1 The Key Idea

Econometrics is about exploring relationships between what we call 'variables'. These are numbers that describe features of an 'individual' – usually a particular person, such as yourself. Variables that relate to you include your age, height, age, income, number of siblings, and years of education. We collect such numbers from a 'sample' of individuals, arrange them nicely into 'datasets', and work with these directly to figure out how the different variables are connected.

We will introduce our main technique for doing this analysis by describing a fictitious situation where it might be applied. The more vividly you can imagine the thought experiment, the more you will get out of it.

2 Wages and Experience

Imagine that you begin working at a very large company. When you join up, the bosses of the company get together in a room (pictured below), where you are briefly interviewed. They then decide on what your annual wage will be.

After meeting you, they determine your wages according to a certain formula, depending on two factors: how much experience you have, and how much the bosses happen to like you based on the interview. Imagine, for example, the formula they use is as follows:

$$Wage = 25,000 + 1,000 \times Experience + U$$

The final component 'U' we will call the 'error term' – which in this example represents how much the bosses like us. This will in turn be based on a complex and unpredictable combination of our other qualities, many of which will be difficult to measure. In general, we think of the error term as representing other factors affecting the left hand side variable that do not appear elsewhere in the formula. As we will see, the error term is not to be ignored and is very important in econometrics.

We assume for the moment that U is completely unrelated to *Experience*, and is zero on average. Given this assumption, then if – for example – we have 10 years of experience, our expected wages will be

$$25,000 + 1,000 \times \mathbf{10} = £35,000.$$

However, if we have 11 years of experience, our expected wages will instead be

$$25,000 + 1,000 \times \mathbf{11} = £36,000.$$

Notice that the increase is exactly $£1,000$. This illustrates an important point: the interpretation of the number $1,000$ in the formula is how much our wage increases with an extra year's worth of experience (that is, ignoring any possible changes in the error term U). This amount is always the same regardless of how much experience you have to start out with. You should ensure you fully understand why this is so from the above calculations; i.e., because the number of years of experience is always multiplied by $1,000$. Meanwhile, the number $25,000$ is our starting or 'base' wage; that is, what we can expect to receive on average if we have yet to acquire any experience – though we may actually receive more or less than $25,000$ depending on U; that is, how popular we are with the bosses.

We emphasise again that these two numbers which the bosses use in their calculations, $1,000$ and $25,000$, never change and are the same for each individual worker. Something similar will be the case in all of our other investigations too.

3 Using Graphs, and Some Useful Notation

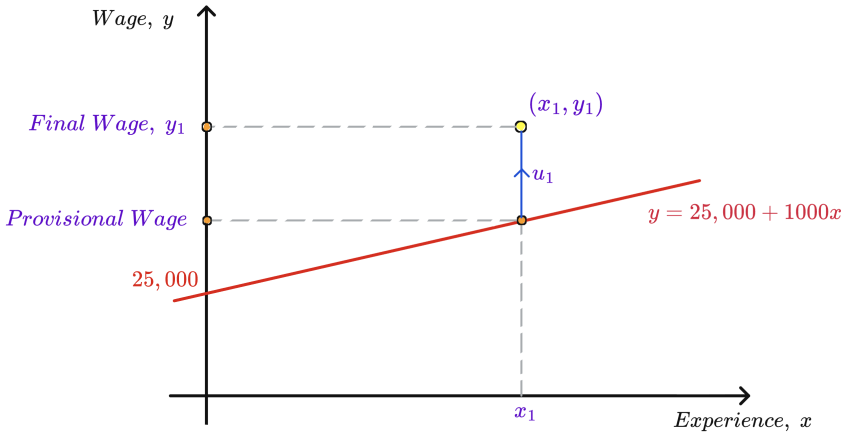
We can illustrate the above setup visually, by drawing a graph with $Wage$ on the vertical axis, and *Experience* on the horizontal axis. We will also introduce some notation to make things easier going forward. However, before proceeding any further with either, we will need to make an important distinction.

When we are thinking of wages and experience as unknown quantities relating to a hypothetical worker who has not yet been selected, we will write $Wage$ as Y and *Experience* as X. That is, these are not the wages of any *particular* person, but are indeterminate. We say that Y and X are 'random' at this stage, prior to having picked an individual to observe. More specifically, we will call them 'population random variables', since they refer to a hypothetical individual drawn at random from the population.

