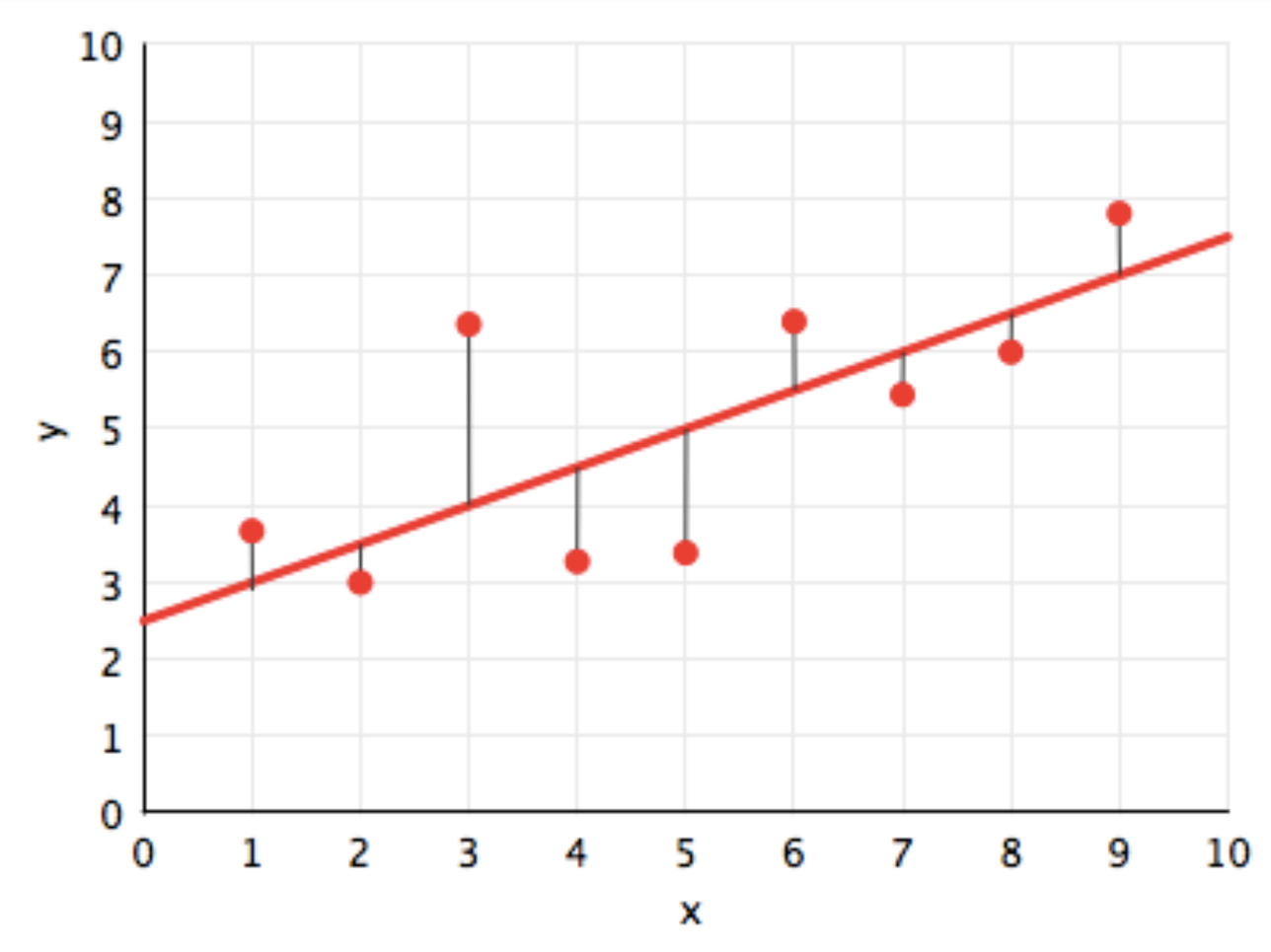
Linear Regression

What is Regression?

- how many dollars will you spend?
- what is your creditworthiness
- how many people will vote for Bernie t days before election
- use to predict probabilities for classification
- causal modeling in econometrics



$X = X_1, \ldots, X_p$

$X_j = x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}$

predictors

features

covariates

$Y = y_1, \ldots, y_n$

outcome

response variable

dependent variable

n observations

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

True vs Statistical Model

We will assume that the measured response variable, y, relates to the predictors, x, through some unknown function expressed generally as:

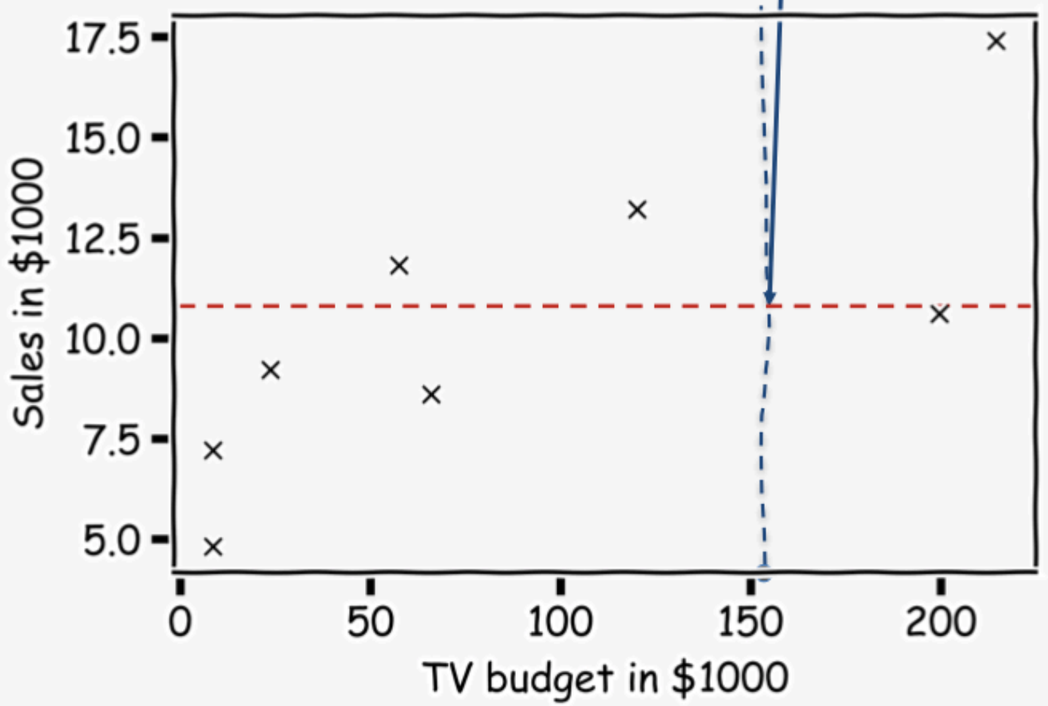
$$y = f(x) + \epsilon$$

Here, f is the unknown function expressing an underlying rule for relating y to x, and ϵ is a random amount (unrelated to x) that y differs from the rule $f(x)$.

In real life we never know the true generating model $f(x)$

Possibly the simplest model: the mean

Simple idea is to take the mean of all y's, $\hat{f}(x) = \frac{1}{n}\sum_1^n y_i$



The next simplest: fit a straight line

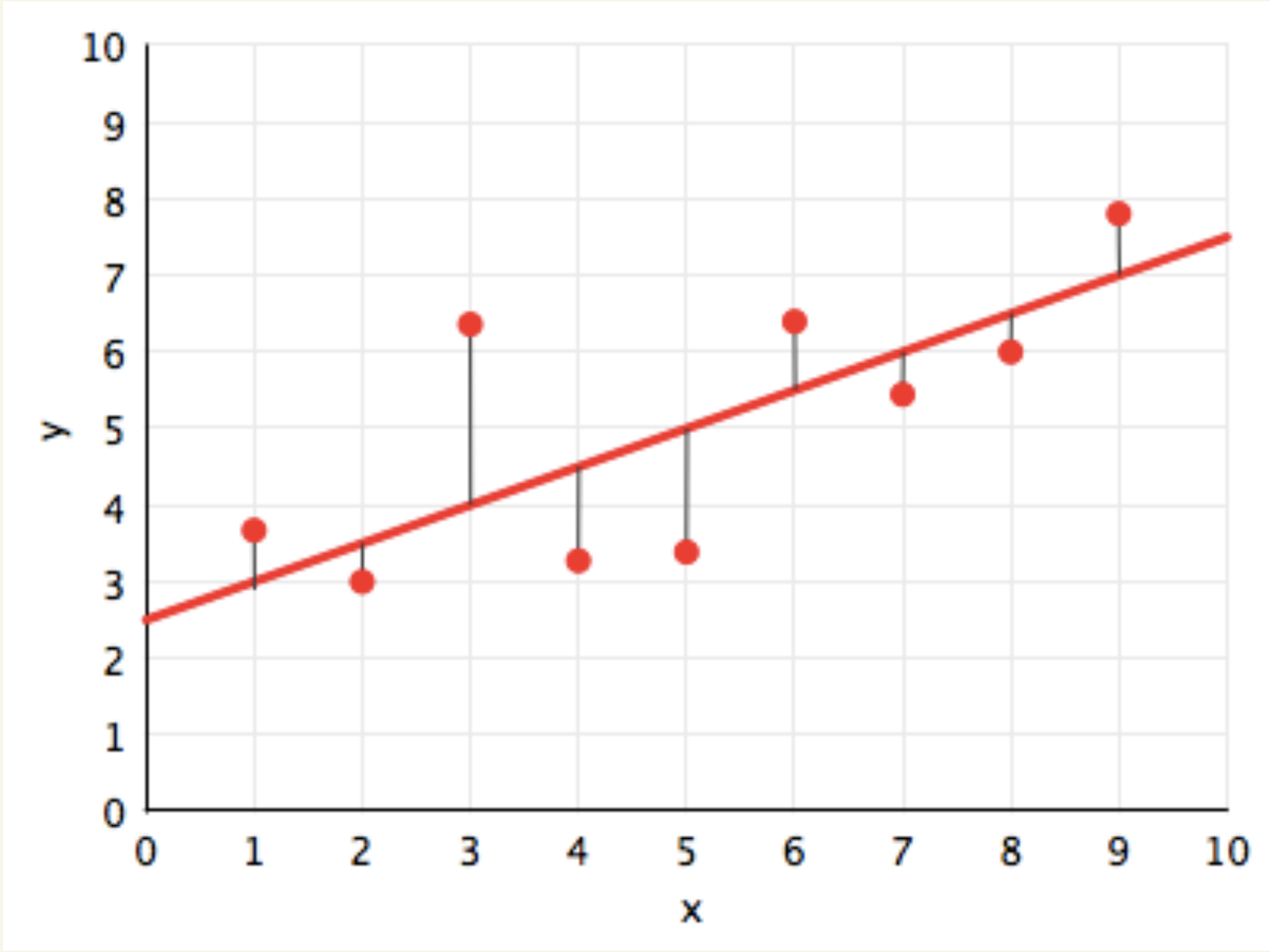
$$\hat{f}(x) = a + bx$$

How? Use the **Mean Squared Error**:

$$MSE = \frac{1}{N}\sum_i (\hat{f}(x_i) - y_i)^2$$

$$= \frac{1}{N}\sum_i (a + bx_i - y_i)^2$$

Minimize this with respect to the *parameters*. (Here the intercept and slope)

