CS M148 –

Data Science Fundamentals

Lecture #4: Model selection & Inference

Baharan Mirzasoleiman

UCLA Computer Science

Announcements

Ethics Survey

- Grab you 3-digit IDs: Canvas->week3-> CSM148_W22_Random_IDs.csv

- Survey: shorturl.at/eflyW

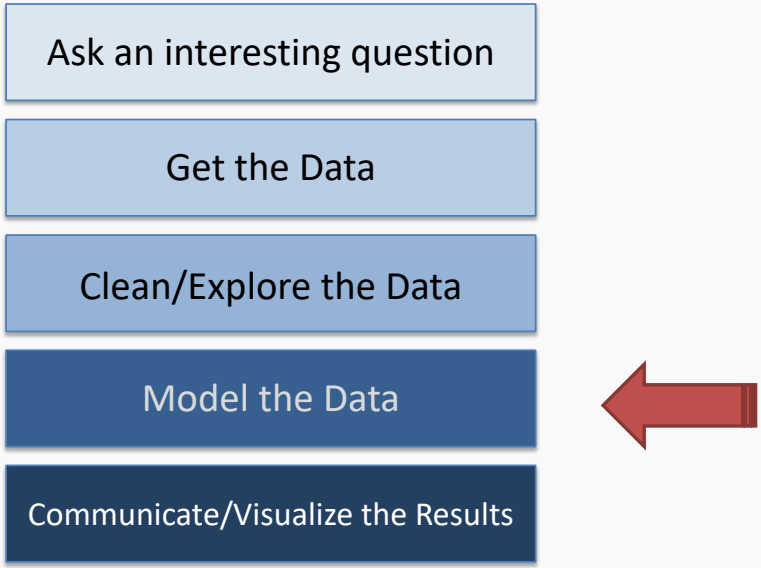
HW 1 is posted

- Canvas->week3->HW1
- Due Wed Jan 26, 2pm

Let’s quickly review what we saw last time

Ready to Model the Data!

The Data Science Process



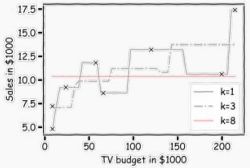
Linear Models

Why do we like them?

Interpretation!

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:

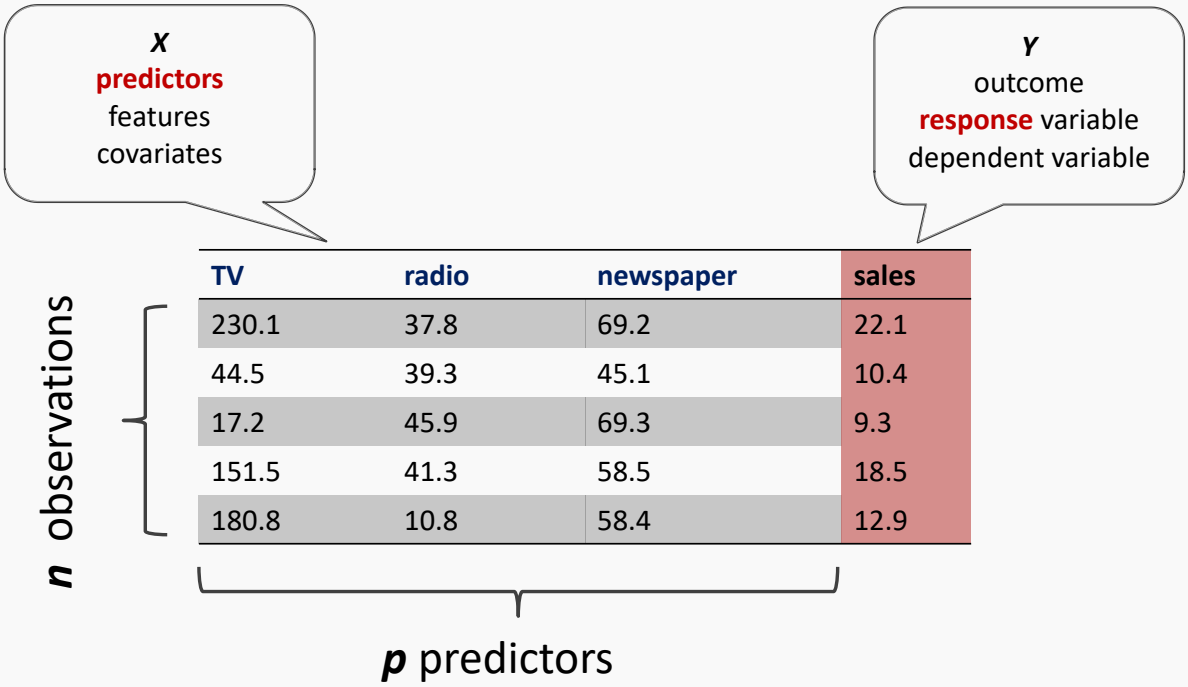
"how much more sales do we expect if we double the TV advertising budget?"

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

$$f(x) = \beta_0 + \beta_1 X$$

Multi & Poly Regression

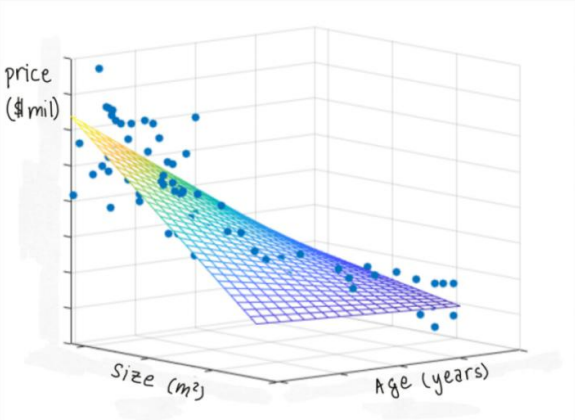
Response vs. Predictor Variables



Multilinear Models

In practice, it is unlikely that any response variable Y depends solely on one predictor *x*. Rather, we expect that is a function of multiple predictors $f(X_1, \dots, X_J)$. Using the notation we introduced in last lecture,

$$Y = y_1, \dots, y_n, \quad X = X_1, \dots, X_J \text{ and } X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$$



In this case, we can still assume a simple form for f -a multilinear form:

$$Y = f(X_1, \dots, X_J) + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_J X_J$$

Hence, $\hat{f}$, has the form

$$\hat{Y} = \hat{f}(X_1, \dots, X_J) + \epsilon = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_J X_J$$

Multiple Linear Regression

Given a set of observations,

$$\{(x_{1,1}, \ldots, x_{1,J}, y_1), \ldots (x_{n,1}, \ldots, x_{n,J}, y_n)\},$$

the data and the model can be expressed in vector/matrix notation,

observation predictor

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,J} \\ 1 & x_{2,1} & \ldots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,J} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Multilinear Model, example

For our data

$$\text{Sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper + \epsilon$$

In linear algebra notation

predictor x data

$$\boldsymbol{Y} = \begin{pmatrix} Sales_1 \\ \vdots \\ Sales_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & TV_1 & Radio_1 & News_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & TV_n. & Radio_n & News_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix}$$

$$Sales_1 = 1 \quad TV_1 \quad Radio_1 \quad News_1 \times \begin{matrix} \beta_0 \\ \vdots \\ \beta_3 \end{matrix}$$

Qualitative Predictors

So far, we have assumed that all variables are quantitative. But in practice, often some predictors are **qualitative**.

Example: The Credit data set contains information about balance, age, cards, education, income, limit , and rating for a number of potential customers.

Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
14.890	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.02	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.59	7075	514	4	71	11	Male	No	No	Asian	580
148.92	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

Qualitative Predictors

If the predictor takes only two values, then we create an **indicator** or **dummy variable** that takes on two possible numerical values.

For example for the gender, we create a new variable:

$$x_i = \begin{cases} 1 & \text{if } i\,\text{th person is female} \\ 0 & \text{if } i\,\text{th person is male} \end{cases}$$

We then use this variable as a predictor in the regression equation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\,\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\,\text{th person is male} \end{cases}$$

Qualitative Predictors

Question: What is interpretation of β_0 and β_1?

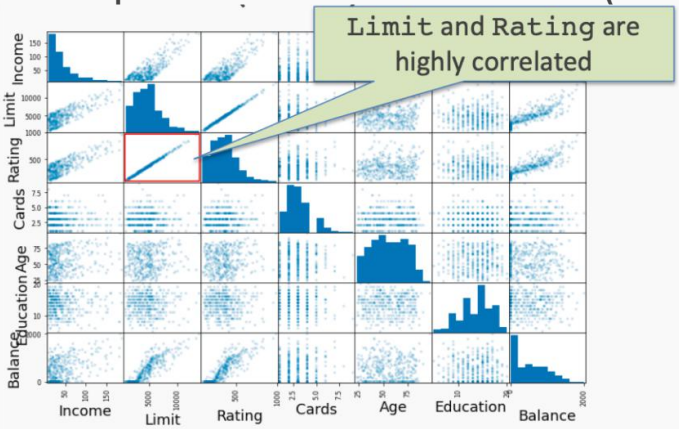
- β_0 is the average credit card balance among males,
- $\beta_0 + \beta_1$ is the average credit card balance among females,
- and β_1 the average difference in credit card balance between females and males.

Example: Calculate β_0 and β_1 for the Credit data.

You should find $\beta_0 \sim \$509, \beta_1 \sim \$19$

Collinearity

Collinearity and multicollinearity refers to the case in which two or more predictors are correlated (related).



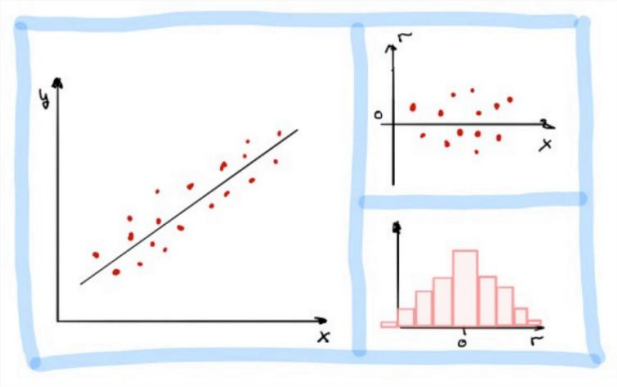
	Columns	Coefficients
0	Income	-7.802001
1	Limit	0.193077
2	Rating	1.102269
3	Cards	17.923274
4	Age	-0.634677
5	Education	-1.115028
6	Gender	10.406651
7	Student	426.469192
8	Married	-7.019100

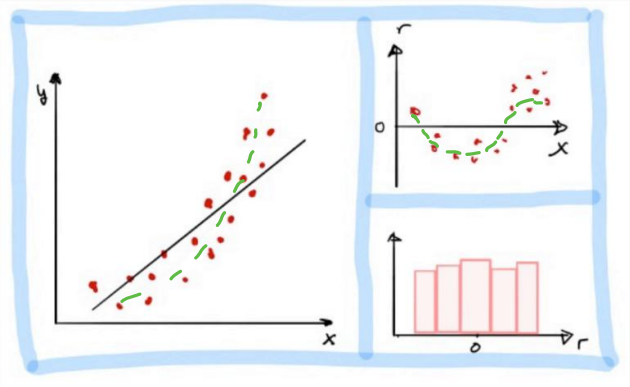
	Columns	Coefficients
0	Income	-7.770915
1	Rating	3.976119
2	Cards	4.031215
3	Age	-0.669308
4	Education	-0.375954
5	Gender	10.368840
6	Student	417.417484
7	Married	-13.265344

The regression coefficients are not uniquely determined. In turn it hurts the interpretability of the model as then the regression coefficients are not unique and have influences from other features.

Both limit and rating have positive coefficients, but it is hard to understand if the balance is higher because of the rating or is it because of the limit? If we remove limit then we achieve almost the same model performance but the coefficients change.

Residual Analysis





Linear assumption is correct. There is no obvious relationship between residuals and x. Histogram of residuals is symmetric and normally distributed.

Linear assumption is incorrect. There is an obvious relationship between residuals and x. Histogram of residuals is symmetric but not normally distributed.

Note: For multi-regression, we plot the residuals vs predicted $\hat{y}$, since there are too many x's and that could wash out the relationship.

Beyond linearity

We also assumed that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

Synergy effect or **interaction effect** states that when an increase on the radio budget affects the effectiveness of the TV spending on sales.

We change

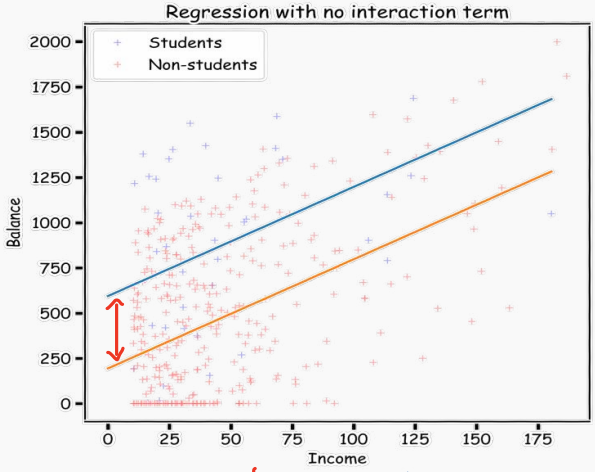
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

To

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

model is linear, but we have X_1, X_2, a nonlinear term

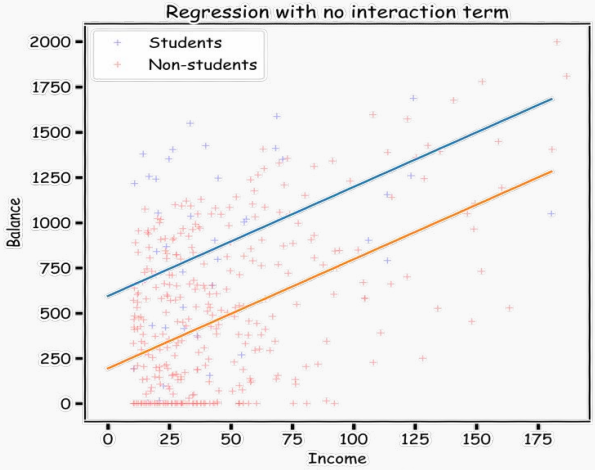
What does it mean?

