Econometrics. A field where the concepts are simple, but the real challenge is making sense of notation so obfuscatory that you wonder if it's done on purpose.

In order to arrive at this statement, I went through a long and confusing journey, one I wish upon no friend. This document's structure takes my journey in reverse order.[1] I start with what I eventually pinned down as the clear mathematical facts. Once armed with this toolkit, I do my best to explain why standard notation is confusing, and attempt to guess, from context, what econometricians actually mean.

In my view, it's a pretty scathing indictment of the field that I spent about *ten times longer* engaging in this interpretative guesswork than I spent understanding the underling concepts.

The facts

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.[2]

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

Two things to note about the facts above.

- 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! Cue ten hours of confusion.
- 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 ...". The way standard econometrics uses the hat is a prime example of obfuscatory notation.

The hermeneutics

Inconsistent hats

Econometrics textbooks, within the same sentence or paragraph, routinely use the hat in two ways which seem to me to be incompatible. I here give my best interpretative guess.

Sample analogues?

In Stock and Watson, p. 158, we have Claim A:

> 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.

Stock and Watson, p. 163 (Claim B):

> 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.

So far so good.

Loss function minimisers?

Stock and Watson, p. 187 (Claim C):

> 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 quote is the biggest culprit. After many conversations, I finally understood that we're supposed to take the quote to mean:

> The OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of b_0 and b_1 that minimise $\sum_{i=1}^{j}(Y_i^{sample} - b_0 - b_1 X_i^{sample})^2$, where j is the number of observations in the sample ($j < n$ if n is the sample size) and Y_i^{sample} and X_i^{sample} are the ith values in the sample.

I swear, I'm not taking this quote out of context! Nowhere, in the entire textbook, would you find a clue that the X_i and Y_i in claim C are *completely different quantities* than X_i and Y_i in claim A. This is criminal negligence. (I'm also not cherry-picking. My lecture notes cheerfully call $\hat{\beta}_0$ and $\hat{\beta}_1$ the 'OLS' solutions, and this usage is standard.)

Of course, I took claim C at face value, and combined it with claim A, to arrive at $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$, which, I gathered from context, was *not* a desirable conclusion.

Confusing hats

The following is not as bad as the above, since it avoids explicit contradiction, but still sows confusion by using the hat to mean different things when put on top of different values.

Claim D, from Stock and Watson p. 163:

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

This is compatible with the loss function mimimiser usage of the hat: claim C, which us $\hat{\beta}_0$ and $\hat{\beta}_1$ are loss function minimisers; claim D then 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.

But, of course, this "predicted value" verbiage is incompatible with the sample analogue usage. $\hat{Y}_i$ can't be both the predicted value (whether in a sample or not) and the *actual* value in a sample. That would imply that predictions are always perfect!

So *even if* we amend claim C as I've done above, we still can't say that the hat is consistently used to mean sample analogue, since in the case of $\hat{Y}_i$ it's apparently used to mean predicted value. (More specifically predicted value *in a sample*, one guesses from context. Hermeneutical ambiguities abound).

Inconsistent causal language

It gets worse. In all of the above we have taken the statement

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

to be an innocuous equality: Y_i is equal to regression intercept, plus regression slope times X_i, plus some remaining difference. Call this this the *algebraic claim*.

But it turns out that the statement is sometimes used to make a completely different, and incredibly strong, causal claim. Econometricians switch between the two usages in a classic case of motte and bailey.

In keeping with the above structure, I'll start with clearly stating the causal claim, then I'll analyse quotes which trade on the ambiguity between the causal and algebraic claims.

The causal claim

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..., ϕ. 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.

Hermeneutics (II)

Stock and Watson p. 158, claim E:

> The term u is the error term [...]. This term contains all the other factors besides X that determine the value of the dependent variable, Y, for a specific observation i

This is a favourite trick: use a word like "determines", which heavily implies a causal claim, but stay just shy of being unambigously causal. That way you can always retreat to the algebraic claim.

Indeed, under the algebraic interpretation, claim E is puzzling. What on earth does it mean for a number to "contain", "factors" that "determine" the value of another number? As far as the mathematics is concerned, we have no concept of "determine", much less of a number "containing" another number.

A causal variant of claim E would be:

> The term u is the further-causes term [...]. This term contains all the other factors besides X that cause the value of the dependent variable, Y, for a specific observation i

Wooldrige, p.92f, claim F:

> Assumption MRL.4:
>
> > $E[u \mid x_1, x_2, x_3, ..., x_k] = 0$
>
> When assumption MLR.4 holds, we often say that we have **exogenous explanatory variables**. If x_j is correlated with u for any reason, then x_j is said to be an **endogenous explanatory variable** [...] Unfortunately, we will never know for sure whether the average value of the unobservables is unrelated to the explanatory variables.

Stock and Watson, p.131, claim G:

> The causal effect of a treatment is the expected effect on the outcome of interest of the treatment as measured in a ideal randomized controlled experiment. This effect can be expressed as the difference of two conditional expectations. Specifically, the causal effect on Y of treatment level x is the difference in the conditional expectations $E[Y \mid X = x] - E[Y \mid X = 0]$ where $E[Y \mid X = x]$ is the expected value of of Y for the treatment group (which received treatment level $X = x$) in an ideal randomized controlled experiment and $E[Y \mid X = 0]$ is the expected value of Y for the control group (which receives treatment level $X = 0$).

Stock and Watson, p. 170, claim H:

> The first of the three least squares assumptions is that the conditional distribution of u_i given X_i has a mean of zero. This assumption is a formal mathematical statement about the "other factors" contained in u_i and asserts that these other factors are unrelated to X_i in the sense that, given a value of X_i, the mean of the distribution of these other factors is zero.

Knowns and unknowns

Wooldridge p. 60, Claim I:

> lorem ipsum

University of Oxford Econometrics lecture slides, Michaelmas Term 2017, claim J:

> The simple regression model
> $y_i = \beta_1 + \beta_2 + u_i$
>
> - y_i and x_i are observable random scalars
> - u_i is the unobservable random disturbance or error
> - β_1 and β_2 are the parameters (constants) we would like to estimate

---

1. For the curious, or those who have to much time on their hands, I include a full version history, showing how this document evolved over the past few weeks. It's an interesting window into my thought process. ↩

2. As a separate gripe from the main one in this post, I note that often what I call $e_i + w_i$ is just written as w_i, by this I mean that in the same document, people will write $Y_i = \beta_0 + \beta_1 X_i + w_i$ *and* $Y_i = E[Y_i \mid X_i] + w_i$. This is either a terrible choice of notation (same name for two different objects) or an implicit and unnecessary (in this case) assumption that $e_i = 0$ and $u_i = w_i$. ↩