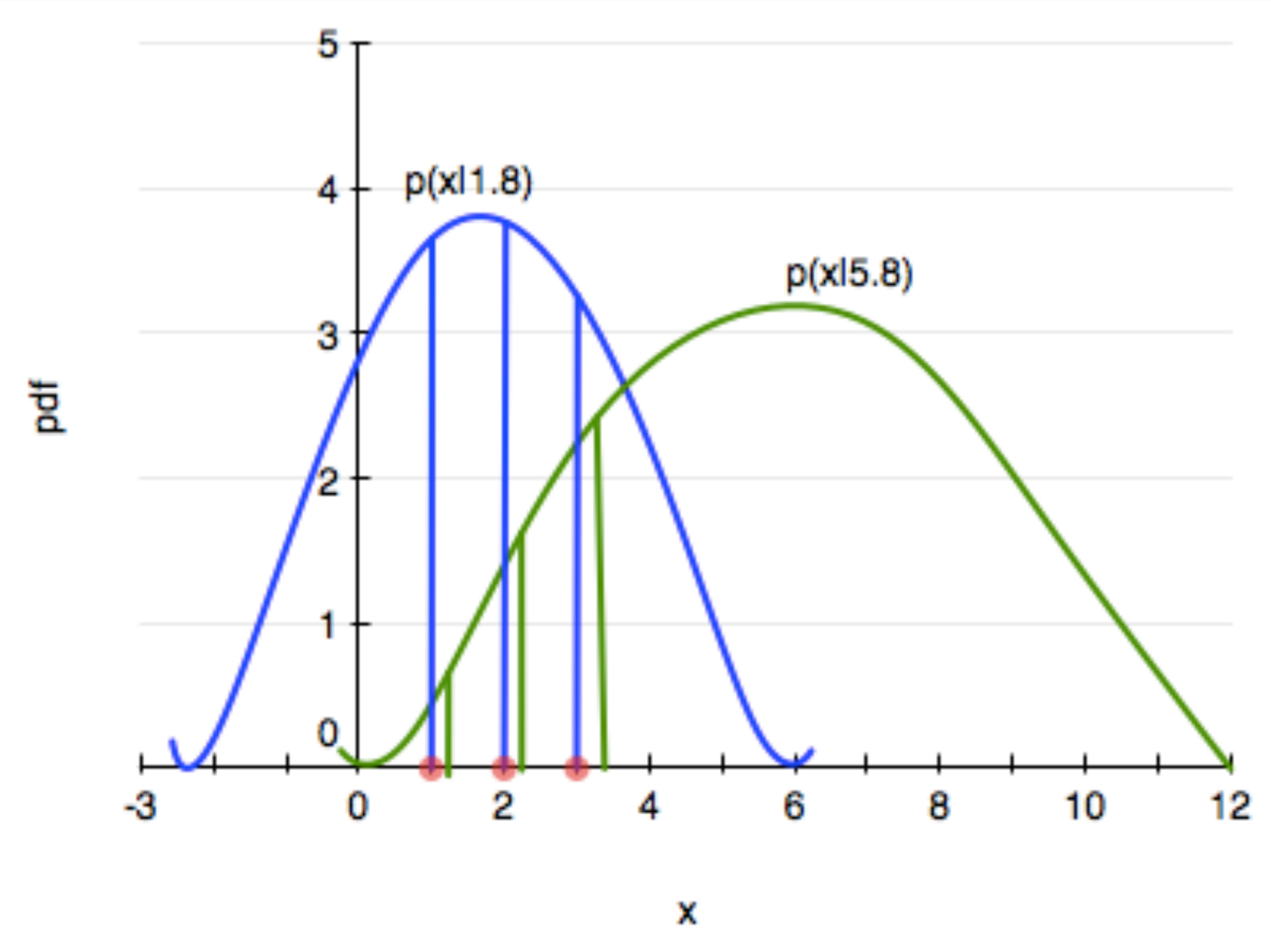


The probabilistic interpretaton: Use MLE

How likely it is to observe values $x_1, \ldots, x_n$ given the parameters λ?

