Lecture 5: Linear Regression

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

ANNOUNCEMENTS

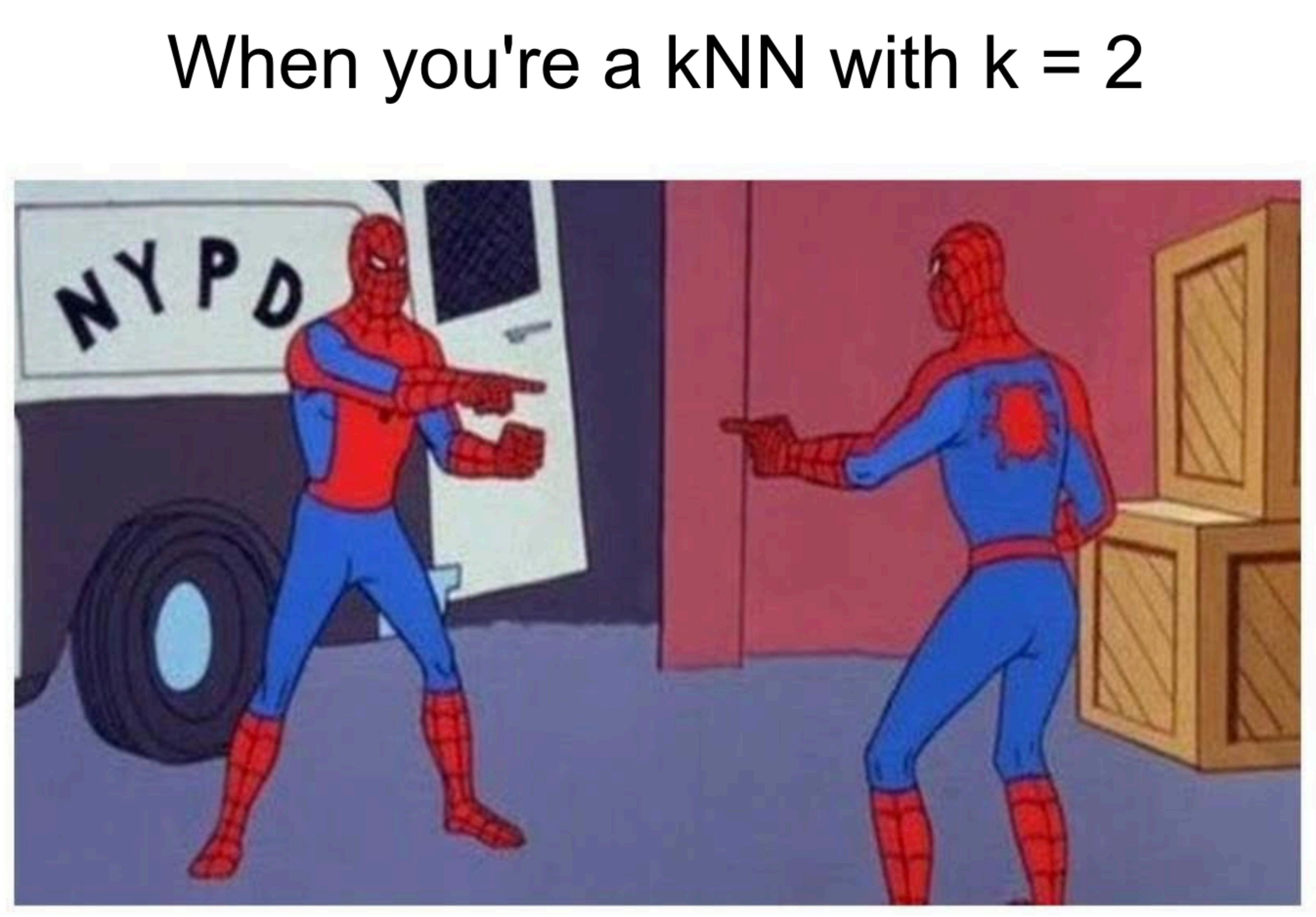
- **Advanced Sections** (A-Sections):

  TODAY @ **4:30pm** (MD G115)

  Linear Algebra and Hypothesis Testing, Pavlos + Kevin

- **ED-Exercises grading:**
  - All exercises together are equivalent to one question for that day’s quiz.
  - We grade for accuracy. You will receive full grade even if it fails the finicky test.
  - We will grade these exercises very leniently.

Summary from last lecture



Summary from last lecture

Model Fitness

How does the model perform predicting?

Comparison of Two Models

How do we choose from two different models?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

This lecture

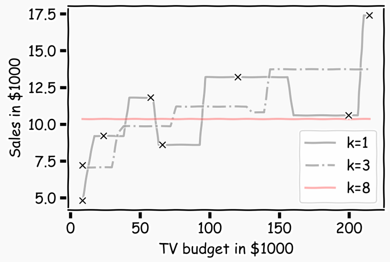
Lecture Outline

- Linear models
- Estimate of the regression coefficients
  - Brute Force
  - Exact method
  - Gradient Descent
- Confidence intervals for the predictors estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$

Lecture Outline

- **Linear models**
- Estimate of the regression coefficients
  - Brute Force
  - Exact method
  - Gradient Descent
- Confidence intervals for the predictors estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$

Linear Models



Note that in building our kNN model for prediction, we did not compute a closed form for $\hat{f}$.

What if we ask the question:

"how much more sales do we expect if we double the TV advertising budget?"

Alternatively, we can build a model by first assuming a simple form of f:

$$\boxed{Y = f(X) + \epsilon = \beta_1 X + \beta_0 + \epsilon.}$$

Linear Regression

... then it follows that our estimate is:

$$\widehat{Y} = \widehat{f}(X) = \widehat{\beta_1} X + \widehat{\beta_0}$$

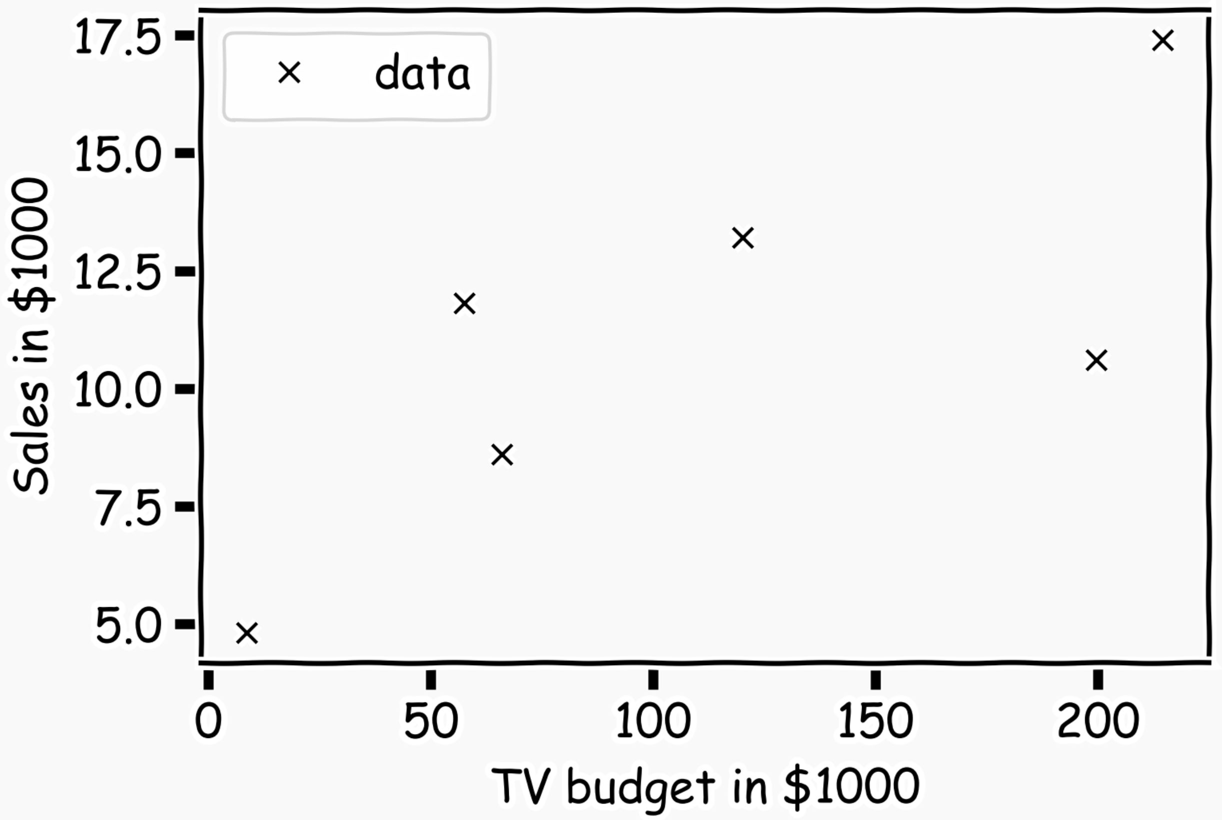
where $\hat{\beta}_1$ and $\hat{\beta}_0$ are **estimates** of β_1 and β_0 respectively, that we compute using observations.

Lecture Outline

- Linear models
- **Estimate of the regression coefficients**
    - Brute Force
    - Exact method
    - Gradient Descent
- Confidence intervals for the predictors estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$

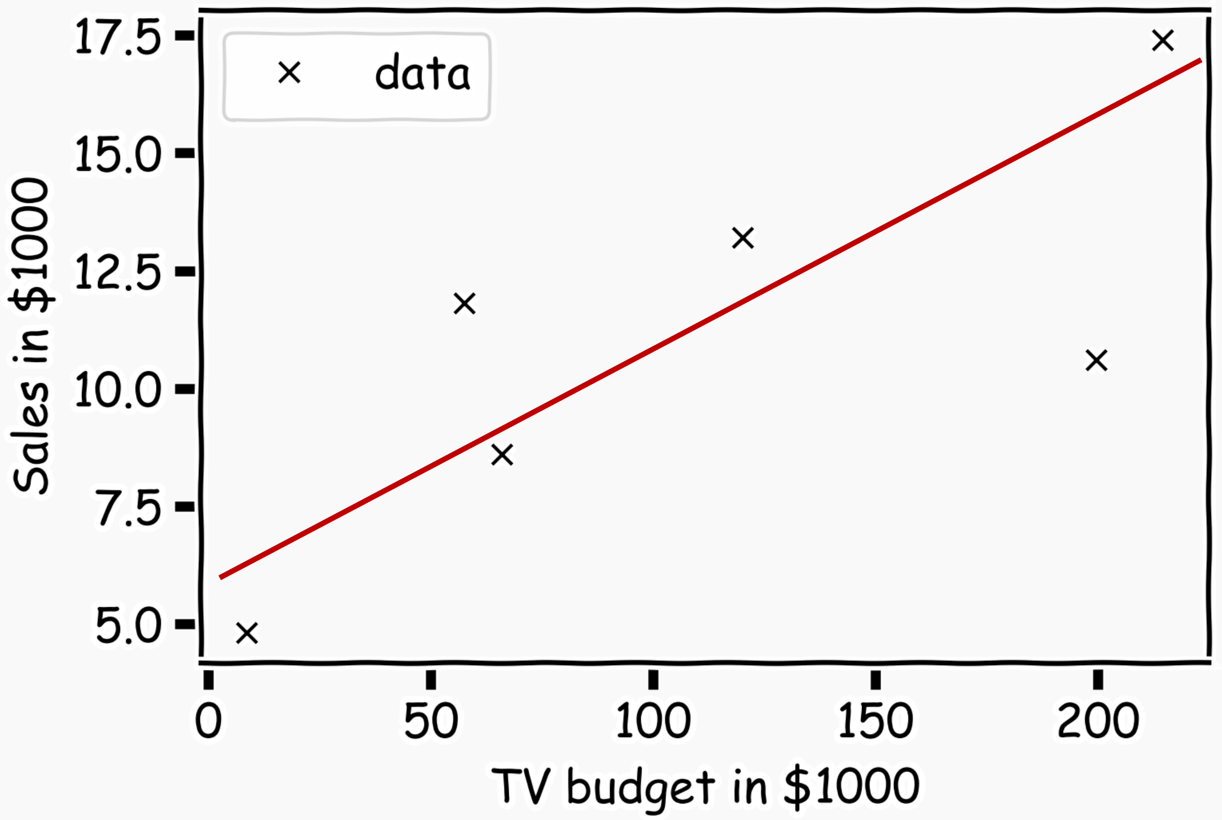
Estimate of the regression coefficients

For a given data set



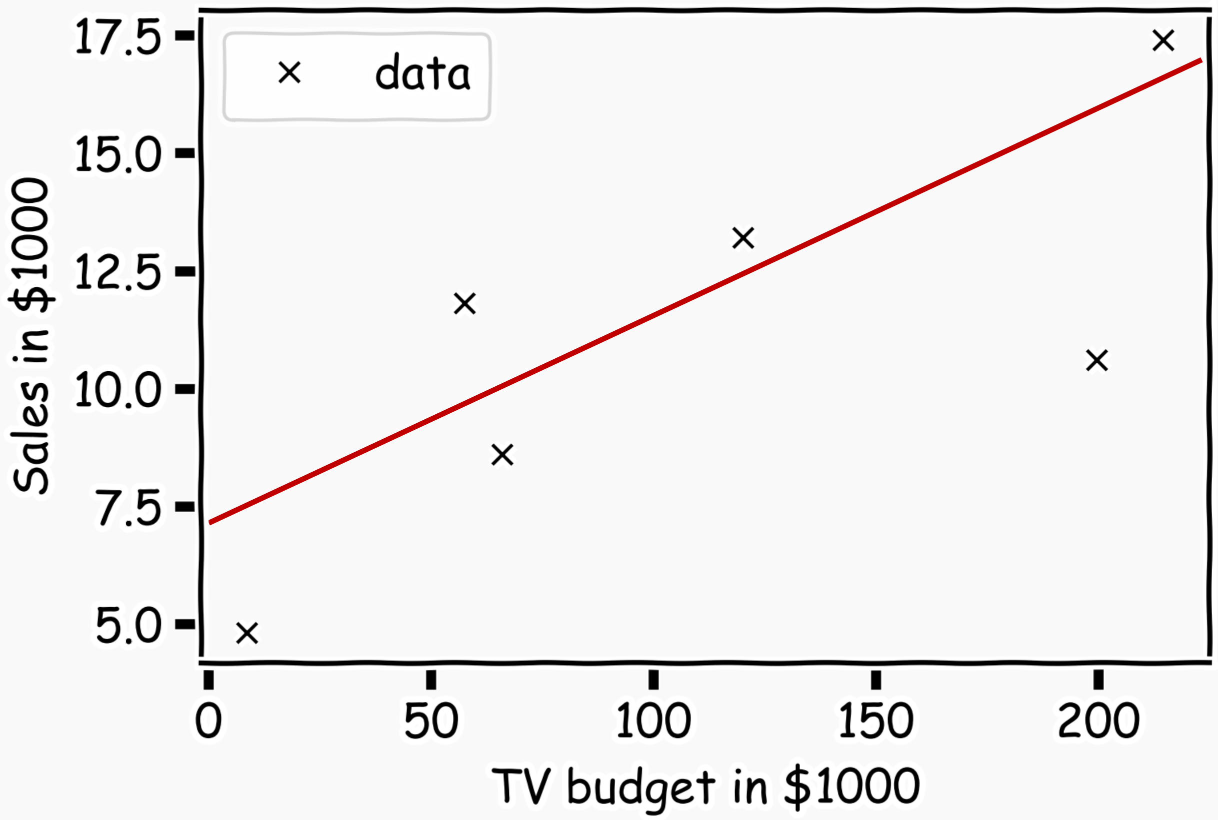
Estimate of the regression coefficients (cont)

Is this line good?



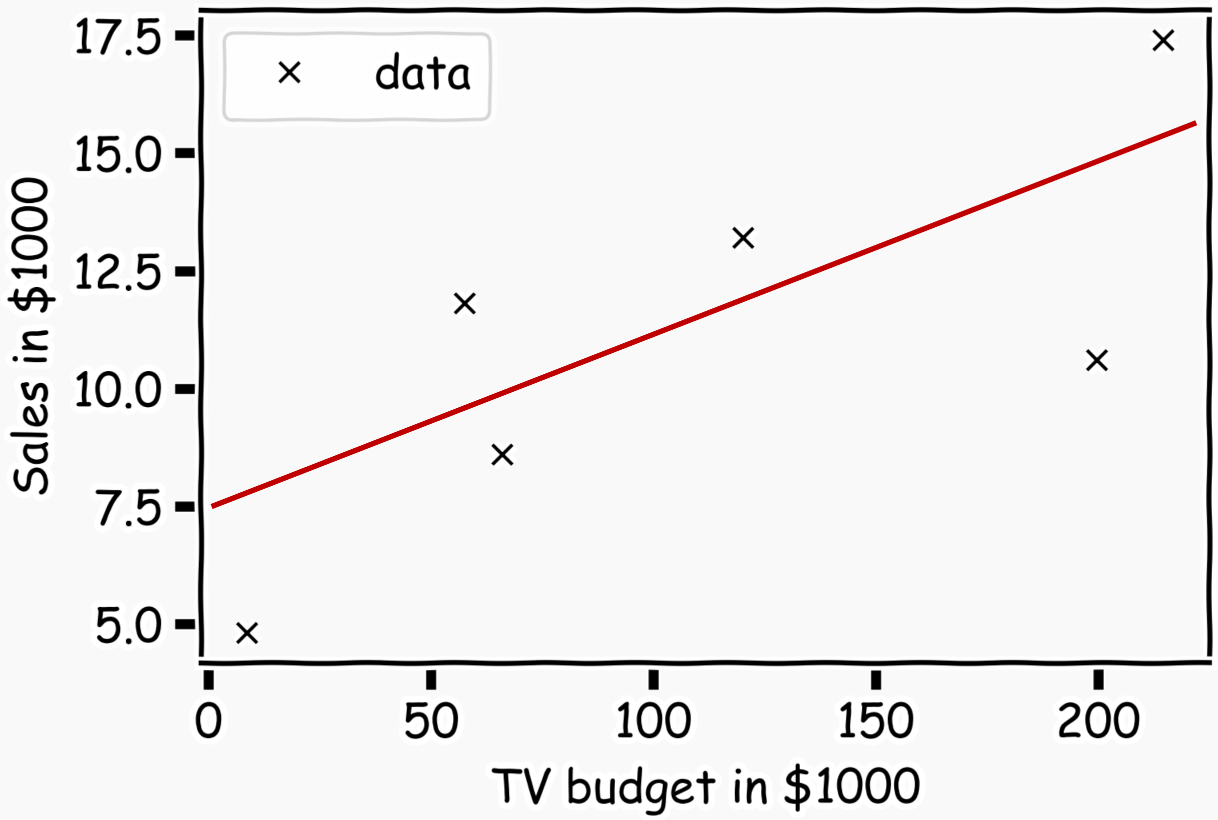
Estimate of the regression coefficients (cont)

Maybe this one?



