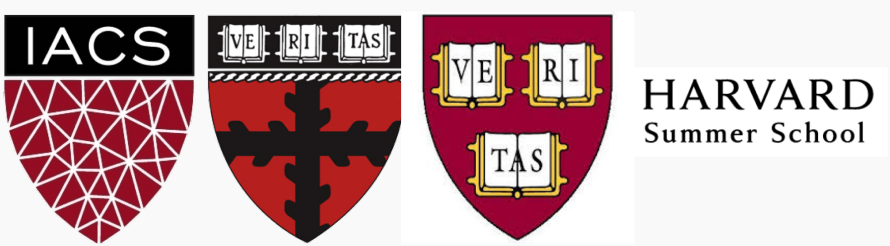
Lecture #3: k-nn and linear regression

CS-S109A: Introduction to Data Science

Kevin Rader



ACADEMY HONESTY

Ethical behavior is an important trait of a Data Scientist, from ethically handling data to attribution of code and work of others. Thus, in CS109 we give a strong emphasis to **Academic Honesty** .

As a student your best guidelines are to be reasonable and fair. We encourage teamwork problem sets, but **you should not split** the homework and you **should work on all the problems together**.

Engaging in not acceptable behavior regarding academic honesty will be handled appropriately.

ACADEMY HONESTY

Engaging in not acceptable behavior regarding academic honesty will be handled harshly.

Please be responsible and **when in doubt ask the CS109 Teaching Staff (Kevin and Chris, presumably)**.

ACADEMY HONESTY

ACCEPTABLE:

➔ Discussing materials and engaging in OH.

➔ Helping debug.

➔ Using a few lines of code found online or other forum as long as you cite the origin and attribute authorship of code.

➔ Searching online to expand your knowledge and for debugging, but not for outright solutions to HW.

➔ Using a tutor, provided the tutor does not do your work for you.

ACADEMY HONESTY

NOT ACCEPTABLE:

➔ Accessing a solution to some problem prior to submitting your own.

➔ Failing to cite the origins of code or techniques that you discover outside of the course's own lessons and integrate into your own work.

➔ Searching for or soliciting outright solutions to problem sets online or elsewhere. Or providing solutions to problem sets to individuals who might take this course in the future.

➔ Paying or offering to pay an individual for work that you may submit as your own.

➔ Splitting a problem set's workload with another individual and combining your work.

Background

Roadmap:

Lecture 1	What is Data Science? Data: types, formats, issues, etc, and briefly visualization
Lecture 2 and Lab1	How to quickly prepare data and scrape the web
This lecture (and next 2 lectures)	**How to model data and evaluate model fitness.** Linear regression, confidence intervals, model selection cross validation, regularization

Lecture Outline

Statistical Modeling

- k-Nearest Neighbors (kNN)
- Simple Linear Regression

Model Fitness

- How does the model perform at predicting?

Comparison of Two Models (train-test splits)

- How do we choose from two different models?

Probabilistic Basis of Regression

Hypothesis Testing and Confidence Intervals ← inferences

Bootstrap Resampling

Predicting a Variable

Let's image a scenario where we'd like to predict one variable using another (or a set of other) variables. Or would like to determine the association of one variable with another (or set of other) variables.

Examples:

- Predicting the amount of view a YouTube video will get next week based on video length, the date it was posted, previous number of views, etc.
- Predicting which movies a Netflix user will rate highly based on their previous movie ratings, demographic data etc.
- Determining what background factors are associated with how much time a student spends on HW1.

Data

The **Advertising data set** consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. Everything is given in units of $1000.

Y = response

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

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

Response vs. Predictor Variables

There is an **asymmetry** in many of these problems:

> The variable we'd like to predict may be more difficult to measure, is more important than the other(s), or may be directly or indirectly influenced by the values of the other variable(s).

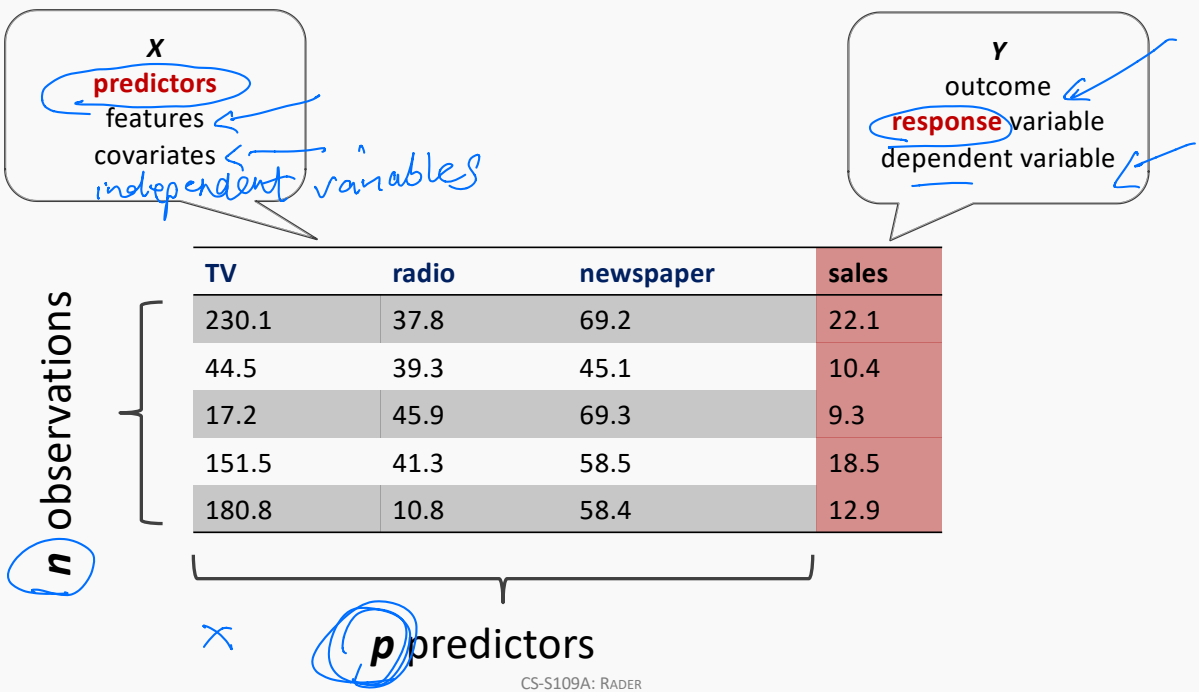
Thus, we'd like to define two categories of variables:

- variables whose value we want to predict (response, Y)
- variables whose values we use to make our prediction (predictors, X_i)

Sales

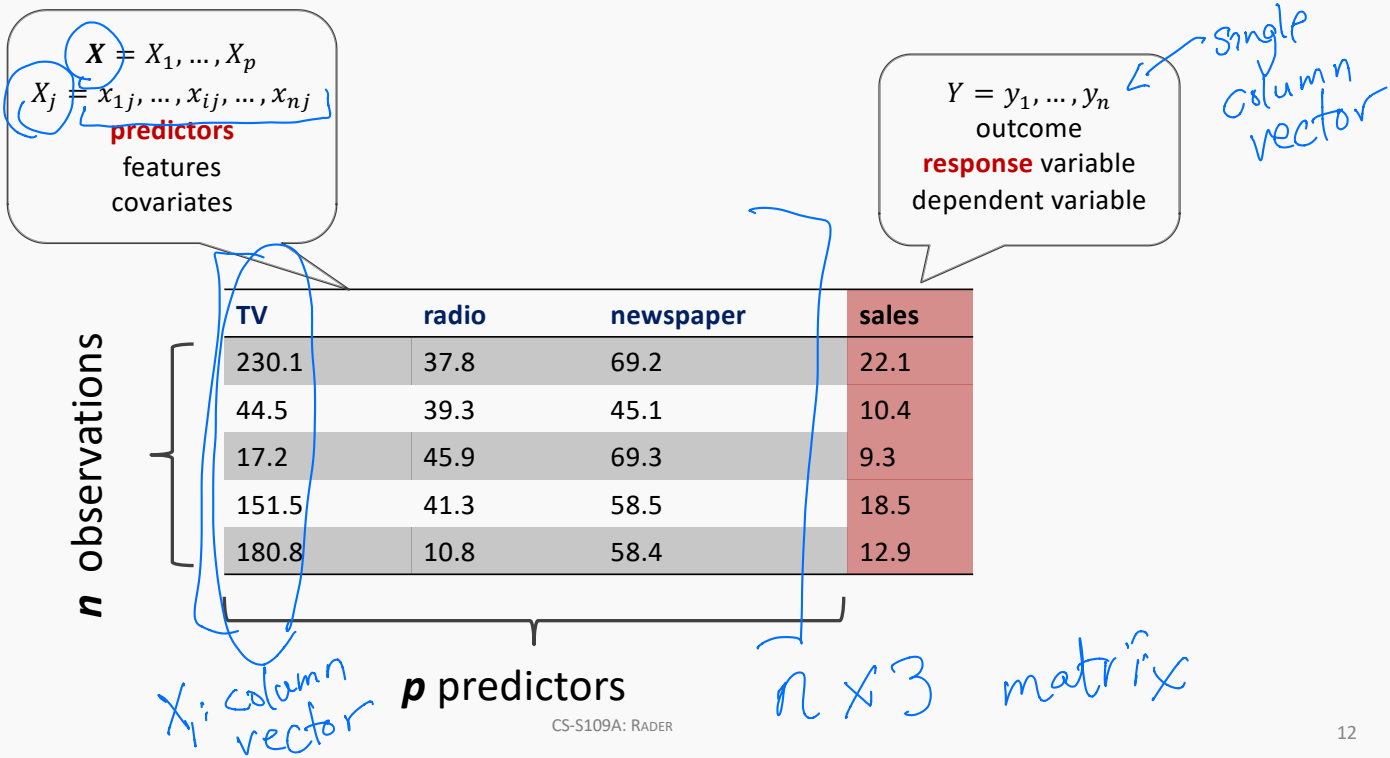
← 3 X's: TV, newspaper, radio

Response vs. Predictor Variables



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

Response vs. Predictor Variables



Definition

We are observing $p + 1$ variables (the p predictors + the response variable)and we are making n sets of observations. We call:

- the variable we'd like to predict the **outcome** or **response variable**; typically, we denote this variable by Y and the individual measurements y_i.
- the variables we use in making the predictions the **features** or **predictor** variables; typically, we denote these variables by $\boldsymbol{X} = X_1, \ldots, X_p$ and the individual measurements $x_{i,j}$.

Note: i indexes the observations ($i = 1, \ldots, n$) and j indexes the value of the j^{th} predictor variable ($j = 1, \ldots, p$).

$x_{3,7}$ ← 3rd observation, in the 7th predictor variable

Statistical Modeling

True vs. Statistical Model

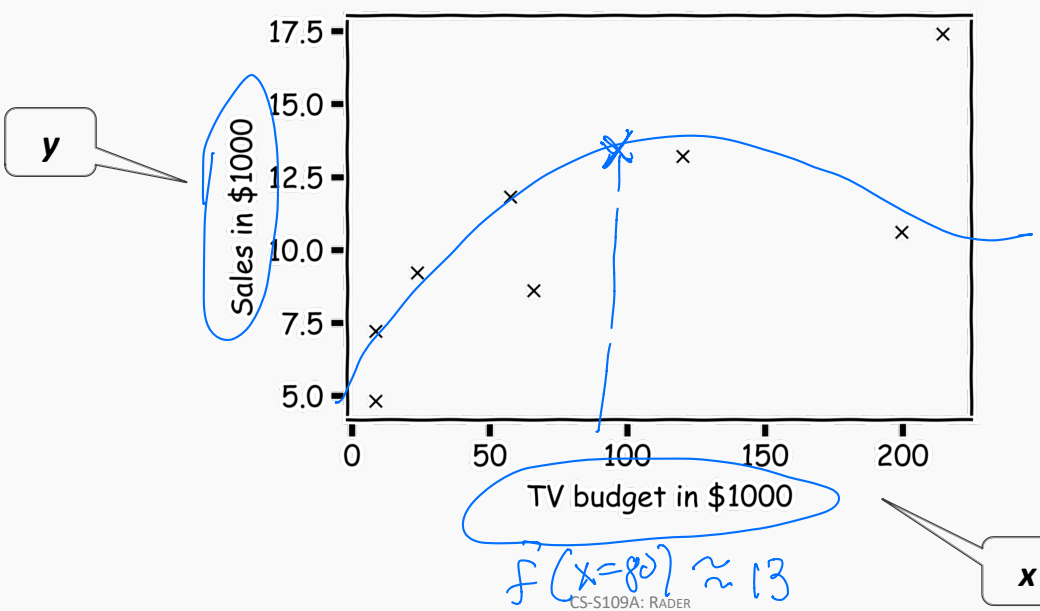
We will assume that the response variable, Y, relates to the predictors, $\boldsymbol{X}$, through some unknown function expressed generally as:

$$Y = f(\boldsymbol{X}) + \varepsilon$$

Here, f is the unknown (true/theoretical) function expressing an underlying rule for relating Y to $\boldsymbol{X}$, ε is the random amount (unrelated to $\boldsymbol{X}$) that Y differs from the rule $f(\boldsymbol{X})$.

