

# **Inference in Linear Regression**

Uncertainty in estimating the linear regression coefficients

Pavlos Protopapas, Ignacio Becker

# Summary so far

---

- Statistical model
- **k-nearest neighbors (kNN)**
- Model fitness and model comparison (MSE)
- Goodness of fit ( $R^2$ )
- **Linear Regression, multi-linear regression and polynomial regression**
- Model selection using validation and cross validation
- One-hot encoding for categorical variables
- What is overfitting
- Ridge and Lasso regressio

# 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 a $f(x)$ 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 evaluate the model and understand what the data tells us

# 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:

$$y = 1.01x + 120$$

Where  $y$  is the sales in 1000\$,  $x$  is the TV budget.

**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?

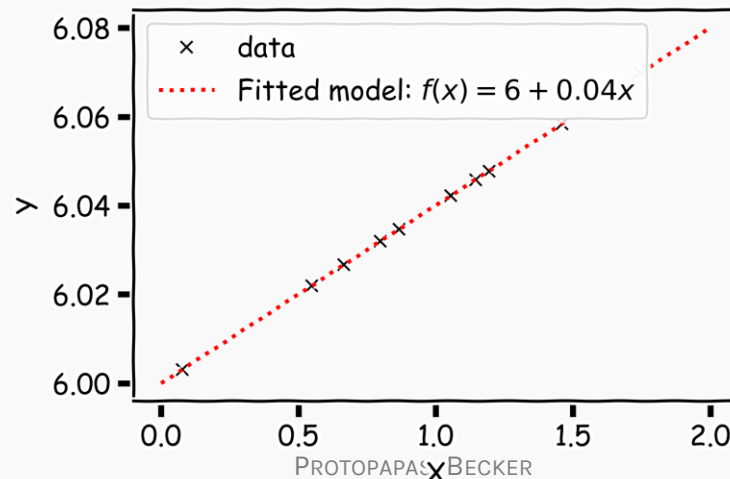
# Confidence intervals for the predictors estimates

We interpret the  $\varepsilon$  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**  $\varepsilon$ .

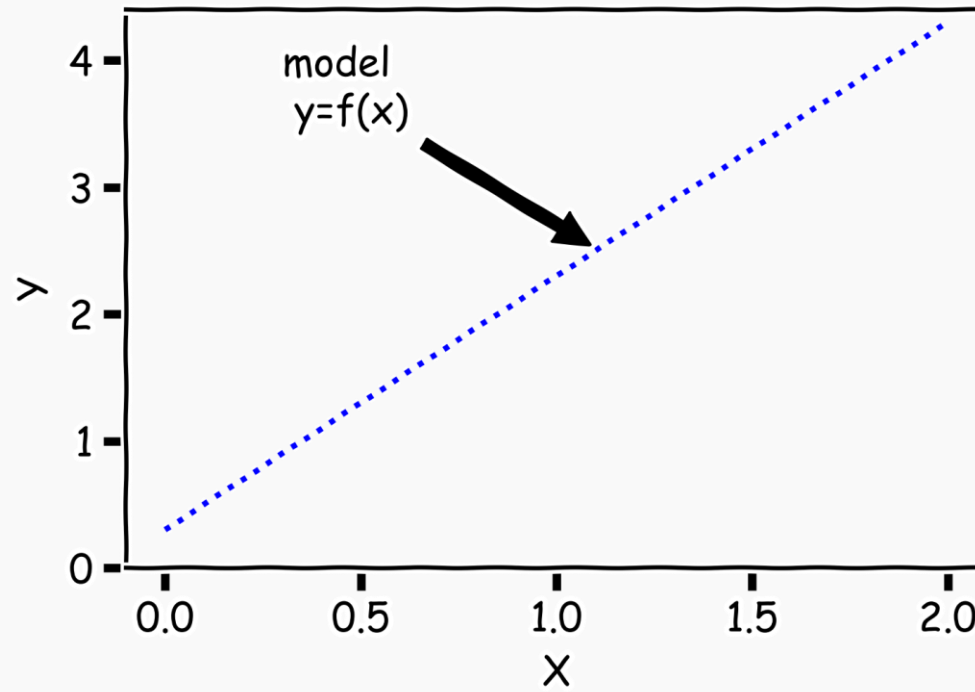
---

Because of  $\varepsilon$ , **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)

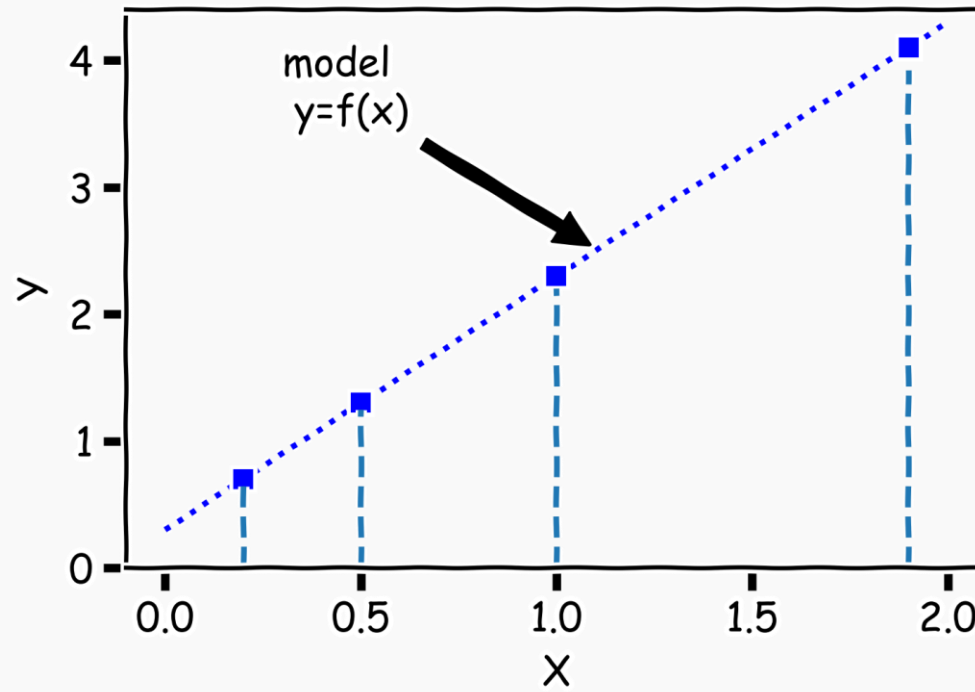
Start with a model  $f(X)$ , the correct relationship between input and outcome.



