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.

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

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 two variables

- Which error are we minimizing?
  - Usually the *squared* error — but not necessarily!
- Now what if we have data on two variables:

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

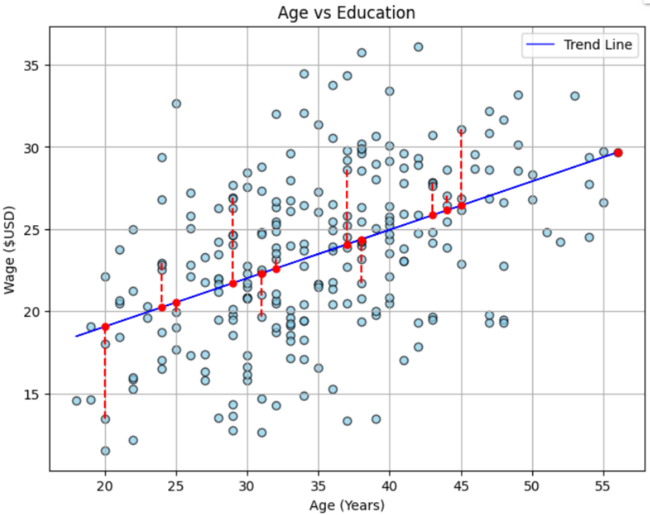
- We want to understand their relationship! Suppose we think that X causes Y: then we might want to know what is our best guess for an arbitrary Y_j, given its X_j?
  - X is years of education and Y is income; if I tell you that someone has 12 years of education, what is your best guess for their income?
  - This will be the *conditional expectation*: $E[Y|X = 12]$

Minimizing error with two variables

- We want to understand their relationship! Suppose we think that X causes Y: then we might want to know what is our best guess for an arbitrary Y_j, given its X_j?
  - X is years of education and Y is income; if I tell you that someone has 12 years of education, what is your best guess for their income?
  - This will be the *conditional expectation*: $E[Y|X = 12]$
- Let's assume a *linear* relationship:

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

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

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

Assumptions

- All we've done so far is draw a line — no assumptions needed!
- What if we want to think about X as having a causal effect on Y, as in the case of more years of school mechanically leading to a higher income?
  - 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

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

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.