Toulouse School of Economics

M1 INTERMEDIATE ECONOMETRICS

Linear regression: Sampling

Koen Jochmans François Poinas

2025 — 2026

The regression slope as a random variable

With data $(Y_1, X_1), \ldots, (Y_n, X_n)$ the fitted regression line has slope coefficient

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{n} (Y_i - X_i' b)^2 = \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \right).$$

When the data change the regression line changes because $\hat{\beta}$ changes.

We see the data as a random sample from the distribution of (Y, X).

This means that the (Y_i, X_i) are all

mutually independent, and

distributed according to the same distribution as (Y, X).

The population regression line

Given the distribution of (Y, X) we can look for the best linear predictor of Y given $X = x$.

With mean squared error loss this is $x'\beta$ for

$$\beta = \arg\min_b \mathbb{E}((Y - X'b)^2) = \mathbb{E}(XX')^{-1}\mathbb{E}(XY).$$

Indeed, the first-order condition for a minimum here is

$$\mathbb{E}(X(Y - X'b)) = 0.$$

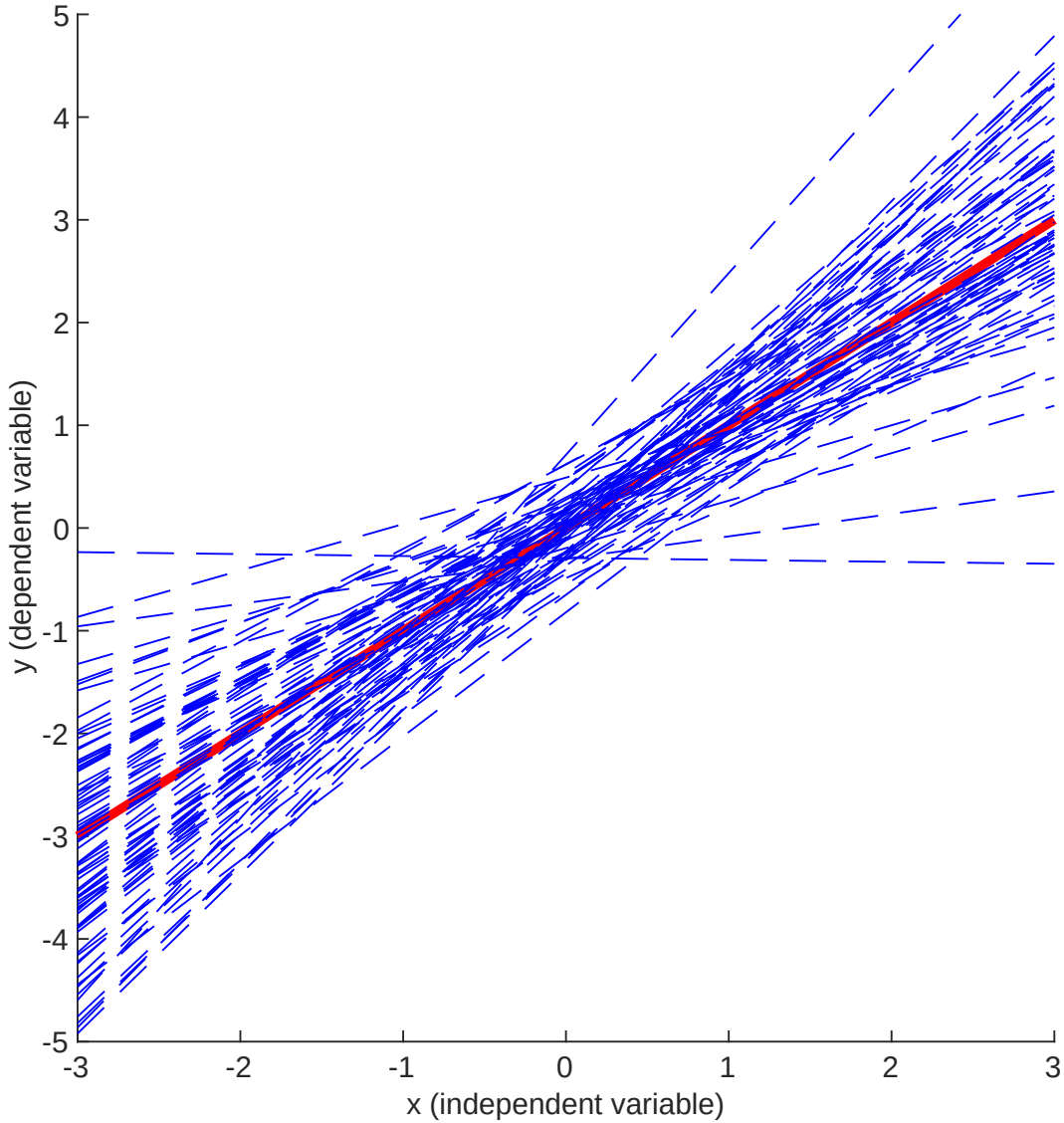
We can define the errors

$$e = Y - X'\beta$$

and write

$$Y = X'\beta + e, \qquad \mathbb{E}(Xe) = 0.$$

Then $\hat{\beta}$ is an estimator of β obtained on solving the sample version of this minimization problem.



