In the 'Intro to Linear Regression' section, we considered how to estimate the coefficients β_0 and β_1 in the linear model:

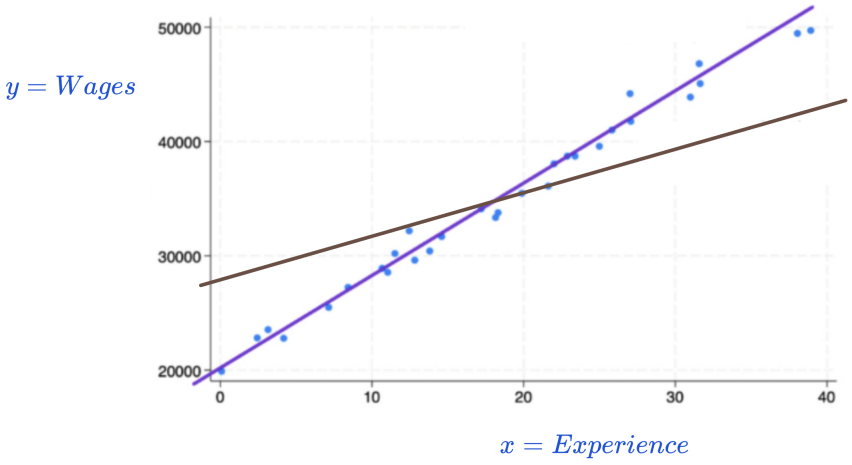
$$Y = \beta_0 + \beta_1 X + U$$

Essentially, we collect some data points (x_i, y_i), and draw a 'line of best fit' (see the previous slides). Here will discussed this idea in much greater depth.

1 Least Squares

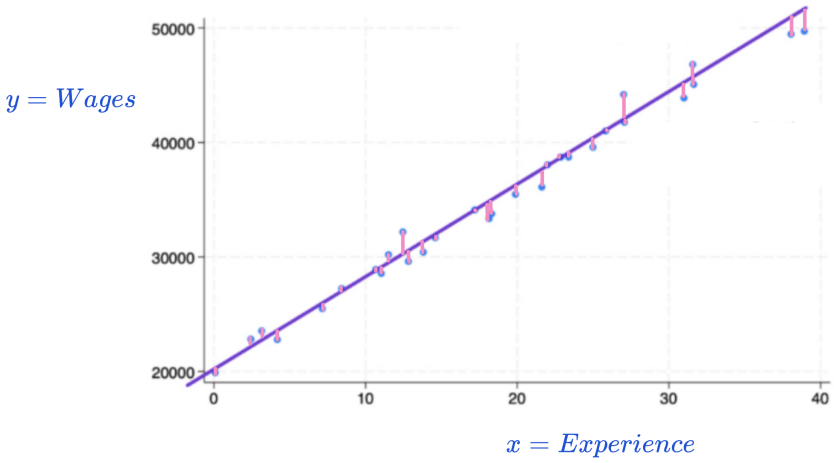
You may have some experience with drawing a 'line of best fit' from school. Most likely, you just drew on a line by eye, based on a vague feeling about what looked 'best'. However, in econometrics, we'll learn to be far more systematic. This will require thinking carefully about what we really mean by a 'line of best fit', which we'll need to make precise and formalise in mathematical terms.

Below picturing our data and our proposed line of best fit. I've also added another line, in brown. But this new line is not as good as our original one, in purple. Why not?



Intuitively, the brown line is a bad choice of 'line of best fit' because it doesn't match the data. What we'd like is for the line to be 'close' to the data.

In formalising this notion, it makes sense to look at the vertical distances of each point, since that is the difference between our predicted wage (according to that candidate line) and the actual wage. When we've produced our final line of best fit, these distances will be precisely *the residuals*. But how can we do this? Let's first draw on these vertical distances (in pink).



What we'd like is for these distances to 'be as small as possible'. Let's try to be even more precise. We can make any *single* pink line length zero by just drawing the line exactly through that point; but then, the other distances might get bigger. So, what we'd really like is for them to 'all be small at the same time'. But there sure are a lot of them to keep track of! Is there a convenient way to summarise 'how big they *all* are', in a single number?

One obvious idea here is to just add them all up (or take an average). However, consider the following example. Suppose for one line drawn with five datapoints, these distances are $10, -5, 15, -5$, and -14. For another, the distances are $-1, 3, 0, -1$, and 2. Then in the former case, the sum of the residuals is 1, but in the later case, it is 3. However, intuitively it seems clear that it is the second line that is actually getting closer to the data! The problem here is that all our residuals are 'cancelling out' in the sum, so just adding them up turns out to be a poor guide to how big they were initially (and likewise with taking the average).