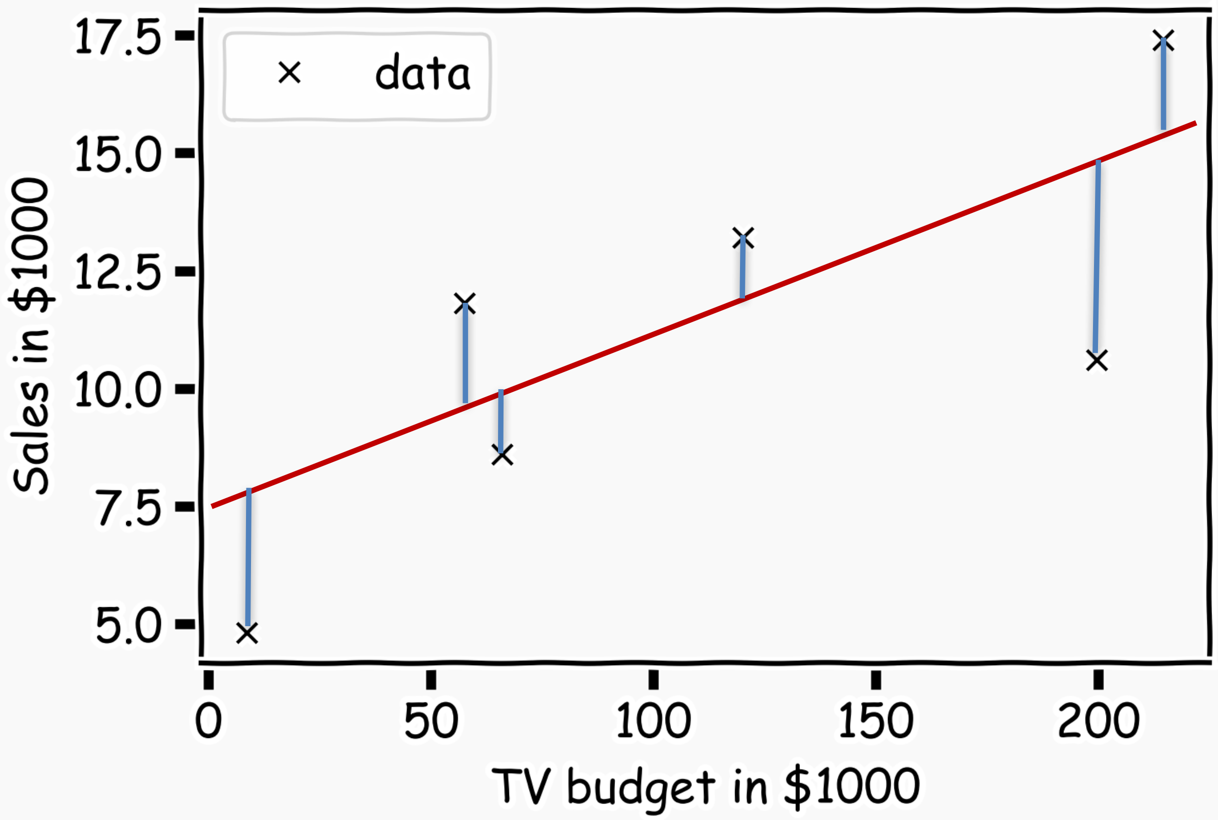
Estimate of the regression coefficients (cont)

Or this one?



Estimate of the regression coefficients (cont)

Question: Which line is the best?
First calculate the residuals



Estimate of the regression coefficients (cont)

Again we use MSE as our **loss function**,

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - (\beta_1 X + \beta_0)\right]^2 .$$

We choose $\hat{\beta}_1$ and $\hat{\beta}_0$ in order to minimize the predictive errors made by our model, i.e. minimize our loss function.

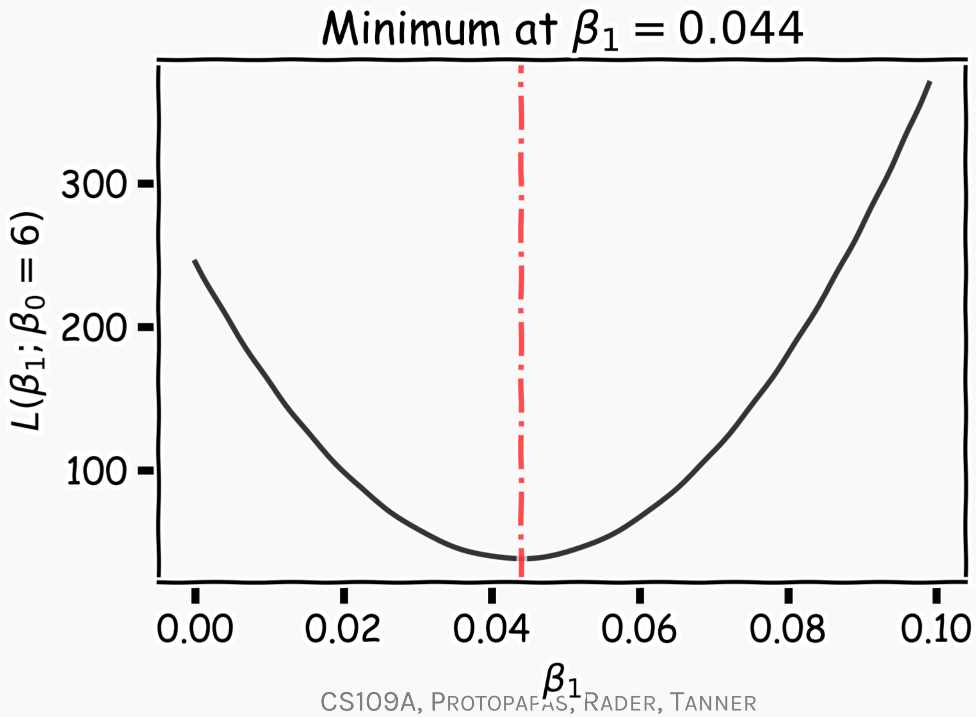
Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

$$\widehat{\beta}_0, \widehat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} L(\beta_0, \beta_1).$$

Estimate of the regression coefficients: brute force

A way to estimate $\text{argmin}_{\beta_0,\beta_1} L$ is to calculate the loss function for every possible β_0 and β_1. Then select the β_0 and β_1 where the loss function is minimum.

E.g. the loss function for different β_1 when β_0 is fixed to be 6:



Estimate of the regression coefficients: exact method

Take the partial derivatives of L with respect to β_0 and β_1, set to zero, and find the solution to that equation. This procedure will give us explicit formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\boxed{\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

where $\bar{y}$ and $\bar{x}$ are sample means.

The line:

$$\widehat{Y} = \widehat{\beta}_1 X + \widehat{\beta}_0$$

is called the **regression line**.

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_i \left[y_i - (\beta_0 - \beta_1 x_i)\right]^2$$

$$\frac{dL(\beta_0, \beta_1)}{d\beta_0} = 0$$

$$\Rightarrow \frac{2}{n}\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \frac{1}{n}\sum_i y_i - \beta_0 - \beta_1 \frac{1}{n}\sum_i x_i = 0$$

$$\boxed{\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}}$$

$$\frac{dL(\beta_0, \beta_1)}{d\beta_1} = 0$$

$$\Rightarrow \frac{2}{n}\sum_i (y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$\Rightarrow -\sum_i x_i y_i + \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 = 0$$

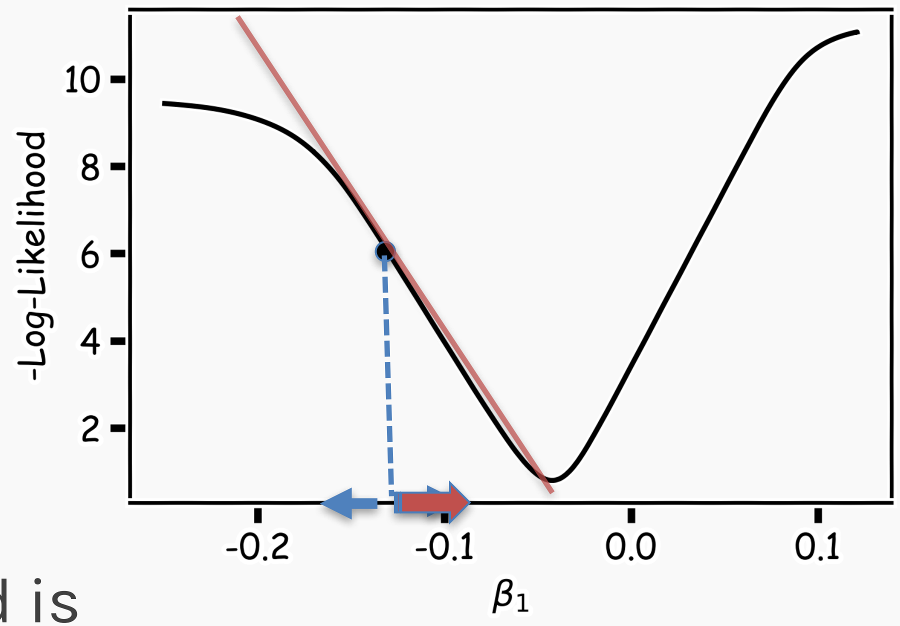
$$\Rightarrow -\sum_i x_i y_i + (\bar{y} - \beta_1 \bar{x}) \sum_i x_i + \beta_1 \sum_i x_i^2 = 0$$

$$\Rightarrow \beta_1 \left(\sum_i x_i^2 - n\bar{x}^2 \right) = \sum_i x_i y_i - n\bar{x}\bar{y}$$

$$\Rightarrow \beta_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

$$\boxed{\Rightarrow \beta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}}$$

Estimate of the regression coefficients: gradient descent



A more flexible method is

- Start from a random point
  1. Determine which direction to go to reduce the loss (left or right)
  2. Compute the slope of the function at this point and step to the right if slope is negative or step to the left if slope is positive
  3. Goto to #1

Estimate of the regression coefficients: gradient descent

Question: What is the mathematical function that describes the slope?

Derivative

Question: What do you think it is a good approach for telling the model how to change (what is the step size) to become better?

If the step is proportional to the slope then you avoid overshooting the minimum

Question: How do we generalize this to more than one predictor?

Take the derivative with respect to each coefficient and do the same sequentially

Estimate of the regression coefficients: gradient descent

We know that we want to go in the opposite direction of the derivative and we know we want to be making a step proportionally to the derivative.

Notation: $w = \beta_0, \beta_1$

Making a step means:

$$w^{new} = w^{old} + step$$

Opposite direction of the derivative and proportional to the derivative means:

Learning Rate

$$w^{new} = w^{old} - \lambda \frac{d\mathcal{L}}{dw}$$

Change to more conventional notation:

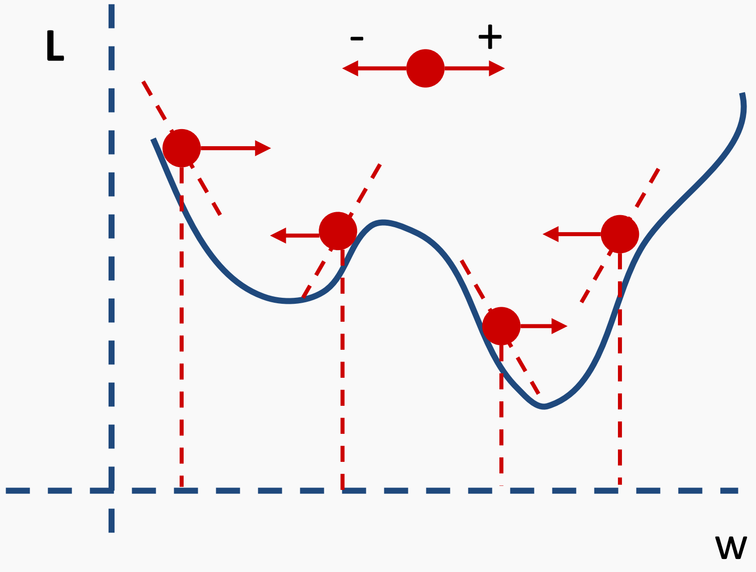
$$w^{(i+1)} = w^{(i)} - \lambda \frac{d\mathcal{L}}{dw}$$

Estimate of the regression coefficients: gradient descent

Summary of Gradient Descent

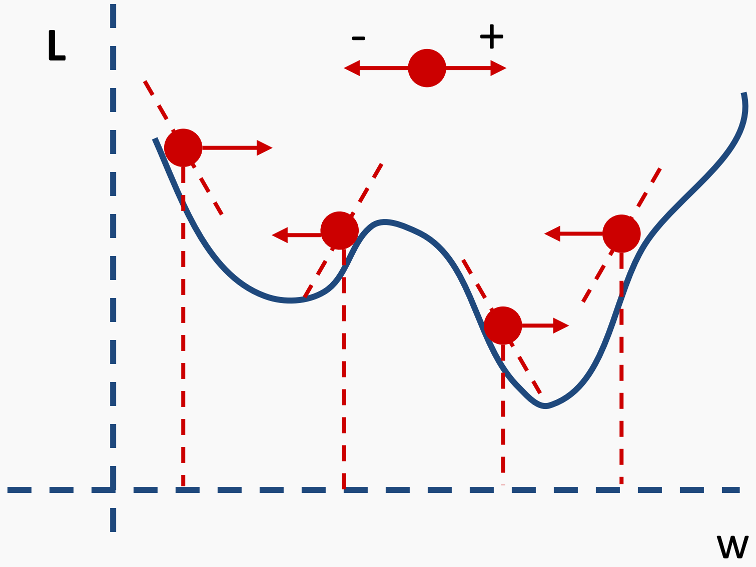
- Algorithm for optimization of first order to finding a minimum of a function.
- It is an iterative method.
- L is decreasing in the direction of the negative derivative.
- The learning rate is controlled by the magnitude of λ.

$$w^{(i+1)} = w^{(i)} - \lambda \frac{d\mathcal{L}}{dw}$$



Gradient Descent: considerations

Gradient Descent Considerations (more in coming lectures)



- We still need to derive or compute the derivatives.
- We need to know what is the learning rate or how to set it.
- We need to avoid local minima.
- Finally, the full loss function includes summing up all individual '*errors*'. This can be hundreds of thousands of examples.

Gradient Descent: considerations

- In linear regression, there are no local minima because the loss function is convex

- In linear regression, we often use the exact formula

- We will talk about optimization again in Neural Network lecture and the last advanced sections

Lecture Outline

- Linear models
- Estimate of the regression coefficients
  - Brute Force
  - Exact method
  - Gradient Descent
- **Confidence intervals for the predictors estimates**
- Bootstrap
- Evaluating Significance of Predictors
- How well we know the model $\hat{f}$

Interpretation of Predictors

Question: What do you think a predictor coefficient means?

$$Sales = 7.5 + 0.04\, TV$$

What does 7.5 mean and what does 0.04 mean?

If we increase the TV by $1000, what would you expect the increase in sales to be?

What if? $$Sales = 7.5 + 1.01\, TV$$

The interpretation of the predictors depends on the values but decisions depend on how much we trust these values.

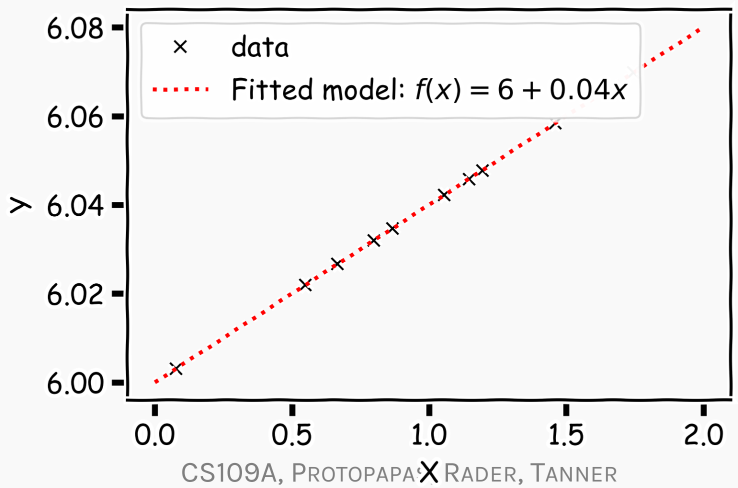
Confidence intervals for the predictors estimates

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.

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



Confidence intervals for the predictors estimates (cont)

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

- $\boldsymbol{\varepsilon}$ is always there
- we do not know the exact form of $f(x)$
- limited sample size

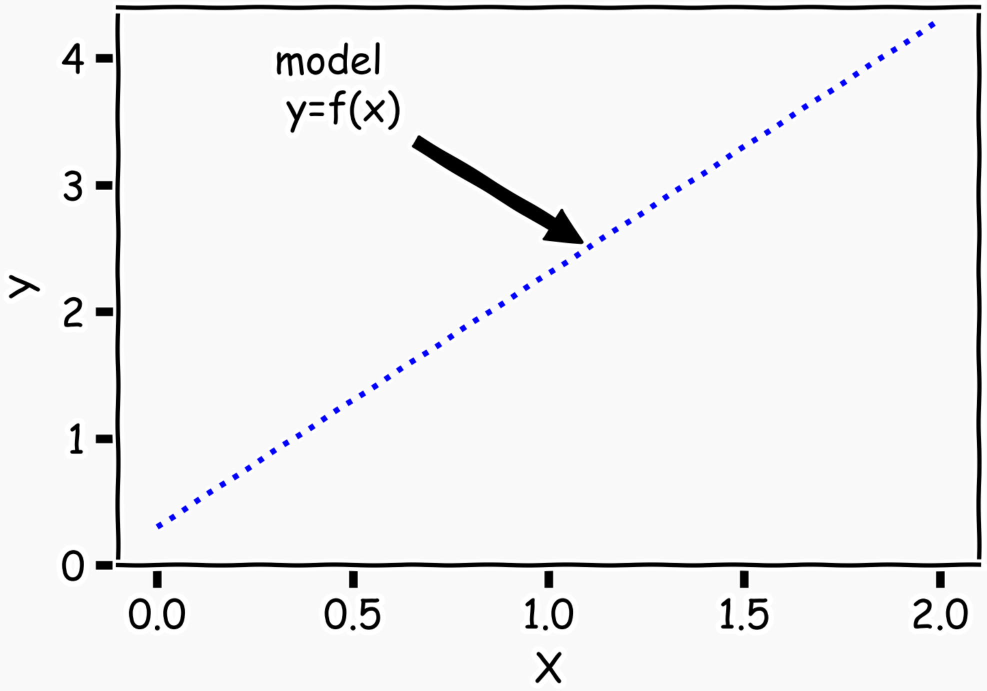
We will first address ε

We call ε the measurement error or ***irreducible error***. Since even predictions made with the actual function f will not match observed values of y.

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 $\hat{\beta}'s$.