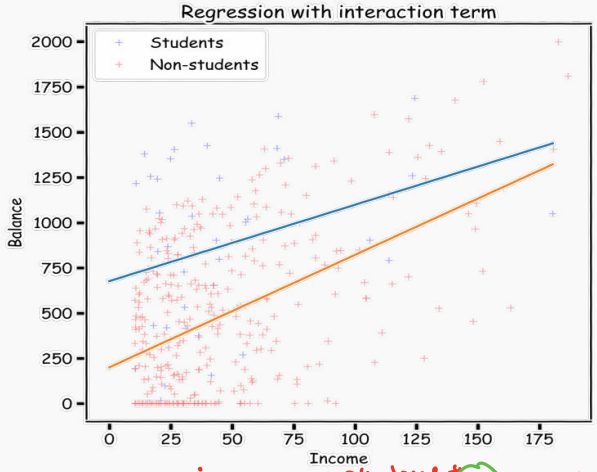


$\beta_0 + \beta_1 \times income + \beta_2 \times student$

$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1) \times Income. \end{cases}$$

What does it mean?





$$\beta_0 + \beta_1 \times Income + \beta_2 \times Student + \beta_3 \times Income \times Student$$

β_3: math

Income × Student: interaction term

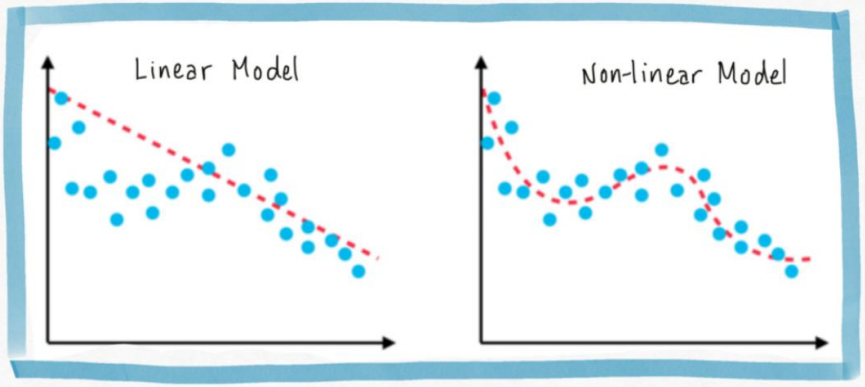
$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1) \times Income. \end{cases}$$

$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times Income \end{cases}$$

Polynomial Regression

Fitting non-linear data

Multi-linear models can fit large datasets with many predictors. But the relationship between predictor and target isn't always linear.



We want a model:

$$y = f_\beta(x)$$

Where f is a non-linear function and β is a vector of the parameters of f.

Polynomial Regression

The simplest non-linear model we can consider, for a response *Y* and a predictor *X*, is a polynomial model of degree *M*,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_M x^M$$

Just as in the case of linear regression with cross terms, polynomial regression is a special case of linear regression - we treat each x^m as a separate predictor. Thus, we can write

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \ldots & x_1^M \\ 1 & x_2^1 & \ldots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

Model Training

Give a dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, we find the optimal polynomial model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

1. We transform the data by adding new predictors:

$$\tilde{x} = [1, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_M]$$

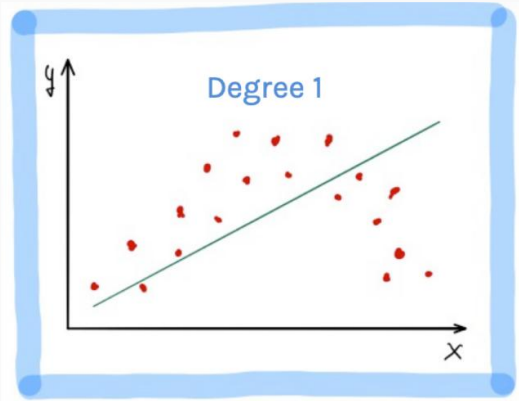
where $\tilde{x}_k = x^k$

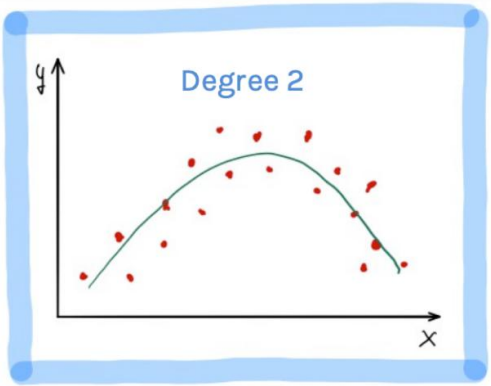
2. Fit the parameters by minimizing the MSE using vector calculus. As in multi-linear regression:

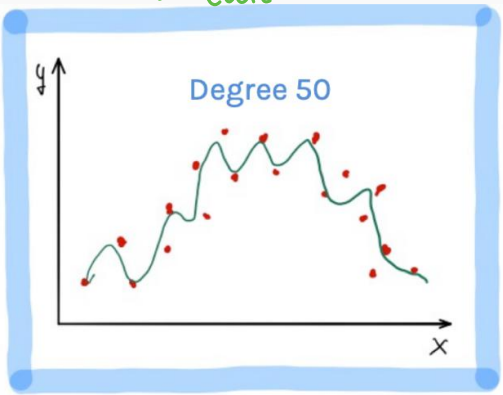
$$\widehat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{y}$$

Polynomial Regression (cont)

Fitting a polynomial model requires choosing a degree. - usually look @ MSE to decide







Underfitting: when the degree is too low, the model cannot fit the trend.

We want a model that fits the trend and ignores the noise.

Overfitting: when the degree is too high, the model fits all the noisy data points.

Feature Scaling

Do we need to scale out features for polynomial regression?

Linear regression, $y = X\beta$, is invariant under scaling. If X is called by some number λ then β will be scaled by $\frac{1}{\lambda}$ and MSE will be identical.

However if the range of X is low or large then we run into troubles. Consider a polynomial degree of 20 and the maximum or minimum value of any predictor is large or small. Those numbers to the 20th power will be problematic.

It is always a good idea to scale X when considering polynomial regression:

$$X^{norm} = \frac{X - \bar{X}}{\sigma_X}$$

Note: sklearn's StandardScaler() can do this.

High degree of polynomial
leads to **OVERFITTING**!

Model Selection

Model Selection

Model selection is the application of a principled method to determine the complexity of the model, e.g. choosing a subset of predictors, choosing the degree of the polynomial model etc.

A strong motivation for performing model selection is to avoid overfitting, which we saw can happen when:

- there are too many predictors:
  - the feature space has high dimensionality
  - the polynomial degree is too high
  - too many cross terms are considered
- the coefficients values are too **extreme (we have not seen this yet)**

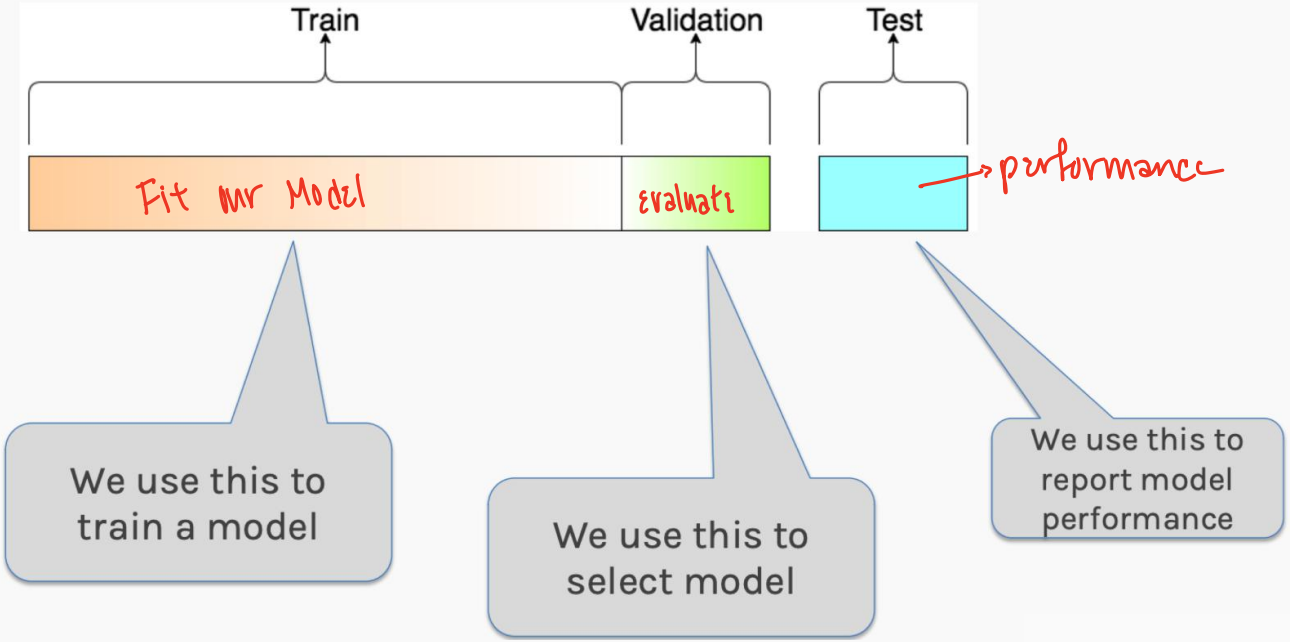
Generalization Error

We know to evaluate the model on both train and test data, because models that do well on training data may do poorly on new data (overfitting).

The ability of models to do well on new data is called generalization.

The goal of model selection is to choose the model that generalizes the best.

Train-Validation-Test



Model Selection

Question: How many different models when considering J predictors (only linear terms) do we have?

Example: 3 predictors (X_1, X_2, X_3)

- Models with 0 predictor:
  - M0:
- Models with 1 predictor:
  - M1: X_1
  - M2: X_2
  - M3: X_3
- Models with 2 predictors:
  - M4: $\{X_1, X_2\}$
  - M5: $\{X_2, X_3\}$
  - M6: $\{X_3, X_1\}$
- Models with 3 predictors:
  - M7: $\{X_1, X_2, X_3\}$

see which works best, & add others & see when is better

2^J Models

Stepwise Variable Selection and Cross Validation

Selecting optimal subsets of predictors (including choosing the degree of polynomial models) through:

slowly "adding" predictors to see which is best

- stepwise variable selection - iteratively building an optimal subset of predictors by optimizing a fixed model evaluation metric each time.
- validation - selecting an optimal model by evaluating each model on validation set.

Stepwise Variable Selection: Forward method

In **forward selection**, we find an 'optimal' set of predictors by iteratively building up our set.

1. Start with the empty set P0, construct the null model M0.

2. For $k = 1, \dots, J$:

2.1 Let M_{k-1} be the model constructed from the best set of $k-1$ predictors, P_{k-1}.

2.2 Select the predictor X_{n_k}, not in P_{k-1}, so that the model constructed from $P_k = X_{n_k} \cup P_{k-1}$ optimizes a fixed metric (this can be p-value, F-stat; validation MSE, R^2, or AIC/BIC on training set).

2.3 Let M_k denote the model constructed from the optimal P_k.

3. Select the model M amongst $\{M_0, M_1, \dots, M_J\}$ that optimizes a fixed metric (this can be validation MSE, R^2; or AIC/BIC on training set)

Stepwise Variable Selection Computational Complexity

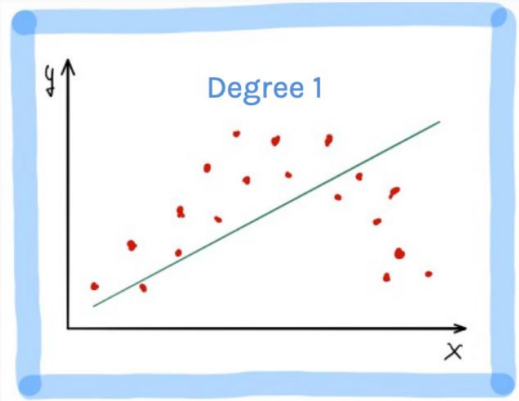
How many models did we evaluate?

