ECON-122
Introduction to Econometrics

Agnieszka Postepska

July 12, 2011

Introduction to estimation

Cook book procedure - suppose we are back in our vitamins-school performance example

- decide which population is of interest (eg. all high school children in the District)
- decide on a sample size n (eg. 15)
- draw a random sample of size n from this population (eg. assign numbers to students and have the computers draw 15 random numbers from the set of all students in the District)
- scatter plot of our data:
- decide how to use our data to estimate the slope and intercept of the regression line

Motivation

- we will now use the two assumptions we saw yesterday
- recall A1: $E(u) = 0$, and A2: $E(u|x) = E(u) = 0$
- Note that A1 imply that $Cov(x, u) = E(xu) = 0$

Proof.

$$\begin{aligned} Cov(x,u) &= E(xu) - E(x)E(u) \\ &= E(xu) \\ &= E[E(xu|x)] \qquad \text{by Law of Iterated Expectations} \\ &= E[xE(u|x)] \\ &= 0 \qquad\qquad \text{QED} \end{aligned}$$

Why are the two assumptions and the 0 covariance so important?

- Rewrite $E(u) = 0$ as $E(y - \beta_0 - \beta_1 x) = 0$
- Rewrite $E(xu) = 0$ as $E[x(y - \beta_0 - \beta_1 x)] = 0$
- these two equations imply two restrictions we can use to estimate the two parameters - in other words: when we take this to the data, the $y's$ and the $x's$ are known - the $\beta's$ are to be estimated. So we will solve the two equations for β_0 and β_1 in terms of $y's$ and $x's$
- however, β_0 and β_1 are the true value of the parameters and the equations above describe the moments of the populations and we do not have the whole population to solve these equations

How do we go from the population moments to the sample?

- So, there is one more step to go before we can do it - the two restrictions above describe the two moments of the population, not the sample. So we need to find the sample analog first - that's basic statistics.

$$E(y - \beta_0 - \beta_1 x) = 0 \quad \Rightarrow \frac{1}{n}\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$E[x(y - \beta_0 - \beta_1 x)] = 0 \quad \Rightarrow \frac{1}{n}\sum_i x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

- this is an example of the *method of moments* estimation
- now we can solve these equations for $\hat{\beta}_0$ and $\hat{\beta}_1$...

How well do you know summation operator?

- Look at the first equation and rewrite the LHS:

$$\begin{aligned}\frac{1}{n}\sum_i(y_i-\hat{\beta}_0-\hat{\beta}_1x_i) &= \frac{1}{n}\sum_i y_i-\frac{1}{n}\sum_i\hat{\beta}_0-\frac{1}{n}\sum_i\hat{\beta}_1x_i\\ &= \frac{1}{n}\sum_i y_i-\frac{1}{n}n\hat{\beta}_0-\frac{1}{n}\sum_i\hat{\beta}_1x_i\\ &= \bar{y}-\hat{\beta}_0-\hat{\beta}_1\bar{x}\end{aligned}$$

so the first restriction becomes:

$$\begin{aligned}\bar{y}-\hat{\beta}_0-\hat{\beta}_1\bar{x} &= 0\\ \Rightarrow\hat{\beta}_0 &= \bar{y}-\hat{\beta}_1\bar{x}\end{aligned}$$

This mathematical formula is an **estimator** for the intercept of the regression function.

How well do you know summation operator? cont.

- Now, look at the second equation and multiply both sides by n:

$$\begin{aligned}
\frac{1}{n}\sum_i x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\sum_i x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\sum_i x_i[y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] &= 0 \\
\sum_i x_i(y_i - \bar{y}) - \hat{\beta}_1 \sum_i x_i(x_i - \bar{x}) &= 0 \\
\Rightarrow \hat{\beta}_1 &= \frac{\sum_i x_i(y_i - \bar{y})}{\sum_i x_i(x_i - \bar{x})}
\end{aligned}$$

This mathematical formula is an **estimator** for the slope of the regression function.