A ***statistical model*** is any algorithm that estimates f. We denote the estimated function as $\hat{f}$.

Statistical Model



scatterplot: is there a relationship?

f or $\hat{f}$ estimated from data

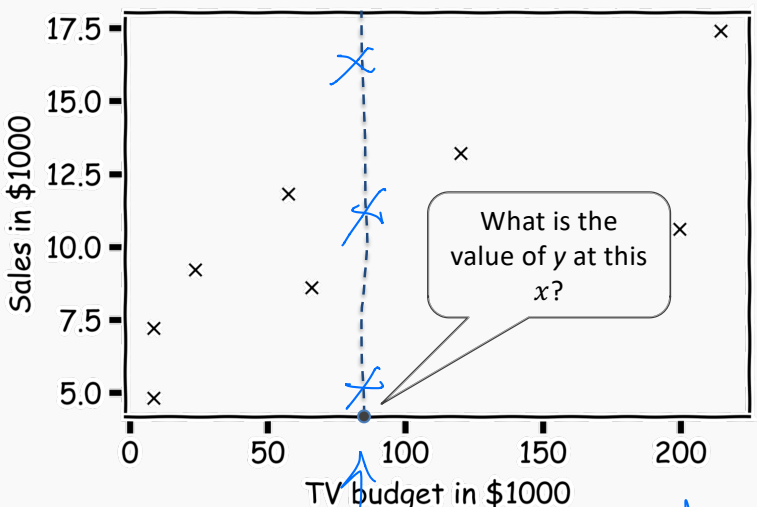
a truth

inputs

$f(x=80) \approx 13$

Statistical Model

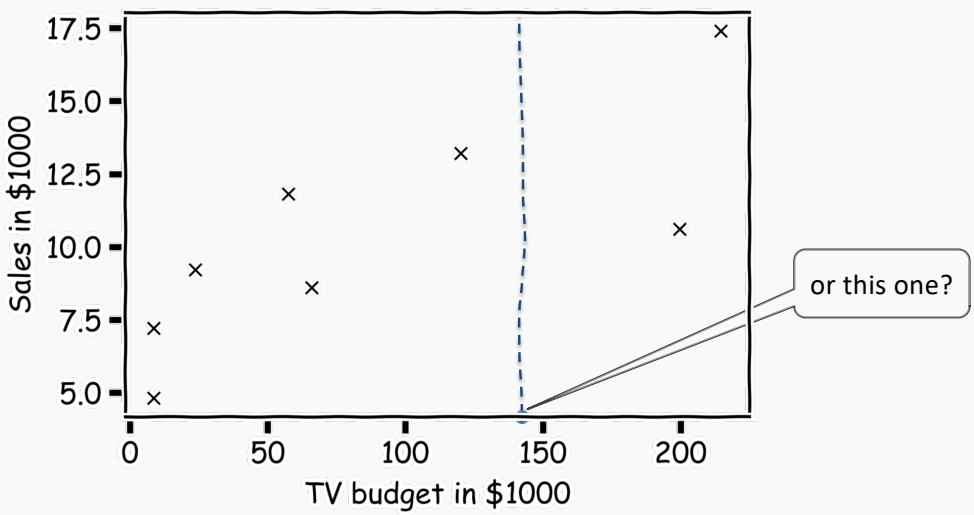
How do we find $\hat{f}(x)$?



$x = 80$, what should $\hat{y} = \hat{f}(x=80)$

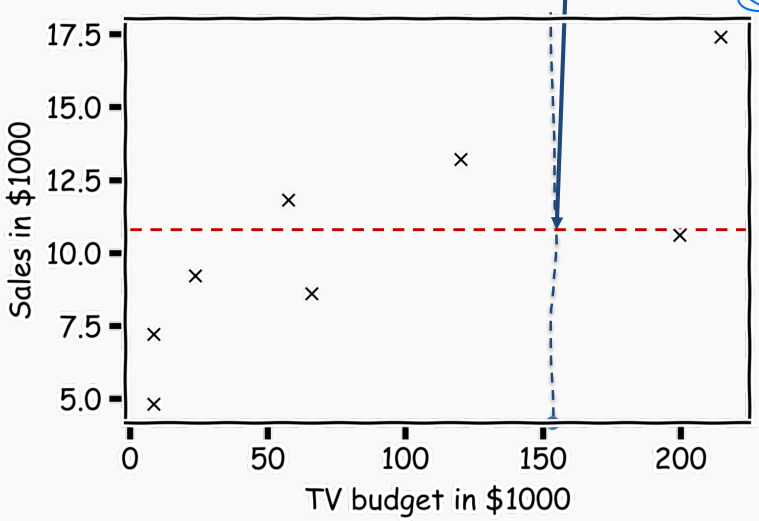
Statistical Model

How do we find $\hat{f}(x)$?



Statistical Model

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



Prediction vs. Estimation

For some problems, what's important is obtaining [the mathematical form of] $\hat{f}$, our estimate of f. These are called ***inference*** problems.

When we use a set of measurements, $(x_{i,1}, \ldots, x_{i,p})$ to predict a value for the response variable, we denote the ***predicted*** value by:

$$\hat{y}_i = \hat{f}(x_{i,1}, \ldots, x_{i,p}).$$

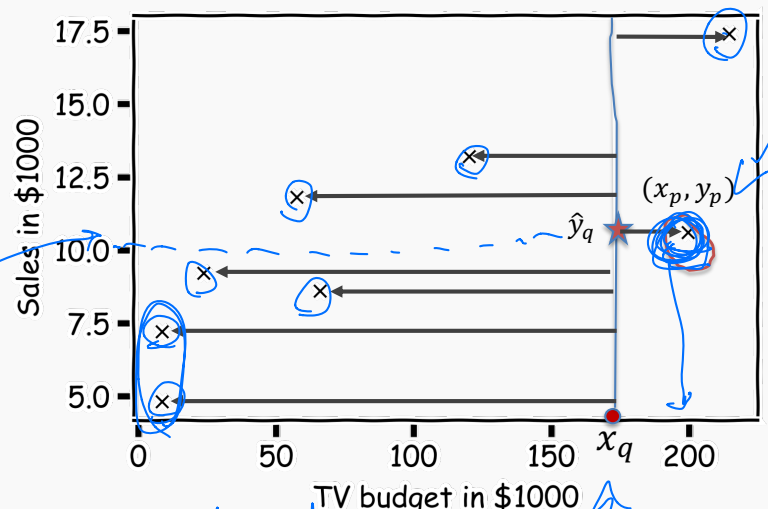
accuracy of how $\hat{y}$ and y relate.

For some problems, we don't care about the specific form of $\hat{f}$, we just want to make our predictions $\hat{y}$'s as close to the observed values y's as possible. These are called ***prediction problems.***

chooses the single closest/nearest neighbor in the X-space.

(k=1)-NN

Simple Prediction Model: use most similar observed observations



most similar observation to $x_q = 170$

What is $\hat{y}_q$ at some x_q ?

Find distances to all other points $D(x_q, x_i)$

Find the single nearest neighbor, (x_p, y_p)

Predict $\hat{y}_q = y_p$

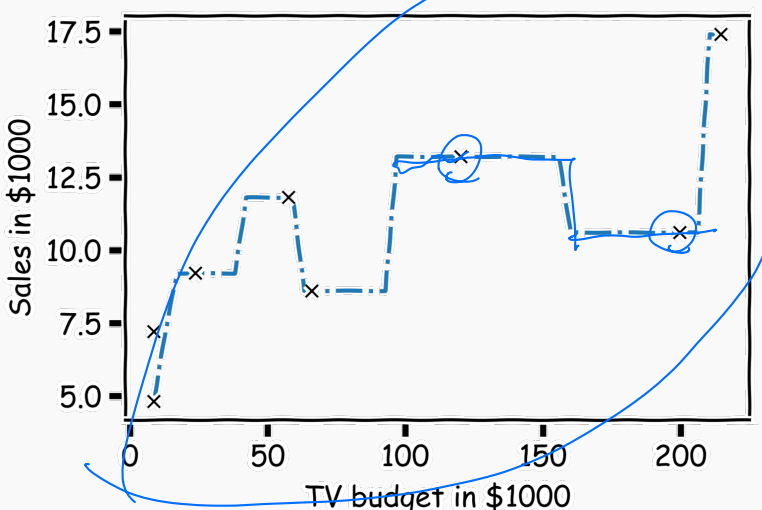
$\hat{y} = 10.5$

all values of X: find the closest neighbor

$x_q = 170$

Simple Prediction Model

Do the same for "all" $x's$



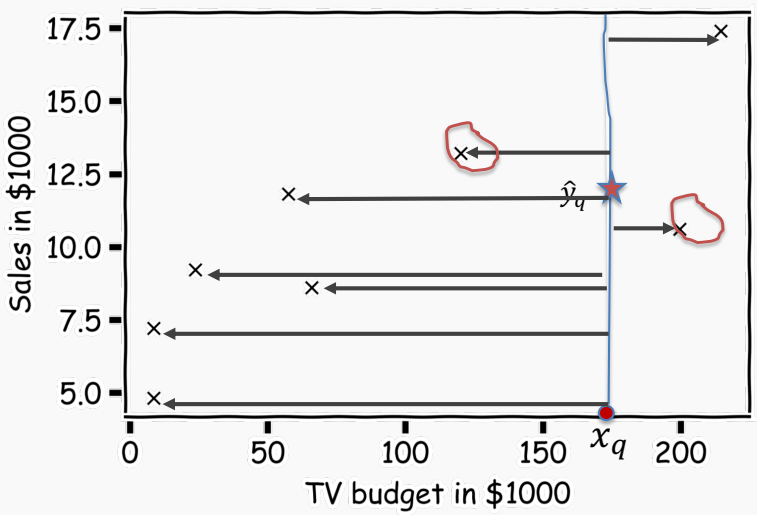
step function: stepping at the midpoint between any 2 consecutive observations in the x-direction

$\hat{f}$ = step function

Extend the Prediction Model

$k = 2, 3,$ etc...

↳ average the responses of the k-nearest neighbors



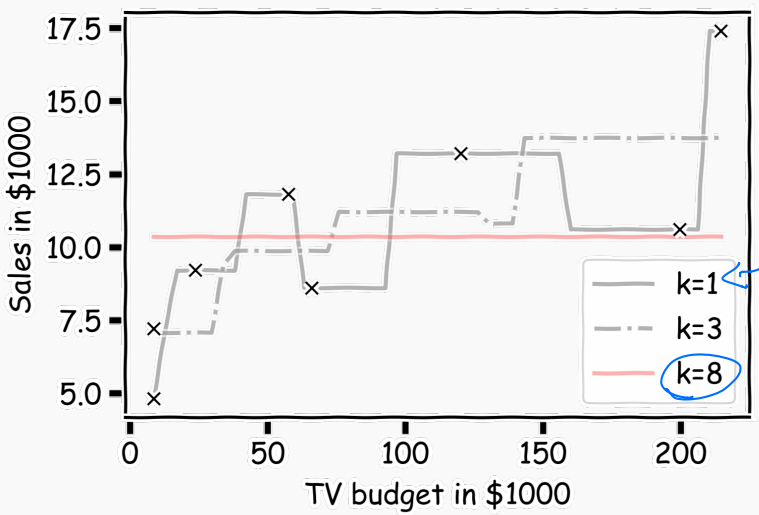
What is $\hat{y}_q$ at some x_q ?

