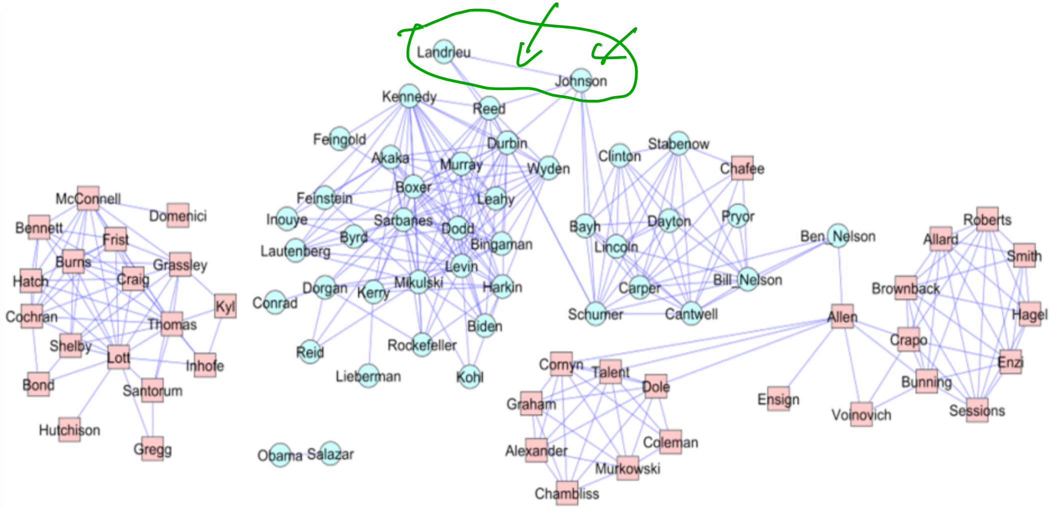
Module 1: Statistical Learning

Lecture 2

Jan 11th, 2023



Banerjee et. al., 2008

- Inputs X are voting records for each Senator
- Output: relationships between Senators
- When training the model, no output is available.
- Unsupervised learning

Unsupervised Learning

- We only have X in the data and want to output something not in the data

	Senate X	Senate Y	Senate XY	Senate Z
Bill 1	0	1	1	1
Bill 2	0	0	1	1
Bill 3	1	1	0	1
Bill 4	1	1	0	1
Bill 5	0	0	1	1
Bill 6	0	0	0	1

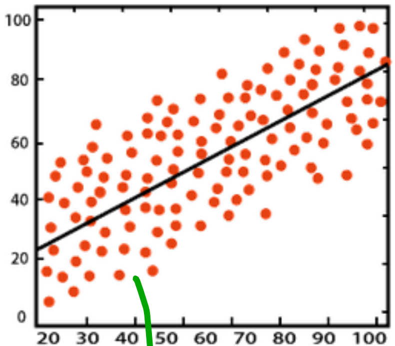
Announcement

MICHIGAN STATE UNIVERSITY

- Go to our **Team** Channel.
  - +1 point on the first HW if you post a gif in the thread!

Recap: Supervised Learning

- Inputs X and output Y both in the data.



Regression

An example data set: Advertising

MICHIGAN STATE UNIVERSITY

X Y

1		TV	Radio	Newspaper	Sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6

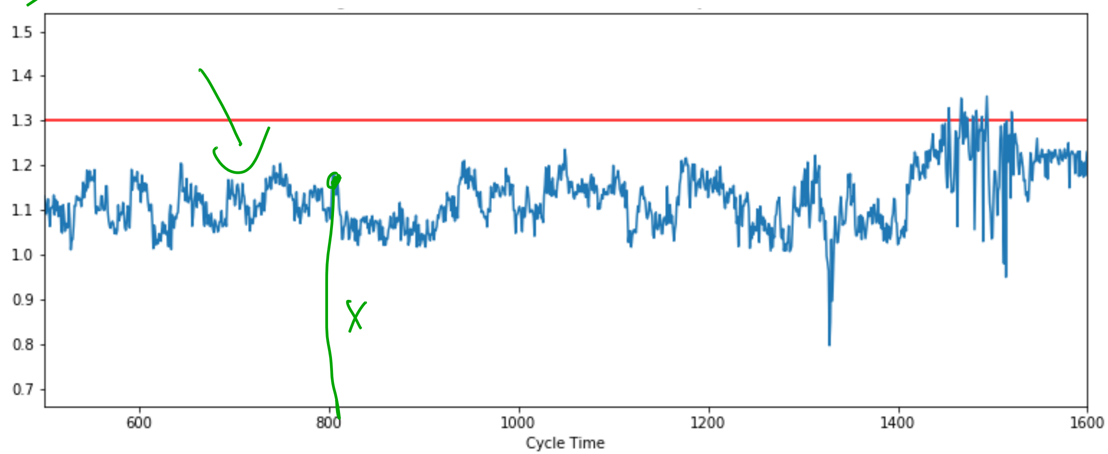
- Sales of a product in 200 markets, along with spent on three type of ad.
- Goal: ? predict Sale

