Econometrics
Deriving the OLS estimator in the univariate and bivariate cases

Nathaniel Higgins
nhiggins@jhu.edu

Primer

Notation

Here are some bits of notation that you will need to understand the derivations below. They are standard in this class, Wooldridge, and (for the most part) econometrics generally.

Referring to variables
When we write x we usually mean to refer to a variable *in general*. When we write x_i we mean the i^{th} observation of the variable x. You can think of x_1 as the first observation of x, e.g.

Summation notation
The equation

$$\sum_{i=1}^{n} x_i$$

is a common example of summation notation. This is just a compact way of writing

$$x_1 + x_2 + \ldots + x_n.$$

When we write

$$\bar{x}$$

we mean the average of x.

$$\bar{x} = \frac{1}{N} \sum_{i} x_i,$$

where N is the number of observations in our sample.

Regression notation
The equation you will see the most as you learn econometrics is this one:

$$y = \beta_0 + \beta_1 x + u.$$

Variables
y: The *dependent* variable (the variable that we are trying to explain)
x: The independent variable (the variable that we think causes y or is associated with changes in y)
u: The unobservables in the model (the stuff that changes y that is not part of our data)

Operators
The one you need to know about for this document is the expectation operator. The expectation of a random variable is, well, what we *expect* to get when we observe a draw of that random variable. So if x is some variable, I *expect* to observe $E(x)$ when I draw an observation of x.

We have a special way of writing the expectation of a random variable: $E(x) = \mu_x$. Basically, μ_x is just a compact way of representing the fact that the expectation of a random variable is a fixed value. When we write μ_x we refer to the *true mean* of x. That is, the expectation of x, $E(x)$, is some fixed number μ_x. The *true mean* of x is not to be confused with the mean of x in any given sample of data. We denote the sample mean of x by $\bar{x}$.

We still need to use the expectation operator even though we have the specialized μ notation, since sometimes we will be taking the expectation of more than just x, e.g. $E(x * y)$ (the expectation of x times y).

For the purposes of trying to understand what the expectation operator does, it may help to think of the expectation as the "average." (The average usually refers to the average of a sample of data, however, whereas the expectation refers to a *true* parameter. But it doesn't hurt to substitute "average" for "expectation" when you are trying to grasp the concept.)

1 Deriving the OLS estimator

1.1 Method of moments

This is the way that Wooldridge does it. I'll do it the same way here, but I'll fill in some of the blanks that Wooldridge leaves for the appendix. The method of moments is an appealing way to derive estimators. It doesn't use a shred of calculus! (A little bit of algebra is absolutely unavoidable)

The method of moments is based on the idea that for every constraint that we impose on the data, we can identify one parameter.

We make an assumption about our model. Then we use this assumption as a rule. Instead of just *assuming* it is true, we *make* it true in our data. That is, we make the

data that we have (our sample) obey this rule. Each rule that we impose gives us the ability to say what one parameter *must be* for the rule to hold.

1.1.1 Univariate model

Start with the simplest possible model: what I call the "univariate" regression model,[1] or a model with no x-variable. A model where y is always equal to a constant:

$$y = \beta_0 + u.$$

There is one unknown parameter in this model: β_0. Therefore we need one assumption — one rule — to identify β_0. We will impose this rule on the data, and this will give us the formula for our best estimate of β_0, which we will denote $\widehat{\beta_0}$. The standard assumption we will use is $E(u) = 0$, the assumption that the model is correct, *on average*. (Not to be confused with the unreasonable assumption that the model is always correct)

The method of moments works by imposing this assumption on the data. So we start with

$$E(u) = 0,$$

and we make this true in the sample of data that we have. Note that

$$u = y - \beta_0,$$

so that

$$E(u) = E(y - \beta_0). \tag{1}$$

What does it mean to "make the assumption true in our data?" It means taking the concept of "expectation," which is a concept about the *universe* of data, and applying the analogous concept to our *sample* of data. What is the sample analog of the expectation? The mean. The average. The sample analog of the expected value of something is the mean of that same thing. So if the theoretical expected value of a variable x is μ_x, then the sample analog of μ_x is $\bar{x}$. Making $E(u) = 0$ true in our data translates to *forcing* the average of u in our data to equal zero.

