Preliminaries

We start with a set of ordered pairs $\{\langle X_1, Y_1\rangle, \langle X_2, Y_2\rangle, \langle X_3, Y_3\rangle, ..., \langle X_n, Y_n\rangle\}$.

You can think of X_i and Y_i as

- real numbers (facts about each of the the n individuals in the population)
- or as random variables (probability distributions over facts about n individuals in a sample),

all the maths will apply equally. (I will return to this fact and comment on it).

The CEF minimises $\sum w_i^2$

Some algebraic facts

We write the equality:

$$Y_i = f(X_i) + w_i$$

Where Y_i and X_i are known, but w_i depends on our choice of f.

Minimisation problem

Suppose we want to solve

$$\min_{f(X_i)} \sum w_i^2 \leftrightarrow \min_{f(X_i)} \sum (Y_i - f(X_i))^2$$

The solution is $f(X_i) = E[Y_i \mid X_i]$. The proof of this is in appendix A. Suppose we specify $f(X_i)$ as such, we then get:

$$Y_i = E[Y_i \mid X_i] + w_i$$

Now f is known and w_i is known (by the subtraction $w_i = Y_i - E[Y_i \mid X_i]$).

The LRM minimises $\sum (e_i + w_i)^2$

Some algebraic facts

Now we write the following equality:

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i + e_i$$

This says that $E[Y_i \mid X_i]$ is equal to a linear function of X_i plus some number e_i.

We then have

$$\begin{aligned} Y_i &= E[Y_i \mid X_i] + w_i \\ &= \beta_0 + \beta_1 X_i + e_i + w_i \end{aligned}$$

As before w_i is known, whereas e_i is a function of β_0 and β_1.

Here e_i is the distance, for observation i, between the LRM and the CEF; while w_i is the distance between the CEF and the actual value of Y_i. We can then call $u_i = e_i + w_i$ the distance between the LRM and the actual value.

We can also see that $E[u_i \mid x_i] = 0$ is equivalent to $e_i = 0$, i.e. the CEF and the LRM occupy the same coordinates.

Minimisation problem

Suppose we want to solve

$$\min_{\beta_0,\beta_1} \sum (e_i + w_i)^2 \leftrightarrow \min_{\beta_0,\beta_1} \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

The solution is

$$\begin{aligned}
\beta_0 &= \bar{y} - \beta_1 \bar{x} \\
\beta_1 &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{aligned}$$

I prove this in appendix B. (It's possible to prove an analogous result in general using matrix algebra, see appendix C.)

Suppose we specify that β_0 and β_1 are equal to these solution values. Now that β_0 and β_1 are known, e_i is known too (by the subtraction $e_i = E[Y_i \mid X_i] - \beta_0 - \beta_1 X_i$). As before, w_i is known.

Thus, in our regression equation,

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + e_i + w_i \\
&= \beta_0 + \beta_1 X_i + u_i
\end{aligned}$$

all of Y_i, X_i, β_0, β_1, e_i and w_i (and thus u_i), are known.

Comments

A few things to note at this point. Whether we are using real numbers of random variables does not matter for anything we've said so far. All we have used are the expectation and summation operators and their properties. Textbooks often warn about the important distinction between the sample and the population, but as far as these algebraic facts are concerned the difference is immaterial! This confused me for a long time before I understood it.

The second thing to note is that I have not used "hat" notation (as in $\hat{\beta}$). Instead I have described the results of optimisation procedures carefully using words, like "the solution to this minimisation problem is ...".

This is because the way standard econometrics uses the hat has been the source of much confusion for me.

Inconsistent hats

Econometrics textbooks, within the same sentence or paragraph, routinely use the hat in two ways which seem to me to be incompatible.

The 'loss function minimiser' usage

Claim A (Stock and Watson (2015), p. 187):

> The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of b_0 and b_1 that minimise $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$.

