CPSC 4300/6300: Applied Data Science

2024

Linear Regression

Dr. Aaron J. Masino
Associate Professor, School of Computing

College of
ENGINEERING, COMPUTING
AND APPLIED SCIENCES

Summary from regression introduction lecture

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How does the model perform predicting?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

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

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

Lecture Outline

- Linear models
- Estimate regression coefficients for a single predictor with bias term
- Confidence intervals for the predictor estimates
- Bootstrap
- Evaluating significance of predictors
- How well do we know the model $\hat{f}$
- What happens with multiple predictors?

Lecture Outline

- Linear models
- Estimate regression coefficients for single predictor with bias term
- Confidence intervals for the predictor estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$
- What happens with multiple predictors?

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:

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

Bias term

Linear Regression

... then it follows that our estimate is:

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

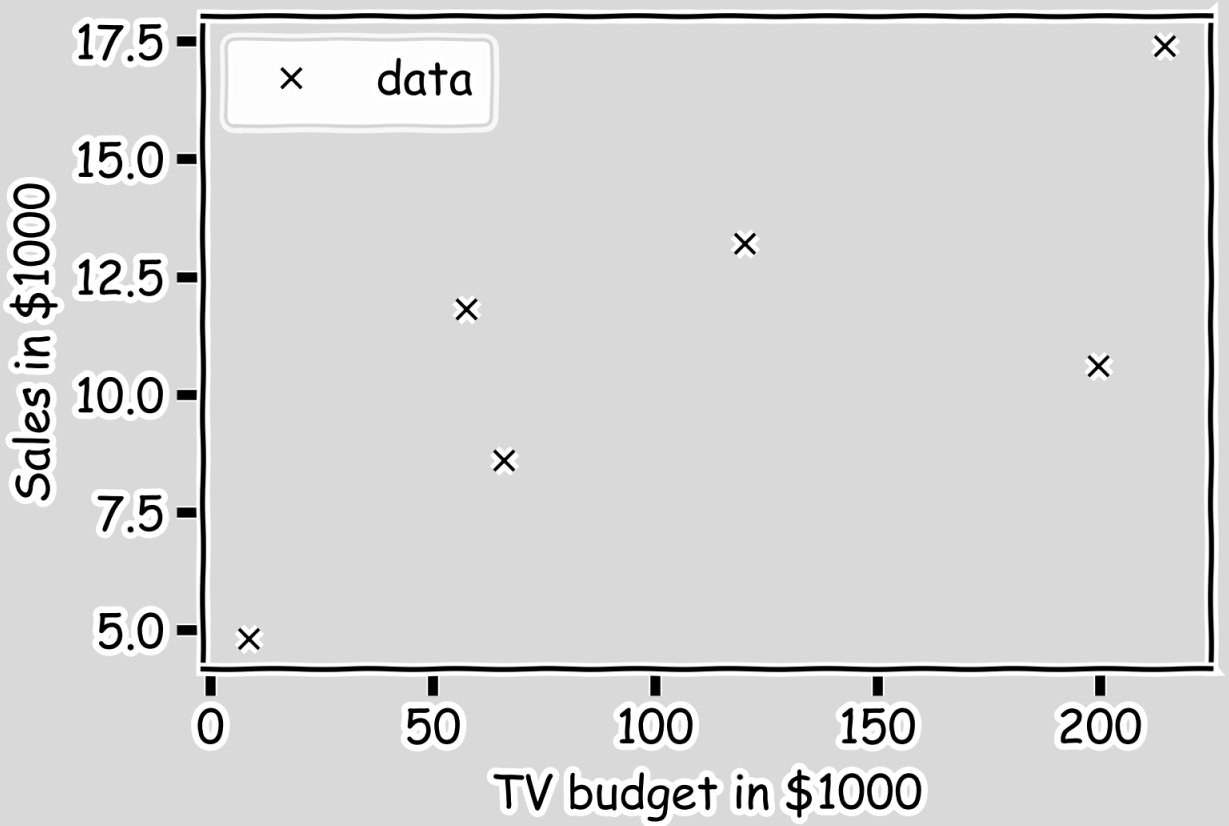
where $\hat{\beta}_1$ and $\hat{\beta}_0$ are **estimates** of β_1 and β_0 respectively, that we compute using observations.

Lecture Outline

- Linear models
- Estimate regression coefficients for single predictor with bias term
- Confidence intervals for the predictor estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$
- What happens with multiple predictors?

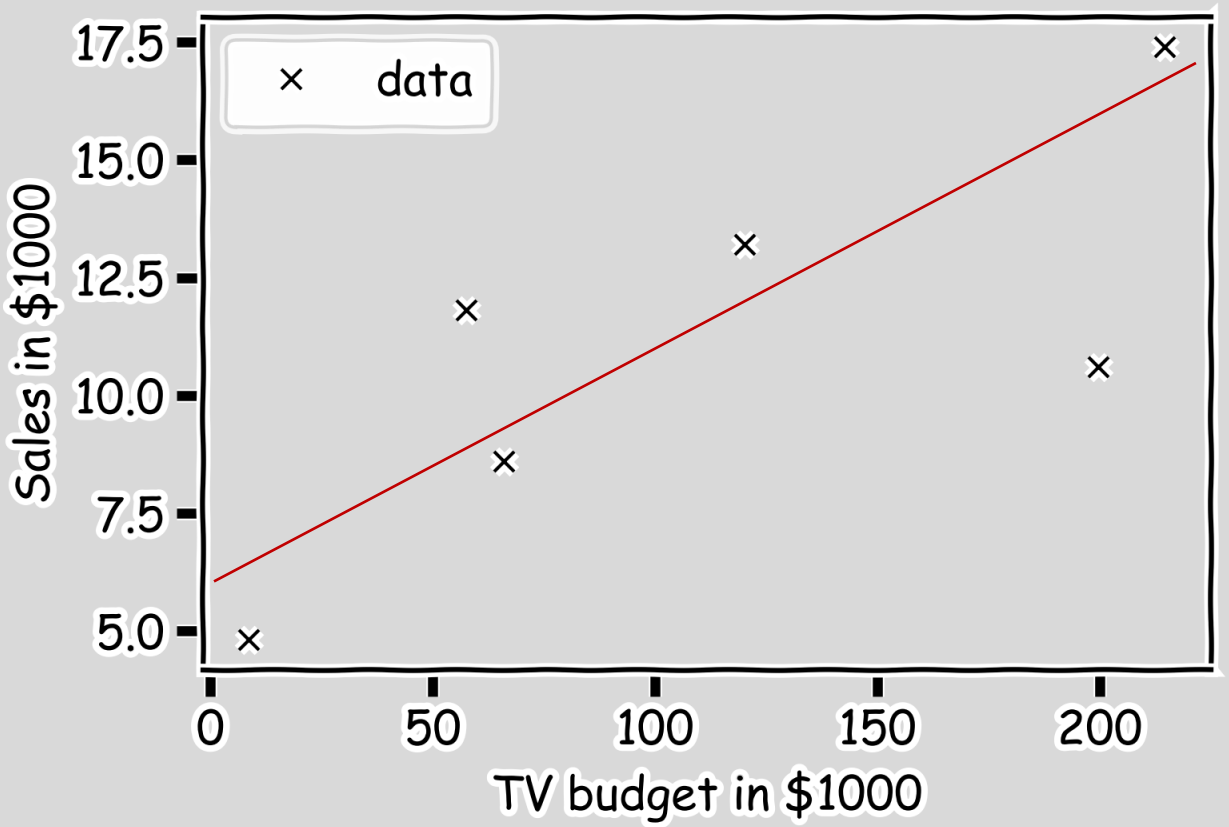
Estimate of the regression coefficients (single predictor)

For a given data set



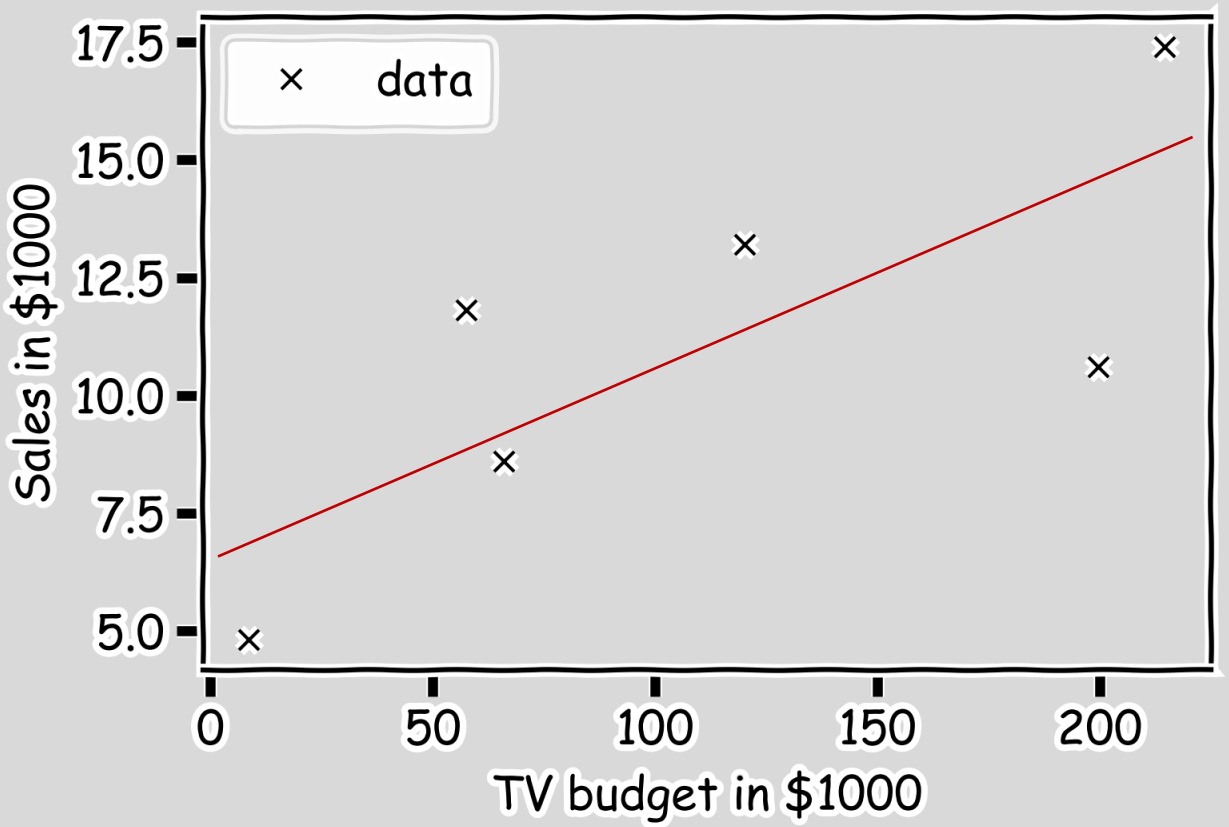
Estimate of the regression coefficients (single predictor)

Is this a good line?



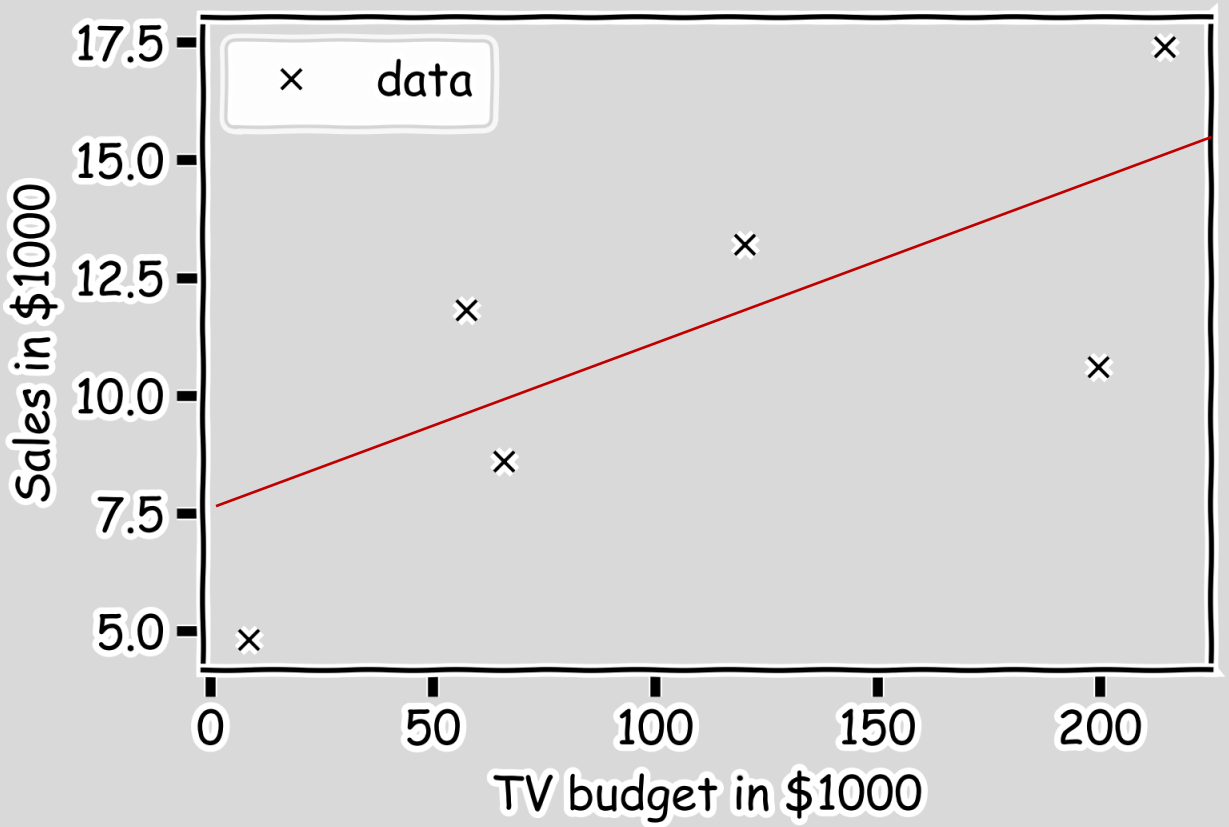
Estimate of the regression coefficients (single predictor)

Or maybe this one?