However, when we are talking about wages and experience as *fixed* numbers that we know and can draw in a particular place on a graph, we will instead use the lowercase letters y and x.

Following this convention, in the graph below we use the line $y = 25,000 + 1000x$ to show the general relationship between wages and experience, where y and x represent generic but fixed values. This line is shown in red. Then, depending on whether the bosses like them or not, points on the line representing a particular individual's provisional wage are randomly *pushed away* from this line in the vertical direction, either up or down. This gives their final wage.

This 'pushing off' is shown also below by a blue arrow, for a particular, lucky person. Of course, since x and y are already the axis labels, we now need a *third* type of notation to indicate their experience and final wage! I have chosen to label this specific worker's individual levels of experience as x_1, and their wages as y_1, rendered in purple. The label '1' indicates this is the first specific individual we have encountered so far. This labelling system will be very helpful later when we will consider investigating multiple individuals' wages and experience at the same time. Naturally, we use the lower-case letter u_1 for the error term for this specific person, represented as a fixed, determinate distance.



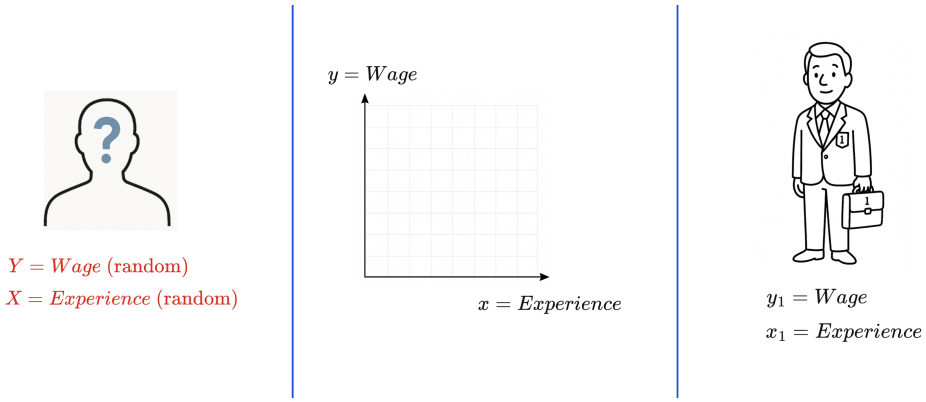
The conceptual distinctions raised here may seem subtle, but these are actually very important differences, as we'll see later on. And in general, notation is very important in econometrics!

To reiterate, we have the following notation and associated distinctions:

- Y, X for the indeterminate (or 'random') wages and experience of a hypothetical worker we haven't yet chosen
- y, x for generic but fixed levels of wages and experience, used as labels for the graph axes
- y_1, x_1 for the wages of one specific individual, representing one particular point on the graph

These distinctions may seem tricky at first, but you will get used to them – and come to appreciate the importance of making them. We illustrate them visually

in the panel below.



4 The Challenge Ahead

In the previous section, we discussed an exact formula for how our wages are determined:

$$Wage = 25,000 + 1,000 \times Experience + U$$

Or, in our more mathematical notation,

$$Y = 25,000 + 1,000X + U$$

Now, the key type of problem we deal with in econometrics is encountering the following situation. Imagine that we know *what sort of formula* is used, but – unlike in this example – we don't know *what the actual numbers in the formula are.*

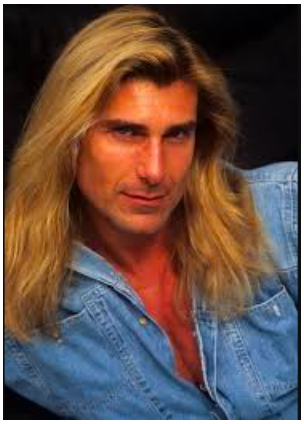
Let's now forget the specific numbers $25,000$ and $1,000$, then, and imagine instead that the real numbers are unknown because the bosses keep them a closely-guarded secret. Nevertheless, for a variety of reasons we may be greatly interested in finding them out.