# Confidence intervals for the predictors estimates (cont)

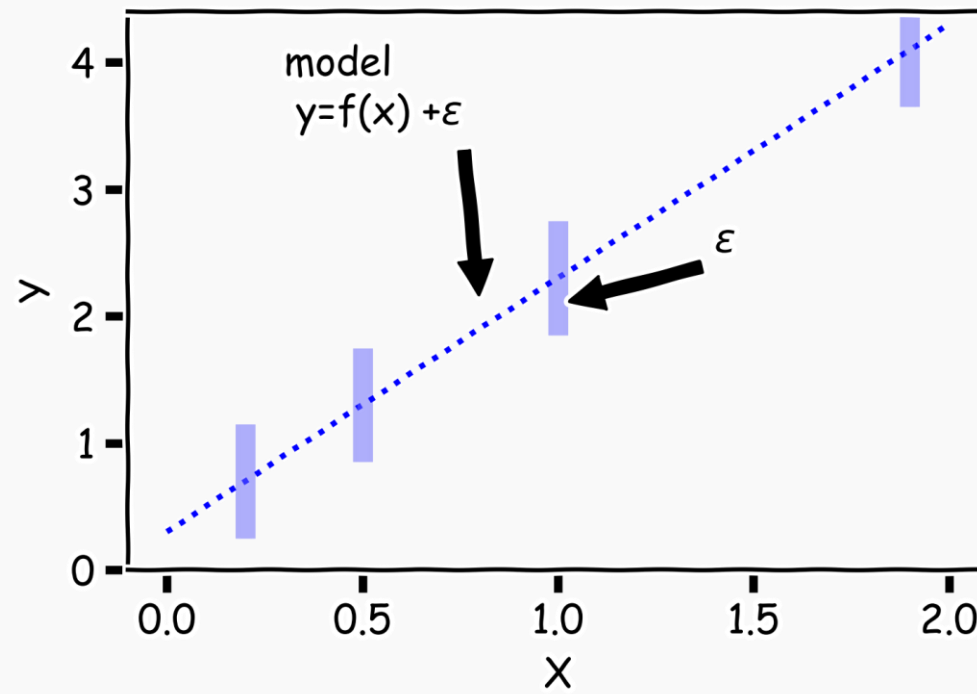


For some values of  $X^*$ ,  $Y^* = f(X^*)$



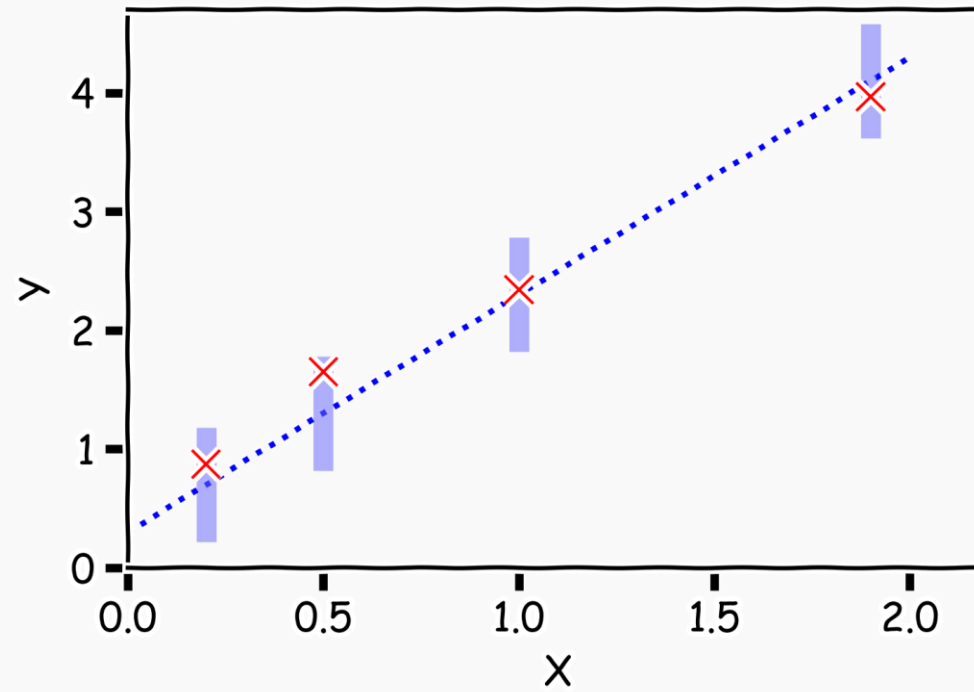
# Confidence intervals for the predictors estimates (cont)