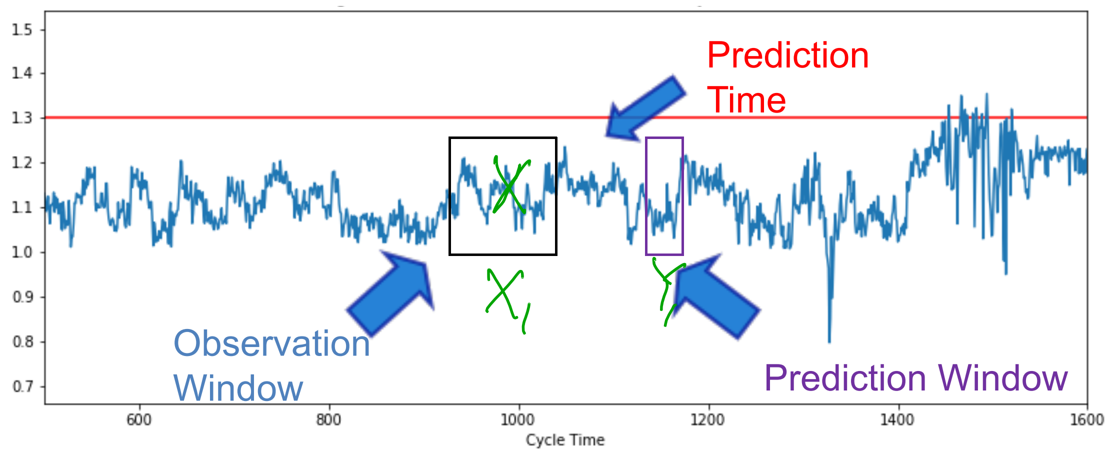
TV, Radio, New

Predicting Failure time for a machine

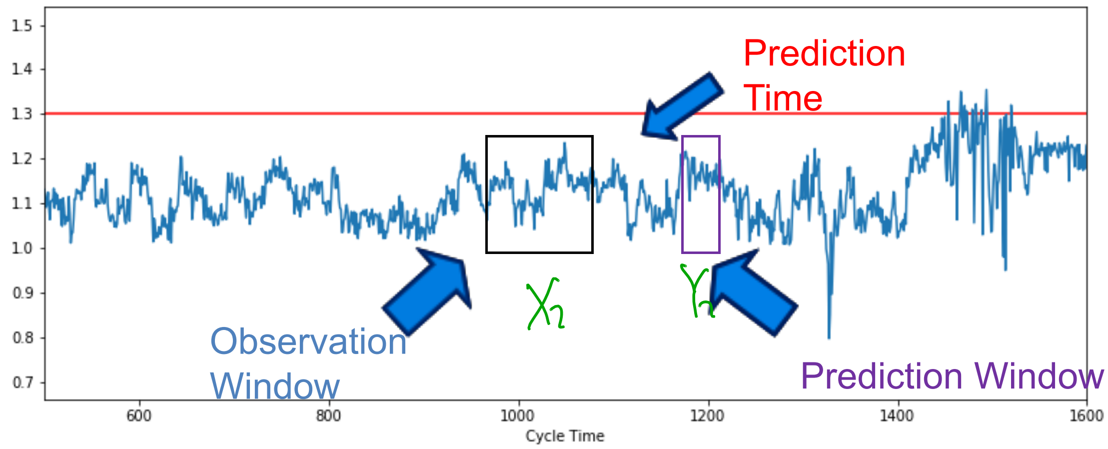
BEET

X

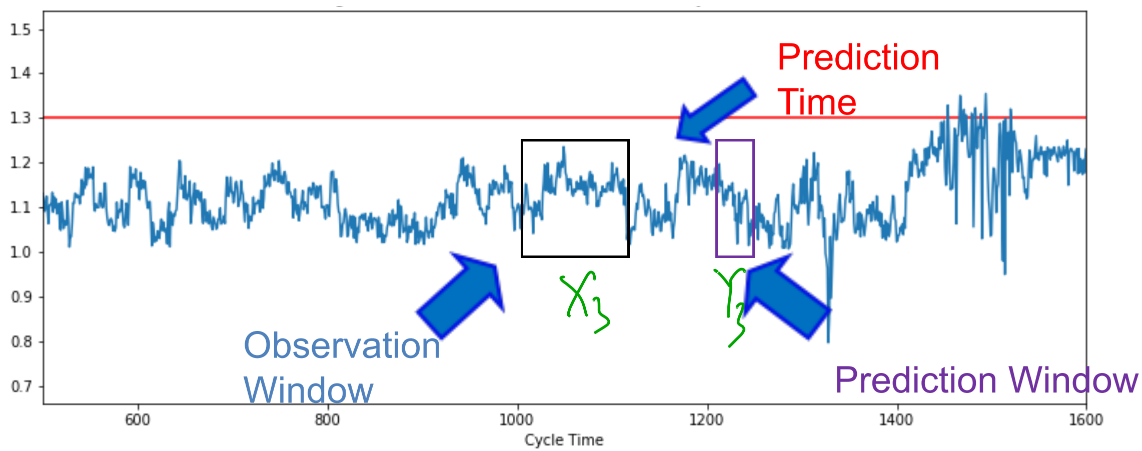


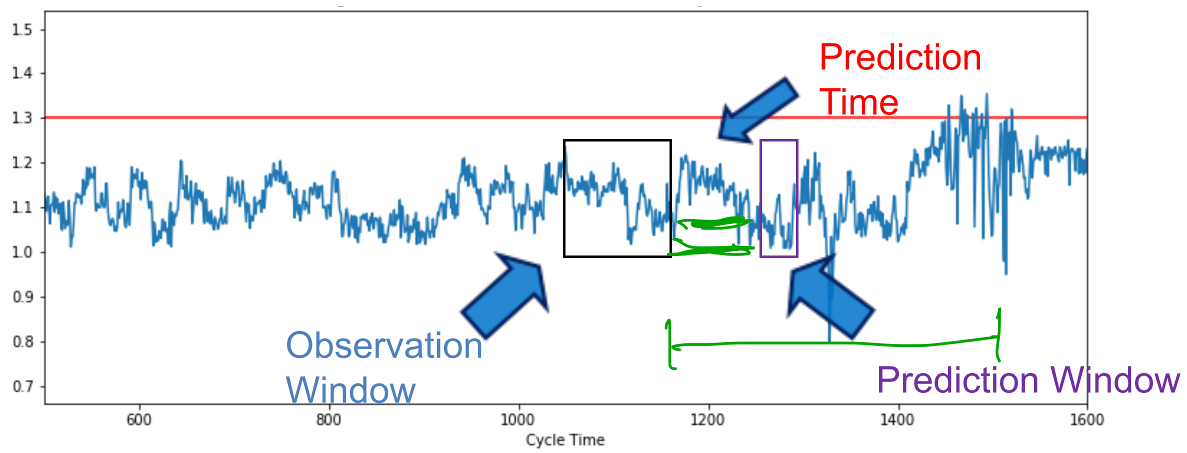


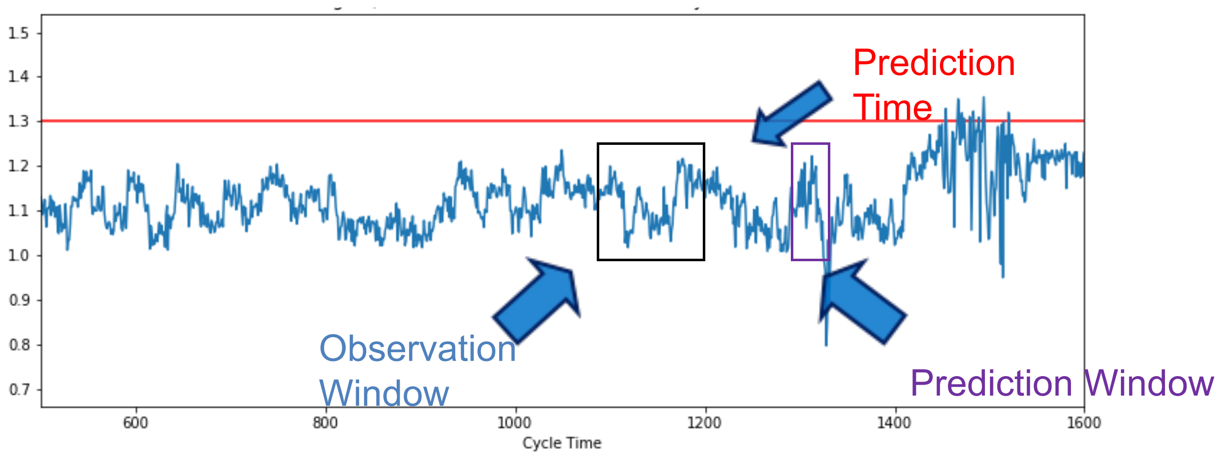
Prediction Time
Observation Window
Prediction Window
X
X1
Y1
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
600
800
1000
1200
1400
1600
Cycle Time



Prediction
Time
Observation
Window
X_2
Y_2
Prediction Window
Cycle Time
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
600
800
1000
1200
1400
1600

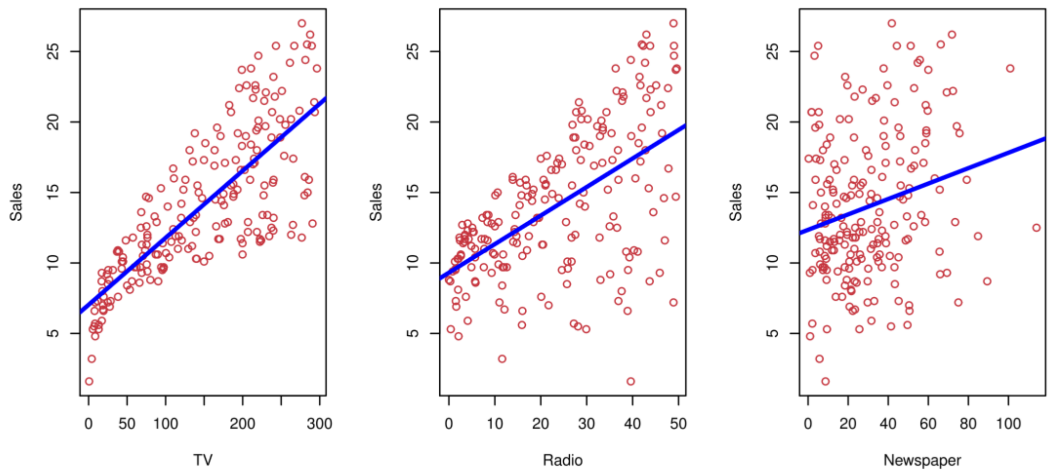

1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