Find distances to all other points $D(x_q, x_i)$

Find the *k*-nearest neighbors, $x_{q_1}, \ldots, x_{q_k}$

Predict $\hat{y}_q = \frac{1}{k}\sum_i^k y_{q_i}$

Simple Prediction Models

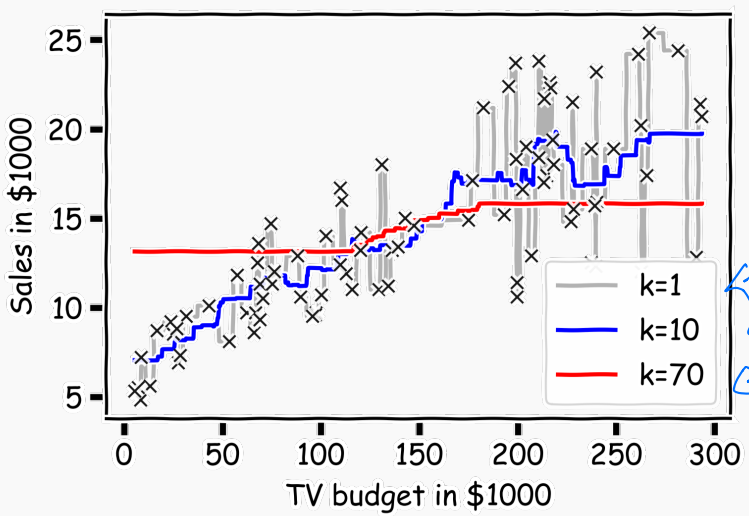


"complex" step function

↓ the function is "simpler"

Simple Prediction Models

We can try different *k*-models on more data



n = 100

k is too small

k is in the "Goldilocks Zone"

k is too big

*Note: the *k*-NN model for each of the 3 different values of *k* here can be thought of as 3 different models all within the class of '*k*-NN models'.

k-Nearest Neighbors (regression)

The ***k*-Nearest Neighbor (*k*-NN) model** is an intuitive way to predict a quantitative response variable:

> *to predict a response for a set of observed predictor values, we use the responses of other observations most similar to it*

Note: this strategy can also be applied in classification to predict a categorical variable. We will encounter *k*-NN again later in the course in the context of classification.

k-Nearest Neighbors – *k*-NN

For a fixed a value of *k*, the predicted response for the i-th observation is the average of the observed response of the *k*-closest observations:

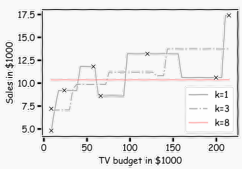
$$\hat{y}_n = \frac{1}{k} \sum_{i=1}^{k} y_{n_i}$$

observations in the neighborhood.

where $\{x_{n1}, \ldots, x_{nk}\}$ are the *k* observations most similar to x_i (*similar* refers to a notion of distance between 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:

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

intercept is β_0

slope β_1

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

Statistical Modeling

- k-Nearest Neighbors (kNN)
- Simple Linear Regression

Model Fitness

- How does the model perform at predicting?

Comparison of Two Models

- How do we choose from two different models?

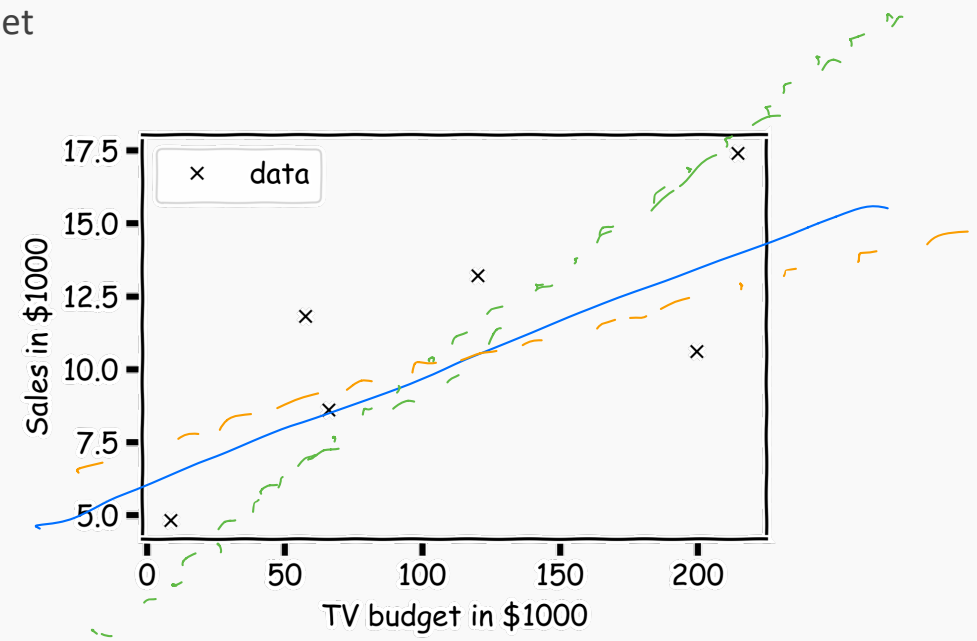
Probabilistic Basis of Regression

Hypothesis Testing and Confidence Intervals

Bootstrap Resampling

Estimate of the regression coefficients

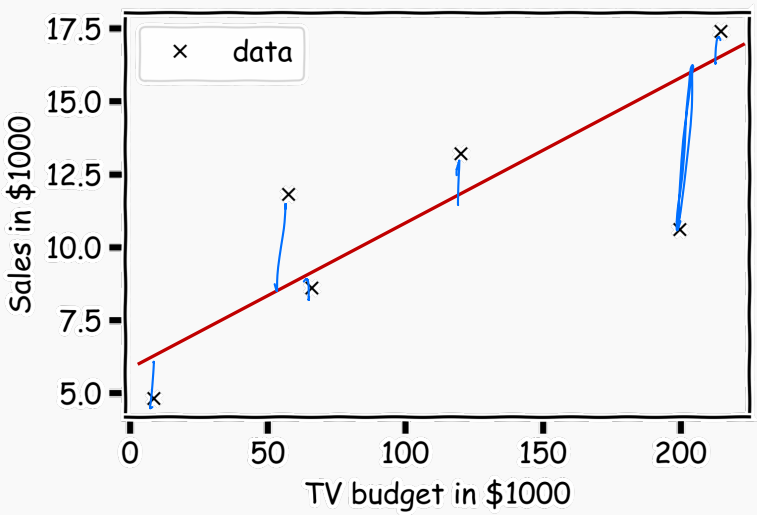
For a given data set



Estimate of the regression coefficients (cont)

Is this line good?

Y is "more important"

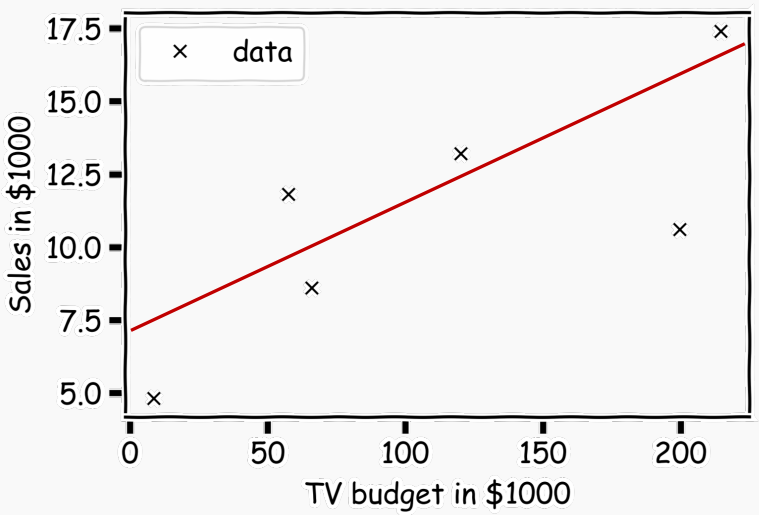


} vertical distances

↑ residuals around the model (line)

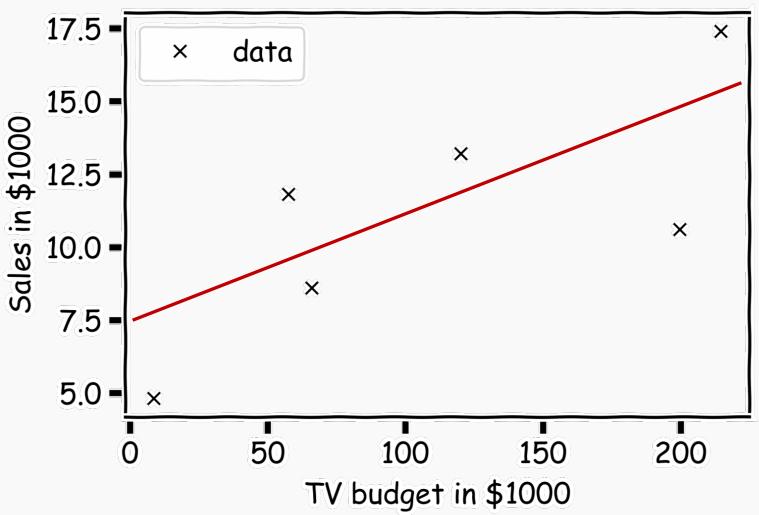
Estimate of the regression coefficients (cont)

Maybe this one?



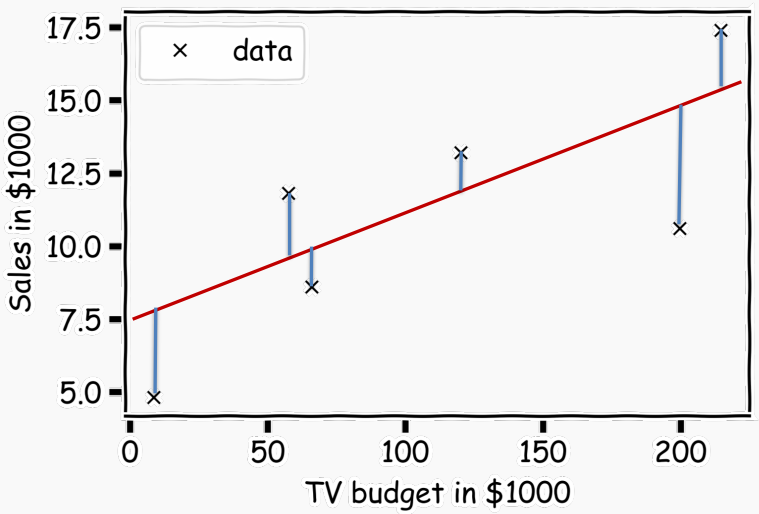
Estimate of the regression coefficients (cont)

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**,

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

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

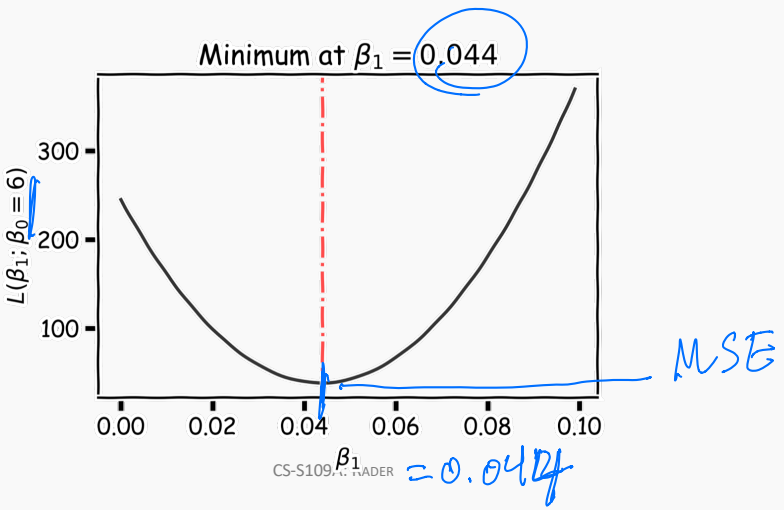
Then the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ should be:

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

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 $\hat{\beta}_1$:

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

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

"OLS" estimates of the linear regression parameters (coefficients)

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

The line:

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

is called the **regression line**.

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

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

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

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

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

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

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

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

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

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

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

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

Interpretation of Coefficients

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 TV advertising expenditures were 0, the estimated Sales would be 7.5 thousand dollars.

- Every 1 unit (thousand dollars) increase in TV advertising expenditures is associated with a 40 dollar increase in Sales on average.