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



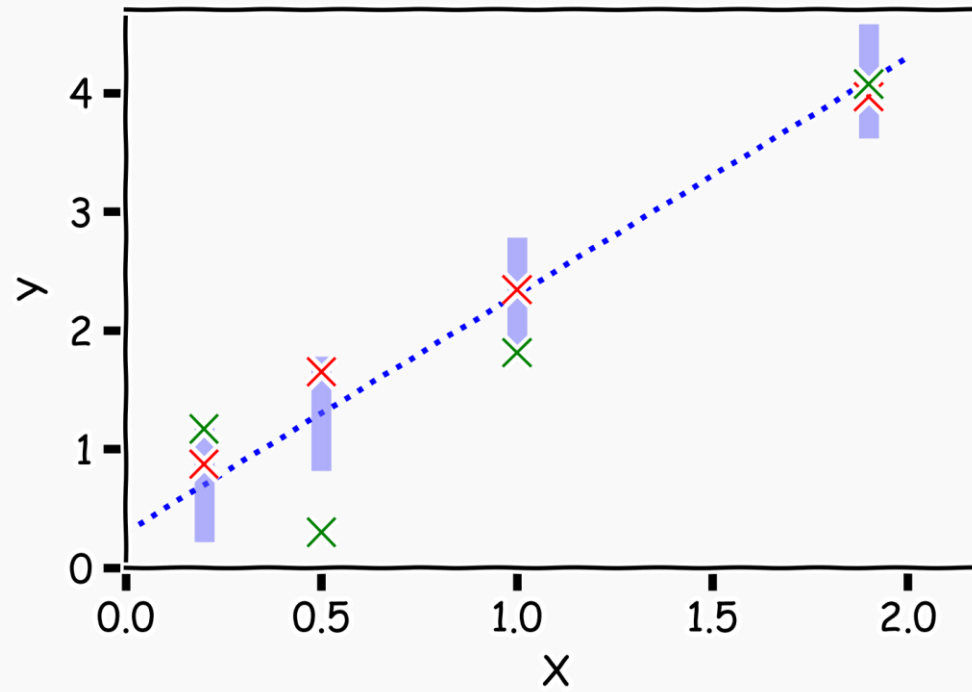
# Confidence intervals for the predictors estimates (cont)

One set of observations, “one realization” yields one set of  $Y$ s (red crosses).



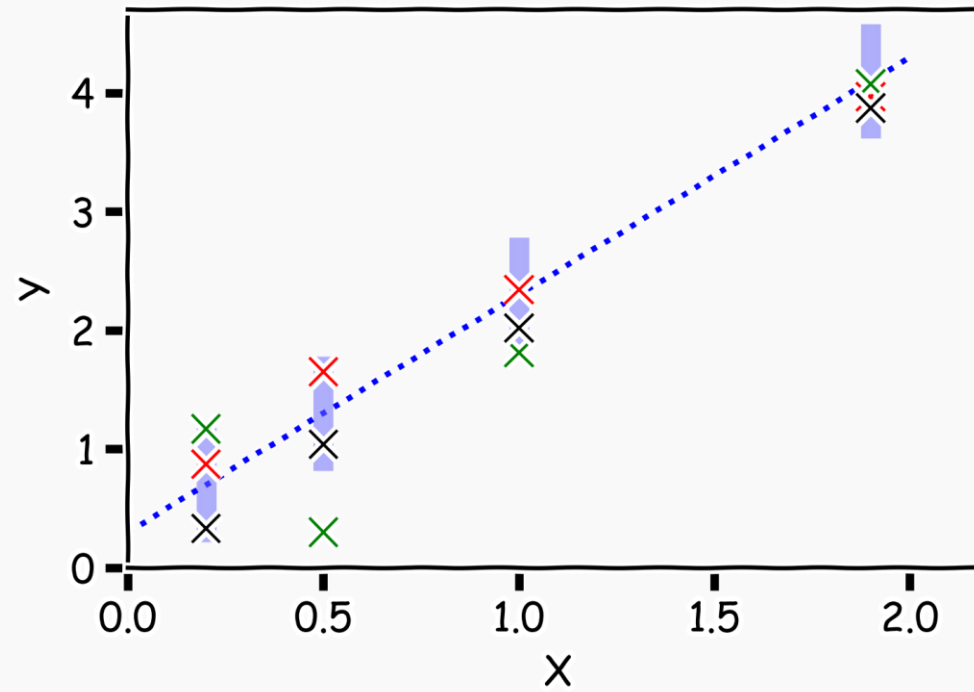
# Confidence intervals for the predictors estimates (cont)

Another set of observations, “another realization” yields another set of  $Y$ s (green crosses).



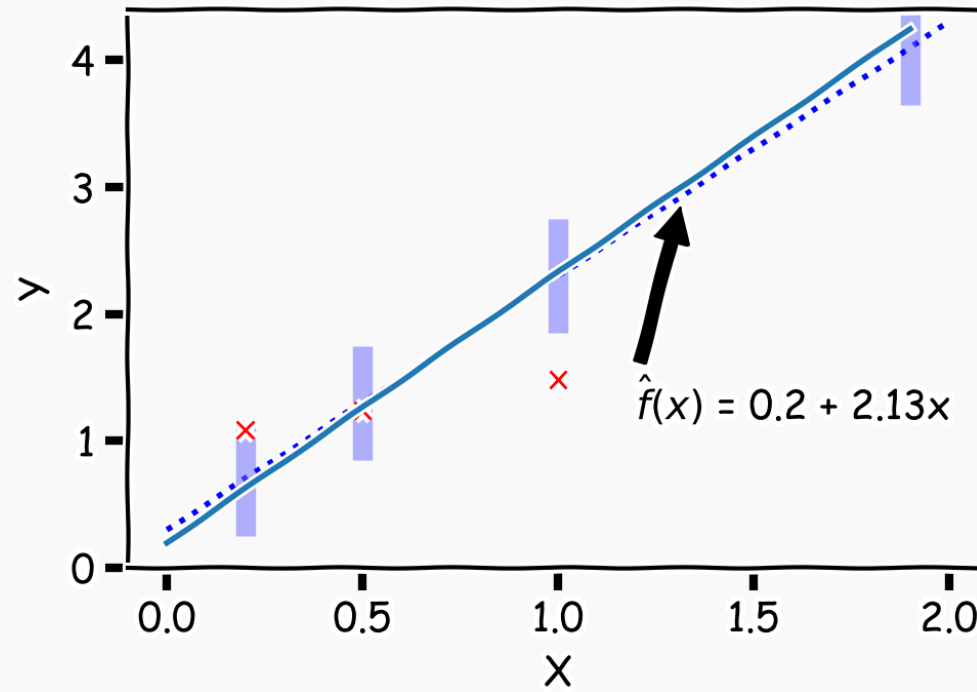
# Confidence intervals for the predictors estimates (cont)

Another set of observations, “another realization”, another set of  $Y$ s (black crosses).



