14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 4-6

Bivariate regression

Motivation for bivariate regression

This course is about the relationship between variables: if X is changed by that many units, how will Y change? For example, if we are studying the effect of class size(X) on student performance(Y), we are interested in the change in test scores due to changes in class size. The parameter of interest would be

$$\beta = \frac{\Delta \ Test\ Score}{\Delta \ Class\ size} = \frac{\Delta Y}{\Delta X}.$$

The simplest model to describe this parameter is

$$Y_i = \alpha + \beta X_i + e_i.$$

Terminology: (α, β) are regression coefficients, α is an intercept, β is a slope, Y is the dependent variable, X- independent variable (regressor), e- error term.

What is included in the error term? Approximation errors, measurement errors, everything else (except X) that directly affects Y. The important assumption so far is the constancy of β (linearity between Y and X). We will discuss how to deal with non-linear relations later on in the course.

Ordinary Least Squares (OLS)

Example Assume one observe an i.i.d. sample $Z_1, ..., Z_n$ and the question arises about how to estimate $\mu = EZ_i$. As we have seen, BLUE is the sample average $\bar{Z}$. This estimate is the solution to the following optimization problem:

$$\bar{Z} = \arg\min_m \sum_{i=1}^n (Z_i - m)^2,$$

which is often called the 'least squares method'. Below we generalize this method to linear regression. □

We have data $(X_i, Y_i), i = 1, .., n$, assumptions about which we will introduce and discuss later. Now we want to estimate regression coefficients in the model

$$Y_i = \alpha + \beta X_i + e_i.$$

Let us define the Ordinary Least Squares (OLS) estimator as the solution to the following optimization problem:

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{a,b} \sum_{i=1}^n (Y_i - a - bX_i)^2 = \arg\min_{a,b} S(a, b).$$

Below, we will derive formulas for the estimates. For this, let us define the predicted (or fitted) values:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} X_i.$$

All points $(X_i, \widehat{Y}_i)$ lie on the fitted line $y = \widehat{\alpha} + \widehat{\beta} x$. Also, define residuals $\widehat{e}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\alpha} - \widehat{\beta} X_i$.

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

Let us write the first-order conditions for the optimization problem:

$$\frac{\partial S(a,b)}{\partial a}|_{(a,b)=(\widehat{\alpha},\widehat{\beta})} = -2\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i) = 0. \tag{1}$$

$$\frac{\partial S(a,b)}{\partial b}|_{(a,b)=(\widehat{\alpha},\widehat{\beta})} = -2\sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)X_i = 0. \tag{2}$$

One can check the second-order condition, which holds as long as $\sum_{i=1}^{n}(X_i - \bar{X})^2 > 0$, which means that not all X_i are the same, and is often referred to as a 'no multicollinearity' assumption. Equation (1) implies

$$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\beta}\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{Y} - \widehat{\beta}\bar{X}. \tag{3}$$

If we plug this into equation (2), we get:

$$\sum_{i=1}^{n}(Y_i - \bar{Y} - \widehat{\beta}(X_i - \bar{X}))X_i = 0.$$

We solve the last equation for $\widehat{\beta}$ (notice that we can solve the equation only when not all X_i are the same – no multicollinearity):

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})X_i}.$$

Notice that $\sum_{i=1}^{n}(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X}) = 0$, thus:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\widehat{cov(X,Y)}}{\widehat{Var(X)}}. \tag{4}$$

Equations (3) and (4) allow us to calculate estimates from a given sample. Also note, that the first-order conditions can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i = 0 \qquad \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i X_i = 0.$$

Discussion: why OLS? Can we do some other type of estimation? Yes, another popular method is median regression, which solves the following optimization:

$$(\widetilde{\alpha},\widetilde{\beta}) = \arg\min_{a,b}\sum_{i=1}^{n}|Y_i - a - bX_i|.$$

This estimate is less sensitive to outliers (it penalizes outliers less). It has different properties than OLS, and apparently estimates somewhat different quantities. The properties of estimators depend on assumptions (which we have not introduced or discussed yet). OLS is the standard method of linear regression mainly because it is analytically easy (simple formulas) and easy to discuss.

