

Machine Learning

Linear and Logistic Regression

Jeff Abrahamson

Cours sur l'année, 2017–2018

Linear models

Problem: $\{(x_i, y_i)\}$.

Given x , predict \hat{y} .

Linear models

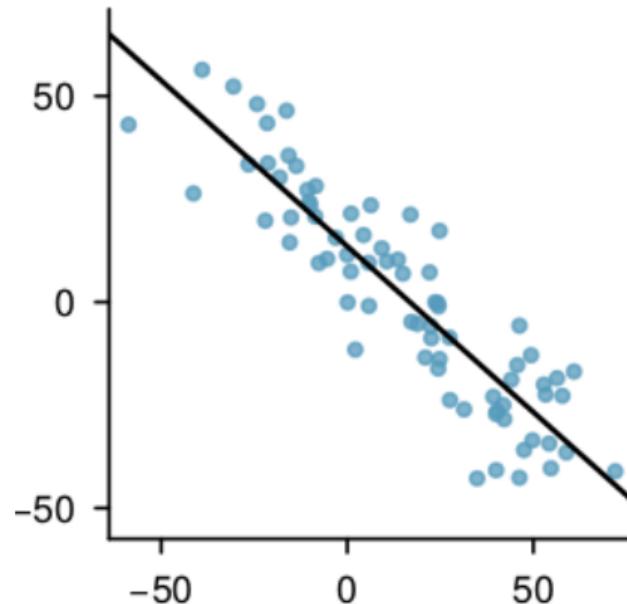
x : **explanatory** or **predictor** variable.

y : **response** variable.

For some reason, we believe a linear model is a good idea.

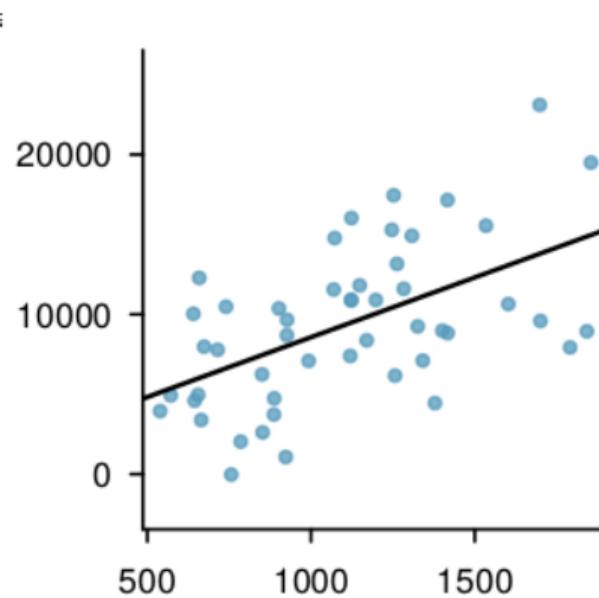
Linear models

Example:



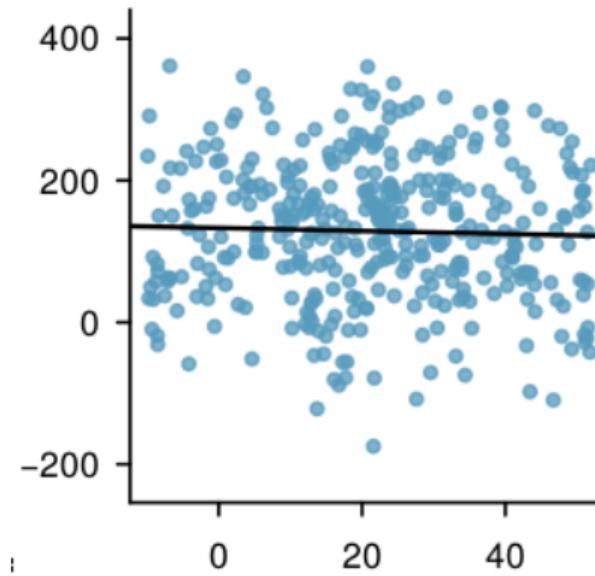
Linear models

Example:



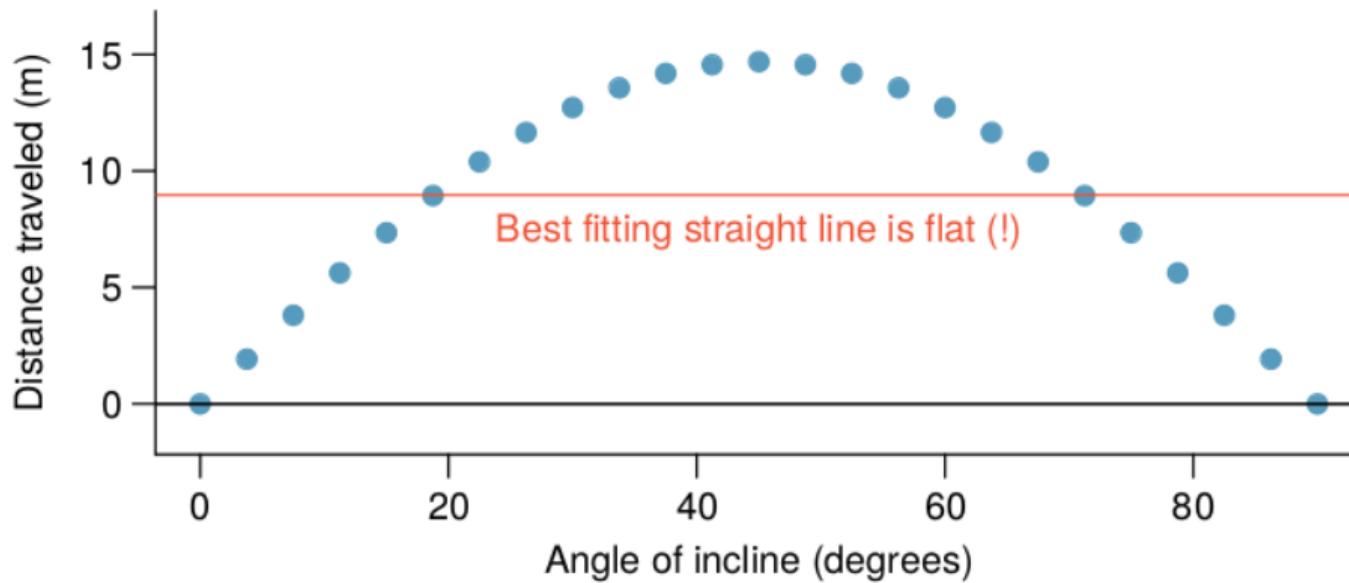
Linear models

Example:



Linear models

Example:



Residuals

What's left over.

$$\text{data} = \text{fit} + \text{residual}$$

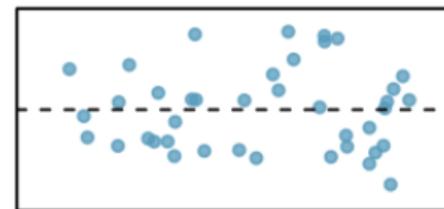
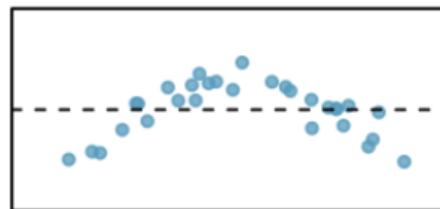
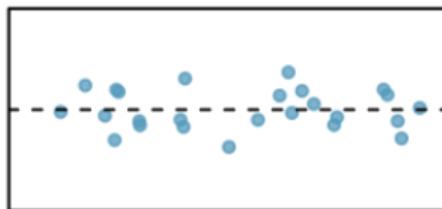
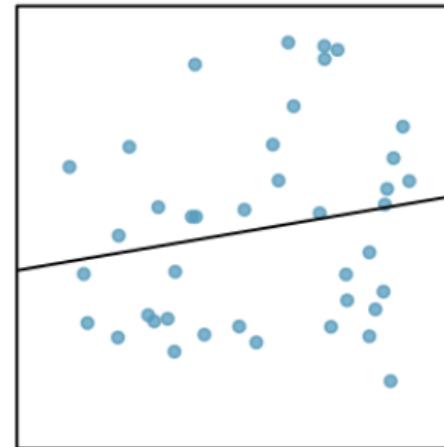
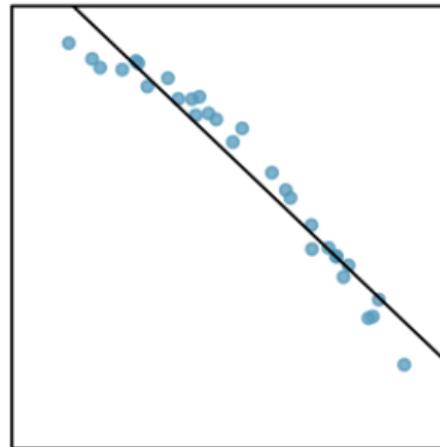
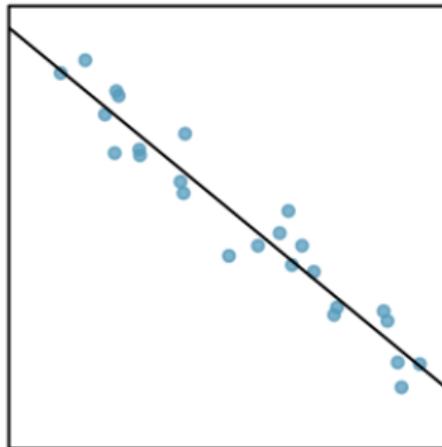
Residuals