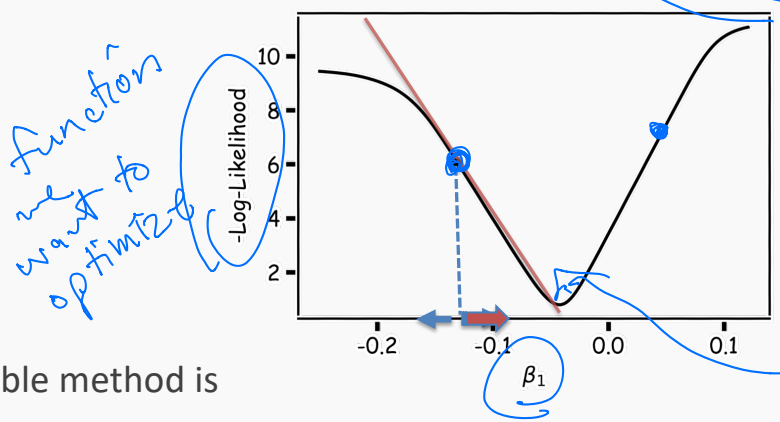
If we increase the TV by \$1000, what would you expect the increase in sales to be?

What if the regression equation was:

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

β_1 measures the magnitude and direction of association

Estimate of the regression coefficients: gradient descent (a sidebar)



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:

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

Learning Rate

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

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

Change to more conventional notation:

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

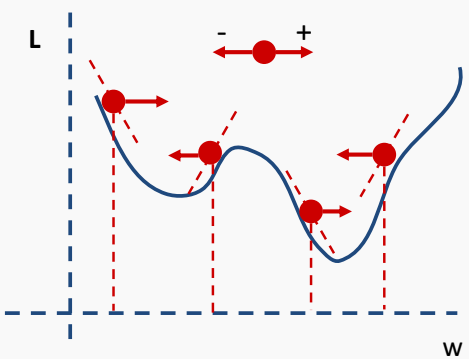
← we'll come back to this in the future (gradient descent in N.N.'s framework)

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.

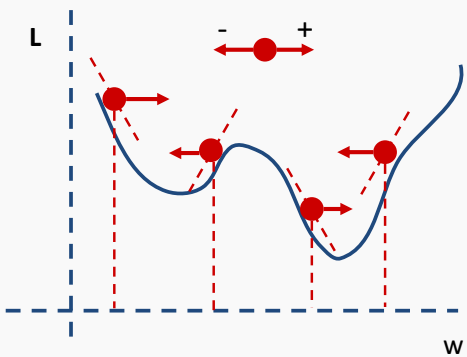
$$w^{(i+1)} = w^{(i)} - \lambda \frac{d\mathcal{L}}{dw}$$



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 is just one single minimum because the loss function is convex.
- In linear regression the best coefficient estimates have a closed form solution, so no need for gradient descent.
- If we were to change our loss function (say to minimize mean absolute error instead of mean square error), then gradient descent would be a very useful approach.
- We will talk about optimization again in future lectures (in the context of Neural Networks, for example).

Things we will Consider

Model Fitness

How does the model perform predicting? How should we measure this?

Comparison of Two Models

How do we choose from two different models? How can we prevent overfitting?

Evaluating Significance of Predictors

Does the outcome truly depend on the predictors? What is a plausible range of values for the true β?

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

What is a plausible range of values for the true $\hat{f}$ at a particular (based on our $\hat{f}(x^*)$)?

Lecture Outline

Statistical Modeling

- *k*-Nearest Neighbors (*k*-NN)
- Simple Linear Regression

Model Fitness

- How does the model perform at predicting?

Comparison of Two Models

- How do we choose from two different models?

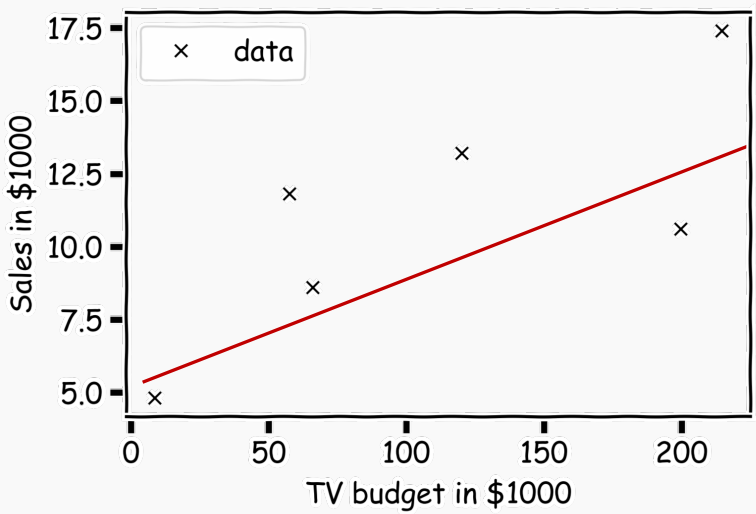
Probabilistic Basis of Regression

Hypothesis Testing and Confidence Intervals

Bootstrap Resampling

Error Evaluation

How can a model be evaluated for accuracy? What should we measure?



Some measurement or summary of the residuals (errors) from the model is a good choice…

Error Evaluation: *MSE*

In order to quantify how well a model performs, we define a ***loss*** or ***error function.***

A common loss function for quantitative outcomes is the **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The quantity $y_i - \hat{y}_i$ is called a ***residual*** and measures the error at the i^{th} prediction.

Note: this is just dividing the loss function we saw before in regression by the sample size, *n*.

Error Evaluation: *RMSE*

Caution: The MSE is by no means the only valid (or the best) loss function!

Question: What would be an intuitive loss function for predicting categorical outcomes?

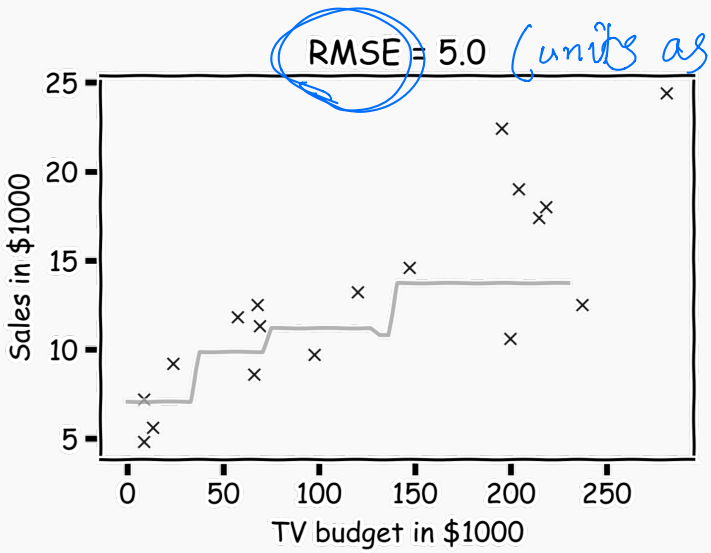
Note: The square **R**oot of the **M**ean of the **S**quared **E**rrors (RMSE) is also commonly used.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

square root of MSE

Model fitness

For a subset of the data, calculate the RMSE for *k=3.* Is RMSE=5.0 good enough?

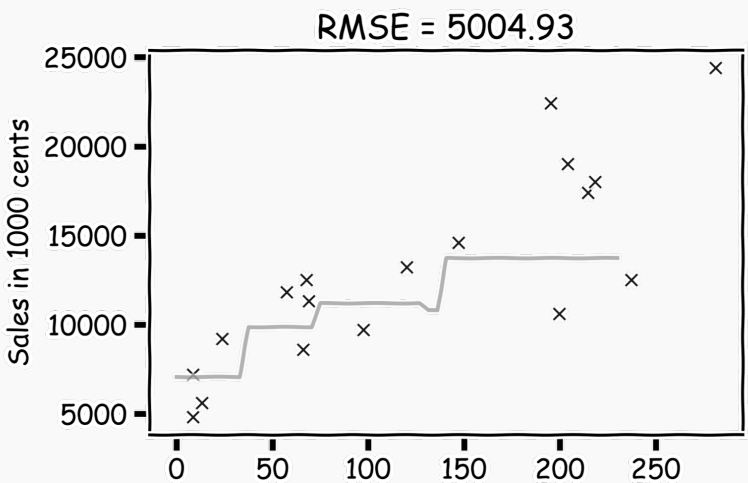


(units as "y")

"average" prediction error of the model

Model fitness

What if we measure the Sales in cents instead of dollars?

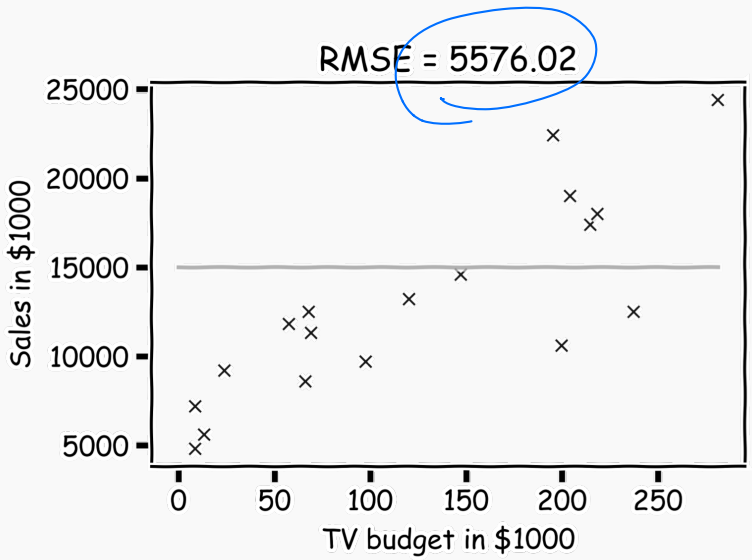


RMSE is now 5004.93.
Is the model now worse?

Still just as good.

Model fitness

It is better if we compare it to something.



We will use the simplest model:

$$\hat{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

Base case model:

Error Evaluation: R^2

- Note: MSE and RMSE are dependent on the units of the response variable: which is both a good and a bad thing.
- R^2 is an alternative metric: it gets rid of units and measures **the percentage of variability in the response that is associated with the predictor(s) based on the model**. This is in reference to the baseline of simply using the sample mean to do the predictions (aka, not using the predictors).

$$R^2 = 1 - \left[\frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}\right]$$

improvement in ratio of squared error of our model to the baseline model.

R^2: some details

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

- If our model is as good as the mean value, $\bar{y}$, for predicting the observed responses, $y_i's$,then $R^2 = 0$.
- If our model is perfect at predicting the $y_i's$, then $R^2 = 1$.
- R^2 can be negative if the model is worse than using $\bar{y}$. When can this happen?

↳ if our model is very wrong (think wrong direction)

or if R^2 is evaluated out-of-sample

Lecture Outline

Statistical Modeling

- *k*-Nearest Neighbors (*k*-NN)
- Simple Linear Regression

Model Fitness

- How does the model perform at predicting?

Comparison of Two Models

- How do we choose from two different models?

Probabilistic Basis of Regression

Hypothesis Testing and Confidence Intervals

Bootstrap Resampling

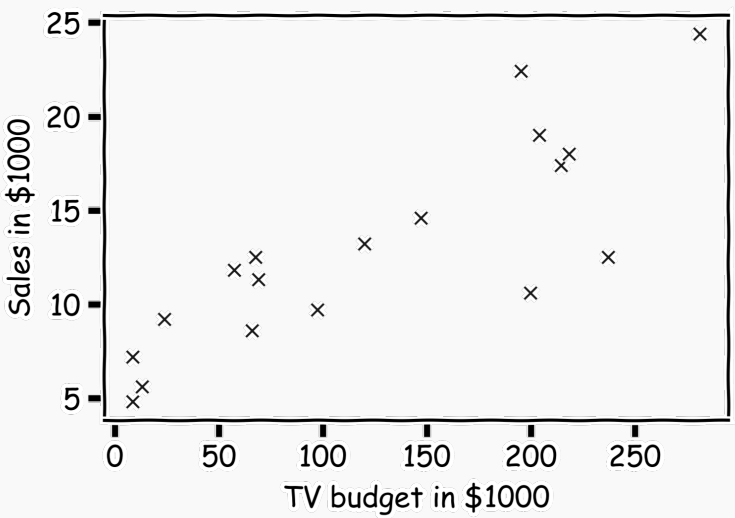
An introduction to overfitting

- The metrics of model fitness/accuracy we've seen (*SSE*, *MSE*, *RMSE*, and R^2) are all prone to the same issue: overfitting.
- Overfitting can occur since a more complex model (like $k = 1$ in k-NN) will *always(*?) lead to more accurate results when evaluated on the same set of data that the model was fit (estimated) to.
- This issue of overfitting can be cured (improved at least) by using a different set of data then what was used when fitting the model.
- This is the basis of creating train-test splits: use the train set to fit/estimate the model, and use the test set to evaluate the model's accuracy.

out-of-sample prediction

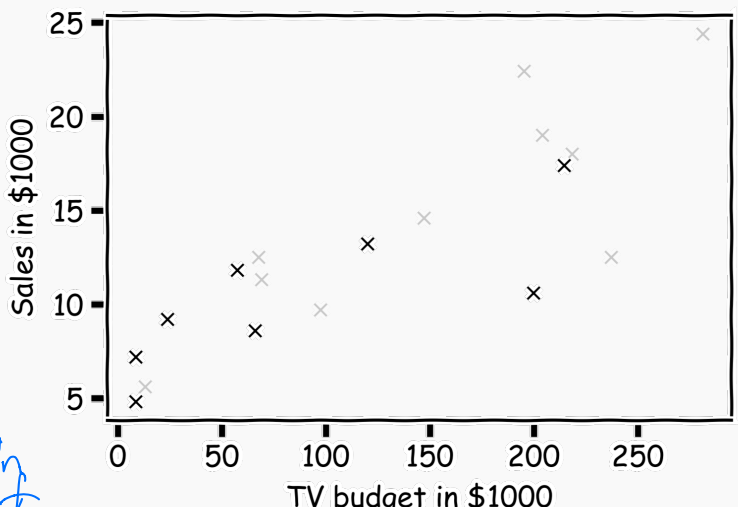
Error Evaluation

Start with some data.