600
800
1000
1200
1400
1600
Cycle Time
Prediction
Time
Observation
Window
Prediction Window
X_3
Y_3


Prediction
Time
Observation
Window
Prediction Window
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
600
800
1000
1200
1400
1600
Cycle Time


Prediction
Time
Observation
Window
Prediction Window
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
600
800
1000
1200
1400
1600
Cycle Time

Sale-Advertising

- Sales of a product in 200 different markets
- Expense on TV, radio, and newspaper in these markets

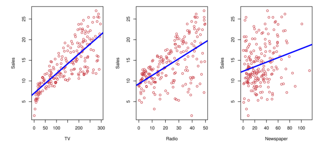


Notation

- We wish to predict sale. We refer it to be the response ***Y***
- TV is a feature (input) which we can control. We denote it **X1**. Similarly, Radio as **X2** ,and so on. We can refer to the input vector collectively as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

3x1



- Now we can write our model as

(1) $\mathrm{Y} = f(\mathrm{X})$

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

$Y = f(X)$

Here f is some fixed but unknown function.

Errors in Machine Learning: World is

MICHIGAN STATE UNIVERSITY

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

$\varepsilon \sim N(0,$

- World is too complex to model precisely
- Measurement error may not be avoidable
- Many features are not captured
- The error is where the statistics kicks in. Confidence interval, etc….