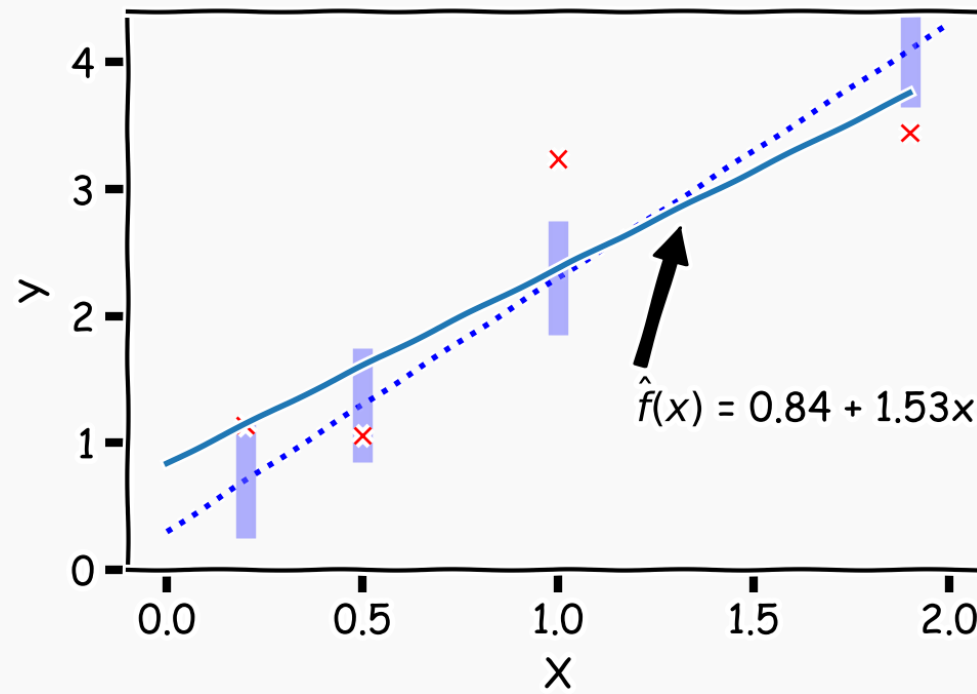
# Confidence intervals for the predictors estimates (cont)

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)

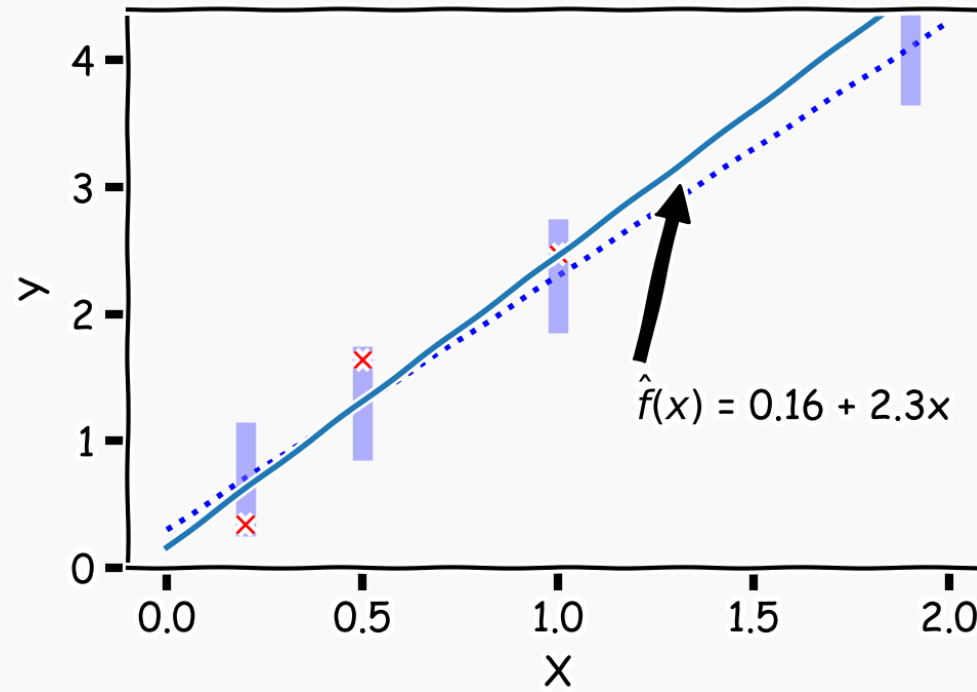
For another “realization”, we fit another model and get different values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .





# Confidence intervals for the predictors estimates (cont)

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



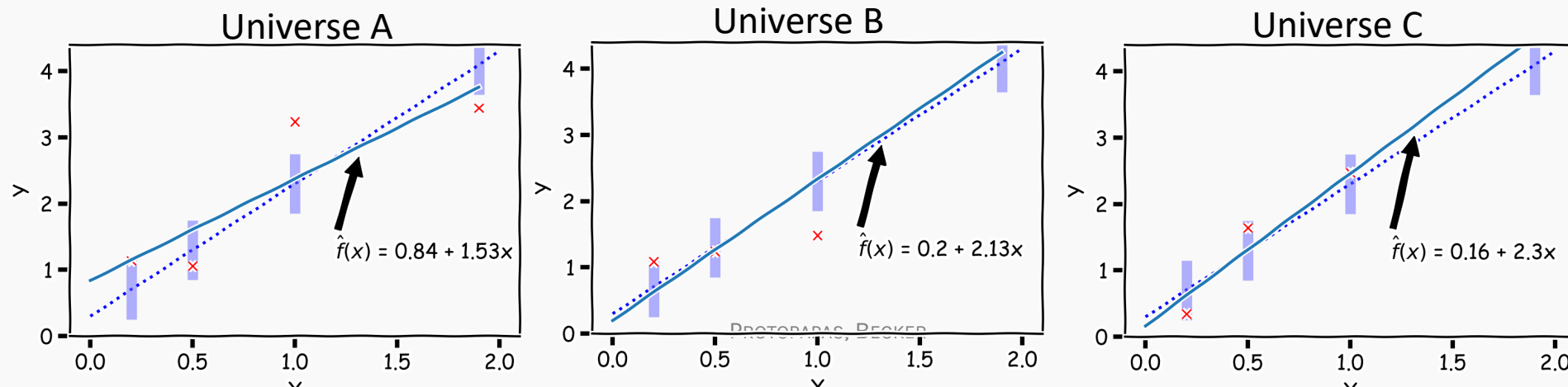
# Confidence intervals for the predictors estimates (cont)