Estimate of the regression coefficients (single predictor)

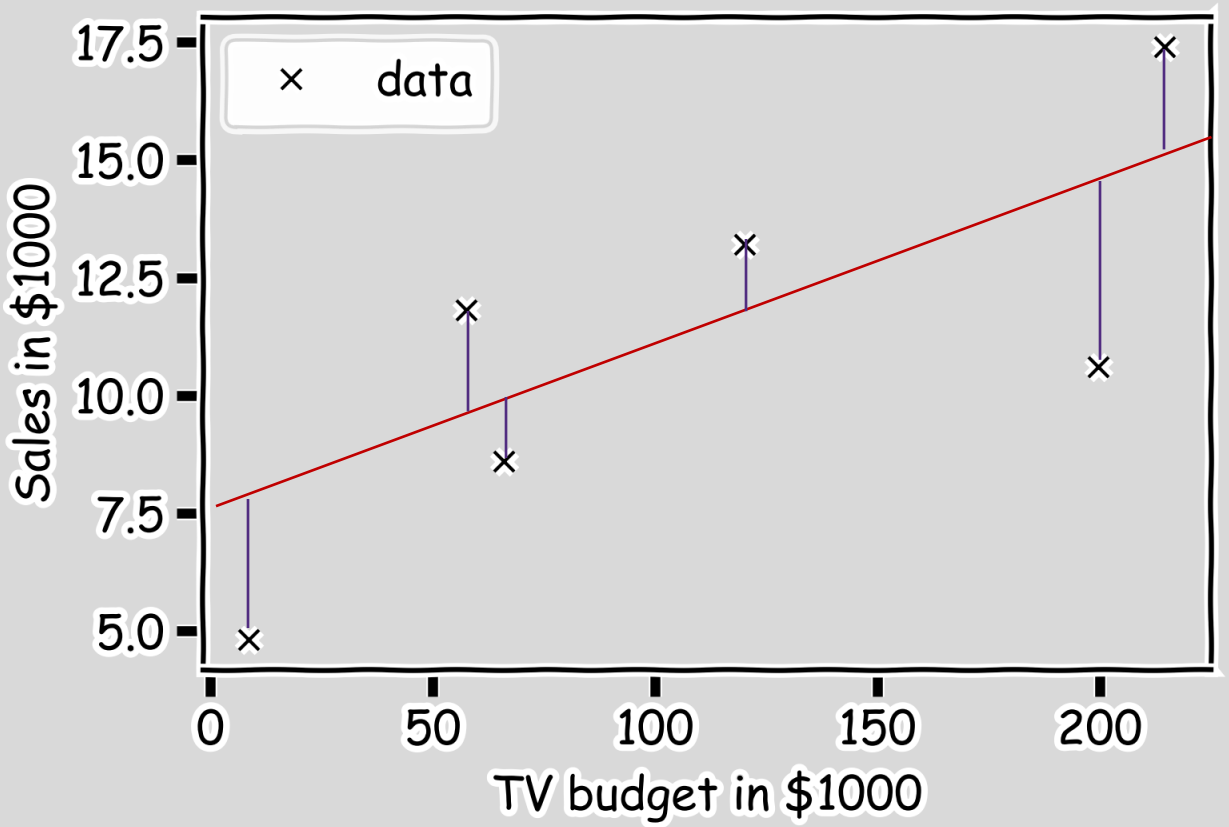
Or this one?



Estimate of the regression coefficients (single predictor)

Question: Which line is the best?

First calculate the residuals



Estimate of the regression coefficients (single predictor)

We use MSE as our **loss function**,

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

We choose $\hat{\beta}_1$ and $\hat{\beta}_0$ that 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 (single predictor): 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 $\hat{\beta}_1$:

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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.

Estimate of the regression coefficients: Exact Method

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

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

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

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

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

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

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

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

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

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

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

Lecture Outline

- Linear models
- Estimate regression coefficients for single predictor with bias term
- Confidence intervals for the predictor estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$
- What happens with multiple predictors?

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.

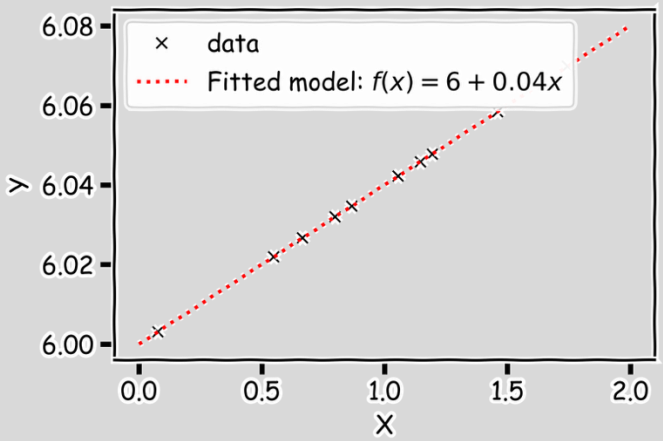
Confidence intervals for the predictor estimates

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

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

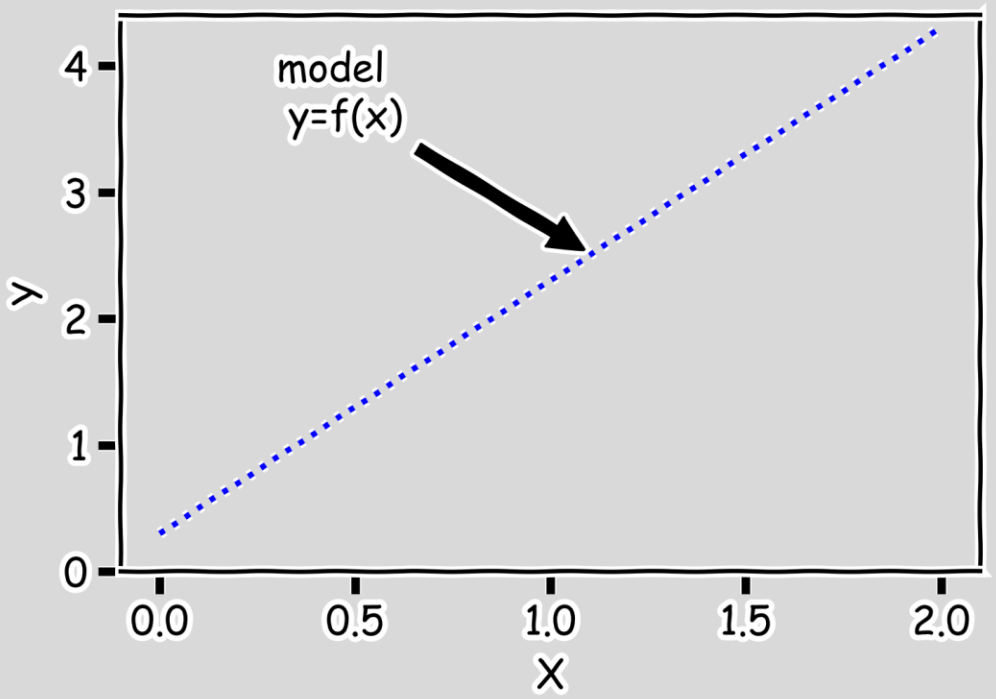
We will first address ε

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 $\hat{\beta}'s$.

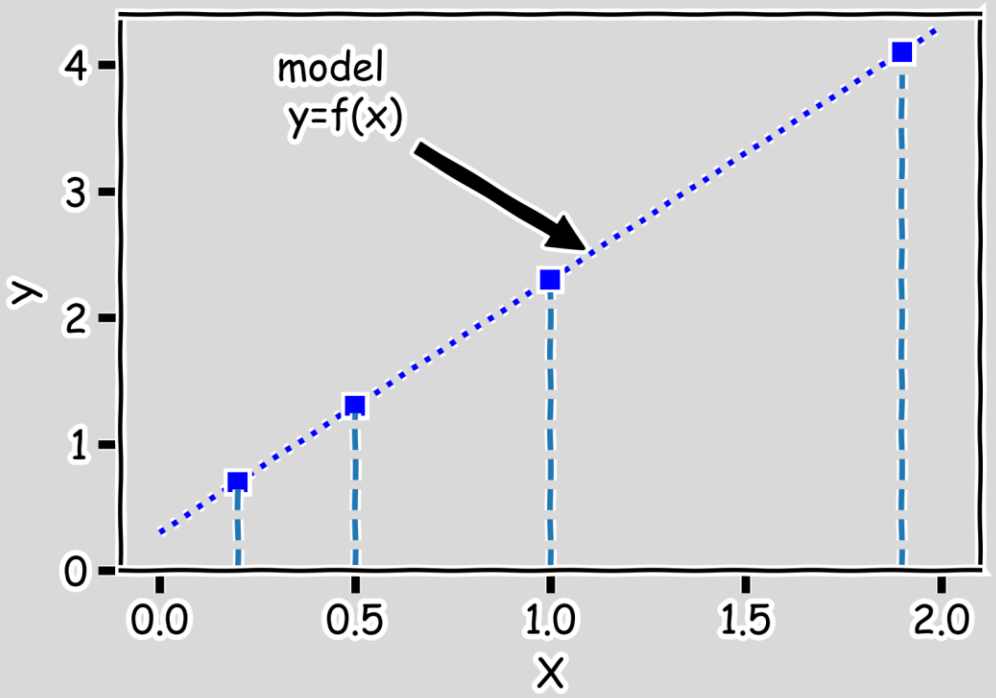
Confidence intervals for the predictor estimates

Start with a model



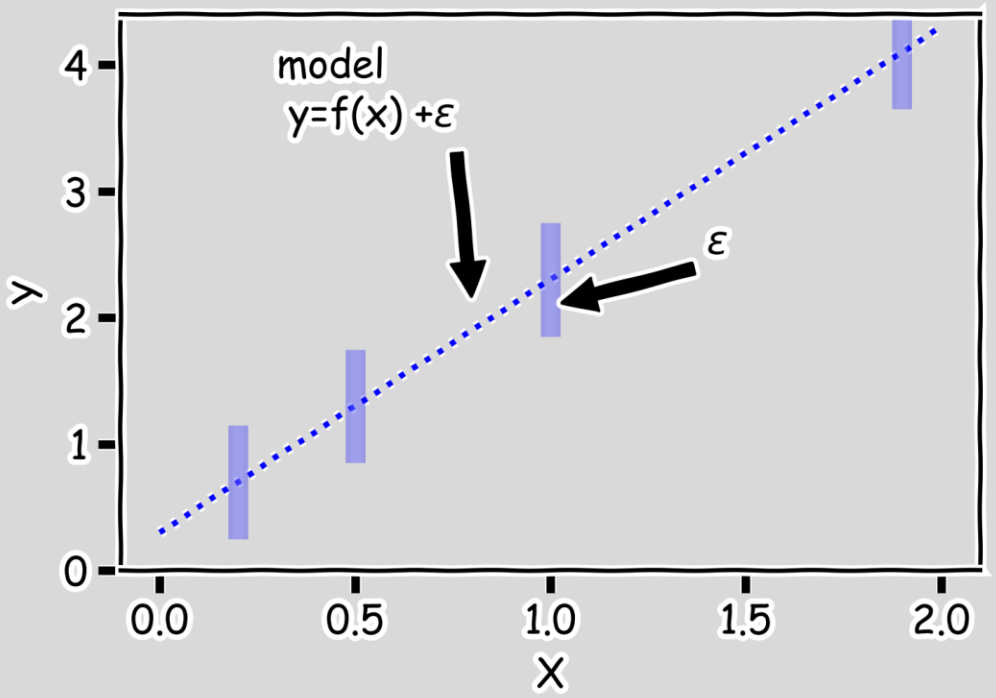
Confidence intervals for the predictor estimates

