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

$$y = \beta_1 + \beta_2 x + u,$$

which I will explain in detail momentarily. But think of this as being the line of best fit for the entire population, and we want to estimate β_1 and β_2 using a sample.

2 Unbiased Estimators

Assumption 1: True Population Model. Again, a regression is just the line of *best* fit – it is not the line of *perfect* fit. When we talk about a specific data point i, we will assume that the true population model is

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

What this says is we use the line $\beta_1+\beta_2 x_i$ to best "predict" what y_i should be for a given value of x_i; but since the regression line doesn't perfectly capture all data points, the prediction will be off by u_i. Accordingly, u_i is called the **error term**, sometimes called the **disturbance term**.

Assumption 2: Zero Conditional Mean. The zero conditional mean assumption states that $E[u_i|x_i] = 0$ for all i. Consider a specific $x_i = x^*$, where x^* is just any old number we plug in for x_i. This allows us to take the true population model and write

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

This is true because β_1 and β_2 are just numbers – there is nothing random about them – so we, uh, expect them to be themselves, regardless of what x_i is. 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.

For more intuition, suppose $E[u_i|x_i] \neq 0$. Then when we plug in a number for x, we expect the model to have some error, on average. This makes for a pretty lousy model if we expect it to be wrong, on average.

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.
- Hence the error term is given by $u_i = y_i - E[y_i|x = x_i]$.

Assumptions 1 and 2 imply that OLS estimates (explained soon) b_1 and b_2 for β_1 and β_2 are unbiased, in other words, we expect the estimates to be their true values. In maths, $E[b_1] = \beta_1$ and $E[b_2] = \beta_2$.

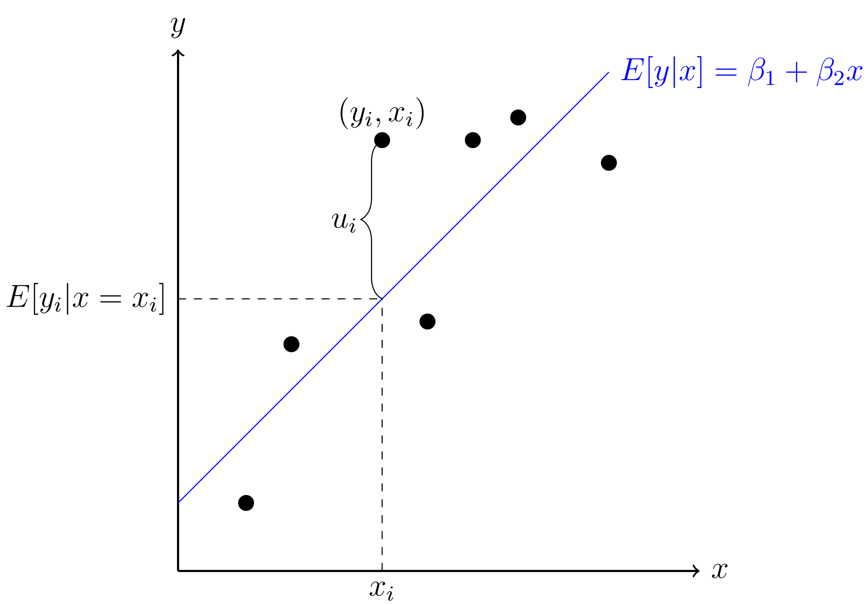

FIGURE 1: Pick some arbitrary data point (x_i, y_i). The regression line tells us $E[y_i|x = x_i]$, that is, what value we expect y_i to be for independent variable x_i. This is the **conditional mean** of y_i given x_i. But the regression line is a line of *best* fit, not a line *perfect* fit, so the 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 y_i actually is and what we expect y_i to be based on the regression, $y_i - \beta_1 - \beta_2 x_i$, is the error term, u_i.

3 BLUE and BUE

3.1 BLUE

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

Assumption 3: Homoskedasticity. The variation of u_i given x_i is the same number σ_u^2 for any x_i. In math,

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

Assumption 4: Uncorrelated Errors. Errors for different observations are uncorrelated: $\text{Cor}(u_i, u_j) = 0$ whenever $i \neq j$.