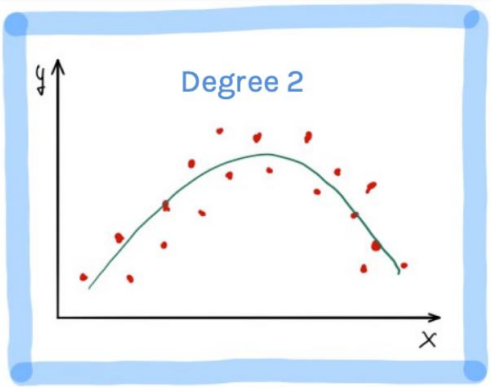
- 1st step, ***J* Models**
- 2nd step, *J-1* **Models** (add 1 predictor out of *J-1* possible)
- 3rd step, *J-2* **Models** (add 1 predictor out of *J-2* possible)
- …

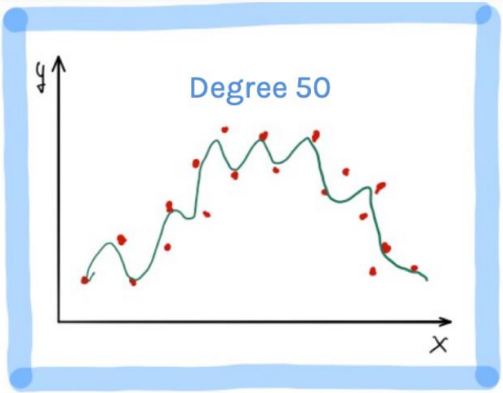
$$O(J^2) \ll 2^J \text{ for large } J$$

Choosing the degree of the polynomial model

Fitting a polynomial model requires choosing a degree.





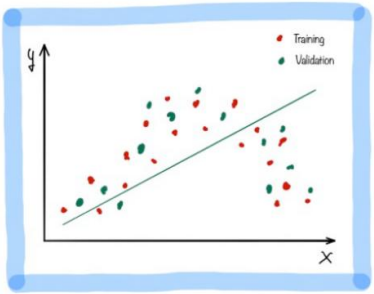


Underfitting: when the degree is too low, the model cannot fit the trend.

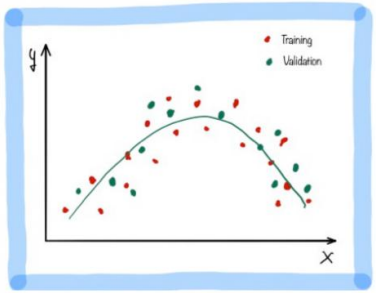
We want a model that fits the trend and ignores the noise.

Overfitting: when the degree is too high, the model fits all the noisy data points.

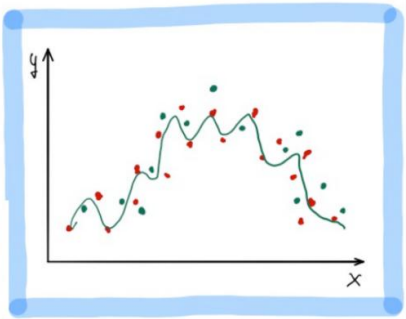
Underfitting: train and validation error is high.

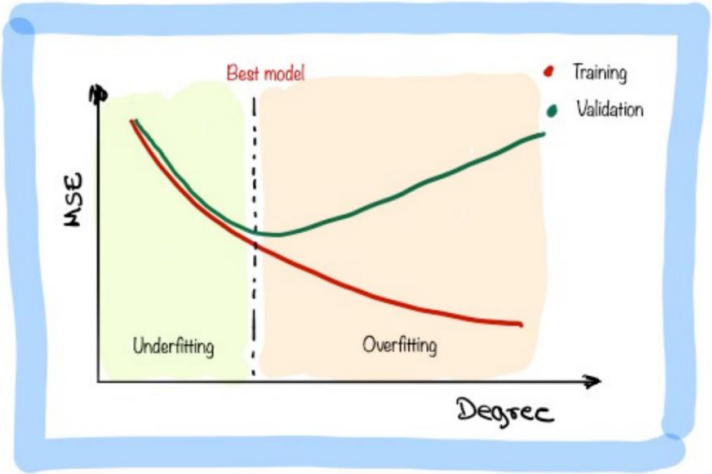


Best model: validation error is minimum.



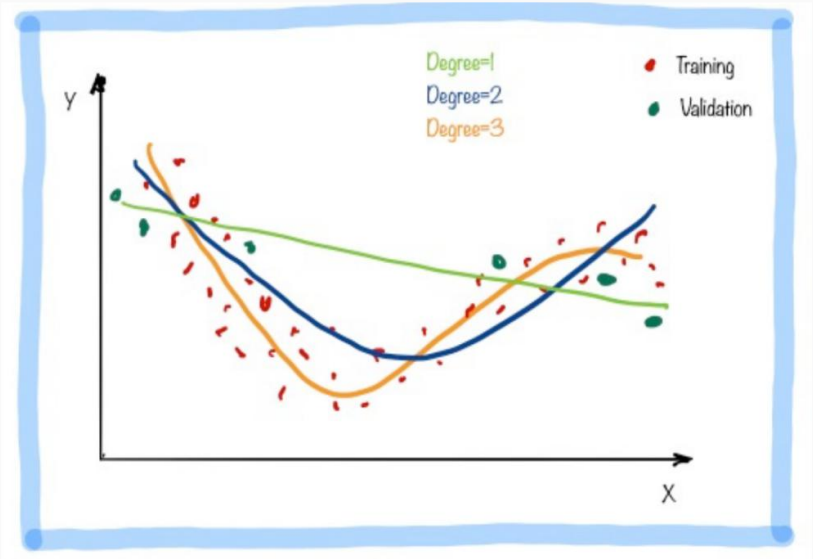
Overfitting: train error is low, validation error is high.





Cross Validation: Motivation

Using a single validation set to select amongst multiple models can be problematic - **there is the possibility of overfitting to the validation set**



It is obvious that degree=3 is the correct model but the validation set by chance favors the linear model.

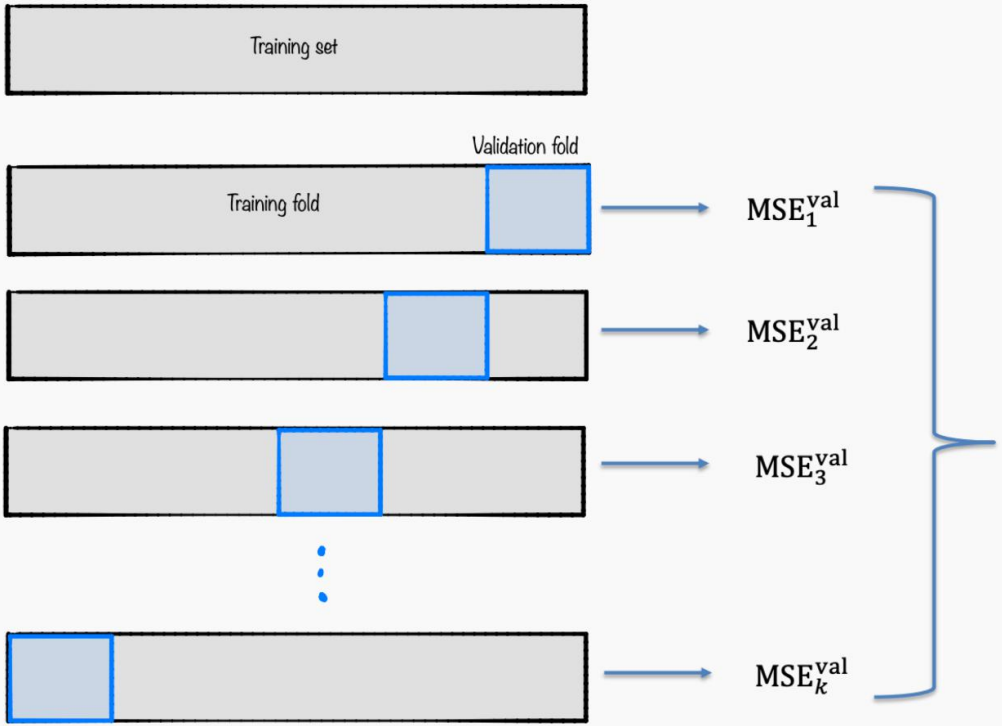
Cross Validation: Motivation

Using a single validation set to select amongst multiple models can be problematic - **there is the possibility of overfitting to the validation set**.

One solution to the problems raised by using a single validation set is to evaluate each model on **multiple** validation sets and average the validation performance.

One can randomly split the training set into training and validation multiple times **but** randomly creating these sets can create the scenario where important features of the data never appear in our random draws.

Cross Validation



$$\text{MSE}^{\text{val}} = \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_i^{\text{val}}$$

K-Fold Cross Validation

Given a data set $\{X_1, \dots, X_n\}$, where each $\{X_1, \dots, X_n\}$ contains *J* features.

To ensure that every observation in the dataset is included in at least one training set and at least one validation set we use the **K-fold validation**:

- split the data into *K* uniformly sized chunks, $\{C_1, \dots, C_K\}$
- we create *K* number of training/validation splits, using one of the *K* chunks for validation and the rest for training.

We fit the model on each training set, denoted $\hat{f}_{C_{-i}}$, and evaluate it on the corresponding validation set, $\hat{f}_{C_{-i}}(C_i)$. The ***cross validation is the performance*** of the model averaged across all validation sets:

$$CV(\text{Model}) = \frac{1}{K}\sum_{i=1}^{K} L(\hat{f}_{C_{-i}}(C_i))$$

where *L* is a loss function.

Leave-One-Out

*Or using the **leave one out*** method:

- validation set: $\{X_i\}$
- training set: $X_{-i} = \{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$

for $i = 1, \dots, n$:

We fit the model on each training set, denoted $\hat{f}_{X_{-i}}$, and evaluate it on the corresponding validation set, $\hat{f}_{X_{-i}}(X_i)$.

The ***cross validation score*** is the performance of the model averaged across all validation sets:

$$CV(\text{Model}) = \frac{1}{n}\sum_{i=1}^{n} L(\hat{f}_{X_{-i}}(X_i))$$

where *L* is a loss function.

Summary so far

Previously in the course

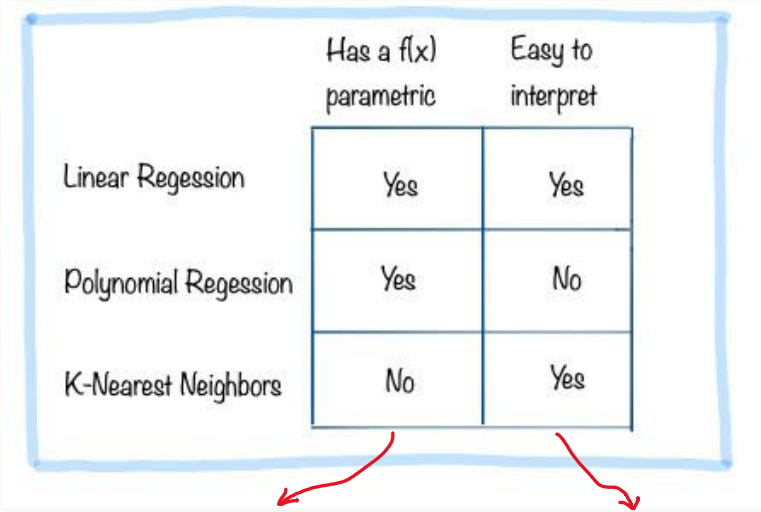
- Statistical model
- k-nearest neighbors (kNN)
- Model fitness and model comparison (MSE)
- Goodness of fit (R2)
- Linear Regression, multi-linear regression and polynomial regression
- Model selection using validation and cross validation
- One-hot encoding for categorical variables
- What is overfitting

Inference in Linear Regression

Uncertainty in estimating the linear regression coefficients

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

Assessing the Accuracy of the Coefficient Estimates

- Bootstrapping and confidence intervals

Evaluating Significance of Predictors

- Does the outcome depend on the predictors?
- Hypothesis testing

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

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

Outline

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, you get 1.01 back in sales, which is 1% net increase.

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

Why aren't we certain?