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



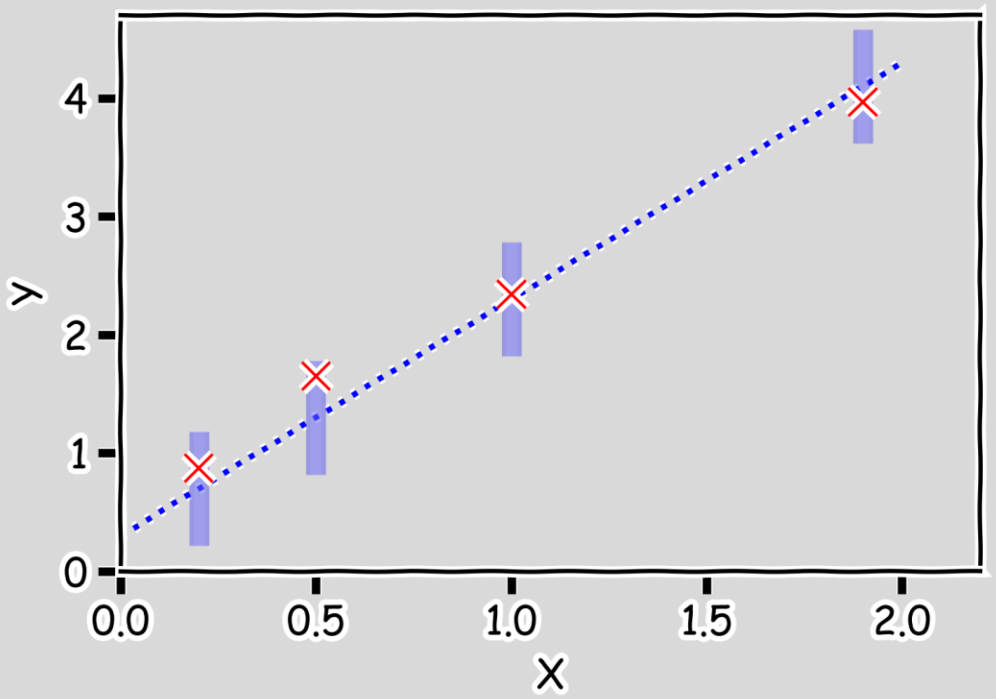
Confidence intervals for the predictor estimates

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



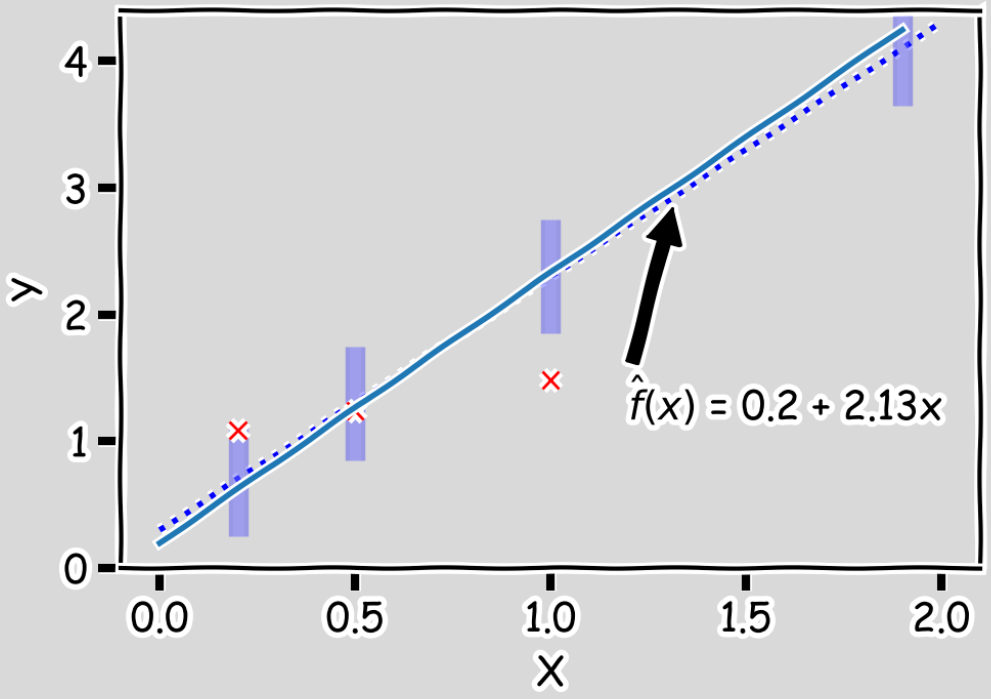
Confidence intervals for the predictor estimates

One set of observations, “one realization” we obtain one set of *Y*s (red crosses).



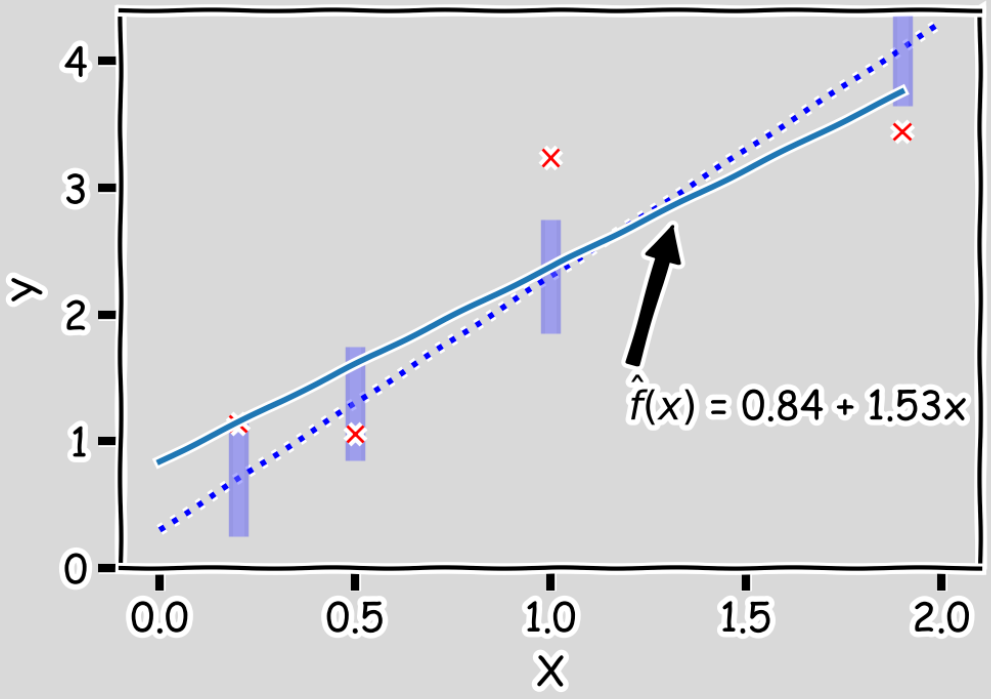
Confidence intervals for the predictor estimates

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



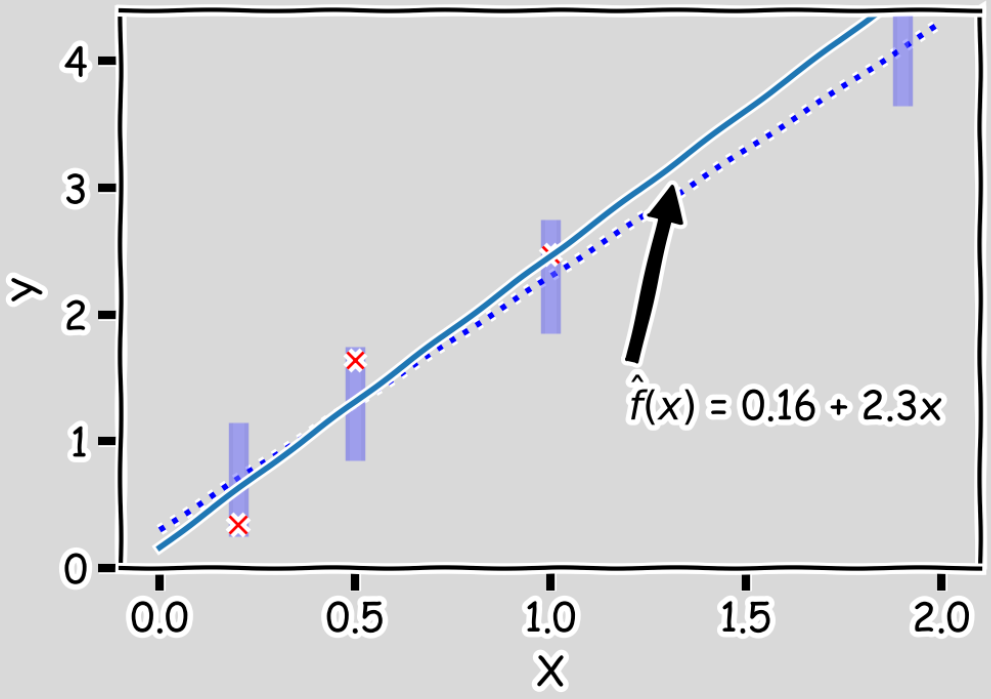
Confidence intervals for the predictor estimates

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



Confidence intervals for the predictor estimates

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



Confidence intervals for the predictor estimates

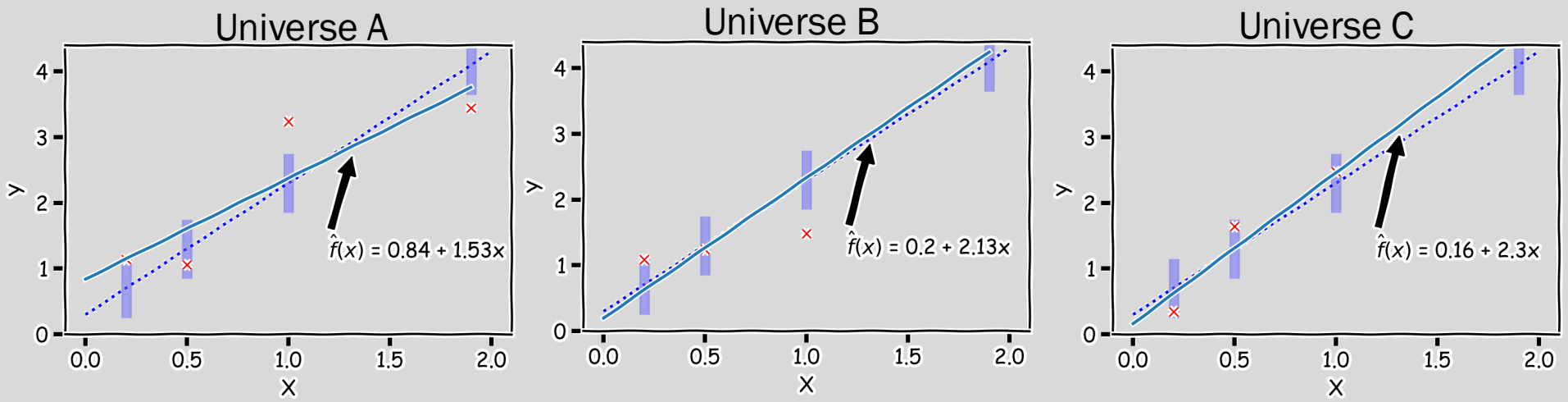
So if we just have one set of measurements of $\{X, Y\}$, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are just for this particular realization.

Confidence intervals for the predictor estimates

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.

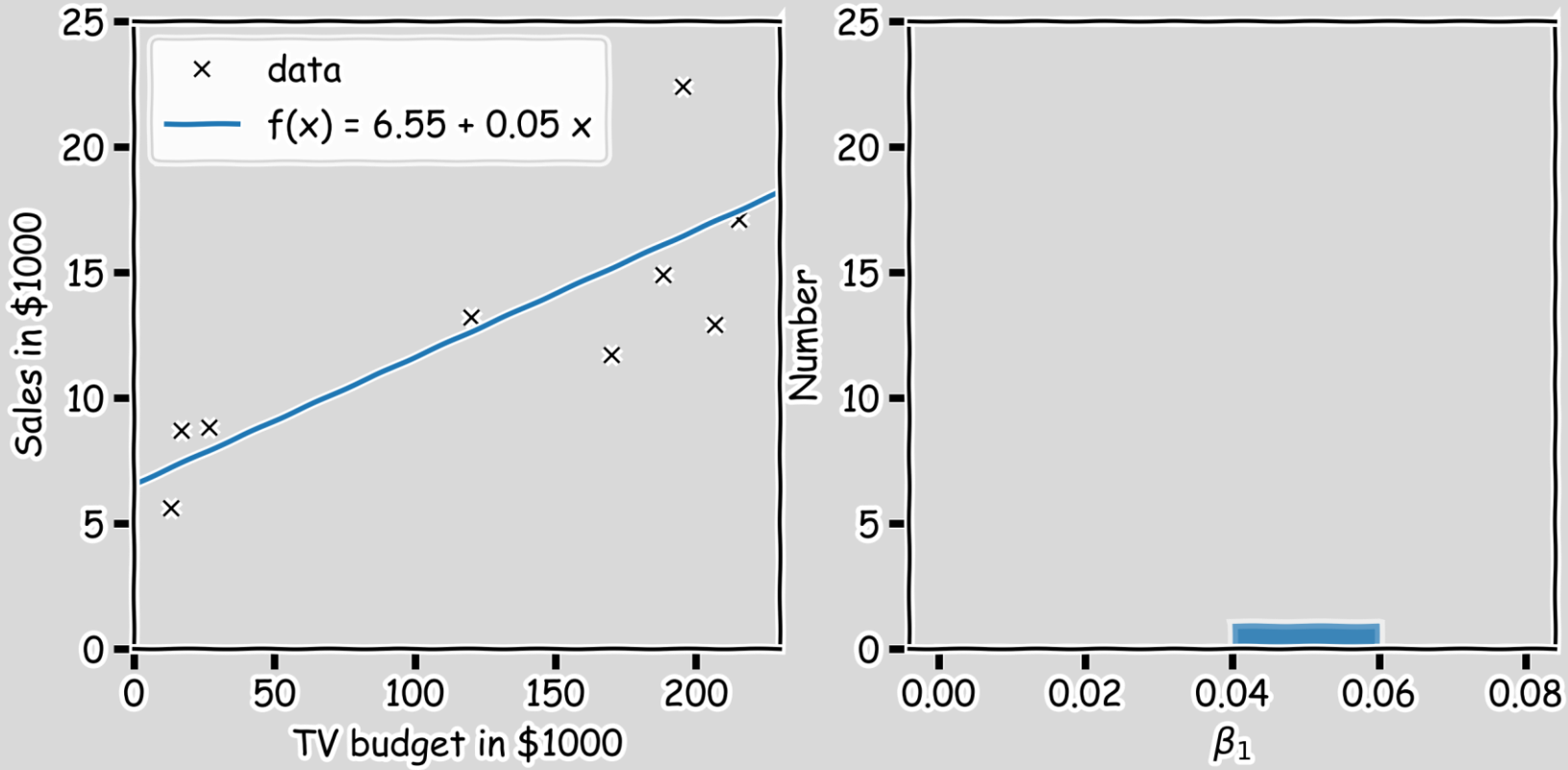
Question: If this is just one realization of the reality how do we know the truth?