Adding assumptions 3 and 4 allows us to say that the variation of y given x is also constant, and specifically, $\text{Var}(y|x) = \sigma_u^2$. OLS assumptions 1-4 imply also imply that estimates are **consistent**, provided the variances of the estimates goes to zero as $n \to \infty$. Put somewhat crudely, this means that our estimates get arbitrarily close (in probability) to their true population values as the sample size increases. In math, we write $b \xrightarrow{p} \beta$.

We can go even further, however. Under OLS assumptions 1-4, the estimates are said to be **BLUE**, which stands for

- **B**est: estimates have the smallest standard errors...
- **L**inear: among linear models...
- **U**nbiased: that give unbiased...
- **E**stimator: um, estimates.

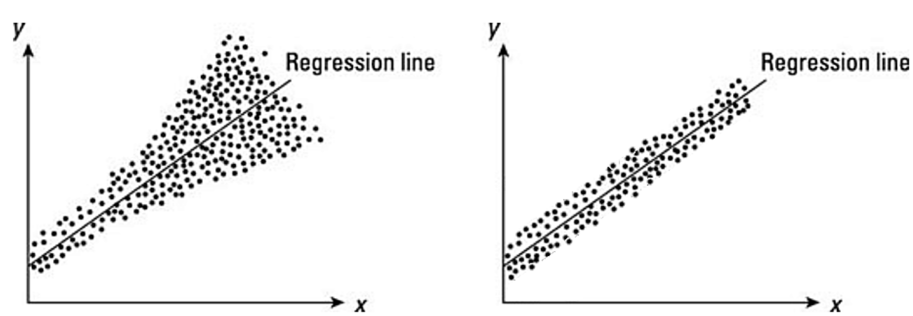

FIGURE 2: The figure on the left is an example of heteroskedasticity; the right an example of homoskedasticity. The left is heteroskedastic because the variation around the regression line gets bigger as x increases.

3.2 BUE

We can make a fifth assumption for one more nice result.

Assumption 5: Normally Distributed Errors. Error terms have normal distribution with some variance σ^2,

$$u_i \sim \mathcal{N}\big(0, \sigma^2\big).$$

This allows us to say that OLS estimates are **BUE**, which means that they have the smallest standard errors among unbiased models, even when compared to nonlinear models. Also note that this is an essential condition if we want to do inference on small sample sizes.

4 OLS Estimation of a Regression

Intuitively, we want a model that makes the fewest mistakes possible. We quantify this by considering the difference between the actual values y_i, and the **fitted values** as predicted by the model, given by $\widehat{y}_i = b_1 + b_2 x_i$; this is referred to as the **residual**, denoted e_i, defined as

$$e_i \equiv y_i - \widehat{y}_i = y_i - [b_1 + b_2 x_i].$$

Now square each residual to ensure that it's positive, then we add the squared terms all up: this is the **residual sum of squares (RSS)**. We want the estimates that *minimize the residual sum of squares.* In math speak, we want to solve

$$\underset{b_1, b_2}{\arg\min} \sum_{i=1}^{n} \big(y_i - [b_1 + b_2 x_i]\big)^2.$$

The solution to this problem is the **ordinary least squares (OLS)** estimation for a linear regression.

I omit the details, but explicitly solving OLS gives formulas

$$b_2 = \frac{s_{xy}}{s_x^2} = r_{xy} \times \frac{s_y}{s_x},$$

$$b_1 = \bar{y} - b_2 \bar{x},$$

where s_{xy} is the **sample covariance** defined by

$$s_{xy} \equiv \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

and r_{xy} is the **sample correlation coefficient** defined by

$$r_{xy} \equiv \frac{s_{xy}}{s_x s_y}.$$

Again, 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 because, well, they're estimates. Hence our estimated regression line will be more or less different than the population regression line, depending on how closely our sample reflects the population.