Applying this notion to equation (1), we get

$$\begin{aligned} E(u) = 0 &= E(y - \beta_0) \\ &= E(y) - E(\beta_0) \\ &= \frac{1}{N}\sum_i y_i - \frac{1}{N}\sum_i \beta_0 \qquad (2) \\ &= \bar{y} - \frac{1}{N} N \beta_0 \\ &= \bar{y} - \beta_0. \qquad (3) \end{aligned}$$

[1] I don't think anybody else in the world calls it this, so don't go throwing the term around in intelligent econometric conversation unless you want people to look at you funny.

We now have a single, simple equation with a single unknown. We choose our estimate of β_0, which we call $\widehat{\beta_0}$, to be the estimate that solves equation (3).

$$0 = \bar{y} - \widehat{\beta_0} \tag{4}$$

$$\widehat{\beta_0} = \bar{y}. \tag{5}$$

The best estimate of a variable y that we can manage when we model y as a constant is $\widehat{\beta_0} = \bar{y}$, the mean of y.

1.1.2 Bivariate model

What I call the bivariate model is a model relating the dependent variable y to an independent (or *explanatory*) variable x. The model looks like this:

$$y = \beta_0 + \beta_1 x + u.$$

We can use the method of moments to derive estimates of $\widehat{\beta_0}$ and $\widehat{\beta_1}$, just as we did in the univariate model above. The only difference is that now we have two unknown parameters that we would like to estimate, instead of one. This means that we need two assumptions (two rules; two constraints) that we can place on the data in order to identify the two unknown parameters.

Just as before, the assumption $E(u) = 0$ will serve as one of the constraints we place on the data. The other that Wooldridge uses is the assumption that $Cov(x, u) = 0$.[2] This condition follows from the assumption that x and u are independent. If two random variables are independent, then they must also have a covariance of zero. We will impose these two conditions on the data.

$$\begin{aligned} E(u) &= 0 \\ &= E(y - \beta_0 - \beta_1 x) \\ &= E(y) - E(\beta_0) - E(\beta_1 x) \\ &= \frac{1}{N}\sum_i y_i - \frac{1}{N}\sum_i \beta_0 - \frac{1}{N}\sum_i \beta_1 x_i \\ &= \bar{y} - \beta_0 - \beta_1 \bar{x} \\ \beta_0 &= \bar{y} - \beta_1 \bar{x}. \end{aligned} \tag{6}$$

[2]Why is the assumption $Cov(x, u) = 0$ a good assumption? Why is it something that we would like to be true? Think of it this way. We want to model the relationship between x and y. If we were running an experiment where we had complete control over x, we could measure the relationship between x and y perfectly. We manipulate x, then simply watch how much y moves. All the movement in y that we observe is caused by movement in x. Now consider an alternative reality. Suppose every time you manipulated x, u moved too. So x is moving, u is moving, and of course, y is moving. We can no longer attribute observed movements in y to x alone — we don't know if it's x or u (or a combination of the two) that is causing y to move. What's the moral of the story? It would be really nice if when x moves, we know that other stuff isn't moving with it.

Once again we have identified the estimator that we will use to estimate β_0.

$$\widehat{\beta_0} = \bar{y} - \beta_1 \bar{x} \tag{7}$$

Only this time the unknown parameter β_1 is part of the equation. We will get the formula for $\widehat{\beta_1}$ using the second condition ($Cov(x, u) = 0$), then plug this estimate back into (7).