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

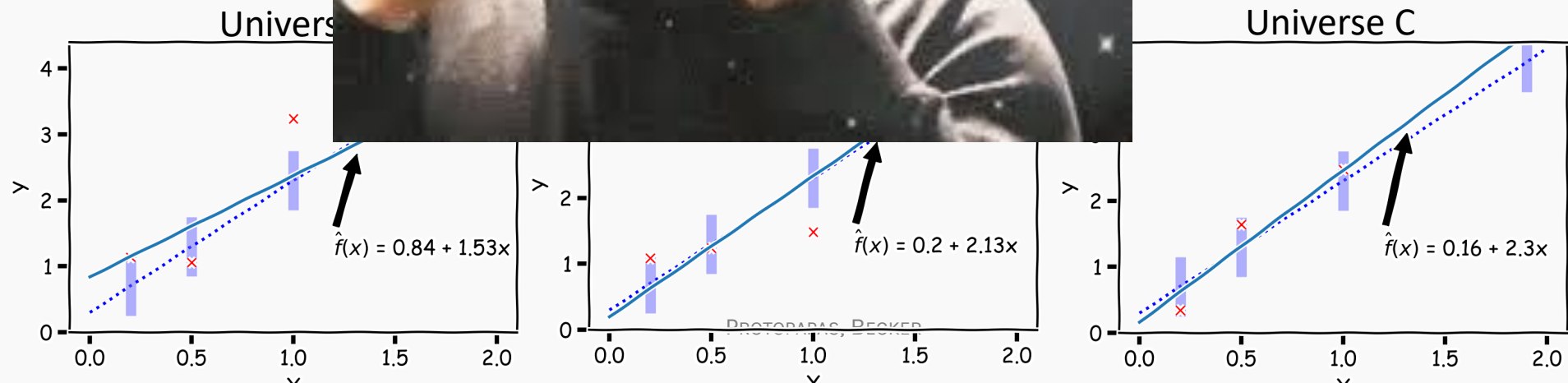


# Confidence intervals for the predictors estimates (cont)

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

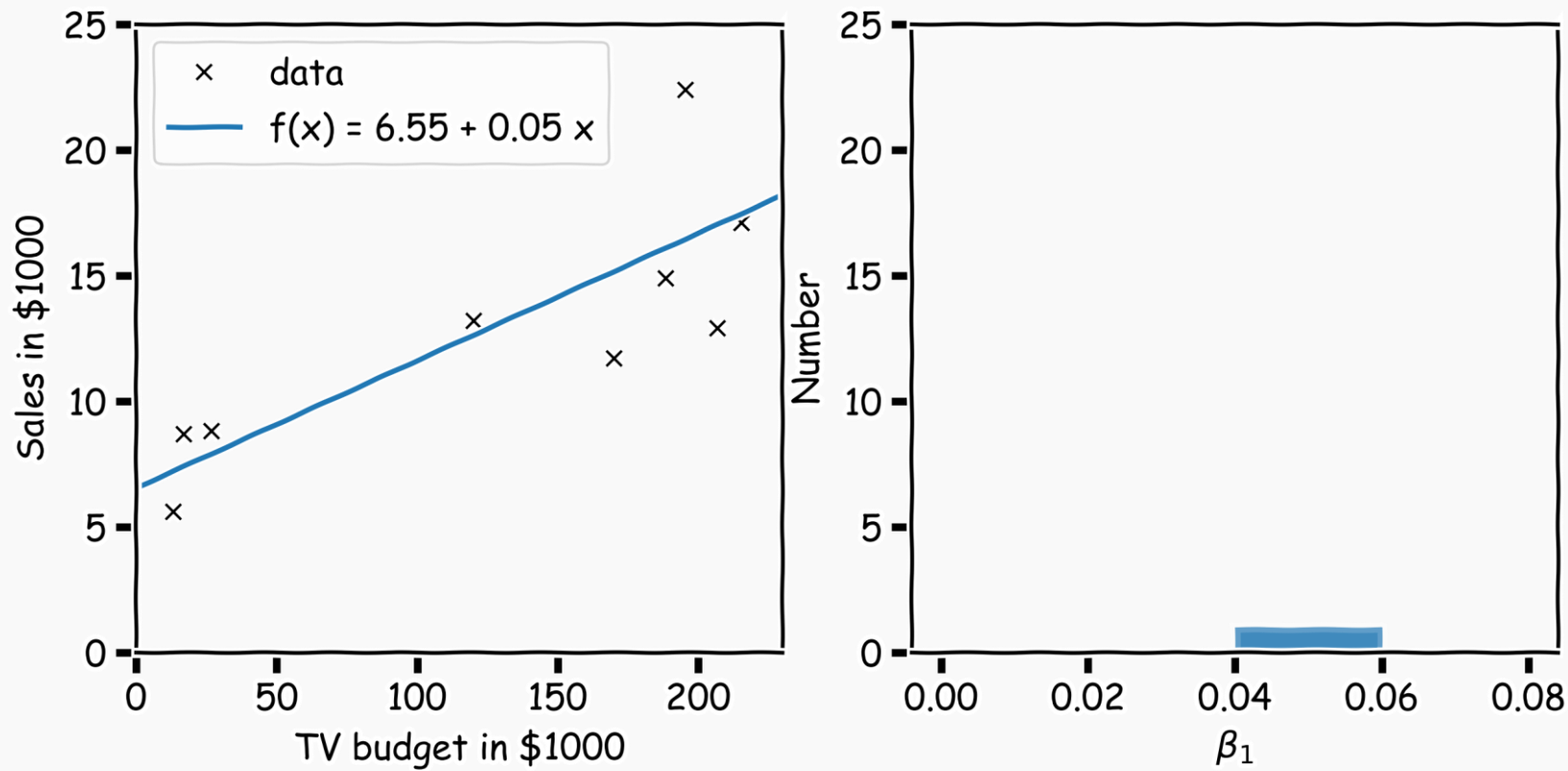
**Question:** If this is the only realization, how do we know the truth? How do we

