14.320: Recitation 5

Martina Uccioli*

March 9, 2023

1 Regression Recap

Let's consider the following linear population regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

The **regression coefficients** are α and β such that

$$E(\epsilon_i) = 0 \text{ and} \tag{1}$$
$$E(X_i \epsilon_i) = 0. \tag{2}$$

These two conditions give you two equations in two unknowns, and are solved by $\alpha = E[Y_i] - \beta E[X_i]$ and $\beta = \frac{COV(Y_i, X_i)}{V(X_i)}$ (COV and V refer to the covariance and variance in the population).

Why do we care about them?

1. *The function $\alpha + \beta X_i$ is the best linear predictor of Y_i given X_i, in the sense of minimizing the mean squared (prediction) error: $MSE_p(a, b) = E\{Y_i - (a + bX_i)\}^2$.*

   Note: this was the original motivation for regression, and this is what it is generally used to *define* population regression coefficients (e.g. in MHE)

   Proof. The first order conditions of the minimization problem are

   $$E\{Y_i - (a + bX_i)\} = 0$$
   $$E\{X_i[Y_i - (a + bX_i)]\} = 0$$

   which are the same as conditions (1) and (2).

*This material draws extensively on Viola Corradini's recitation notes - many thanks to her and to previous generations of 14.320 TAs.

2. *If the CEF is linear, then $\alpha + \beta X_i$ is the CEF.*

   Proof. First, recall that the CEF $E[Y_i|X_i]$ has the following properties[1]:

   $$E\{Y_i - E[Y_i|X_i]\} = 0 \text{ and } E\{X_i(Y_i - E[Y_i|X_i])\} = 0 \ .$$

   If the CEF is linear, they become $E\{Y_i-(a+bX_i)\} = 0$ and $E\{X_i(Y_i-(a+bX_i))\} = 0$, which are the same as conditions (1) and (2).

3. *Even if the CEF is not linear, $\alpha + \beta X_i$ provides the best linear approximation to it, in the sense of minimizing the mean squared approximation error: $MSE_a(a,b) = E\{E[Y_i|X_i] - (a+bX_i)\}^2$.*

   Proof. Note that

   $$\begin{aligned} E\{Y_i - (a+bX_i)\}^2 =& E\{Y_i - E[Y_i|X_i] + E[Y_i|X_i] - (a+bX_i)\}^2 \\ =& E\{Y_i - E[Y_i|X_i]\}^2 + E\{E[Y_i|X_i] - (a+bX_i)\}^2 \\ &+ 2E\{(Y_i - E[Y_i|X_i])(E[Y_i|X_i] - (a+bX_i))\} \end{aligned}$$

   The last term is equal to zero by Law of Iterated Expectations (see footnote 1). a and b do not appear in the first term, so whichever a, b minimize $E\{Y_i - (a+bX_i)\}^2$ are the same a, b that minimize $E\{E[Y_i|X_i] - (a+bX_i)\}^2$. Since we have shown above that the regression coefficients α, β minimize $MSE_p(a,b) = E\{Y_i - (a+bX_i)\}^2$, the regression coefficients α, β are also the solution to the minimization of $MSE_a(a,b) = E\{E[Y_i|X_i] - (a+bX_i)\}^2$.

Quoting MM: "To summarize: if the CEF is linear, regression finds it; if it is nonlinear, regression finds a good approximation to it".

2 Unbiasedeness of OLS

Aside

Let's review what are the CEF, the population regression and the OLS estimator of the population regression parameters.

CEF It is $E[Y_i|X_i]$. It summarizes the relationship between Y and X in the population, and this is generally the object we are interested in. Since it is a population object, we typically don't know the shape of the CEF. But it need not be linear. For instance, Y and X might be related in the following way: $Y_i = 3X_i^2 + u_i$, where u_i is just "noise" around the CEF, that is $E[u_i|X_i] = 0$. In this case the CEF is not a linear function of X_i. In fact, $E[Y_i|X_i] = 3X_i^2$.

