is (sometimes) trickier
- 3 main tools to take advantage of:
  - Law of large numbers: as $n \to \infty$, $\bar{x} \xrightarrow{p} \mu_x$
  - Central limit theorem: as $n \to \infty$, $\sqrt{n}(\bar{x} - \mu_x) \xrightarrow{d} N(0, \sigma^2)$ eg $\bar{x} \xrightarrow{d} N(\mu_x, \frac{\sigma^2}{n})$
  - Continuous mapping theorem: if $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} Y$, then for continuous functions g, $g(X_n, Y_n) \xrightarrow{p} g(X, Y)$
- Nothing complicated: just take things piece by piece and ask yourself what happens when n gets big

Sample mean

- Random variable $X_i \sim F$, let $\mathbb{E}[X_i] = \mu_x$ and $\text{Var}(X_i) = \sigma^2$
- Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Is it biased?

Sample mean

- Random variable $X_i \sim F$, let $\mathbb{E}[X_i] = \mu_x$ and $\text{Var}(X_i) = \sigma^2$
- Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Is it biased?

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] = \frac{1}{n} n\mu_x = \mu_x$$

  Unbiased!

- Is it consistent?

Sample mean

- Random variable $X_i \sim F$, let $\mathbb{E}[X_i] = \mu_x$ and $\text{Var}(X_i) = \sigma^2$
- Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Is it biased?

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] = \frac{1}{n}n\mu_x = \mu_x$$

  Unbiased!

- Is it consistent?
  - Of course – this is exactly what the LLN tells us

$$\bar{x} \xrightarrow{p} \mu_x$$

Consistency and bias

- Does bias → consistency? **No!**

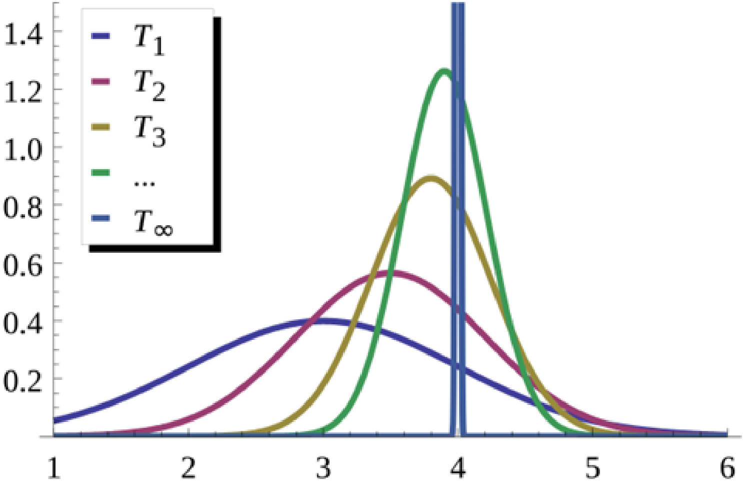
Consistency and bias

- Does bias → consistency? **No!**
- Does consistency → bias?

Consistency and bias

- Does bias → consistency? **No!**
- Does consistency → bias? **Not true either.**

Consistent but biased



Consistent but biased

- Remember that our estimator for the sample mean is:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- This is because the naive estimator is biased:

$$\hat{\sigma}^2_{naive} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\rightarrow$$

$$\mathbb{E}[\hat{\sigma}^2_{naive}] = \frac{n-1}{n}\sigma^2$$

- So there is a bias term of $\frac{n-1}{n}$
- But is this estimator consistent?

Consistent but biased

- The naive estimator is biased:

$$\hat{\sigma}^2_{naive} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\rightarrow$$

$$\mathbb{E}[\hat{\sigma}^2_{naive}] = \frac{n-1}{n}\sigma^2$$

- So there is a bias term of $\frac{n-1}{n}$
- But is this estimator consistent?
  - Yes, since $\lim_{n\rightarrow\infty}\frac{n-1}{n} = 0$

Unbiased but not consistent

- What if we use an extremely naive estimator for μ_x: X_1
  - Take a sample, take the first value, that's our estimator
- Is this biased?

Unbiased but not consistent

- What if we use an extremely naive estimator for μ_x: X_1
  - Take a sample, take the first value, that's our estimator
- Is this biased?
  - No: $\mathbb{E}[X_1] = \mu_x$
- Is it consistent?

Unbiased but not consistent

- What if we use an extremely naive estimator for μ_x: X_1
    - Take a sample, take the first value, that's our estimator
- Is this biased?
    - **No**: $\mathbb{E}[X_1] = \mu_x$
- Is it consistent?
    - **No**: $X_1 \sim F$ and does not depend on n; so as $n \to \infty$, X_1 does not converge to μ_x

Bias and consistency

- To sum up: in general, we like unbiased and consistent estimators
- Take expectation to check for bias; use CLT $+$ CMT for consistency
- Bias $!=$ consistency, and consistency $!=$ bias
- Our favorite estimators, eg $\frac{1}{n}\sum_i x_i$ and $\hat{\beta}^{OLS}$, are both consistent and unbiased

Prediction

- Oftentimes, we are just concerned with predicting a Y-variable:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Will always have some error in our prediction; but important to think about where it comes from
- In general (and in particular with the OLS model) we have at least two sources of error:
  - The intrinsic error in predicting an uncertain outcome ("oracle" error; V_1 in slides); $\mathbb{E}[u_i^2]$
  - Additional error introduced because we are estimating our model with a random sample ("approximation" error; V_2 in slides); $\mathbb{E}\left[\left(\left(\hat{\beta}_0 - \beta_0\right) + x_0\left(\hat{\beta}_1 - \beta_1\right)\right)^2 \mid X = x_0\right]$

Prediction

- We want to account for both of these when we make a prediction (see example)
- If we only account for sampling error, we will be overconfident in our prediction
- So, what to do if we want to make good predictions?
- OLS estimator will not be best for out of sample predictions
- Later on, will explore the “bias-variance” tradeoff: we may intentionally introduce some bias into our model in order to reduce the variance due to the estimated coefficients, thus getting the error in our model closer to the oracle prediction