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



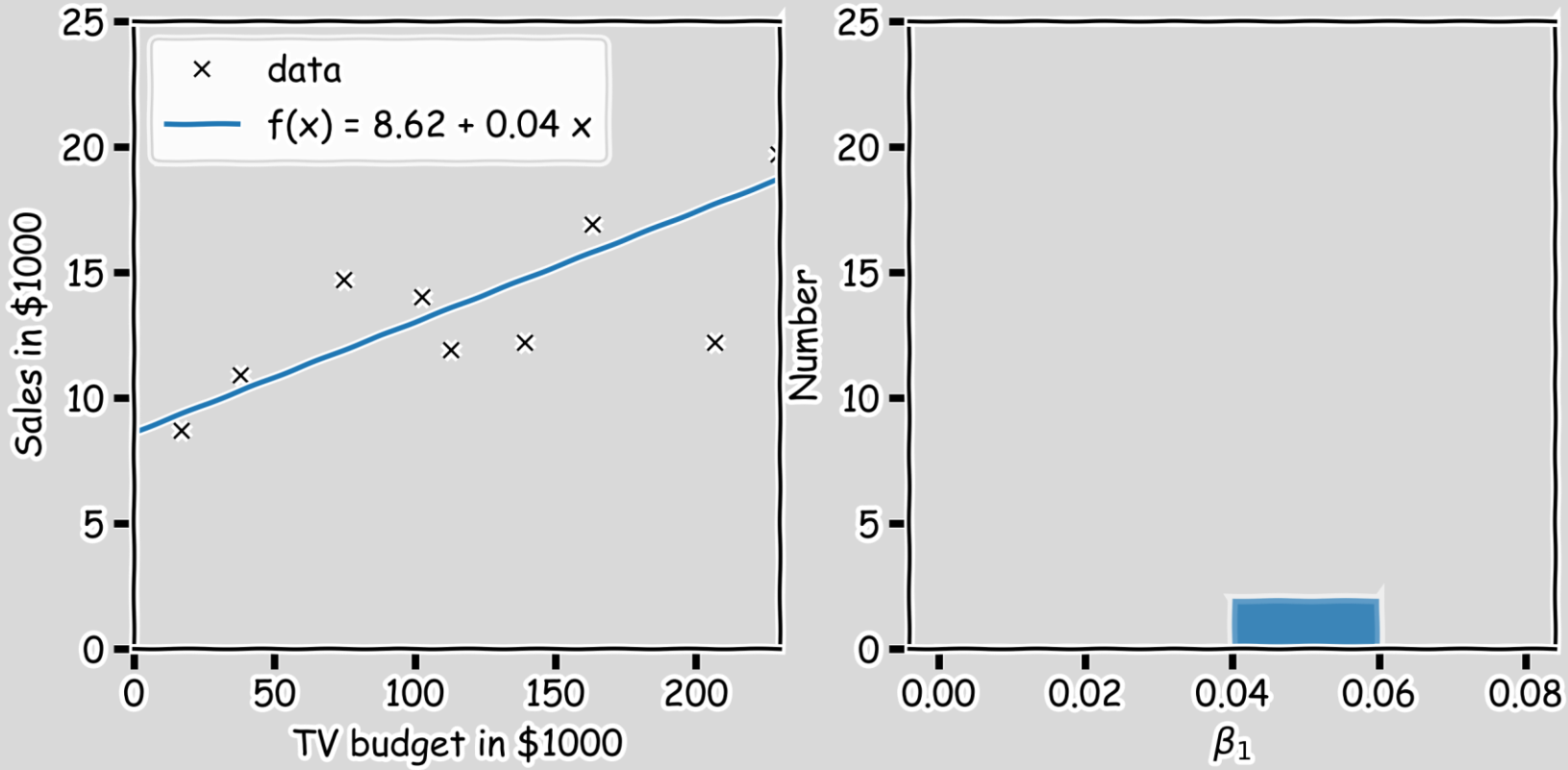
Confidence intervals for the predictor estimates

In our magical realisms, we can now sample multiple times



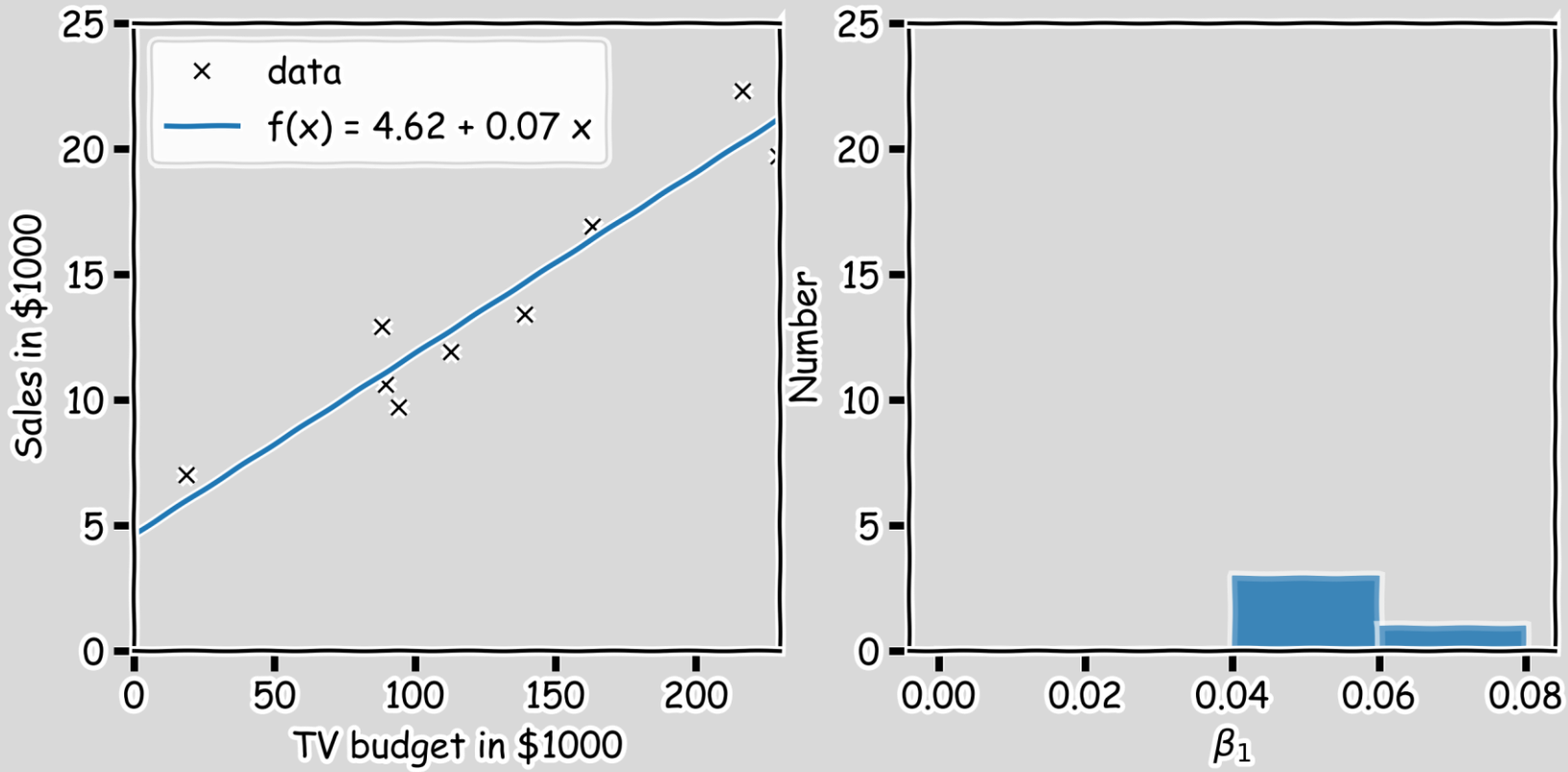
Confidence intervals for the predictor estimates

Another sample



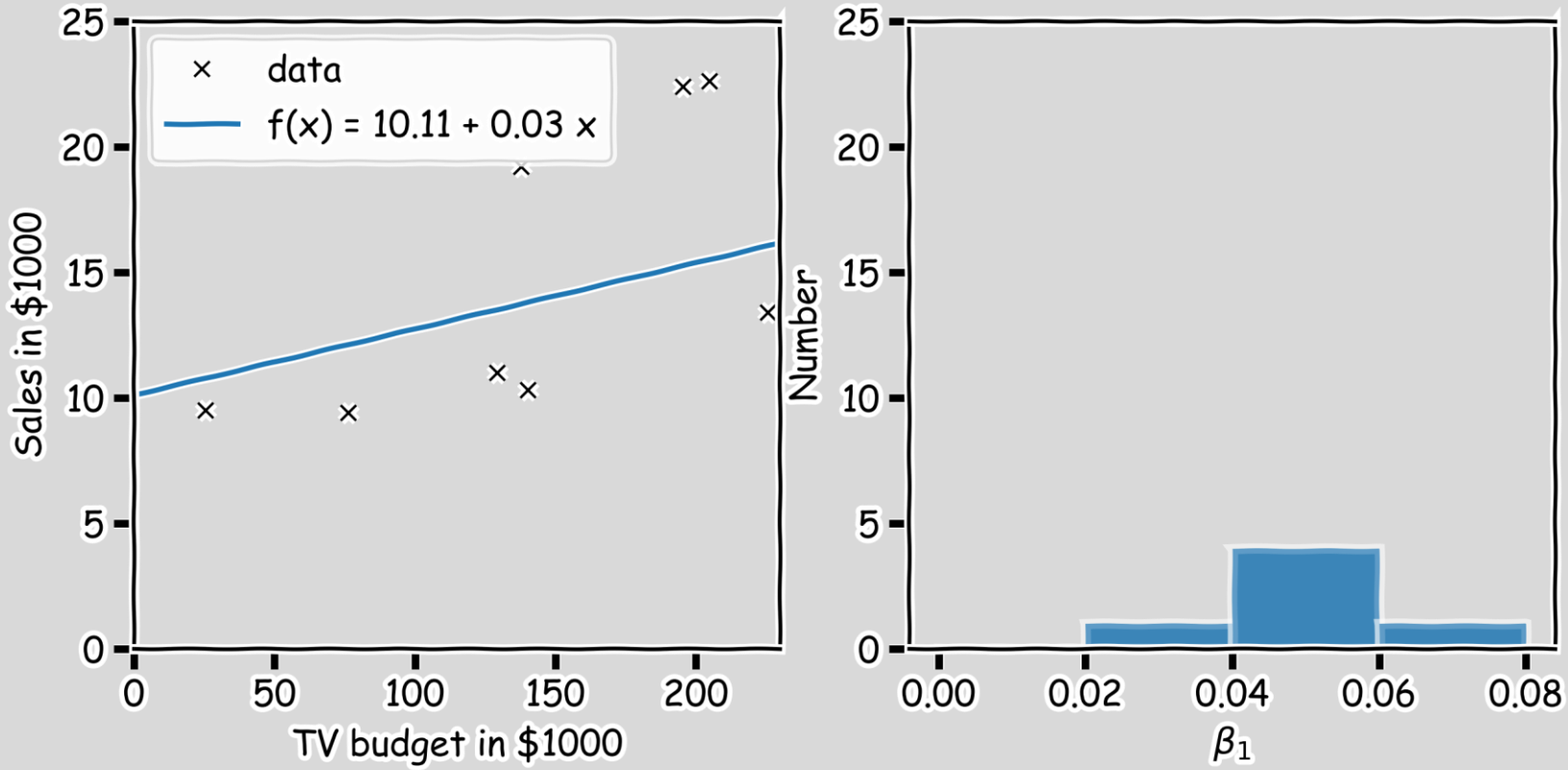
Confidence intervals for the predictor estimates