Error Evaluation

Hide some of the data from the model. This is called **train-test** split.



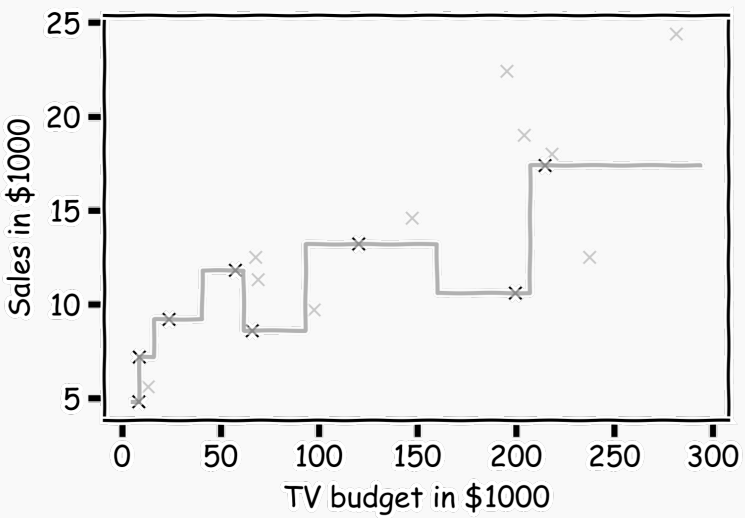
black: train set

grey: test set

We use the train set to estimate $\hat{f}$, and the test set to evaluate the model (using *MSE*, R^2, etc.).

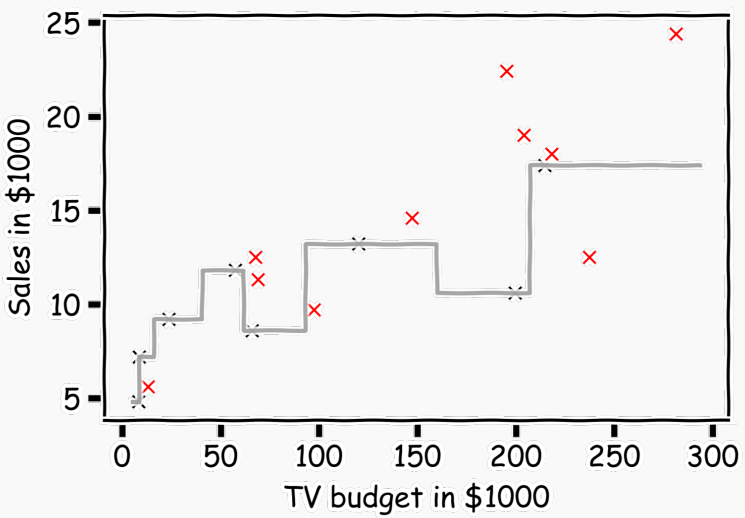
Error Evaluation

Estimate $\hat{y}$ for *k=1* .



Error Evaluation

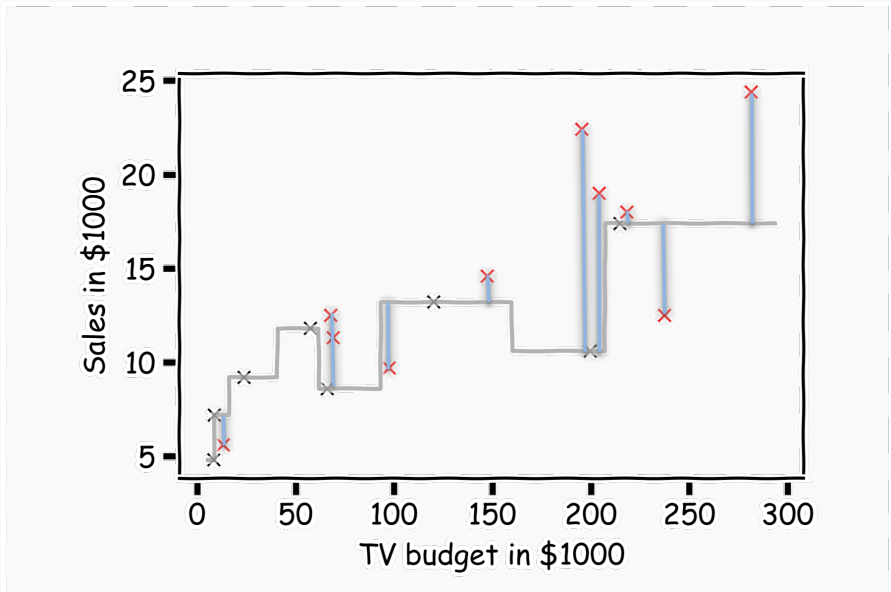
Now, we look at the data we have not used, the **test data** (red crosses).



Error Evaluation

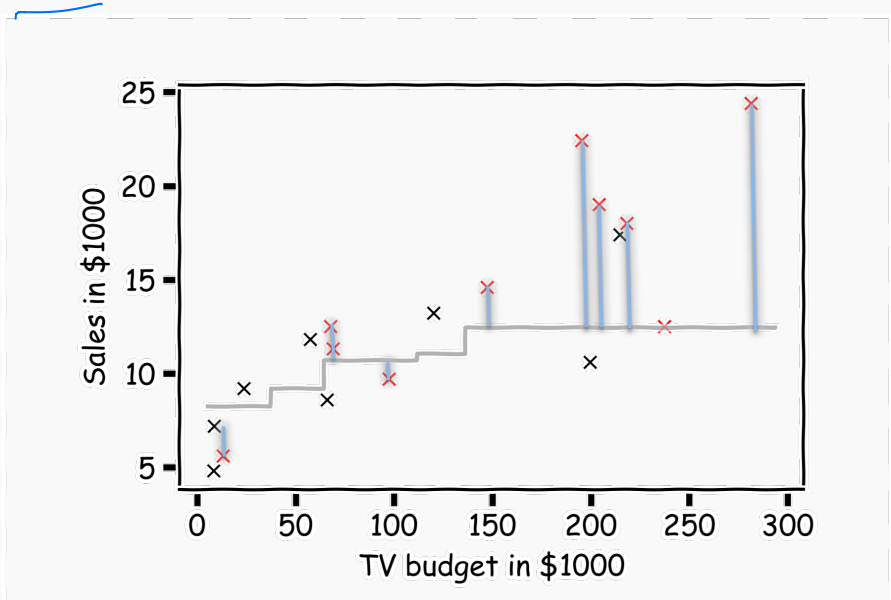
of the test set (out-of-sample)

Calculate the **residuals** $(y_i - \hat{y}_i)$.



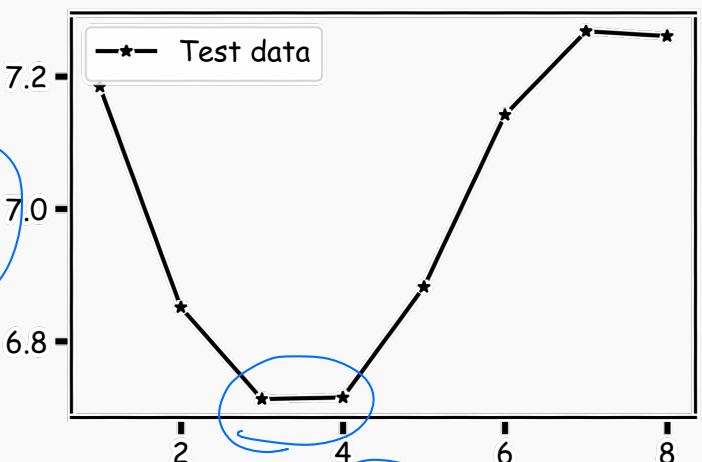
Error Evaluation

Do the same for *k=3*.



Model Comparison

Do the same for all *k*'s and compare the *RMSE*s (or *MSE*s or R^2). *k = 3 or* 4 seem to be the best model. Which should we choose?



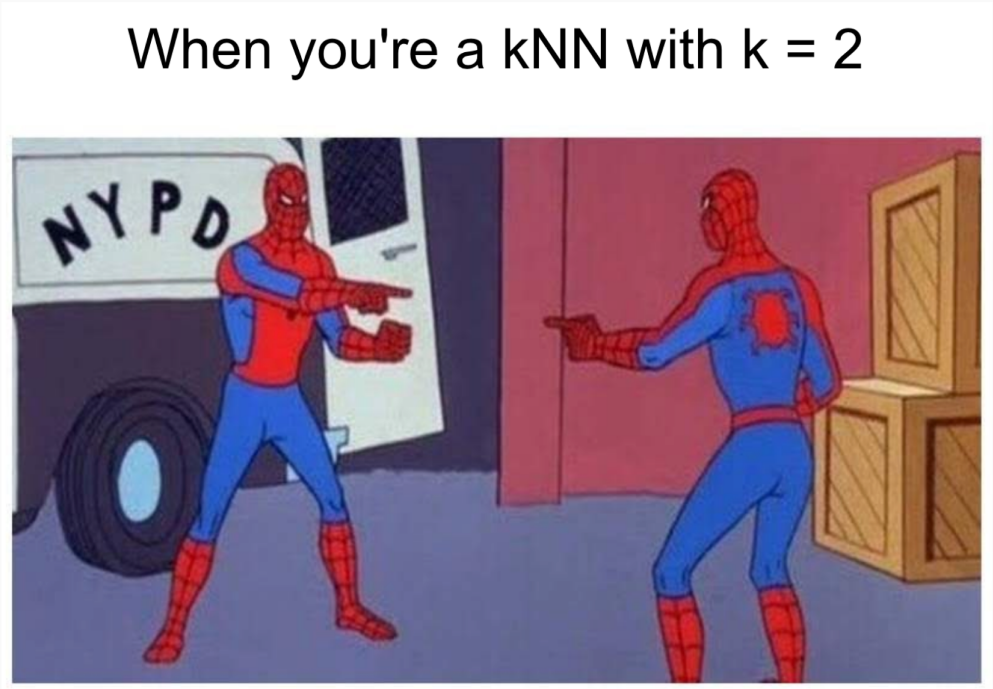
k=4 is a "simpler" or "smoother" model so that's better than k=3

- minimize RMSE or MSE in the testset
- maximize R^2

R^2 or MSE

too complex ← → too simple

What!?!?



Lecture Outline

Statistical Modeling

- *k*-Nearest Neighbors (*k*-NN)
- Simple Linear Regression

Model Fitness

- How does the model perform at predicting?

Comparison of Two Models

- How do we choose from two different models?

Probabilistic Basis of Regression

Hypothesis Testing and Confidence Intervals

Bootstrap Resampling

Uncertainty of our coefficient estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$

We interpret the ε term in our observation

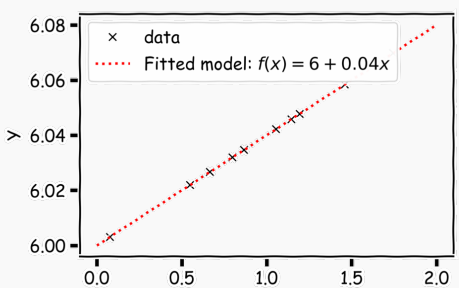
$$y = f(x) + \epsilon$$

signal

noise

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?).



Uncertainty of our coefficient estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ (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 $\hat{\beta}'s$.