We can avoid this issue with cancellation by first *squaring* the distances before adding them up. This will work because the square of any number is never negative. Then in the former case, we get $10^2 + (-5)^2 + 15^2 + (-5)^2 + (-14)^2 = 571$, whereas in the latter case we get $(-1)^2 + 3^2 + 0^2 + (-1) + 2^2 = 15$. So, this 'summary' gives a much less misleading picture of what is going on.[1]

This, then, is essentially what we mean by a 'line of best fit': it is a line such that the sum of the *squared* distances from the data (i.e. the squares of the line lengths shown in pink above) are as small as possible. The process of systematically constructing the line of best fit according to this technical definition is called an 'ordinary least squares' regression, or OLS. Reflect on the choice of name here, noting that it is termed 'ordinary' because that's the way

[1] We could have simply ignored the minus signs instead; however, in doing it this way the maths turns out to be easier, and large errors are also emphasised proportionally more.

we ordinarily do things around here. We go through the maths for this process over the next couple of sections.

2 The Quadratic 'Loss Function'.

We continue our hunt for the line of best fit, having now clarified what this means. This will involve some dense algebra; it's important to at least get the gist of this, even if you can't reproduce all of it yourself.

Let's introduce some notation for *any* generic candidate line, writing $y = c + mx$, with intercept c and gradient m. We are acquiring a lot of notation, but this is essential to keep track of the differences we need to attend to. To reiterate: β_0 and β_1 are the *true* values we're trying to estimate; $\hat{\beta}_0$ and $\hat{\beta}_1$ will be our *final, best* guesses that we're carefully working towards, and these new c and m are *any generic* guesses we might make (not necessarily the best ones).

The vertical distance of the datapoint (x_i, y_i) from the generic line $y = c+mx$ is equal to y_i–$(c + mx_i)$. Expanding out the bracket, this is $y_i - c - mx_i$. For any particular line, we evaluate its performance by looking at the expression formed by squaring these distances and adding them up:

$$L(c, m) = \sum_{i=1}^{n} (y_i - c - mx_i)^2$$

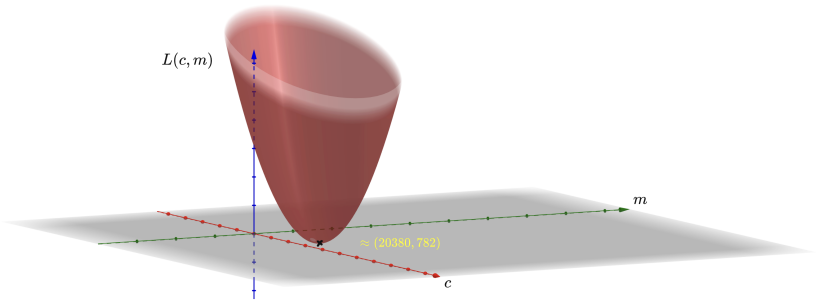
The bigger this function $L(c, m)$ is, the worse the line $y = c + mx$ is performing in matching the data. We sometimes call this a 'loss function', following general usage in economics (it's like the opposite of our 'utility function', if you know what that is). Hence the choice of the letter 'L' in the definition of the function.

The expression above looks rather complex; however, we should learn not to be afraid of intricate notation. Indeed, armed with our dataset, all of the y_i and x_i values are *just numbers* that we *know*. So, subbing these in, we'll be left with a quadratic function of c and m. This is a bit like the quadratics you might be familiar with from school; but we have two variables rather than one.

For the above dataset, we can sub in the values 30 pairs of values (x_i, y_i) for each individual and simplify to get

$$L(c, m) = 30c^2 + 13,299.37m^2 + 1,103.57cm - 2,086,290.76c - 43,301,427.38m + 3,8225,114,630.89$$

Some of these coefficients are quite big – but again, they are still just numbers! We can now plot a 3D graph of $L(c, m)$, as a function of the intercept c and gradient m:



Inspecting the graph visually, the 'best' values seem to be around $c = 20,380$ and $m = 782$. This is an improvement on our earlier efforts. But what if we wanted to be even more accurate – or didn't have 3D graphing software?

