Slides for the course

Statistics and econometrics

Part 4: The simple regression model

European University Institute

Andrea Ichino

September 8, 2014

Outline

What a regression can do for us

Regression and the CEF

The Population Regression Function

Sample Regression Function and Population Regression Function

Three equivalent ways to estimate the Population Regression Function with the Sample Regression Function

Method of Moments

Ordinary Least Squares

Maximum Likelihood

Section 1

What a regression can do for us

The problem

- an outcome variable y: e.g. *labor earnings*;
- a variable x which we consider as a possible determinant of y in which we are interested: e.g. *years of education*;
- a variable e which describes all the other determinants of y that we do not observe.

The general notation for the model that relates y, x and e is

$$y = f(x, e) \tag{1}$$

We are interested in the relationship between x and y in the population, which we can study from two perspectives:

1. To what extent knowing x allows to “predict something" about y.
2. Whether Δx “causes" Δy given a proper definition of causality.

Subsection 1

Regression and the CEF

Regression and the conditional expectation function

Following Angrist and Pischke (2008) *Regression* is a useful tool because of its link with the *Conditional Expectation Function*.

We can always decompose (1) in the following way:

$$y = E(y|x) + \epsilon \tag{2}$$

where $E(y|x)$ is the CEF of y given x and $\epsilon = y - E(y|x)$ is:

- mean independent of x:

$$E(\epsilon|x) = E(y - E(y|x)|x) = E(y|x) - E(y|x) = 0 \tag{3}$$

- is uncorrelated with any function of x, i.e. for any h:

$$E(h(x)\epsilon) = E(h(x)E(\epsilon|x)) = 0 \tag{4}$$

An example of Conditional Expectation Function

Figure : The CEF of labor earnings given education in the US

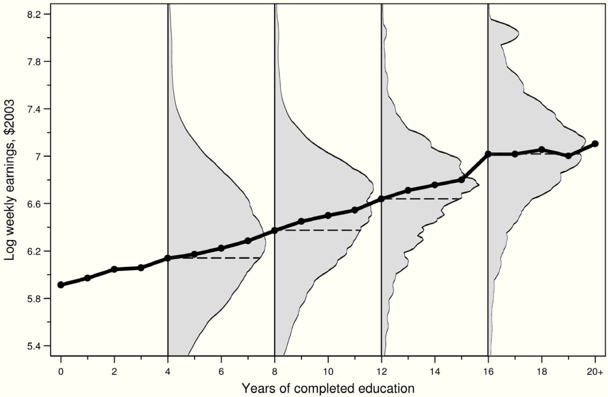


Figure 2.1.1 - CEF of Weekly earnings as a function of schooling.

The sample includes white men aged 40-49. The data are from the 1980 IPUMS 5% sample.

Interesting properties of the CEF

1. Let $m(x)$ be any function of x. The CEF solves

$$E(y|x) = \arg\min_{m(.)} E\left[(y - m(x))^2\right] \tag{5}$$

   and minimizes the Mean Square Error of the prediction of Y given X.

2. The variance of y can be decomposed in the variances of the CEF and of ϵ.

$$\begin{aligned} V(y) &= V(E(y|x)) + V(\epsilon) \\ &= V(E(y|x)) + E(V(y|x)) \end{aligned} \tag{6}$$

Exercise: prove the two properties.

Subsection 2

The Population Regression Function

The Population Regression Function

We do not know the CEF but we can show that the Population Regression Function (PRF) is a “good" approximation to the CEF:

$$y_p = \beta_0 + \beta_1 x \tag{7}$$

such that β_0 and β_1 minimize the square of the residual distance $u = y - y_p$ in the population, i.e. the “distance" between y and the PRF line itself:

$$(\beta_0, \beta_1) = \arg\min_{b_0, b_1} E\left[(y - b_0 - b_1 x)^2\right] \tag{8}$$

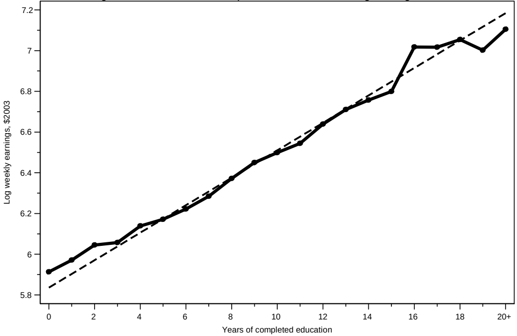
The First Order conditions of problem 8 are:

$$\begin{aligned} E\left[x(y - b_0 - b_1 x)\right] &= 0 \qquad (9) \\ E\left[(y - b_0 - b_1 x)\right] &= 0 \end{aligned}$$