5 Further Useful Notation

Introducing the right notation can take us a long way. A helpful step at this point is to introduce the labelled greek letters β_0 and β_1 to represent the secret numbers in the formula – which we don't know, but would like to learn about. We can then write the formula like this:

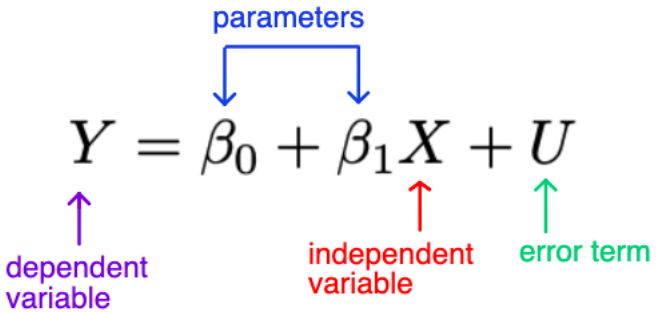
$$Wage = \beta_0 + \beta_1 Experience + U$$

As above, β_0 represents the average wage for someone with no experience, whereas β_1 tells us the impact of studying for a further year. In econometrics we call a formula such as this a 'model' – in this case, a 'structural model', which emphasises how the variables are related. This is quite different to what normal people use the term 'model' to mean, which is perhaps something more like this:



You should try your best not to get the two types of model confused.

As with human models, it is also important not to get mixed up between the different parts of these mathematical models. Doing so causes students a great many problems. Abstracting from the present context slightly, in general we can write our sort of model as follows:



The 'parameters' (or 'coefficients') β_0 and β_1 in this structural model are fixed numbers, which we don't know (although the bosses do). Again, these are always the same for everyone and never change. Y is our 'dependent variable' – in our main example, $Wage$, which is determined for each worker in a manner that *depends on* the other things in the model. The X is our 'independent variable', in our case $Experience$, which is down to that individual and their personal history. The mysterious variable U is the 'error term', which we continue to assume is independent from $Experience$ and zero on average.

The fact that we never actually know the value of this error term infuriates econometricians no end, and will continue to bother us throughout the book as we go about our primary task. This task is so important that it deserves its own special box:

Econometrics is all about guessing what β_0 and β_1 are!

In an effort to make us sound more professional, I will try to say 'estimating' rather than 'guessing' from now on. How might this estimation proceed?

6 Working with Data

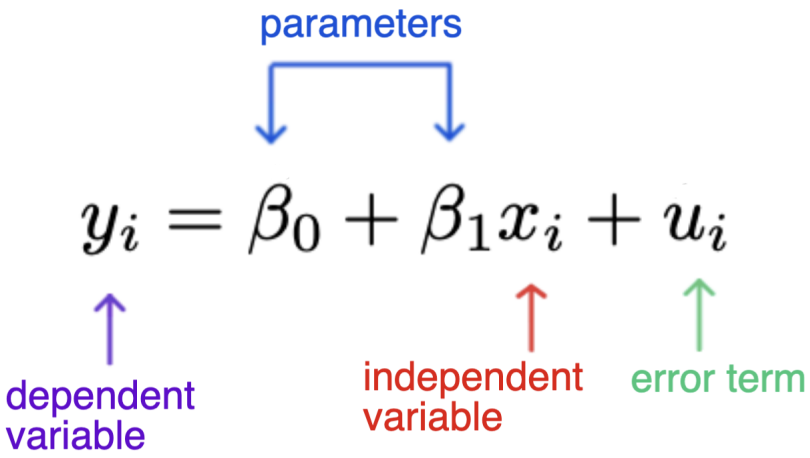
A natural idea for estimating β_0 and β_1 is to find some coworkers (pictured below) and ask them what their levels of wage and experience are.