[1] Proof using the Law of Iterated Expectations:

$$E\{Y_i - E[Y_i|X_i]\} = E\{E\{Y_i - E[Y_i|X_i]\}|X_i\} = E\{E\{[Y_i|X_i] - E[Y_i|X_i]\}\} = 0$$

$$E\{X_i[Y_i - E[Y_i|X_i]]\} = E\{E\{X_i[Y_i - E[Y_i|X_i]]\}|X_i\} = E\{X_i(E\{[Y_i|X_i] - E[Y_i|X_i])\}\} = 0$$

Population regression Let's consider the following linear population regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Of course, the linear model may not be correct, in the sense that the true relationship between Y_i and X_i need not be linear. This is a linear approximation of the relationship between Y and X in the population. We have shown above that the best linear approximation of the CEF is given by $\alpha + \beta X_i$, where $\alpha = E[Y_i] - \beta E[X_i]$ and $\beta = \frac{COV(Y_i, X_i)}{V(X_i)}$.
When the CEF is linear, the CEF is actually identical to $\alpha + \beta X_i$ with α and β defined as above. When the CEF is not linear, then the population regression is the best linear approximation to the CEF.
Importantly, when the CEF is not linear, then it is not the case that $E[\epsilon_i|X_i] = 0$. Why is that? Consider for instance the example above, in which the true population relationship between Y and X is $Y_i = 3X_i^2 + u_i$, with $E[u_i|X_i] = 0$. Then,

$$\epsilon_i = Y_i - \alpha - \beta X_i = 3X_i^2 - \alpha - \beta X_i + u_i.$$

Therefore $E[\epsilon_i|X_i] \neq 0$ in this case, because the conditional expectation of the "misspecification" part of the error (due to the fact that we are using a linear model when in fact the true relationship is not linear), $E[3X_i^2 - \alpha - \beta X_i|X_i]$ need not be zero: it will be systematically higher for certain values of X and lower for other values of X. This will be important to keep in mind for our proof of the OLS unbiasedeness.
However, $E[\epsilon_i] = E[Y_i - \alpha - \beta X_i] = 0$. This simply follows mechanically from the formula of α and β.

OLS estimator The OLS estimator is an estimator of the parameters of the population regression model. It uses data from our sample to produce an estimate of the population parameters α and β above.
What does it mean that $\hat{\beta}_{OLS}$ is unbiased? It means that when I take the expectation of the estimator over repeated samples of the data I get the population regression parameter, that is: $E[\hat{\beta}_{OLS}] = \beta$. Intuitively, if $\hat{\beta}_{OLS}$ is unbiased, then if I could repeatedly draw a very large number of different samples from the population, compute the OLS estimate of β on each sample, and then take an average over all these estimates, I would get something very close to the population regression parameter $\beta = \frac{COV(Y_i, X_i)}{V(X_i)}$. Obviously we cannot do what I just said since in practice we typically only have one sample. But we can still think about and show whether/when $\hat{\beta}_{OLS}$ is unbiased.

Proof that $\hat{\beta}_{OLS}$ is unbiased when X is fixed in repeated samples

What does it mean for X_i to be fixed vs. random? X_i is fixed in repeated samples if it has the exact same distribution in each sampling draw.
Here's an example: suppose we are randomly assigning 100 subjects to treatment in a randomized control trial so that $X_i = 1$ if treated and $X_i = 0$ if control. One way to randomize is to assign every subject a random number, sort them, and then assign the first 50 subjects to treatment and the next 50 subjects to control. Then any time we re-ran the randomization, we will have exactly 50 treated and 50 control and we will know that $X_i = 1$ if $i \leq 50$

and $X_i = 0$ if $i > 50$. So in this case, the distribution of X_i is fixed. In this sense, X_i may be treated as a constant when taking expectations.
Another way to do the randomization is just to flip a coin for each individual. Then the proportion of treatment and control could change if we re-ran the randomization (e.g. 47 treated on one draw, 51 treated on the next, etc.). We will also not know with certainty whether the first observation is treated or not. In this sense, X_i would be stochastic.
Let's see how this distinction affects the math. To show that $\hat{\beta}$ is unbiased, substitute