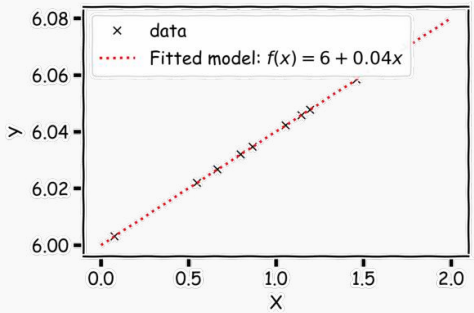
Confidence intervals for the predictors estimates

We interpret the ε term in our observation

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

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

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



Confidence intervals for the predictors estimates (cont)

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

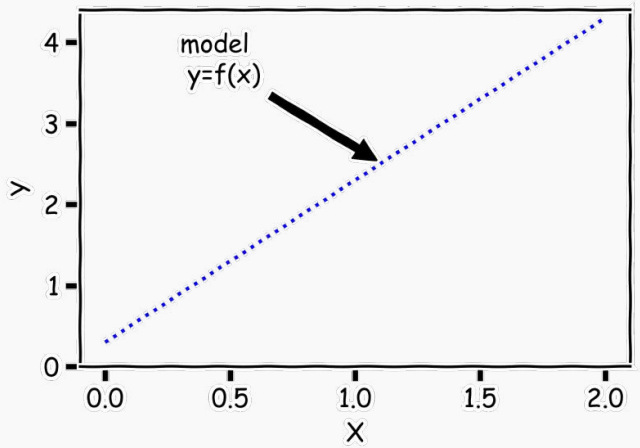
- observational error is always there – this is called ***aleatoric*** error, or ***irreducible*** error.
- we do not know the exact form of $f(x)$ - this is called ***misspecification*** error and it is part of the *epistemic* error

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

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

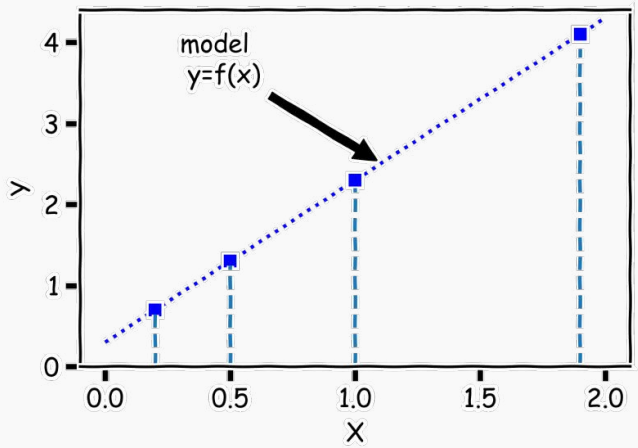
Confidence intervals for the predictors estimates (cont)

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



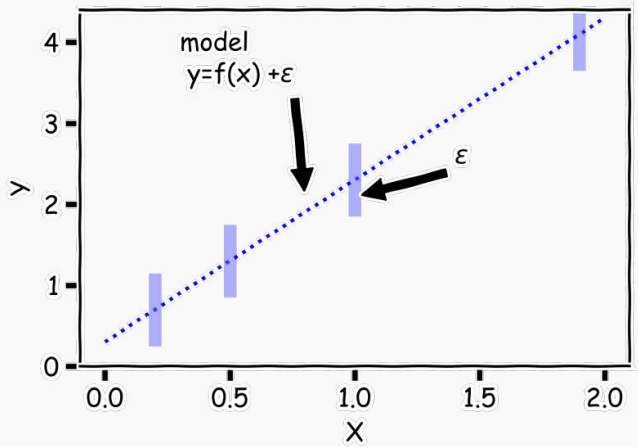
Confidence intervals for the predictors estimates (cont)

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



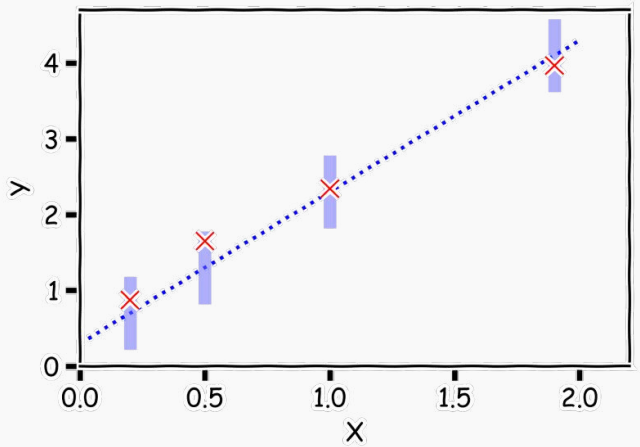
Confidence intervals for the predictors estimates (cont)

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



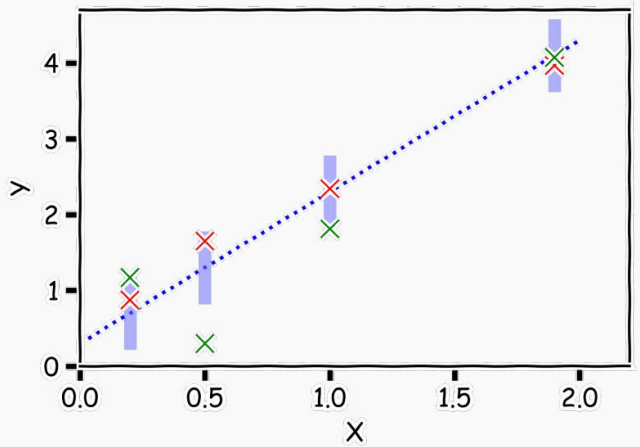
Confidence intervals for the predictors estimates (cont)

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



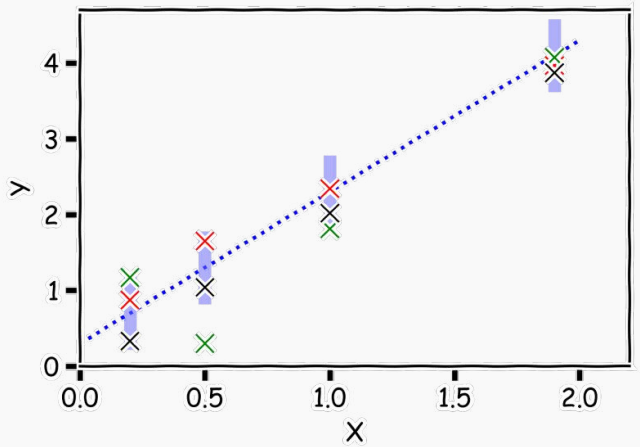
Confidence intervals for the predictors estimates (cont)

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



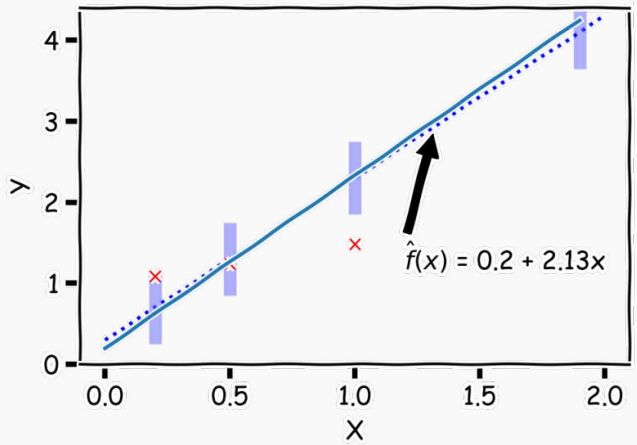
Confidence intervals for the predictors estimates (cont)

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



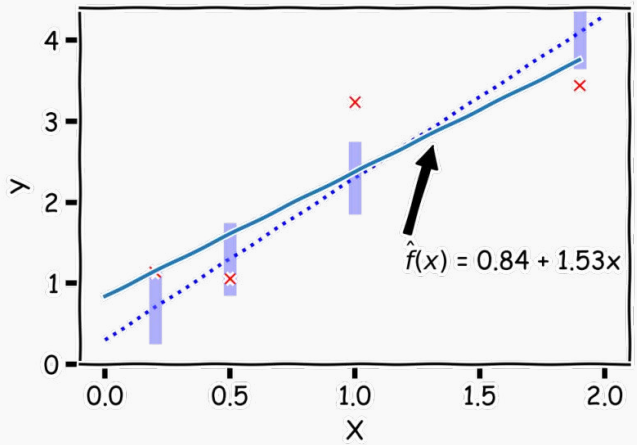
Confidence intervals for the predictors estimates (cont)

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



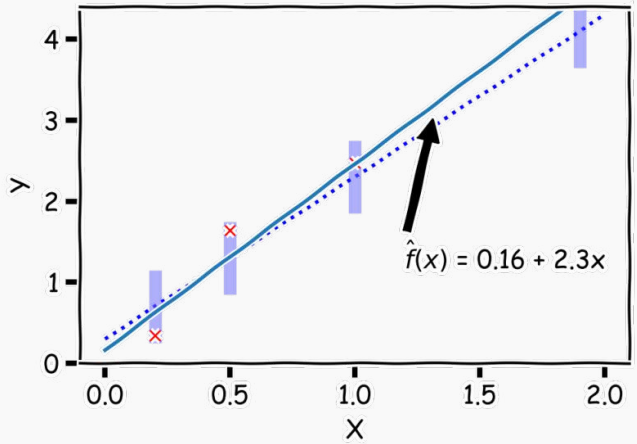
Confidence intervals for the predictors estimates (cont)

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 $\hat{\beta}_1$.

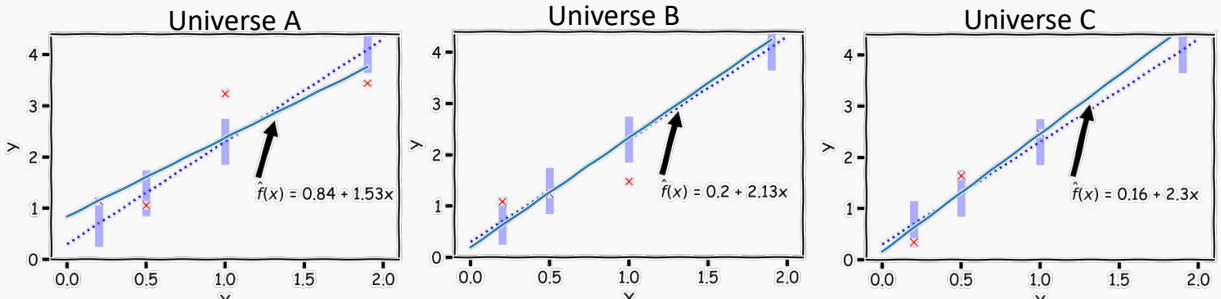


Confidence intervals for the predictors estimates (cont)

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.

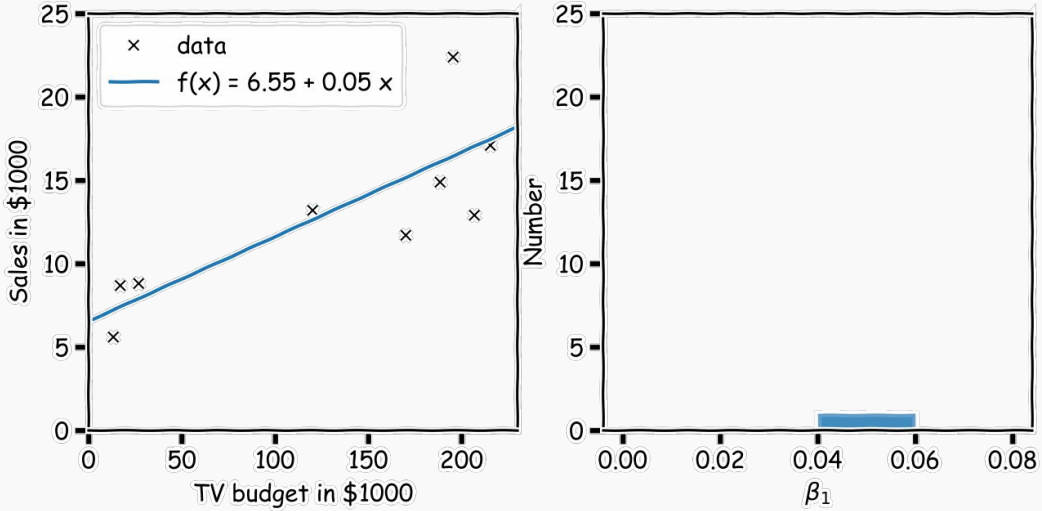
Question: If this is just one realization of reality, how do we know the truth? How do we deal with this conundrum?

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)

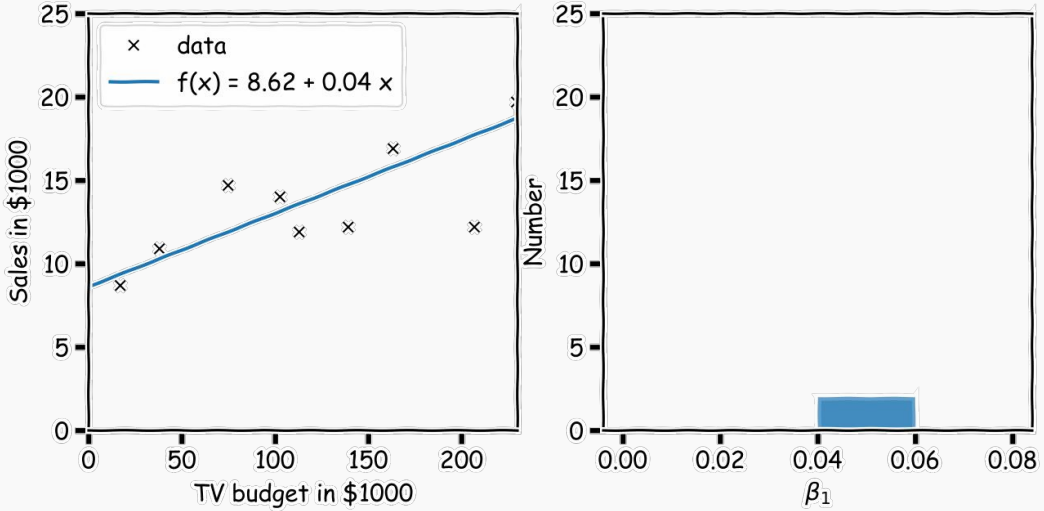
In our magical realisms, we can now sample multiple times. One universe, one sample, one set of estimates for $\hat{\beta}_0, \hat{\beta}_1$