Another sample



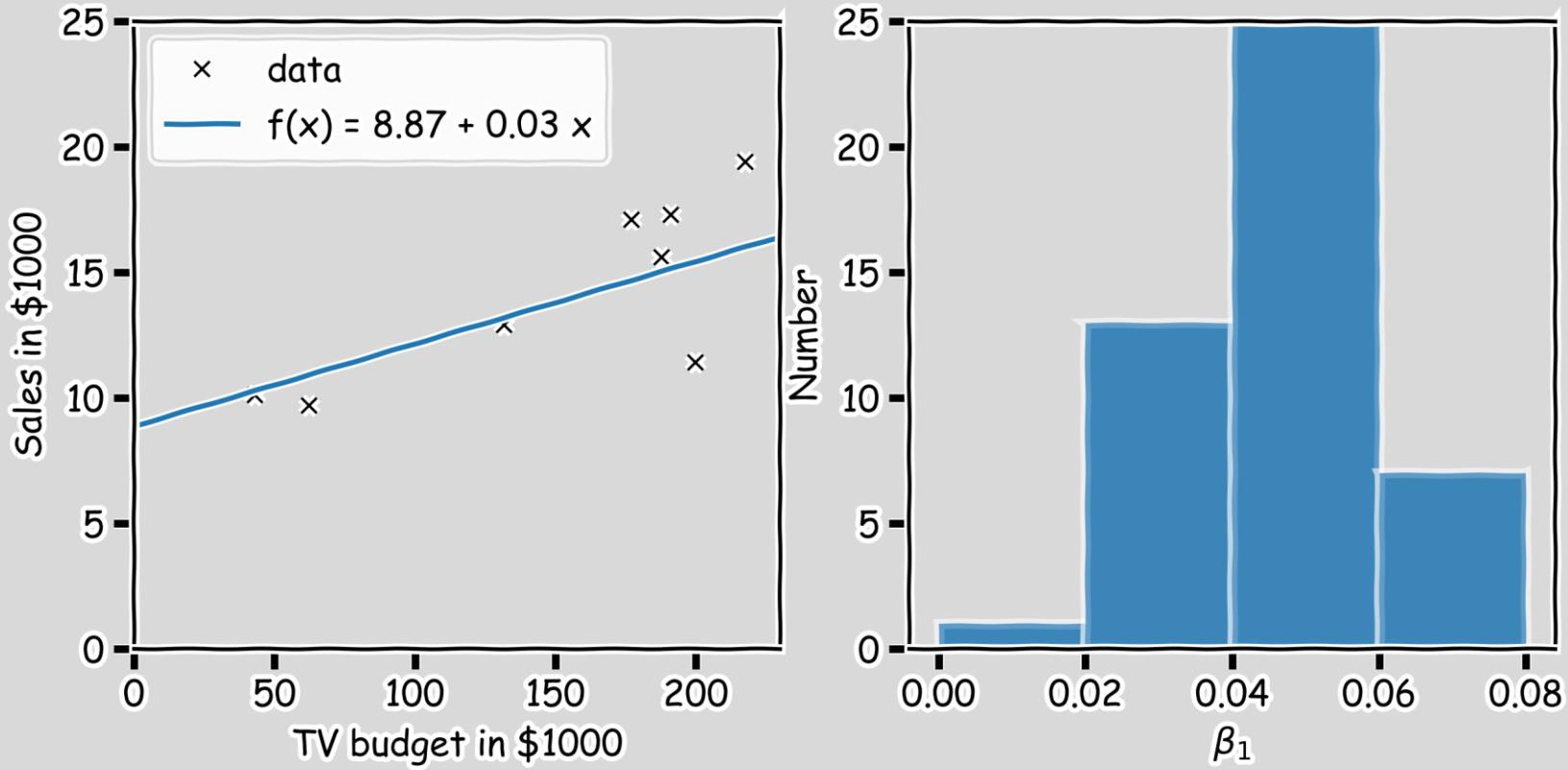
Confidence intervals for the predictor estimates

And another sample



Confidence intervals for the predictor estimates

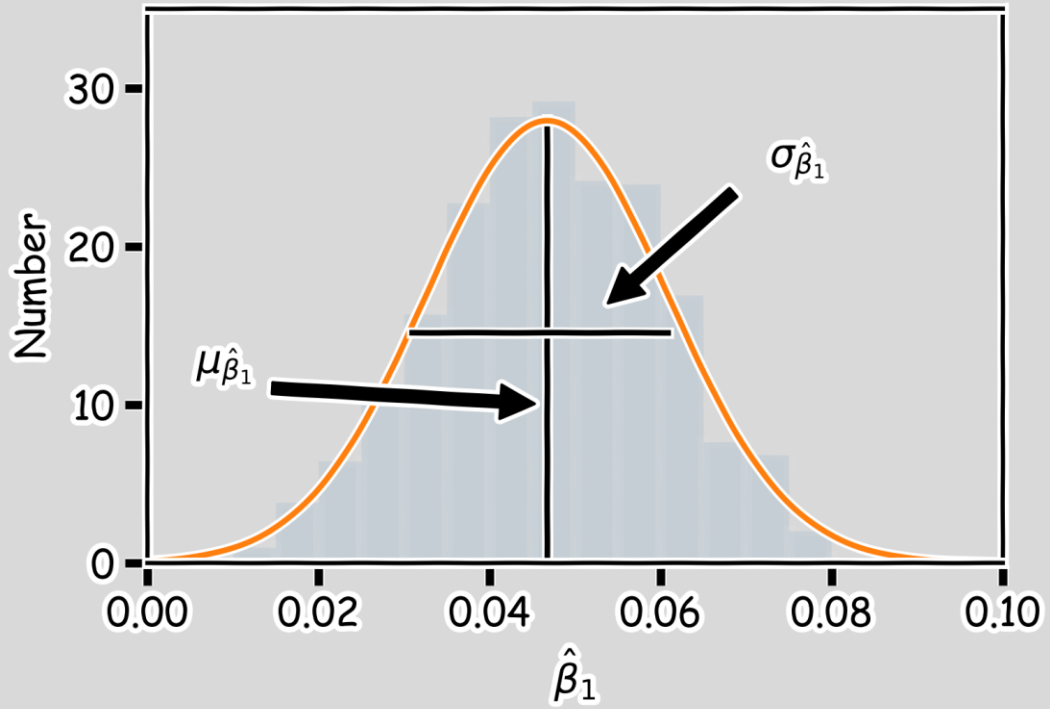
Repeat this 100 times



Confidence intervals for the predictor estimates

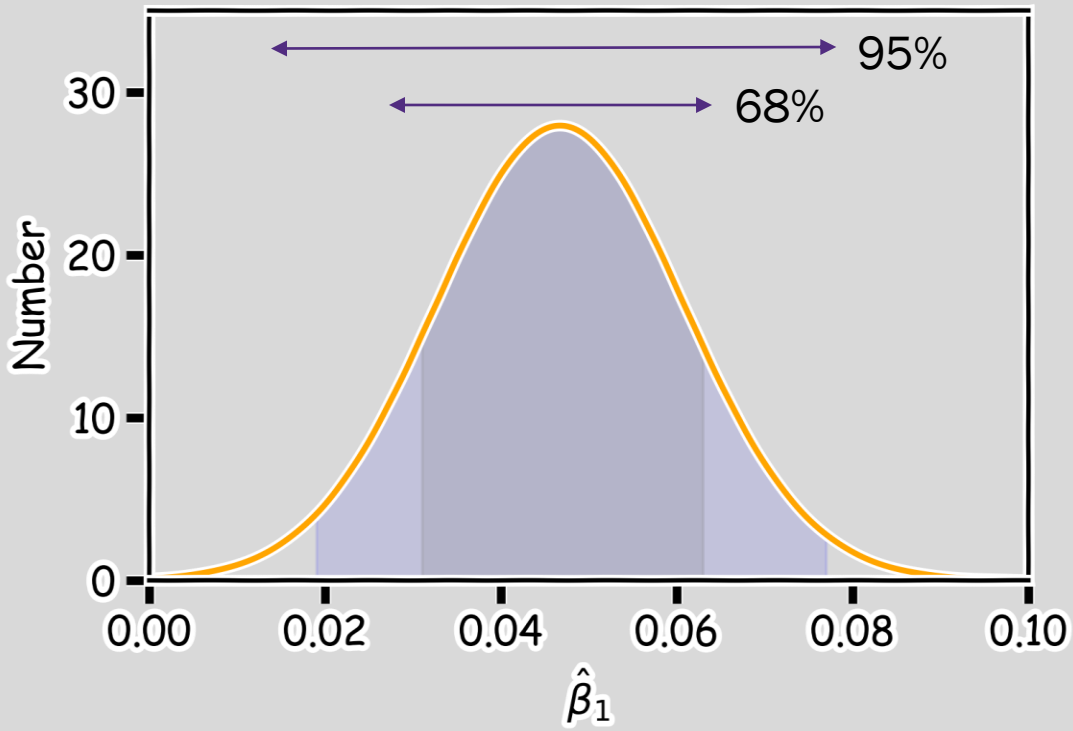
We can now estimate the mean and standard deviation of all the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.

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 predictor estimates

We can now 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.



The 95% confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$ are approximately given by:

$$\left[\hat{\beta}_0 - 2 \cdot SE(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot SE(\hat{\beta}_0)\right]$$

$$\left[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)\right]$$

How do we get these values? Stay tuned.

Lecture Outline

- Linear models
- Estimate regression coefficients for single predictor with bias term
- Confidence intervals for the predictor estimates
- **Bootstrap**
- Evaluating significance of predictors
- How well we know the model $\hat{f}$
- What happens with multiple predictors?

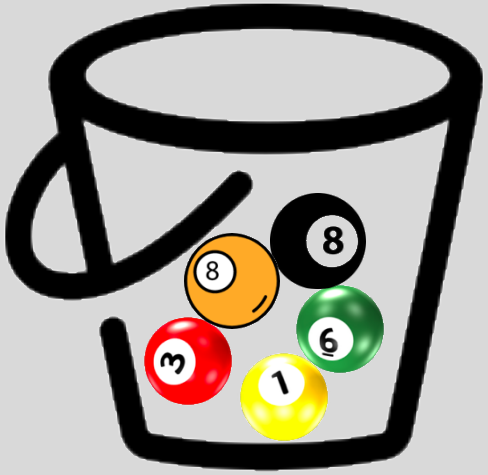
Bootstrap

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.

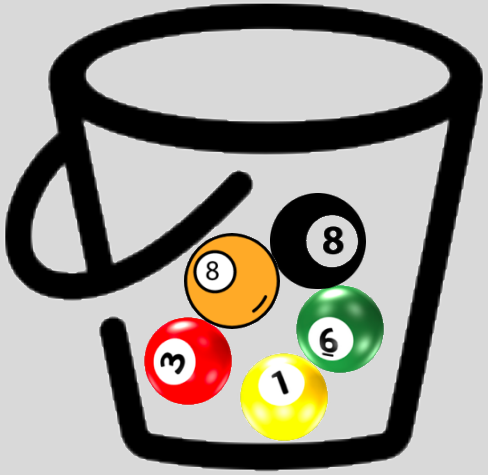
Bootstrap

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



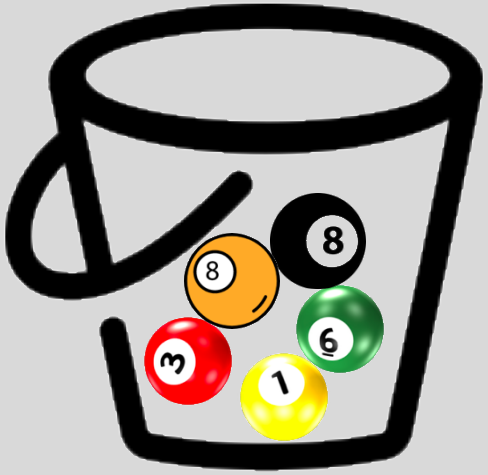
Bootstrap

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



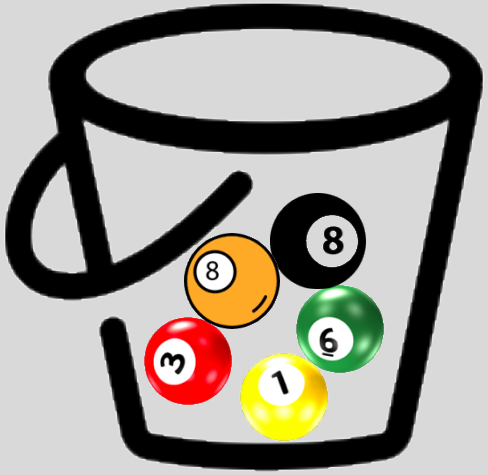
Bootstrap

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



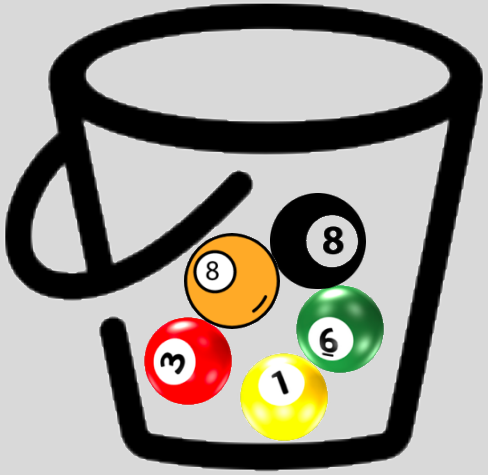
Bootstrap

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



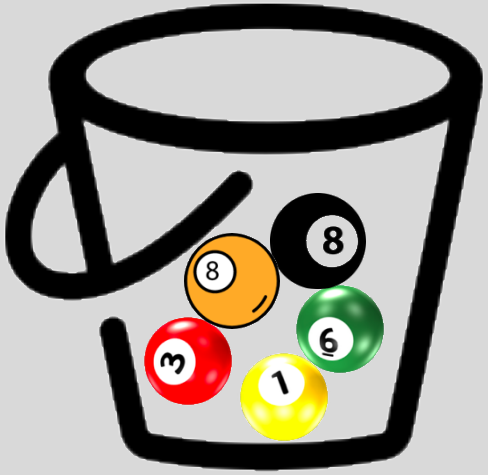
Bootstrap

We repeat this process...



