14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 10-12

Non-linear regression

One underlying assumption of linear regression is that the effect of change in X on Y is constant and does not depend on the level of X or any other regressor. This assumption is that the conditional expectation of Y given regressors is linear in X. This lecture discusses what we can do if that is not true.

Polynomials

Any continuous function of a compact set can be approximated arbitrarily well by a polynomial. Thus, it often makes sense to try to estimate a polynomial relation between X and Y. For example,

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + ... + \beta_r X^r + \gamma W + e,$$

where W contains controls.

Observation 1. We can estimate $(\alpha, \beta_1, ..., \beta_r, \gamma)$ via OLS by regressing Y_i on $1, X_i, X_i^2, ..., X_i^r, W_i$. We just need to create technical regressors (powers of X). The usual properties of OLS are applicable to this case.

The main difficulty is in interpreting the coefficients.

Example. Consider the example of the returns to education. Assume we estimated OLS regression:

$$Wage = 2.5 - 0.37\ Education + 0.041\ Education^2 + 0.019\ Experience + 0.17\ Tenure.$$

Imagine a person who has 12 years of education. For him:

$$Wage_{12} = 2.5 - 0.37 \cdot 12 + 0.041 \cdot 12^2 + 0.019\ Experience + 0.17\ Tenure,$$

and compare to a similar person with 13 years of education:

$$Wage_{13} = 2.5 - 0.37 \cdot 13 + 0.041 \cdot 13^2 + 0.019\ Experience + 0.17\ Tenure.$$

Thus

$$\Delta Wage = -0.37(13 - 12) + 0.041(13^2 - 12^2) \approx 0.66.$$

If we do the same exercise, changing years of education from 8 to 9 years, we would obtain:

$$\Delta Wage = -0.37(9 - 8) + 0.041(9^2 - 8^2) \approx 0.33.$$

What we see is that the effect of one additional year of education depends on the current level of education.

From looking at the relation between wage and education, we see that it is an upward-looking parabola with the minimum achieved at $0.37/(2 \cdot 0.041) \approx 4.5$. Given that almost all observations are on the right from that point, we are on the increasing part of the parabola. Thus, the effect of each additional year of education on the wage in the relevant range is positive. We also have:

$$\frac{\partial Wage}{\partial Education} = -0.37 + 0.082 \cdot Education,$$

thus there is an increasing return to education. The average level of education in our population is 12.56, thus we have the marginal effect at the average is $-0.37 + 0.082 \cdot 12.56 \approx 0.66$.

Cite as: Anna Mikusheva, course materials for 14.32 Econometric Data Science, Fall 2019. Massachusetts Institute of Technology.

The important points of this discussion: we cannot interpret coefficients of polynomials one-by-one, they can only be interpreted as a group. The discussion should be not about individual coefficients, but about the functional relation they produce as a group.

If we wish to decide if we want to use a quadratic function rather than a linear, one needs to test whether the quadratic term is zero. The standard errors on the quadratic term in the current example is 0.009, which indicates significance. Thus, we have sufficient statistical evidence to claim that the quadratic function fits better than the linear. We may also try to see if there is a need for a cubic polynomial, and for this, we will run a cubic regression and look at significance of the cubic term.

Logarithms

Another common way to account for non-linearity is to use logarithms. The tricky part is to interpret the coefficients. Logarithmic transformation permits modelling in relative (percentage) terms. Notice that for small Δ we have

$$\ln(x+\Delta)-\ln(x)=\ln\left(\frac{x+\Delta}{x}\right)=\ln\left(1+\frac{\Delta}{x}\right)\approx\frac{\Delta}{x},$$

where the last approximation holds for very small $\frac{\Delta}{x}$. The approximation appears as a result of a Taylor expansion and $\frac{d}{dx}\ln(x)=\frac{1}{x}$. Numerically the approximation works well for changes of order up to 10%:

$$\ln(1.01)=0.00995...\approx 0.01;\quad \ln(1.1)=0.0953\approx 0.1.$$

Linear-log case

Assume one runs the regression

$$Y=\alpha+\beta\ln(X)+\gamma W+e.$$

Imagine an intervention in which X is changed by ΔX holding W constant, then we have

$$Y+\Delta Y=\alpha+\beta\ln(X+\Delta X)+\gamma W+e.$$

By subtracting two equations we get:

$$\Delta Y=\beta\left(\ln(X+\Delta X)-\ln(X)\right)\approx\beta\frac{\Delta X}{X},$$

thus

$$0.01\beta\approx\frac{\Delta Y}{100\frac{\Delta X}{X}}.$$

So, the interpretation is: a 1% increase in X is associated with an expected change in Y of 0.01β units, holding controls W constant.

Log-linear case

Consider the following regression

$$\ln(Y)=\alpha+\beta X+\gamma W+e.$$

Then the changed equation is:

$$\ln(Y+\Delta Y)=\alpha+\beta(X+\Delta X)+\gamma W+e,$$

and after differencing:

$$\frac{\Delta Y}{Y}\approx\ln(Y+\Delta Y)-\ln(Y)=\beta\Delta X.$$

Thus

$$100\beta \approx \frac{100\frac{\Delta Y}{Y}}{\Delta X}.$$

The interpretation is: a one unit change in X induces $100\beta\%$ change in Y controlling for W. For example, imagine

$$\ln(Wage) = 0.583 + 0.08 Education + e.$$

The interpretation of the coefficient is : an additional year of education increases wages on average by 8%.

Log-log

If the specification is

$$\ln(Y) = \alpha + \beta \ln(X) + \gamma W + e.$$

Then by differencing we have:

$$\frac{\Delta Y}{Y} \approx \ln(Y + \Delta Y) - \ln(Y) = \beta \left[\ln(X + \Delta) - \ln(X)\right] \approx \beta \frac{\Delta X}{X}.$$

Thus

$$\beta \approx \frac{100\frac{\Delta Y}{Y}}{100\frac{\Delta X}{X}}.$$

In this case β can be interpreted as elasticity: by how many percent on average Y would change due to a 1% change in X keeping W constant.

