Lecture 5 Notes: Regression Analysis

Econ 4523: Economics of Education
Prof. Tyler Ransom
University of Oklahoma

Fall 2019

Last time

- Controlled vs. randomized experiments
- Why can't we do controlled experiments in econ or other social sciences?
- How to test if randomization worked?
- Internal vs. external validity
- Why can't we do experiments sometimes?

Activity

- Read https://www.theguardian.com/society/2017/jul/16/stressful-experiences-can-age-
- Is this experimental or observational data?
- What is the "treatment"? What is the outcome?
- Do you think the results are causal? What do you think about measurement issues?

Today's Big Questions

- What is regression analysis?
- How does a regression tell us about correlation between the treatment and outcome variables?
- Under what conditions does regression tell us about causal impacts of treatment on the outcome variable?

1 What is regression analysis?

Regression analysis is a statistical technique to *estimate* numbers from a statistical model (called *parameters*) that best fit the data.[1]

The statistical model can be as simple or complex as the researcher would like it to be. For now, let's follow the example in Lovenheim and Turner:

$$Y_i = \alpha + \beta Ed_i + \varepsilon_i$$

The basic components of a regression are:

- **Independent variable(s)** [a.k.a. explanatory variables, a.k.a. right-hand side variables, a.k.a. covariates]. In this case, Ed_i is the independent variable and measures the years of education for observation i in our data set.

- **Dependent variable** [a.k.a. left-hand side variable, a.k.a. outcome variable]. In this case, Y_i is the dependent variable and measures the earnings for observation i in our data set.

- **Error term** [a.k.a. disturbance term]. ε_i (lower case Greek letter "epsilon") simply represents all other factors (besides years of education) that determine earnings. Depending on what our data set looks like, ε can be other characteristics about observation i that we chose not to include in the regression model, or it can be characteristics about i that are not included in the data (e.g. how abrasive i's personality is, how often i shows up late for appointments, etc.).

- **Parameters** [a.k.a. coefficients]. (α, β) are *statistical parameters* that we are interested in knowing the values of. These are the "estimates of interest" and constitute the main takeaway from the data. Understanding how to interpret the coefficients is how researchers influence policy!

The interpretation of β in the statistical model above is as follows: "one extra unit of education is associated with an *average* increase in earnings of β." Note that the interpretation always depends on the units of both the explanatory and outcome variables!

[1] "Best fit" here is defined to be the sum of the squared residuals from the line. The way to think about this process is as follows:

1. Plot your data on an X-Y plane
2. Draw a line that looks like it has equal number of observations above and below the line
3. Compute the vertical distance from each point to that line, and square it
4. Add up all of the squared vertical distances
5. Try another line, and repeat until you've found the line that has the smallest sum of squared distances

The process described above is called "Ordinary Least Squares" or OLS and is by far the most popular technique for doing regression analysis.

2 Regression as correlation

It turns out that regression is the most natural way to think about correlation in a multivariate setting.

1. In the model listed above, the formula for the estimate of β (which we denote $\hat{\beta}$ and read as "beta hat") is:
$$\hat{\beta} = \frac{\text{cov}\left(Y_i, Ed_i\right)}{\text{var}\left(Ed_i\right)}$$
This formula closely resembles the formula for correlation. Specifically, it can be shown that
$$\hat{\beta} = \rho\sqrt{\frac{\text{var}\left(Y_i\right)}{\text{var}\left(Ed_i\right)}}$$
(In an upcoming problem set you will prove the relationship between the two.)

2. A statistic called R^2 (pronounced "R squared") is actually the measure of correlation. Specifically,
$$\rho^2 = R^2$$
We won't get into the details of where the R^2 formula comes from. Just know that it is an often-used metric of how well the regression model fits the data. Even if there are 100 explanatory variables in the model, the R^2 gives the multivariate analog to the basic correlation formula.

2.1 Why correlation in a regression might not be causation

We know from previous lectures that correlation does not imply causation. How does this work in the regression setting?

The short story is that there is something in the error term of the regression that is correlated with both the right-hand side and left-hand side variables. If we think about the original example of earnings and education, we might think that there are things in ε_i that explain education *and* that explain earnings.

For example, suppose that earnings are heavily determined by parental income. But, because education costs so much money, children of rich parents are also more likely to get more education. So if in fact it is parental income that determines earnings, but we only include education in our regression model, then we would conclude that education causes higher earnings, when in fact it is parental income that does. (Remember that this is a simple illustrative example!)

We call this phenomenon **omitted variable bias**. It is basically the regression analog to selection bias: there is some reason why the treated group is systematically different from the control group. In the previous example, the omitted variable bias is caused by the selection of rich children into higher levels of education.

The model that would give us the correct parameters in this scenario would be:

$$Y_i = \alpha + \beta Ed_i + \gamma ParentalIncome_i + \varepsilon_i$$

And based on the example, we should expect the estimated coefficient $\hat{\gamma}$ to be large, and the coefficient $\hat{\beta}$ to be zero.

[Go through example on the board with "rich" and "poor" kids educational level correlated with their parental income and future earnings

2.2 Estimating the average treatment effect from experimental data using a regression

Regressions can be used to estimate the ATE from an experiment. To see how this can be done, consider a variation from the model above:

$$Y_i = \alpha + \beta treatment_i + \varepsilon_i$$

where now our lone explanatory variable, *treatment$_i$*, indicates that observation *i* was in the treatment group of our randomized experiment. The average treatment effect in this case is simply

$$\begin{aligned}
ATE = &\underbrace{E\left[Y_i | treatment = 1\right]}_{\hat{\alpha}+\hat{\beta}} - \underbrace{E\left[Y_i | treatment = 0\right]}_{\hat{\alpha}} \\
= &\hat{\alpha} + \hat{\beta} - \hat{\alpha} \\
= &\hat{\beta}
\end{aligned}$$

Q: In a perfectly randomized experiment, should we include additional right-hand side variables in the regression? Why or why not?

3 When can regression parameters from non-experimental data give causal effects?

Omitted variable bias (or selection bias) can be circumvented in three ways:

1. Include in the regression model every possible variable that is correlated with treatment and explains the outcome
2. Assume that including enough explanatory variables makes treatment "effectively" random, in which case we would get the treatment effect as in the experimental case above.
3. Find some additional source of variation in the treatment that is unrelated to treatment status.

The approach in (1) above is not likely to be possible in most cases. Why?

The approach in (2) is called **selection on observables** or **the unconfoundedness assumption**. It amounts to assuming that there are enough explanatory variables in the regression model to account for all differences that could explain treatment and outcome.

The approach in (3) requires a different type of data called **quasi-experimental** or **quasi-random** data, or data from a **natural experiment**. We will talk about this approach in the next two lectures.

Video links

- https://www.youtube.com/watch?v=R0LeLaR-17U
- https://www.youtube.com/watch?v=zZtL7cWN-3c
- https://www.youtube.com/watch?v=iHZLHdTPc6E
- https://www.youtube.com/watch?v=FQ1CB0Svh8I

References

Lovenheim, Michael and Sarah E. Turner. 2017. *Economics of Education*. New York: Worth Publishers.