We organise this information into three columns. The first column, i, is a label that enables us to keep track of each individual we ask. In this case, since we have asked 30 individuals, it runs from 1 to 30. The other two give $Wage$ and $Experience$ for everyone in the sample. We add an 'i' to label these according to the individual, as with y_1 and x_1 above. So, each *column* represents a *variable* – including the variable i used just for labelling. Meanwhile, each *row* represents a particular *individual*, giving the values of all variables for them – except for u_i, which we cannot include because it is unknown. Take a minute to glance at the table to familiarise yourself with how it is structured.

i	$y_i =$ $Wage_i$	$x_i =$ $Experience_i$
1	29617.29	12.8
2	38723.32	23.3
3	31687.21	14.5
4	43884.69	30.9
5	22818.9	2.43
6	41765.89	27.0
7	30418.98	13.8
8	34108.48	17.1
9	35470.93	19.8
10	46811.88	31.5
11	32185.52	12.4
12	33775.32	18.3
13	44192.16	27.0
14	41005.11	25.8
15	38722.23	22.8
16	36103.19	21.6
17	23538.19	3.14
18	45072.11	31.6
19	28904.37	10
20	49718.41	38.9
21	33352.14	18.1
22	27241.85	8.42
23	28556.63	11.0
24	38051.21	21.9
25	30202.33	11.5
26	49462.76	38.0
27	19896.85	.08
28	39588.68	24.9
29	25484.49	7.14
30	22784.26	4.17

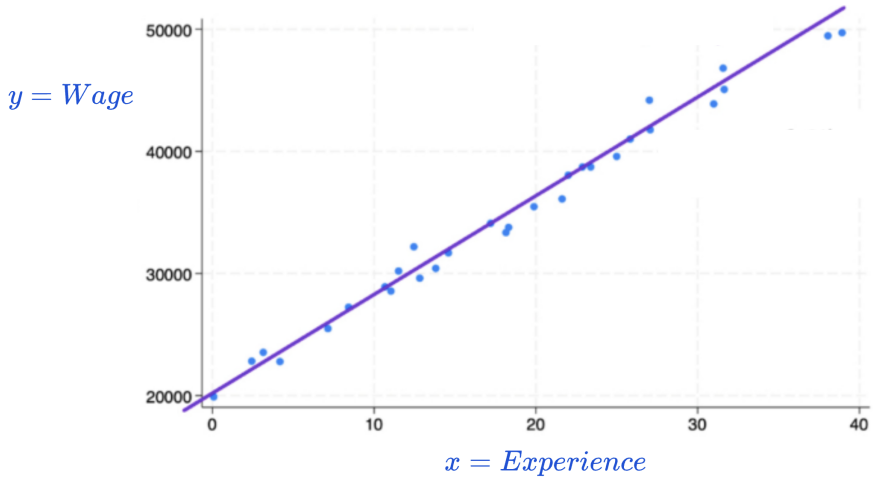
Notice again that now we're working with *concrete* data, based on this particular group of coworkers we've surveyed, I have used *lowercase letters* for the values of the variables they have reported. In contrast, previously we used *capital letters* when talking about *indeterminate* wages and experience levels, before we'd collected any data. As above, in this situation, prior to identifying which particular workers we will ask, we will say these indeterminate values are 'random'. Again, this tricky distinction is very important.

We assume that the numerical values of our variables are also related according to the 'structural model' equation given in the last section:



This statement looks similar to the structural model itself; but as well as now being written in lowercase to indicate concrete values from the sample, the variables are also labelled for the individual i. In this case, the 'index' or labelling variable i runs from $i = 1$ to $i = 30$. So, we actually have 30 separate equations here! If we had asked n individuals instead of 30, then the index i would of course have to run from 1 to n, the size of our sample.

As a next step from here, we could produce a scattergraph of the data, showing both x_i and y_i for each individual. We might then think of drawing on a 'line of best fit', an idea which will probably be familiar from school. Both the data and the resulting line of best fit are shown below:



Now comes the **most important idea in the whole of econometrics**.

Consider the straight line $y = \beta_0 + \beta_1 x$. This line represents our structural model $Y = \beta_0 + \beta_1 X + U$ visually on a graph. Alternatively, it shows provisional wages before the error term is applied. Let's call it the 'structural model line'. It was illustrated graphically above.