**Imagine** (magic machine) that generates many realizations, and we repeat this experiment on each



# Confidence intervals for the predictors estimates (cont)

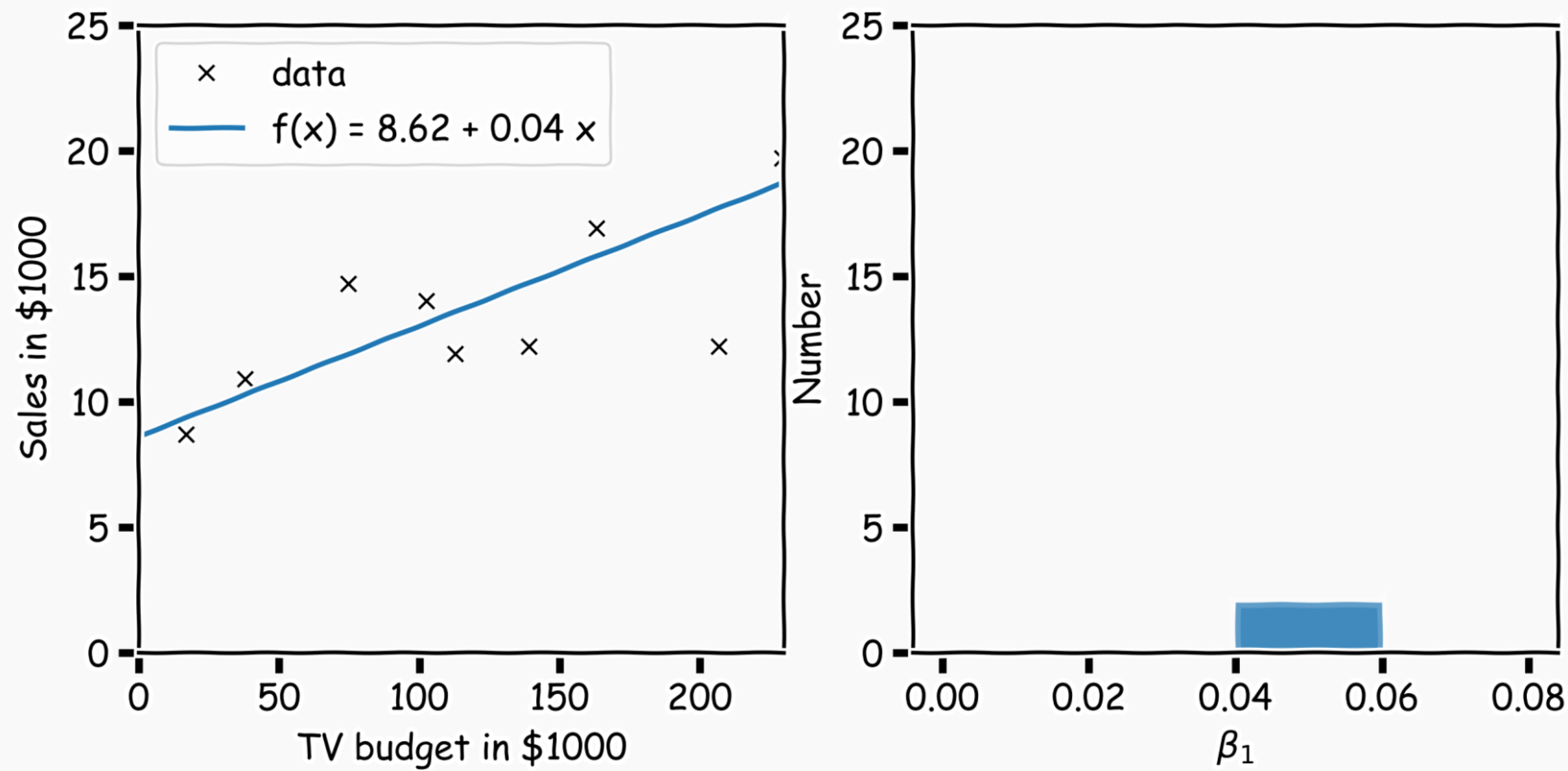
In our magical realisms, we can now sample multiple times. One universe, one sample, one set of estimates for  $\hat{\beta}_0, \hat{\beta}_1$



