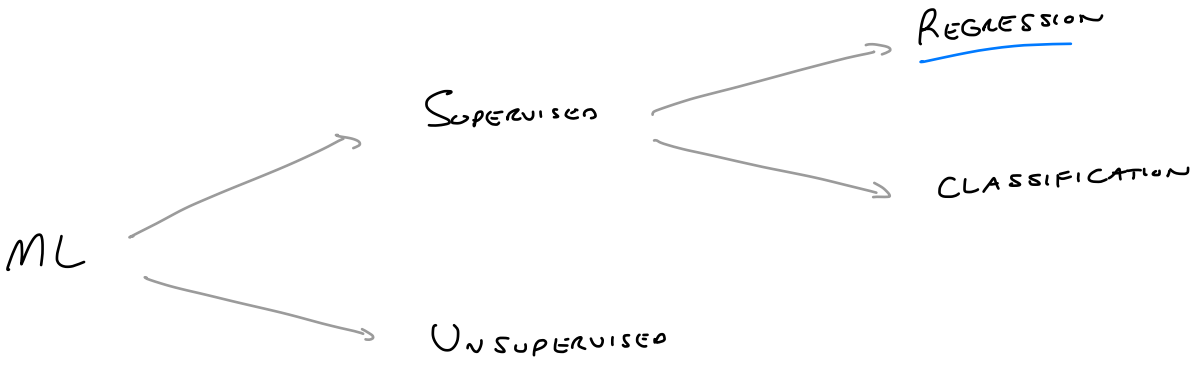
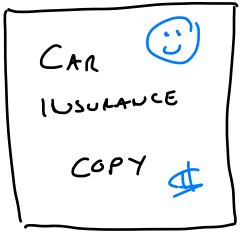
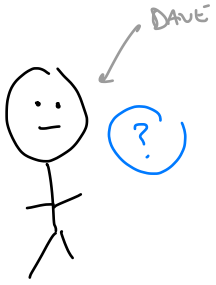
LINEAR REGRESSION