Assumption for OLS

Assume that data (X_i, Y_i) comes from a model

$$Y_i = \alpha + \beta X_i + e_i,$$

with no multicollinearity. We impose the following assumptions:

Assumption 1 $E(e_i|X_i) = 0$.

Assumption 2 (X_i, Y_i) is an i.i.d. sample.

Assumption 3 $EX_i^4 < \infty$, $Ee_i^4 < \infty$.

First, notice that the model with no assumptions is not restrictive at all, since for any α, β one can define $e = Y - \alpha - \beta X$. It is the assumptions that add content to the model.

Assumption 1 is the most important. First of all, it implies $E(Y|X) = \alpha + \beta X$ – that the conditional mean is a linear function.

Assumption 1 refers to exogeneity, which is related to causal interpretation – most of the course will be about this. Assumption 1 means that for any function f we have $E[ef(X)] = 0$.

We will maintain Assumption 2 for most of the course, though we will discuss some other sampling settings like panel data and time series in the later part of the course.

Assumption 3 is mostly technical and is needed for some Law of Large Numbers and central limit theorems. It is often called 'no (too many) outliers'.

Properties of the OLS estimator

Lemma 1. *If assumptions 1-3 are true then $(\widehat{\alpha}, \widehat{\beta})$ is an unbiased estimate for (α, β).*

Proof. For the proof we will use the following properties of conditional expectation: for any two random vectors η, ξ

$$E[E(\xi|\eta)] = E[\xi],$$
$$E(f(\eta)\xi|\eta) = f(\eta)E(\xi|\eta).$$

Now, the proof of the lemma:

$$Y_i - \bar{Y} = \alpha + \beta X_i + e_i - \alpha - \beta\bar{X} - \bar{e} = \beta(X_i - \bar{X}) + (e_i - \bar{e}),$$

$$\widehat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n \left[\beta(X_i - \bar{X}) + e_i - \bar{e}\right](X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta + \frac{\sum_{i=1}^n (e_i - \bar{e})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

$$E(\widehat{\beta} - \beta|X) = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} E\left(\sum_{i=1}^n (e_i - \bar{e})(X_i - \bar{X})|X\right) = \frac{\sum_{i=1}^n (X_i - \bar{X})E(e_i - \bar{e}|X)}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0.$$

Thus,

$$E\widehat{\beta} = E\left[E(\widehat{\beta}|X)\right] = \beta.$$

As for $\widehat{\alpha}$:

$$\widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} = \alpha + \beta\bar{X} + \bar{e} - \widehat{\beta}\bar{X} = \alpha + (\beta - \widehat{\beta})\bar{X} + \bar{e},$$
$$E(\widehat{\alpha}|X) = \alpha + \bar{X}E(\beta - \widehat{\beta}|X) + E(\bar{e}|X) = \alpha.$$

□

Lemma 2. *If assumptions 1-3 hold, then $(\widehat{\alpha}, \widehat{\beta})$ is a consistent estimate of (α, β).*

Proof. As shown before:

$$\widehat{\beta} - \beta = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})e_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

We have $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}s_X^2$, where s_X^2 is the sample variance which, as we proved before, is consistent. Thus the denominator converges to $Var(X)$. As for the numerator, by the Law of Large Numbers:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})e_i = \frac{1}{n}\sum_{i=1}^{n}X_i e_i - \bar{X}\bar{e} \to^p E[eX] - E[X]E[e] = 0,$$

the last equality is implied by Assumption 1. Indeed, $E[e] = E[E(e|X)] = 0$ and $E[eX] = E[E(eX|X)] = E[XE(e|X)] = 0$. Thus,

$$\widehat{\beta} - \beta \to^p \frac{0}{Var(X)} = 0.$$

Now, about the intercept:

$$\widehat{\alpha} = \alpha + (\beta - \widehat{\beta})\bar{X} + \bar{e}.$$

By the Law of Large Numbers $\bar{X} \to^p E[X]$ and $\bar{e} \to^p E[e] = 0$. Thus,

$$\widehat{\alpha} \to^p \alpha.$$

Lemma 3. *Under assumptions 1-3, the OLS is asymptotically Gaussian.*

Proof.

$$\sqrt{n}(\widehat{\beta} - \beta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \bar{X})e_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

The denominator converges (as we showed before) to $Var(X)$. As for the numerator, the Central Limit Theorem leads to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \bar{X})e_i \Rightarrow N(0, E[(X - EX)^2e^2]).$$

Thus,

$$\sqrt{n}(\widehat{\beta} - \beta) \Rightarrow N\left(0, \frac{E[(X - EX)^2e^2]}{(Var(X))^2}\right).$$

We will not do the proof of asymptotic gaussianity of $\widehat{\alpha}$.