$$L(\lambda) = \prod_{i=1}^{n} P(x_i|\lambda)$$

How likely are the observations if the model is true? Maximize this



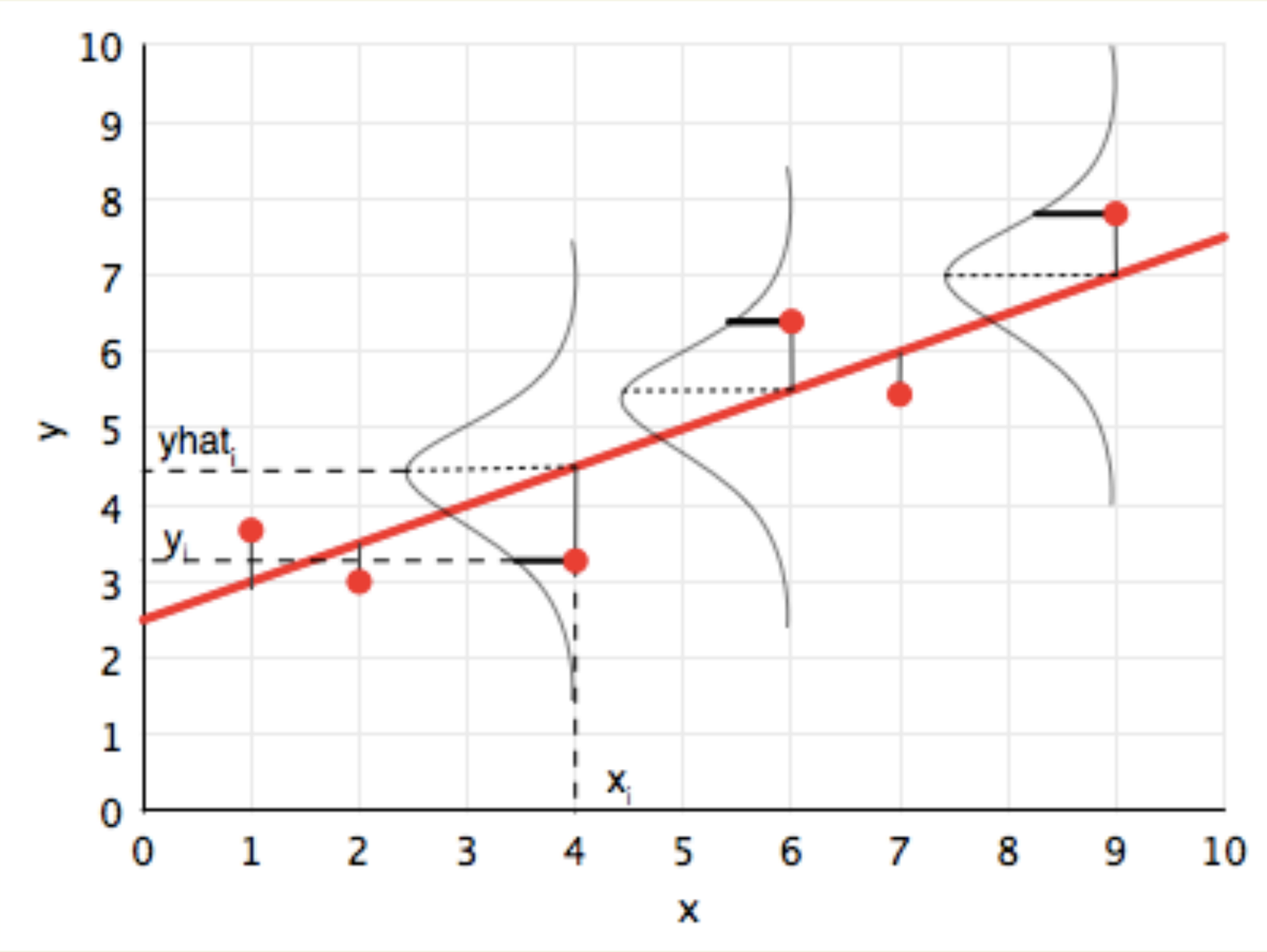
MLE + Gaussian Distribution

Each y_i is gaussian distributed with mean $a + bx_i$ (the value of the regression line) and variance σ^2:

$y_i \sim N(\hat{f}(x_i), \sigma^2)$, where

$$N(\hat{f}(x), \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\hat{f}(x))^2/2\sigma^2}, \text{ and}$$

$$\hat{f}(x) = a + bx$$



We can then write the likelihood:

$$\mathcal{L} = p(\{y\}|\{x\}, a, b, \sigma) = \prod_i p(y_i|x_i, a, b, \sigma)$$

$\mathcal{L} = (2\pi\sigma^2)^{(-n/2)} e^{\frac{-1}{2\sigma^2}\sum_i (y_i-(a+bx_i))^2}$. The log likelihood ℓ then is given by:

$$\ell = \frac{-n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - (a + bx_i))^2.$$

Maximize it! Does it look familiar? Looks like a negative distance!