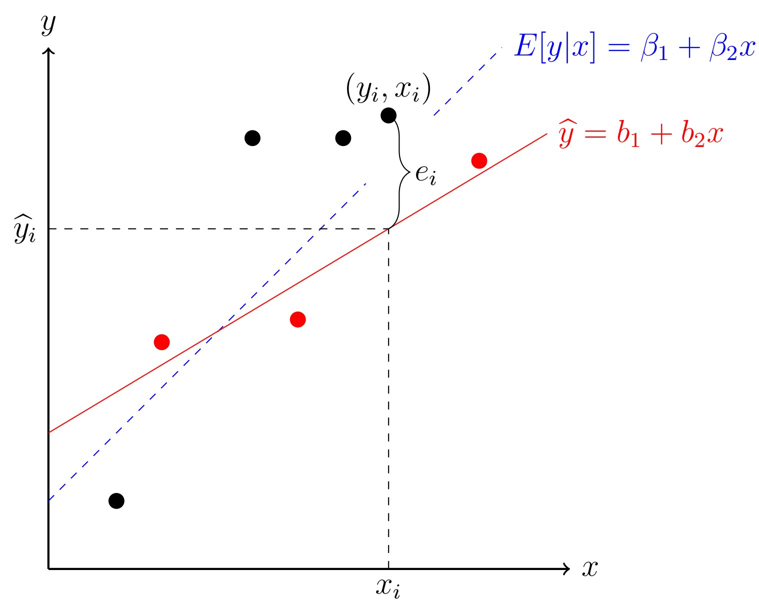

FIGURE 3: Suppose our sample consists of only the red dots. Thus the estimated regression line is in red, which is different than the true population regression line, in blue. For x_i, it gives us a prediction for y_i, i.e. the fitted value $\widehat{y}_i$. The fitted value will not in general be exactly the true value y_i, and the difference between the true value and the fitted value is the residual, $e_i = y_i - \widehat{y}_i$.

Furthermore, assumptions 3 and 4 imply that the variance of the slope estimate 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 practice, 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**.

5 Explained and Unexplained Variation

To reiterate, we define the **residual sum of squares** to be

$$\text{RSS} \equiv \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

This captures the total error of the estimated regression line, squared so that the errors are positive. Dividing this by $n-2$ and taking the square root gives the **standard error of the regression**,

$$s_e \equiv \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2},$$

which is the sample analogue of σ_u. You can think of this as being the variation of data around $\bar{y}$ that cannot be explained by x. This is sometimes called the **standard error of the residuals** or **root mean squared error (RMSE)**.

On the other hand, the variation of data around $\bar{y}$ that can be explained by x is the **explained sum of squares**,

$$\text{ESS} \equiv \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2.$$

Finally, the total variation of data around $\bar{y}$ is given by the **total sum of squares**,

$$\text{TSS} \equiv \sum_{i=1}^{N}(y_i - \bar{y})^2.$$

Based on the intuition it should not be surprising, and it is not difficult to show either, that

$$\text{TSS} = \text{ESS} + \text{RSS}.$$

Total variation is explained variation plus unexplained variation. Great.

The proportion of explained variation around $\bar{y}$ is called the **R-squared** or **coefficient of determination**, defined as

$$R^2 \equiv \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

If R^2 is high, then x explains a lot about what's going on with y; if R^2 is low, then it doesn't. There is no cutoff for what should be considered "high" or "low," however. Note that R^2 also equals the squared correlation between y and x, that is, $R^2 = r_{xy}^2$. Also note that R^2 is only valid if the regression includes the intercept.

Note that to test whether R^2 is zero or not in a bivariate regression, we use the test statistic

$$F \equiv \frac{R^2}{(1-R^2)/(n-2)} \sim F(1, n-2),$$

or equivalently,

$$F \equiv \frac{\text{ESS}}{\text{RSS}/(n-2)} \sim F(1, n-2),$$

6 Estimator Properties

Under assumptions 1-4, our slope estimator b_2 has expected value of β_2 because it is unbiased; and it also has variance $\sigma^2_{b_2}$. Thus we can write

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

For sufficiently large sample size (greater than 30), the z-score is approximately standard normal, that is,

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

But we don't actually know what σ_{b_2} is because we don't know what σ_u is. Instead we must use the sample estimate of σ_u, given earlier as s_e. This then allows us to conclude that the sample standard error of b_2 is

$$\text{se}(b_2) = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

So under assumptions 1-4, for sufficiently large sample size (which does *not* have a clear cut rule-of-thumb in this case), we use the t-statistic

$$t \equiv \frac{b_2 - \beta_2}{\text{se}(b_2)} \sim T(n-2),$$

where the distribution is approximate. If we add an additional assumption that the error terms are normally distributed, or if $n \to \infty$, then we can say that $t \sim T(n-2)$ exactly.