Closer look at the estimator for the slope

Note that we can rewrite the expression for $\hat{\beta}_1$ in the following way:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i x_i(y_i - \bar{y})}{\sum_i x_i(x_i - \bar{x})} \\
&= \frac{\sum_i x_i y_i - \bar{y}\sum_i x_i}{\sum_i x_i^2 - \bar{x}\sum_i x_i} \\
&= \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}
\end{aligned}$$

Also note the following equalities:

$$\begin{aligned}\sum_i (x_i - \bar{x})(y_i - \bar{y}) &= \sum_i x_i y_i - \bar{y}\sum_i x_i - \bar{x}\sum_i y_i + \sum_i \bar{x}\bar{y} \\ &= \sum_i x_i y_i - \bar{y} n\bar{x} - \bar{x} n\bar{y} + n\bar{x}\bar{y} \\ &= \sum_i x_i y_i - \bar{y} n\bar{x}\end{aligned}$$

which is what we have in the denominator above

And:

$$\begin{aligned}
\sum_i x_i(x_i - \bar{x}) &= \sum_i x_i^2 - \bar{x}\sum_i x_i \\
&= \sum_i x_i^2 - n\bar{x}^2 \\
\sum_i (x_i - \bar{x})^2 &= \sum_i x_i^2 - 2\bar{x}\sum_i x_i + \sum_i \bar{x}^2 \\
&= \sum_i x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_i x_i^2 - n\bar{x}^2 \\
\text{therefore: } \sum_i x_i^2 - n\bar{x}^2 &= \sum_i x_i(x_i - \bar{x}) = \sum_i (x_i - \bar{x})^2
\end{aligned}$$

Finally, we can rewrite the estimator for the slope as:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

which is the sample covariance divided by the sample variance and β_1 has the sign of the correlation between x and y

Motivating the name OLS

- Define **fitted value** for y when $x = x_i$ as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Define the **residual** for observation i as the difference between the actual value of y and its fitted value:

$$\hat{u} = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Note that residuals ARE NOT the same as errors

- Let's pick β_0 and β_1 so that the sum of squared residuals is as small as possible:

$$\min_{\beta_0,\beta_1} \sum_i^n \hat{u}_i^2 \Rightarrow \min_{\beta_0,\beta_1} \sum_i^n (y_i - \hat{\beta}_0 - \hat{\beta}_i x_i)^2$$

- The first order conditions are the following:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})} \end{aligned}$$

which is exactly what we had before

Introduction to Algebraic properties of OLS estimator

- the following properties are sample not population properties
- they hold "by construction", so they are true regardless of the sample we draw
- keep in mind that the regression line is representation of the fitted, not actual values of y given values of $x \longrightarrow$ none of the actual data points must actually lie on the OLS line $\longrightarrow$ none of the residuals must be 0

Algebraic properties of OLS

- the sum of the OLS residuals is 0 - on average residual is equal to 0

$$\sum_i \hat{u}_i = 0$$

- the sample covariance between the $x's$ and the residuals, $\hat{u}'s$, is 0

$$\sum_i x_i \hat{u}_i = 0$$

- the point $(\bar{x}, \bar{y})$ lies always on the regression line; in other words the predicted value for y, $\hat{y}$, when $x = \bar{x}$ is $\bar{y}$

Decomposition of y_i

- note that we can write y_i as:

$$y_i = \underbrace{\hat{y}_i}_{\text{fitted value}} + \overbrace{\hat{u}_i}^{\text{residual}}$$

- define the following:

$$\begin{aligned} SST &\equiv \sum_{i=1}^{n}(y_i - \bar{y})^2 \rightarrow \text{ total sum of squares} \\ SSE &\equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \rightarrow \text{ explained sum of squares} \\ SSR &\equiv \sum_{i=1}^{n}\hat{u}_i^{\,2} \rightarrow \text{ residual sum of squares} \\ \text{and } SST &= SSE + SSR \end{aligned}$$

Goodness of fit R^2

- in the linear regression model we can actually measure how good does the independent variable, x, explain the dependent variable y

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- interpretation: fraction of the sample variation in y that is explained by x (if multiplied by 100, it has the interpretation in terms of percentages)
- note that $0 \leq R^2 \leq 1$ (by the equality from decomposition of y_i) - the model cannot explain more variation than actually exists
- when does $R^2 = 0$ occur?
- no threshold value to judge whether the model is good/bad - do not overuse it!!!!
- R^2 also tells us nothing about the correctness of the model itself