Introductory Econometrics
Lecture 3: OLS Assumptions
Suggested Solution

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.

Solution: It seems we really observe a positive relationship between height and weight. Each additional 1 cm of height explains an increase in weight by 1.14 kg. Here the estimated intercept has real meaning; it represents an estimated weight when somebody is precisely 150 cm tall (perhaps a kid).

(b)

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

Solution: The regression seems to work well as you now have gained USD 6.70, and you lost only three times.

(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?

Solution: Customers 3, 4, and 20 according to highest residuals and the occurrence of losses. If we drop these three observations from the sample, we naturally obtain a mathematically better fitting estimated regression line as potential 'outliers' rotate the OLS predicted line. See also how the sum and sample average of the squared residuals change. But a great caution here, we lose 15% of information and violate the random sampling! Outliers are exceptions from the 'typical' real relationship. Still, it is often unclear (such as in this example), and no rigorous rule exists, how to decide where a border between 'typical' observations and outliers should be. So the answer is ambiguous. The 'issue' of outliers was discussed in detail during the seminar and can be studied in the attached Excel file.

(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?

Solution: We can get some general knowledge about average values for the US males via Wikipedia here and here: $\text{height}_{\text{US}} = 175.3\text{cm}$, $\text{weight}_{\text{US}} = 90.6\text{kg}$. However, in the attached Excel file, list 1, we observe that our sample averages are $\text{height}_n = 176.15\text{cm}$ but (!) $\text{weight}_n = 76.25\text{kg}$, even if 2 from 3 'outliers' are overweight cases. Our prediction for an average US male is then: $\widehat{\text{weight}} = 46.49 + 1.14 \cdot 26.15 = 76.3\text{kg}$. Members of the sample are thus abnormally slim! It seems that our sample is likely not random (why can we expect a self-selection bias?) and thus does not represent the average US male population. This, however, does not deteriorate the usefulness of our regression results as it seems that mostly slim men visit this attraction, so we need a regression equation for them, not for average US men.

(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?

Solution: We can think of, e.g., age (older people tend to gain weight; thus, a positive relationship is expected) or people doing office jobs (a positive relationship expected again as they do not move much on average).

(f)

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

Solution: First, it is crucial to understand/remember that regression gives evidence, but does not prove the 'real world' economic causality. The regression results only tell us whether a significant quantitative relationship exists and the strength and direction (+/ sign) of this relationship. However, the 'if-then direction' of the causal effect needs to be set in advance by the researcher via the structure of an econometric model (dependent vs. independent variables), which should be based on economic theory, common sense, etc. The real-world causal relationships (cause → consequence) are often unclear even to philosophers and thus can hardly be proven using regression. The important 'issue' of causality in econometrics was discussed in detail during the seminar. A highly recommended reading about causality in econometrics can be found in Studenmund (2016 & 17), 1.2 (1 page of text) and Wooldridge (2016), 1-4 (4 pages of text only).

To answer this question finally, the regression results would capture the suggested (by the structure of the model) causal relationship if there were no other unobserved factors related to weight that can also influence height (think of DNA inheritance, for instance). If such factors exist, and we do not control for them, their impact is captured by the error term, and the CA 3. is likely to be violated. Our estimator is expected to be biased and inconsistent.

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.

Solution: gnuplot price sqft –output=display

Figure 1: Caption

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.

Solution: If the size of the house increases by one unit, price increases by 13.4 thousand dollars.

	coefficient	std. error	t-ratio	p-value	
const	-115.424	13.0882	-8.819	1.95e-017	***
sqft	13.4029	0.449164	29.84	5.92e-113	***

Mean dependent var	250.2369	S.D. dependent var	171.4765
Sum squared resid	5262847	S.E. of regression	102.8006
R-squared	0.641317	Adjusted R-squared	0.640596
F(1, 498)	890.4114	P-value(F)	5.9e-113
Log-likelihood	-3024.863	Akaike criterion	6053.726
Schwarz criterion	6062.155	Hannan-Quinn	6057.033

Figure 2: Caption

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.

Solution: Take a derivative with respect to `sqft` in $\text{PRICE} = \alpha_1 + \alpha_2 \text{SQFT}^2 + e$:

$$\frac{\partial \text{PRICE}}{\partial \text{SQFT}} = 2 \cdot \alpha_2 \cdot \text{sqft}$$

If sqft = 2000 then

$$\frac{\partial \text{PRICE}}{\partial \text{SQFT}} = 2 \cdot \alpha_2 \cdot 2000 = 4000 \cdot 0.18 = 720$$

If you increase the size of the house with 2000 square feet by 100 square feet, price will increase by 720 thousand dollars (initial condition matters).

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?

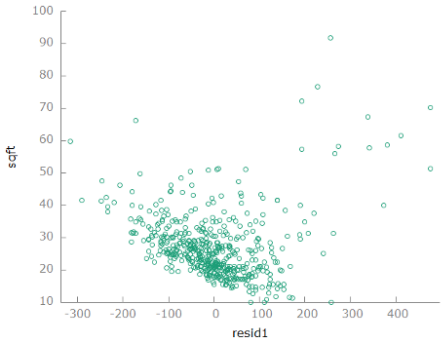

Figure 3: Caption

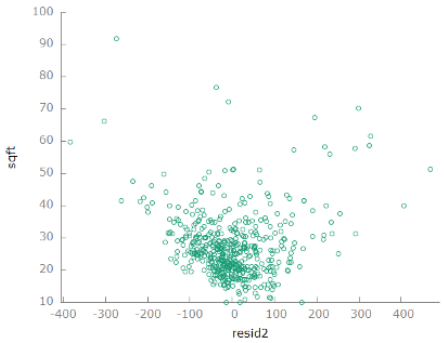

Figure 4: Caption

Solution:

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?

Solution: The second model has lower SSR. Lower SSR means that there is less variation unexplained in the model. SSR is tightly related with the goodness of fit measure in fact $R^2 = 1 - SSR/SST$, therefore, larger SSR will deliver worse goodness of fit.