

# 1 Population Regression

When we estimate things, our estimation is going to depend on whatever sample we happen to have obtained. That sample is usually not going to be a perfect representation of the population, and hence any given sample will differ from the population in random ways.

To illustrate, suppose you have a population of 100 people and you want to estimate their income. You take 20 random samples, someone else takes 20 random samples. Chances are you won't sample the exact same 20 people and hence your estimates will be a bit different. We need to account for that sampling variability.

In the context of regressions, we'd like a regression that best fits the population data. It will be given by the formula

$$y_i = \beta_1 + \beta_2 x_i,$$

which I will explain momentarily.

**Assumption 1.** Again, this is just the line of best fit – it is not the of perfect fit. In generality there will be no line of perfect fit. So when we talk about a specific data point  $i$ , the true population model is

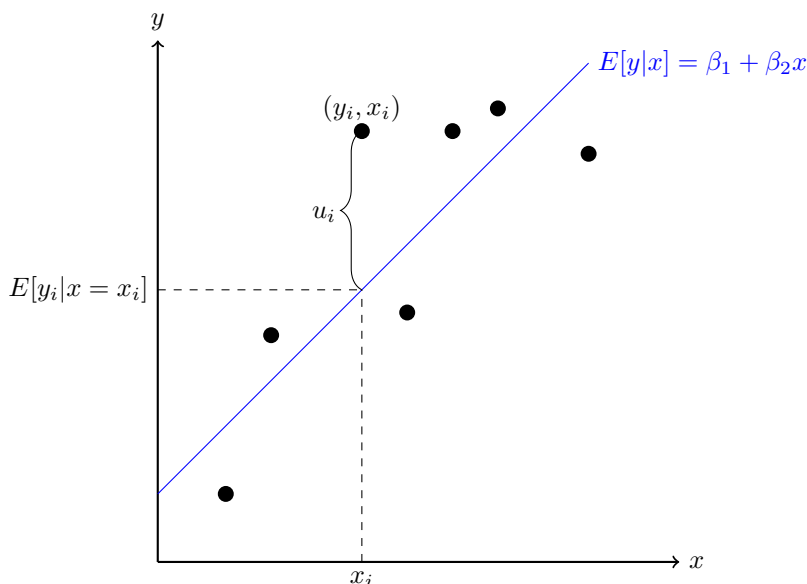
$$y_i = \beta_1 + \beta_2 x_i + u_i,$$

where  $u_i$ , called the **error**, is the difference between the actual point  $y_i$  given  $x_i$  and the regression line's estimate with  $x_i$ .

**Assumption 2.** Now use assumption 2, the zero conditional mean:  $E[u_i|x_i] = 0$ . This allows us to take the true population model and write

$$\begin{aligned} E[y_i|x = x_i] &= E[\beta_1|x = x_i] + E[\beta_2 x_i|x = x_i] + E[u_i|x = x_i] \\ &= \beta_1 + \beta_2 x_i. \end{aligned}$$

This is true because  $\beta_1$  and  $\beta_2$  are just numbers – there is nothing random about them – so we, uh, expect them to be themselves. And because of our zero conditional mean assumption, the error term drops out. Thus, the regression line is what we expect  $y_i$  to be, given  $x_i$ .



Pick some arbitrary data point  $(x_i, y_i)$ . The regression line tells us  $E[y_i|x = x_i]$ , that is, what value we expect  $y_i$  to be for independent variable  $x_i$ . This is the **conditional mean** of  $y_i$  given  $x_i$ . But the regression line is a line of *best* fit, not a line *perfect* fit, so the actual value of  $y_i$  will in general be different than what we expect it to be based on the regression line. The difference between what we expect  $y_i$  to be based on the regression and what  $y_i$  actual is is called the **error**, denoted  $u_i$ .

To summarize the population characteristics:

- The actual value  $y_i$  is given by  $y_i = \beta_1 + \beta_2 x_i + u_i$ .
- The regression line is what we expect  $y_i$  to be, given  $x_i$ . Expressed in the maths,  $E[y_i|x = x_i] = \beta_1 + \beta_2 x_i$ . This is a consequence of assumptions 1 and 2 combined.
- And hence the error term is given by  $u_i = y_i - E[y_i|x = x_i]$ .

We can throw down two more assumptions to make analysis easier.

- **Assumption 3: homoskedasticity.** The variation of  $u$  given  $x_i$  is the same for any  $x_i$ . In math,

$$\text{Var}(u_i|x_i) = \sigma_u^2 \quad \forall i.$$

- **Assumption 4: independent errors.** Errors for different observations are statistically independent:  $u_i$  is independent of  $u_j$  whenever  $i \neq j$ .

Assumptions 3 and 4 allow us to say that the variation of  $y$  given  $x$  is also constant, and specifically,  $\text{Var}(y|x) = \sigma_u^2$ .