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

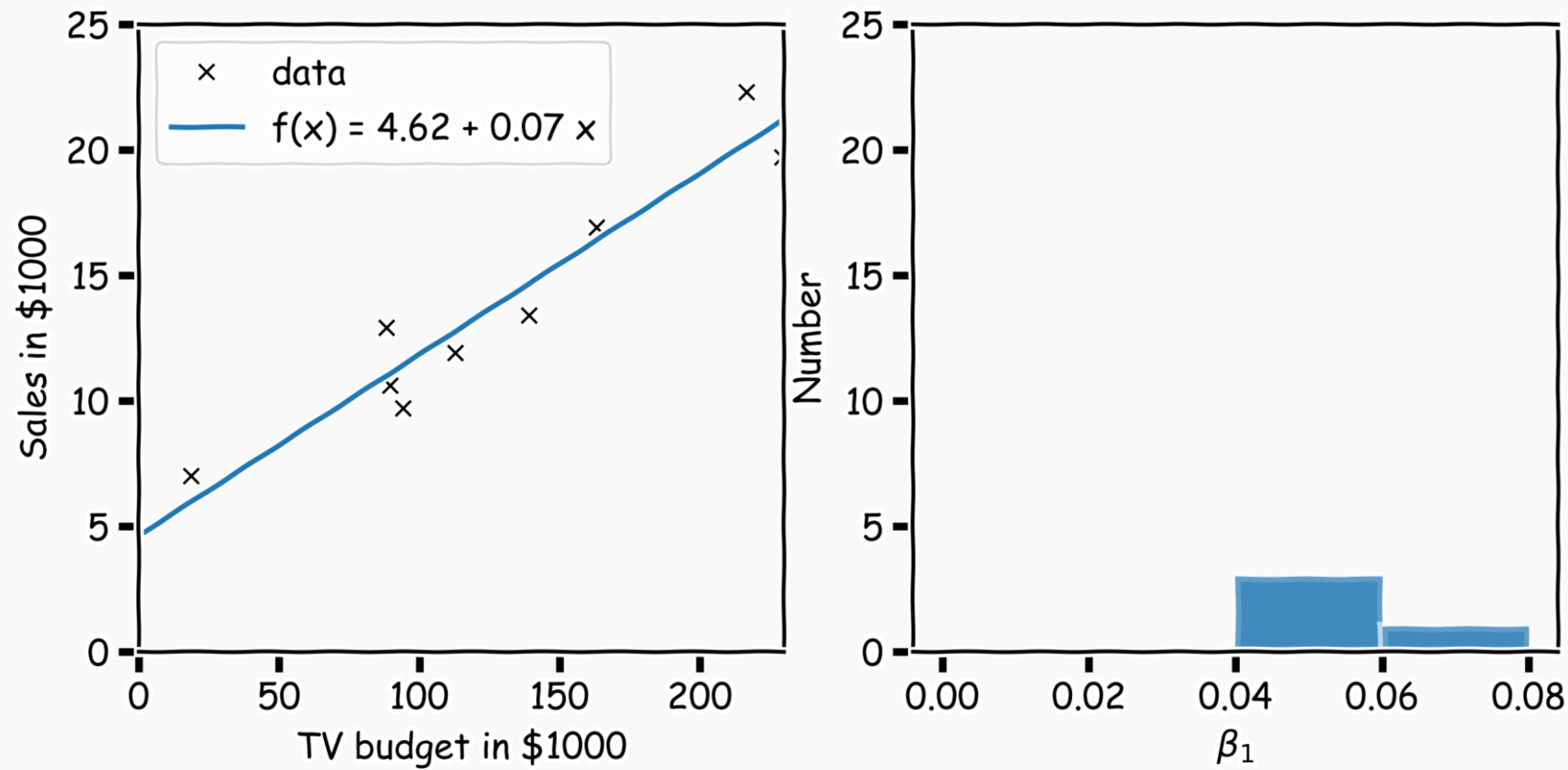
# Confidence intervals for the predictors estimates (cont)

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



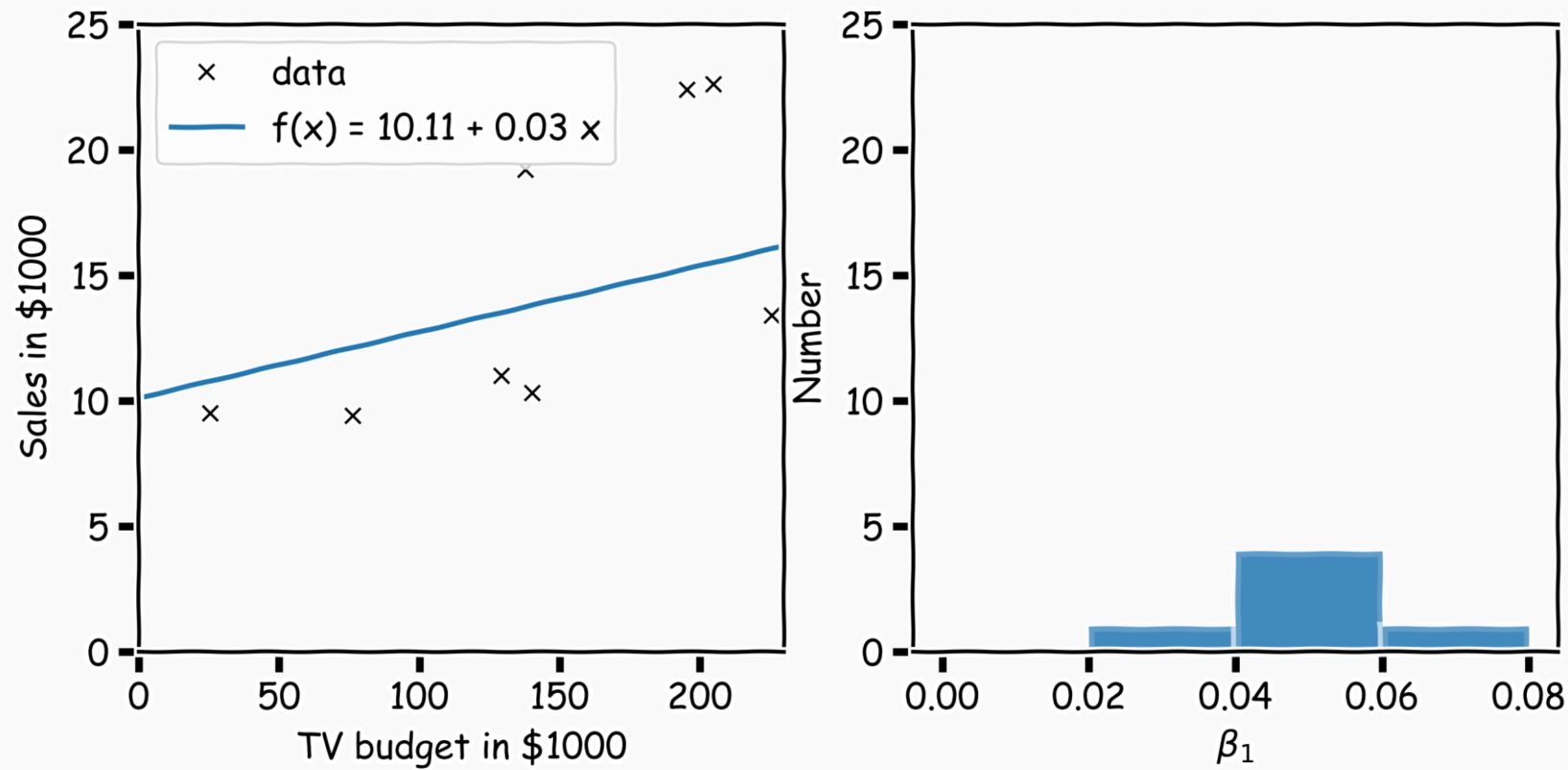
# Confidence intervals for the predictors estimates (cont)