Now, hopefully, the line of best fit we just drew above is actually pretty close to this true 'structural model' line. But this means that **we can use it to guess what the original β_0 and β_1 are!**

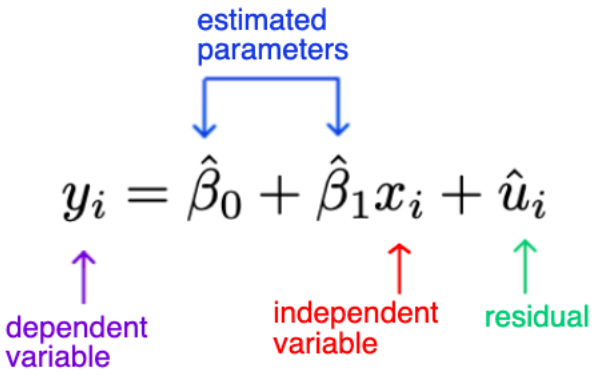
In more detail: we think of the y-intercept of *our* line of best fit (i.e. the point when $x = Experience = 0$) as our estimate of the base wage, β_0, which is the intercept of the *true* structural line, representing the formula the bosses actually use. And likewise, we think of the slope or gradient of *our* line as our estimate of gradient of the true line, β_1 – that is, the increase in wage when experience goes up by one year. These estimates will be accurate insofar as the two lines are indeed close together.

For the intercept and slope of our line of best fit, we will use the notation $\hat{\beta}_0$ and $\hat{\beta}_1$, with 'hats' on. The hats make it clear that these are just *estimates*: not the actual 'secret' numbers the bosses are using, β_0 and β_1, but merely our best *guesses* of them based on the data. This is easy to forget, but doing so makes much of econometric theory incomprehensible. Our line of best fit can therefore be written as $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

In the above graph, the intercept is about $20,400$, and the gradient is about 780. Our line of best fit therefore gives us the estimates $\hat{\beta}_0 = 20,400$ and $\hat{\beta}_1 = 780$. I am assuming familiarity with straight line graphs here; look this topic up online if you're not confident with it or have forgotten it – especially how to draw straight line graphs, and how to find the gradient and intercept. Check you can see how I obtained these numbers.

Returning to the data, the numerical values of our variables for our 30 indi-

viduals are also related like this:



Where i varies from 1 to 30. So again, this is really a *series* of 30 different equations; one for each individual.

Let's consider the final term of the estimated model, the 'residual' $\hat{u}_i$, which has taken the place of the error term u_i. Notice that now it has a hat on. Why is this?

Recall that the *real* values β_0 and β_1 from the *true* equation the bosses use are different to our *guesses* $\hat{\beta}_0$ and $\hat{\beta}_1$ from our *estimated* equation, as represented by the line of best fit. So, although y_i and x_i are the same in both, the relationships between them given by the structural model ($y_i = \beta_0 + \beta_1 x_i + u_i$) and by the estimated model ($y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$) are different.

In particular, rearranging the structural relationship, we see that we have the following expression for the error terms:

$$u_i = y_i - \beta_0 - \beta_1 x_i$$

for each person i in our sample. However, the residual is given by rearranging the estimated relationship:

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

That is, the 'real' *error term* u_i for individual i is the difference between their actual wage y_i and $(\beta_0 + \beta_1 x_i)$, whereas the *residual* $\hat{u}_i$ is instead the difference between their wage y_i and $(\hat{\beta}_0 + \hat{\beta}_1 x_i)$. The **residual** is therefore **different** to the **error term**.

As well as differing numerically, there is an important conceptual difference between u_i and $\hat{u}_i$ as well. That is, since we actually know the value of everything in the expression $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, we see that $\hat{u}_i$ is something we can

actually calculate! Indeed, we think of $\hat{u}_i$ as a bit like a *guess* of what u_i is, based on the sample. Hence the 'hat' on top, as with our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. This idea may be useful if we'd like to discover something about how much our bosses like us! In general, we often use the residuals $\hat{u}_i$ to try to learn things about the true error terms u_i, which sadly we can never observe directly.

7 Estimated Models

Using our estimates $\hat{\beta}_0 = 20,400$ and $\hat{\beta}_1 = 780$, we can also give the following general expression for the estimated relationship between $Wages$ and $Experience$ for any generic worker:

$$Wage = 20,400 + 780 \times Experience + \hat{U}$$

Here we define $\hat{U}$ as the difference $Wage - 20,400 + 780 \times Experience$ for *any* randomly-chosen individual, not necessarily in our original sample of 30 workers. We now don't need the label i, and use a capital letter as the value is indeterminate. I could have put 'Wage' and 'Experience' in all capitals too, to emphasise the same distinction, but I didn't want it to seem like I was shouting at you.