What's left over.

$$y_i = \hat{y}_i + e_i$$

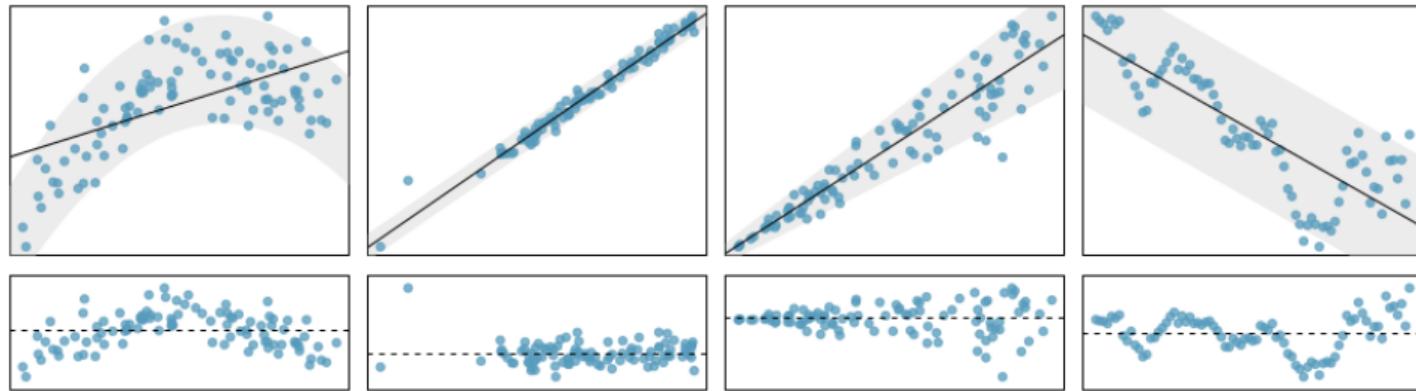
Residuals

What's left over.



Residuals

What's left over.



Residuals

What's left over.

Goal: small residuals.

$$\sum | e_i |$$

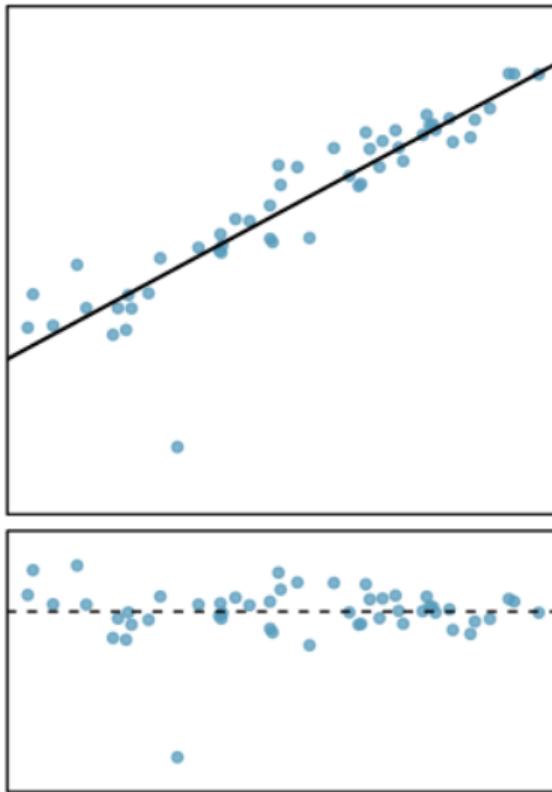
Residuals

What's left over.