Dataset:

	Y	X_1	X_2	X_3
Market 1	10	101	20	35
Market 2	20	66	41	85
Market 3	11	101	43	78
Market 4	25	25	10	61
Market 5	5	310	51	11

rows are samples

Supervised Machine Learning Algorithm

- **Input:**

  Training data-set with features (X) and targets (Y)

- **Output:**
  - Prediction function $\boxed{f}$

f f

estimated true funct
known

Prediction vs Inference

Prediction

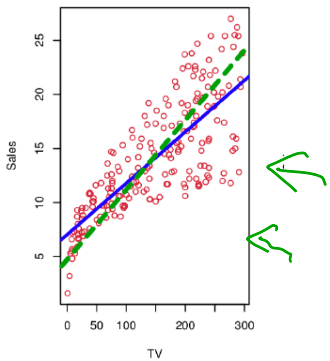
Prediction: Make predictions about future:

Inputs X are readily available, but output Y is hard to obtain

Build a model:

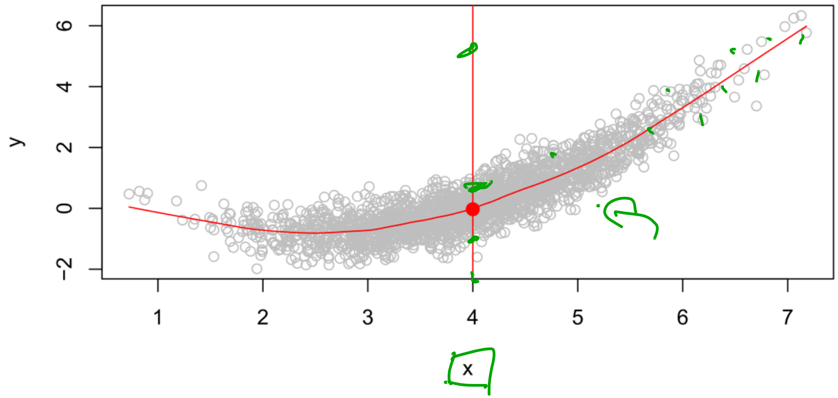
$$\hat{Y} = \hat{f}(X)$$

Example: If we spend $150 on TV advertising, what will we make in sales?



- Want to get a good guess for f, which is unknown blue
- Model is $\hat{f}$ is green dashed lines

Is there an Ideal *f*(X)?



- Given X = 4, what is 'the best' prediction for Y? or what can an Oracle say?

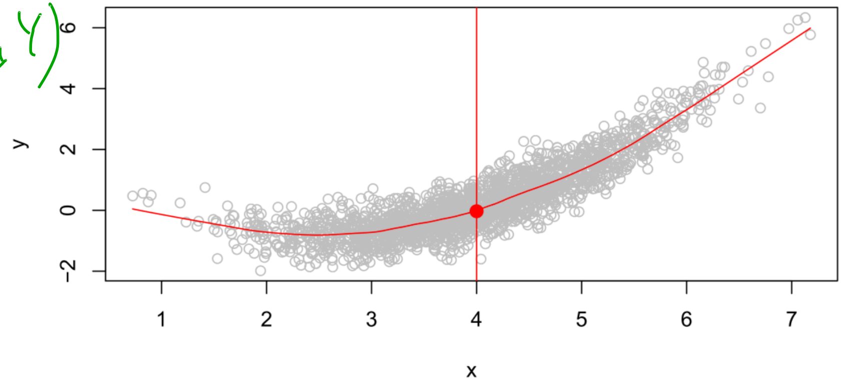
$$f \Rightarrow \min \sum_{i=1}^{n} (f - y_i)^2$$

Is there an Ideal $f(X)$?

MICHIGAN STATE UNIVERSITY

$(X,Y) \sim f(X,Y)$

$E(Y)$



$EY = \int_{\Omega_Y} y \cdot f(y)\, dy$

- Given X = 4, what is 'the best' prediction for Y? or what can an Oracle say?
- A good value is

$f(x) = E(Y|X=x) \Rightarrow E(Y|X)$

$$\boxed{f(4) = E(Y|X=4)} \Leftarrow$$

$E(Y|X=2) = E(Y|X=2)$

where E(Y|X = 4) is the expected value of Y given (condition on) X = 4.

- This $\boxed{f(x) = E(Y|X=x)}$ is called the **regression function or Oracle function**.

$E(Y|X=4) = \int y \cdot f_{Y|X}(y, x=4)\, dy$

The regression function *f(x)*

MICHIGAN STATE UNIVERSITY

- The regression function is also defined for vector X as

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

- $f(x) = E(Y|X = x)$ is the best predictor of Y given x in what sense?

