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:

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

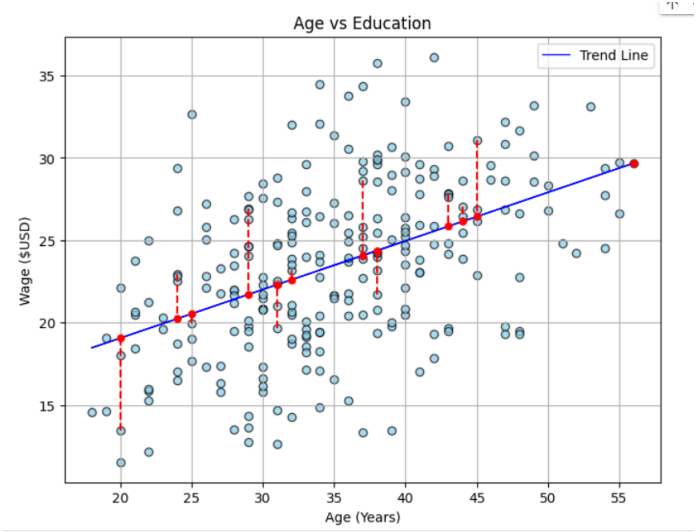
- We want to understand their relationship!
- What is our best guess for an arbitrary Y_j, given X_j?
  - X is age and Y is income; if I tell you that someone is 20 years old, what is your best guess for their income?
  - This will be the *conditional expectation*: $E[Y|X = 20]$

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

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.