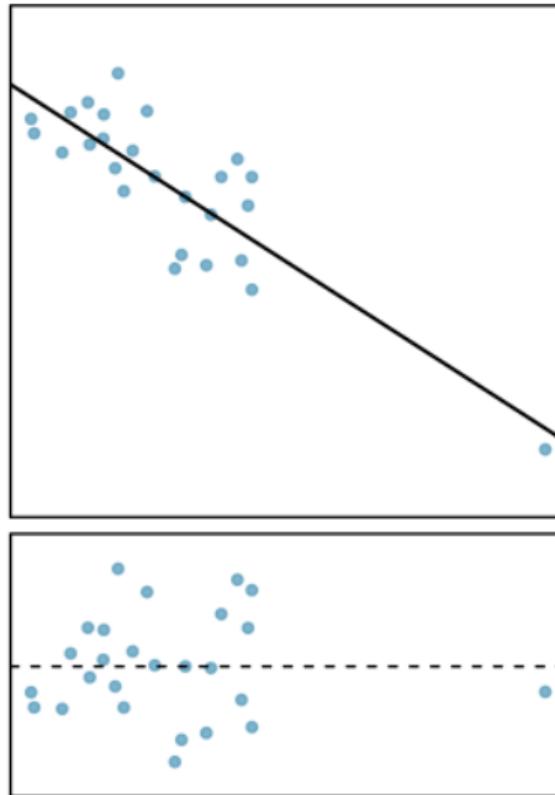
Goal: small residuals.