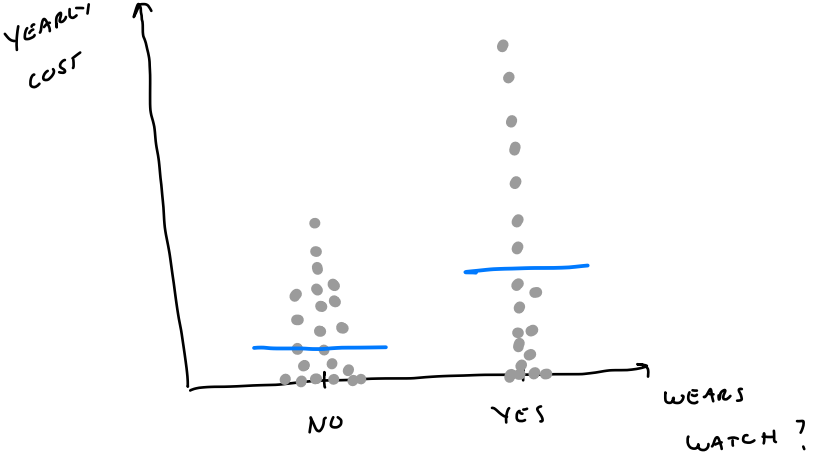
FOR PREDICTION

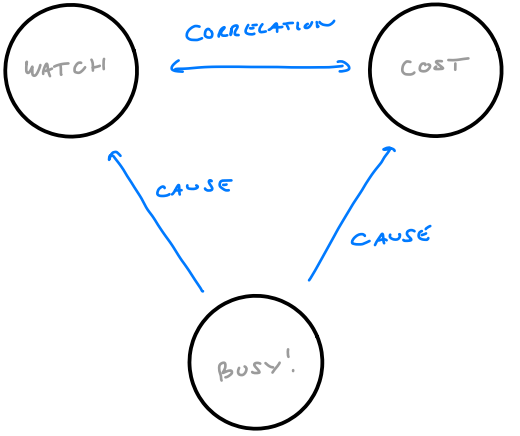


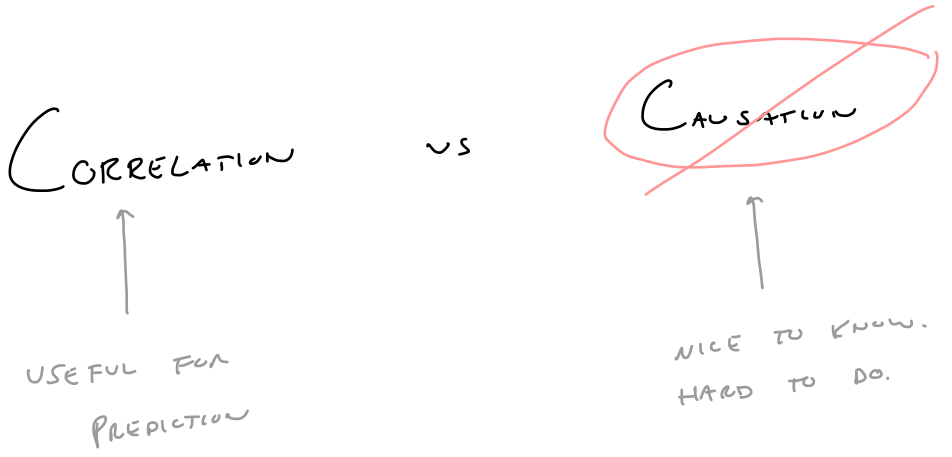
EXPLANATION VERSUS PREDICTION










CORRELATION
vs
CAUSATION
USEFUL FOR PREDICTION
NICE TO KNOW. HARD TO DO.

The Regression Task

"Data view"

Numeric

Response

Features

Grade	GPA	Major	Year
89	3.23	STAT	3
100	3.51	MATH	3
72	2.61	ECON	4
93	4.00	STAT	4
95	3.62	CS	2

New Student

GPA = 3.42

Major = MATH

Year = 4

Grade = ?

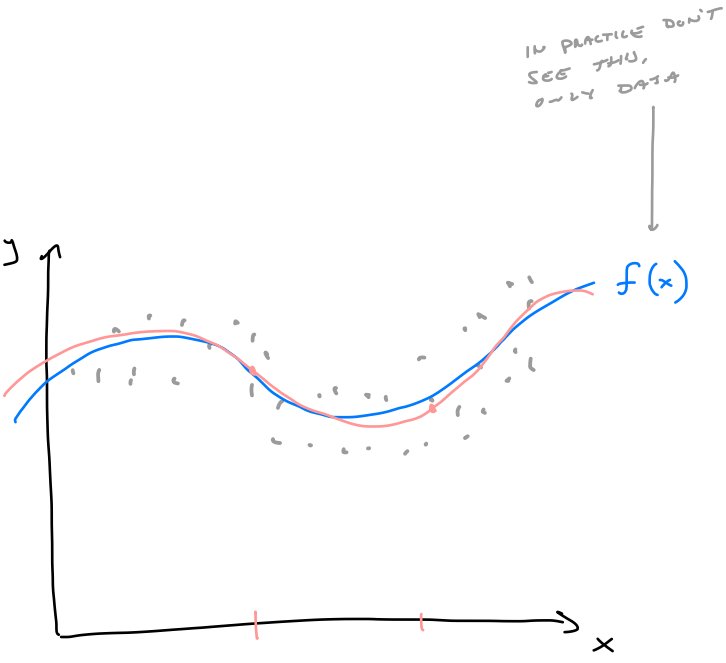
The Regression Task

"Math view"

$$Y = f(X) + \varepsilon$$

Want to learn this function

Response | Signal | Noise



Making Good Predictions

We want $f(x)$ close to Y.

In particular we want $(Y - f(x))^2$ to be small

This happens when

$$f(x) = \mathbb{E}[Y \mid X = x] = \mu(x)$$

The Regression Function

Linear Regression Models

Assume

LINEAR COMBINATION OF FEATURES

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

OR

CONDITIONAL MEAN

$$Y \mid X = x \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2)$$

THAT IS

REGRESSION FUNCTION →

$$\mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

How to "fit" linear models?

"train"

using data (x_i, y_i)

$i = 1, \ldots n$

"least squares"