The Linear Regression Model

Recall that the linear regression model can be written as (for each individual observation):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

To simplify the mathematics and uncertainty to perform inferences, a probabilistic model is often placed on the irreducible error (true residuals) term, ε_i:

$$\varepsilon_i \sim N(0, \sigma^2)$$

This can be listed as 4 assumptions based on the true residuals:

1. Residuals are normally distributed
2. Residuals have mean zero no matter the value of the x_i (sometimes the linearity assumption)
3. Variance of the residuals is constant no matter the value of the x_i
4. Residuals are independent

Consequences of the Probabilistic Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

If this model is correct (or at least reasonable), then the sampling distribution (aka, the distribution of values we would expect for the estimated β_0 and β_1 whenever a sample of data is taken) have well-known distributions:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)\right)$$

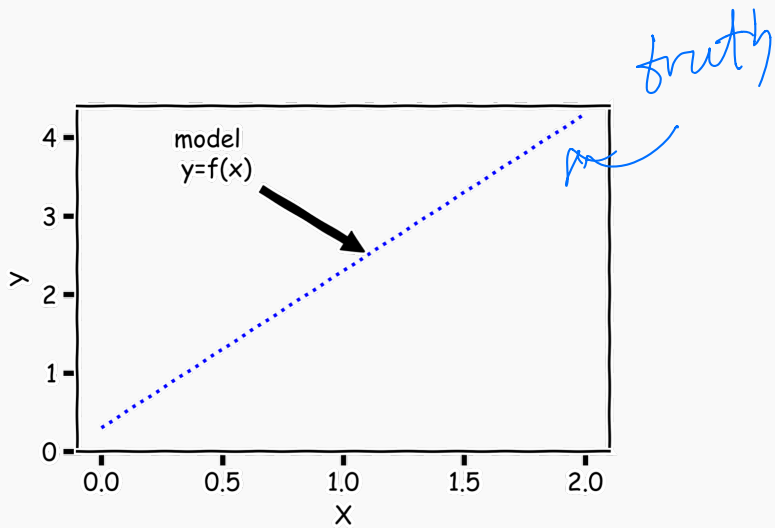
$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

slope's sampling distribution

We will see that these formula will generalize when we get to multiple linear regression (multiple predictors), we just need to use linear algebra ☺

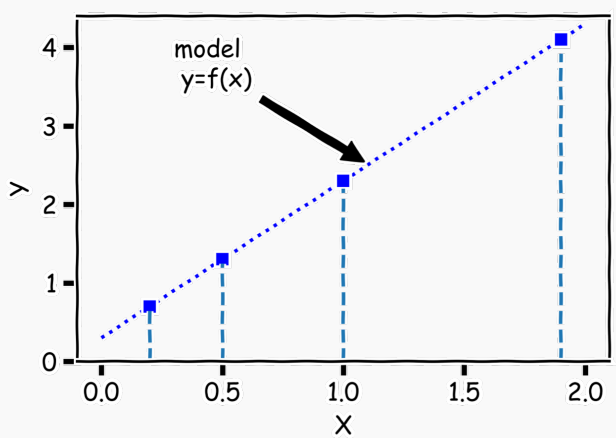
Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

Start with a model



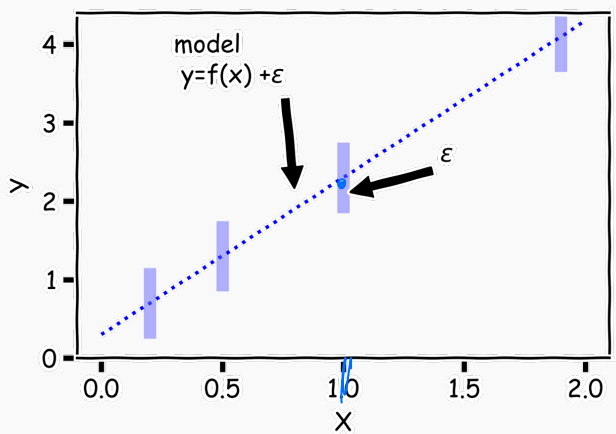
Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

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



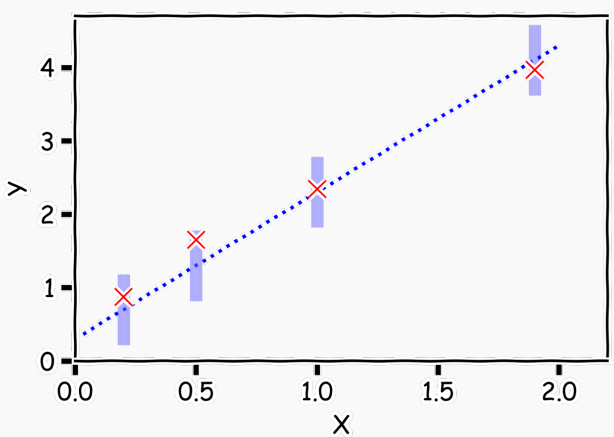
Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

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



Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

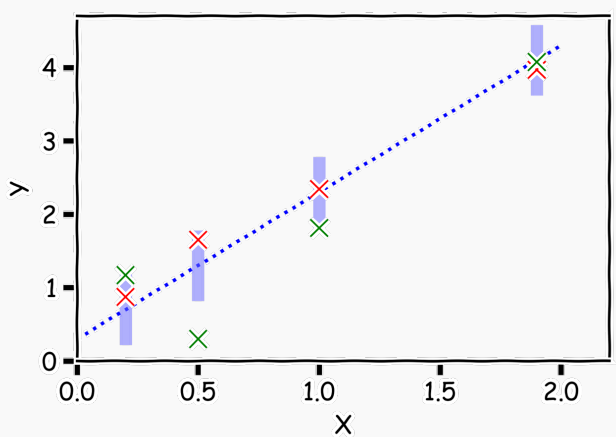
One set of observations, "one realization" we obtain one set of *Y*s (red crosses).



red x's are many sample data points

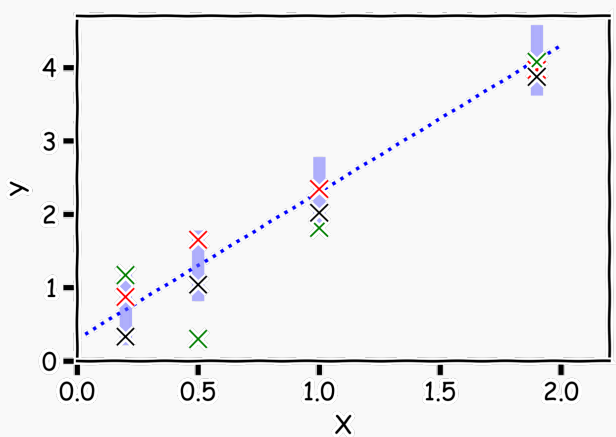
Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

Another set of observations, "another realization" we obtain another set of *Y*s (green crosses).



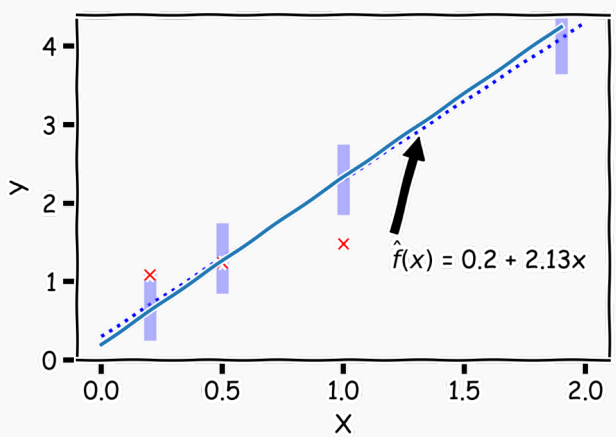
Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

Another set of observations, "another realization" we obtain another set of *Y*s (black crosses).



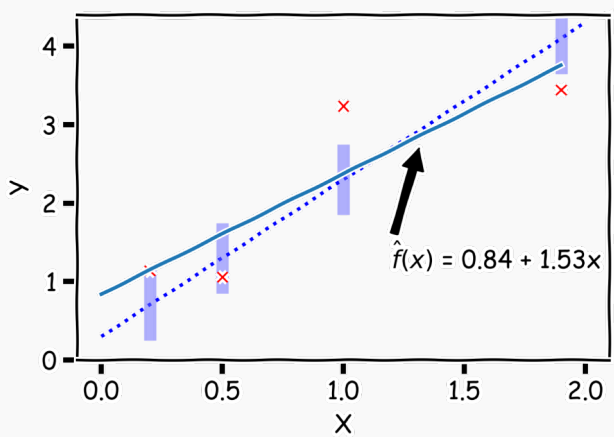
Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

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



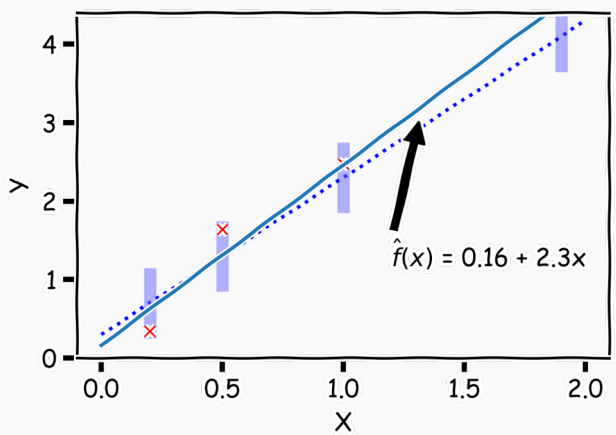
Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

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



Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (cont.)

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



Building the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (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.

Lecture Outline

Statistical Modeling

k-Nearest Neighbors (*k*-NN)

Simple Linear Regression

Model Fitness

How does the model perform at predicting?

Comparison of Two Models

How do we choose from two different models?

Probabilistic Basis of Regression

Hypothesis Testing and Confidence Intervals

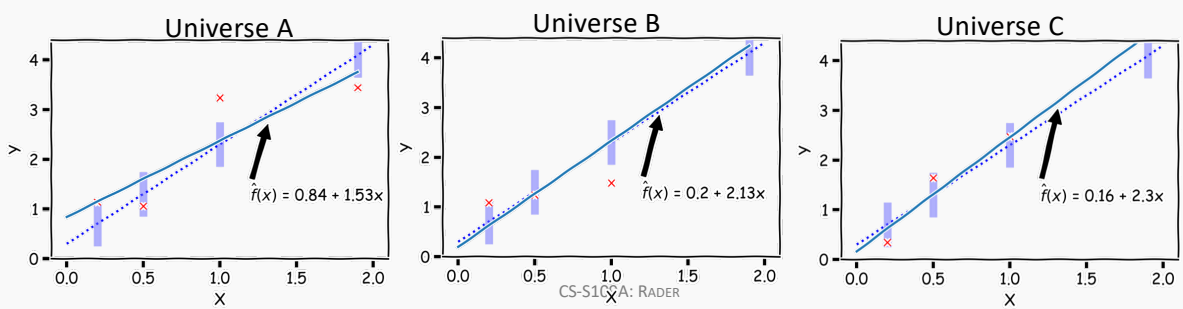
Bootstrap Resampling

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.

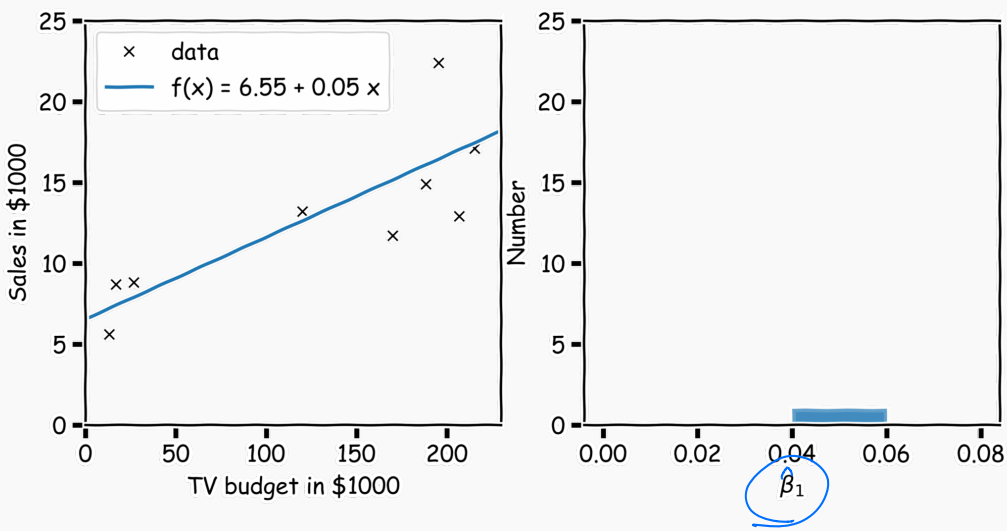
Question: If this is just one realization of the reality how do we know the truth? How do we deal with this conundrum?