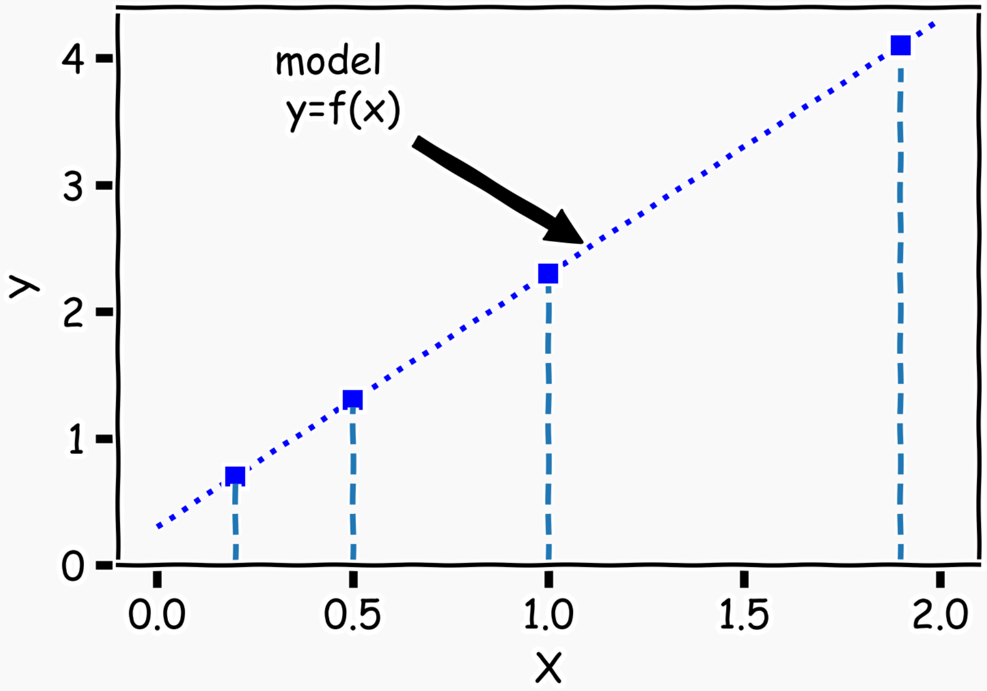
Confidence intervals for the predictors estimates (cont)

Start with a model



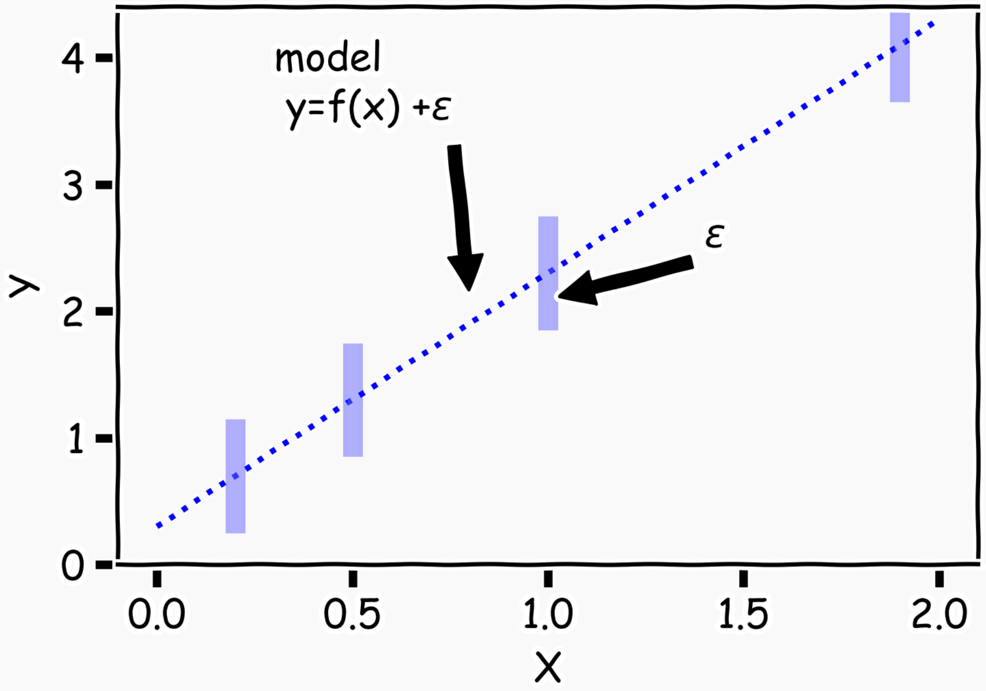
Confidence intervals for the predictors estimates (cont)

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



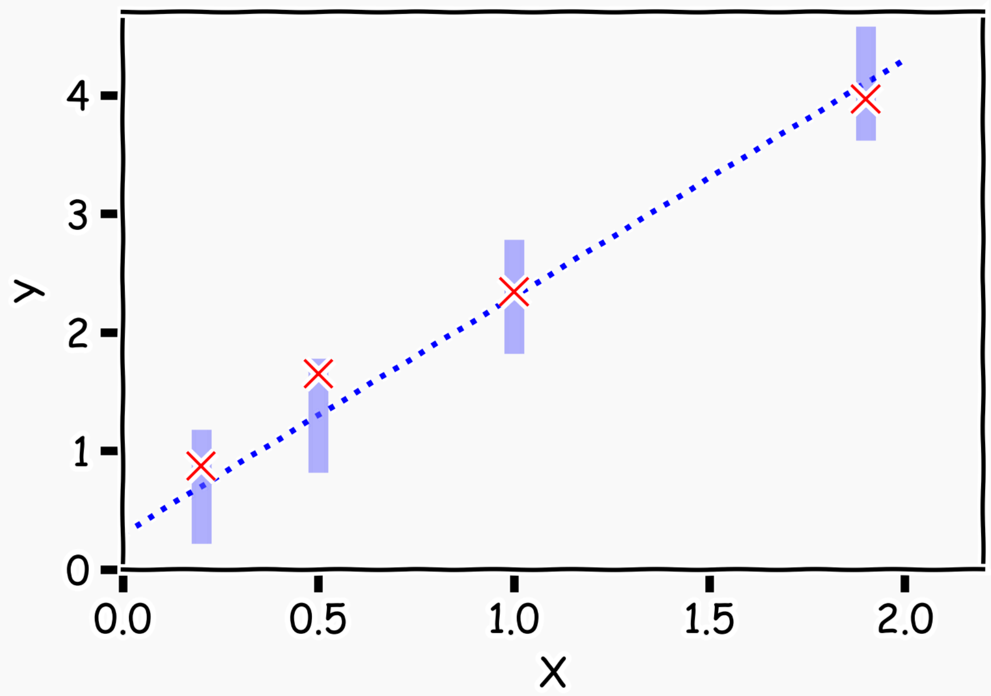
Confidence intervals for the predictors estimates (cont)

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



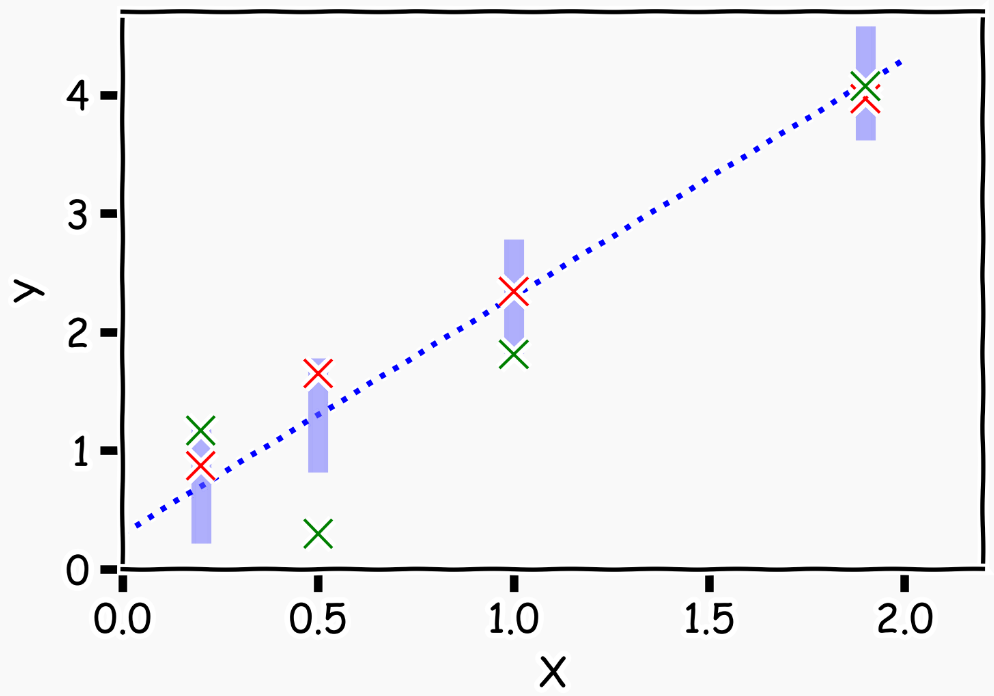
Confidence intervals for the predictors estimates (cont)

One set of observations, "one realization" we obtain one set of Ys (red crosses).



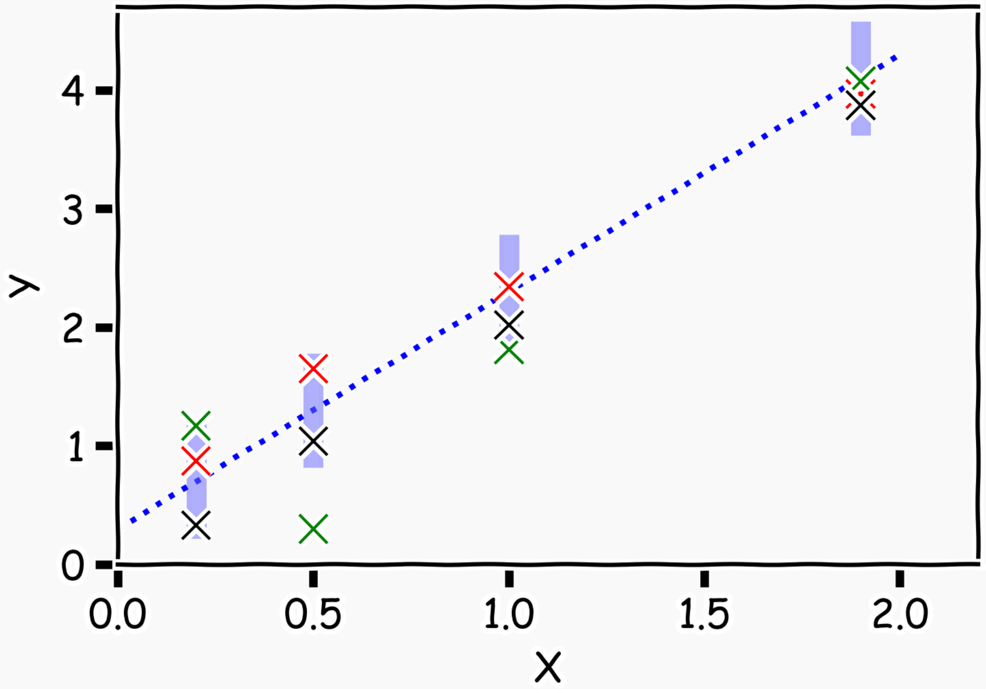
Confidence intervals for the predictors estimates (cont)

Another set of observations, “another realization” we obtain another set of Ys (green crosses).



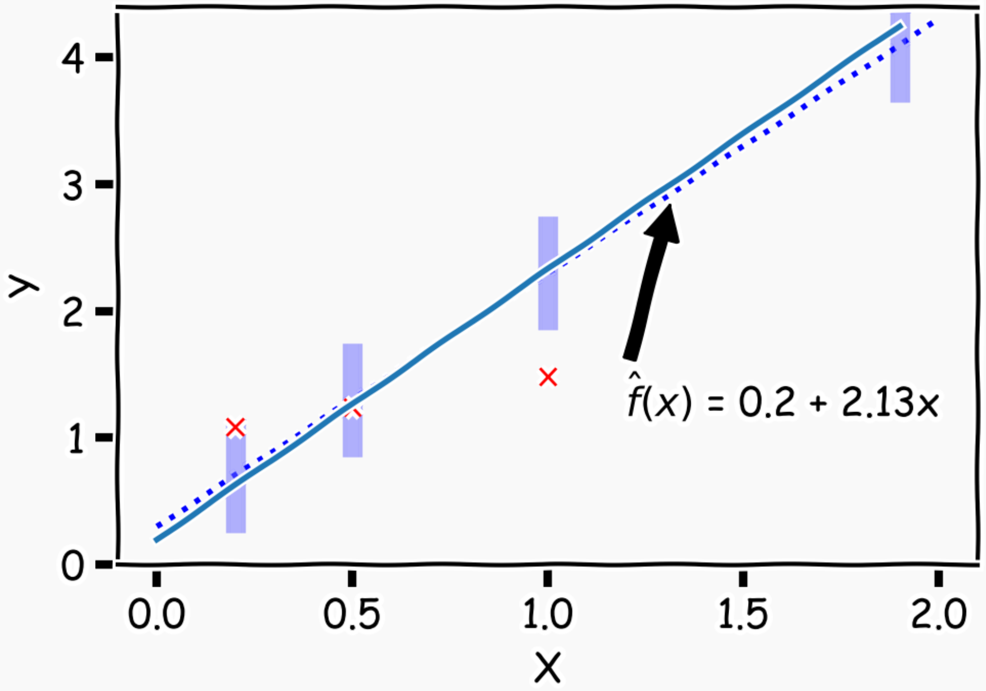
Confidence intervals for the predictors estimates (cont)

Another set of observations, "another realization" we obtain another set of Ys (black crosses).



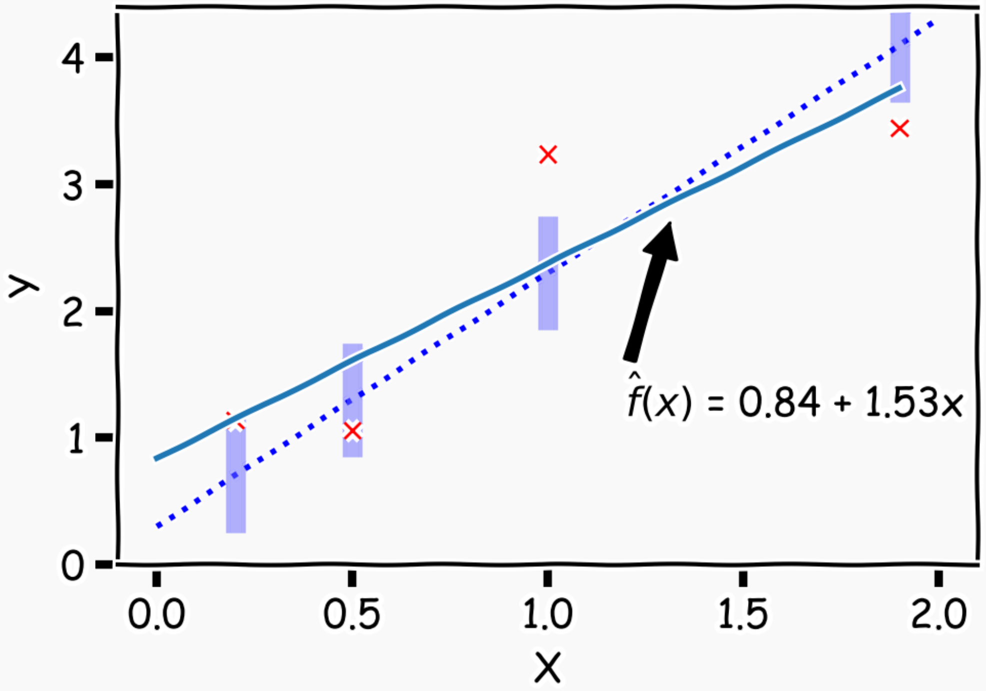
Confidence intervals for the predictors estimates (cont)

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



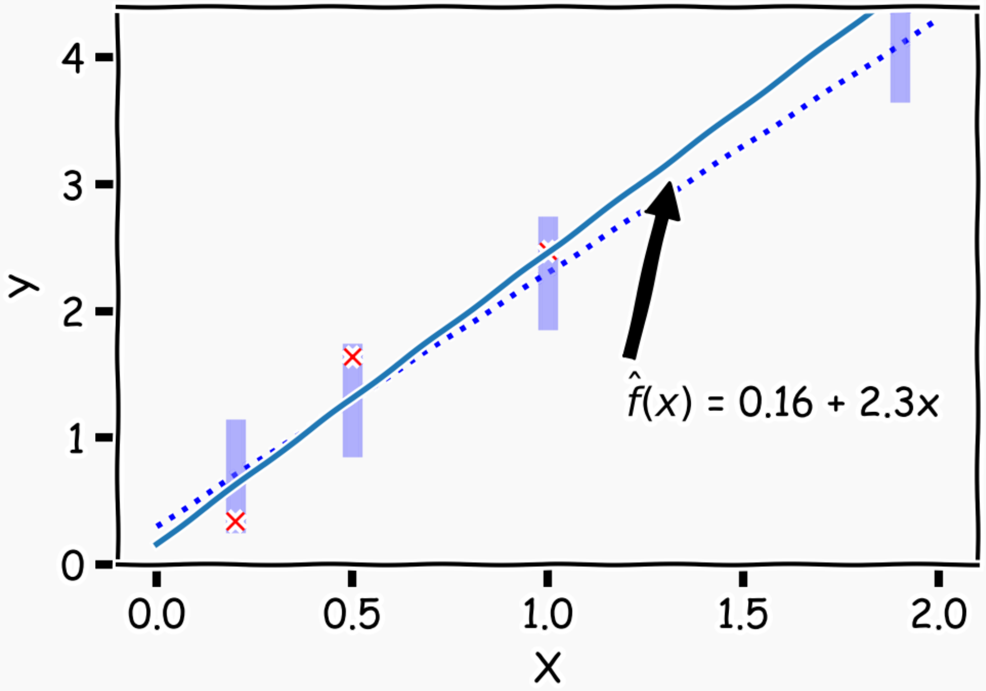
Confidence intervals for the predictors estimates (cont)

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



Confidence intervals for the predictors estimates (cont)

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



Confidence intervals for the predictors estimates (cont)

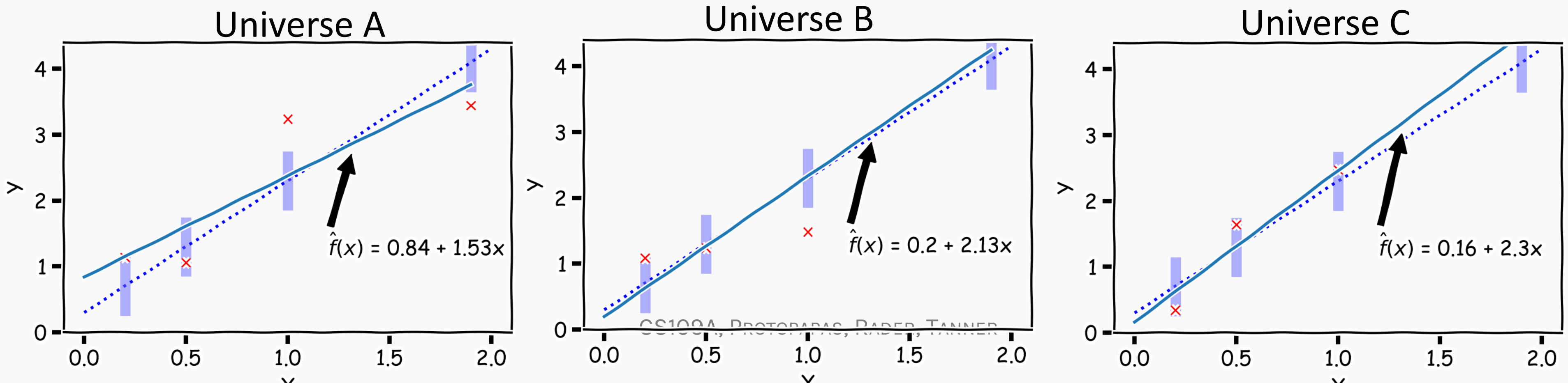
So if we just have one set of measurements of $\{X, Y\}$, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are just for this particular realization.