Regression lines from many different samples (blue) and the population regression line (red).

Quantifying uncertainty

If we would know the sampling distribution of $\hat{\beta}$ we could calculate measures of spread, such as the variance, or of extreme events via tail probabilities.

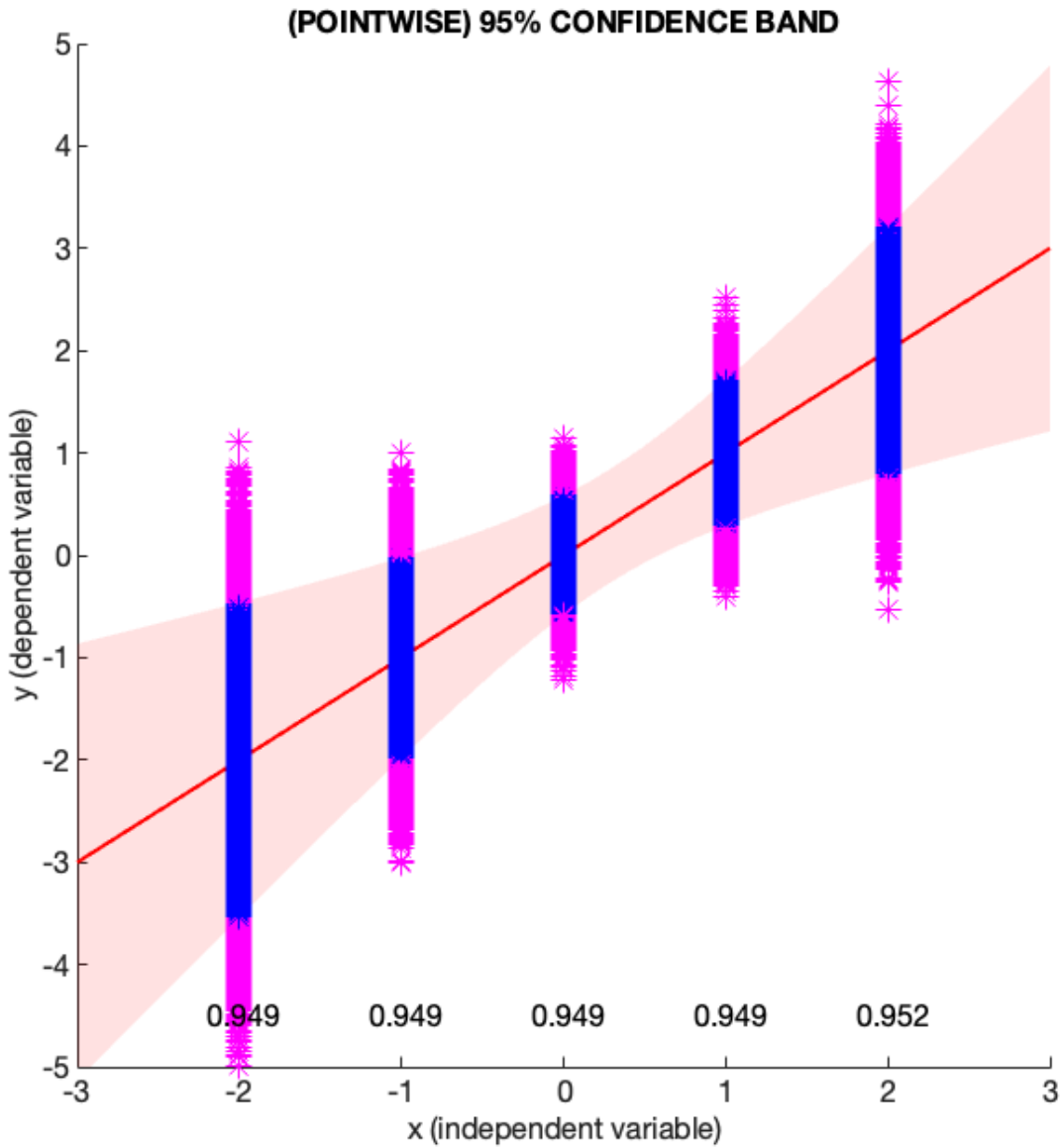
This is not possible, in general.

An exception is the classical linear regression model, where we have that

$$Y|X = x \sim N(x'\beta, \sigma^2).$$

More generally, we will need to work with an approximation to this distribution that becomes better as the sample size grows large (an asymptotic argument as $n \to \infty$).

Classical linear regression model



Here

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}, \qquad \boldsymbol{e}|\boldsymbol{X} \sim N(0, \sigma^2 \boldsymbol{I}_n)$$

and so

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} = \beta + \boldsymbol{A}\boldsymbol{e}.$$

The source of randomness is the second term.

We condition on $\boldsymbol{X}$, so that $\boldsymbol{A}$ is no longer random.