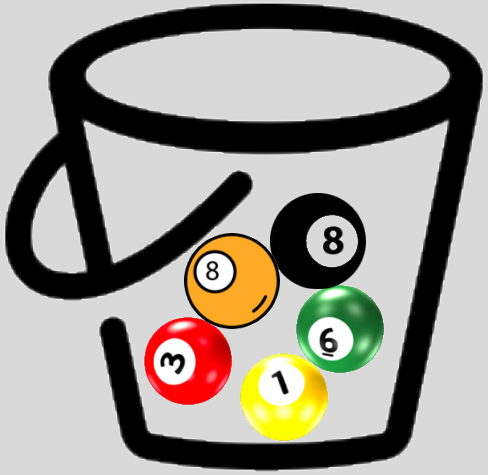
Bootstrap

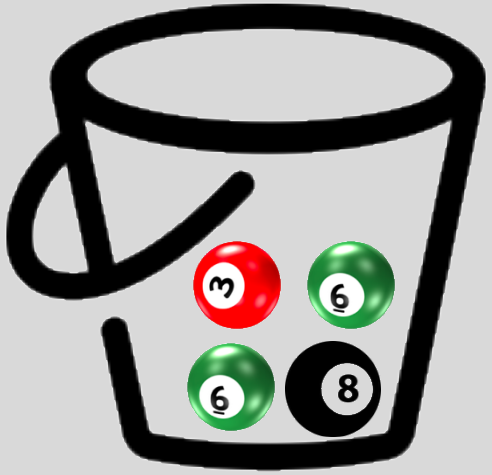
again



Bootstrap

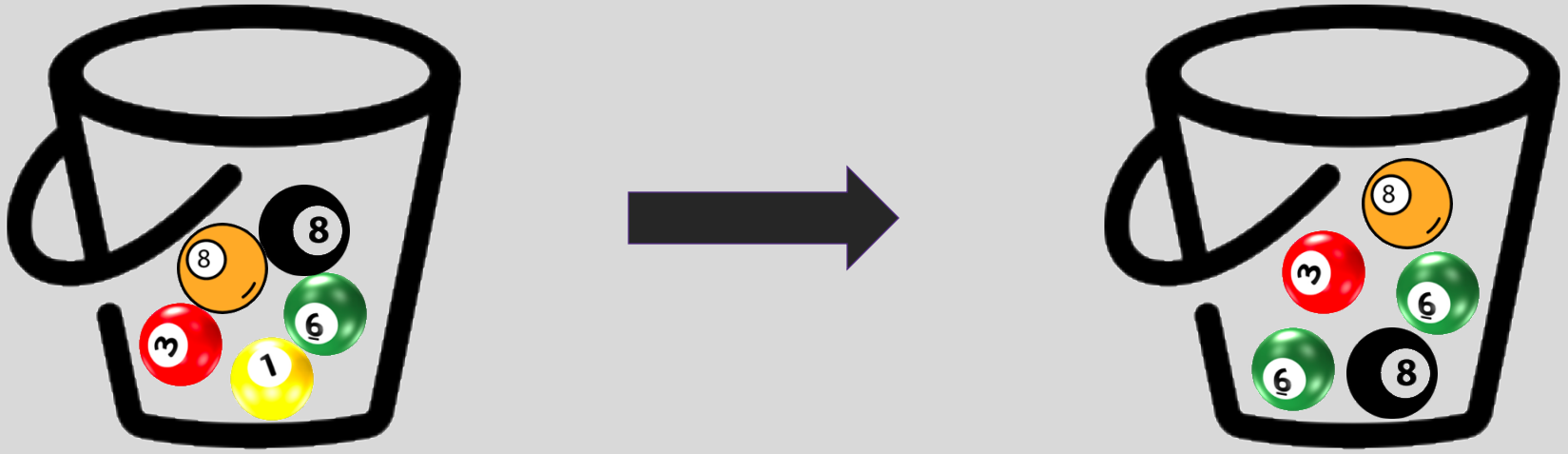
And again





Bootstrap

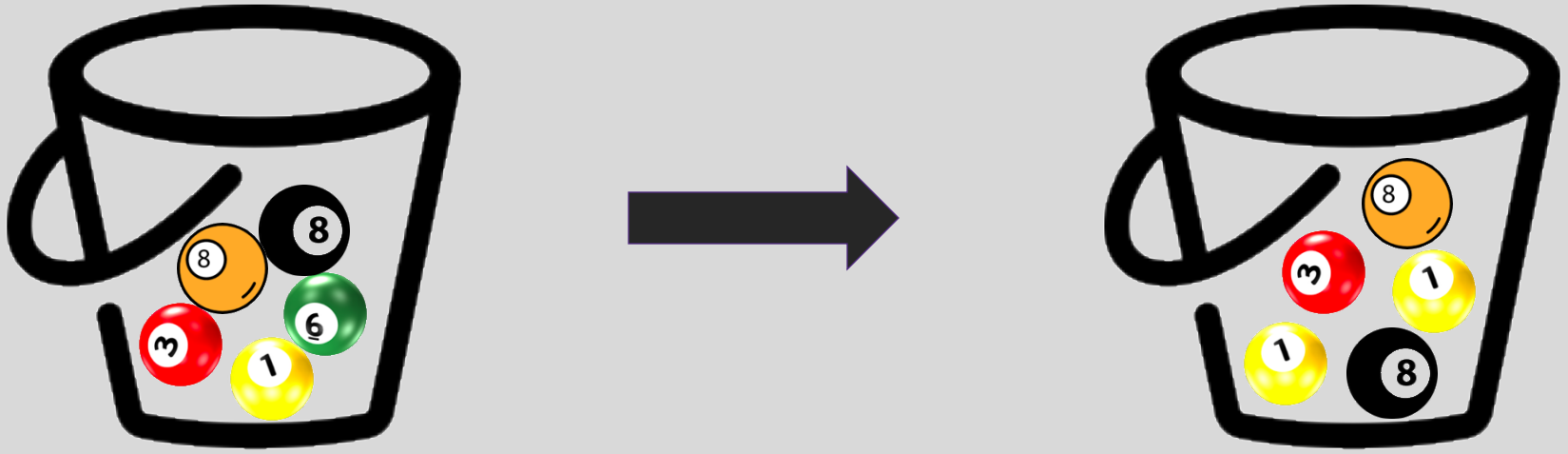
Until the “other” bucket has the same number of balls as the original one.



This new bucket represents a new parallel universe

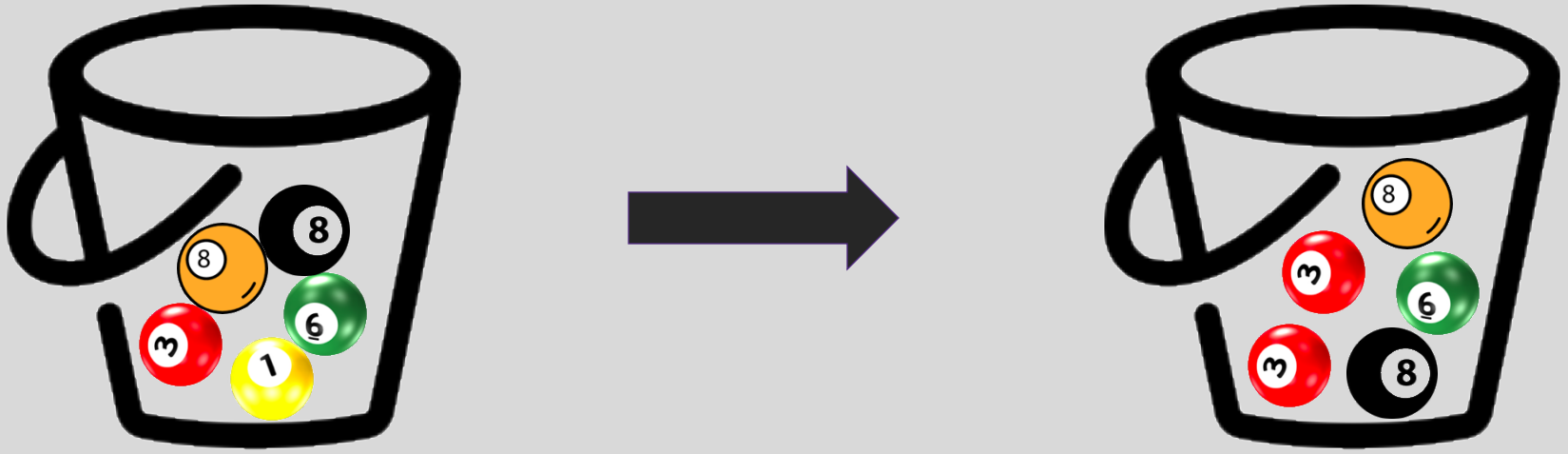
Bootstrap

We repeat this same process and acquire another sample.



Bootstrap

Until the “other” bucket has the same number of balls as the original one.



These new buckets represent 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 $\hat{\beta}_1$.

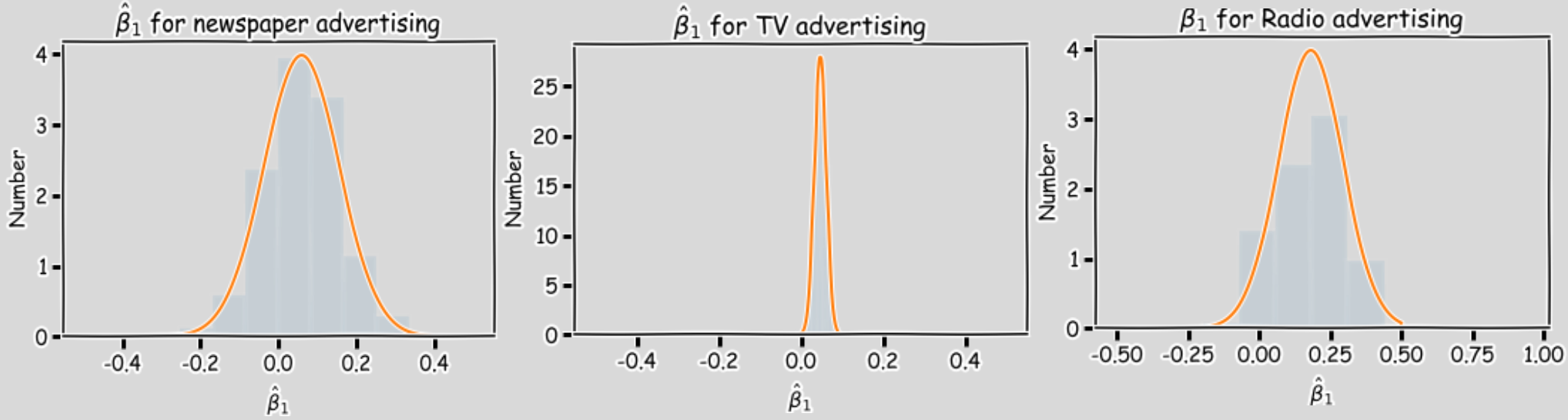
Lecture Outline

- Linear models
- Estimate regression coefficients for single predictor with bias term
- Confidence intervals for the predictor estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$
- What happens with multiple predictors?

Significance of the predictor estimates

We want to know if the predictor estimates are significant? Is there really a relationship between the predictor and the outcome?

How do we answer this question?



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. How likely is it that we would observe a value of the test-statistic as large (or small) as the one we got? This is a probability referred to as the p-value, ρ.
   a) We reject the null hypothesis if $\rho < \alpha$
   b) α=0.05 or α=0.01 are common – though there are caveats
   c) Importantly, α is selected BEFORE computing the p-value

Significance of the predictor estimates

Given a single estimated value of $\hat{\beta}_1$ and "magical" knowledge of its ***standard error***, $SE(\hat{\beta}_1)$, we state:

- H_0 : the true value of $\hat{\beta}_1$ is 0 indicating there is no relation between the predictor and the outcome
- H_1 : the true value of $\hat{\beta}_1$ is not 0
- Our test statistic is the *t-statistic*

We still don't know how to get this. Stay tuned.