Confidence intervals for the predictors estimates (cont)

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

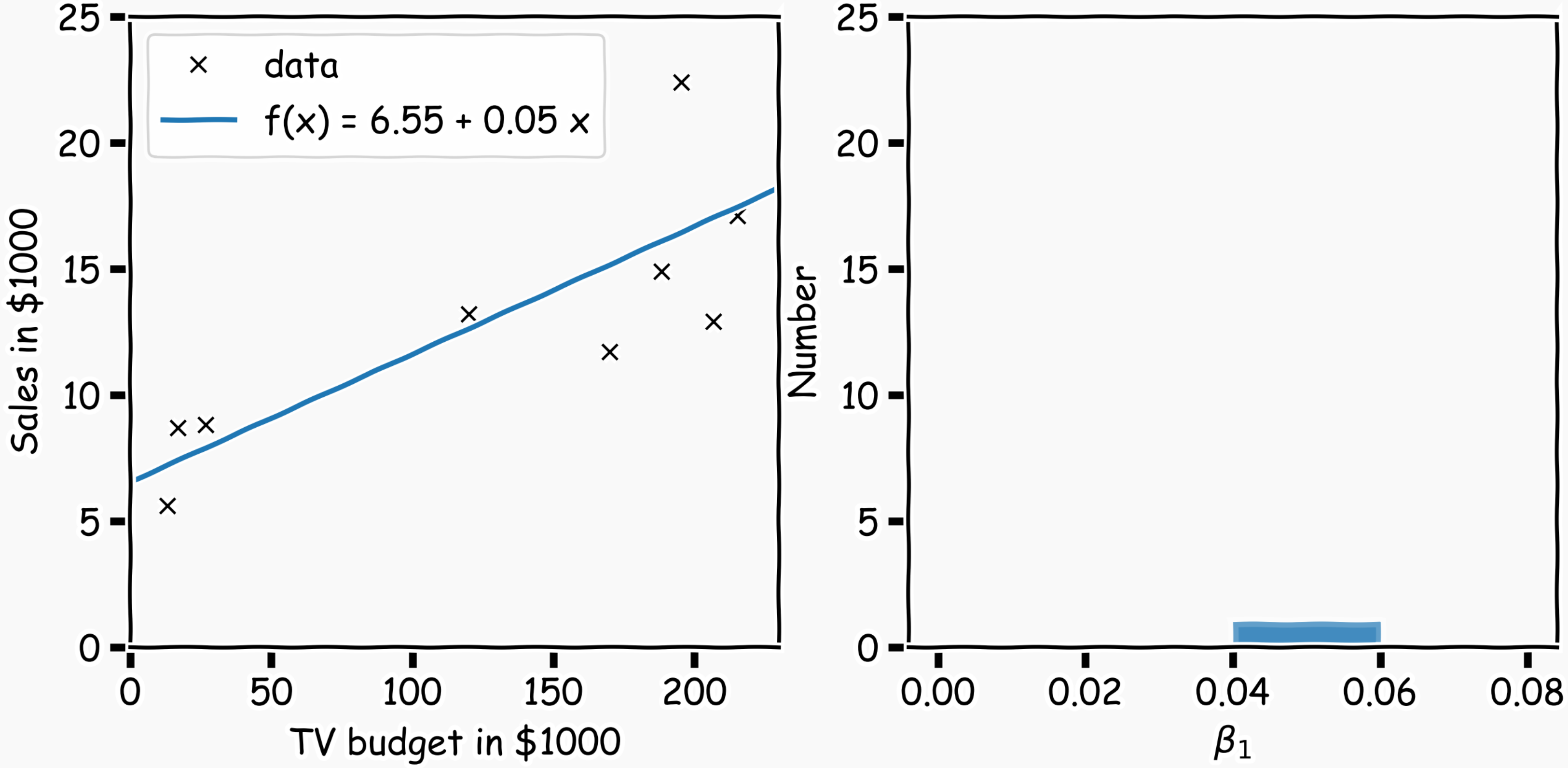
Question: If this is just one realization of the reality how do we know the truth? How do we deal with this conundrum?

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



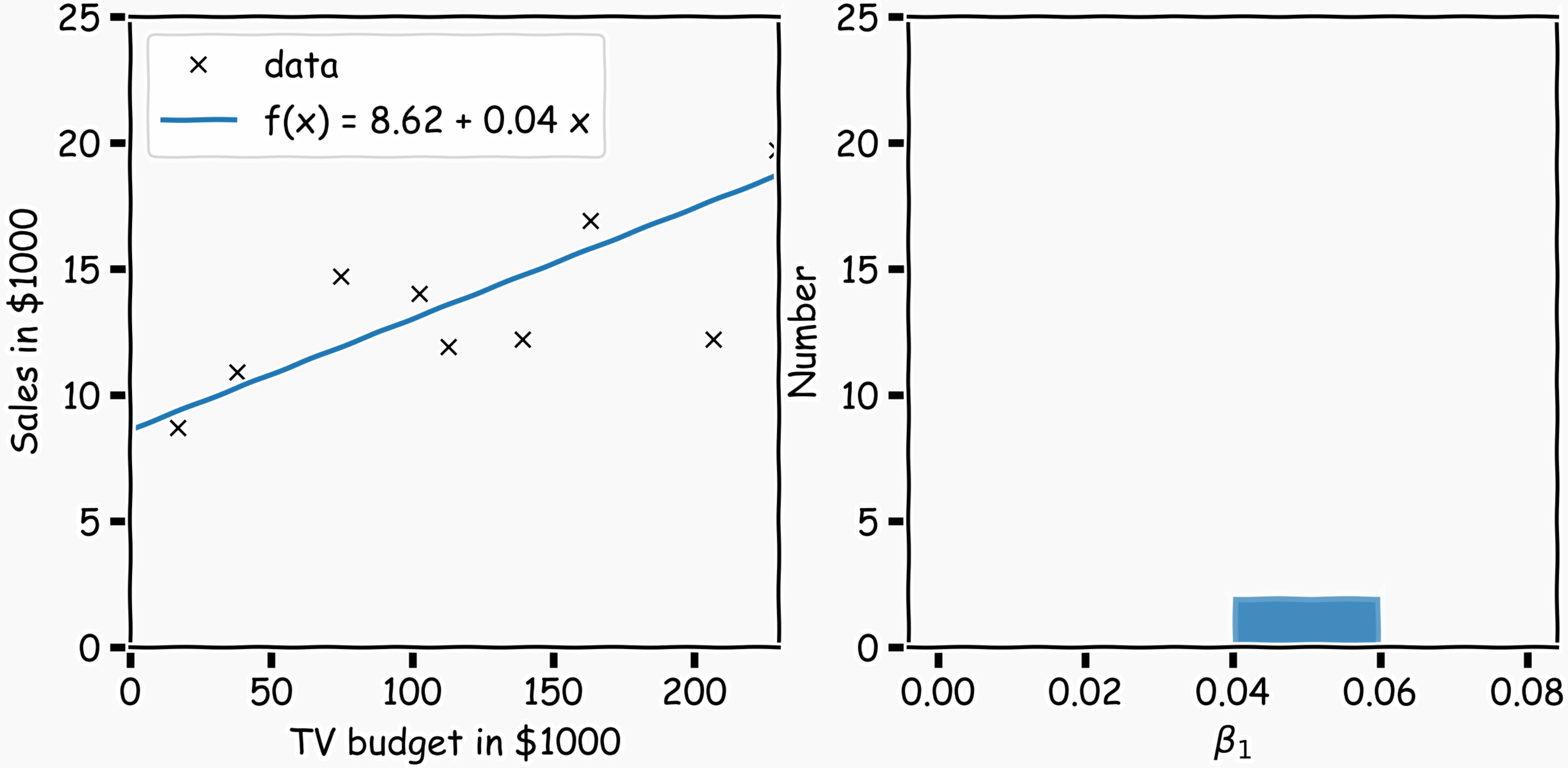
Confidence intervals for the predictors estimates (cont)

In our magical realisms, we can now sample multiple times



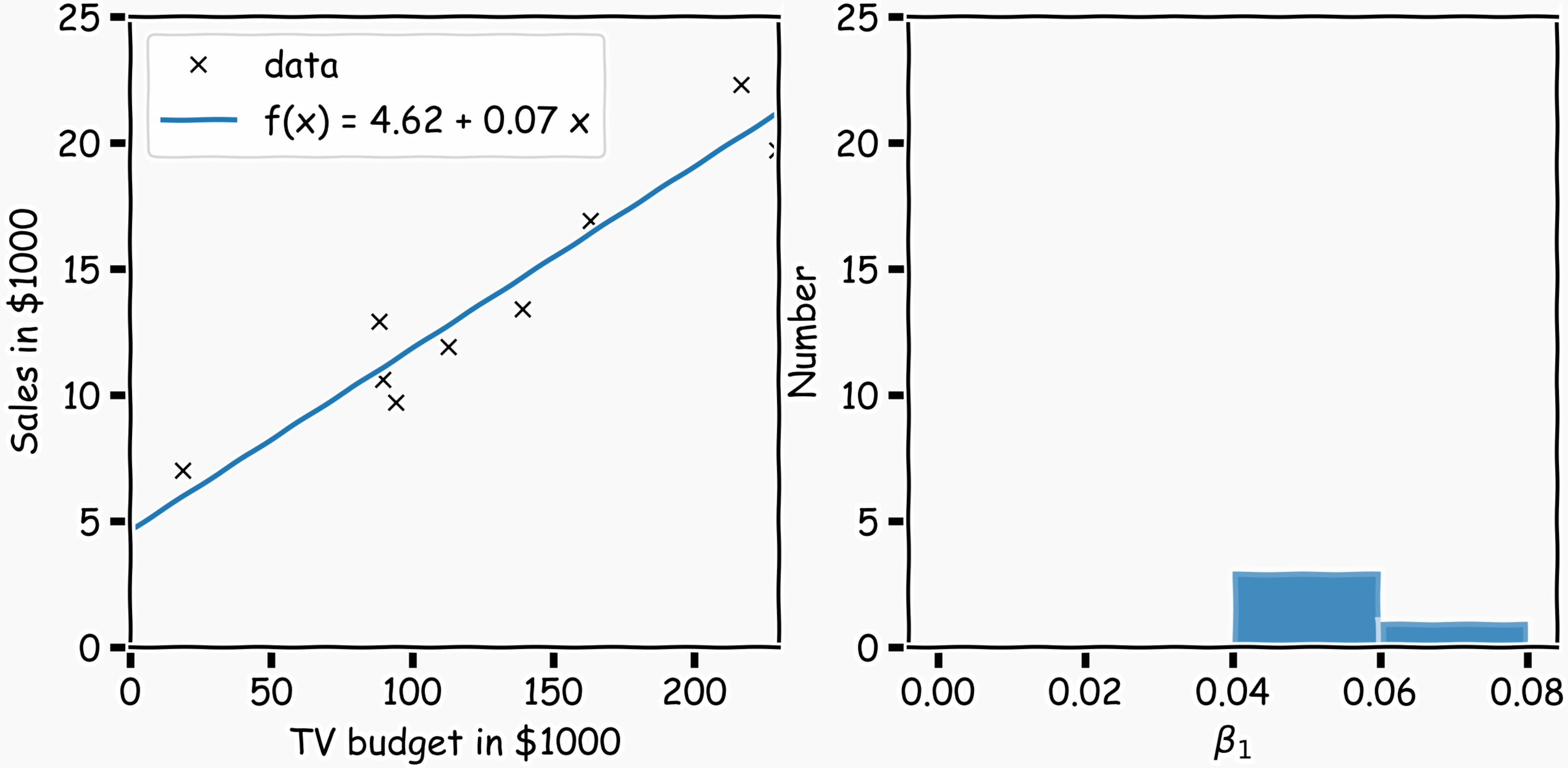
Confidence intervals for the predictors estimates (cont)

Another sample



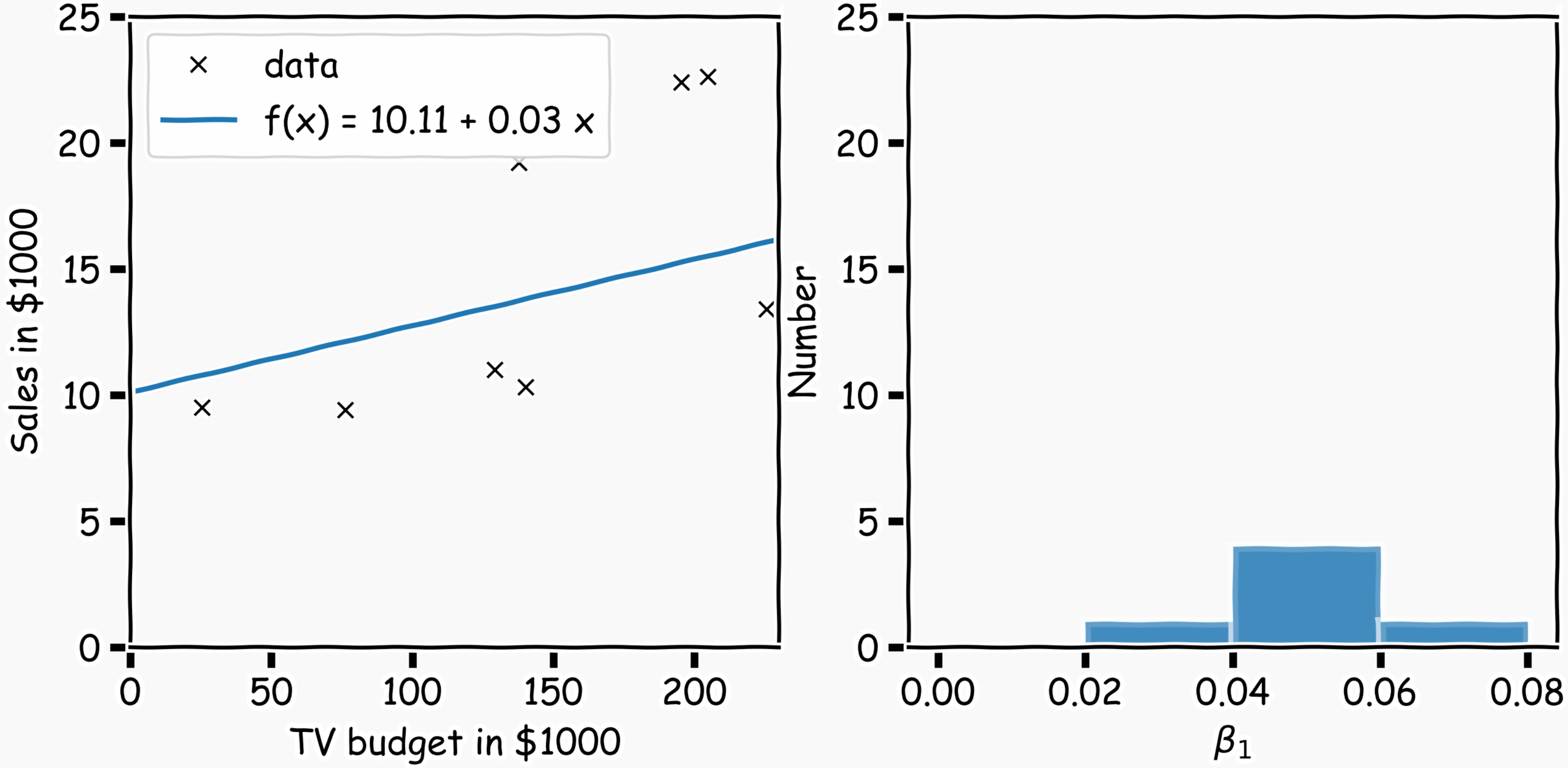
Confidence intervals for the predictors estimates (cont)

Another sample



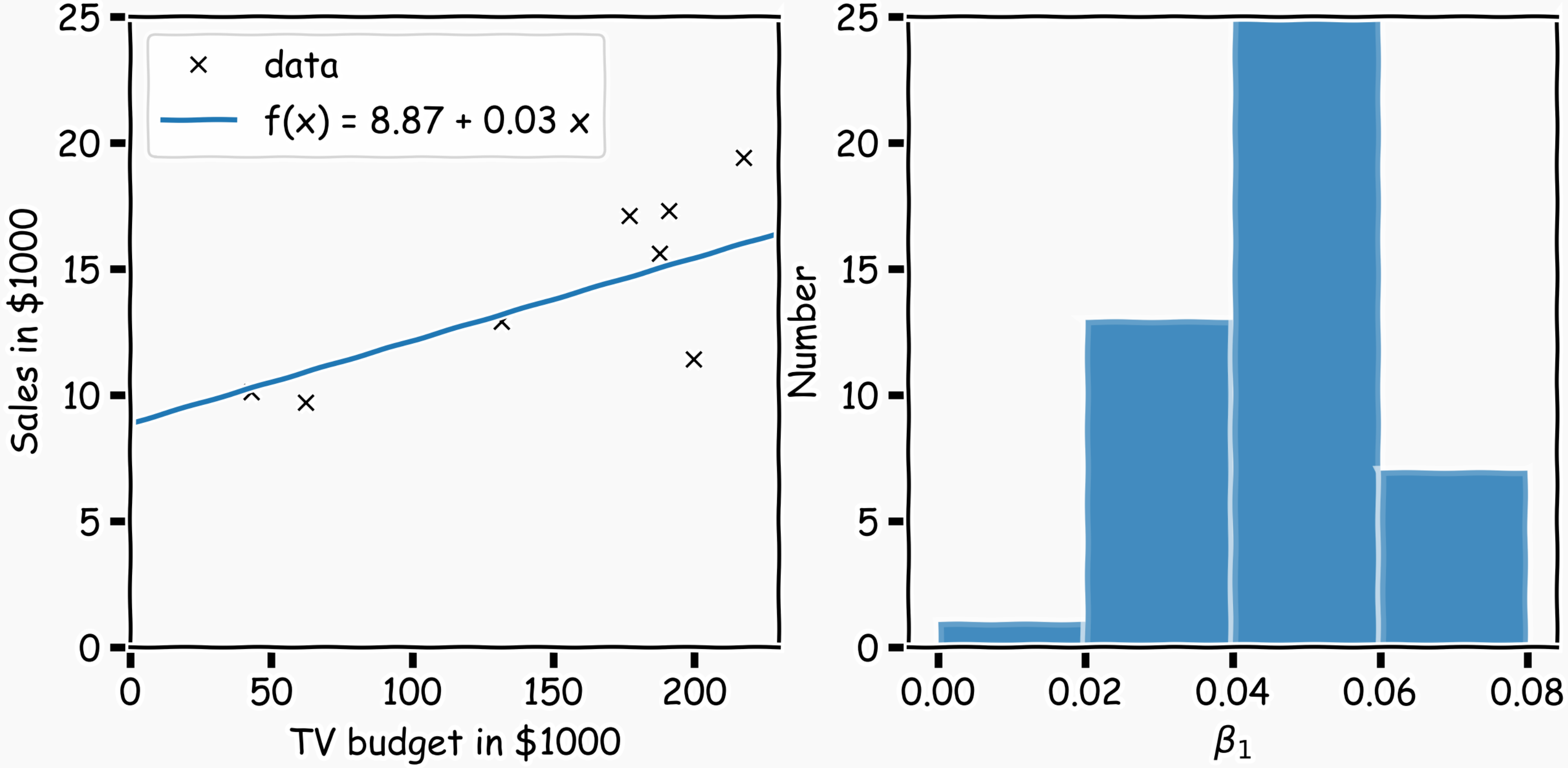
Confidence intervals for the predictors estimates (cont)

And another sample



Confidence intervals for the predictors estimates (cont)