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

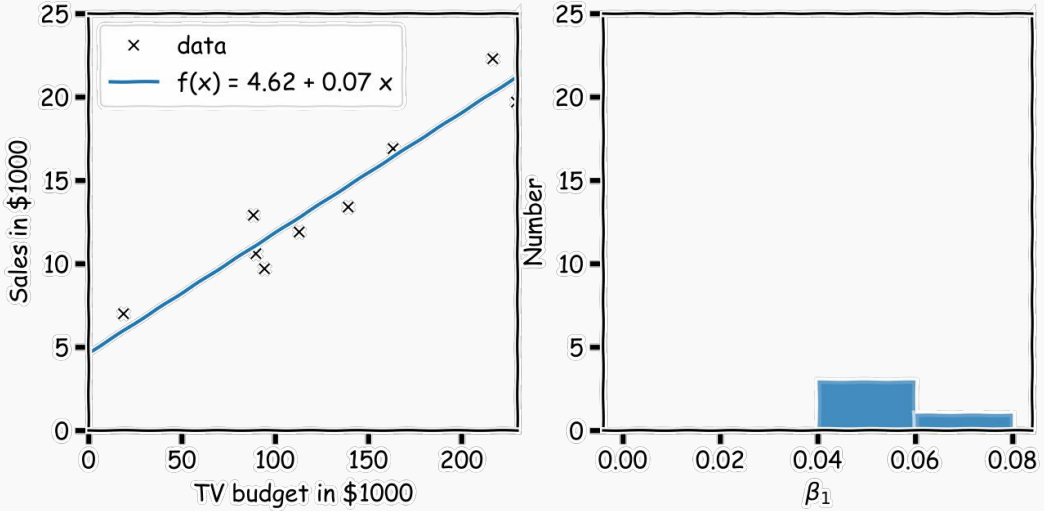
Confidence intervals for the predictors estimates (cont)

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



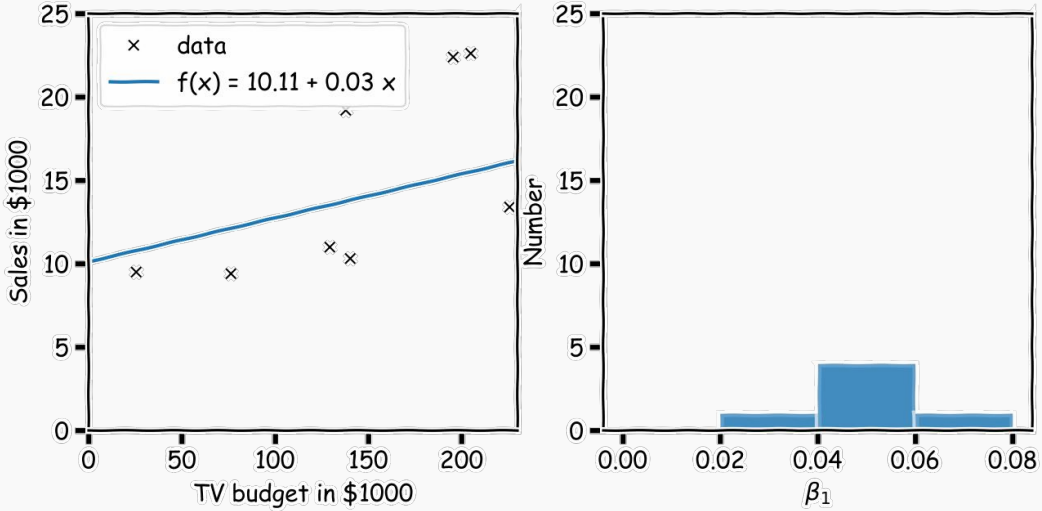
Confidence intervals for the predictors estimates (cont)

Again



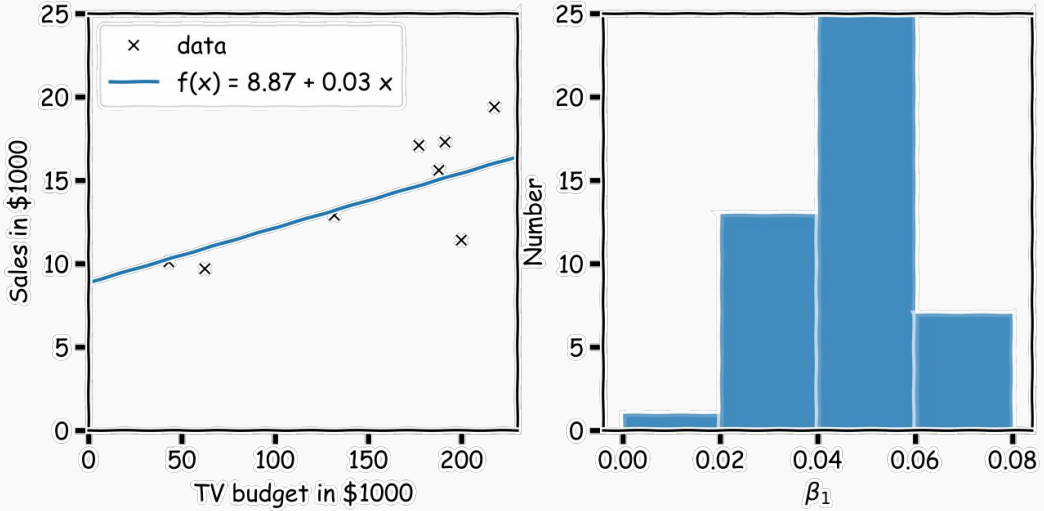
Confidence intervals for the predictors estimates (cont)

And again



Confidence intervals for the predictors estimates (cont)

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



Bootstrapping and Confidence Intervals

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.

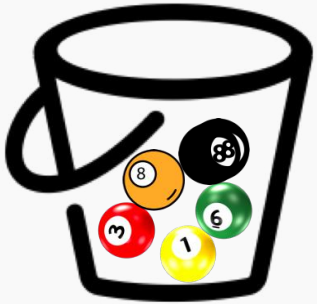
Bootstrapping is the practice of sampling from the observed data (X,Y) in estimating statistical properties.

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.

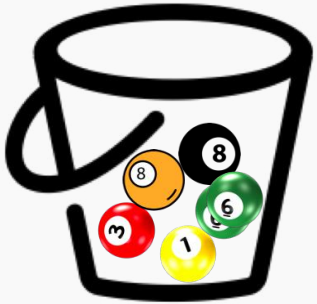


Bootstrap

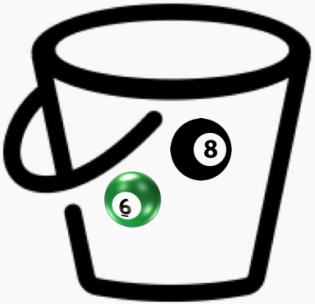
Bootstrap

We then randomly pick another ball and again we replicate it.

As before, we move the replicated ball to the other bucket.



Bootstrap

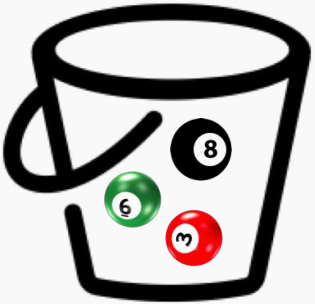


Bootstrap

We repeat this process.

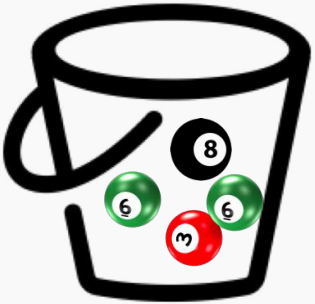
Bootstrap

Again



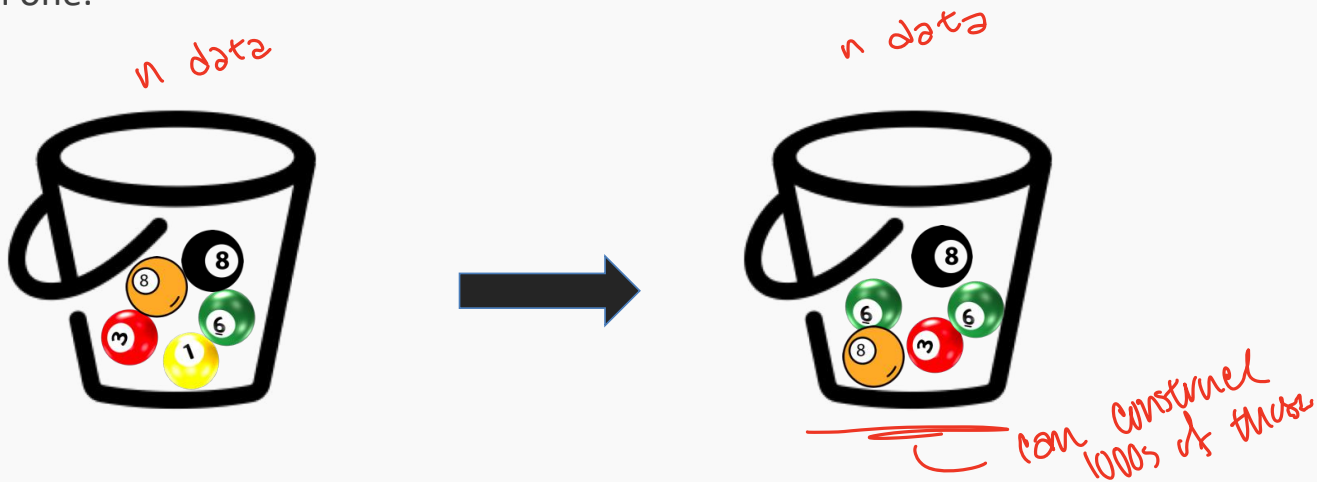
Bootstrap

And again



Bootstrap

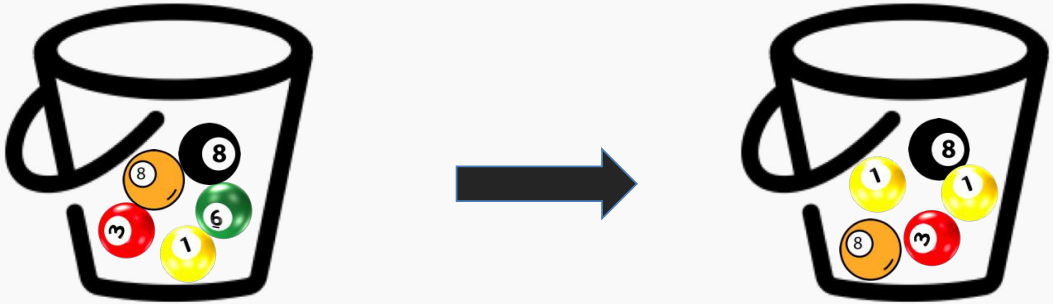
We continue until the “other” bucket has **the same number of balls** as the original one.



This new bucket represents a new parallel universe

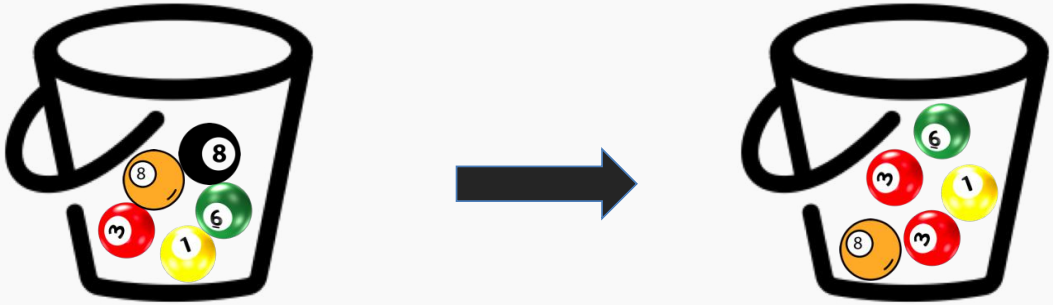
Bootstrap

We repeat the same process and acquire another sample.

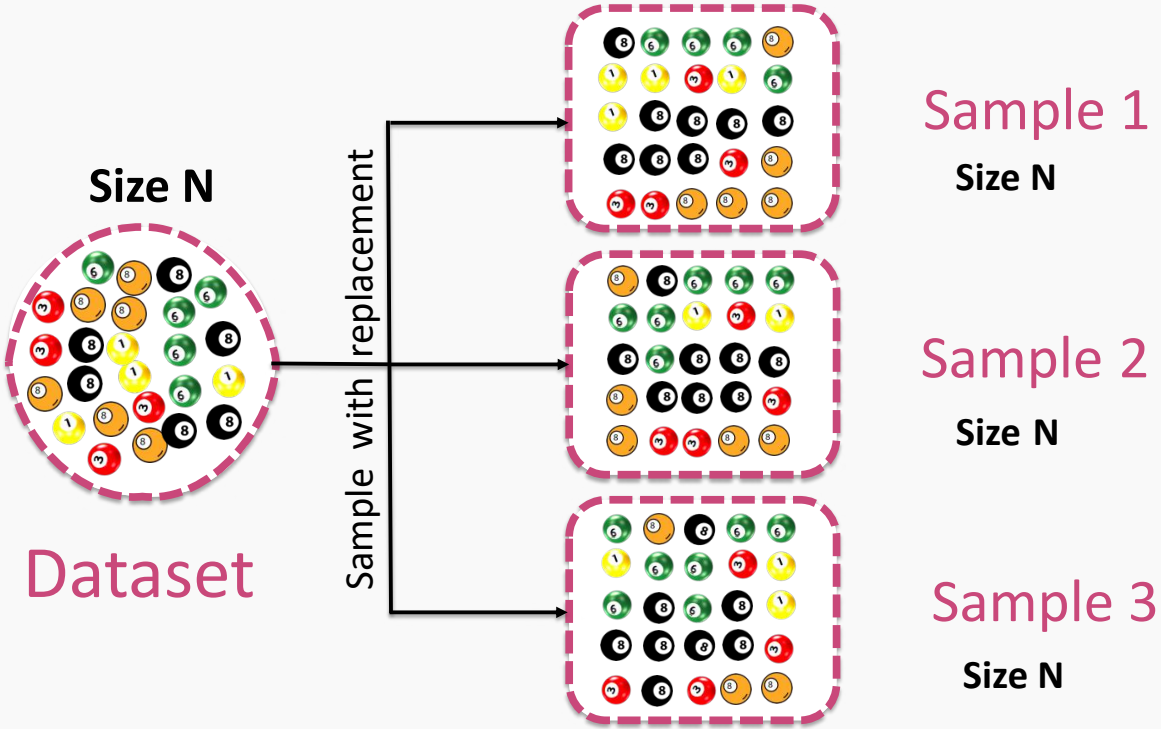


Bootstrap

We repeat the same process and acquire another sample.



