Inference in Linear Regression

Uncertainty in estimating the linear regression coefficients

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai

Photo: **Shivesh Sood**

Summary so far

- Statistical model
- k-nearest neighbors (kNN)
- Model fitness and model comparison (MSE)
- Goodness of fit (R2)
- Linear Regression, multi-linear regression and polynomial regression
- Model selection using validation and cross validation
- One-hot encoding for categorical variables
- Overfitting
- Ridge and Lasso regression

Skip ▶I

Comparison of Models

We have seen already 3 models. Choosing the right model isn't about minimizing the test error. We also want to understand and get insights from our models.

	Has f(x) parametric	Easy to interpret
Linear Regression	Yes	Yes
Polynomial Regression	Yes	No
K-Nearest Neighbors	No	Yes

Having an explicit functional form of f(x) makes it easy to store.

Interpretation is important to evaluating the model and understanding what the data tells us

Take home message

By taking a probabilistic approach to linear regression and assuming the residuals are normally distributed, we see that **maximizing the likelihood** for this model is equivalent to **minimizing mean squared error** around the line!

So, if we believe our residuals are normally distributed, then minimizing mean square error is a natural choice.

But by choosing this specific probability model, we get much more than just motivation for our loss function. We get instructions on how to perform inferences as well ☺

Checking the assumptions of linear regression models:

The probabilistic model of linear regression leads to 4 main assumptions that can be checked with the data (the first 3 at least):

1. <u>Linearity</u>: relationships are linear and there is no clear non-linear pattern around the line (as evidenced by the residuals).
2. <u>Normality</u>: the residuals are normally distributed.
3. <u>Constant Variance</u>: the vertical spread of the residuals is constant everywhere along the line.
4. <u>Independence</u>: the observations are independent of each other.

Note: Collinearity is not a violation of an assumption but can certainly muck up the model.

Outline

Part A and B: Assessing the Accuracy of the Coefficient Estimates

Bootstrapping and confidence intervals

Part C: Evaluating Significance of Predictors

Does the outcome depend on the predictors?

Hypothesis testing

🎥 **Part D: How well do we know $\hat{f}$**

The confidence intervals of $\hat{f}$

Outline

Part A and B: Assessing the Accuracy of the Coefficient Estimates

Bootstrapping and confidence intervals

Part C: Evaluating Significance of Predictors

Does the outcome depend on the predictors?

Hypothesis testing

🎥 **Part D: How well do we know $\hat{f}$**

The confidence intervals of $\hat{f}$

How reliable are the model interpretation

Suppose our model for advertising is:

$$y = 1.01x + 5$$

where y is the sales in 1000 units and each unit sales for $1, x is the TV budget in $1000.

Interpretation: for every dollar invested in advertising gets you 1.01 back in sales, which is 1% net increase.

But how certain are we in our estimation of the coefficient 1.01?

Why aren't we certain?

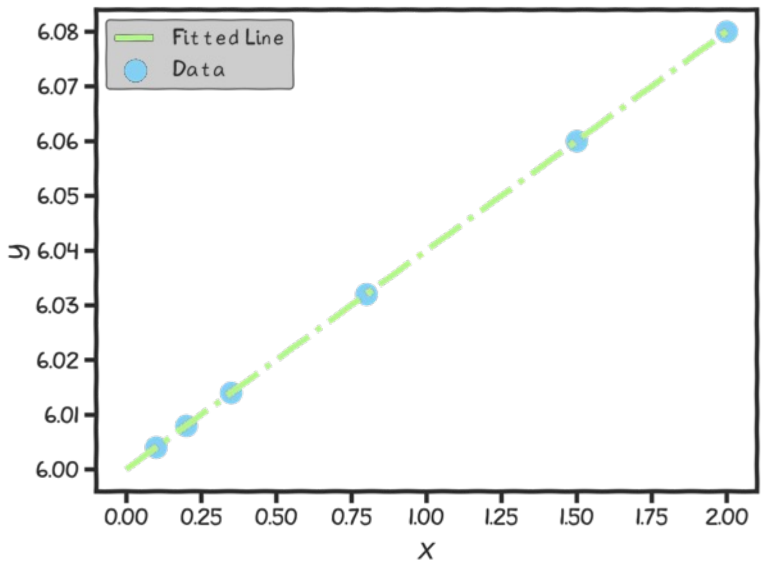
Confidence intervals for the predictors estimates

We interpret the ε term in our observation

$$y = f(x) + \epsilon$$

to be noise introduced by random variations in natural systems or imprecisions of our scientific instruments and everything else.

If we knew the exact form of $f(x)$, for example, $f(x) = \beta_0 + \beta_1 x$, and there was no noise in the data , then estimating the $\hat{\beta}'s$ would have been exact (so is 1.01 worth it?).



Confidence intervals for the predictors estimates (cont.)

However, two things happen, which result in mistrust of the values of $\hat{\beta}'s$:

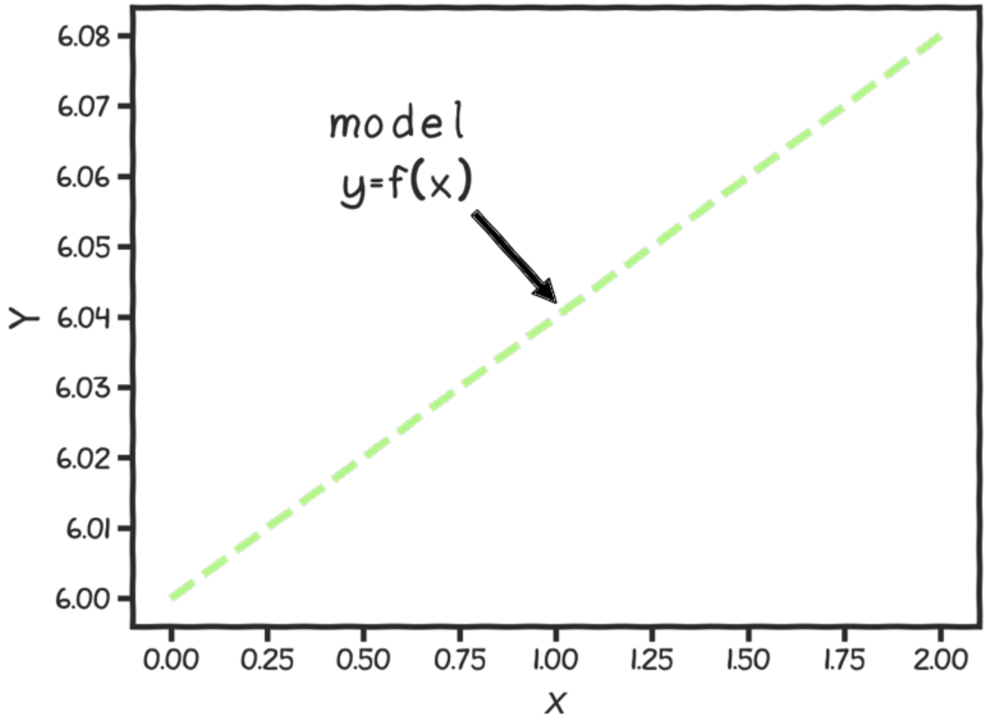
- observational error is always there – this is called ***aleatoric*** error, or ***irreducible*** error.
- we do not know the exact form of $f(x)$ - this is called ***misspecification*** error and it is part of the ***epistemic*** error