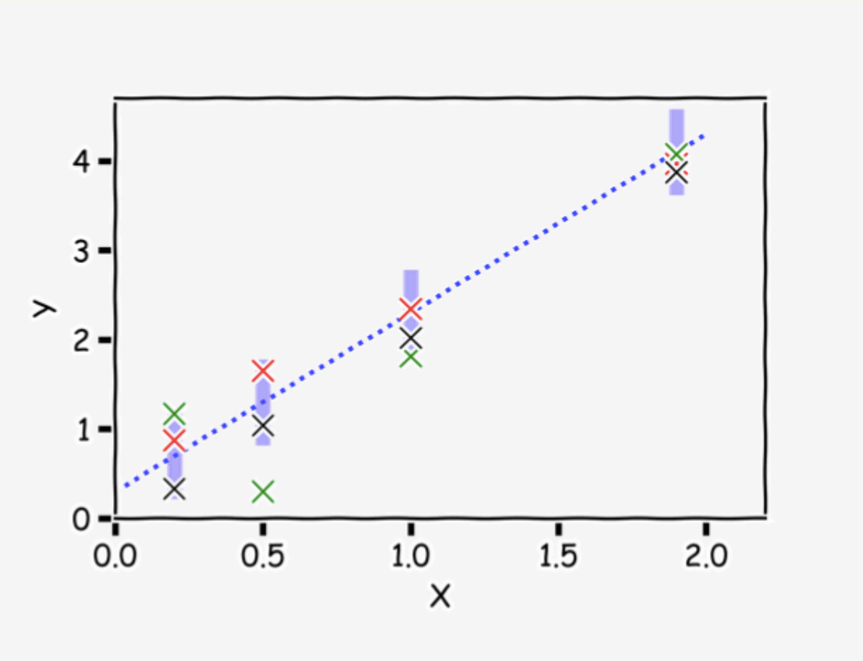
Regression Noise Sources

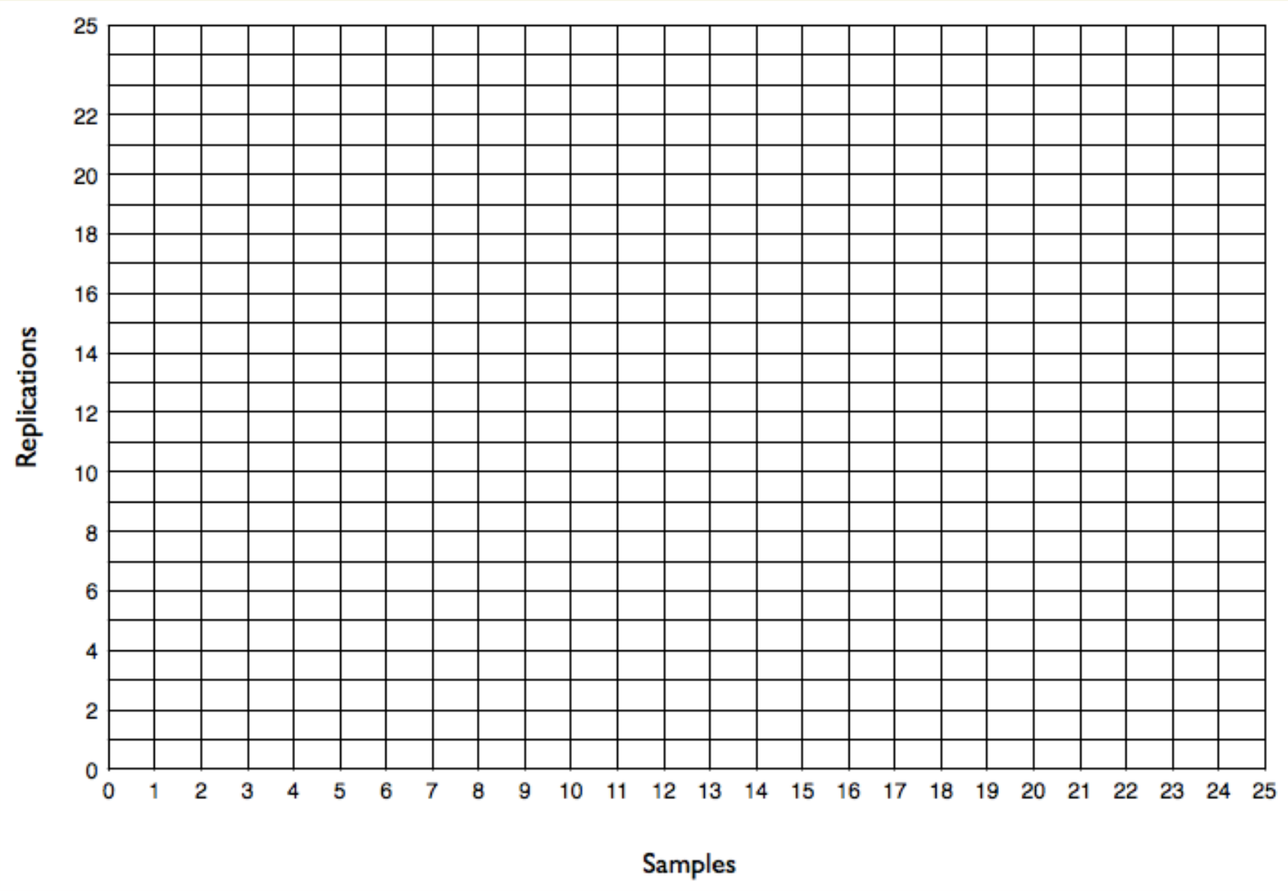
- Specification Error: lack of knowledge of the true generating process (we cant distinguish this from the others)
- The irreducible error ϵ: This is the Normal Distribution!. Applies even on the population as well and this error does not go away. Because of ϵ, every time we measure the response y for a fixed value of x we will obtain a different observation based on the Normal distribution: and this will give different values of slope and intercept.
- sampling

Sampling: Magic Realism

Why have we been using all these hats? Remember we have only one sample.

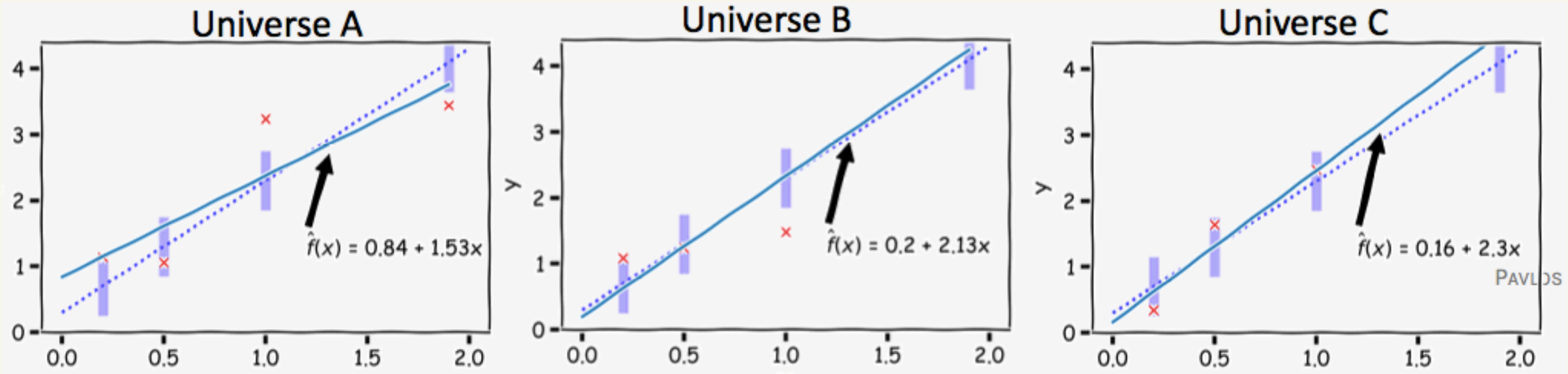
Now, imagine that God gives you some M data sets **drawn** from the population. This is a hallucination, a magic realism..