Homoskedasticity.

We do talk about this because of the issue of historic stickiness. Historically, there was another assumption often invoked, called homoskedasticity:

$$Var(e|X) = \sigma^2.$$

This assumption says that the unexplained factors in Y have the same spread at all levels of X. This is not a very applicable assumption; indeed, consider a question of how years of education (X) affects income (Y). It is a known empirical fact that the spread of wages for highly educated people is much wider than the spread of wages for less-educated individuals. This contradicts homoskedasticity. The absence of homoskedasticity is called heteroskedasticity. Absence of an assumption is always more general than the presence of an assumption. The method that does not use any additional assumption will always work in more settings than the one employing an additional assumption.

Previously, an assumption of homoskedasticity was employ to calculate the asymptotic variance of the OLS estimate. Notice that under homoskedasticity $E[(X - EX)^2e^2] = Var(X)Var(e)$, and the formula for variance simplifies: $\lim Var(\sqrt{n}\widehat{\beta}) = \frac{Var(e)}{Var(X)}$. This formula is correct only under assumptions of homoskedasticity, but typically fails under heteroskedasticity.

Based on the two formulas for variance of $\widehat{\beta}$, there are two estimators for variance of $\widehat{\beta}$: one is homoskedasticity-only and one is heteroskedasticity-robust. The heteroskedasticity-robust variance estimate is:

$$\widehat{\sigma}^2_{\beta} = \left(\frac{1}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}\right)^2 \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \widehat{e}_i^2.$$

I do not provide the homoskedasticity-only variance estimate because I strongly prefer that you do not use it (with few exceptions).

WARNING!!!! STATA still uses the homoskedasticity-only variance as the default, mostly due to historic stickiness. In almost all application it is advisable to use the heteroskedasticity-robust option which, in STATA, is done by adding (', robust').

Testing and confidence set construction

Assume one considers an example of the effect of the class performance (X) on the test scores of students (Y). Thus we run a regression:

$$Test\ score_i = \alpha + \beta class\ size_i + e_i.$$

The hypothesis of interest, that class size has any effect whatsoever, is:

$$H_0 : \beta = 0 \quad vs \quad H_1 : \beta \neq 0.$$

By Lemma 3 we have

$$\frac{\sqrt{n}(\widehat{\beta} - \beta)}{\sigma_{\beta}} \Rightarrow N(0, 1).$$

The formula above allows us to calculate a consistent estimate of the variance σ^2_{β}. Let us define standard errors $s.e.(\widehat{\beta}) = \widehat{\sigma}_{\beta}/\sqrt{n}$. In order to test our hypothesis of interest, let us form a t-statistic: $t = \frac{\widehat{\beta}-0}{s.e.(\widehat{\beta})}$. We accept the null at the 95% level if $|t| < 1.96$ and reject it if $|t| > 1.96$. In the latter case the coefficient β is referred to as *significant*. The p-value for the given statistics is calculated as $p - value = 2\Phi(-|t|)$.

The confidence set for β is $[\widehat{\beta} - 1.96 s.e.(\widehat{\beta}), \widehat{\beta} + 1.96 s.e.(\widehat{\beta})]$.