To find the minimum of the function $L(c, m)$, we can also proceed algebraically. We take the partial derivates of $L(c, m)$ with respect to both c and m, and set each of these equal to zero:

$$\frac{\partial L}{\partial c} = 60c + 1,103.57m - 2,086,290.76 = 0$$

$$\frac{\partial L}{\partial m} = 26,598.74m + 1,103.57c - 43,301,427.38 = 0$$

That is, we have two simultaneous equations in c and m:

$$60c + 1,103.57m = 2,086,290.76$$

$$26,598.74m + 1,103.57c = 43,301,427.38$$

Finally, these can be solved using school methods. For instance, I could rearrange the first one to make c the subject; then substitute this expression into the second to give a single equation in m only. Solving this to find m and then c, we finally obtain precise answers for our 'best estimates of β_0 and β_1:

$$c = \hat{\beta}_0 = 20,384.59, \;\; m = \hat{\beta}_1 = 782.20$$

Phew! These calculations were a pain to do by hand; but fortunately, it won't be necessary to do nearly as much work as that. At a higher level, we usually don't spend time working through actual calculations; we now have computers to do that for us.

In the next section, we proceed a bit more systematically, aiming to get a general *formula* for $\hat{\beta}_0$ and $\hat{\beta}_1$. Then we (or more likely, one of our computer slaves) can always just plug in the data (x_i, y_i) and get the answers immediately,

with very little fuss.

3 Formulae for the OLS Estimates

Recall that when we find our line of best fit $y = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we start with a generic line $y = c + mx$, and then find values of c and m which make the following expression as small as possible:

$$L(c, m) = \sum_{i=1}^{n} (y_i - c - mx_i)^2$$

Let's now proceed by taking partial derivatives right away. This time we won't expand out the brackets, but it may help to first write out the terms, to give:

$$(y_1 - c - mx_1)^2 + (y_2 - c - mx_2)^2 + ... + (y_n - c - mx_n)^2$$

If we differentiate term-by-term with respect to c, using the chain rule, you can check that we get

$$2(y_1 - c - mx_1)(-1) + 2(y_1 - c - mx_1)(-1) + ... + 2(y_1 - c - mx_1)(-1)$$

Moving back to sigma notation, this is

$$\frac{\partial L}{\partial c} = \sum_{i=1}^{n} 2(y_i - c - mx_i)(-1)$$

Doing the exact same thing for m, we have:

$$\frac{\partial L}{\partial m} = \sum_{i=1}^{n} 2(y_i - c - mx_i)(-x_i)$$

We now set both derivatives equal to zero, giving two equations characterising our OLS estimates; that is, our 'best guesses' $c = \hat{\beta}_0$ and $m = \hat{\beta}_1$:

$$\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$
$$\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

Essentially, we must just solve these simultaneously for $\hat{\beta}_0$ and $\hat{\beta}_1$, recalling that the x_i and y_i are just known numbers from our dataset. There are lots of ways to do this, but let's proceed as follows.

Firstly, dividing each equation through by -2, we have:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$$

Now, let's focus on the first equation. We can separate the terms in the sum, to give

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

Where we note that $\hat{\beta}_0$ appears n times in the previous sum, and that since $\hat{\beta}_1$ does not depend on i, it can be taken outside. Now, we can divide each term through by n, and using the notation $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ for the averages:

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

Moving the terms over, we get $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. This means that, on our line of best fit, when $x = \bar{x}$, we also have $y = \bar{y}$. That is, our line of best fit goes through the point $(\bar{x}, \bar{y})$; for the dataset above, the average values of Wage and Experience. This should seem like a reasonable thing to happen.

Now we return to the first equation. Separating the terms again, we get:

$$\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

From the first equation, we had that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Substituting this in here, we get:

$$\sum_{i=1}^{n} y_i x_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

Now, we move the terms that depend on $\hat{\beta}_1$ to the right hand side:

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

Factoring out $\hat{\beta}_1$:

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

Moving the averages inside the sums:

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

Combining the sums, and factoring out x_i:

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} (x_i - \bar{x}) x_i \right)$$

Finally, dividing through to find $\hat{\beta}_1$:

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

This formula is all well and good as it stands. However, there is an equivalent formula that we usually prefer to use (see the exercises at the end of the chapter for the proof these are indeed the same). Using this formula instead, the general expressions for our OLS estimates of β_0 and β_1 are:

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

Of course, as above, we don't have to go through the above process every time we estimate a model; now we have the general formulae, next time we can just use this right away! Indeed, you'll probably never even have to actually work out these numbers by hand. We now have computers to do mundane things like sticking lots of numbers into a formula. However, it's important for you to understand what is going on, so try to grasp as much of the derivation as you can.