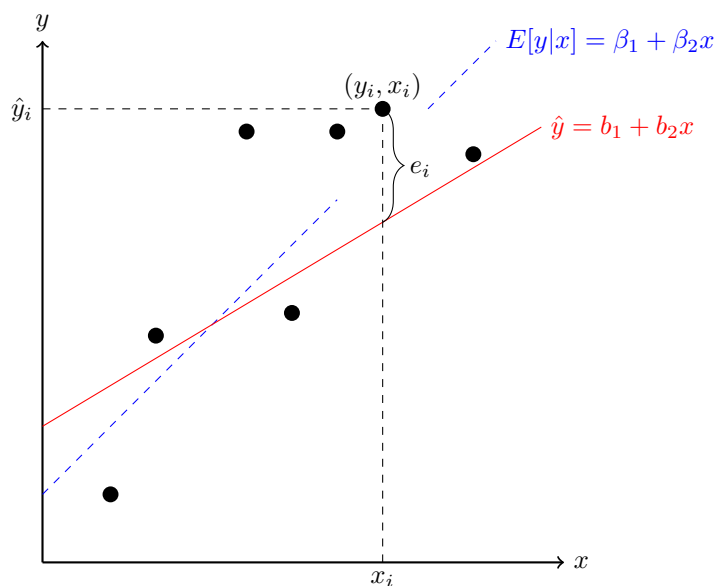
## 2 Estimation Regression

Now we use sample data to estimate  $\beta_1$  and  $\beta_2$  using the ordinary least squares (OLS) technique. Call these estimates  $b_1$  and  $b_2$ , respectively. Under **assumptions 1 and 2**, the estimates will be unbiased:  $E[b_1] = \beta_1$  and  $E[b_2] = \beta_2$ . That said, they will be different in generality than their population analogues. Hence our estimated regression line will be more or less different than the population regression line, depending how well we can estimate them.

For our estimated regression, our prediction of  $y_i$  given  $x_i$  is called the **fitted value** and is given by

$$\hat{y}_i = b_1 + b_2 x_i.$$

Much like in the population case, this will in generality be different than the actual value  $y_i$ . We call the difference between the actual value  $y_i$  and our fitted value  $\hat{y}_i$  the **residual**, denoted  $e_i$ . (I find this confusing as hell – why not denote the error term with  $e$  instead? Sigh, economics.)

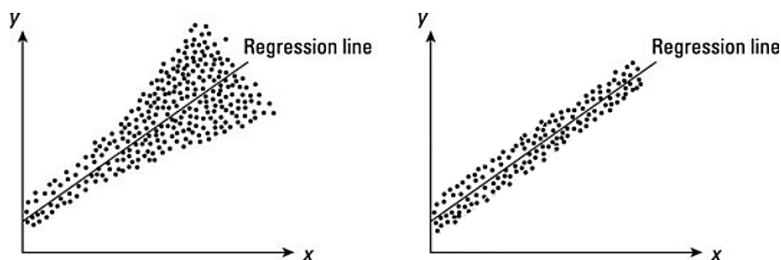


Our estimated regression line is in red. It will almost always be a bit different than the true population regression line, in blue. For  $x_i$ , it gives us a prediction for  $y_i$ , i.e. the fitted value  $\hat{y}_i$ . The fitted value will not in general be exactly the true value  $y_i$ , and the difference between the true value and the fitted value is the residual  $e_i = y_i - \hat{y}_i$ .

**Assumptions 3 and 4** imply that the variance of the slope estimator  $b_2$  will be

$$\text{Var}(b_2) = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \equiv \sigma_{b_2}^2.$$

**Assumption 3** is most likely to break down, in which case we will have **heteroskedasticity** – the variance of  $u_i$  will depend on  $x_i$ . In this case we need to use **heteroskedasticity-robust standard errors**, which are given by the Stata option `vce(robust)`.



The figure on the left is an example of heteroskedasticity; the right an example of homoskedasticity. The left is heteroskedastic because the variation around the regression line gets bigger as  $x$  increases.

### 3 Estimation Properties

Under **assumptions 1-4**, our slope estimator  $b_2$  has expected value of  $\beta_2$  because it is unbiased; and it also has variance  $\sigma_{b_2}^2$ . Thus we can write

$$b_2 \sim (\beta_2, \sigma_{b_2}^2).$$

The  $z$ -score is standard normal, i.e.

$$\frac{b_2 - \beta_2}{\sigma_{b_2}} \sim \mathcal{N}(0, 1).$$

But we don't actually know  $\sigma_{b_2}$ , so we have to divide by the standard error  $se(b_2)$  instead. Under **assumptions 1-4**,

$$T = \frac{b_2 - \beta_2}{se(b_2)} \sim \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty.$$

If we add an **additional assumption** that the error terms are normally distributed with mean zero, i.e.  $u_t \sim \mathcal{N}(0, \sigma_u^2)$ , then we can say that  $T \sim t(n-2)$  exactly. But we can still use  $t(n-2)$  as an approximation even if errors are not normally distributed (which we will when doing hypothesis testing and confidence intervals).