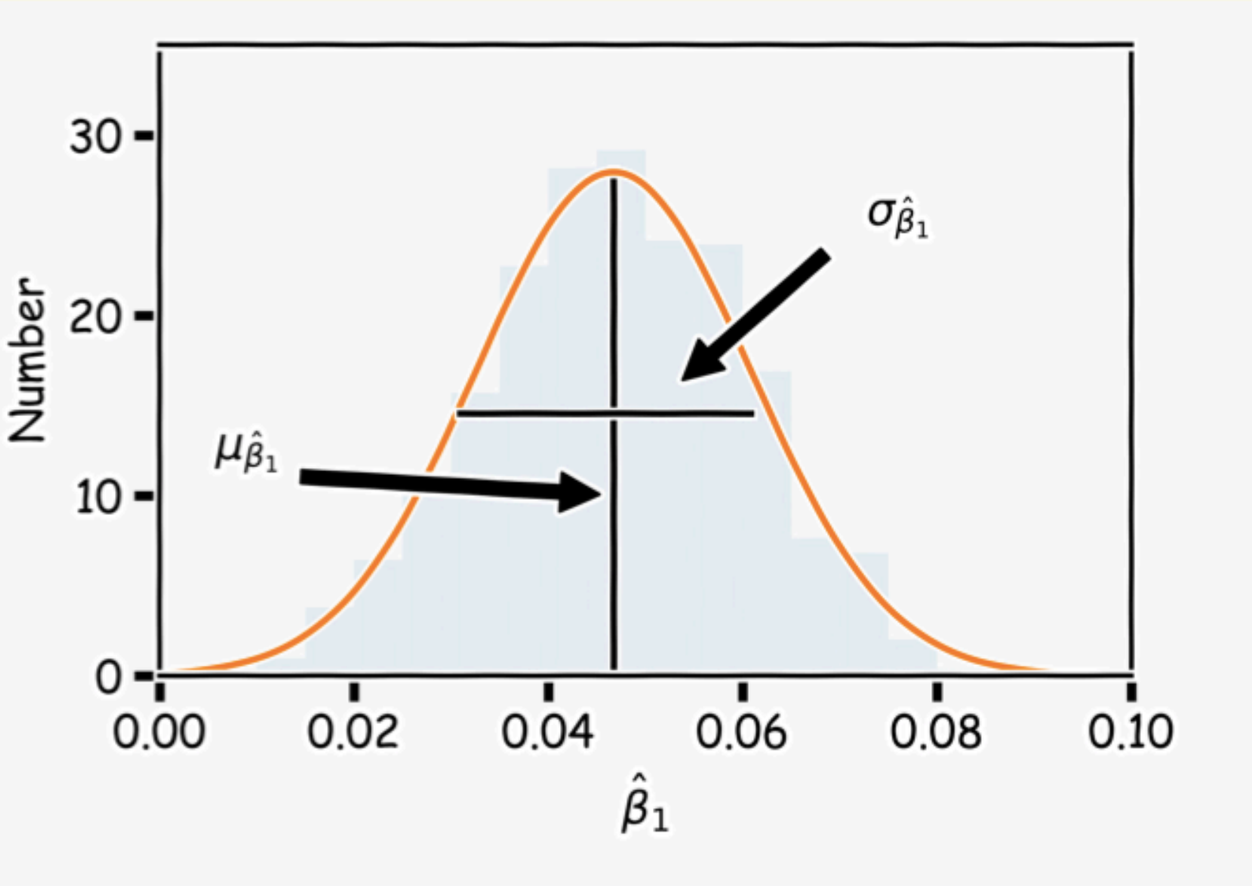


Multiple Fits

..and you can now find the regression on each such dataset (because you fit for the slope and intercept). So, we'd have M estimates of the slope and intercept.



Sampling Distributions of parameters



As we let $M \to \infty$, the distributions induced on the slope and intercept are the empirical **sampling distribution of the parameters**.

We can use these sampling distribution to get confidence intervals on the parameters.

The variance of these distributions is called the **standard error**.

But we dont have M samples. What to do?

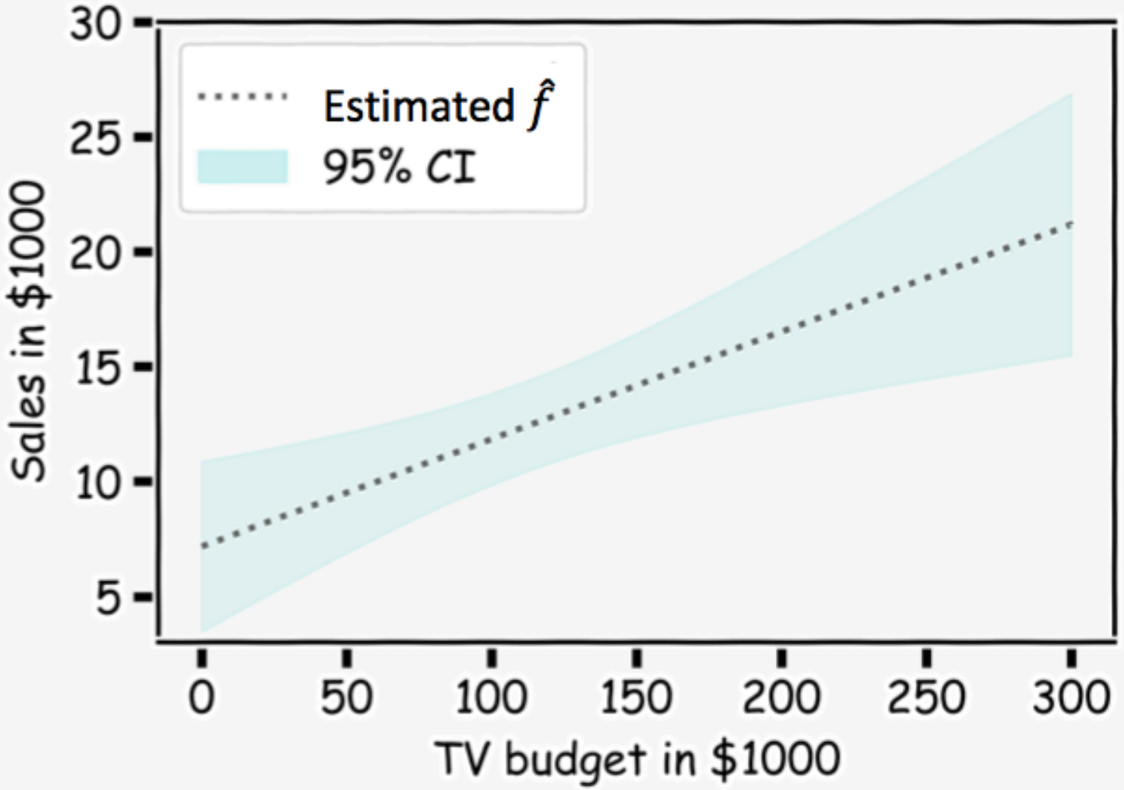
Bootstrap

- If we knew the true parameters of the population, we could generate M fake datasets.
- we dont, so we use our existing data to generate the datasets
- this is called the Non-Parametric Bootstrap

Sample with replacement the x from our original sample D, generating many fake datasets.