The second condition is $Cov(x, u) = 0$. This condition follows from the assumption that x and u are statistically independent. If they are independent, then their covariance is zero. So we impose zero covariance on the data in our sample. Here is the definition of covariance between x and u (i.e. what follows is simply an expansion of the definition of covariance):

$$\begin{aligned} Cov(x, u) &= E[(x - E(x))(u - E(u))] \\ &= E[(x - \mu_x)(u - \mu_u)] \\ &= E[x * u - x * \mu_u - \mu_x * u + \mu_x * \mu_u] \\ &= E(x * u) - E(x * \mu_u) - E(\mu_x * u) + E(\mu_x * \mu_u). \end{aligned} \tag{8}$$

The first line of (8) expresses the definition of covariance between two variables (see p. 729 - 730 of Wooldridge). The second line simply substitutes μ_x for $E(x)$ and μ_u for $E(u)$. Recall that $E(x)$ is a theoretical construct — the expected value of the random variable x is equal to some fixed value, which we call μ_x. Likewise with u. The third line expands the expression using the FOIL method. The fourth line uses the fact that the expectation operator (which you could think of roughly as the "average operator") is linear.[3]

Let's pick up where we left off, with the final line of (8). Now we will apply the population analog of the expectation operator, just like we did in (6) above.

$$\begin{aligned} E(x * u) - E(x * \mu_u) - E(\mu_x * u) + E(\mu_x * \mu_u) & \\ &= E(x * u) - \mu_u * E(x) - \mu_x * E(u) + \mu_x * \mu_u \\ &= E(x * u) - \mu_u * \mu_x - \mu_x * \mu_u + \mu_x * \mu_u \\ &= E(x * u) - \mu_x * \mu_u \\ &= E(x * u) \end{aligned} \tag{9}$$

The second line of (9) uses the fact that μ_u and μ_x are constants. The expectation of a constant times a variable is the same as the constant times the expectation of the variable (if that seems like math-ey jargon to you, just read the previous sentence again, substituting the word "average" for "expectation"). Having applied this rule to obtain line 2, we calculate $E(x)$ and $E(u)$ to obtain line 3. At this point, it is straightforward to collect terms and see that the fourth line of (9) is the simplest way to express $Cov(x, u)$.

[3]This is just like saying that the average of $q + r$, $\overline{q + r}$, is equal to the sum of the averages, $\bar{q} + \bar{r}$. Try it with a few numbers if you need to convince yourself. Notice! The expectation operator is linear, so you can expand it over the four terms in (8) above, but it is *not* true that $E(q * r) = E(q) * E(r)$. Just a note to be sure you don't misunderstand

Finally, since the expectation of u is zero by assumption (that was our first condition we used in (7), remember?), the whole thing simplifies to just $E(x * u)$.

Now we take this simple expression, and substitute in our regression equation, which includes the two unknown parameters β_0 and β_1. We do this in order to use the condition $E(x * u) = 0$ to identify the two parameters.

$$\begin{aligned} E(x * u) &= E(x * (y - \beta_0 - \beta_1 x)) \\ &= E(x * y - \beta_0 * x - \beta_1 * x^2) \\ &= E(x * y) - E(\beta_0 * x) - E(\beta_1 * x^2) \\ &= \frac{1}{N}\sum_i x_i * y_i - \beta_0 \frac{1}{N}\sum_i x_i - \beta_1 \frac{1}{N}\sum_i x_i^2 \end{aligned}$$

Using only the rules we have already discussed above, we now have an expression for the sample value of $E(x * u)$. We now set this equal to zero (recall that the condition we are imposing on the data is $Cov(x, u) = 0$) and solve for β_1.

$$\frac{1}{N}\sum_i x_i * y_i - \beta_0 \frac{1}{N}\sum_i x_i - \beta_1 \frac{1}{N}\sum_i x_i^2 = 0 \tag{10}$$

But notice that the unknown term β_0 is part of the expression in (10). Before we solve for β_1, then, we need to substitute in our estimator for β_0 (the expression we got in (7) above). That is, when we see β_0 in (10), we will substitute in the identity $\beta_0 = \bar{y} - \beta_1 \bar{x}$.

$$\begin{aligned} 0 &= \frac{1}{N}\sum_i x_i * y_i - \beta_0 \frac{1}{N}\sum_i x_i - \beta_1 \frac{1}{N}\sum_i x_i^2 \\ &= \frac{1}{N}\sum_i x_i * y_i - (\bar{y} - \beta_1 \bar{x}) \frac{1}{N}\sum_i x_i - \beta_1 \frac{1}{N}\sum_i x_i^2 \end{aligned} \tag{11}$$