We will put everything into catch-it-all term ε.

Because of ε, every time we measure the response y for a fix value of x, we will obtain a different observation, and hence a different estimate of $\hat{\beta}'s$.

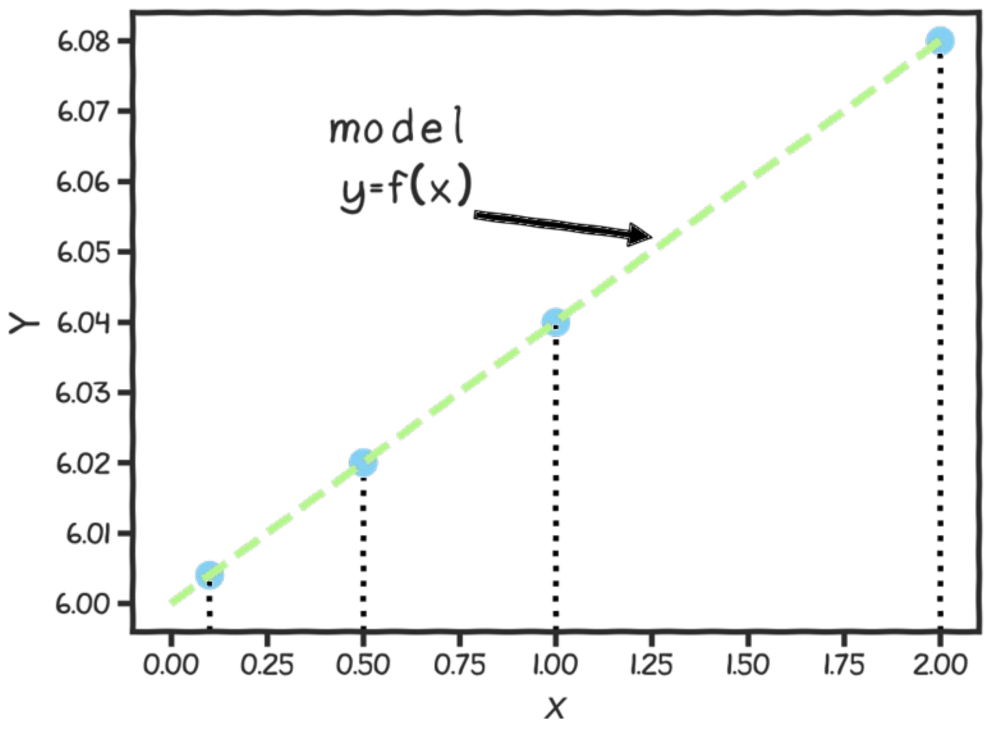
Confidence intervals for the predictors estimates (cont.)

Start with a model $f(X)$, the correct relationship between input and outcome.



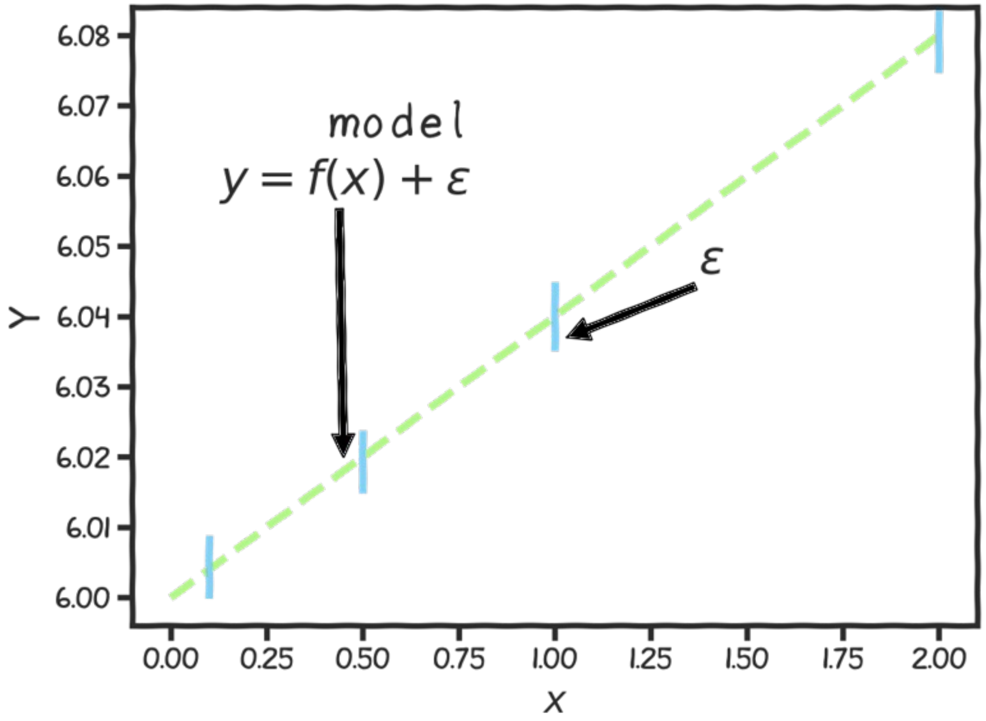
Confidence intervals for the predictors estimates (cont.)