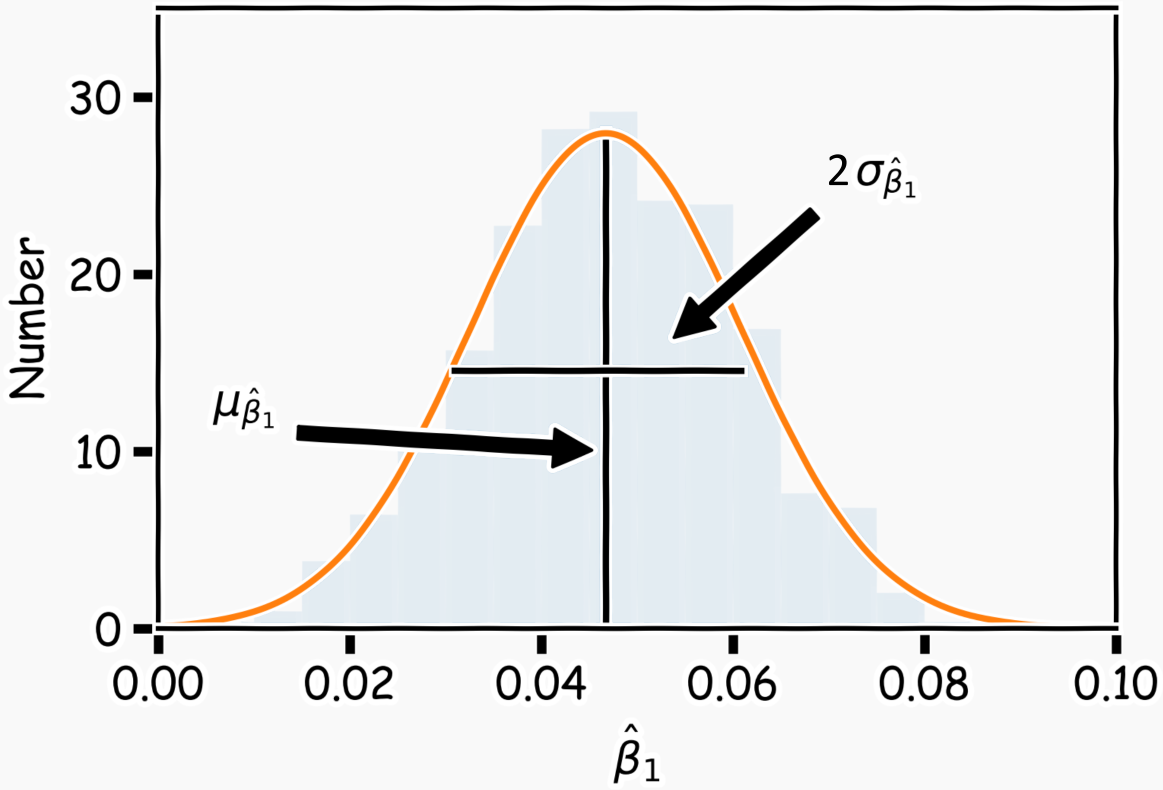
Repeat this for 100 times



Confidence intervals for the predictors estimates (cont)

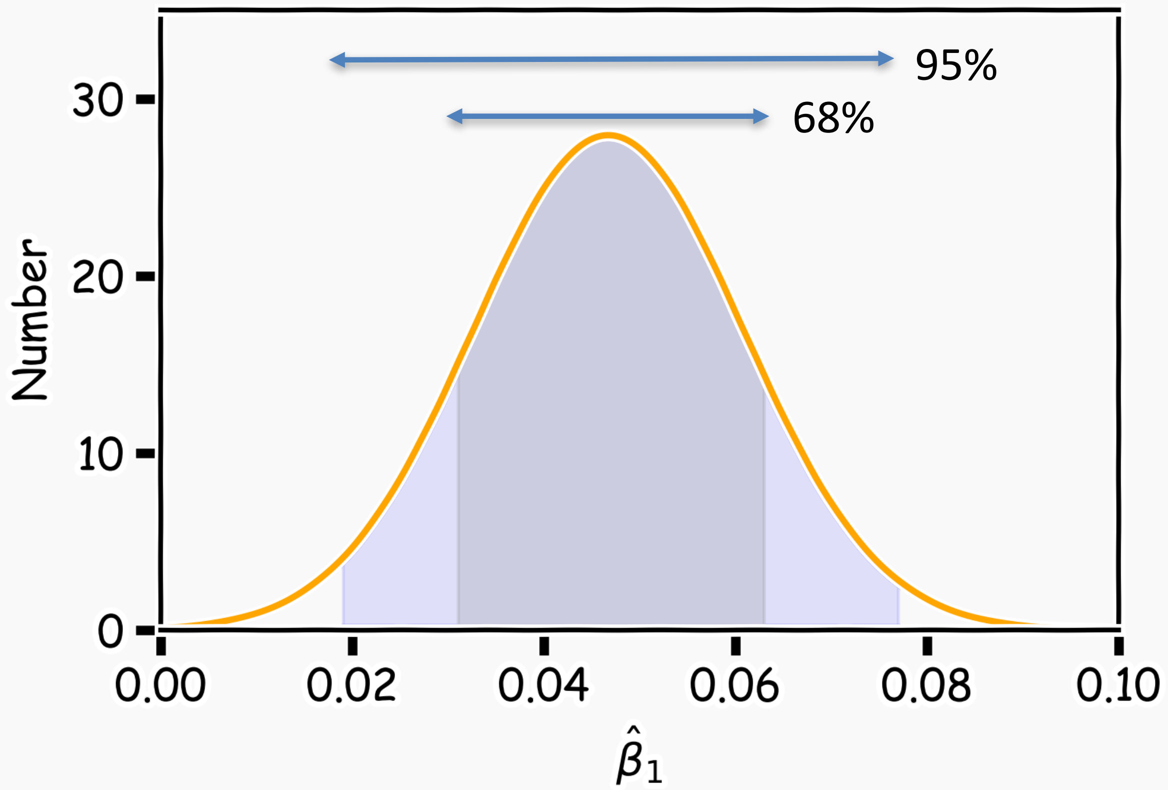
We can now estimate the mean and standard deviation of all the estimates $\hat{\beta}_1$.

The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are also called their ***standard errors***, $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$.



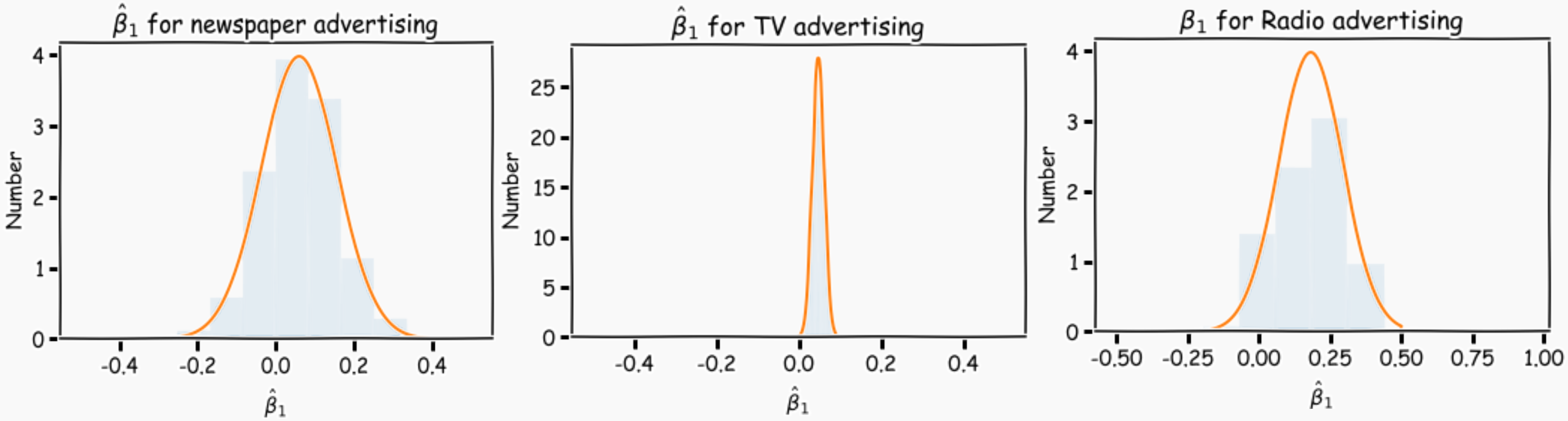
Confidence intervals for the predictors estimates (cont)

Finally we can calculate the confidence intervals, which are the ranges of values such that the **true** value of β_1 is contained in this interval with *n* percent probability.



And also we can answer the question, 'how significant are the predictors?' Here we show the same analysis for all three predictors.

Question: Which ones are important?



Before we answer this question, we need to answer another question.

Lecture Outline

- Linear models
- Estimate of the regression coefficients
  - Brute Force
  - Exact method
  - Gradient Descent
- Confidence intervals for the predictors estimates
- **Bootstrap**
- Evaluating Significance of Predictors
- How well we know the model $\hat{f}$

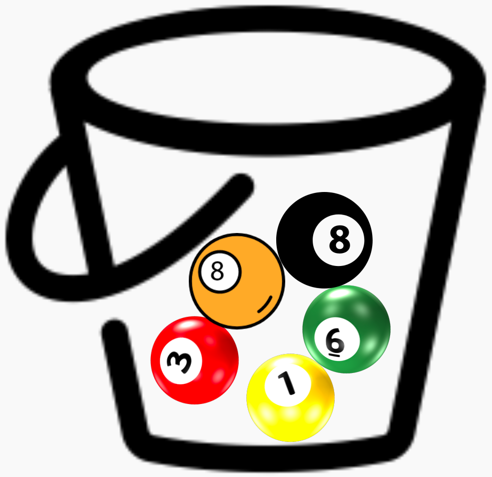
Bootstrap

In the lack of active imagination, parallel universes and the likes, we need an alternative way of producing fake data set that resemble the parallel universes.

Bootstrapping is the practice of sampling from the observed data (X,Y) in estimating statistical properties.

Bootstrap

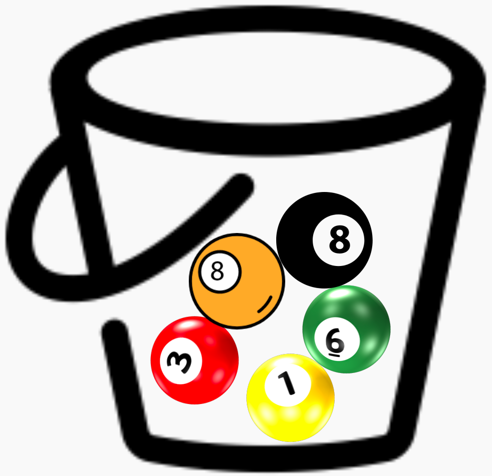
Imagine we have 5 billiard balls in a bucket.



Bootstrap

We first pick randomly a ball and replicate it. This is called **sampling with replacement.** We move the replicated ball to another bucket.

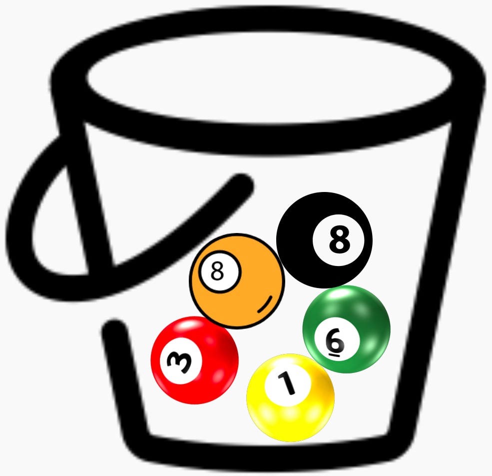
Bootstrap

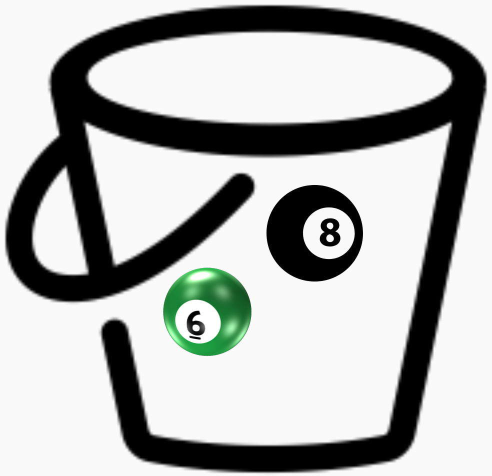


Bootstrap

We then randomly pick another ball and again we replicate it.
As before, we move the replicated ball to the other bucket.

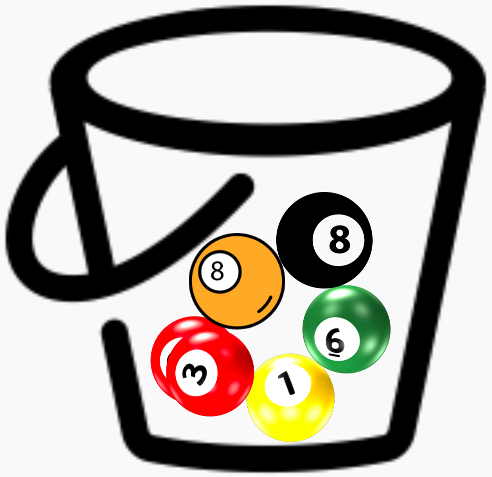
Bootstrap

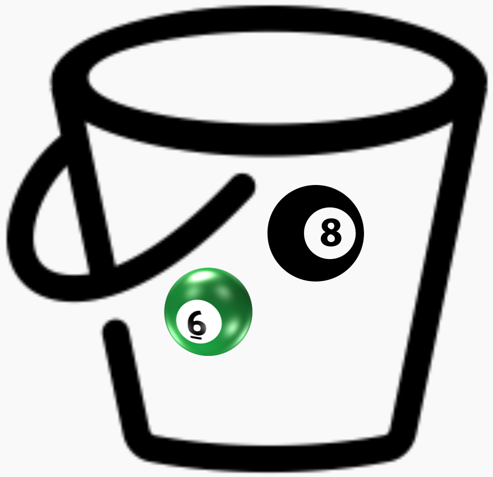




Bootstrap

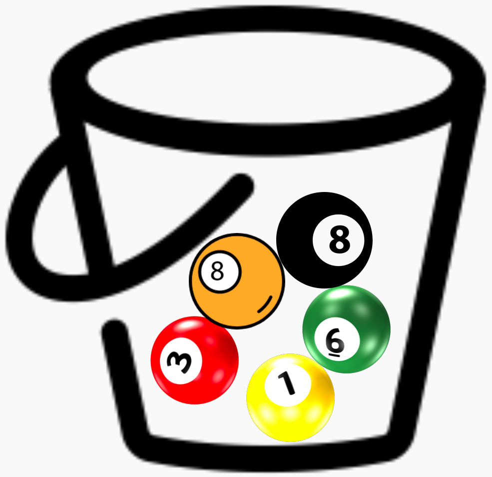
We repeat this process.

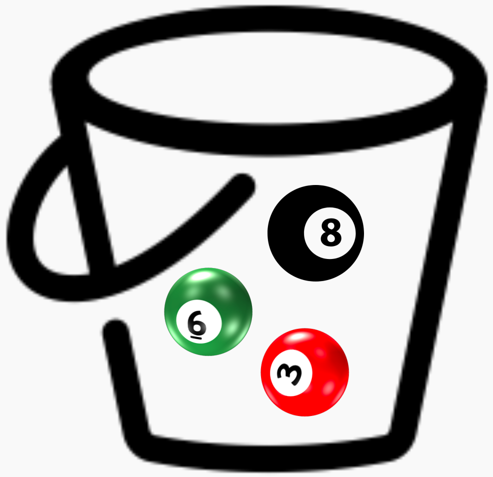




Bootstrap

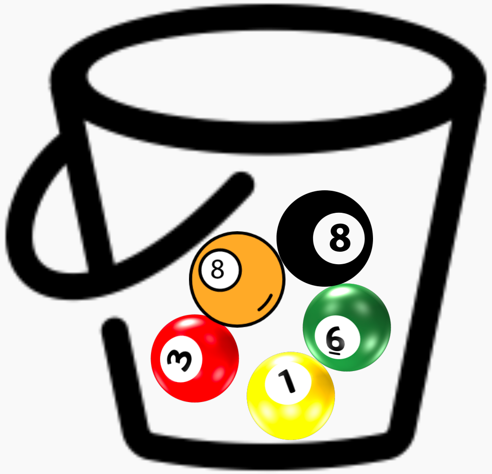
Again

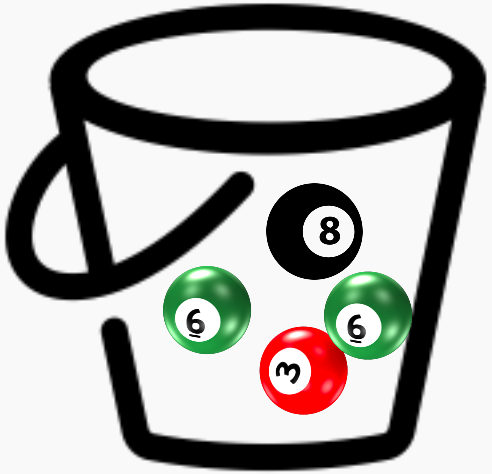




Bootstrap

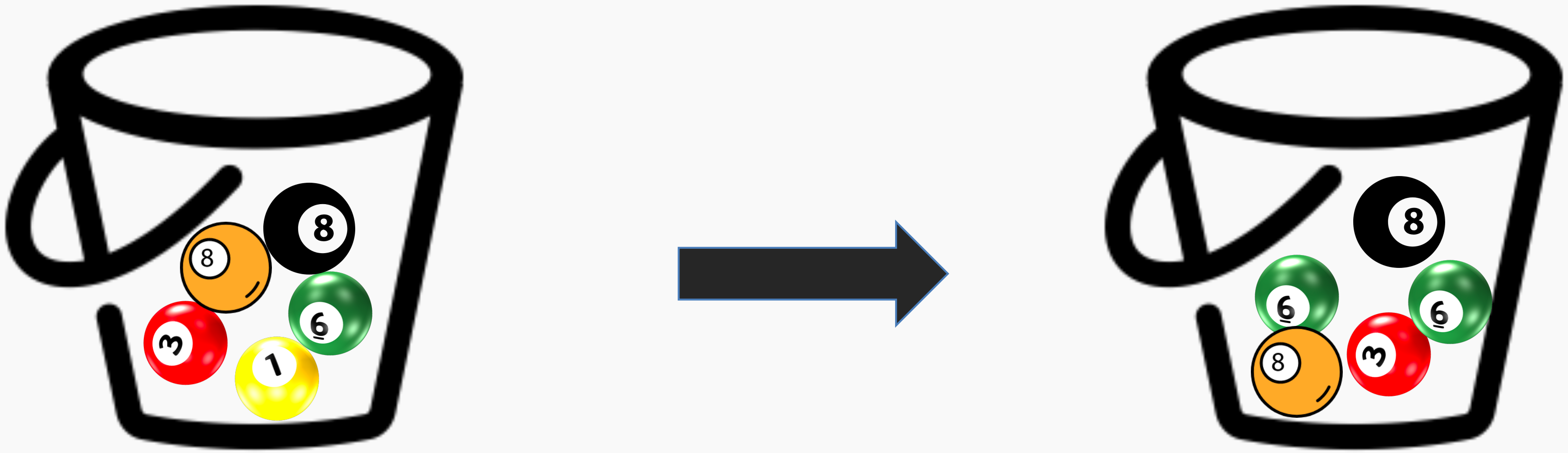
And again





Bootstrap

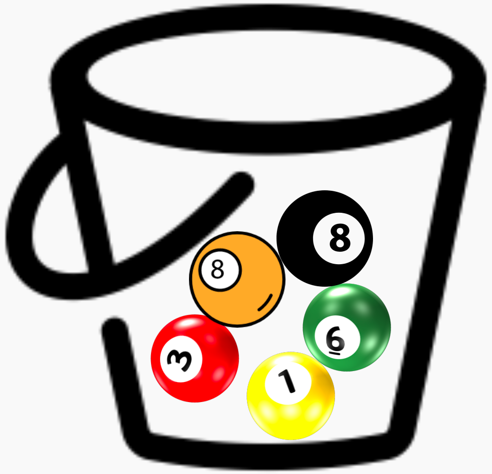
Until the "other" bucket has **the same number of balls** as the original one.