An example of Population Regression function

Figure : The PRF of labor earnings given education in the US

Figure 3.1.2 - A conditional expectation function and weighted regression line



Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

The parameters of the PRF

The solutions are:

$$\beta_1 = \frac{E[x(y-\beta_0)]}{E(x^2)} = \frac{Cov(y,x)}{V(x)} \tag{10}$$

$$\beta_0 = E(y) - \beta_1 E(x) \tag{11}$$

Note that by definition of β_0 and β_1:

$$y = y_p + u = \beta_0 + \beta_1 x + u \tag{12}$$

and

$$E(xu) = E[x(y - \beta_0 - \beta_1 x)] = 0 \tag{13}$$

In words, the PRF is the linear function of x that makes the residuals u uncorrelated with x in the population.

Properties of the PRF

1. If the CEF is linear then the PRF is the CEF. This happens, specifically:
    - when y and x are jointly normally distributed;
    - in a fully saturated model (to be defined below in the context of multiple regression)

2. The PRF is the best linear predictor of y in the sense that it minimizes the Mean Square Error of the prediction.

3. The PRF is the best linear approximation to the CEF in the sense that it minimizes the Mean Square Error of the approximation.

Exercise: prove these properties.

Parenthesis: an informative exercise

Take any dataset and assume that this is your entire population

Define the variables of interest y and x.

Estimate the linear regression of y on x.

Compute $\bar{y} = E(y|x)$ and estimate the linear regression $\bar{y}$ on x.

Compare the results of the two estimations and comment on your findings.

In which sense the properties of the CEF and the PRF are relevant for your findings?

Could this result be useful whenever data providers do not want to release individual observations?

What have we accomplished so far

If we are simply interested in predicting y given x it would be useful to know the correspondent CEF because of its properties.

We do not know the CEF but the PRF is the best linear approximation to the CEF and the best linear predictor of y given x .

If we had data for the entire population we could then use the PRF, which we can characterize precisely, to predict y given x.

Usually, we have (at best) a random sample of the population.

We now have to show that the Sample Regression Function (SRF) is a “good" estimate of the PRF according to some criteria.

This is an inference problem.

Repetita juvant: again on the orthogonality condition

By saying that our goal is to estimate the PRF defined as:

$$y_p = \beta_0 + \beta_1 x \tag{14}$$

where the parameters satisfy by construction:

$$(\beta_0, \beta_1) = arg \min_{b_0, b_1} E\left[(y - b_0 - b_1 x)^2\right] \tag{15}$$

the orthogonality condition

$$E(xu) = E[x(y - \beta_0 - \beta_1 x)] = 0 \tag{16}$$

is *not a necessary assumption* for regression to make sense.

It follows instead from the definition β_0 and β_1 and ensures that:

- The OLS-MM estimator is by definition consistent for the PRF
- and unbiased in some important special cases.

Subsection 3

Sample Regression Function and Population Regression Function

The starting point: a random sample

Now suppose that we have a random sample of the population

Definition

If $\{z_1 ... z_i ... z_n\}$ are independent draws from a population with density function f(z,θ), then $\{z_1 ... z_i ... z_n\}$ is a random sample from the population defined by f(z,θ). Note that each draw is a random variable.

Exercise: make sure that you understand the meaning of random sampling.

The sample analog of the PRF

We want to know whether the sample analogs of

$$\beta_1 = \frac{Cov(y,x)}{V(x)} \qquad \text{and} \qquad \beta_0 = E(y) - \beta_1 E(x) \tag{17}$$

which are:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{18}$$

where we denote sample averages as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{and} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{19}$$

can be considered as “good" estimators of β_1 and β_0 under some criteria to be defined.

Why do we focus on the sample analog of β_1 (or β_0)?

Recall that an “estimator" is a function (a “recipe") of the sample which originates an “estimate" (a “cake") when the actual draws (the "ingredients") are combined in the way suggested by the estimator.

The “quality" of the estimate (the “cake") depends on the properties of the estimator (the “recipe") and on the characteristics of the actual sample (the “ingredients").

The “cakes" we want are the parameters β_0 and β_1 of the PRF which is the fitted line that minimizes residuals from y in the population.

We want to know whether the slope $\hat{\beta}_1$ of the *sample fitted line* (SRF) “approaches" the “cake we want", which is β_1. (Same for β_0)

We consider three justifications for using the SRF

Section 2

Three equivalent ways to estimate the Population Regression Function with the Sample Regression Function

Subsection 1

Method of Moments

The “Method of Moment" justification of $\hat{\beta}_0$ and $\hat{\beta}_1$

The “Methods of Moments" constructs estimators using restrictions imposed by population moments that should hold also in the sample (under random sampling).