Again



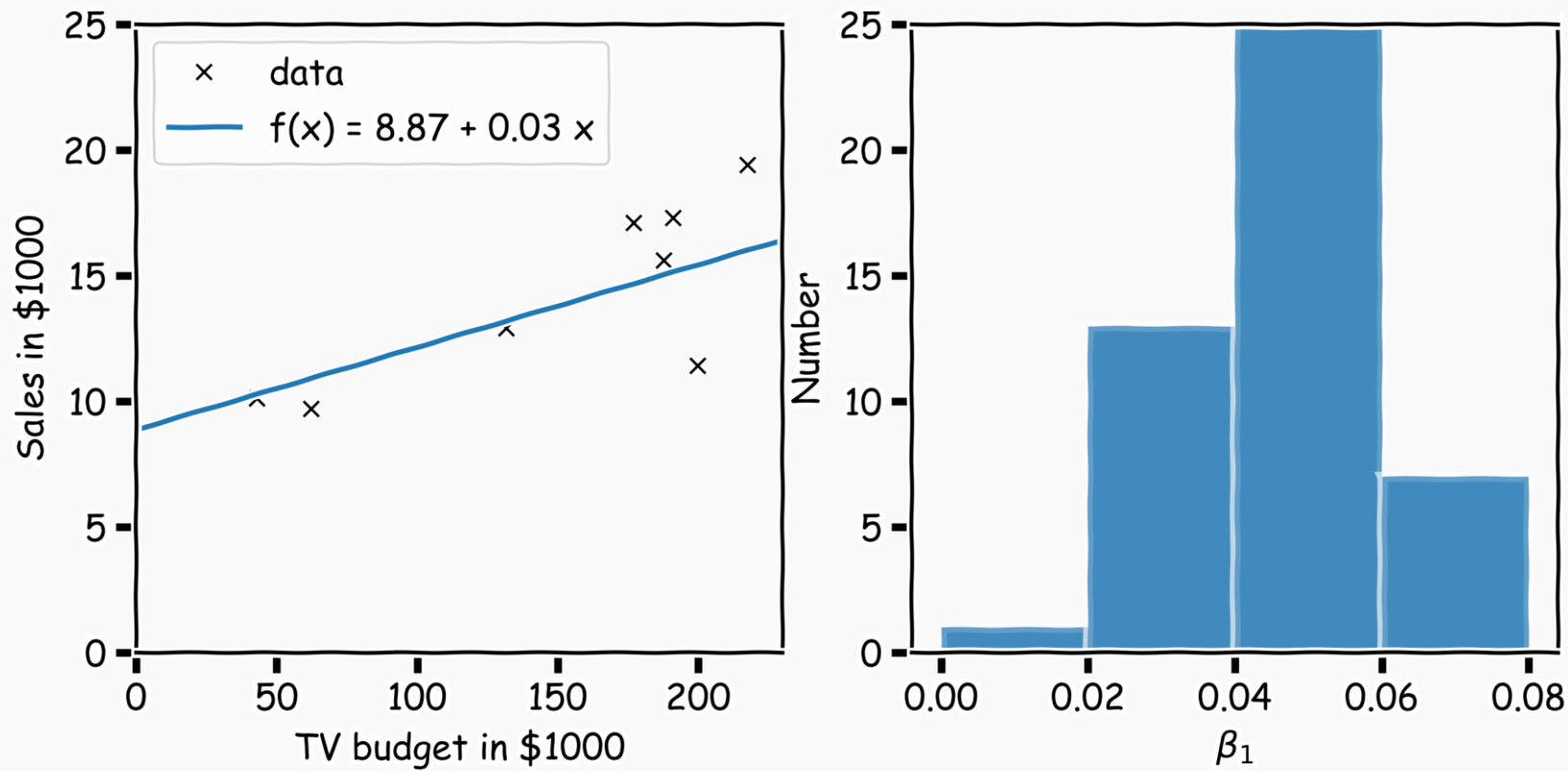
# Confidence intervals for the predictors estimates (cont)

And again

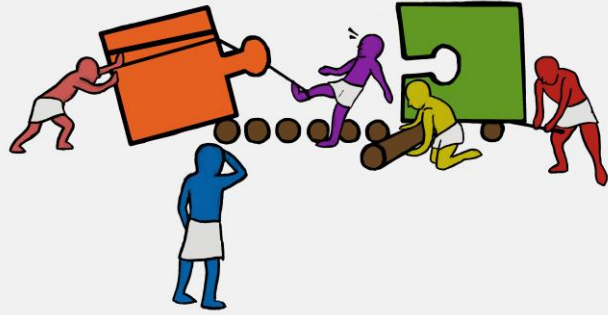


# Confidence intervals for the predictors estimates (cont)

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

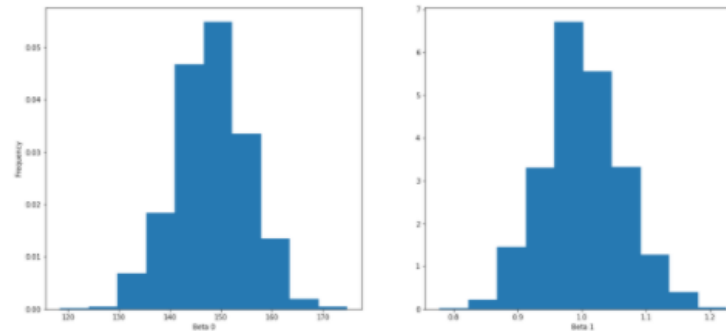






## 🧑 Exercise: Beta Values for Data from Random Universe

Given a `RandomUniverse(dataframe) -> dataframe` function that gives a new dataset from a "parallel" universe, calculate the  $\beta_0$ 's and  $\beta_1$ 's and plot a histogram like the one below.



### Instructions:

- Get a new dataframe using the RandomUniverse function already provided in the exercise
- Calculate  $\beta_0$ ,  $\beta_1$  for that particular dataframe
- Add the calculated  $\beta_0$  and  $\beta_1$  values to a python list
- Plot a histogram using the lists calculated above