R^2 and measure of fit

The notion of 'measure of fit' answers the question of how well the data fits a linear relation. There are several quantitites we will work with:
Sum of Squared Residuals

$$SSR = \sum_{i=1}^n \widehat{e}_i^2,$$

Explained Sum of Squares

$$ESS = \sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2,$$

and Total Sum of Squares

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

Lemma 4.

$$TSS = SSR + ESS.$$

Explanation. TSS is a measure of the variation of the dependent variable (if normalized properly, it is a good estimate of valiance of Y). It is decomposed into an explained part (ESS), which describes the variation due to X – since $\widehat{Y}_i$ are different only because the X_i are different, and the unexplained part- which is measured by the distance of the observations from the fitted line ($y = \widehat{\alpha} + \widehat{\beta}x$).

Proof.

$$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}((Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y}))^2 =$$
$$= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \bar{Y}).$$

We will prove that the last sum is zero. Indeed, the last sum is equal to:

$$\sum_{i=1}^{n}\widehat{e}_i(\widehat{\alpha} + \widehat{\beta}X_i - \bar{Y}) = \widehat{\alpha}\sum_{i=1}^{n}\widehat{e}_i + \widehat{\beta}\sum_{i=1}^{n}\widehat{e}_i X_i - \bar{Y}\sum_{i=1}^{n}\widehat{e}_i = 0.$$

Here we used the first order conditions from the optimization. □

Definition 5. $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \in [0, 1]$ is the fraction of the variation of Y explained by the regression.

R^2 is a measure of the *predictive* strength of the regression, but it has no direct relation to causality. Maximizing R^2 is **not** an objective of causal regression design.

The standard error of regression (SER) is defined as Ee_i^2 and estimated as

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{SSR}{n-2}.$$

It measures the 'typical' spread of the data around the regression line. The use of $n-2$ in the denominator of the estimator corresponds to the degree correction (two parameters are estimated to fit the regression line).

Slides with an example

Here is the list of questions we have discussed during lecture:

- What effect do we want to measure? What mental experiment describes it? If you wanted to run an experiment measuring this effect, what would you do?
- Describe the context of the available data set. Read the data description. Do you understand how different variables are measured?
- Look at the summary statistics. Do they look plausible? Everything as expected?
- Linear regression on education: what is the value of β? What does it mean? Is it of the right sign? Does the magnitude seem plausible? Is it statistically large (significant)? How would you test $H_0 : \beta = 0$?
- Look at the fit. What is R^2? SER? What does it mean?
- Look at the graphed data. Do they look homoskedastic? heteroskedastic?
- Look at the regression with heteroskedasticity-robust standard errors. Compare whether the following are the same: $\widehat{\beta}$, standard errors, t-statistics, confidence sets, R^2? Do you understand why?

Bivariate regression with a regressor-dummy

A dummy variable is variable that takes a 0-1 value, for example, $Female = 1$ for individuals who are female and 0 otherwise. Imagine a regression with a dummy regressor X. An example can be the regression:

$$Income_i = \alpha + \beta Female_i + e_i.$$

In such a case it is not useful to think about β as a slope, but rather to consider different cases. For example, for a male individual we will have $Income_i = \alpha + e_i$, that is, α is the average income of a male; while for a female we have $Income_i = \alpha + \beta + e_i$. That is, $\alpha + \beta$ is the average income of a female. Then β is the average income gap between genders.

What does OLS?

$$S(a,b) = \sum_{i=1}^{n}(Y_i - a - bX_i)^2 = \sum_{i \in Female}(Y_i - a - b)^2 + \sum_{i \in Male}(Y_i - a)^2 \to \min_{(a,b)}.$$

The optimization problem is equivalent to solving the following optimization problem, where we substitute $\gamma = \alpha + \beta$, and where we will use c for this in optimization:

$$S(a,c) = \sum_{i \in Female}(Y_i - c)^2 + \sum_{i \in Male}(Y_i - a)^2 \to \min_{(a,c)}.$$

This is equivalent to two problems about estimating means, then

$$\widehat{\alpha} = \frac{1}{\#Male}\sum_{i \in Male} Y_i, \quad \widehat{\gamma} = \frac{1}{\#Female}\sum_{i \in Female} Y_i, \quad \widehat{\beta} = \widehat{\gamma} - \widehat{\alpha}.$$

The problem of equality of the two means then is the same thing as the testing of $H_0 : \beta = 0$ in a regression with a binary regressor. Heteroskedasticity in such a regression corresponds to different variances for two sub-populations (males and females).