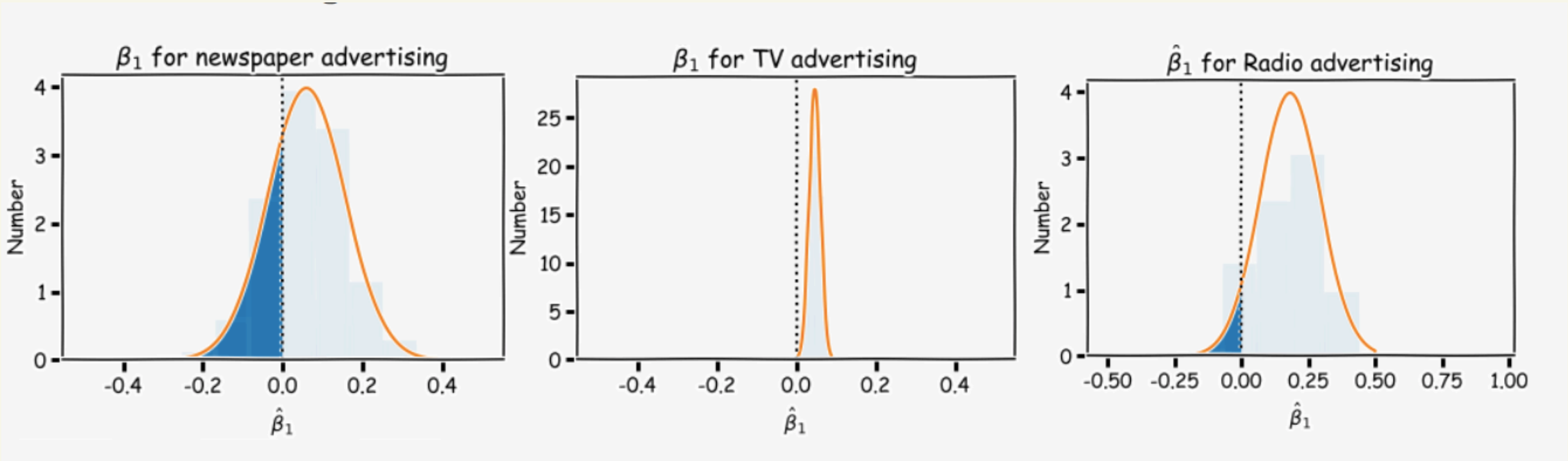
Sales in $1000
30
25
20
15
10
5
0
50
100
150
200
250
300
TV budget in $1000
12
14
16
18
20
22

Confidence Intervals on the "line"



- Each line is a rendering of $\mu = a + bx$, the mean value of the MLE Gaussian at each point x
- Thus the sampling distributions on the slope and intercept induce a sampling distribution on the lines
- And then the estimated $\hat{f}$ is taken to be the line with the mean parameters