In our more mathematical notation, we can write this as:

$$Y = 20,400 + 780X + \hat{U}$$

For a hypothetical worker with wages Y and experience X. We call this an 'estimated model'. Here it represents our attempt to replicated the structural model, based on the 30 workers we asked.

Now, let's suppose that a friend is planning to join the company. We'd like to help them predict their wage, armed with our estimated model. Let's imagine further that they have 5 years of experience. Of course, we don't know what the value of their residual $\hat{U}$ will be. But since the real error U term is zero on average, it seems the best we can do is simply ignore it. We therefore estimate that their wage will be $20,400 + 780 \times 5 = £24,300$. We call this a 'fitted value' of Wage, and in general write

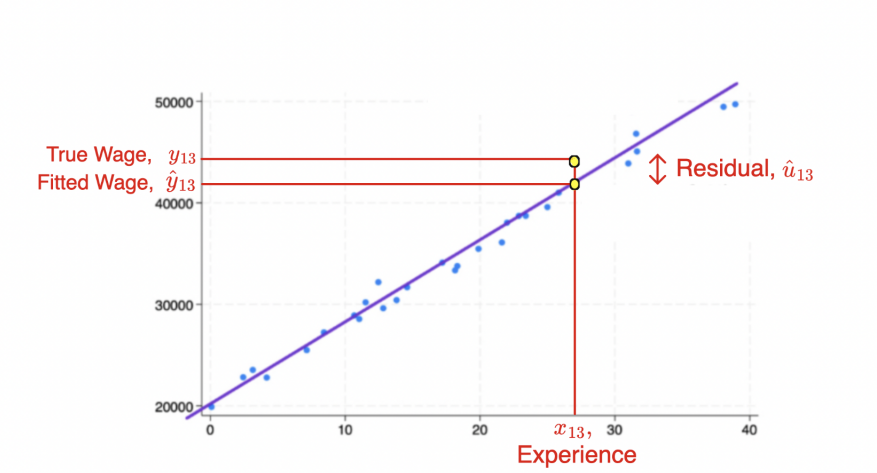
$$\widehat{Wage} = 20,400 + 780 \times Experience$$

This is all well and good as a guess, but we would be wise not to promise our friend a wage of *exactly* $£24,300$. For as well as just ignoring the residual, there is a second source of uncertainty here: we have **not used the real values** of β_0 and β_1, but **only our estimates** $\hat{\beta}_0$ and $\hat{\beta}_1$ based on the data we had. If we were to ask another 30 people, we'd probably get a similar but slightly different line of best fit. So, we'd also get slightly different estimates of β_0 and β_1, and thus of our friend's wage too.

We can also calculate the fitted values of *Wage* for the individuals in our original sample – even though we actually know what their *true* wages are. These are given by:

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

These fitted values are simply the points on our line of best fit that are vertically inline with our original datapoints. We illustrate this below for individual 13 in our sample ($i = 13$). Their experience is $x_{13} = 27.0$ years; the fitted value of their wage is therefore $\hat{y}_{13} = 20,400 + 780 \times 27.0 = £41,460$ based on the estimated model. Their true wage is surprisingly high in comparison; $y_{13} = £44,192.16$. Their residual is just the difference between these values; $\hat{u}_{13} = y_{13} - \hat{y}_{13} = £2,732.16$. Though this is not the same as the true error term u_{13}, it seems likely that they made a good impression on the bosses! Perhaps the number 13 is not so unlucky after all.



Despite actually knowing their true wages, finding the fitted values explicitly for individuals in our sample is useful for evaluating how closely our line of best fit matches the datapoints. Indeed, when unpacking this basic approach more carefully, we will actually *choose* our line of best fit so that the match between these fitted values and the true wages is 'as good as possible'; or equivalently, so that 'the residuals are small'. Once the basics are understood, the next thing to try to understand is what exactly this means.