Now we distribute terms by multiplying $(\bar{y} - \beta_1 \bar{x})$ and $\frac{1}{N}\sum_i x_i$.

$$\begin{aligned} 0 &= \frac{1}{N}\sum_i x_i * y_i - (\bar{y} - \beta_1 \bar{x}) \frac{1}{N}\sum_i x_i - \beta_1 \frac{1}{N}\sum_i x_i^2 \\ &= \frac{1}{N}\sum_i x_i * y_i - \bar{y} \frac{1}{N}\sum_i x_i + \beta_1 \bar{x} \frac{1}{N}\sum_i x_i - \beta_1 \frac{1}{N}\sum_i x_i^2 \end{aligned}$$

Finally, we do two thing: (1) we divide everything in the equation by $1/N$ to eliminate it from the expression (since $1/N$ currently multiplies everything in the equation, it has no effect, so we can get rid of it); (2) factor out β_1 so that we can solve the equation for β_1.

$$\begin{aligned} 0 &= \sum_i x_i * y_i - \bar{y}\sum_i x_i + \beta_1 \bar{x}\sum_i x_i - \beta_1 \sum_i x_i^2 \\ &= \sum_i x_i * y_i - \bar{y}\sum_i x_i + \beta_1 (\bar{x}\sum_i x_i - \sum_i x_i^2) \\ \beta_1 &= \frac{\sum_i x_i * y_i - \bar{y}\sum_i x_i}{\sum_i x_i^2 - \bar{x}\sum_i x_i} \end{aligned} \tag{12}$$

We are now "done." This expression for β_1 gives us our method-of-moments estimator of β_1, which we call $\widehat{\beta_1}$.

$$\widehat{\beta_1} = \frac{\sum_i x_i * y_i - \bar{y}\sum_i x_i}{\sum_i x_i^2 - \bar{x}\sum_i x_i} \tag{13}$$

You will notice, however, that this expression is different than expression 2.19 on p. 29 of Wooldridge. Both expressions are fully correct. Wooldridge has simply re-arranged the terms above in (13) to make them more ... *interpretable*. By taking what you see in (13) and making it look like expression 2.19 on page 29, you can see more clearly exactly what $\widehat{\beta_1}$ is in terms of the data.

For your own edification, I'll show you below how (13) and expression 2.19 are completely identical. I'll do this in two parts. First I'll show that the numerator of (13) is equal to the numerator in expression 2.19, then the denominator.

I'll work backwards from Wooldridge's expression of the numerator.

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

All I have done is use the fact that $\bar{x} = 1/N \sum x_i$, which implies that $\sum x_i = N\bar{x}$ (and likewise for y). Now the denominator.

$$\begin{aligned}
\sum_i (x_i - \bar{x})^2 &= \sum_i x_i^2 - \sum_i x_i\bar{x} - \sum_i \bar{x}x_i + \sum_i \bar{x}^2 \\
&= \sum_i x_i^2 - 2\bar{x}\sum_i x_i + N\bar{x}^2 \\
&= \sum_i x_i^2 - 2\bar{x} * N\bar{x} + N\bar{x}^2 \\
&= \sum_i x_i^2 - N\bar{x}^2 \\
&= \sum_i x_i^2 - \bar{x}\sum_i x_i
\end{aligned}$$

As you can see, the Wooldridge expression and the expression that we derived, working step-by-step, are totally identical. Wooldridge's expression contains some intuitive information: our estimate of the slope parameter β_1 is equal to the sample covariance between x and y, divided by the sample variance of x. Said another way, our estimate of the relationship between x and y is equal to the covariance between x and y, normalized by the variance of x. What does it mean to "normalize," and why does this make sense?

Why does it help us to think this way? The covariance of any two variables (here x and y) tells us something about what happens to one variable when the other moves. But covariance is a term without scale. That is, there are no sensible "units" of covariance. This is because the covariance of two variables depends on the units in which these two variables were measured. In the case of $\widehat{\beta}_1$, we can interpret the estimate as the amount of covariance between x and y, *measured in* units of variation in x.