The definition of the PRF parameters implies that the following two moment conditions should hold in the data

$$E(u) = E\left[y - \beta_0 - \beta_1 x\right] = 0 \tag{20}$$

$$E(xu) = E\left[x(y - \beta_0 - \beta_1 x)\right] = 0 \tag{21}$$

If the sample is a scaled down but perfect image of the population (a random sample), these two conditions should hold also in the sample.

The moment conditions in the sample

The analogs of the population moment conditions in the sample are:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad (22)$$

With simple algebra one can derive the MM estimators for β_1 and β_0:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad (23)$$

Note an important necessary condition :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0 \qquad (24)$$

What does this mean for your research question and empirical work?

Subsection 2

Ordinary Least Squares

The “Least Squares" justification of $\hat{\beta}_0$ and $\hat{\beta}_1$

$\hat{\beta}_0$ and $\hat{\beta}_1$ can also be chosen to minimizes the sum of squared residuals in the sample.

The PRF minimizes the sum of squared residual in the population, and the SRF should do the same in the sample

The Ordinary Least Square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are constructed as

$$(\hat{\beta}_0, \hat{\beta}_1) = arg \min_{\hat{b}_0, \hat{b}_1} \sum_{i=1}^{n} \left[(y_i - \hat{b}_0 - \hat{b}_1 x_i)^2 \right] \tag{25}$$

It is easy to check that the FOCs of this problem are identical to (22):

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{26}$$

The OLS estimators

Since the OLS conditions (26) and the MM conditions (22) are the same, they deliver the same estimators:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \tag{27}$$

The second order conditions of problem (25) are satisfied.

The way to do it is to add and subtract $\hat{\beta}_0 + \hat{\beta}_1 x_i$ within the squared parentheses in the minimand (25) to get

$$\sum_{i=1}^{n}\left[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) + (\hat{\beta}_0 - \hat{b}_0) + (\hat{\beta}_1 x_i - \hat{b}_1 x_i)\right]^2 \tag{28}$$

Developping the square one can show that the minimum occurs for $\hat{b}_0 = \hat{\beta}_0$ and $\hat{b}_1 = \hat{\beta}_1$.

Subsection 3

Maximum Likelihood

The “Maximum Likelihood" justification of $\hat{\beta}_0$ and $\hat{\beta}_1$

There is a third way to justify the $\hat{\beta}_0$ and $\hat{\beta}_1$ estimators based on the logic of Maximum Likelihood (ML).

This justification requires the assumption that y is distributed normally.

Thanks to this distributional assumption, in addition to the MM and OLS desirable properties that we will discuss below, $\hat{\beta}_0$ and $\hat{\beta}_1$ acquire also the properties of ML estimators.

We will discuss the additional properties of ML estimators later.

Now we just want to show that $\hat{\beta}_0$ and $\hat{\beta}_1$ can also be interpreted as ML estimates, under the assumption of normality.

The likelihood function

Consider the model

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{29}$$

$$u_i \sim f(u_i|0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u_i)^2}{2\sigma^2}} \tag{30}$$

which implies

$$y_i \sim f(y_i|0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \tag{31}$$

Given a random sample $\{y_i\}$ and $\{x_i\}$, the likelihood function is:

$$L(y|x,\beta_0,\beta_1,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \tag{32}$$

i.e.: the probability of observing the sample given β_0, β_1 and σ^2.

The “Recipe" of maximum likelihood estimation

The ML estimator chooses $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ as the values of β_0, β_1 and σ^2 that maximize the likelihood, given the observed sample.

$$\{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\} = \underset{\beta_0, \beta_1, \sigma^2}{\text{argmax}}\, L(y|x, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} \tag{33}$$

Computations simplify if we maximize the log likelihood:

$$Log[L(y|x, \beta_0, \beta_1, \sigma^2)] = \sum_{i=1}^{n} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}\right] \tag{34}$$

$$= -\frac{N}{2} log(2\pi) - \frac{N}{2} log(\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}$$

First Order Conditions for β_0 and β_1

Maximization of the log likelihood with respect to β_0 and β_1 implies:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\text{argmax}} \sum_{i=1}^{n} \left[(y_i - \beta_0 - \beta_1 x_i)^2 \right] \tag{35}$$

ML FOC, are identical to the contitions for MM and OLS :

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{36}$$

Solving the FOC we get the same estimator:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{37}$$

Second Order Conditions can be checked as for OLS.