For some values of $X^{\star}, Y^{\star} = f(X^{\star})$



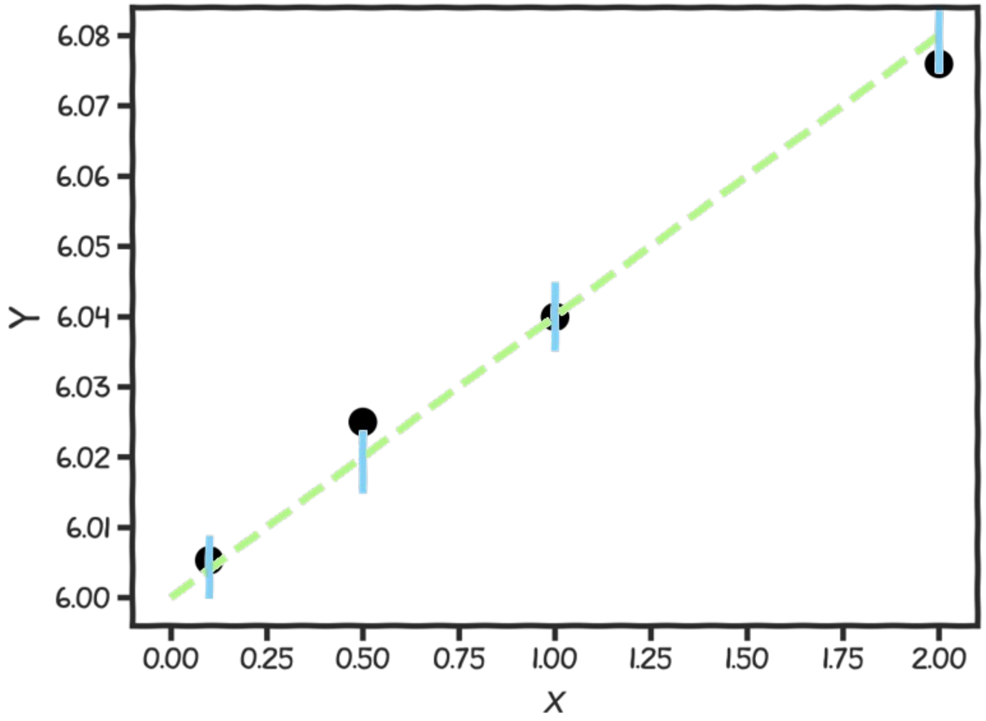
Confidence intervals for the predictors estimates (cont.)

But due to error, every time we measure the response Y for a fixed value of $X^{\star}$ we will obtain a different observation.



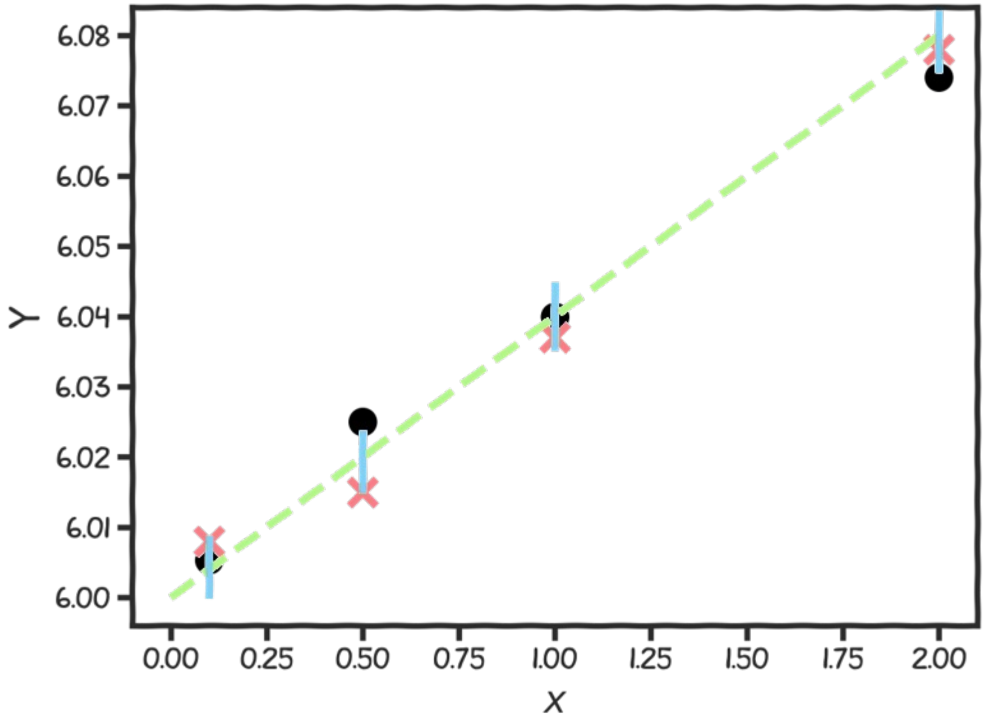
Confidence intervals for the predictors estimates (cont.)

One set of observations, "one realization" yields one set of Ys (Circles: ●).