Sampling Distributions and Significance

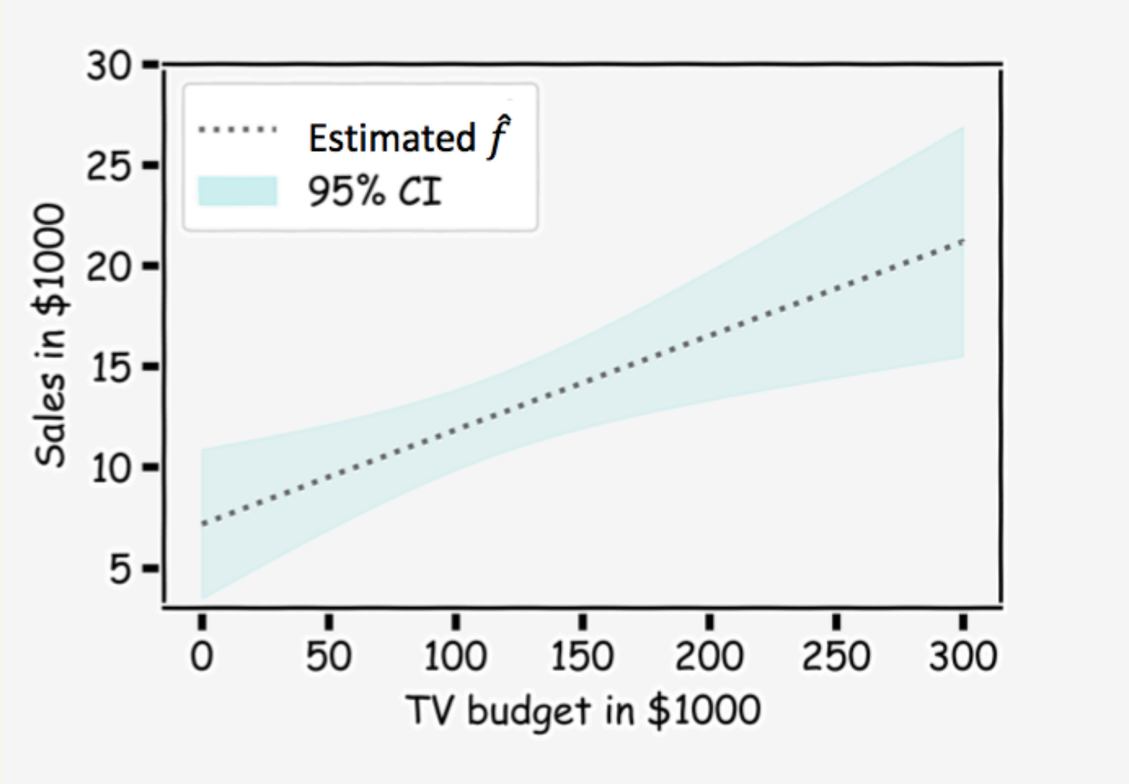


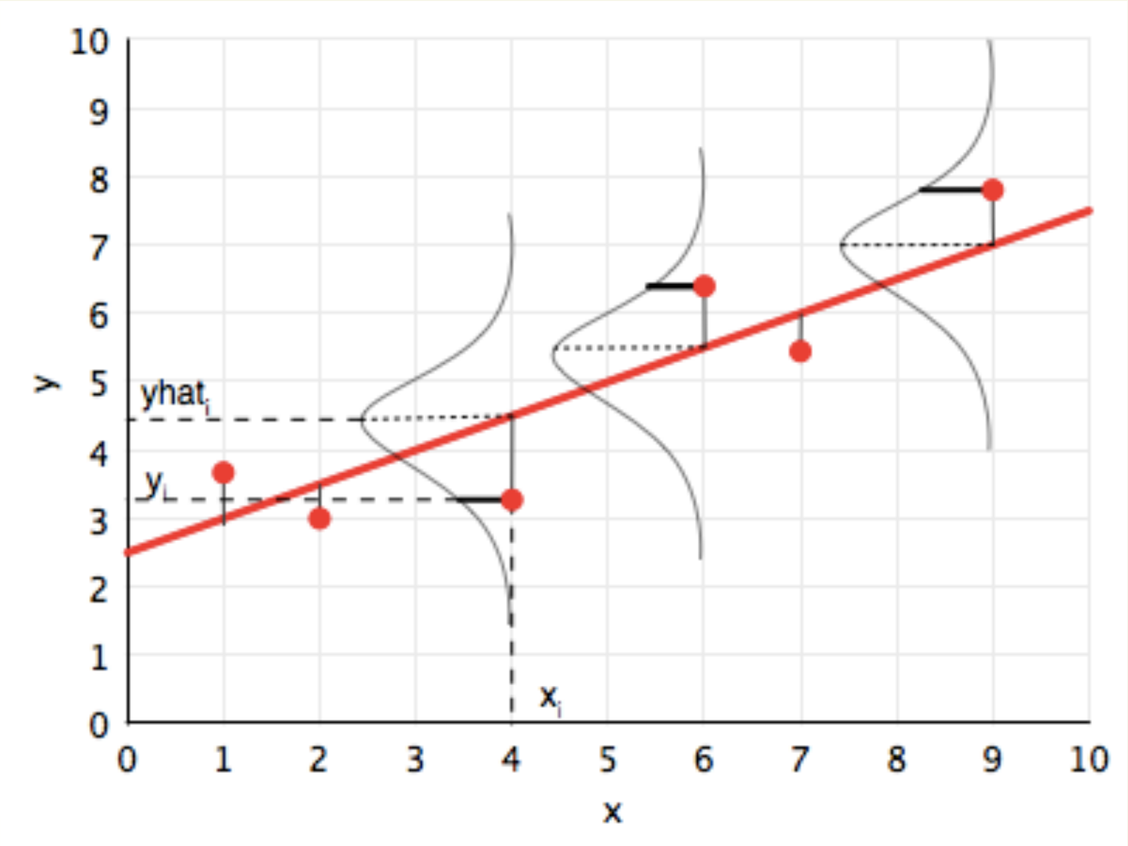
You want parameters to have their sampling distributions as far away from 0 as possible. But consider the "Null Hypothesis": a given parameter has no effect. We can do this by re-permuting just that column (HW :-)).

Prediction is more uncertain than the mean

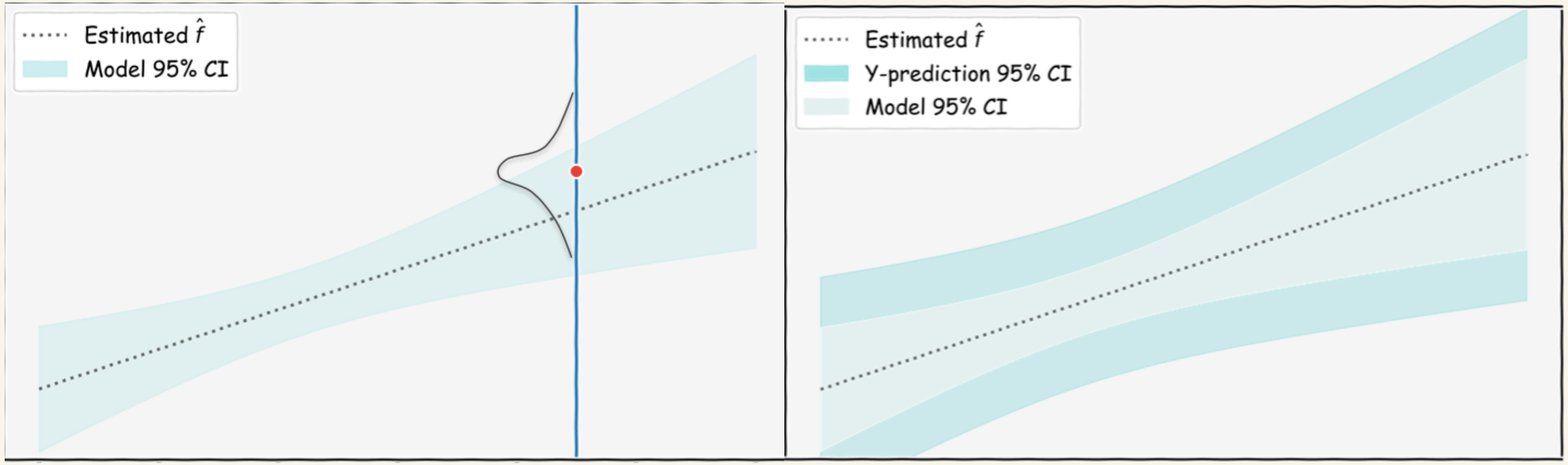
- In machine learning we do not care too much about the functional form of our prediction $\hat{y} = \hat{f}(x)$, as long as we predict "well"
- Remember however our origin story for the data: the measured y is assumed to have been a draw from a gaussian distribution $p(y|\mathbf{x}, \mu_{MLE}, \sigma^2_{MLE})$ at each x: this means that our prediction at an as yet not measured x should also be a draw from such a gaussian
- While we use the mean value of the gaussian as the value of the "prediction", the actual data will be gaussian draws at each point of the prediction

Combining Sampling and ϵ





Mean vs Prediction



$p(y^*|\mathbf{x}^*, \{\mathbf{x}_i, y_i\}, \mu_{MLE}, \sigma^2_{MLE})$ is the predictive distribution for as yet unseen data y^* at $\mathbf{x}^*$ at the covariates $\mathbf{x}^*$. This is a wider band.