This is consistent with **Claim B** (Stock and Watson p. 163):

> The predicted value of Y_i, given X_i, based on the OLS regression line, is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

Claim A tells us $\hat{\beta}_0$ and $\hat{\beta}_1$ are loss function minimisers. Claim B tells us that $\hat{Y}_i$ is the value obtained when you compute the values of b_0 and b_1 which minimise a loss function, and plug them into the regression function.

All very well so far.

The 'sample analogue' usage

However, elsewhere (Stock and Watson p. 158) we have **Claim C**:

> The linear regression model is:
>
> $$Y_i = \beta_0 + \beta_1 X_i + u_i$$

> Where $\beta_0 + \beta_1 X$ is the population regression line or population regression function, β_0 is the intercept of the population regression line, and β_1 is the slope of the population regression line.

So far as I can tell, the "population regression intercept" and "population regression slope" are *defined* as the values that minimise $\sum u_i^2$, i.e. β_0 and β_1 are also the solutions which minimise $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$. So, by the loss function minimiser usage above, we get: $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$. (Otherwise, if we take the population LRM to be simply some equality in the population, where β_0, β_1 and u_i can take any consistent values, I don't see what sense can be made of the expressions "sample analogue of β_1", say.)

However, $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$ is not compatible with **Claim D**, which makes the sample analogue usage (Stock and Watson, p. 163):

> The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, are sample counterparts of the population coefficients β_0 and β_1. Similarly, the OLS regression line $\hat{\beta}_0 + \hat{\beta}_0 X$ is the sample counterpart of the population regression line $\beta_0 + \beta_1 X$ and the OLS residuals $\hat{u}_i$ are sample counterparts of the population errors u_i.

This quote implies that $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables, which couldn't possibly be equal to the real numbers β_0 and β_1.

Furthermore, if the hat is supposed to mean "sample analogue", we would expect $\hat{Y}_i$ to be the sample counterpart of Y_i, that is, we would expect $\hat{Y}_i$ to be the ith value of Y in a sample. Yet we have seen above that the loss function minimiser usage defines $\hat{Y}_i$ as a "predicted value" of Y_i (given the loss-function-minimising $\hat{\beta}_0$ and $\hat{\beta}_1$).

One (bad) solution

The following interpretation makes sense of *some* of the claims economists make. Some rather important ones will have to go. I haven't found a way to make sense of all the claims, they seem incompatible to me.

We think of

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Not as a regression equation, but as a complete causal account of everything causally affecting Y. For example, if there are ϕ things causally affecting Y, we have:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + ... + \beta_\phi \phi_i$$

We can think of this claim as equivalent to an infinite lists of counterfactuals, giving the potential values of Y for every combination of values of the causal factors $X, A, B..., \phi$. It also makes the claim that nothing else has a causal effect on Y.

(if we think the world is non-deterministic, the claim becomes $Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + ... + \beta_\phi \phi_i + \varepsilon_i$, where ε_i are i random variables, and we have a list of counterfactuals giving the potential *distributions* of Y for every combination of values of the causal factors.)

That's a rather huge claim. In any realistic case, causal chains are incredibly long and entangled, so that basically everything affects everything else in some small way. So the claim often amounts to an entire causal model of the world. This makes sense of why economists keep calling the β coefficients "unobservable" - they truly are nearly impossible to observe under this interpretation.

Let's go back to $Y_i = \beta_0 + \beta_1 X_i + u_i$ (1), which, remember, is a huge causal claim and *not* the linear regression function. We can keep Claim A and Claim B (though it's not clear what use they will be). Claims C and D immediately go out the window, since (1) has nothing to do with regression coefficients, let alone the sample counterparts of regression coefficients.

Under this interpretation, the claim:

$$\beta_0 = E[Y] - \beta_1 E[X]$$
$$\beta_1 = \frac{cov(X_i, Y_i)}{var(X_i)}$$

is completely false. It would be shocking (!) if the true causal effects β_0 and β_1 were equal to some simple function of the moments of Y and X.

Claims A and B imply:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Now, the assumption $E[u_i \mid x_i] = 0$, which is needed to relate $\hat{\beta}$ to β, is a huge and unlikely causal claim.