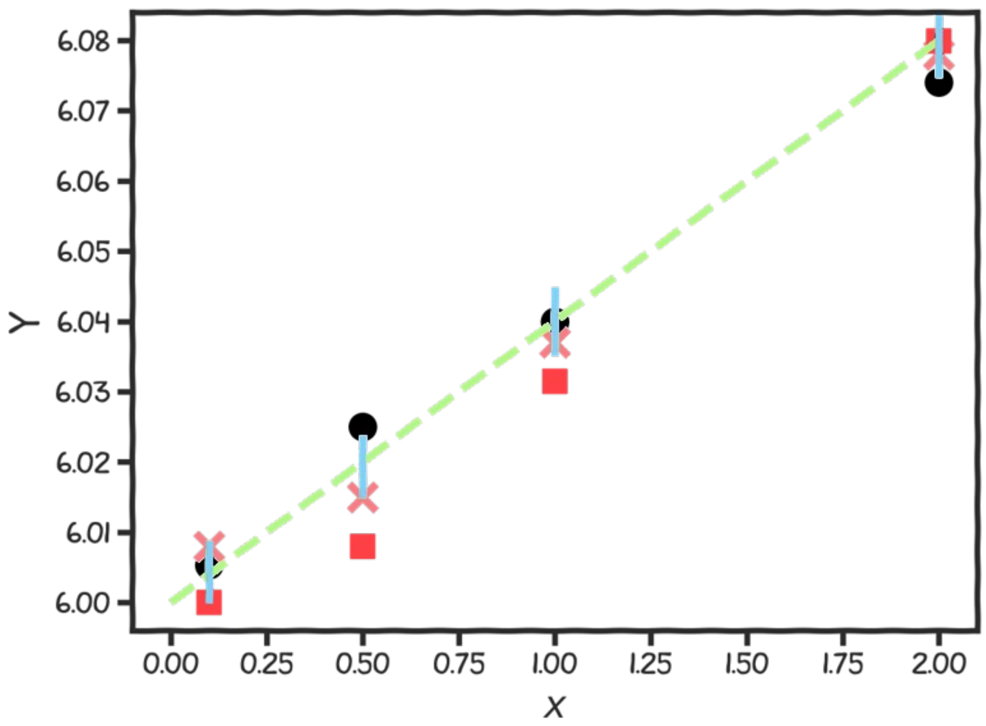
Confidence intervals for the predictors estimates (cont.)

Another set of observations, “another realization” yields another set of Ys (Crosses:).



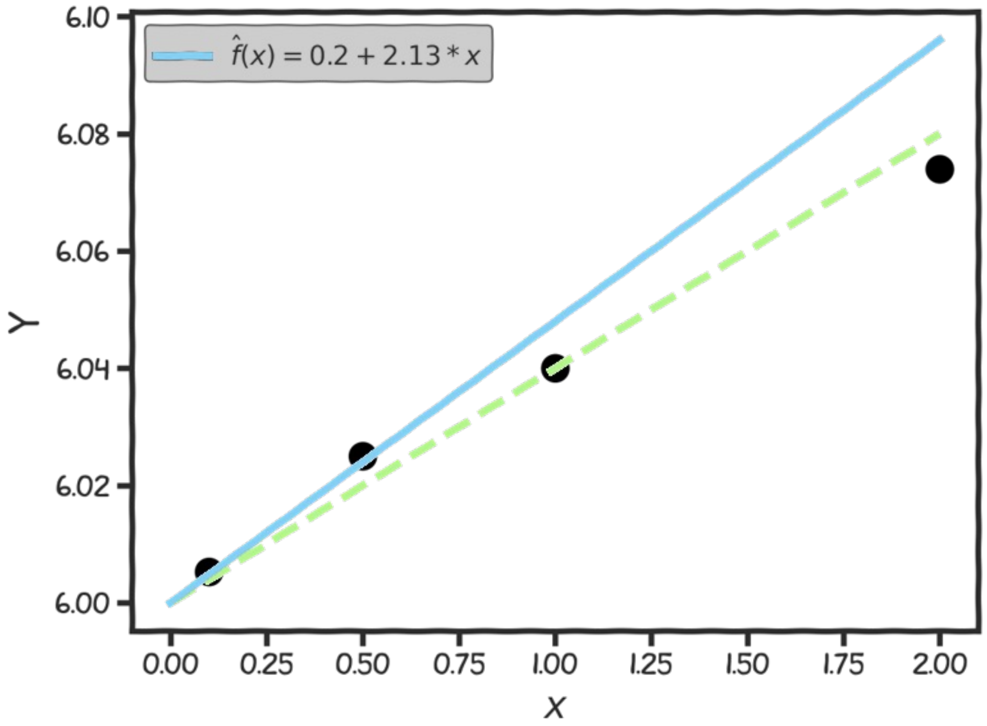
Confidence intervals for the predictors estimates (cont.)

Another set of observations, “another realization”, another set of Ys (Squares: ■).



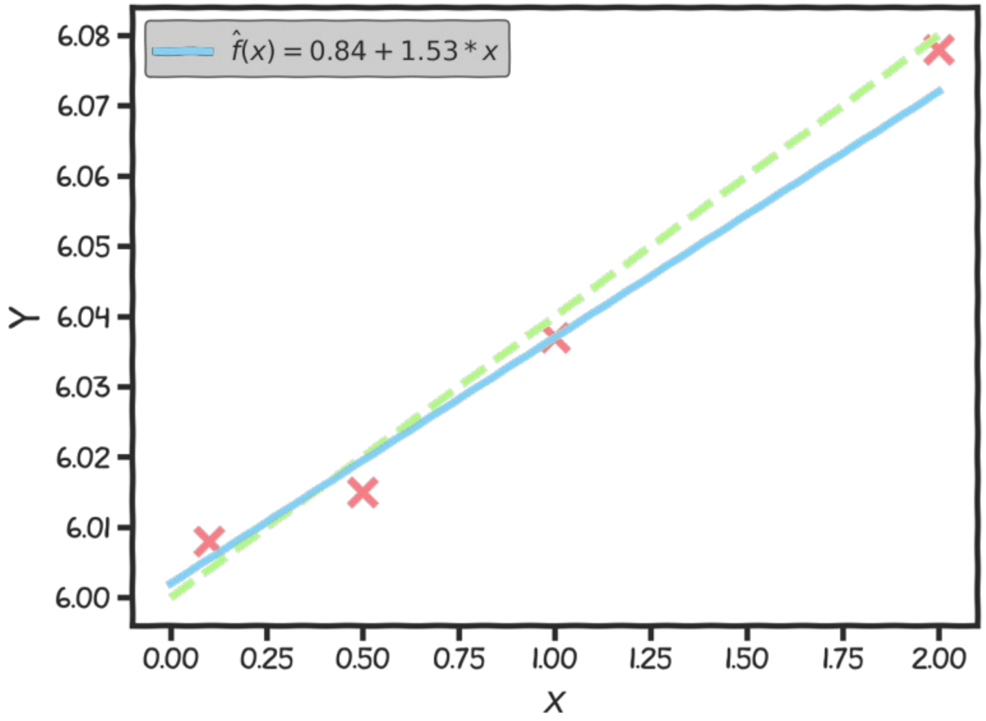
Confidence intervals for the predictors estimates (cont.)