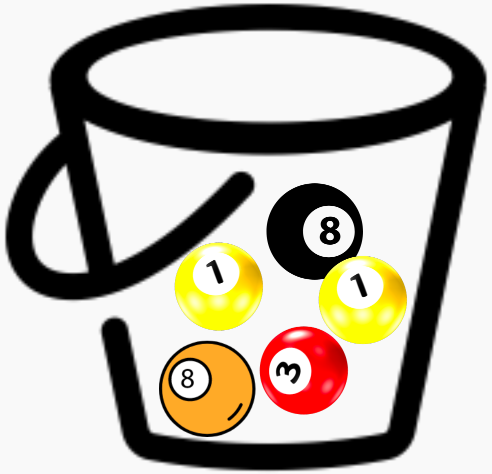


This new bucket represents a new parallel universe

Bootstrap

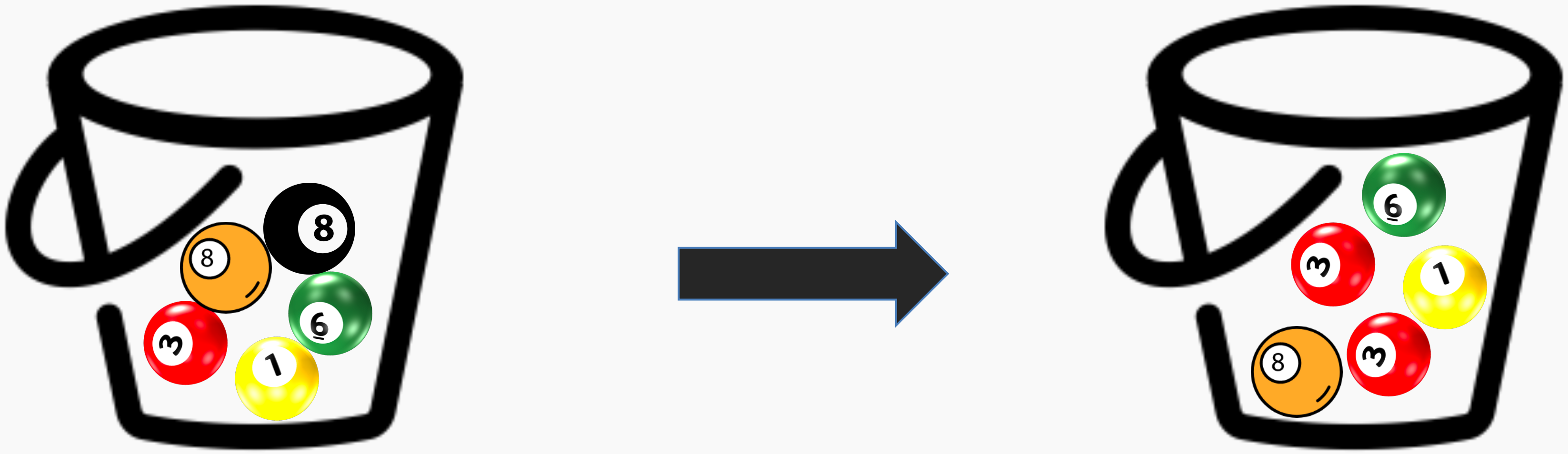
We repeat the same process and acquire another sample.





Bootstrap

We repeat the same process and acquire another sample.



These new buckets represents the parallel universes

Bootstrapping for Estimating Sampling Error

Definition

Bootstrapping is the practice of estimating properties of an estimator by measuring those properties by, for example, sampling from the observed data.

For example, we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ multiple times by randomly sampling from our data set. We then use the variance of our multiple estimates to approximate the true variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Confidence intervals for the predictors estimates: **Standard Errors**

We can empirically estimate the **standard errors**, $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ of β_0 and β_1 through bootstrapping.

If for each bootstrapped sample the estimated betas are: $\hat{\beta}_{0,i}, \hat{\beta}_{1,i}$, then

$$SE(\hat{\beta}_0) = \sqrt{\mathrm{var}(\widehat{\beta_0})}$$

$$SE(\hat{\beta}_1) = \sqrt{\mathrm{var}(\widehat{\beta_1})}$$

Confidence intervals for the predictors estimates: **Standard Errors**

Alternatively:

If we know the variance σ_ϵ^2 of the noise ϵ, we can compute $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ analytically using the formulae below (no need to bootstrap):

$$SE\left(\widehat{\beta}_0\right) = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

$$SE\left(\widehat{\beta}_1\right) = \frac{\sigma_\epsilon}{\sqrt{\sum_i \left(x_i - \overline{x}\right)^2}}$$

Standard Errors

$$SE\left(\widehat{\beta}_0\right) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

$$SE\left(\widehat{\beta}_1\right) = \frac{\sigma}{\sqrt{\sum_i \left(x_i - \overline{x}\right)^2}}$$

More data: $n \uparrow$ and $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Larger coverage: $var(x)$ or $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Better data: $\sigma^2 \downarrow \Rightarrow SE \downarrow$

In practice, we do not know the theoretical value of σ since we do not know the exact distribution of the noise ϵ.

Standard Errors

However, if we make the following assumptions,

- the errors $\epsilon_i = y_i - \hat{y}_i$ and $\epsilon_j = y_j - \hat{y}_j$ are uncorrelated, for $i \neq j$,
- each ϵ_i has a mean 0 and variance σ_ϵ^2,

then, we can empirically estimate σ^2, from the data and our regression line:

$$\sigma_\epsilon \approx \sqrt{\frac{n \cdot \text{MSE}}{n-2}} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}}$$

Remember:

$$y_i = f(x_i) + \epsilon_i \Longrightarrow \epsilon_i = y_i - f(x_i)$$

Standard Errors

More data: $n \uparrow$ and $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Larger coverage: $var(x)$ or $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Better data: $\sigma^2 \downarrow \Rightarrow SE \downarrow$

$$SE\left(\widehat{\beta}_0\right) = \sigma \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$

$$SE\left(\widehat{\beta}_1\right) = \frac{\sigma}{\sqrt{\sum_i \left(x_i - \overline{x}\right)^2}}$$

Better model: $(\hat{f} - y_i) \downarrow \Longrightarrow \sigma \downarrow \Longrightarrow SE \downarrow$

$$\sigma \approx \sqrt{\sum \frac{(\hat{f}(x) - y_i)^2}{n-2}}$$

Question: What happens to the $\widehat{\beta_0}$, $\widehat{\beta_1}$ under these scenarios?

Standard Errors

The following results are for the coefficients for TV advertising:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0061
Bootstrap	0.0061

The coefficients for TV advertising but restricting the coverage of x are:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0068
Bootstrap	0.0068

This makes no sense?

The coefficients for TV advertising but with added **extra** noise:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0028
Bootstrap	0.0023

Standard Errors

Exercise: Duplicate the following results for the coefficients for TV advertising.

Method	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$
Analytic Formula	0.353	0.0028
Bootstrap	0.328	0.0023

Lecture Outline

- Linear models
- Estimate of the regression coefficients
  - Brute Force
  - Exact method
  - Gradient Descent
- Confidence intervals for the predictors estimates
- Bootstrap
- **Evaluating Significance of Predictors**
  - **Hypothesis Testing**
- How well we know the model $\hat{f}$

Interpretation of Predictors

Question: What do you think a predictor coefficient means?

$$Sales = 7.5 + 0.04\ TV$$

What does 7.5 mean and what does 0.04 mean?

If we increase the TV by $1000, what would you expect the increase in sales to be?

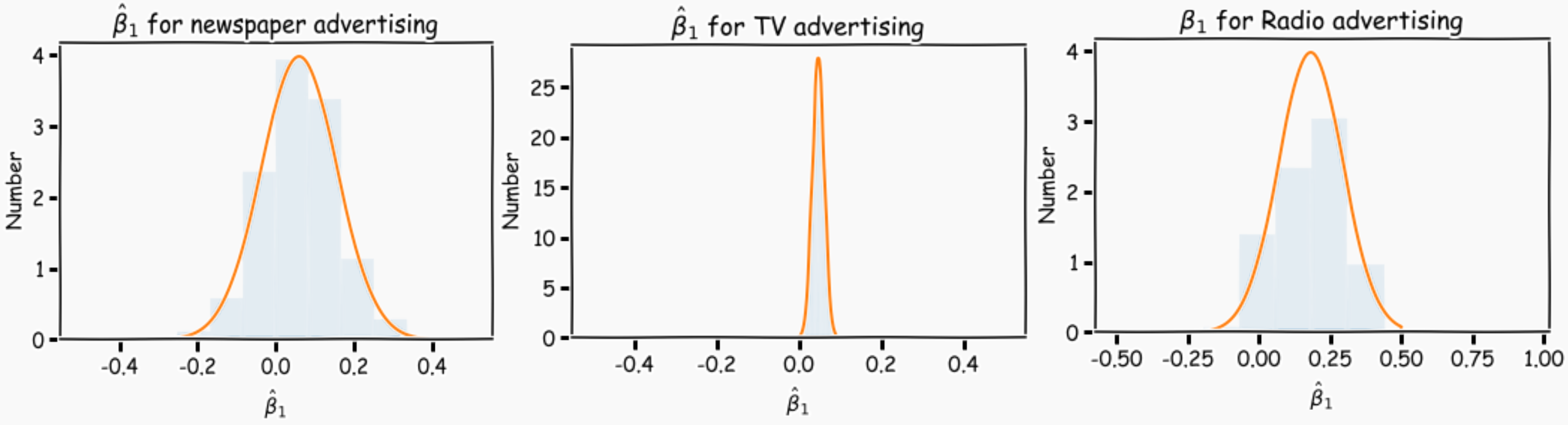
What if?

$$Sales = 7.5 + 1.01\ TV$$

The interpretation of the predictors depends on the values but decisions depend on how much we trust these values.

And also we can answer the question, 'how significant are the predictors?' Here we show the same analysis for all three predictors.

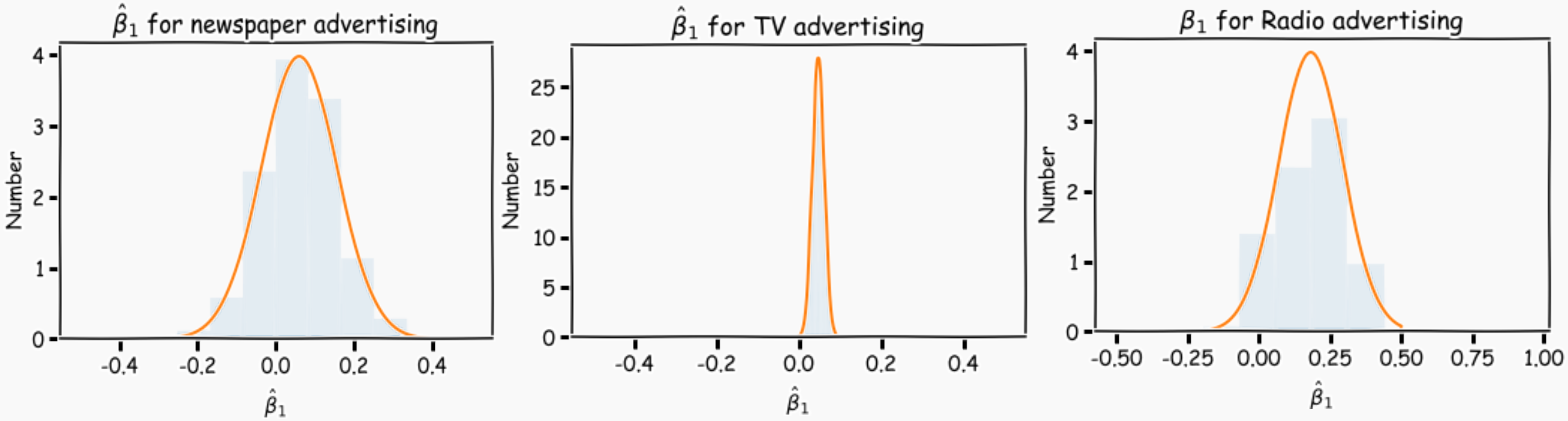
Question: Which ones are important?



Before we answer this question, we need to answer another question.

And also we can answer the question, ’how significant are the predictors?’ Here we show the same analysis for all three predictors.

Question: Which ones are important?



Now we know how to generate these distributions we are ready to answer ***’how significant are the predictors?’***

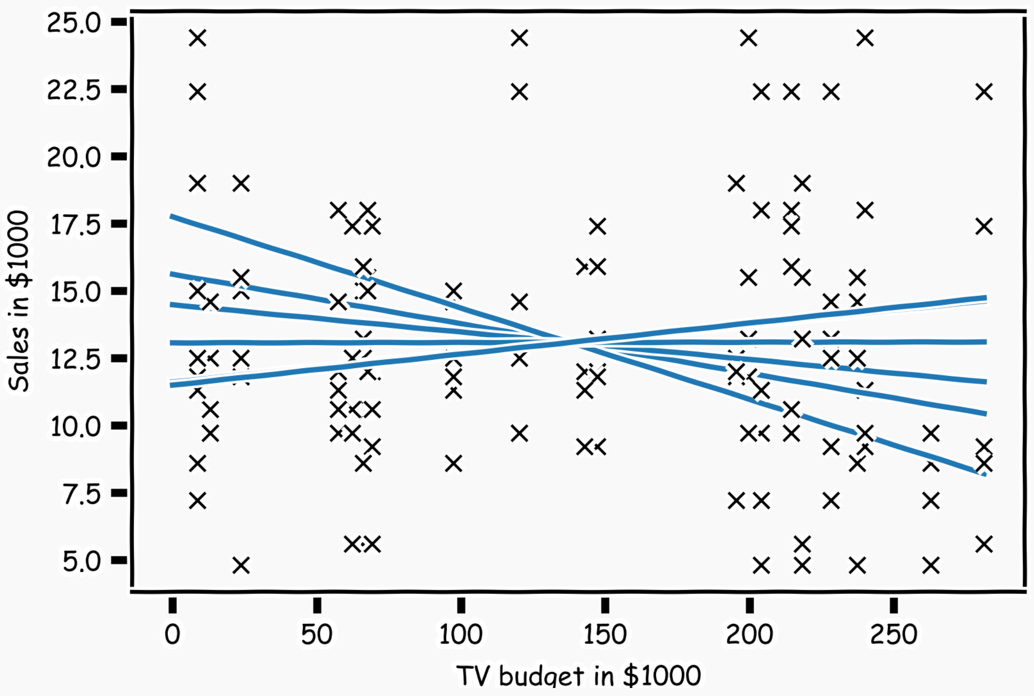
Hypothesis Testing

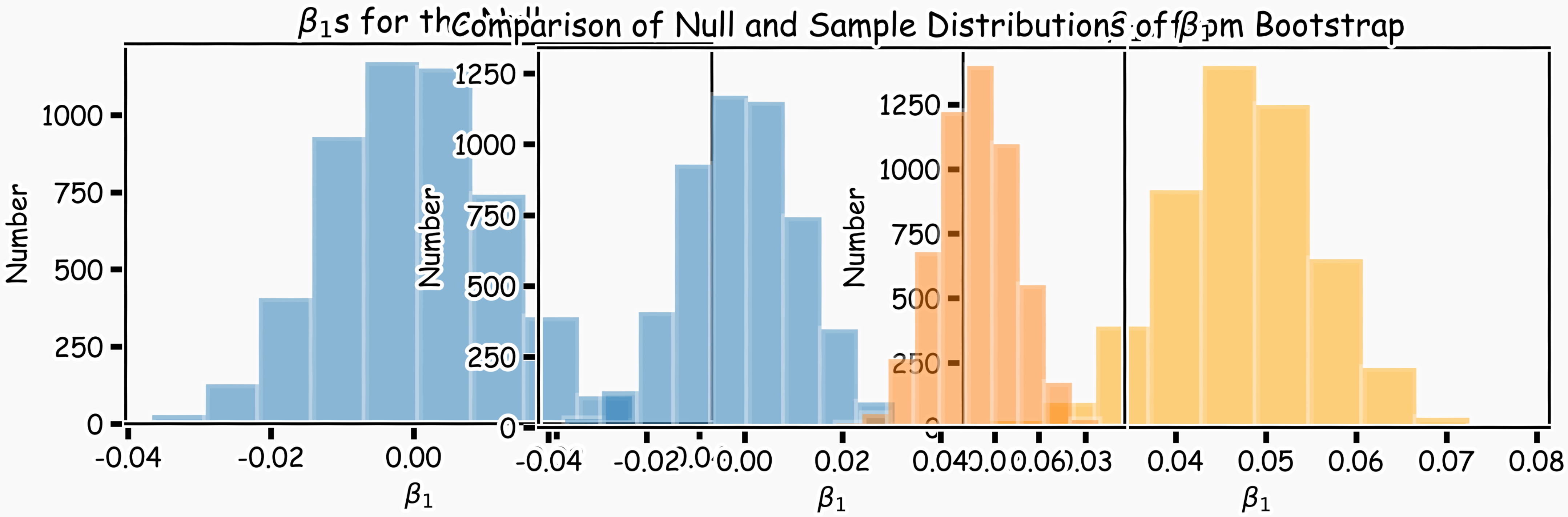
Hypothesis testing is a formal process through which we evaluate the validity of a statistical hypothesis by considering evidence **for** or **against** the hypothesis gathered by ***random sampling of the data***.

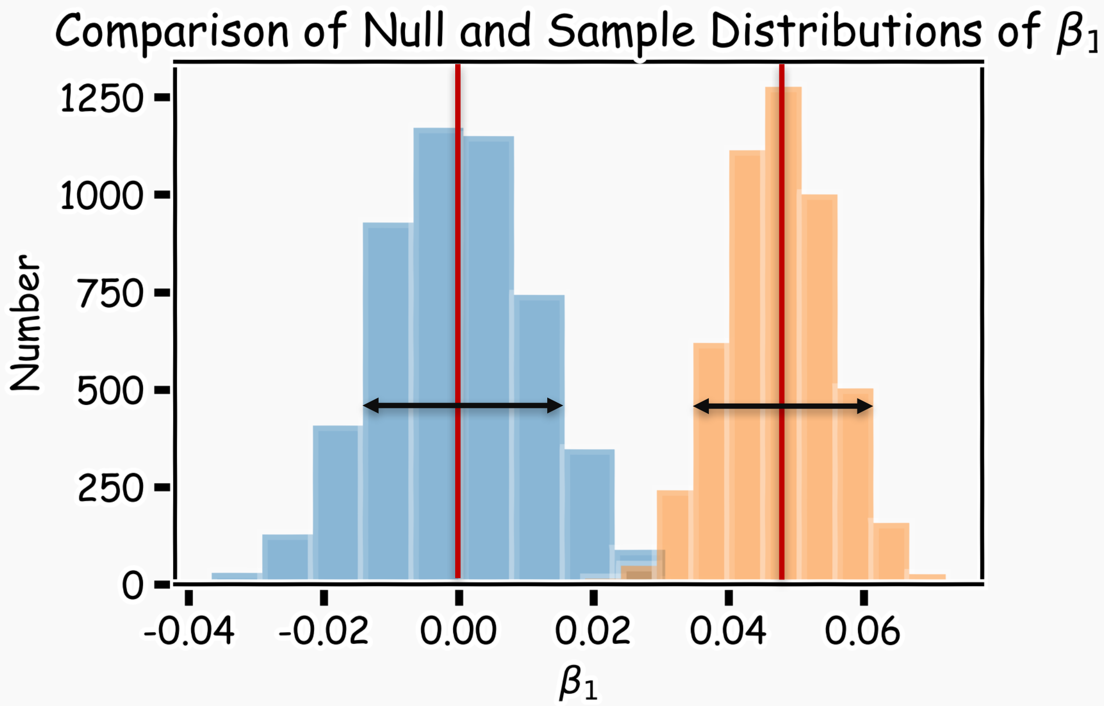
TV	sales
[illegible]	22.1
[illegible]	10.4
[illegible]	9.3
[illegible]	18.5
[illegible]	12.9
[illegible]	7.2
[illegible]	11.8
[illegible]	13.2
[illegible]	4.8
[illegible]	10.6
[illegible]	8.6
[illegible]	17.4
[illegible]	9.2
[illegible]	9.7
[illegible]	19.0
[illegible]	22.4
[illegible]	12.5
[illegible]	24.4

Random sampling of the data

Shuffle the values of the predictor variable




Comparison of Null and Sample Distributions of
Number
β_1



$$\mu_{Null} = 0$$

$$\mu_{\hat{\beta}} = \mu_{boot}$$

$$\sigma_{\hat{\beta}} = SE(\hat{\beta}) = \sigma_{boot}$$

$$\sigma_{Null} \approx \sigma_{\hat{\beta}}$$

Translate this to the significance. Let's look at the distance of the estimated value of the coefficient in units of SE($\hat{\beta}_1$) = $\sigma_{\hat{\beta}_1}$.

$$\boxed{D = \frac{\mu_{\hat{\beta}} - \mu_{Null}}{\sqrt{\sigma_{Null}^2 + \sigma_{\hat{\beta}}^2}} = \frac{\mu_{\hat{\beta}}}{\sqrt{\sigma_{Null}^2 + \sigma_{\hat{\beta}}^2}} = \frac{\mu_{\hat{\beta}}}{\sqrt{2}\sigma_{\hat{\beta}}}}$$

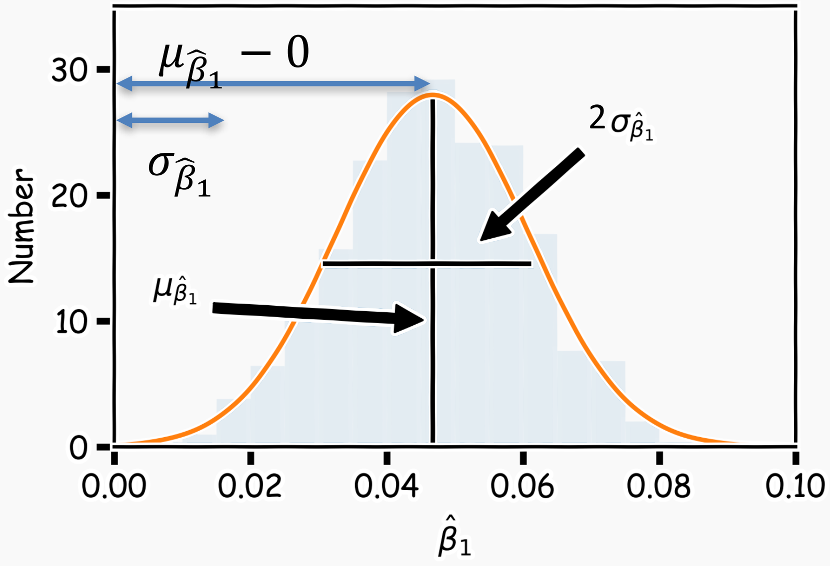
Importance of predictors

In practice, we do not need the distribution for Null.