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



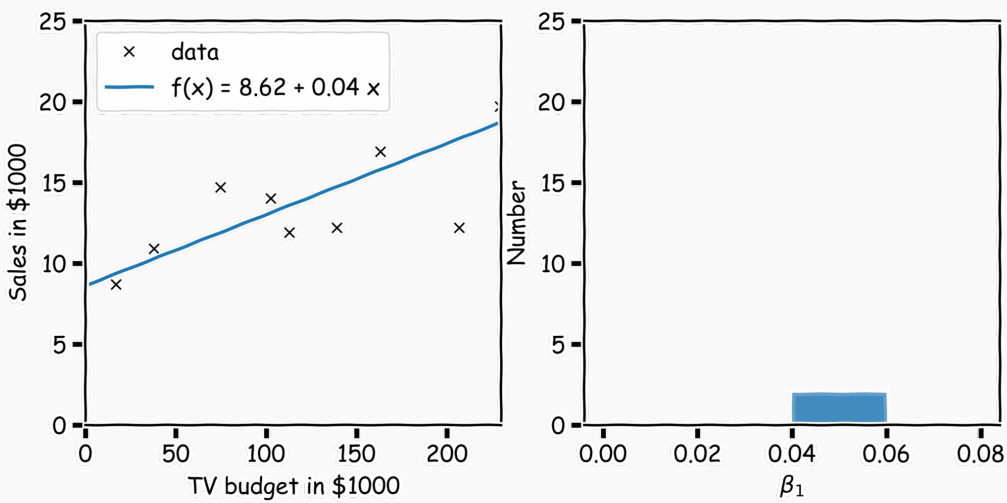
Confidence intervals for the predictors estimates (cont)

In our magical realisms, we can now sample multiple times



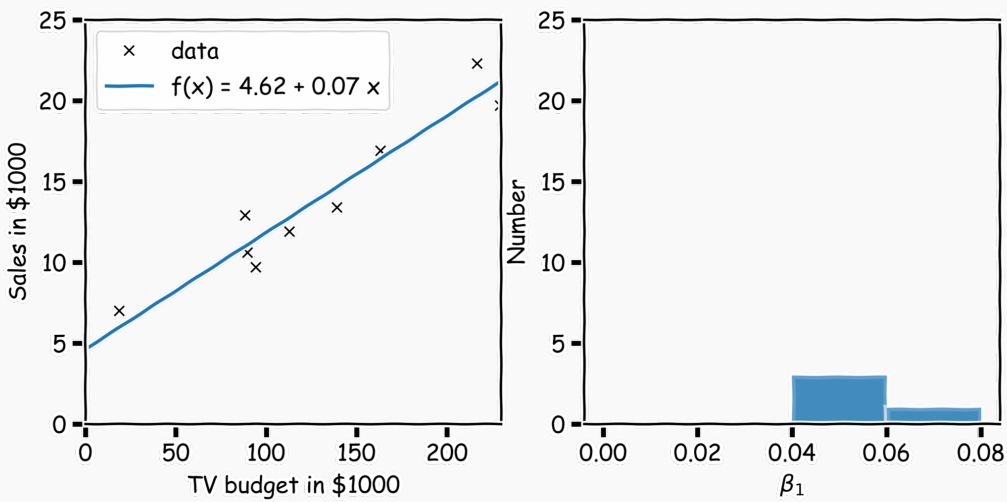
Confidence intervals for the predictors estimates (cont)

Another sample



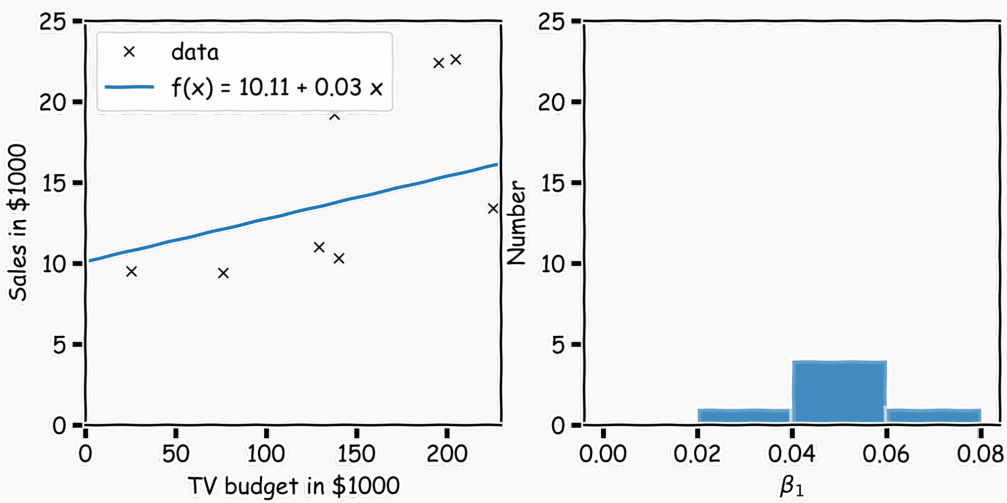
Confidence intervals for the predictors estimates (cont)

Another sample



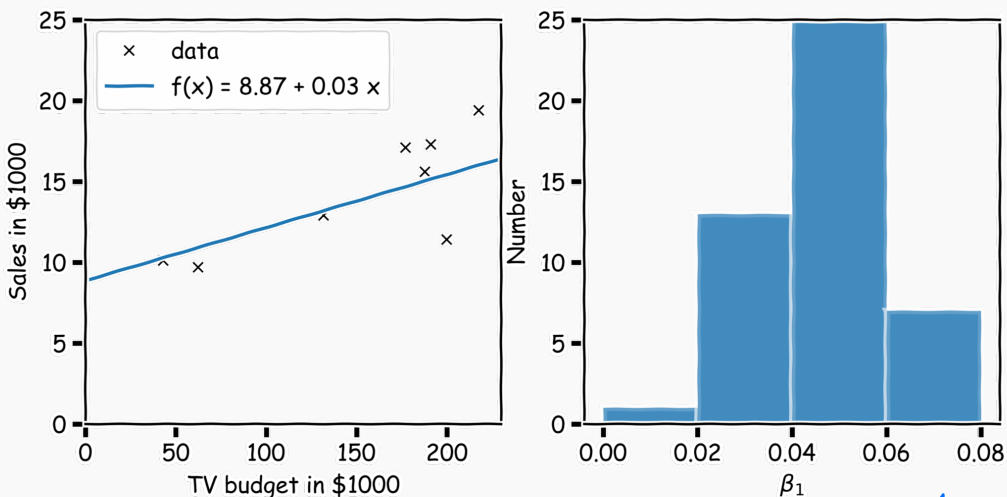
Confidence intervals for the predictors estimates (cont)

And another sample



Confidence intervals for the predictors estimates (cont)

Repeat this for 100 times (or better yet: 1,000 times)



theory says this should be normally distributed

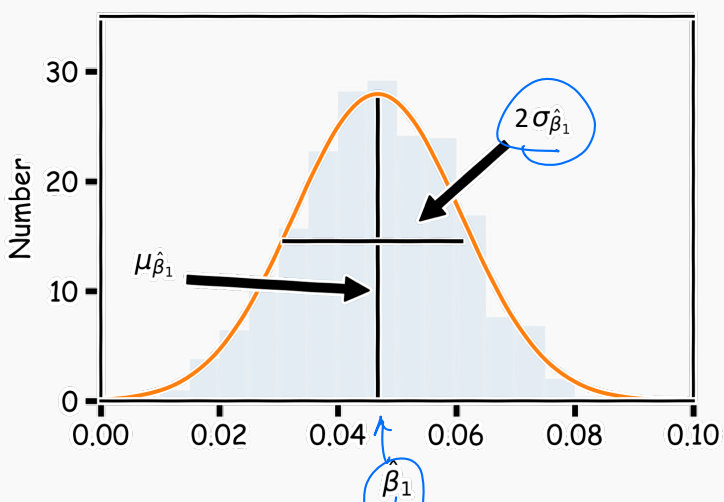
sampling distribution of $\hat{\beta}_1$

$\hat{\beta}_1 \pm$ margin of error bounds

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***, $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$.

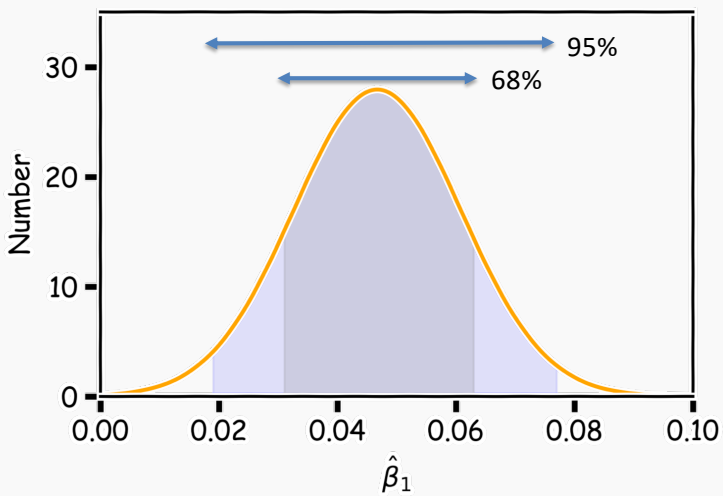


95% C.I,

$\hat{\beta}_1 \pm 2 * SE(\hat{\beta}_1)$

Confidence intervals for the predictors estimates (cont)

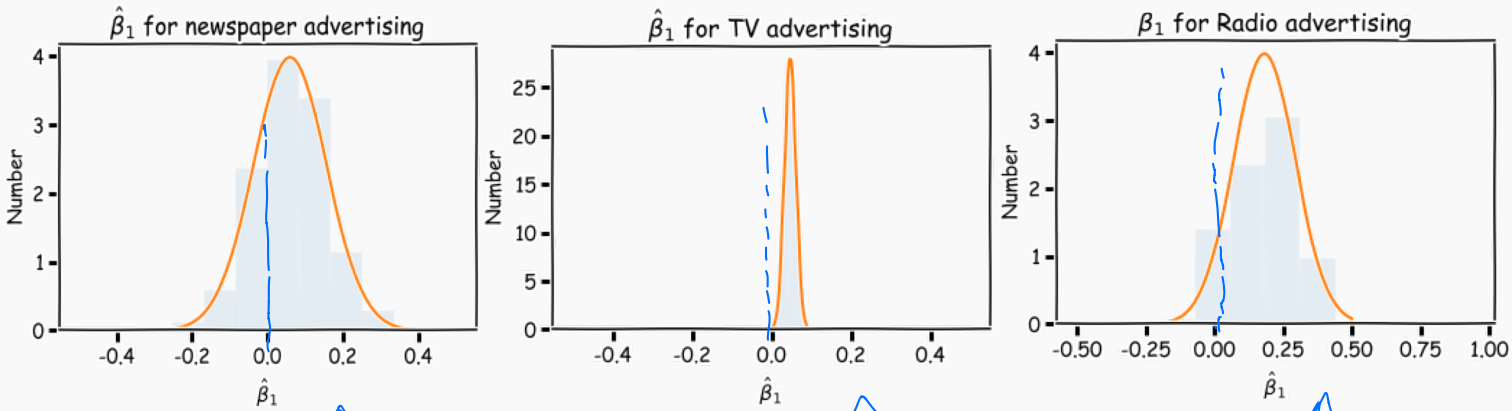
Finally we can calculate the confidence intervals, which are the ranges of values such that the **true** value of β_1 is contained in this interval with *n* percent probability.



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?

: what value of $\hat{\beta}_1$ indicates no association?