For each one of those "realizations", we fit a model and estimate $\hat{\beta}_0$ and $\hat{\beta}_1$.



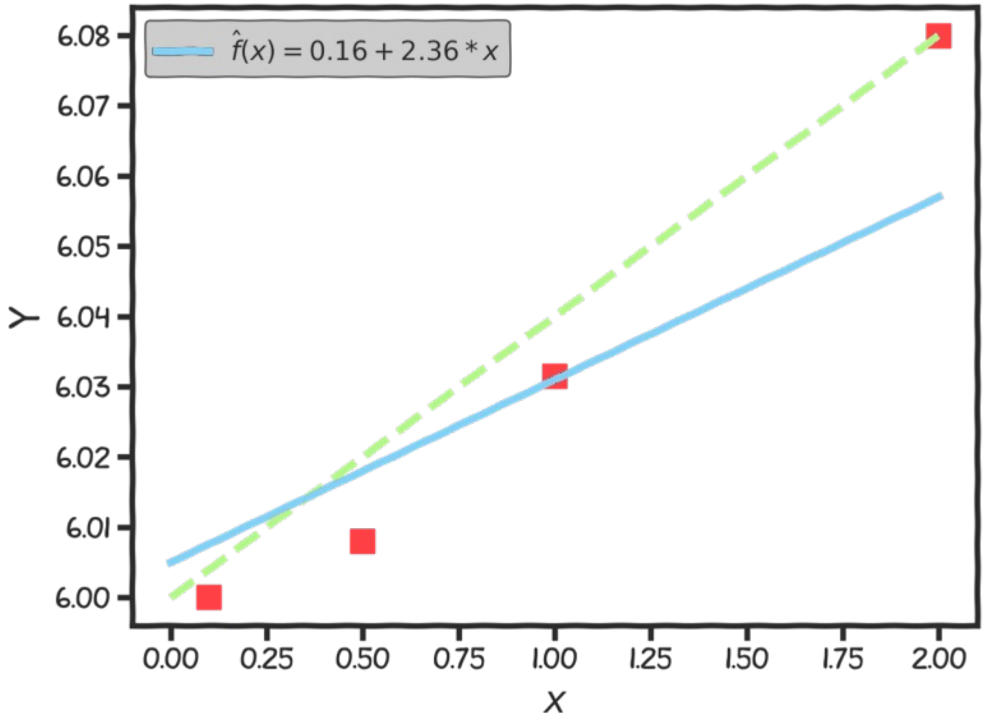
Confidence intervals for the predictors estimates (cont.)

For another "realization", we fit another model and get different values of $\hat{\beta}_0$ and $\hat{\beta}_1$.



Confidence intervals for the predictors estimates (cont.)

For another "realization", we fit another model and get different values of $\hat{\beta}_0$ and $\hat{\beta}_1$.

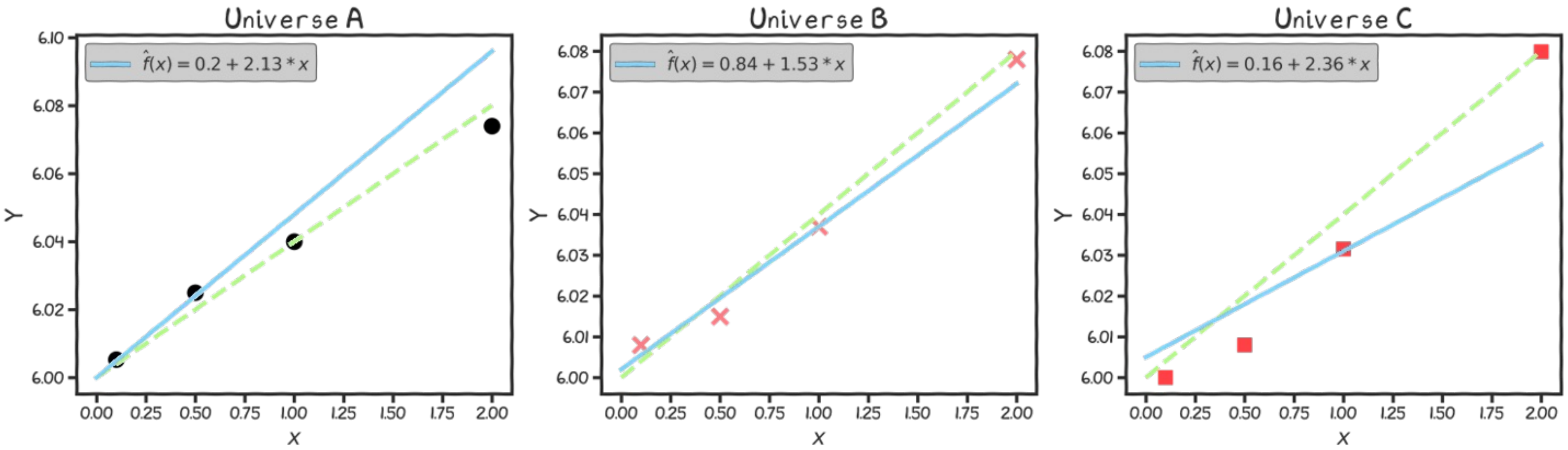


Confidence intervals for the predictors estimates (cont.)

So if we have one set of measurements of $\{X, Y\}$, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are just for that particular realization.

Question: If this is just one realization of reality, how do we know the truth? How do we deal with this conundrum?

Imagine (magic realism) we have parallel universes, and we repeat this experiment on each of the other universes.