7 Regression by Hand

Consider the following data:

$$(y_1, x_1) = (2, 0),$$
$$(y_2, x_2) = (3, 3),$$
$$(y_3, x_3) = (4, 3).$$

Let us regress this entire thing manually.

Step 1: Estimate Regression Coefficients. We need the following sample estimates:

$$\begin{aligned}
\bar{y} &= \frac{1}{3}\left[2+3+4\right] = 3, \\
s_y^2 &= \frac{1}{2}\left[(2-3)^2 + (3-3)^2 + (4-3)^2\right] = 1, \\
\bar{x} &= \frac{1}{3}\left[0+3+3\right] = 2, \\
s_x^2 &= \frac{1}{2}\left[(0-2)^2 + (3-2)^2 + (3-2)^2\right] = 3 \\
s_{xy} &= \frac{1}{2}\left[(2-3)(0-2) + (3-3)(3-2) + (4-3)(3-2)\right] = 1.5, \\
r_{xy} &= \frac{1.5}{\sqrt{1}\sqrt{3}} \approx 0.866.
\end{aligned}$$

Now we can estimate the regression coefficients, given by

$$\begin{aligned}
b_2 &= \frac{s_{xy}}{s_x^2} = \frac{1.5}{3} = 0.5 \qquad \text{or} \quad r_{xy} \times \frac{s_y}{s_x} = 0.866 \times \frac{1}{\sqrt{3}} = 0.5, \\
b_1 &= \bar{y} - b_2\bar{x} = 3 - 0.5(2) = 2.
\end{aligned}$$

Therefore our estimated regression is

$$\widehat{y}_i = 2 + 0.5x_i.$$

Step 2: Calculate Residuals. We can find the standard error of the regression by finding the fitted values, i.e. by plugging each x_i into the regression formula and finding the corresponding $\widehat{y}_i$. Doing so gives

$$\widehat{y}_1 = 2 + 0.5(0) = 2,$$

$$\widehat{y}_2 = 2 + 0.5(3) = 3.5,$$

$$\widehat{y}_3 = 2 + 0.5(3) = 3.5.$$

The residuals are the difference between the actual y_i and what the regression line expects y_i to be based on x_i, which in our case are

$$e_1 = y_1 - \widehat{y}_1 = 2 - 2 = 0,$$

$$e_2 = y_2 - \widehat{y}_2 = 3 - 3.5 = -0.5,$$

$$e_3 = y_3 - \widehat{y}_3 = -3.5 = 0.5.$$

Step 3: Calculate Standard Error of Residual. The residual sum of squares (RSS), uh, squares each residual and sums them up, so

$$\text{RSS} = (0)^2 + (-0.5)^2 + (0.5)^2 = 0.5.$$

Now we can find the standard error of the residuals using formula

$$s_e \equiv \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} = \sqrt{\frac{0.5}{3-2}} = 0.707.$$

Step 4: Calculate Standard Error of Slope Coefficient. The slope coefficient has standard error

$$\text{se}(b_2) = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

So we gotta figure out the denominator. Not a big deal, it's pretty much just the calculation for the standard deviation of x but without the division. The formula yields

$$\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{(0-2)^2 + (3-2)^2 + (3-2)^2} = \sqrt{6}.$$

Therefore the standard error of b_2 is

$$\text{se}(b_2) = \frac{0.707}{\sqrt{6}} \approx 0.289,$$

Hence for a hypothesis test you would calculate the t-statistic

$$t \equiv \frac{b_2 - \beta_2^0}{\mathrm{se}(b_2)} \sim T(n-2),$$

where β_2^0 is the hypothesized value.