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

This follows an approximately normal distribution. The libraries (statsmodels) will provide the value and the p-value for each predictor.

Standard Errors Estimation via Bootstrapping

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

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

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

Standard Errors (without bootstrapping)

Alternatively, if we don't or can't using bootstrapping:

If we know the variance σ_ϵ^2 of the noise ϵ (which we usually do not), we can compute $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ analytically using the formulae below (no need to bootstrap):

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

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

Standard Errors (without bootstrapping)

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: $\sigma_\epsilon^2 \downarrow \Rightarrow SE \downarrow$

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

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

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

Standard Errors (without bootstrapping)

However, if we assume the samples are independent and identically distributed (so-called **IID assumption**),

- 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: $\sigma \downarrow \Rightarrow SE \downarrow$

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

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

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

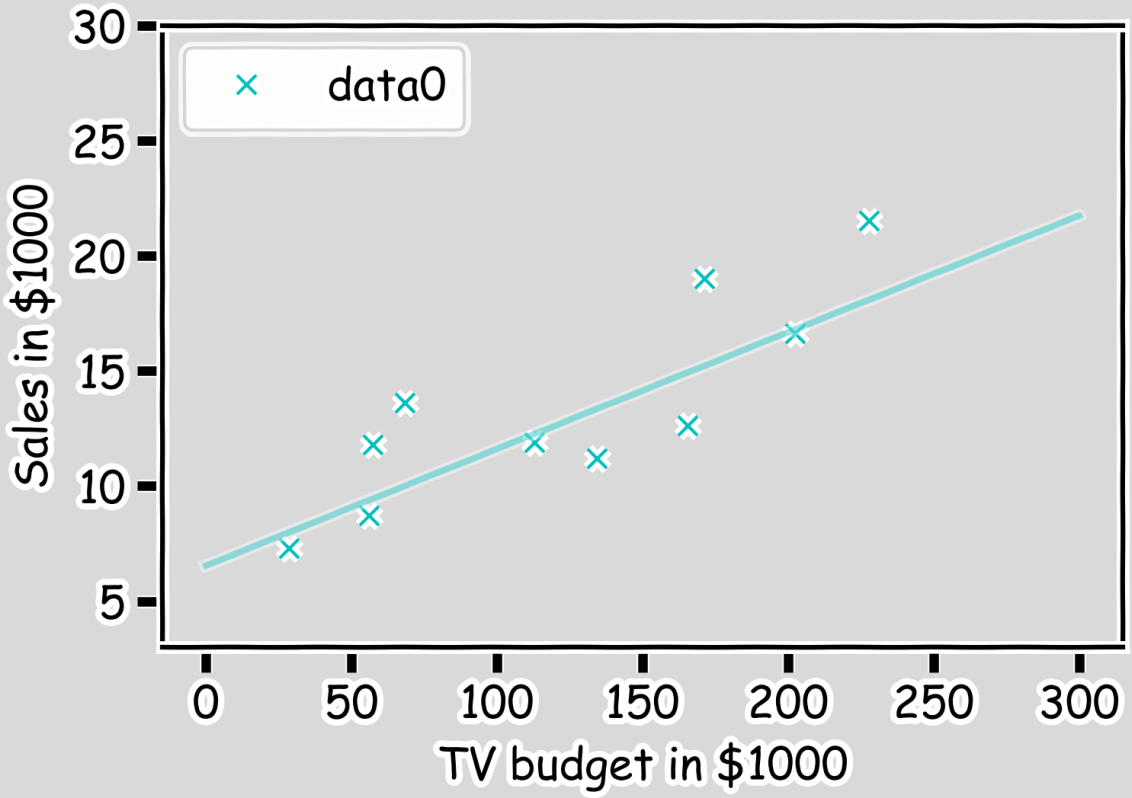
$$\sigma_\epsilon \approx \sqrt{\frac{n \cdot \mathrm{MSE}}{n-2}} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}}$$

Lecture Outline

- Linear models
- Estimate regression coefficients for single predictor with bias term
- Confidence intervals for the predictor estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$
- What happens with multiple predictors?

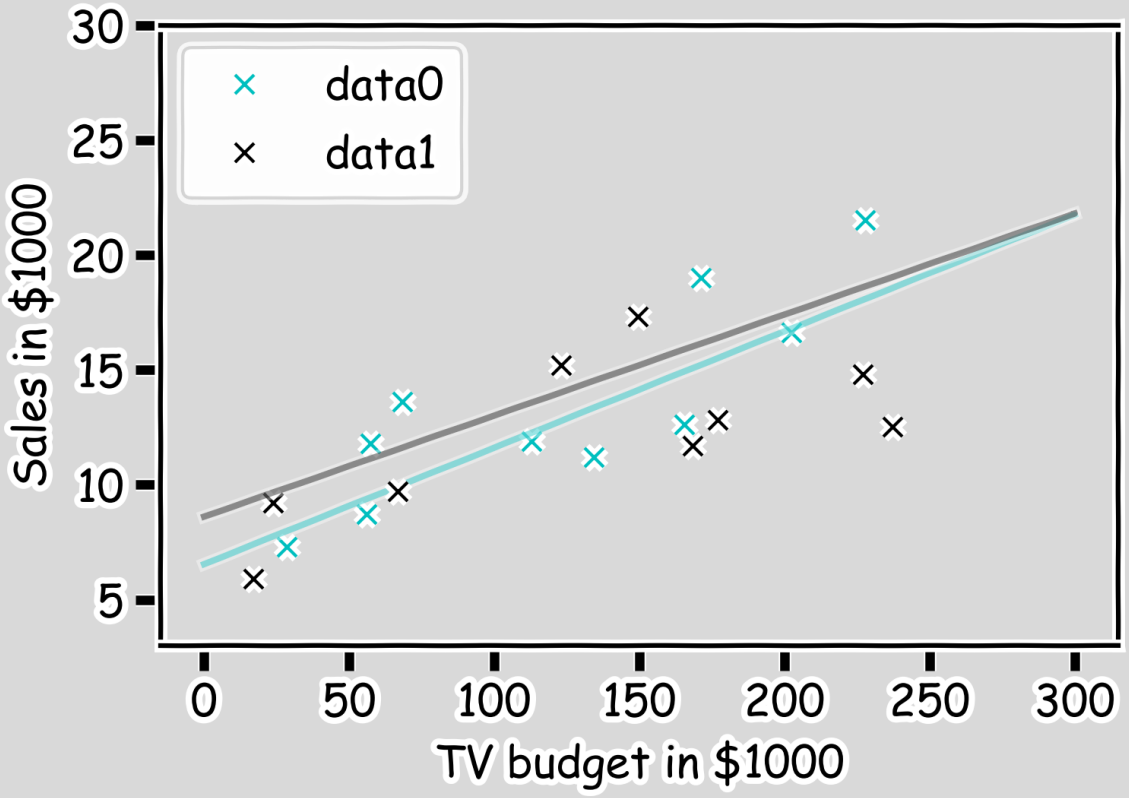
How well we know the model $\hat{f}$?

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.



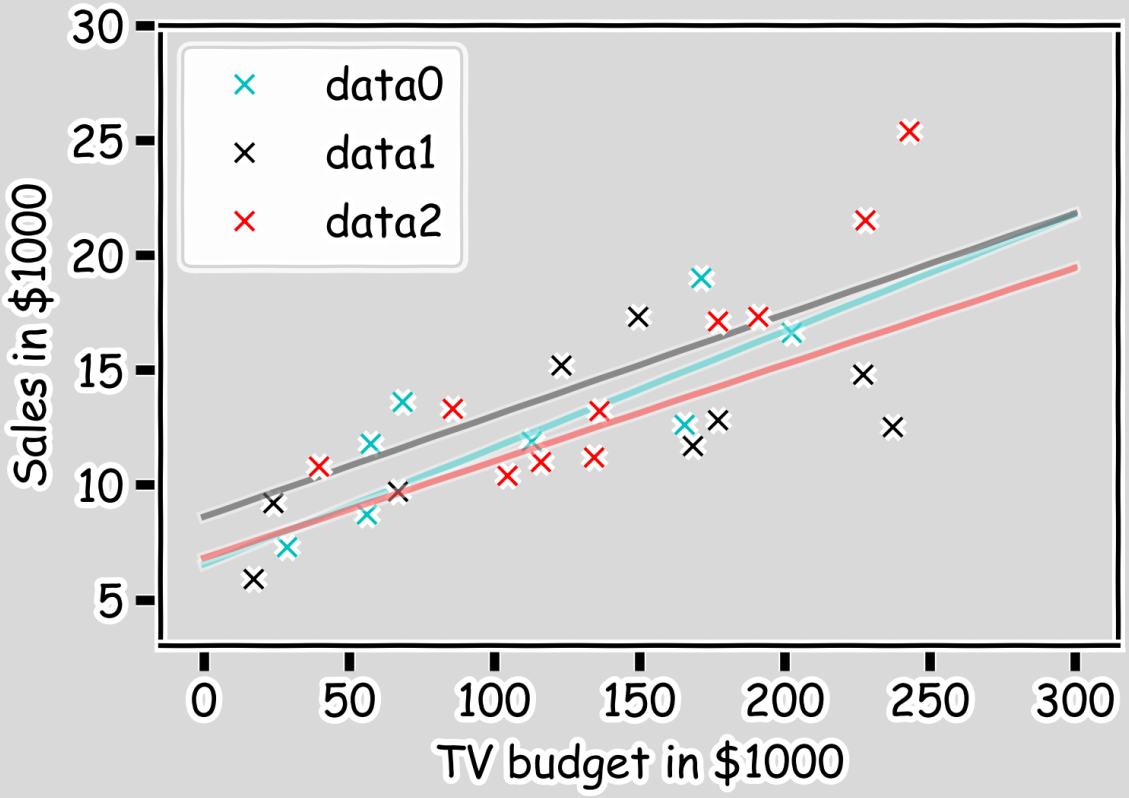
How well we know the model $\hat{f}$?

Here we show two models given the fitted coefficients.



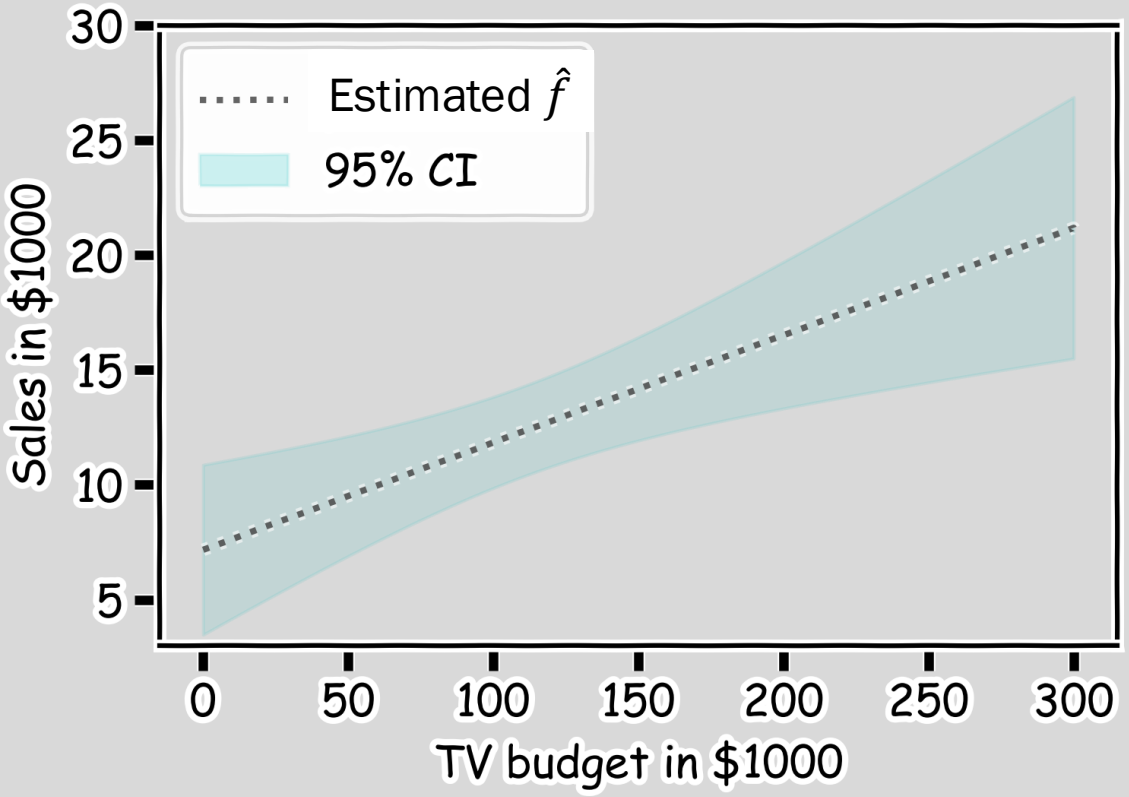
How well we know the model $\hat{f}$?

There is one such regression line for every bootstrapped sample.



How well we know the model $\hat{f}$?

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).