Important note. One should not compare R^2 between two regressions with different left-hand sides, say when one has Y, while another, $\ln(Y)$. One should choose between these two regressions based on sound economic reasoning: whether it is proper to discuss absolute or relative changes in Y.

Interactions

Interactions can be used when an effect of X on Y depends on the level of another variable.

Interactions: two binary variables

Example. Let us consider a regression which describes the value of a college diploma:

$$\ln(Wage) = \beta_0 + \beta_1 Female + \beta_2 College + e.$$

Here, $Female$ is the dummy for female, while $College$ is the dummy for whether one has a college degree. Then $100\beta_1$ is the average wage gap in percents between genders, while $100\beta_2$ is the average percentage increase in wages due to having a college diploma, controlling for gender. This regression assumes that the value of a college diploma is the same for males and females, though we may reasonably expect it to be different. To capture this phenomenon, we run:

$$\ln(Wage) = \beta_0 + \beta_1 Female + \beta_2 College + \beta_3 Female \times College + e.$$

In order to interpret any coefficient in this regression one should consider all combinations of dummies:

$$E(\ln(Wage)|Female = x, College = 0) = \beta_0 + \beta_1 x,$$

$$E(\ln(Wage)|Female = x, College = 1) = \beta_0 + \beta_1 x + \beta_2 + \beta_3 x.$$

Thus,

$$E(\ln(Wage)|Female = x, College = 1) - E(\ln(Wage)|Female = x, College = 0) = \beta_2 + \beta_3 x.$$

That is, coefficient β_2 describes effect of a college diploma on $\ln(Wage)$ for males, while the effect for females is described by $\beta_2 + \beta_3$.

Interpretation: $100\beta_2\%$ is the average increase in wages from acquiring a college degree for a male. While $100\beta_3\%$ is the average difference in value of a college degree for wages between genders. If β_3 is negative, then the wage gap between genders narrows for more educated individuals. As for the other coefficients:

$$\beta_1 = E(\ln(Wage)|Female = 1, College = 0) - E(\ln(Wage)|Female = 0, College = 0).$$

Thus, β_1 characterises the gender wage gap for people without a college degree.

$$\beta_0 = E(\ln(Wage)|Female = 0, College = 0).$$

While β_0 is connected to average $\ln(Wage)$ for males without a college degree.

Interactions: a continuous and a binary variables

Consider a regression:

$$\ln(Wage) = \beta_0 + \beta_1 Education + \beta_2 Female + e,$$

here, *Education* is the full years of education. In this regression β_1 captures the returns to education controlling for gender, while β_2 captures the gender gap while controlling for education. As in the subsection above we may expect that the returns to education may be different for men and women. To capture this, consider the following regression:

$$\ln(Wage) = \beta_0 + \beta_1 Education + \beta_2 Female + \beta_3 Education \times Female + e.$$

In order to interpret the coefficients, let us again consider different values for the dummy:

$$E(\ln(Wage)|Female = 0, Education = x) = \beta_0 + \beta_1 x,$$

$$E(\ln(Wage)|Female = 1, Education = x) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x.$$

See that we have two different curves describing the relation between wage and education- one for each gender.

Thus, we see that β_1 characterizes the return to education for males: every additional year of education increases the wage of a male on average by $100\beta_1\%$. While for females, the return to education is characterised by $\beta_1 + \beta_3$. Thus β_3 is the difference in value of each additional year of education between women and men. If $\beta_3 < 0$, then the wage gap narrows with education.

Notice that coefficients β_0 and β_2 do not have a meaningful interpretation, as $Education = 0$ falls outside of the relevant data range.

Discussion of slides

Slides are on the course web-site.

- Discuss interpretation of coefficients in the baseline regression: is the return to education of the expected sign? size? Is it statistically significant?
- Why there are 3 regional dummies when there are 4 regions? Interpret the coefficients.
- Regression (2) allows for two distinct linear relations (one for each gender) between education and wages. Do you think this is statistically justified? Is it economically meaningful? Did you expect the result?

- Regression (3) describes two relations as non-linear. Remember, you cannot interpret coefficients individually, only as a group(!). Do you think using quadratics is justified? Do you think using different curves for genders is justified? Which regression would you prefer?
- Look at the graph. Discuss the curves. Does the gender gap close with education?

Interaction: two continuous variables

Let us consider the following regression:

$$\ln(Wage) = \beta_0 + \beta_1 Education + \beta_2 Experience + \beta_3 Education \times Experience + e;$$

then the marginal effect of education is:

$$\frac{\partial \ln(Wage)}{\partial Education} = \beta_1 + \beta_3 Experience.$$

That is, the return to education depends on the level of work experience a person has. If $\beta_3 < 0$ then the return to education cancels out with experience.

Summary about non-linear regression

- Effect(slope of population function) depends on the level of a regressor.
- Be careful with interpretation: often coefficients can only be interpreted as groups.
- Taking a before-after difference can be helpful.
- When one has dummies, it is useful to try different combinations of turning them ‘on’ and ‘off’.
- Log-transformations bring interpretation of relative change.

How to select a non-linear specification:

(1) Think. What may the effect depend on? Is relative change the proper way of discussing this variable?

(2) If choosing between nested models (polynomials, interactions), do a formal statistical test.

(3) If choosing between non-nested models with the same left-hand-sides, one can use R^2_{adj}, but also think about the resulting easiness to interpret when making a choice.

(4) When comparing models with different left-hand-sides, think about resulting economic interpretation.