Define a test statistic, which we we call t-test statistic

$$t = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

Which measures the distance from zero in units of standard deviation.



We evaluate how often a particular value of t can occur by accident. We expect that t will have a *t-distribution with n-2 degrees* of freedom.

To compute the probability of observing any value equal to $|t|$ or larger, assuming $\hat{\beta}_1 = 0$ is easy. We call this probability the **p-value**.

a small p-value (<0.05) indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.

Hypothesis Testing

Hypothesis testing is a formal process through which we evaluate the validity of a statistical hypothesis by considering evidence for or against the hypothesis gathered by **random sampling** of the data.

1. State the hypotheses, typically a **null hypothesis**, H_0 and an **alternative hypothesis**, H_1, that is the negation of the former.
2. Choose a type of analysis, i.e. how to use sample data to evaluate the null hypothesis. Typically this involves choosing a single test statistic.
3. **Sample** data and compute the test statistic.
4. Use the value of the test statistic to either reject or not reject the null hypothesis.

Hypothesis testing

1. State Hypothesis:

Null hypothesis:

H_0: There is no relation between X and Y

The alternative:

H_a: There is some relation between X and Y

2: **Choose test statistics**

To test the null hypothesis, we need to determine whether, our estimate for $\hat{\beta}_1$, is sufficiently far from zero that we can be confident that $\hat{\beta}_1$ is non-zero. We use the following test statistic:

$$t = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

Hypothesis testing

3. **Sample**:

   Using bootstrap we can estimate $\hat{\beta}_1' s$, and therefore $\mu_{\hat{\beta}_1}$ and $\sigma_{\hat{\beta}_1}$.

4. **Reject or not reject the hypothesis:**

   If there is really no relationship between X and Y, then we expect that will have a *t-distribution with n-2 degrees* of freedom.

   To compute the probability of observing any value equal to $|t|$ or larger, assuming $\hat{\beta}_1 = 0$ is easy. We call this probability the p-value.

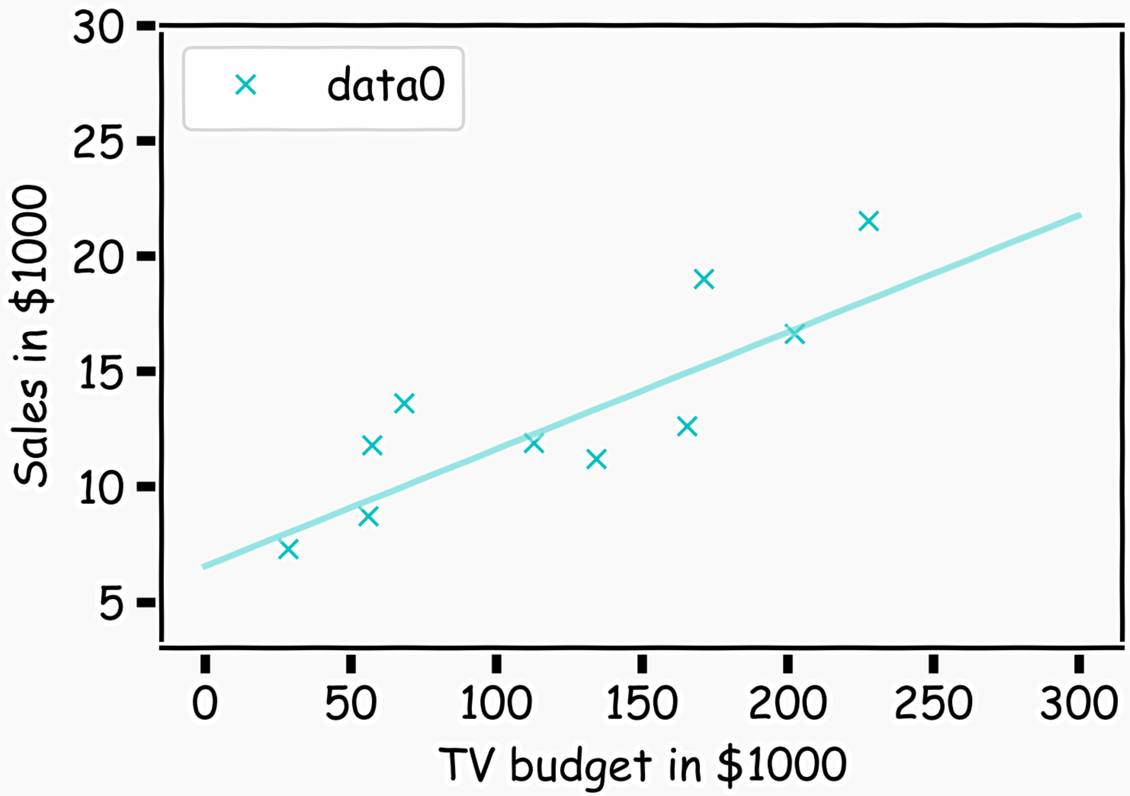
   a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance

Lecture Outline

- Linear models
- Estimate of the regression coefficients
  - Brute Force
  - Exact method
  - Gradient Descent
- Confidence intervals for the predictors estimates
- Bootstrap
- Evaluating Significance of Predictors
  - Hypothesis Testing
- **How well we know the model $\hat{f}$**

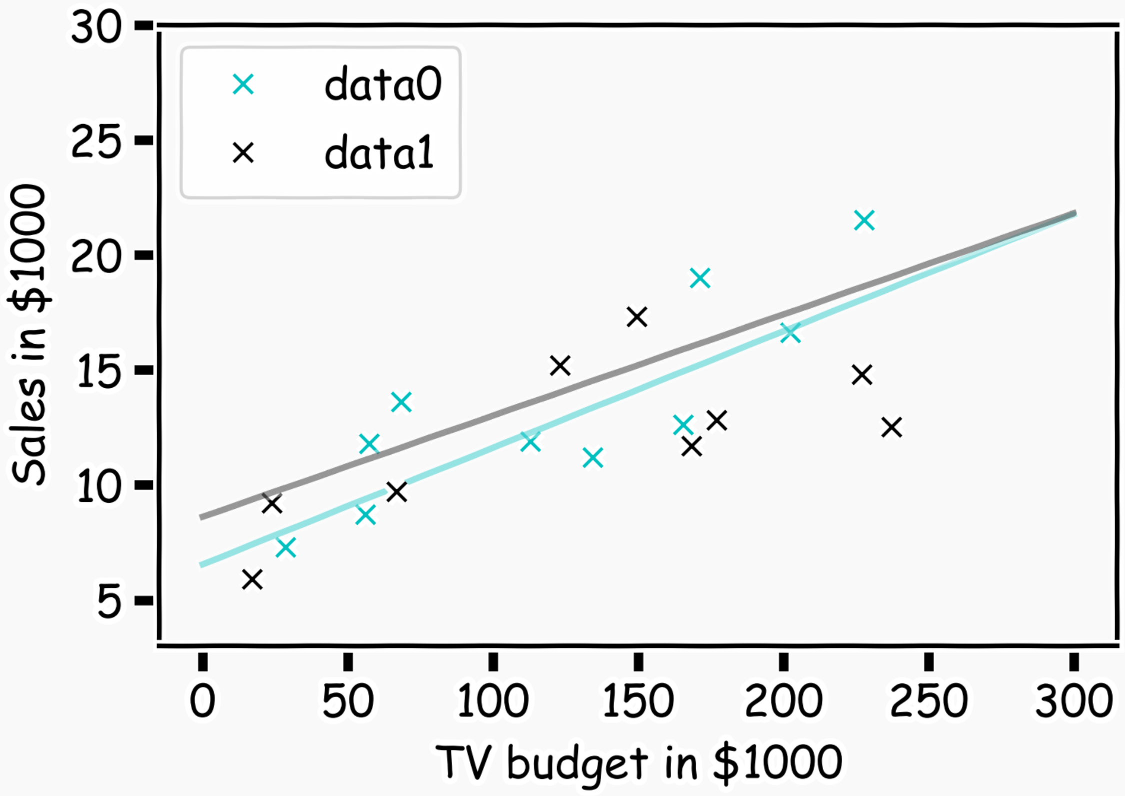
How well do we know $\hat{f}$?

Our confidence in f is directly connected with the confidence in βs. So for each bootstrap sample, we have one β which we can use to determine the model.



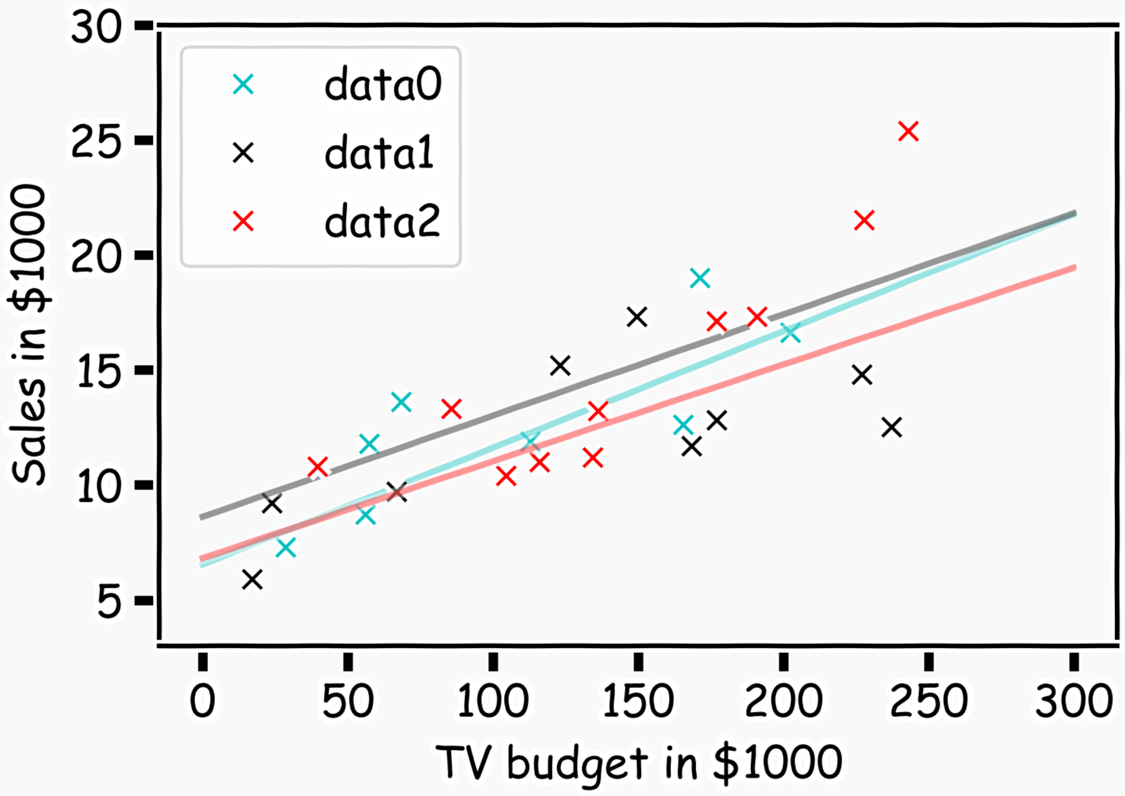
How well do we know $\hat{f}$?

Here we show two difference set of models given the fitted coefficients.



How well do we know $\hat{f}$?

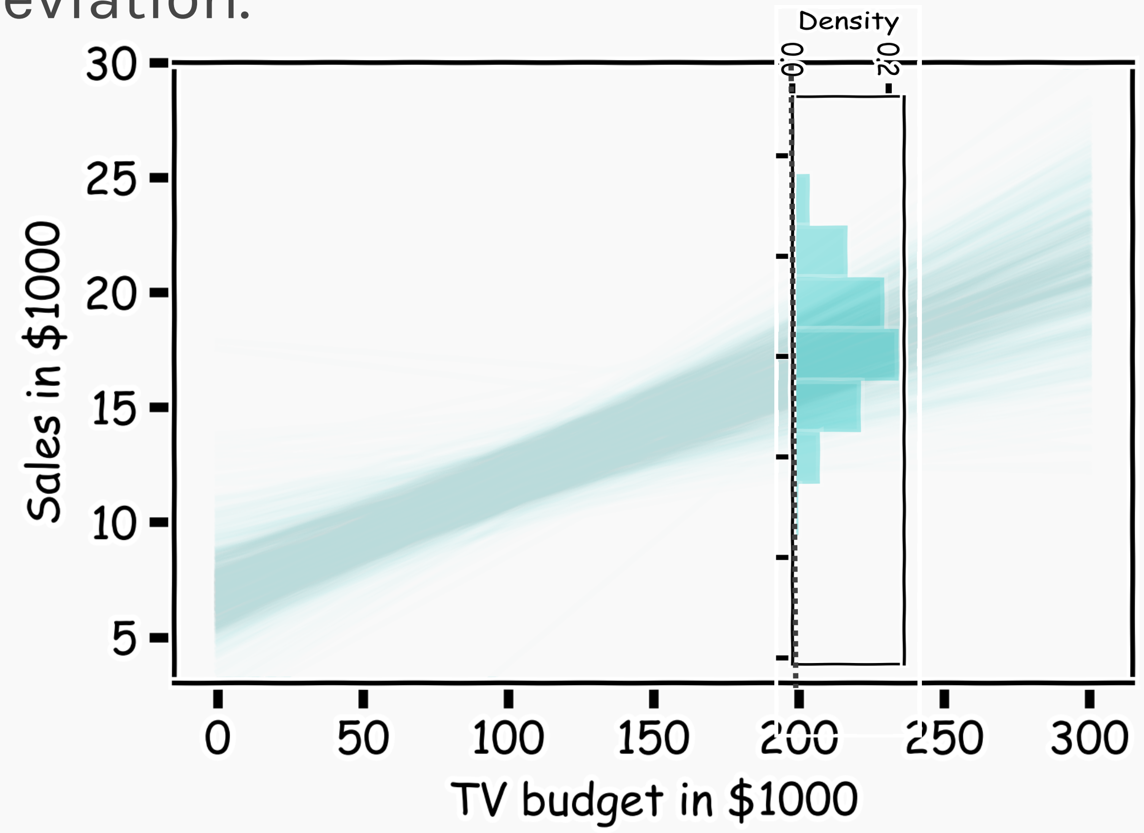
There is one such regression line for every bootstrapped sample.



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such bootstrapped samples.

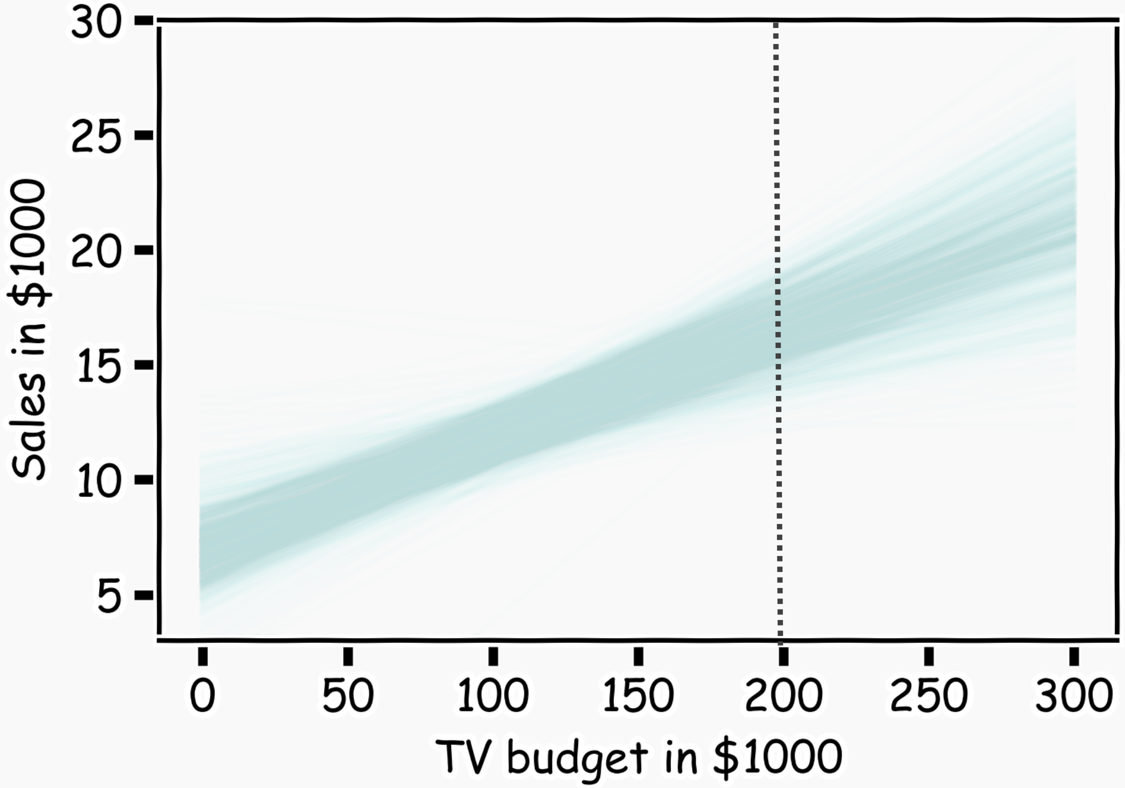
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.