For a given model, $\hat{f}$, the 95% confidence interval is approximately given by:

$$\left[\hat{f}(x) - 2 \cdot SE\left(\hat{f}(x)\right), \hat{f}(x) + 2 \cdot SE\left(\hat{f}(x)\right)\right]$$

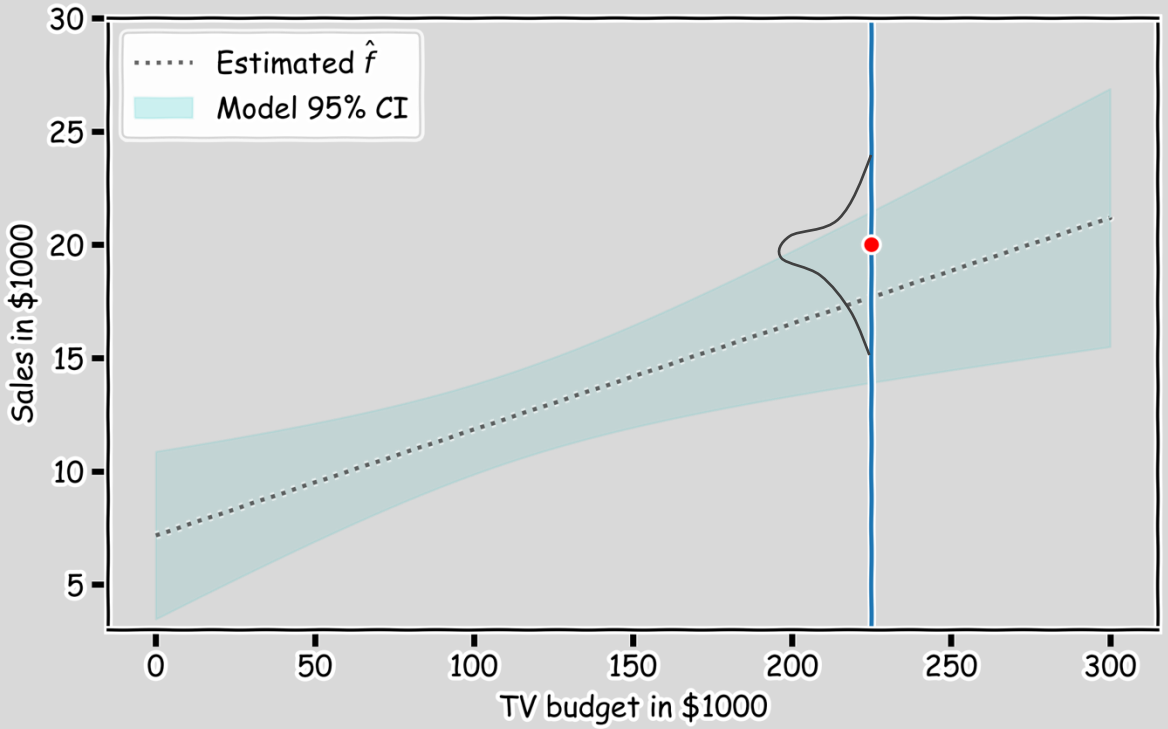
The *model* confidence interval indicates how well the model $\hat{f}$ represents the true relation f in:

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

This model CI only accounts for reducible error, it does not account for the noise, ε

Confidence in predicting $\hat{y}$ - the prediction interval

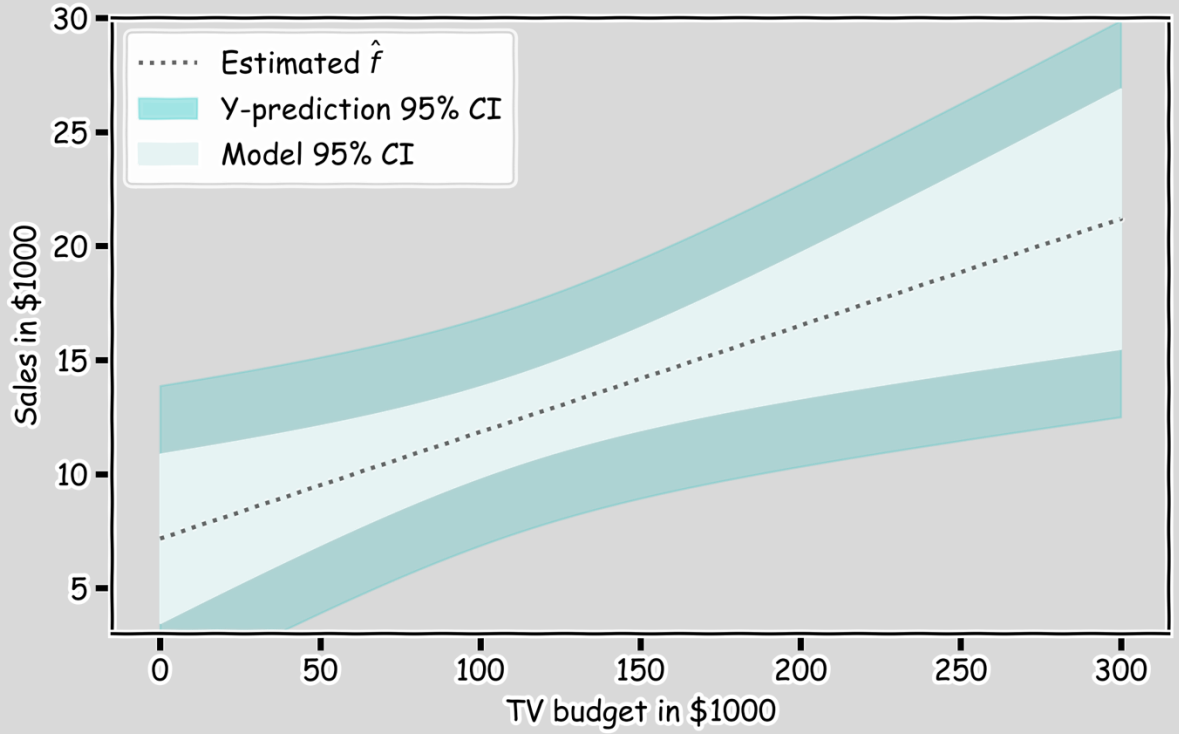
- for a given x, we have a distribution of models $\hat{f}(x)$
- for each of these $\hat{f}(x)$, a new observation of the outcome is $y \sim N(\hat{f}, \sigma_\epsilon)$ - why?



For a given x and $\hat{f}$, there is also a 95% confidence interval for the estimate of the response, y, called the prediction interval.

Prediction interval

- for a given x, we have a distribution of models $f(x)$
- for each of these $f(x)$, a new observation at x yields $y \sim N(f, \sigma_\epsilon)$
- The prediction confidence intervals are always wider than the mode confidence intervals because they also account for irreducible error, ε



Lecture Outline

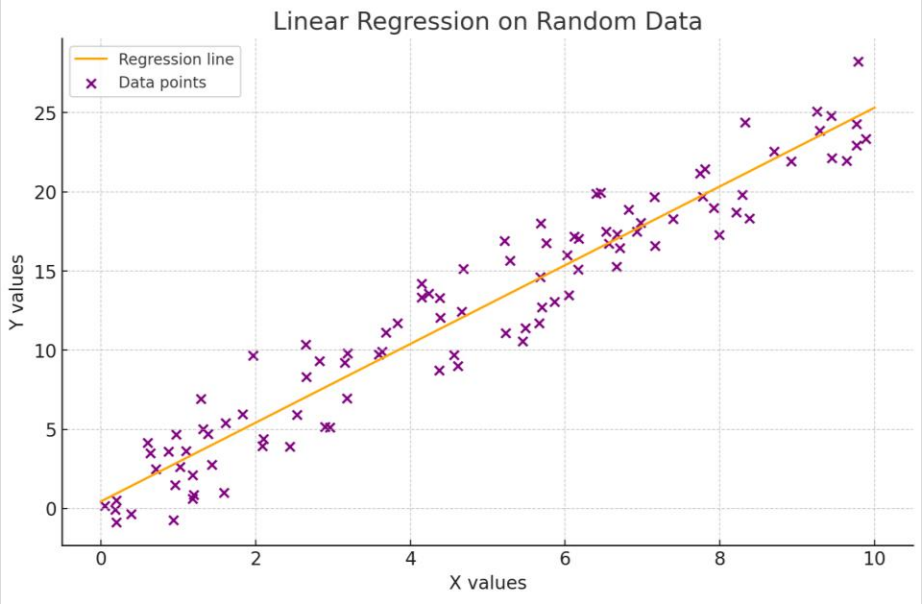
- Linear models
- Estimate regression coefficients for single predictor with bias term
- Confidence intervals for the predictor estimates
- Bootstrap
- Evaluating significance of predictors
- How well we know the model $\hat{f}$
- What happens with multiple predictors?

Multiple linear regression

- Assumes the relationship is given by

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon$$

- where β_0 represents the bias (constant offset), β_i are the coefficients (slopes), and ε is the error term
- Model characteristics
  a) Additive – the change in *y* due to a change in x_i is not affected by the values of $\boldsymbol{X}/x_i$
  b) Linear – the change in y due to a unit change in x_i is constant



Multiple linear regression (MLR) solution

Suppose we observe n samples of y, then the MLR problem in matrix form is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are $n \times 1$, $\mathbf{X}$ is $n \times p$, and $\boldsymbol{\beta}$ is $p \times 1$.

Ignoring the error term, the coefficients are given by the solution to

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}.$$

This system will rarely have an exact solution (why?). Instead, we find $\widehat{\boldsymbol{\beta}}$ that minimizes the MSE objective:

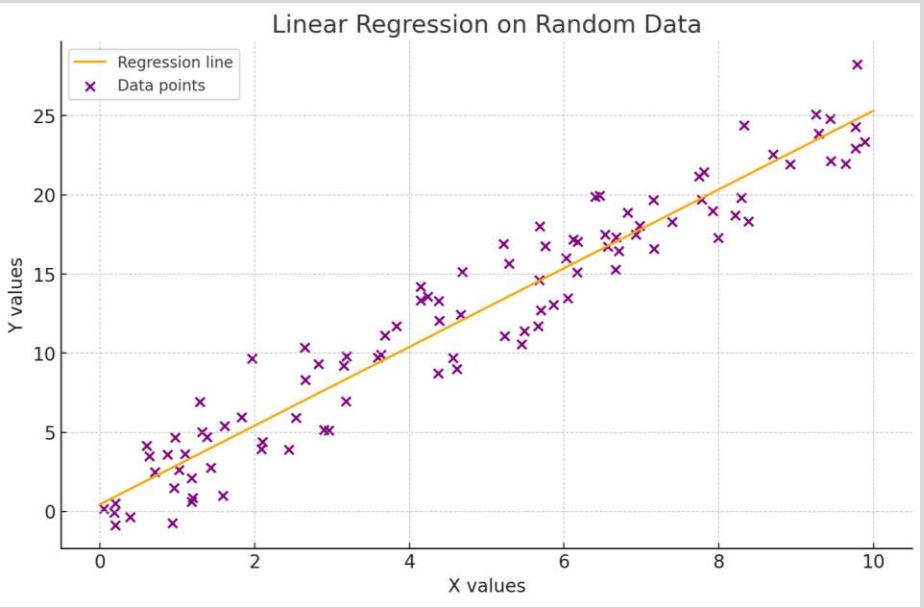
$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} [y_i - \boldsymbol{x}_i\boldsymbol{\beta}]^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

which has a unique solution (if columns of $\mathbf{X}$ are linearly independent):

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

MSE estimate of $\boldsymbol{\beta}$

Moore-Penrose pseudoinverse



Multiple linear regression – assessing the coefficients

F-statistic – used to assess likelihood of the null hypothesis, H_0, that there is no relation between the predictors and the outcome

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Typically, we examine p(F), the p-value of F, where a value near 0 suggests we should reject H_0 in favor of the alternative hypothesis that at least 1 coefficient is not zero

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

$$TSS = \sum_i (y_i - \bar{y})^2$$

$$RSS = \sum_i (y_i - \hat{y})^2$$

Coefficient standard error – average deviation of coefficient estimate from the actual coefficient

t-statistic - For a given coefficient, it is a test of the null hypothesis that the individual coefficient is zero

Multiple linear regression – assessing the model fit

R^2 – the proportion of explained variance

- Represents the fraction of the variance (departure from the mean) in *y* explained by the model
- Values near 1 indicate nearly all the variance is explained by the model
- Adding predictors ALWAYS increases R^2, hence usually consider an *adjusted* R^2 that adds a penalty relative to the number of predictors
- What is a good R^2? It depends on the context.

$$R^2 = \mathrm{Cor}(Y, \hat{Y})^2 = \left[\frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 (\hat{y}_i - \bar{\hat{y}}_i)^2}} \right]^2$$

- Other measures of model fit include *Akaike information criteria (AIC)* and *Bayesian information criteria (BIC)*

Potential problems

1. Data non-linearity – the relation between the outcome and predictors is not linear
2. Error term correlation – model errors may be correlated. Often seen in temporal data.
3. Heteroscedasticity – non-zero variance of error terms (i.e., error is dependent on the value of the outcome)
4. Outliers – values of y that are further from $\hat{y}$ than most
5. High leverage points – input points that are unlikely given the joint distribution

These are usually identified with visual inspection of residuals and fitted values.
We will explore these in more detail in the next lab.

Questions?