Introductory Econometrics
Lecture 3: OLS Assumptions

by Hieu Nguyen

Fall 2024

1.

Suppose you have accepted a summer job as a weight guesser at the local amusement park, Magic Hill. Customers pay 50 cents each, which you get to keep if you guess their weight within 5 kilograms. If you miss by more than 5 kilograms, then you have to give the customer a small prize that you buy from Magic Hill for 60 cents each. Luckily, the friendly managers of Magic Hill have arranged a number of marks on the wall behind the customer so that you can accurately measure the customer's height. Unfortunately, there is a 150 cm wall between you and the customer so that you can tell little about the person except for height and (usually) gender.

On your first day on the job, you do so poorly that you work all day and somehow lose two dollars, so on the second day, you decide to collect data to run a regression to estimate the relationship between weight and height (above 150 cm). Since most of the participants are male, you decide to limit your sample to males. You hypothesize the following theoretical relationship:

$$\text{weight}_i = \beta_0 + \beta_1 \text{height}_i + \varepsilon_i.$$

The next day you collect the data summarized in the following table:

Customer	Height (cm)	Weight (kg)
1	170	65
2	180	75
3	175	80
4	160	60
5	185	85
6	155	55
7	165	70
8	170	72
9	175	78
10	180	83
11	185	90
12	190	95
13	195	100
14	160	55
15	155	50

Then you run your regression on the Magic Hill computer, obtaining the following estimates:

$$\hat{\beta}_0 = 46.49, \quad \hat{\beta}_1 = 1.14.$$

(a)

Interpret the estimated coefficients.

(b)

When you observe the table, how well do you think the regression works?

(c)

Identify the three customers who seem to be quite a distance from the estimated regression line. Would you have a better regression equation if we dropped these customers from the sample?

(d)

Look over the sample with the thought that it might not be randomly drawn. Does the sample look abnormal in any way? Would this affect the regression results and estimated weights if the sample is not random?

(e)

Think of at least one other factor besides height that might be a good choice as a variable in the weight/height equation. What would the expected sign of this variable's coefficient be if the variable was added to the equation?

(f)

Does this simple regression capture a causal relationship between height and weight? Explain.

2.

The data file `collegetown` contains observations on 500 single-family houses sold in Baton Rouge, Louisiana during 2009–2013. The data include sale price (in thousands of dollars) `PRICE` and total interior area of the house in hundreds of square feet `SQFT`.

a.

Plot house price against house size in a scatter diagram.

b.

Estimate the linear regression model

$$\text{PRICE} = \beta_1 + \beta_2 \text{SQFT} + e.$$

Interpret the estimates. Draw a sketch of the fitted line.

c.

Estimate the quadratic regression model

$$\text{PRICE} = \alpha_1 + \alpha_2 \text{SQFT}^2 + e.$$

Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

d.

For the regressions in (2) and (3) compute the least squares residuals and plot them against `SQFT`. Do any of our assumptions appear violated?

e.

One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (SSR) from the models in (2) and (3). Which model has a lower SSR? How does having a lower SSR indicate a "better-fitting" model?