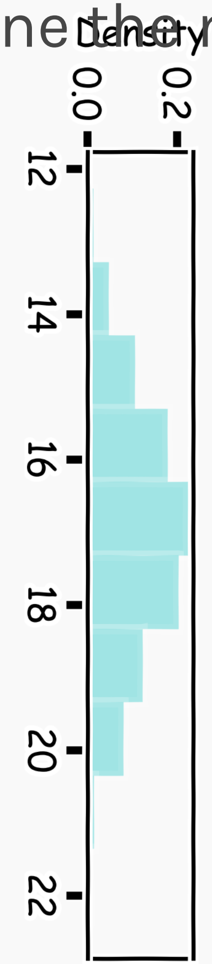


How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.

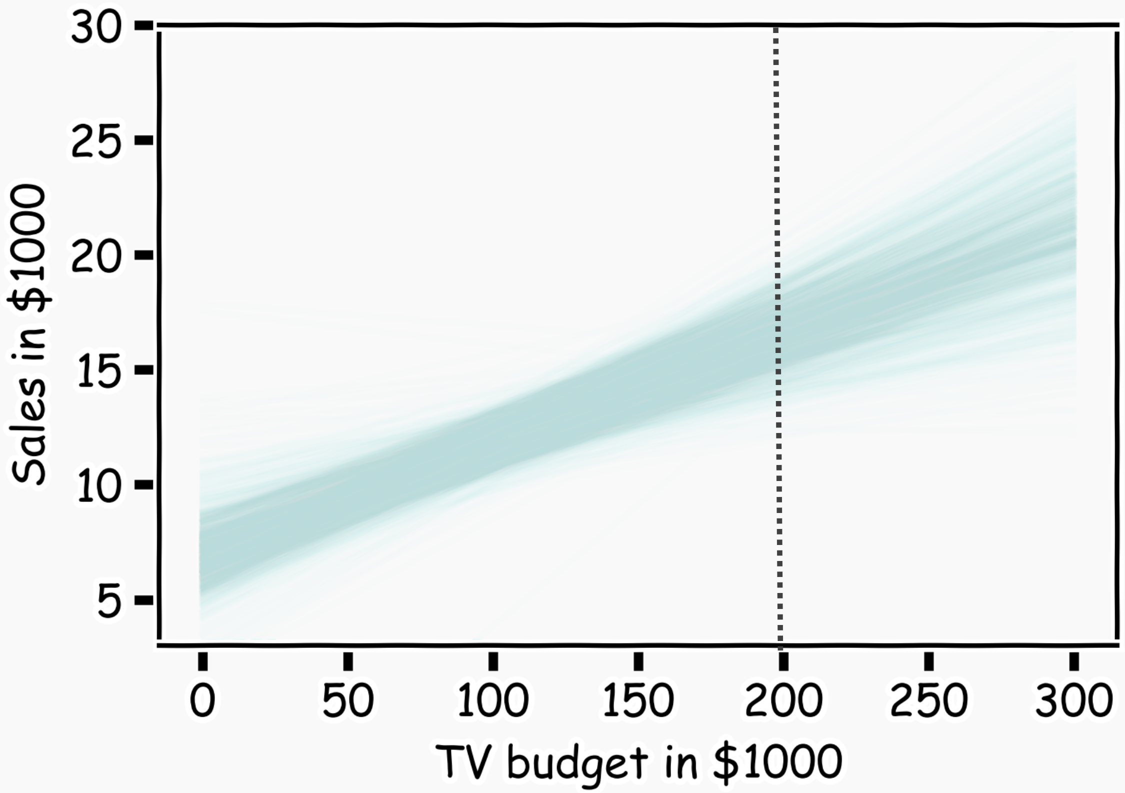


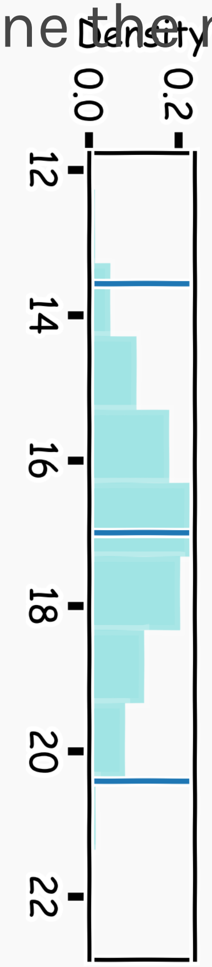


How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such sub-samples.

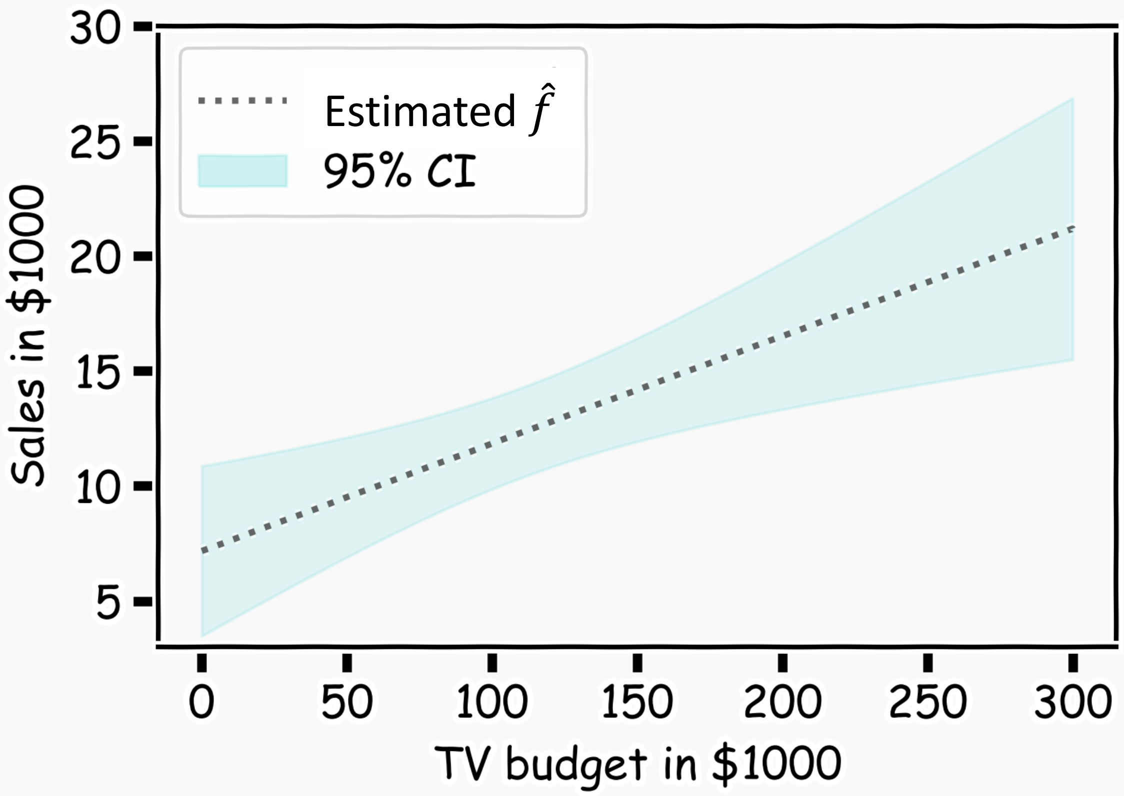
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.



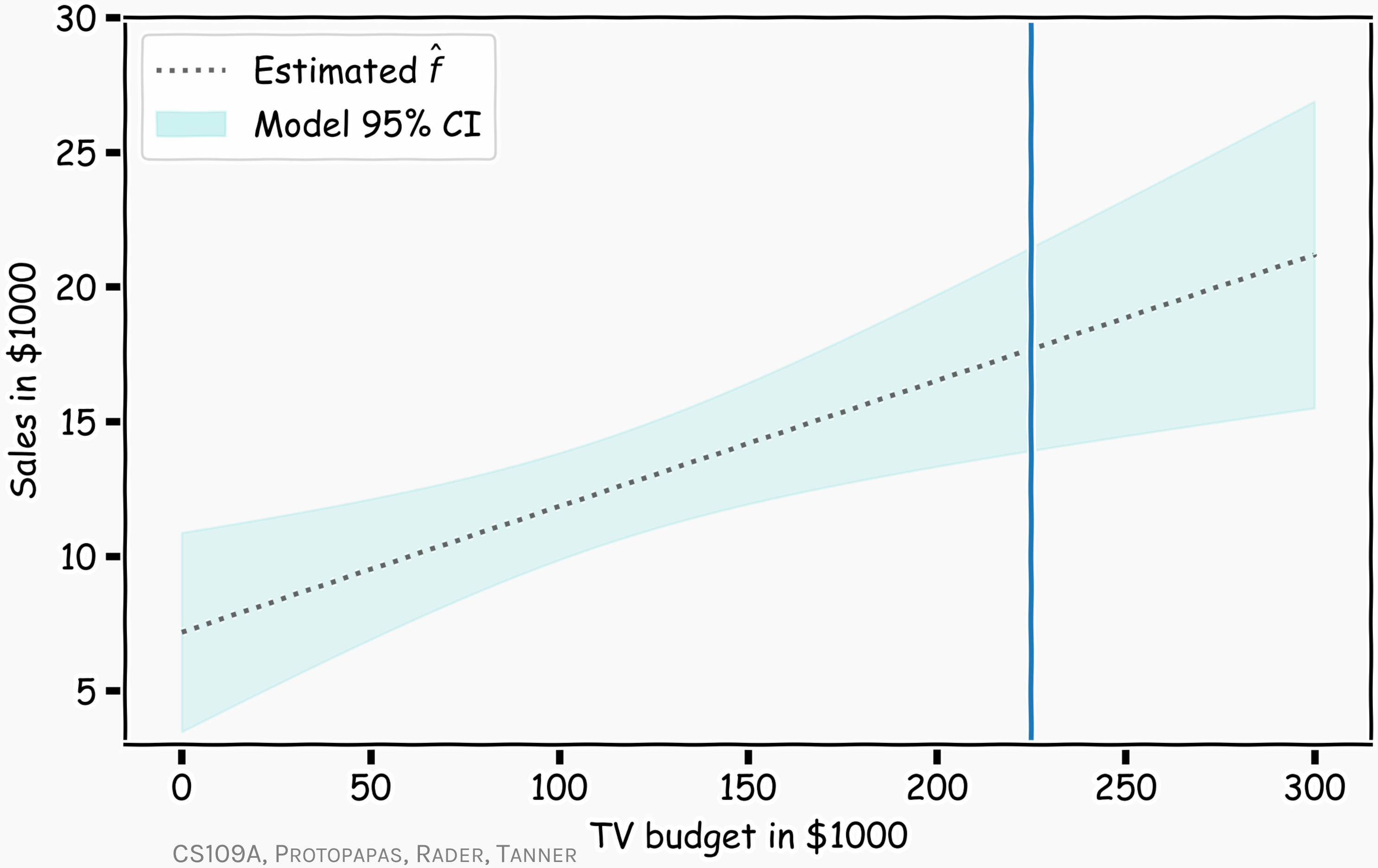


How well do we know $\hat{f}$?

For every x, we calculate the mean of the models, $\hat{f}$ (shown with dotted line) and the 95% CI of those models (shaded area).

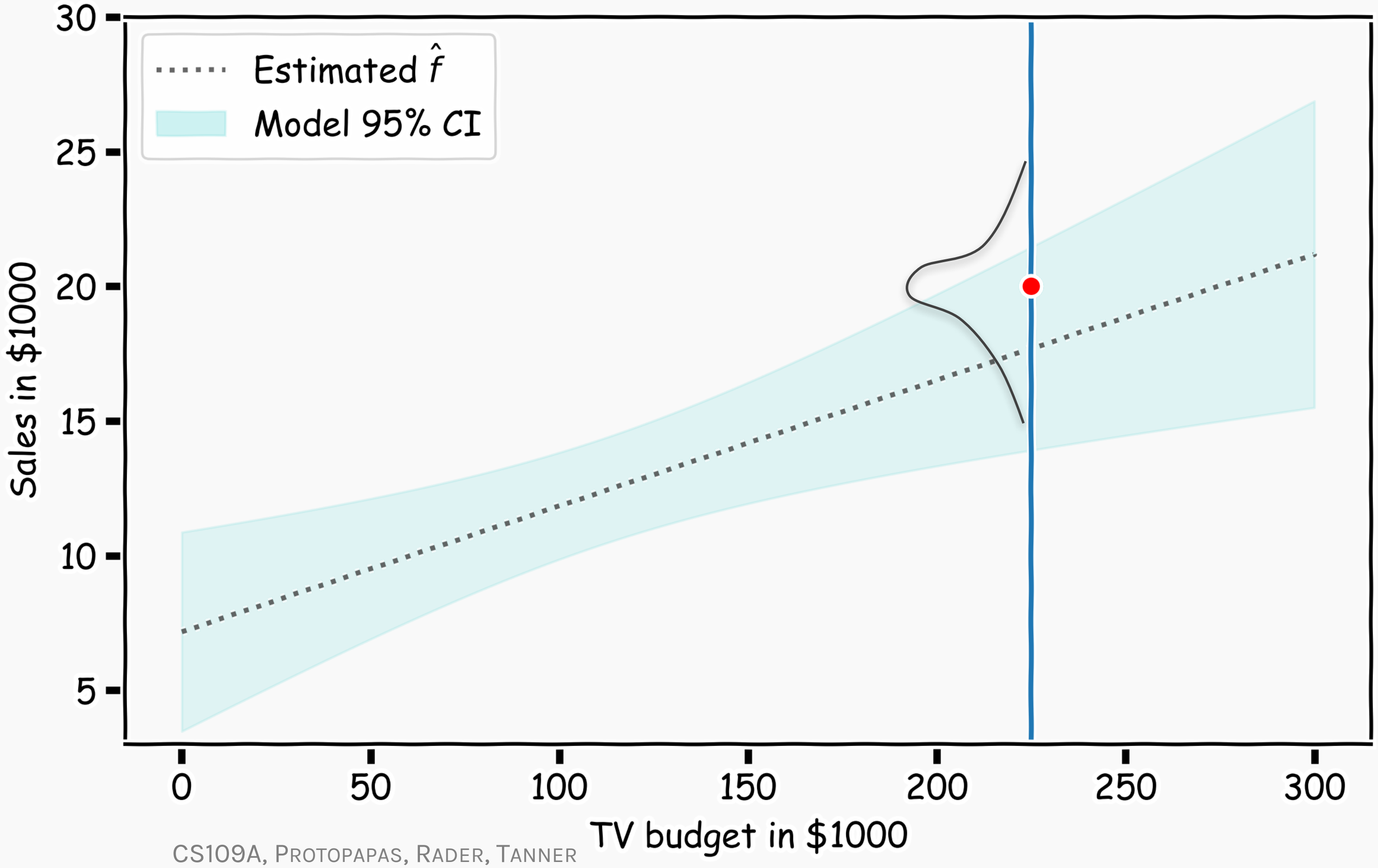


Confidence in predicting $\hat{y}$



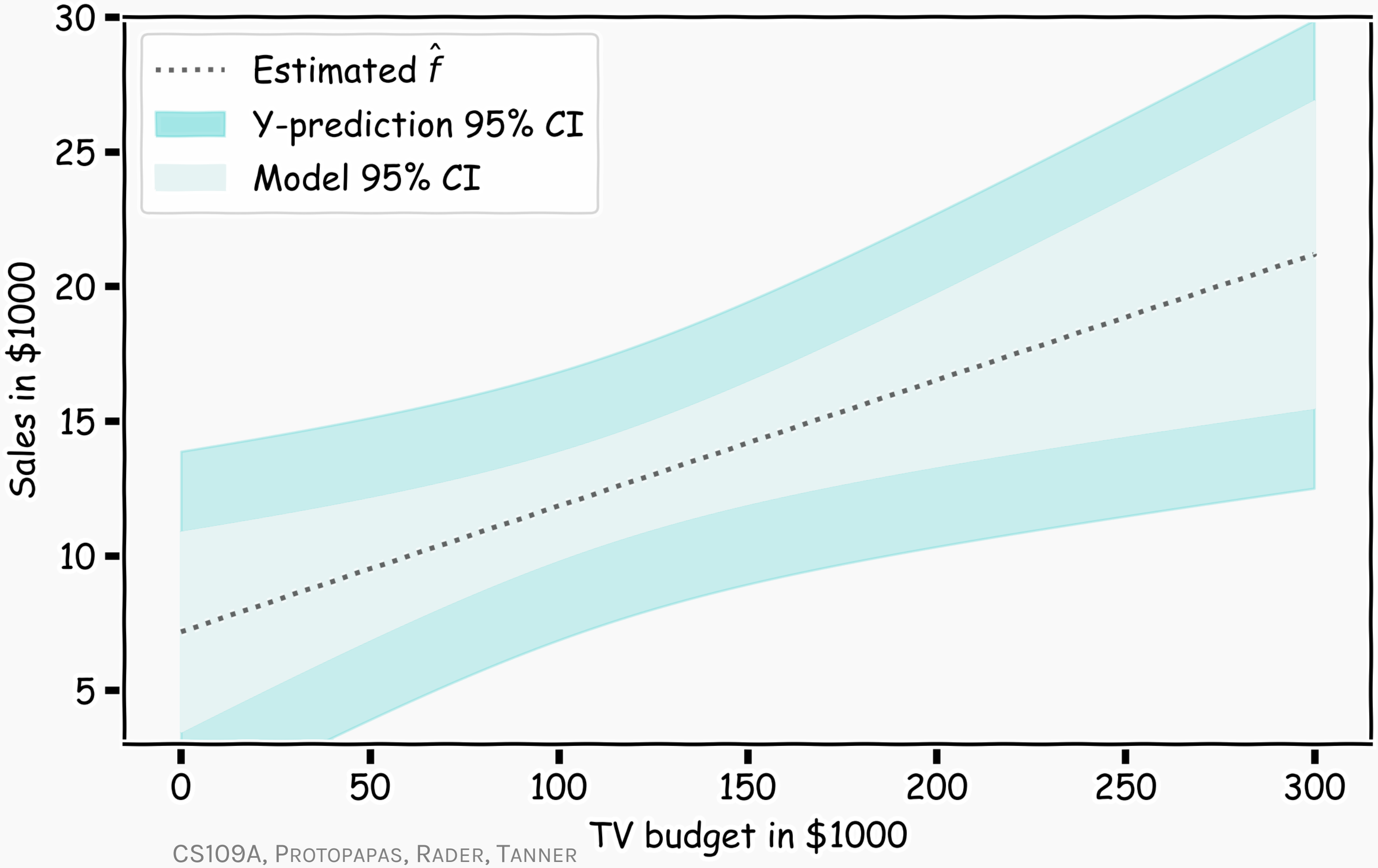
Confidence in predicting $\hat{y}$

- for a given x, we have a distribution of models $f(x)$
- for each of these $f(x)$, the prediction for $y \sim N(f, \sigma_\epsilon)$



Confidence in predicting $\hat{y}$

- for a given x, we have a distribution of models $f(x)$
- for each of these $f(x)$, the prediction for $y \sim N(f, \sigma_\epsilon)$
- The prediction confidence intervals are then



Summary

- Linear models
- Estimate of the regression coefficients
  - Brute Force
  - Exact method
  - Gradient Descent
- Confidence intervals for the predictors estimates
- Bootstrap
- Evaluating Significance of Predictors
  - Hypothesis Testing
- How well we know the model $\hat{f}$

Summary so far

Model Fitness

How does the model perform predicting?

Comparison of Two Models

How do we choose from two different models?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

What's next?

Multiple predictors

Collinearity

Categorical variables

Polynomial regression

Interaction terms

Afternoon Exercises

Quiz - to be completed in the next 10 min:

Sway: Lecture 5: Linear Regression

Programmatic – to be completed by lab time tomorrow:

Lessons: Lecture 5: Linear Regression – three (3) exercises