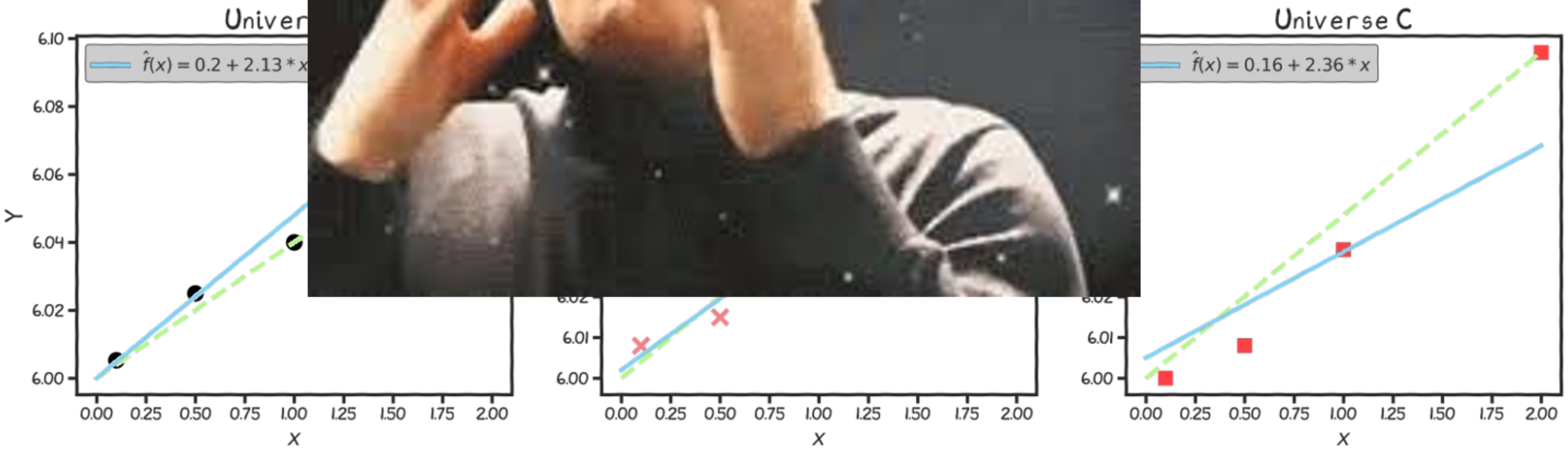


Confidence intervals for the predictors estimates (cont.)

So if we have one set of measurements of $\{X, Y\}$, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are just for that particular realization.

Question: If this is ... ow do we know the truth? How do we d...

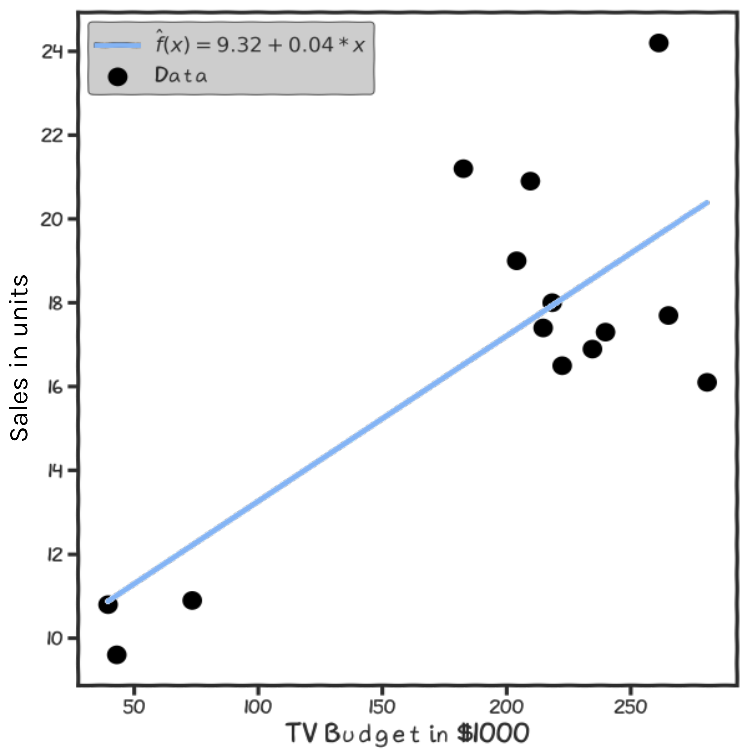
Imagine (magic rea... es, and we repeat this experiment on...

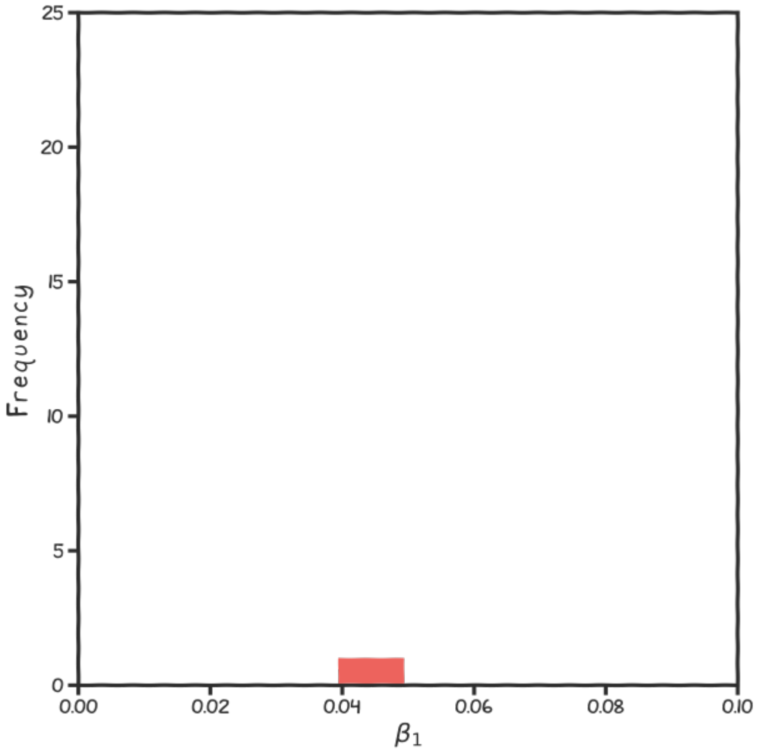


Confidence intervals for the predictors estimates (cont.)

In our magical realisms, we can now sample multiple times.