Bootstrap


Size N
Dataset
Sample with replacement
Sample 1
Size N
Sample 2
Size N
Sample 3
Size N

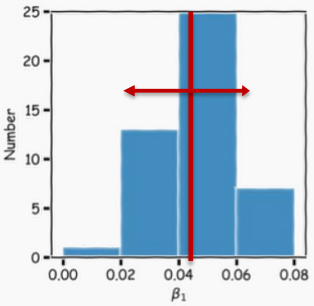
Bootstrapping for Estimating Sampling Error

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 $\hat{\beta}_1$.

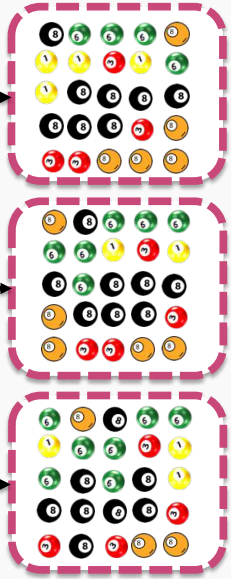
Bootstrap



Size N

Dataset

Sample with replacement



train

Model 1: $\hat{y} = \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x$

train

Model 2: $\hat{y} = \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x$

train

Model s: $\hat{y} = \hat{\beta}_0^{(s)} + \hat{\beta}_1^{(s)} x$

$$\mu_{\hat{\beta}} = \frac{1}{s}\sum_{i=1}^{s} \hat{\beta}^{(i)}$$

$$\sigma_{\hat{\beta}} = \sqrt{\frac{1}{s}\sum_{i=1}^{s} \left(\hat{\beta}^{(i)} - \bar{\beta}\right)^2}$$

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

We can empirically estimate the standard deviations $\sigma_{\hat{\beta}}$ which are called the **standard errors,** $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ through bootstrapping.

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 formula below (no need to bootstrap):

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

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

Where n is the number of observations

$\bar{x}$ is the mean value of the predictor.

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

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

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

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

$$\sigma(\epsilon) = \sqrt{\sum \frac{\left(\hat{f}(x) - y_i\right)^2}{n - 2}}$$

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

unbiased estimates of β_0, β_1

Standard Errors

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

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:

$$\sigma_\epsilon = \sqrt{\frac{n \cdot MSE}{n-2}} = \sqrt{\sum \frac{\left(\hat{f}(x) - y_i\right)^2}{n-2}}$$

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

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

SE increase

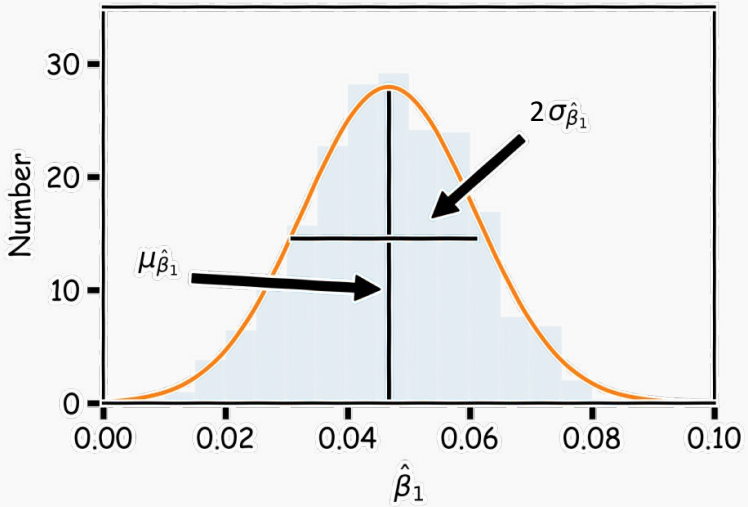
The coefficients for TV advertising but with added **extra** noise:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.028
Bootstrap	0.023

SE increase

Confidence intervals for the predictors estimates (cont)

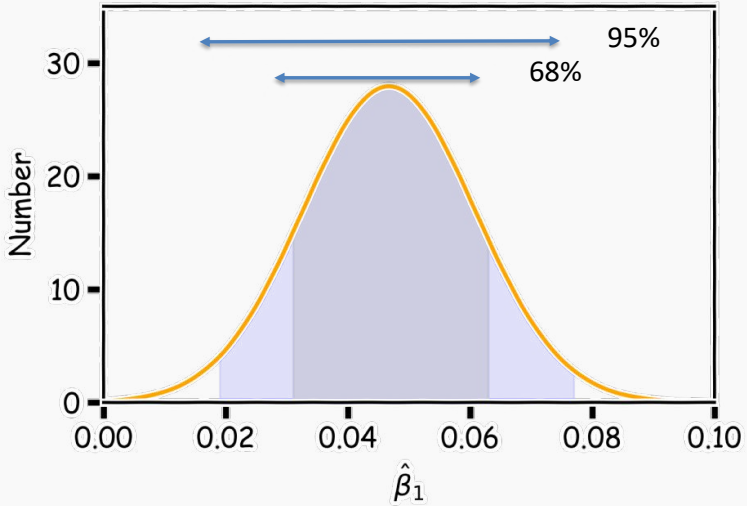
We can now estimate the mean and standard deviation of the estimates of $\hat{\beta}_0, \hat{\beta}_1$.



Confidence intervals for the predictors estimates (cont)

The standard errors give us a sense of our uncertainty over our estimates.

Typically we express this uncertainty as a 95% confidence interval, which is the range of values such that the **true** value of β_1 is contained in this interval with *95%* percent probability.



$$CI_{\hat{\beta}} = (\hat{\beta} - 2\sigma_{\hat{\beta}}, \hat{\beta} + 2\sigma_{\hat{\beta}})$$

Estimating Significance of Predictors
Hypothesis testing

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.

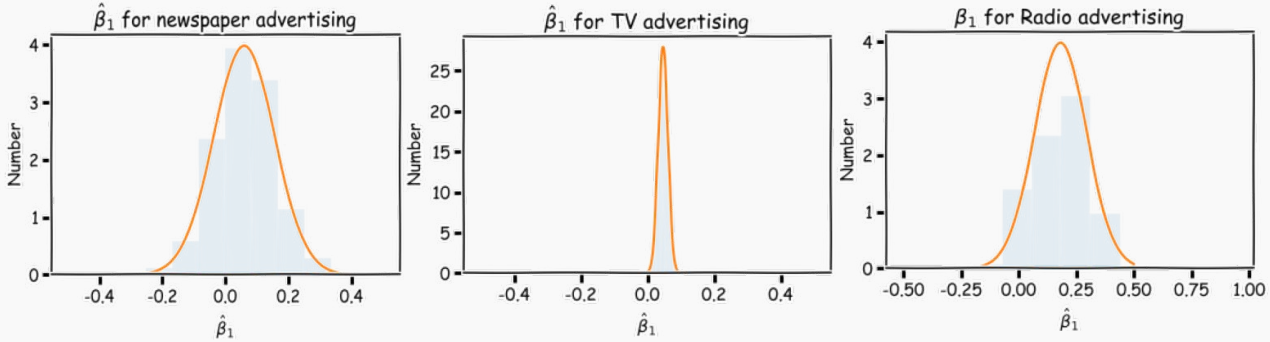
But how certain are we in our estimation of the coefficient 1.01?

Now you know how certain you are in your estimates, will you want to change your answer?

Feature importance

Now we know how to generate these distributions we are ready to answer ***two important questions:***

A. Which predictors are most important?
B. And which of them really affect the outcome?

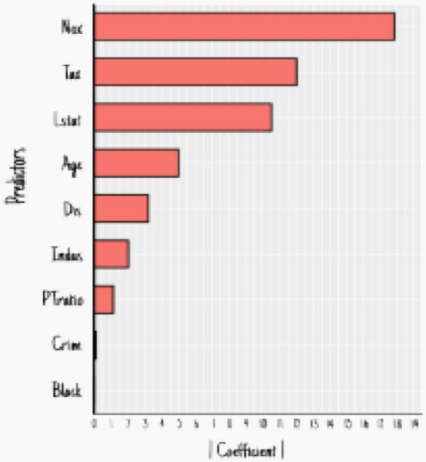


$\mu_{\beta_1} = 0.03$
σ_{β_1}=0.13

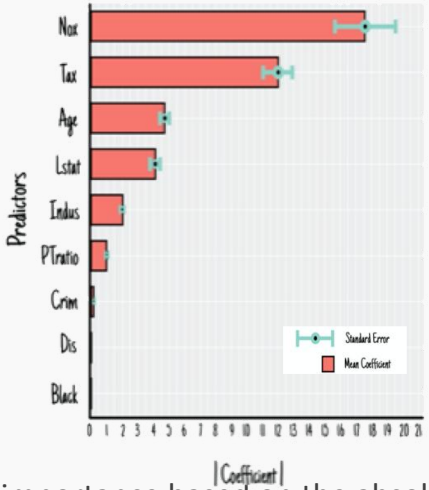
$\mu_{\beta_1} = 0.033$
σ_{β_1}=0.01

$\mu_{\beta_1} = 0.23$
σ_{β_1}=0.25

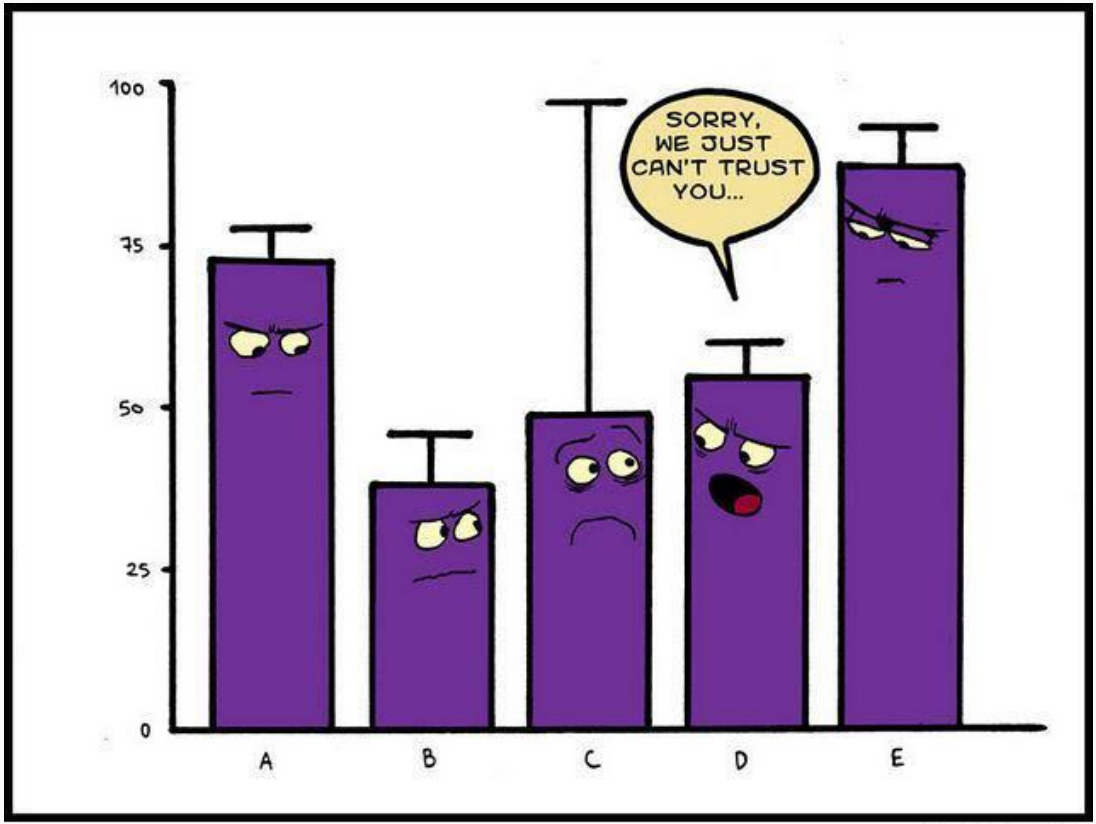
The example below is from Boston housing data. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston. The coefficients below are from a model that predicts prices given house size, age, crime, pupil-teacher ratio, etc.



Feature importance based on the absolute value of the coefficients.



Feature importance based on the absolute mean value of the coefficients over multiple bootstraps and includes the uncertainty of the coefficients.