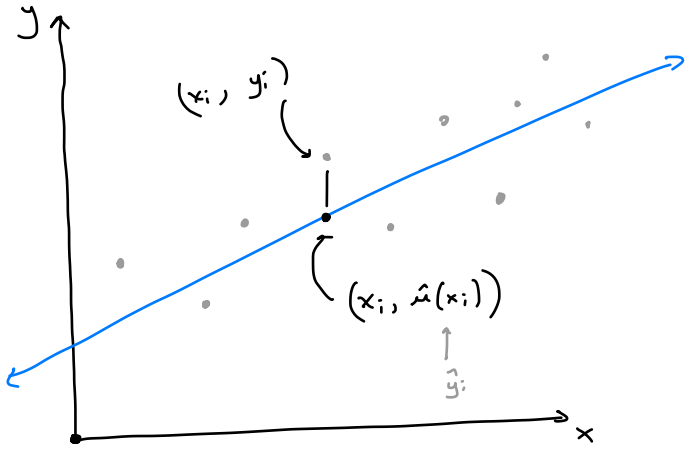
Find βs that minimize

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \right) \right)^2$$

→ call these $\hat{\beta}$s

Gives $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

as estimate of $\mu(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

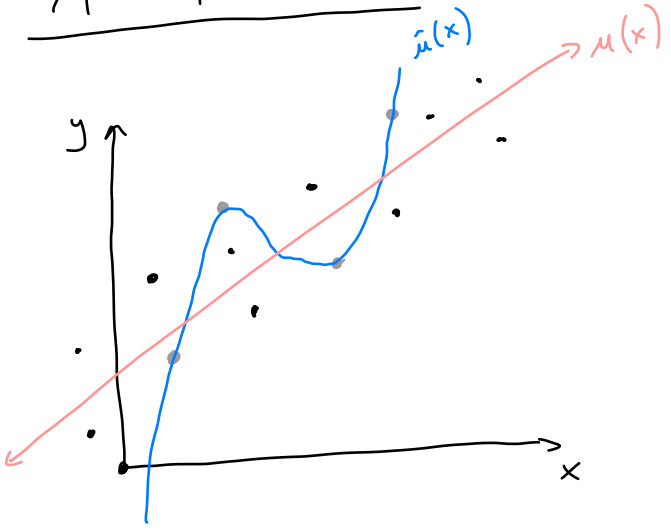


$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{m}(x_i)\right)^2}$$

RMSE

ROOT MEAN SQUARED ERROR

A PROBLEM



Assume

$$\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3^3$$

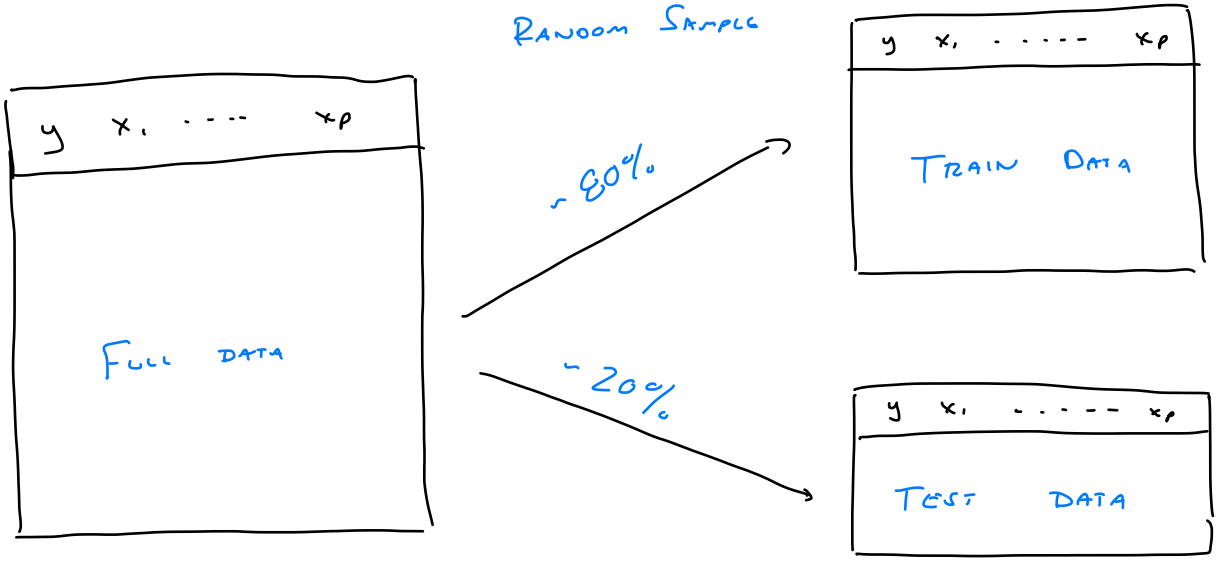
PERFECT!

?