$$\sum e_i^2$$

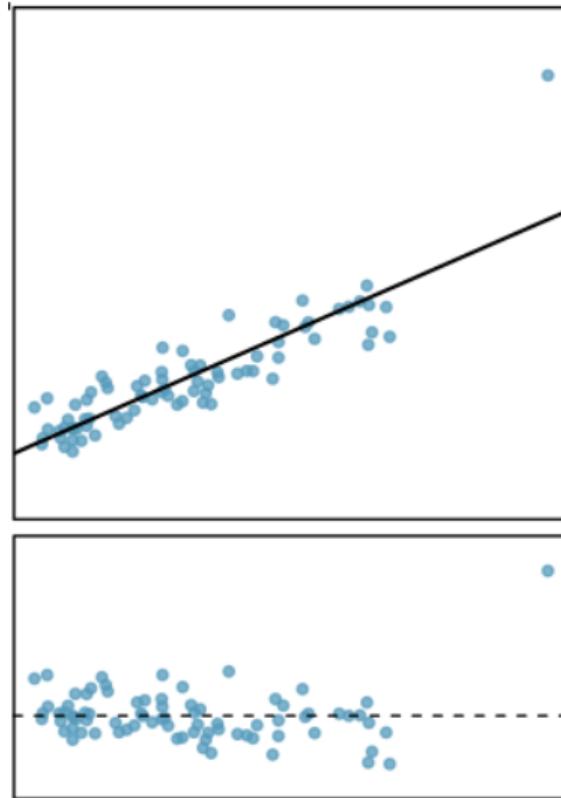
Outliers



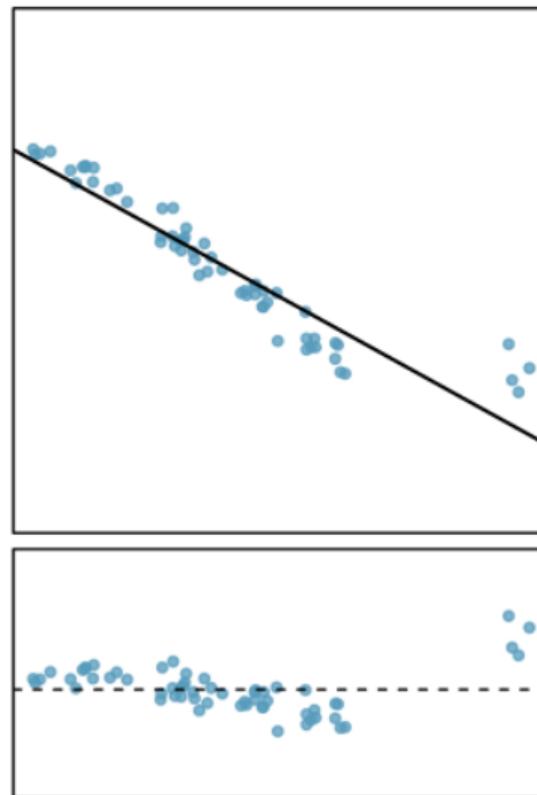
Outliers



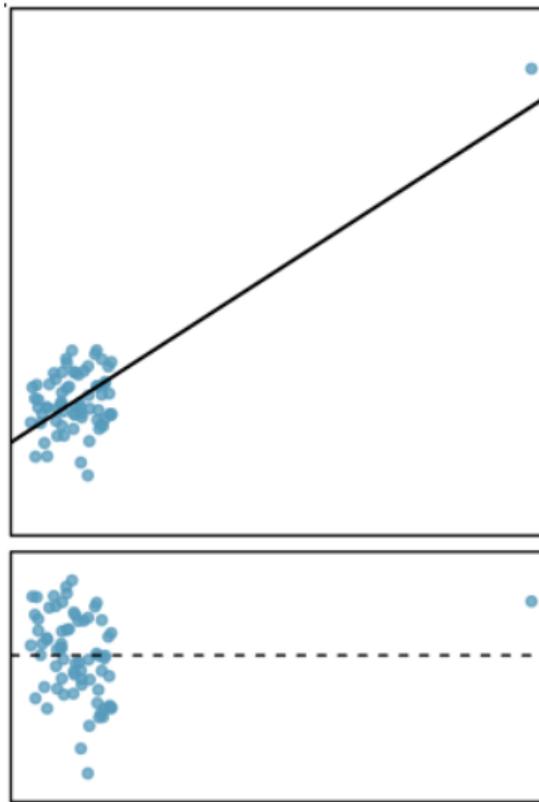
Outliers