One universe, one sample, one set of estimates for $\hat{\beta}_0, \hat{\beta}_1$

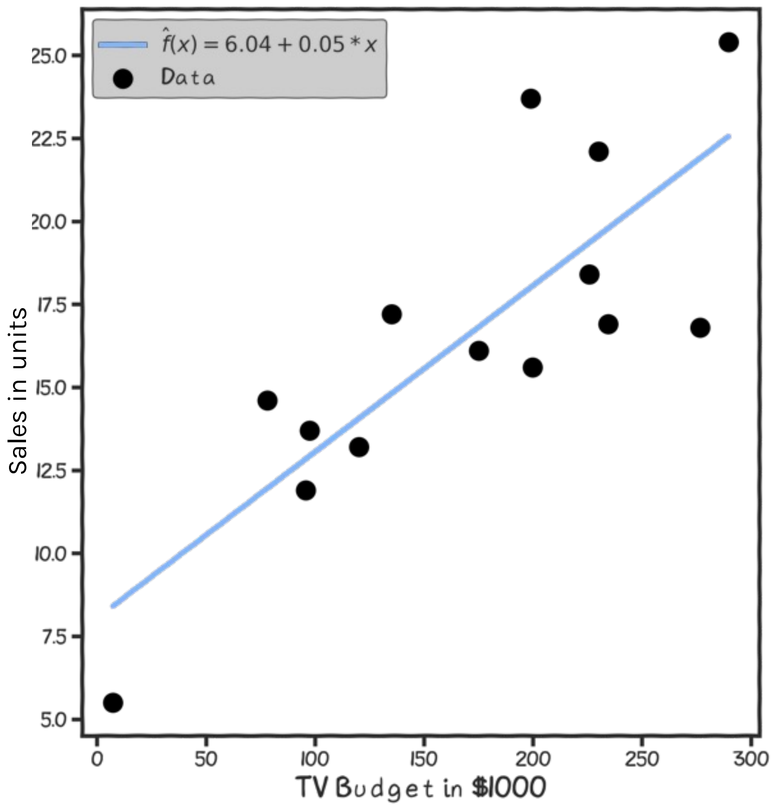


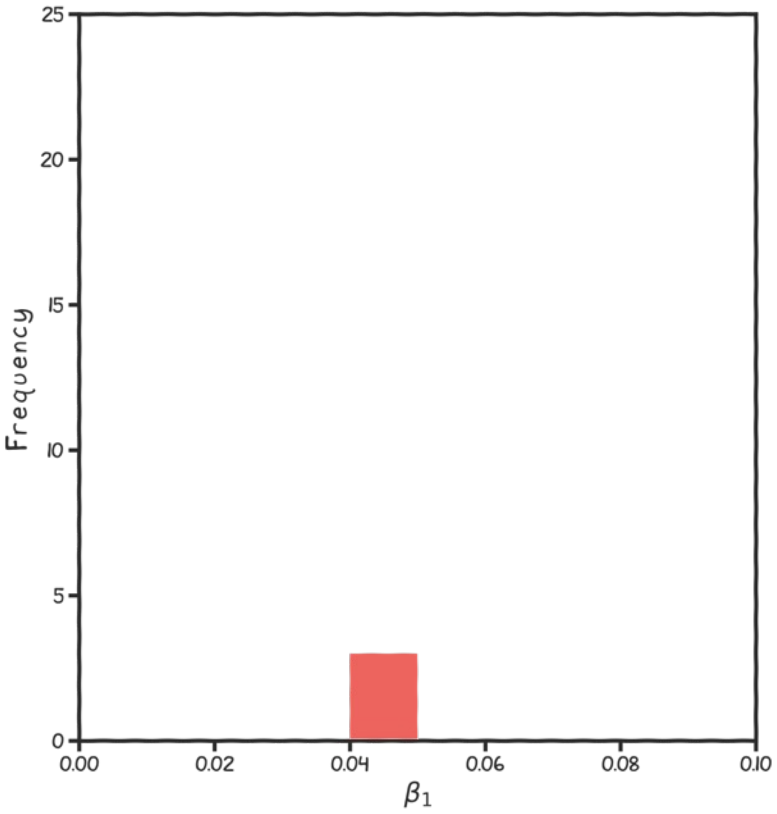


There will be an equivalent plot for $\hat{\beta}_0$ which we don't show here for simplicity

Confidence intervals for the predictors estimates (cont.)

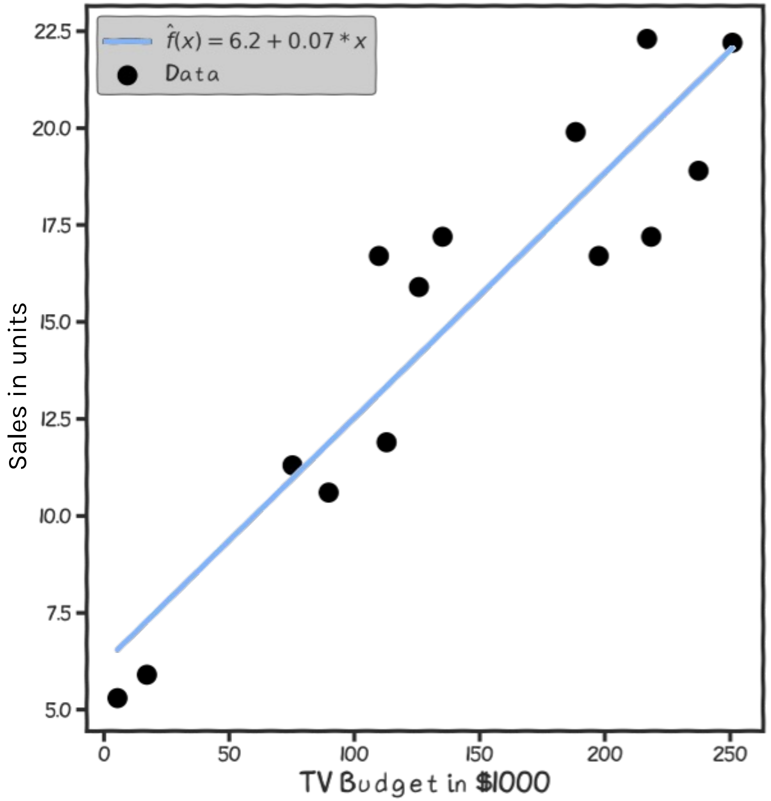
Another sample, another estimate of $\hat{\beta}_0, \hat{\beta}_1$