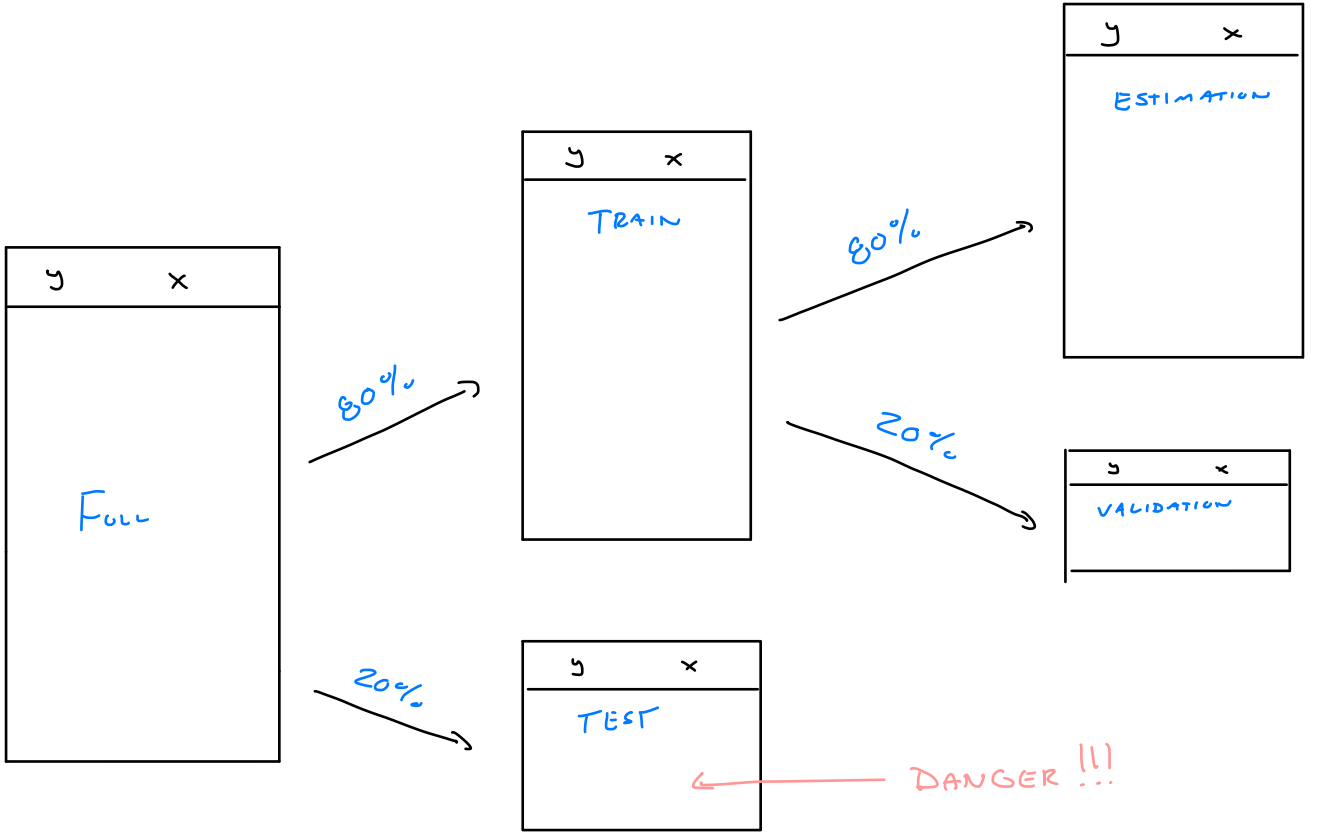
OVERFITTING

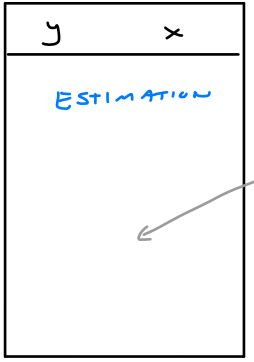
A SOLUTION

USE "UNSEEN" DATA TO EVALUATE MODELS

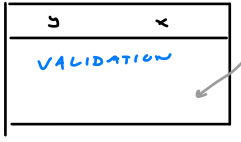


More Complicated Solution





FIT "CANDIDATE" MODELS HERE



EVALUATE "CANDIDATE" MODELS HERE

FIT TO ESTIMATION DATA

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{u}(x_i)\right)^2}$$

FROM VALIDATION DATA

VALIDATION RMSE

MODEL SELECTION →

- PICK MODEL WITH SMALLEST VALIDATION RMSE

ERROR ESTIMATION →

- RE-FIT TO TRAIN DATA
- REPORT METRIC WITH THIS MODEL ON TEST DATA

FRAMEWORK

- "LOOK AT THE DATA"
- SPLIT DATA TST-TRN EST-VAL
- PICK CANDIDATE MODELS
- FIT MODELS TO ESTIMATION DATA
- EVALUATE AND SELECT ON VALIDATION DATA
- FIT CHOSEN MODEL TO TRAIN DATA
- ESTIMATE ERROR OF CHOSEN MODEL ON TEST DATA
- USE!