It is the best for **the mean-squared prediction error** over all function *g(.)* at all points X = x.

$$f(x) = \underset{g}{\operatorname{argmin}} \mathrm{E}\left[\left(Y - g(X)\right)^2 | X = x\right]$$

$g(x) = E(Y|X)$

- **Q: Let's prove it.**

Recap

MICHIGAN STATE UNIVERSITY

- Given two random variables X and Y with joint probability density function $f_{X,Y}(x,y)$

- E(Y) = $\int_{\Omega_Y} y f_Y(y)dy$

- E(Y|X = x) =?

$$E(Y|X=x) = \int_{\Omega_Y} y f_{Y|X}(y,x)dy$$

- E(*f*(Y) – g(X) | X = x) = ?

$$E(g(X) \mid X = x)$$
$$= E(g(x) \mid X = x)$$
$$= g(x)$$

$$= E(f(Y) \mid X = x) - E(g(X) \mid X = x)$$
$$= E(f(Y) \mid X = x) - g(x)$$

The regression function $f(x)$

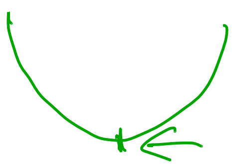
MICHIGAN STATE UNIVERSITY

$$g(x) \text{ s.t. } \min E\left(\left(Y - g(X)\right)^2 \mid X = x\right)$$

$$\Rightarrow \min E\left(\left(Y - g(x)\right)^2 \mid X = x\right)$$

$$\Rightarrow \min \int \left(y - g(x)\right)^2 f_{Y|X}(y, x)\, dy$$

$$\Rightarrow \frac{\partial \int (y - g)^2 f_{Y|X}(y, x)\, dy}{\partial g} = 0$$



$$\Rightarrow \int \cancel{2}\,(y - g)\,\cancel{(-1)}\, f_{Y|X}(y, x)\, dy = 0$$

$$\boxed{\int y\, f_{Y|X}(y, x)\, dy} - \int g\, f_{Y|X}(y, x)\, dy = 0$$

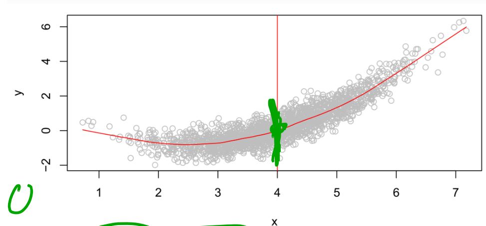
$E(Y|X=x) = g \cdot \left(\int f_{Y|X}(y, x)\, dy\right) = g$

The regression function *f(x)*

- It is the best predictor of Y with regards to mean-squared prediction error over all function g at all points X = x.

$$f(x) = E(Y|X=x) = \underset{f}{\operatorname{argmin}}\, \mathrm{E}\left[(Y - g(X))^2 | X = x\right]$$

- $\epsilon = Y - f(x)$ is the $\boxed{\text{irreducible error}}$. Even if we knew *f(x)*, we will still make errors in prediction. What cause this?



- For any estimate $\hat{f}(x)$ of $f(x)$, we have

$$E[(Y - \hat{f}(X))^2 | X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{Reducible} + \underbrace{\boxed{\mathrm{Var}(\epsilon)}}_{Irreducible}$$

Inference

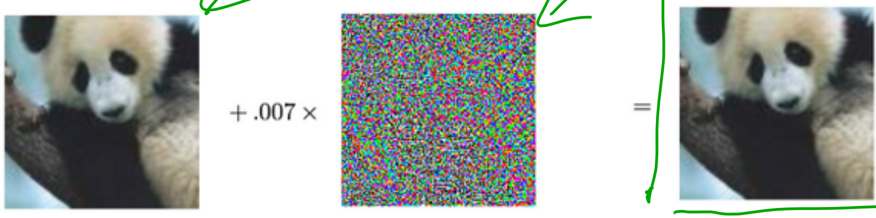
- Inference: Understand the relationship between X and Y within f

  what kind of ads work? Why?

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation? Is it more complicated?

f

Inference is important



"panda"
57.7% confidence

noise

"gibbon"
99.3% confidence

Plan for the lab

- Find a group of 4 or so.
- Download the jupyter notebook and the csv file from github.
- Get started!