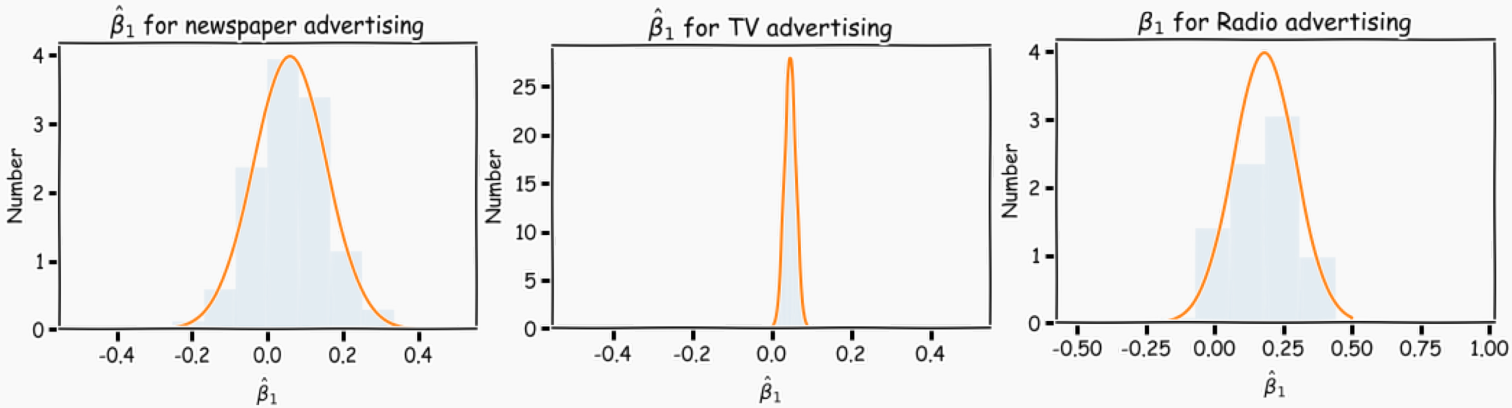


Before we answer this question, we need to answer another question.

3 separate regression models

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

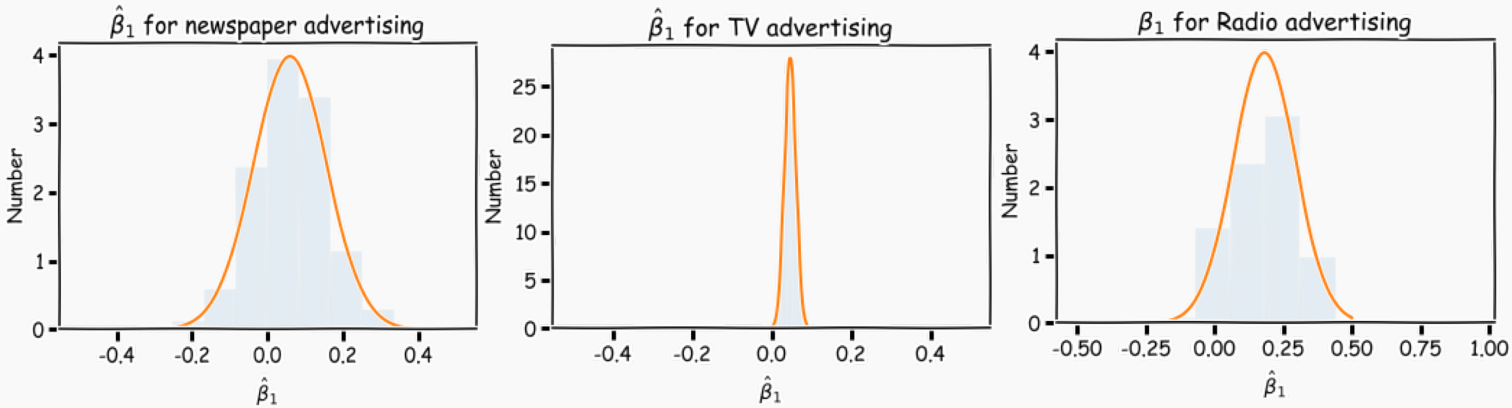
Question: Which ones are important?



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?'***

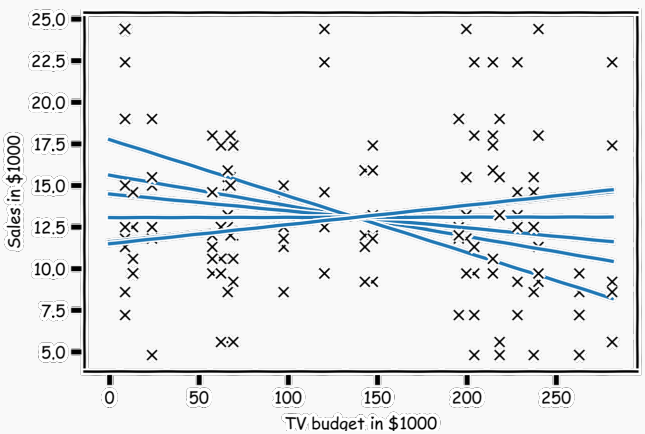
Hypothesis Testing

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

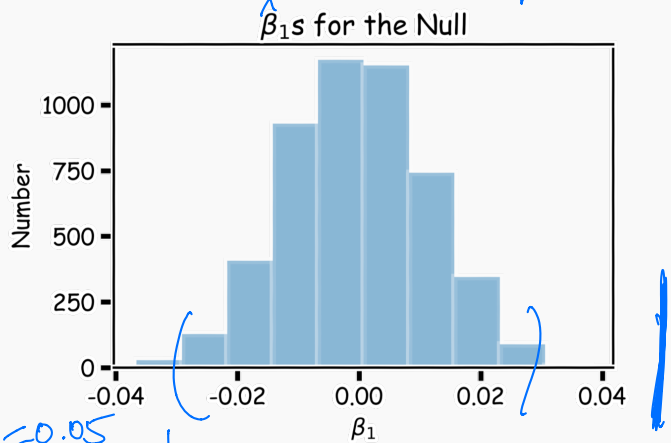


formal hypothesis test approach.

Conf. int approach.

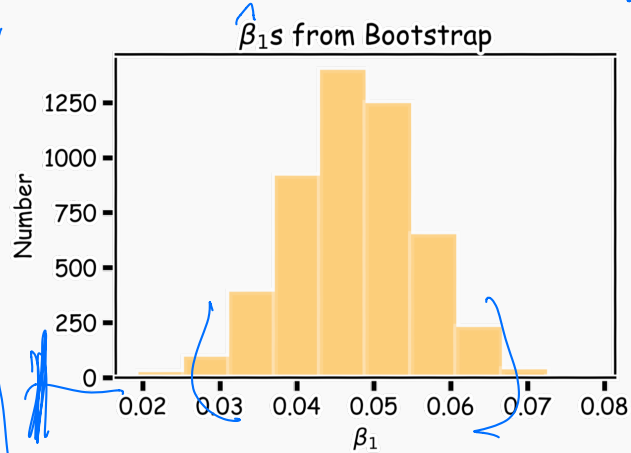
sampling distribution of $\hat{\beta}_1$'s if $\beta_1 = 0$

sampling distribution of $\hat{\beta}_1$'s if $\hat{\beta}_1$ is accurate



$\hat{\beta}_1 = 0.05$ close to zero

centered at 0



0 close to $\hat{\beta}_1$

centered at observed $\hat{\beta}_1$

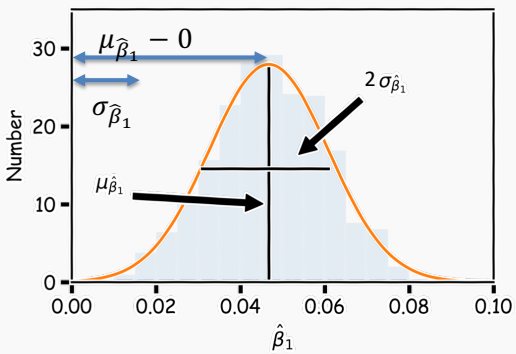
Importance of predictors

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

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

$$t = \frac{\mu_{\hat{\beta}_1}}{\hat{\sigma}_{\hat{\beta}_1}}$$

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.

p-value: probability of observing our result (for $\hat{\beta}_1$) or more extreme if $H_0: \beta_1 = 0$ is true.

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.

Lecture Outline

Statistical Modeling

k-Nearest Neighbors (*k*-NN)

Simple Linear Regression

Model Fitness

How does the model perform at predicting?

Comparison of Two Models

How do we choose from two different models?

Probabilistic Basis of Regression

Hypothesis Testing and Confidence Intervals

Bootstrap Resampling

{ were used to build inferences for β_1 - and we used probabilistic theory (Normal and t) to come to these conclusions

↖ we'll do next time.

{ alternative approach to probability theory

Bootstrap

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

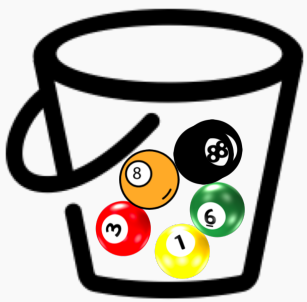
> Bootstrapping is the practice of sampling from the observed data (X, Y) in estimating statistical properties.

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.



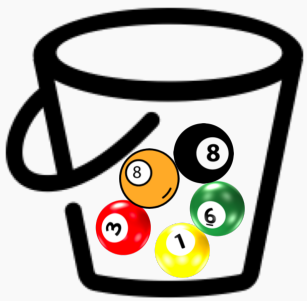
Bootstrap

Bootstrap

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

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

Bootstrap

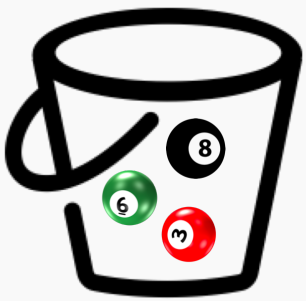


Bootstrap

We repeat this process.

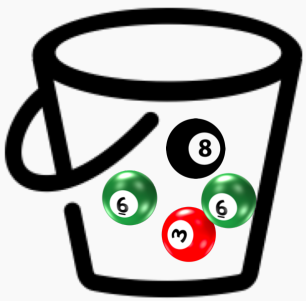
Bootstrap

Again



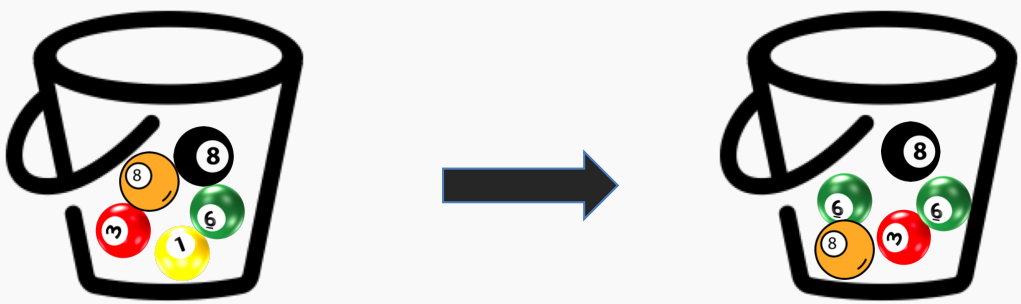
Bootstrap

And again



Bootstrap

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



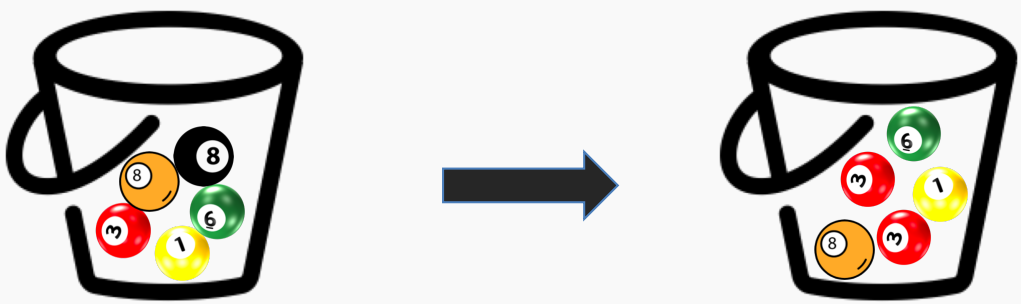
This new bucket represents a new parallel universe

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, resampling from the observed data.

For example, we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ multiple times by randomly resampling from our data set (the same sample size to maintain the same amount of uncertainty in each resample). 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

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

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

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):

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

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

*Note: these formulas do not rely on the normality assumption. But they do rely on independence, constant variance, and linearity.

Standard Errors

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

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

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:

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

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

$$SE\left(\widehat{\beta_0}\right) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$

$$SE\left(\widehat{\beta_1}\right) = \frac{\sigma}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

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

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

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

Does this make sense?

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 quasi-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

What's next?

Multiple Regression (aka, multiple predictors)

- Collinearity
- Categorical predictors
- Polynomial regression
- Piecewise Linear Regression
- Interaction terms
- Variable (predictor) selection