100
75
50
25
0
A
B
C
D
E
SORRY, WE JUST CAN'T TRUST YOU...
WeKnowMemes

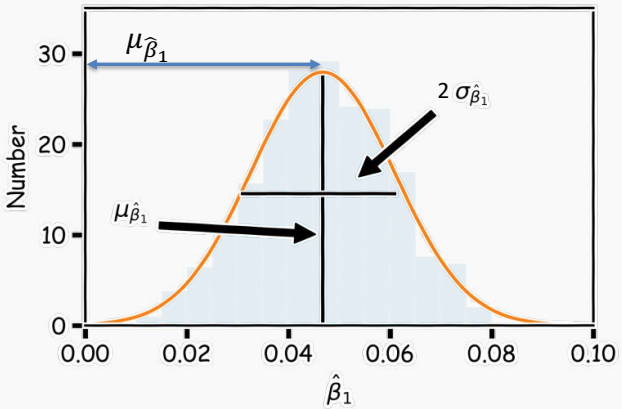
Feature Importance

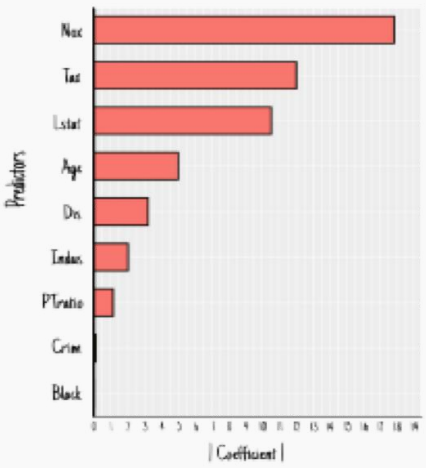
To incorporate the coefficients' uncertainty, we need to determine whether the estimates of $\beta's$ are sufficiently far from zero.

To do so, we define a new metric, which we call t-test statistic:

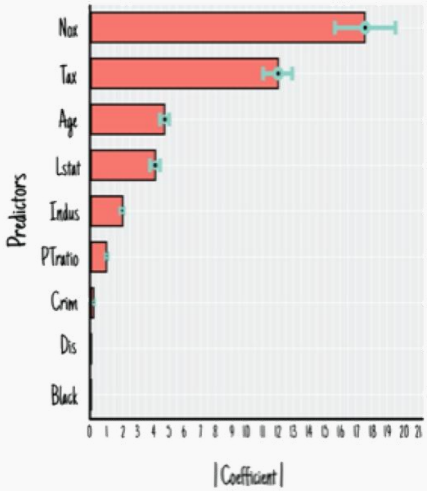
$$t = \frac{\mu_{\widehat{\beta}_1}}{\sigma_{\widehat{\beta}_1}}$$

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

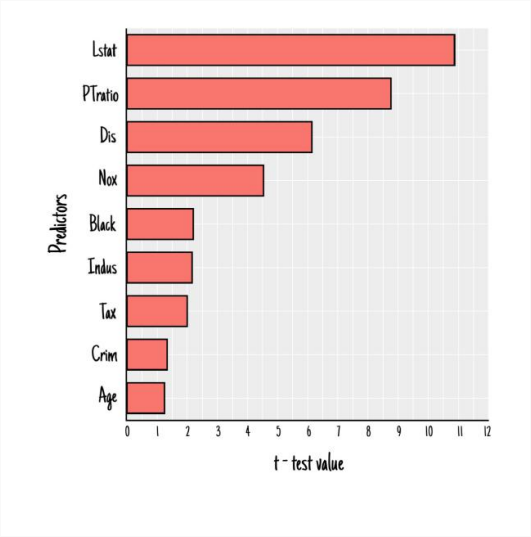




Feature importance base on the absolute value of the coefficients.



Feature importance base on the absolute value of the coefficients over multiple bootstraps and includes the uncertainty of the coefficients.



Feature importance base on t-test. Notice the rank of the importance has changed.

Feature Importance

Because a predictor is ranked as the most important, it does not necessarily mean that the outcome depends on that predictor.

How do we assess if there is a true relationship between outcome and predictors?

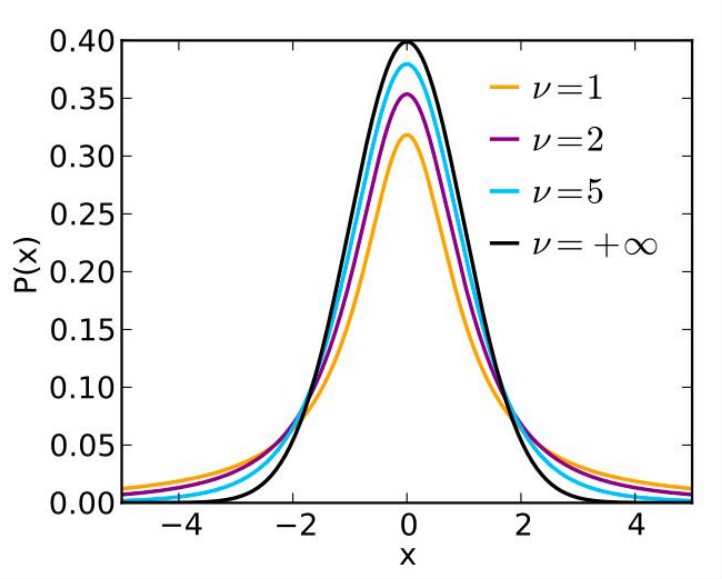
As with R-squared, we should compare its significance (t-test) to the equivalent measure from a dataset where we know that there is no relationship between predictors and outcome.

We are sure that there will be no such relationship in data that are randomly generated. Therefore, we want to compare the t-test of the predictors from our model with t-test values calculated using random data.

1. For n random datasets fit n models.
2. Generate distributions for all predictors and calculate the means and standard errors $(\mu_{\hat{\beta}}, \sigma_{\hat{\beta}})$.
3. Calculate the t-tests.

Repeat and create a probability density function (pdf) for all the t-tests.

It turns out we do not have to do this, because this is a known distribution called student-t distribution.



Student-t distribution, where ν is the degrees of freedom (number of data points minus number of predictors).

To learn more about why student-t, what are degrees of freedom and more details see https://en.wikipedia.org/wiki/Student%27s_t-test

P-value

To compare the t-test values of the predictors from our model, $|t^*|$, with the t-tests, calculated using random data, $|t^R|$, we estimate the probability of observing $|t^R| \geq |t^*|$.

We call this probability the p-value.

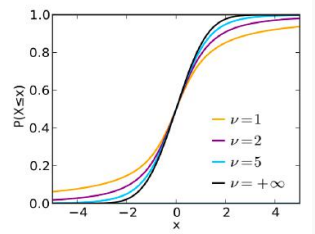
$$p - value = P(|t^R| \geq |t^*|)$$

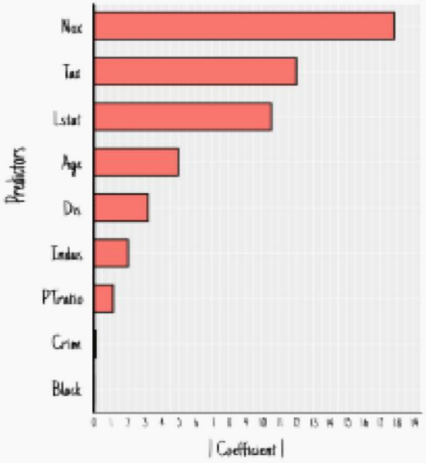
small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.

It is common to use p-value<0.05 as the threshold for significance.

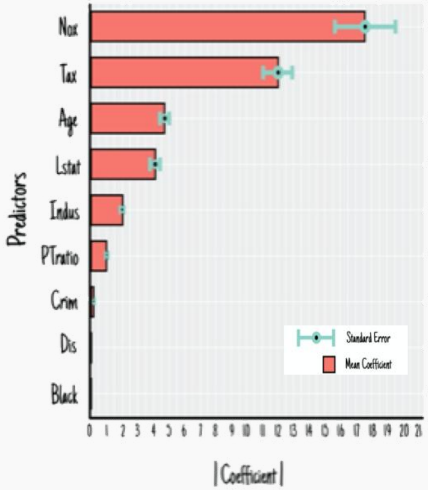
To calculate the p-value we use the cumulative distribution function (CDF) of the student-t.

`stats model` a python library has a build-in function `stats.t.cdf()` which can be used to calculate this.

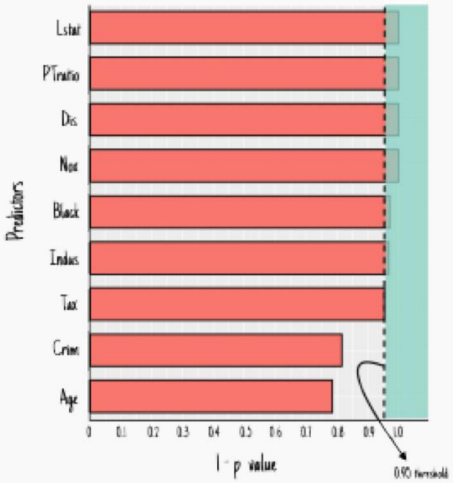




Feature importance based on the absolute value of the coefficients over multiple bootstraps and includes the coefficients' uncertainty.



Feature importance based on t-test. Notice the rank of the importance has changed.



Feature importance using p-value.

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**

t-test

3. Sample:

Using bootstrap we can estimate $\hat{\beta}_1'$s, and $\mu_{\hat{\beta}_1}$ and $\sigma_{\hat{\beta}_1}$ and the t-test.

Hypothesis testing

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

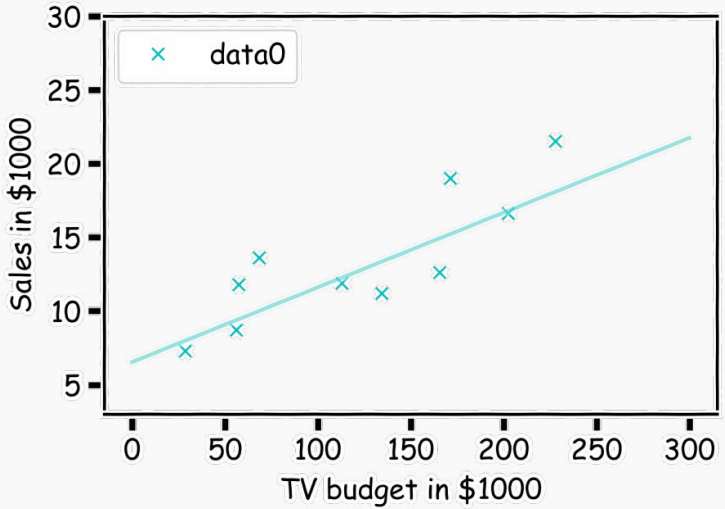
   We compute ***p-value ,*** the probability of observing any value equal to $|t|$ or larger, from random data.

   p-value < p-value-threshold we reject the null.

Prediction Intervals

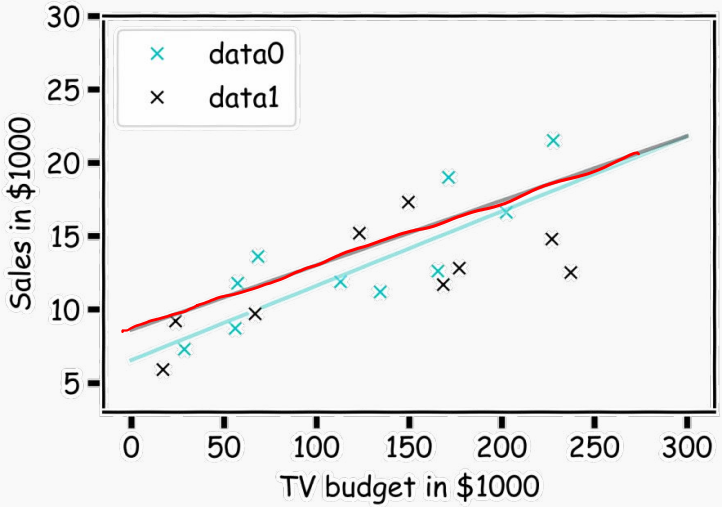
How well do we know $\hat{f}$?

Our confidence in f is directly connected with our confidence in βs. For each bootstrap sample, we have one β, which we can use to determine the model, $f(x) = X\beta$.



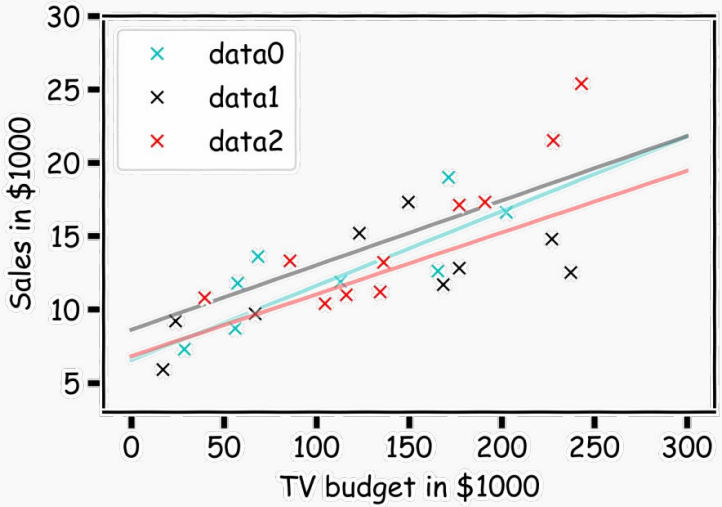
How well do we know $\hat{f}$?