Then

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X} \sim \boldsymbol{A}\boldsymbol{e}|\boldsymbol{A} \sim N(0, \sigma^2 \boldsymbol{A}\boldsymbol{A}').$$

Because

$$\boldsymbol{A}\boldsymbol{A}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

the variance is simply $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

The independent normal errors and the conditioning on the regressors are both important for this result to go through. The key is that, then, $\hat{\beta} - \beta$ is a linear combination of zero-mean i.i.d. normals, which is again normal.

It follows that the regression line at x, $x'\hat{\beta}$, is (conditionally) unbiased for the population regression line $x'\beta$, with normal deviations around it.

Let $\hat{\theta} = x'\hat{\beta}$ and $\theta = x'\beta$ for a chosen value x.

Then

$$\hat{\theta} - \theta | \boldsymbol{X} \sim x'(\hat{\beta} - \beta) | \boldsymbol{X} \sim x' \boldsymbol{A} \boldsymbol{e} | \boldsymbol{X} \sim N(0, \sigma^2 x' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} x).$$

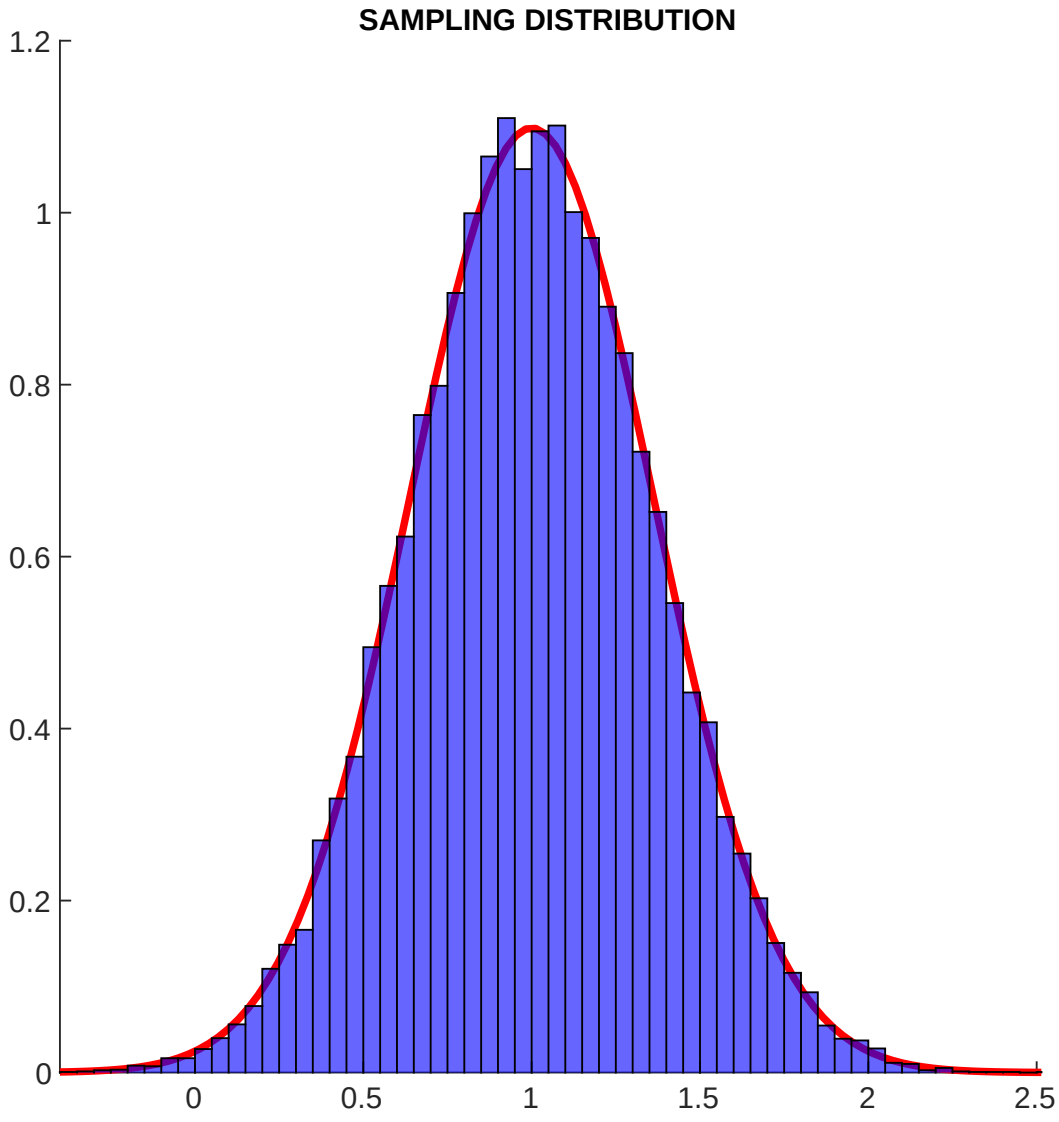
This can be used to compute the confidence bands in the previous figure (where we have taken σ^2 to be known).

The standard deviation of this distribution,

$$\sigma \sqrt{x' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} x},$$

is a measure of spread of $\hat{\theta}$ around θ.

Note how the spread depends on the point x at which the regression line is evaluated.


SAMPLING DISTRIBUTION
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5