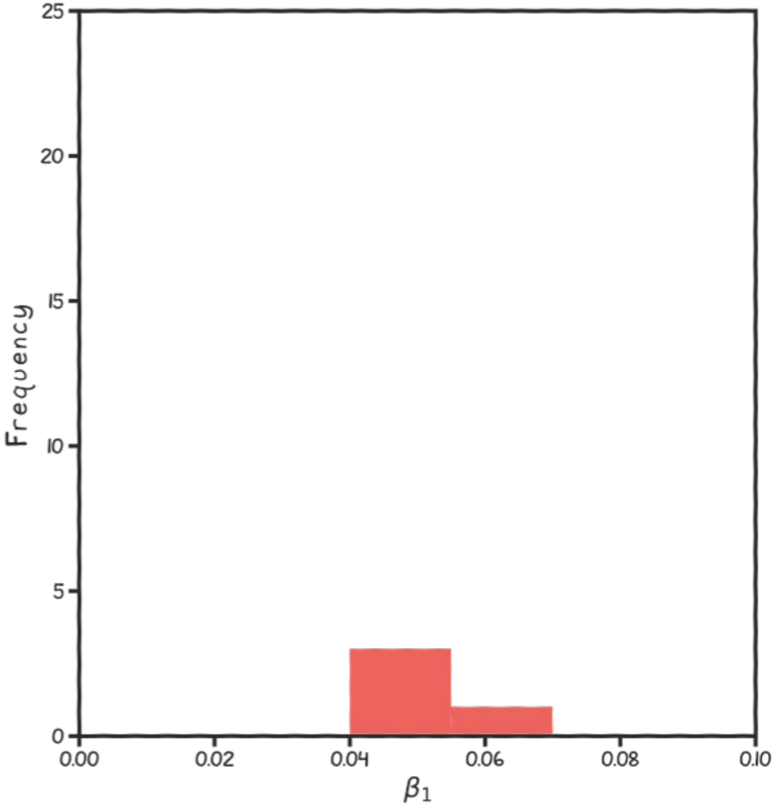




Confidence intervals for the predictors estimates (cont.)

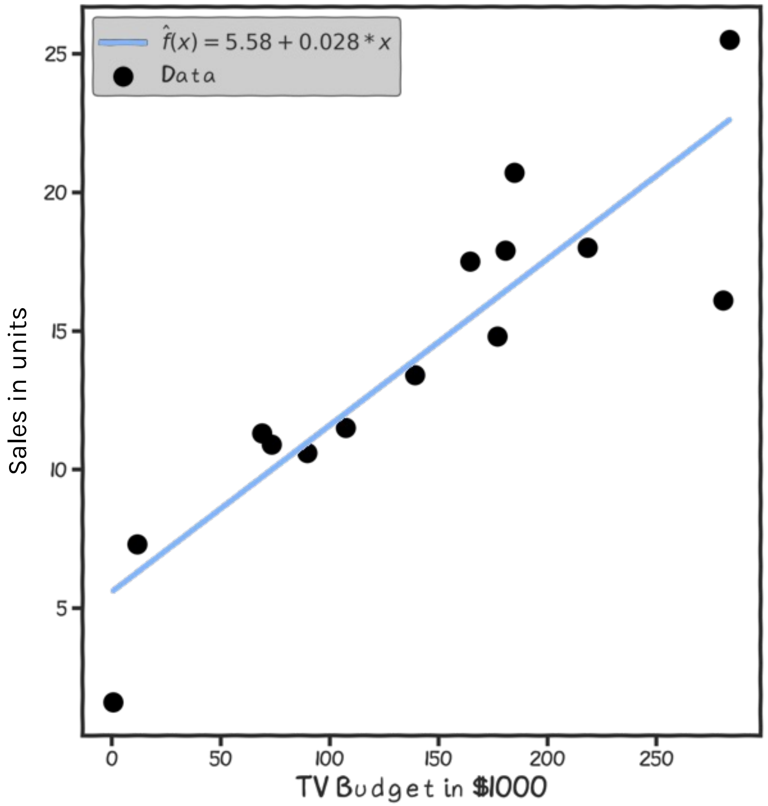
Again

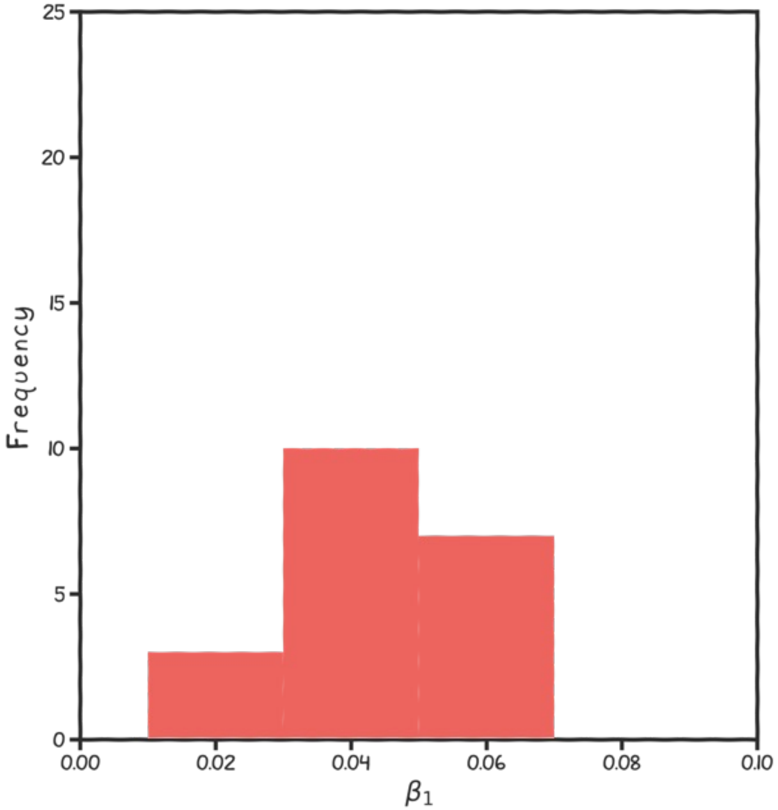




Confidence intervals for the predictors estimates (cont.)

And again





Confidence intervals for the predictors estimates (cont.)

Repeat this for 100 times, until we have enough samples of $\hat{\beta}_0, \hat{\beta}_1$.