1.2 Method of least-squares

Now we will derive the exact same estimators, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, using the least-squares procedure. You should see this at least once. After all, the least-squares procedure is how the OLS estimator got its name!

We have some data, x and y. We have a model

$$y = \beta_0 + \beta_1 x + u.$$

Our goal is to choose estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that "fit" the data well. There are many possible ways to measure "fit," but one that is particularly appealing (and it is the one that is most widely used in practice) is to minimize a measure of how wrong our predictions of y are. Think of it this way. If the true model is

$$y = \beta_0 + \beta_1 x + u,$$

and we pick some estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, then we can use these estimates to *predict* y:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

But these predictions of y that we call $\widehat{y}$ are not always going to be right. If our model is good, they should be close, but they won't be exactly correct. They will be off by some amount $\widehat{y} - y$. This amount by which our predictions will be off is a measure of the unobserved determinants of y. That is, it is a measure of the *unexplained* part of y — the variation in y that our model $\widehat{\beta}_0 + \widehat{\beta}_1 x$ does not explain. So we call it $\widehat{u}$. We want to choose the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that make the amount by which our predictions are "off," $\widehat{u}$, small.

We could make $\widehat{u}$ small my minimizing $\widehat{u}$. That is, we could choose $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize

$$\widehat{u} = y - \widehat{\beta}_0 - \widehat{\beta}_1 x.$$

But think of what this will do. If we want to make something really *small*, all we have to do is make it really negative. To make the expression for $\widehat{u}$ really negative, all we need to do is under-predict y all the time. This isn't very informative, and it isn't in keeping with our goal to find estimators that "fit" the data well. So we shouldn't try to minimize $\widehat{u}$. We want to minimize a measure that weights mis-predictions *above* y and mis-predictions *below* y equally. There are two obvious candidate measures. We could minimize the absolute value of $\widehat{u}$, or we could minimize the square of $\widehat{u}$. Both are legitimate. As the name "least-squares" suggests, the standard estimator of $\widehat{\beta}_0$ and $\widehat{\beta}_1$

is the estimator that minimizes $\widehat{u}^2$. There end up being some nice properties that result from using the square of $\widehat{u}$ rather than the absolute value of $\widehat{u}$. We will explore these in the future. But for now, if you need a justification — an idea of what using $\widehat{u}^2$ instead of $|\widehat{u}|$ gets us — realize that minimizing the square of $\widehat{y} - y$ penalizes large errors more. That is, when our predictions $\widehat{y}$ are close to y, the squared difference $\widehat{u}^2$ is small. When our predictions $\widehat{y}$ are far from y, the squared difference is much larger. Said another way, the measure $\widehat{u}^2$ weights prediction errors more the bigger they are. So choosing our estimates of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ to minimize $\widehat{u}^2$ should result in estimates that minimize really big prediction errors. Not a bad deal.

Now for the derivation of $\widehat{\beta_0}$ and $\widehat{\beta_1}$. We want to choose $\widehat{\beta_0}$ and $\widehat{\beta_1}$ to minimize the sum of the squared differences between $\widehat{y}$ and y

$$\sum_i \widehat{u}^2 = \sum_i (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2. \tag{14}$$

To do this, we need to employ some calculus (don't worry, it's not hard-core calculus, just the regular (soft-core?) variety). We take the derivative of (14) with respect to $\widehat{\beta_0}$ and set the expression equal to zero. We want to find the $\widehat{\beta_0}$ where changing $\widehat{\beta_0}$ a little bit doesn't change the value of $\widehat{u}^2$ (if the derivative of $\widehat{u}^2$ were negative instead of zero, this would mean that making $\widehat{\beta_0}$ a little bigger would *decrease* $\widehat{u}^2$, telling us that our choice of $\widehat{\beta_0}$ was too small; likewise, if the derivative of $\widehat{u}^2$ were positive, this would mean that making $\widehat{\beta_0}$ a little smaller would decrease $\widehat{u}^2$, telling us that our choice of $\widehat{\beta_0}$ was too big). We then take the derivative of (14) with respect to $\widehat{\beta_1}$ and set the expression equal to zero (for the same reason given above for $\widehat{\beta_0}$).