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

into the numerator of the expression for $\hat{\beta}$:

$$\begin{aligned}
\hat{\beta} &= \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2} \\
&= \frac{\sum \left(X_i - \bar{X}\right) Y_i}{\sum \left(X_i - \bar{X}\right)^2} - \frac{\bar{Y} \sum \left(X_i - \bar{X}\right)}{\sum \left(X_i - \bar{X}\right)^2} \\
&= \frac{\sum \left(X_i - \bar{X}\right) Y_i}{\sum \left(X_i - \bar{X}\right)^2} \\
&= \frac{\sum \left(X_i - \bar{X}\right)\left(\alpha + \beta X_i + \epsilon_i\right)}{\sum \left(X_i - \bar{X}\right)^2} \\
&= \frac{\beta \sum \left(X_i - \bar{X}\right) X_i + \sum \left(X_i - \bar{X}\right) \epsilon_i}{\sum \left(X_i - \bar{X}\right)^2} \\
&= \beta + \frac{\sum \left(X_i - \bar{X}\right) \epsilon_i}{\sum \left(X_i - \bar{X}\right)^2}
\end{aligned}$$

and to show that $\hat{\beta}$ is unbiased, we take the expectation of both sides:

$$E\left[\hat{\beta}\right] = E\left[\beta\right] + E\left[\frac{\sum \left(X_i - \bar{X}\right) \epsilon_i}{\sum \left(X_i - \bar{X}\right)^2}\right]$$

Suppose that X_i is fixed.

$$\begin{aligned}
E\left[\hat{\beta}\right] &= E\left[\beta\right] + E\left[\frac{\sum \left(X_i - \bar{X}\right) \epsilon_i}{\sum \left(X_i - \bar{X}\right)^2}\right] \\
&= E\left[\beta\right] + \frac{\sum \left(X_i - \bar{X}\right)}{\sum \left(X_i - \bar{X}\right)^2} E[\epsilon_i] \\
&= \beta + \frac{\sum \left(X_i - \bar{X}\right)}{\sum \left(X_i - \bar{X}\right)^2} \times 0 = \beta
\end{aligned}$$

Recall that we're taking the expectation of the estimator over repeated samples of the data. In this case, X_i has the same distribution with each sampling draw, so the sums involving X_i are constant. Therefore we can pull out the sums from the expectation. That leaves us with $E[\epsilon_i]$ which is identically zero. Therefore $E[\hat{\beta}] = \beta$ and the estimator is unbiased.

Proof that $\hat{\beta}_{OLS}$ is unbiased when the CEF is linear

Suppose instead that X_i is random but that the CEF is linear. Then

$$\begin{aligned} E\left[\hat{\beta}\right] &= E\left[\beta\right] + E\left[\frac{\sum\left(X_i - \bar{X}\right)\epsilon_i}{\sum\left(X_i - \bar{X}\right)^2}\right] \\ &= \beta + E\left[E\left[\frac{\sum\left(X_i - \bar{X}\right)\epsilon_i}{\sum\left(X_i - \bar{X}\right)^2}|X_i\right]\right] \\ &= \beta + E\left[\frac{\sum\left(X_i - \bar{X}\right)E\left[\epsilon_i|X\right]}{\sum\left(X_i - \bar{X}\right)^2}\right] \end{aligned}$$

Now we can no longer treat X_i as a constant distribution. It has to stay inside the expectation. But we can use the Law of Iterated Expectations. Once we've conditioned on X_i in the inner expectation, we can pull out the sums involving X_i. We're left with $E[\epsilon_i|X_i]$. When the CEF is ***linear,* this term is equal to zero.**

$$\begin{aligned} E[\epsilon_i|X_i] &= E[(Y_i - \alpha - \beta X_i)|X_i] \\ &= E[Y_i|X_i] - \alpha - \beta X_i \\ &= \alpha + \beta X_i - \alpha - \beta X_i \\ &= 0 \end{aligned}$$

Therefore $E[\hat{\beta}] = \beta$ and the estimator is unbiased.

Side note: even when the CEF is nonlinear, $\hat{\beta}_{OLS}$ is still an unbiased estimator of the *slope of the best linear approximation* to the CEF. So, you should not worry too much about nonlinear CEFs and biasedness of OLS. As long as what you care about is the best linear approximation to the CEF, you are fine!