Here we show two difference models predictions given the fitted coefficients.



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

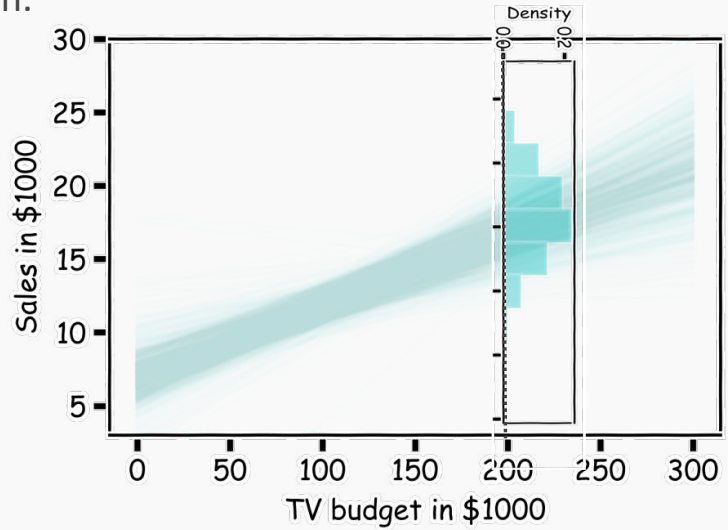
There is one such regression line for every bootstrapped sample.



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

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

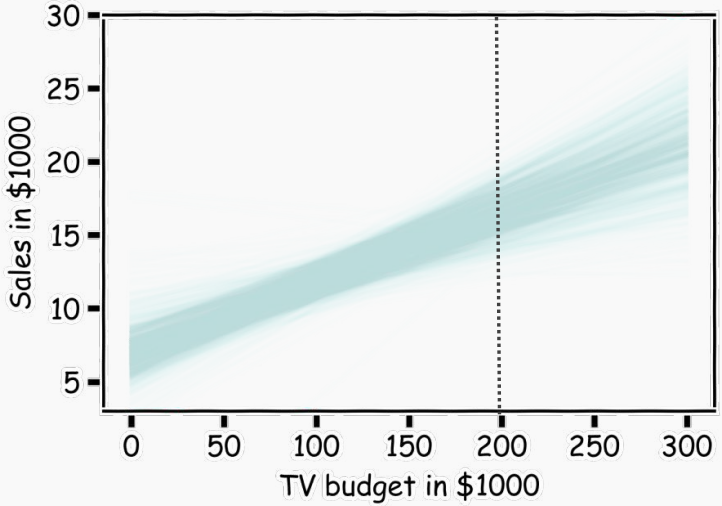
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.

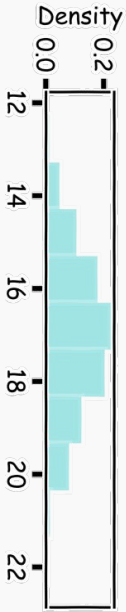


How well do we know $\hat{f}$?

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

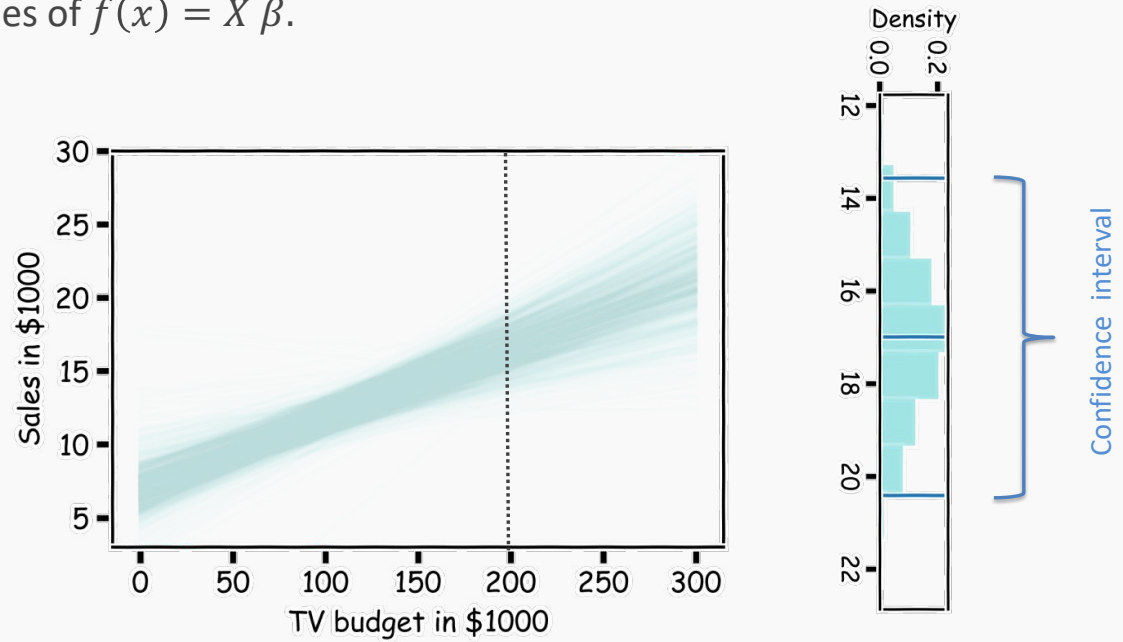
For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.





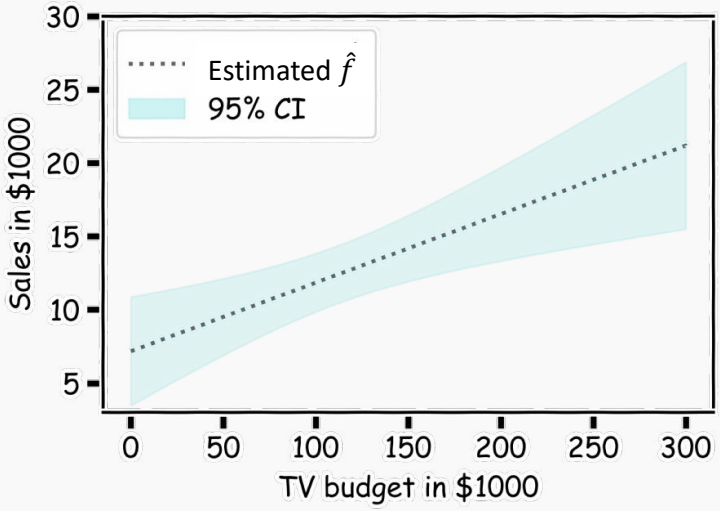
How well do we know $\hat{f}$?

We determine the confidence interval of $\hat{f}$ by selecting the region that contains 95% of the samples of $\hat{f}(x) = X\,\hat{\beta}$.



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

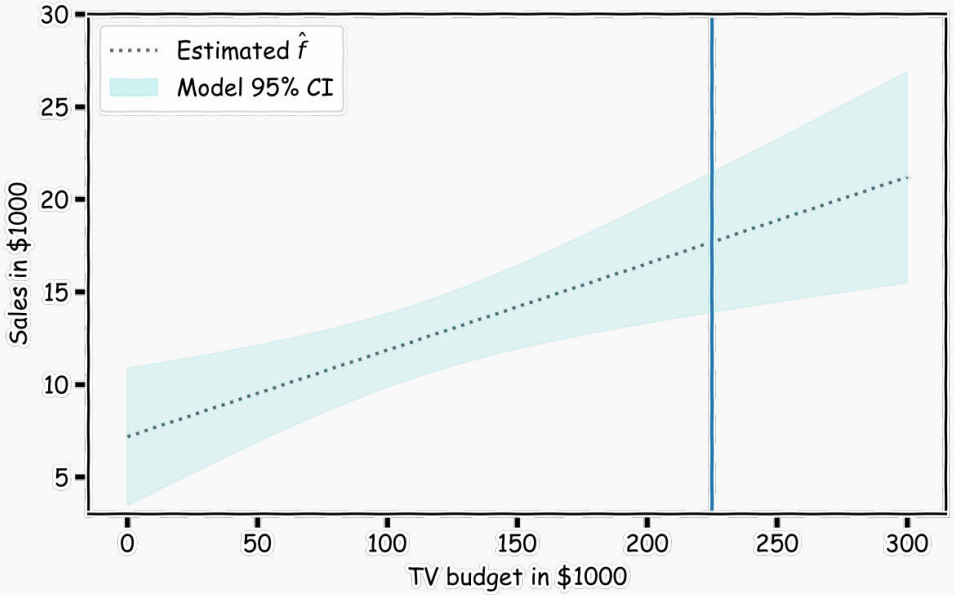
For every x, we calculate the mean of the models, $\widehat{\mu_f}$ (shown with dotted line) and the 95% CI of those models (shaded area).



Confidence in predicting $\hat{y}$

Even if we knew $f(x)$ —the response value cannot be predicted perfectly because of the random error in the model (irreducible error).

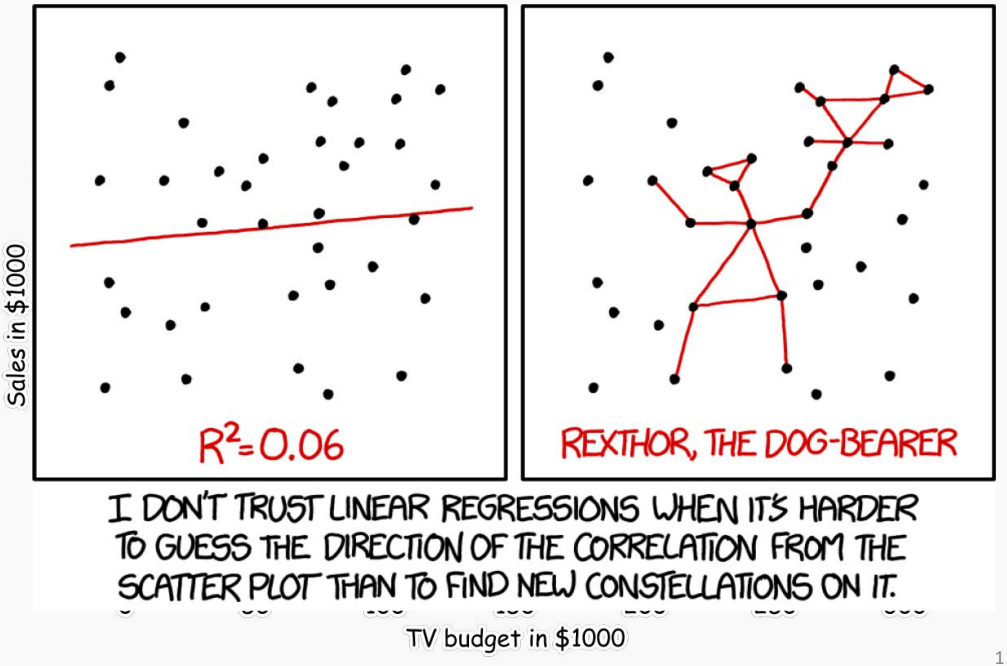
How much will Y vary from $\hat{Y}$? We use prediction intervals to answer this question.



Confidence in predicting $\hat{y}$

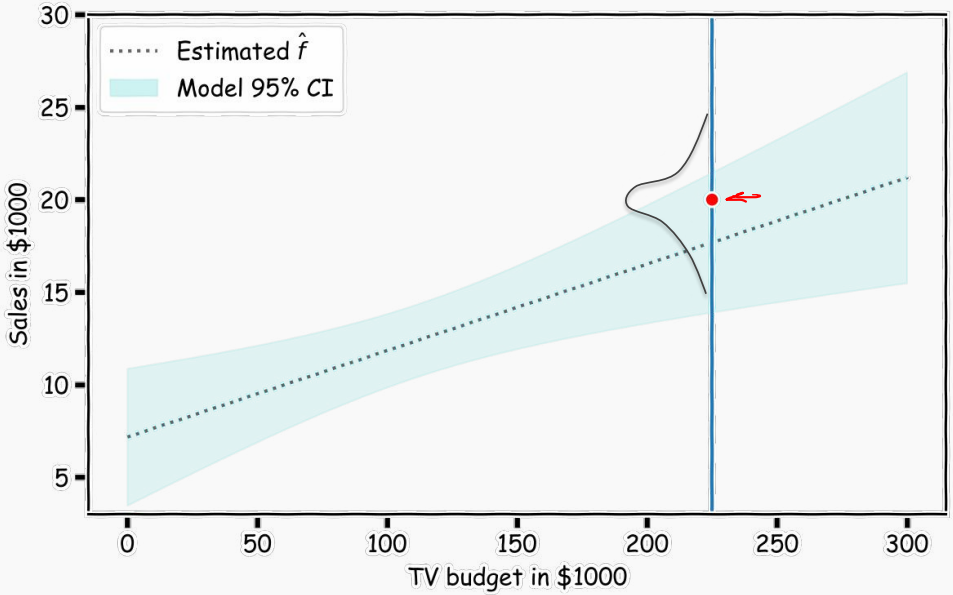
Even if we knew $f(x)$ —the response value cannot be predicted perfectly because of the random error in the model (irreducible error).

How much will Y vary from $\hat{Y}$? We use prediction intervals to answer this question.



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 $y \sim N(f(x), \sigma_\epsilon)$



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 $y \sim N(f(x), \sigma_\epsilon)$
- The prediction confidence intervals are then …