$$\begin{aligned}
\frac{d}{d\widehat{\beta_0}} \sum_i \widehat{u_i}^2 &= \frac{d}{d\widehat{\beta_0}} \sum_i (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 \\
&= \frac{d}{d\widehat{\beta_0}} \sum_i (y_i^2 - 2\widehat{\beta_0} y_i + \widehat{\beta_0}^2 - 2\widehat{\beta_1} x_i y_i + 2\widehat{\beta_0}\widehat{\beta_1} x_i + \widehat{\beta_1}^2 x_i^2) \\
&= \sum_i (-2y_i + 2\widehat{\beta_0} + 2\widehat{\beta_1} x_i)
\end{aligned}$$

Because everything in the expression is multiplied by 2, we can divide the whole expression by 2 and eliminate that term. We then set the expression equal to zero and solve

for $\widehat{\beta}_0$.

$$\begin{aligned}
\sum_i(-2y_i + 2\widehat{\beta}_0 + 2\widehat{\beta}_1 x_i) &= 0 \\
&= \sum_i(-y_i + \widehat{\beta}_0 + \widehat{\beta}_1 x_i) \\
&= -\sum_i y_i + \sum_i \widehat{\beta}_0 + \sum_i \widehat{\beta}_1 x_i \\
&= -\sum_i y_i + N\widehat{\beta}_0 + \sum_i \widehat{\beta}_1 x_i \\
\widehat{\beta}_0 &= \frac{1}{N}\sum_i y_i - \frac{1}{N}\widehat{\beta}_1 \sum_i x_i \\
\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x}
\end{aligned}$$

Look familiar? This is exactly the expression we got for $\widehat{\beta}_0$ when we used the method of moments. We now only need to take the derivative of (14) with respect to $\widehat{\beta}_1$, set it equal to zero, and solve for $\widehat{\beta}_1$.

I will skip the explanation of some of the steps that are identical to the case above for $\widehat{\beta}_0$.

$$\begin{aligned}
\frac{d}{d\widehat{\beta}_1}\sum_i \widehat{u}_i{}^2 &= 0 \\
&= \frac{d}{d\widehat{\beta}_1}\sum_i(y_i^2 - 2\widehat{\beta}_0 y_i + \widehat{\beta}_0{}^2 - 2\widehat{\beta}_1 x_i y_i + 2\widehat{\beta}_0\widehat{\beta}_1 x_i + \widehat{\beta}_1{}^2 x_i^2) \\
&= \sum_i(-2x_i y_i + 2\widehat{\beta}_0 x_i + 2\widehat{\beta}_1 x_i^2) \\
&= \sum_i(-x_i y_i + \widehat{\beta}_0 x_i + \widehat{\beta}_1 x_i^2) \\
&= -\sum_i x_i y_i + \widehat{\beta}_0 \sum_i x_i + \widehat{\beta}_1 \sum_i x_i^2
\end{aligned}$$

Substitute in $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$.

$$\begin{aligned}
0 &= -\sum_i x_i y_i + \widehat{\beta}_0 \sum_i x_i + \widehat{\beta}_1 \sum_i x_i^2 \\
&= -\sum_i x_i y_i + (\bar{y} - \widehat{\beta}_1 \bar{x})\sum_i x_i + \widehat{\beta}_1 \sum_i x_i^2 \\
&= -\sum_i x_i y_i + \bar{y}\sum_i x_i - \widehat{\beta}_1 \bar{x}\sum_i x_i + \widehat{\beta}_1 \sum_i x_i^2
\end{aligned} \tag{15}$$

Notice what we have here. The last line of (15) is *exactly* the first line of (12). When we use the logic of least-squares, we quickly end up with the same estimator that we

obtained by using the logic of the method of moments. The rest of the math needed to derive the expressions for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ is literally exactly the same as the math used above in the method-of-moments section.