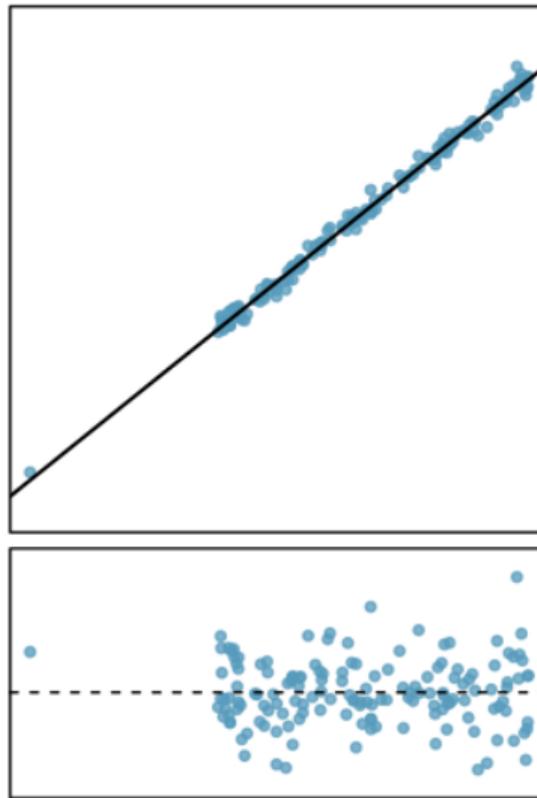
Outliers



Outliers



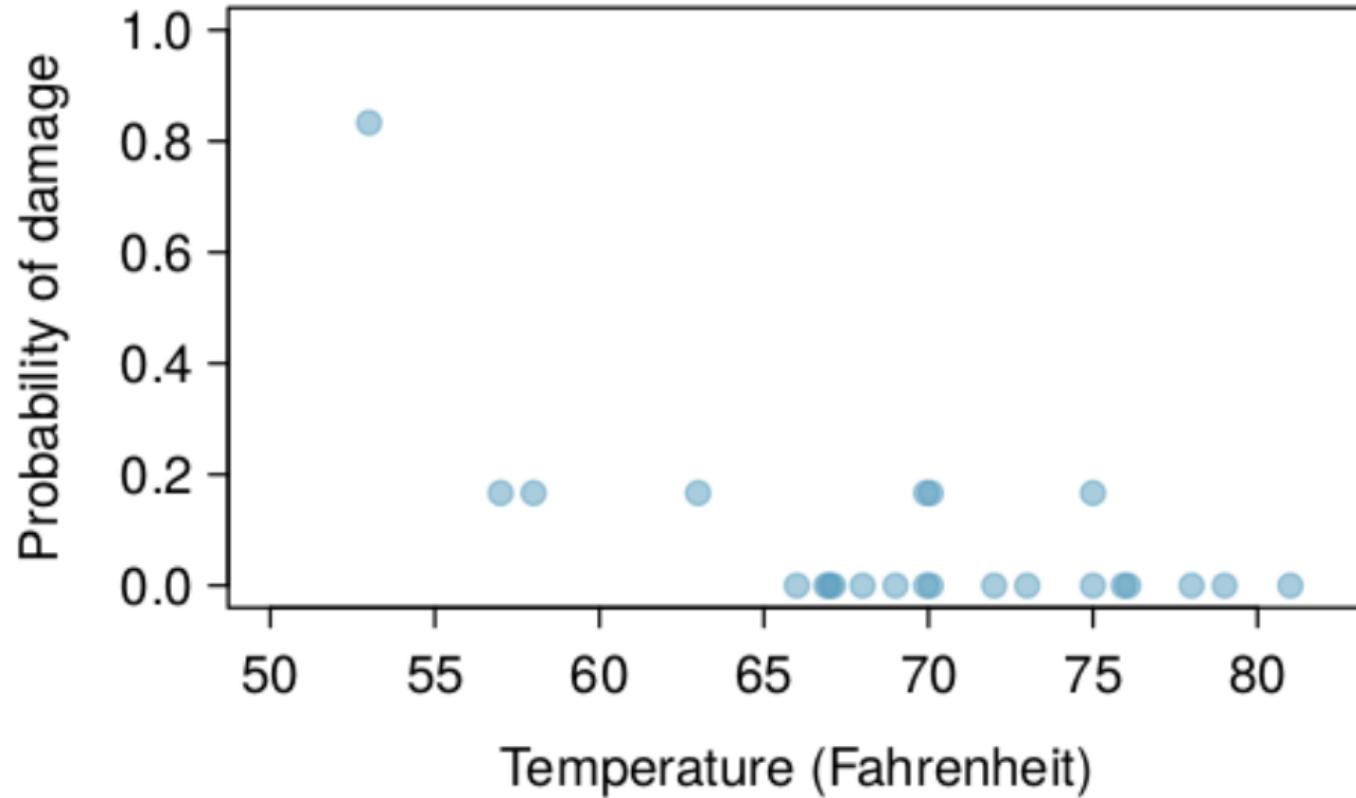
Outliers



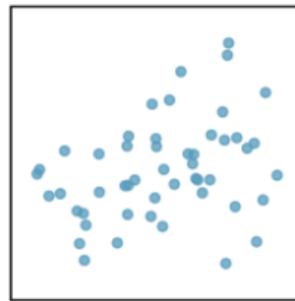
Outliers

Don't ignore outliers.

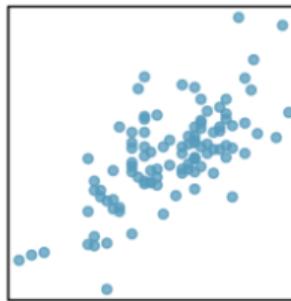
Outliers



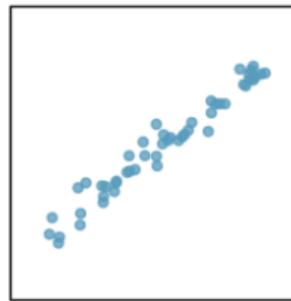
Correlation



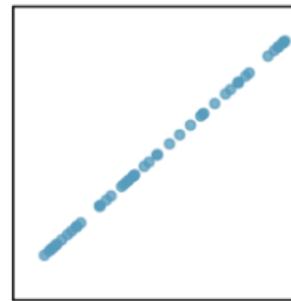
$R = 0.33$



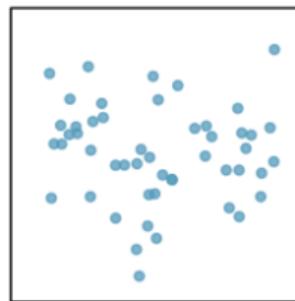
$R = 0.69$



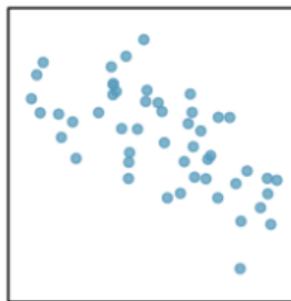
$R = 0.98$



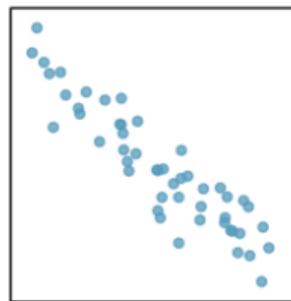
$R = 1.00$



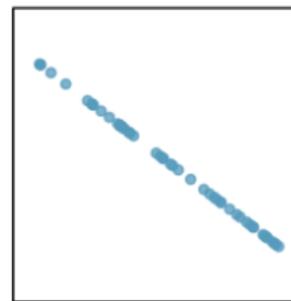
$R = -0.08$



$R = -0.64$

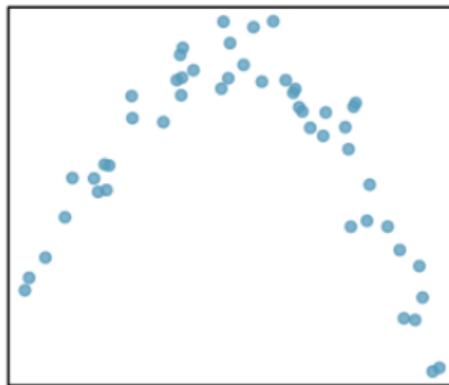


$R = -0.92$

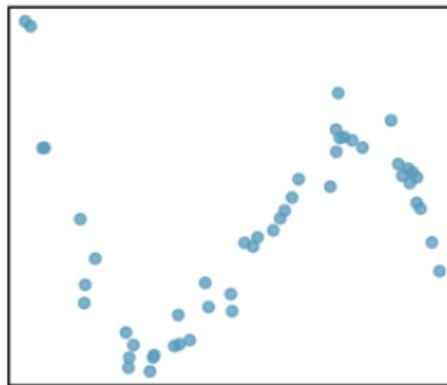


$R = -1.00$

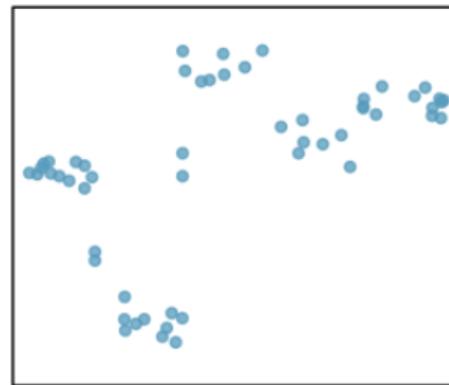
Correlation



$R = -0.23$



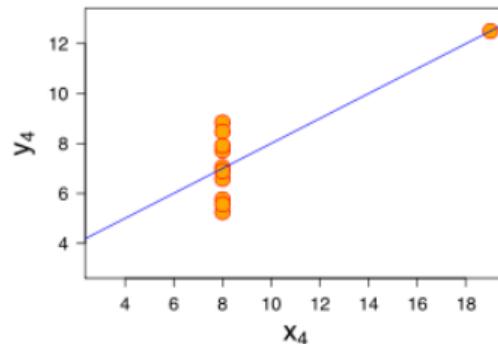
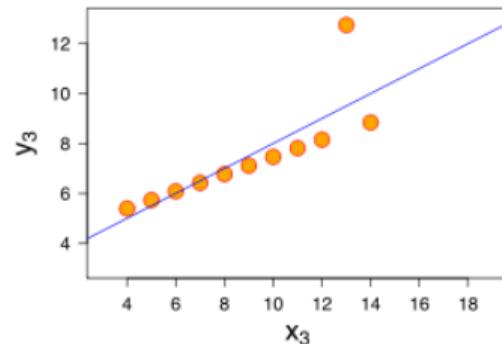
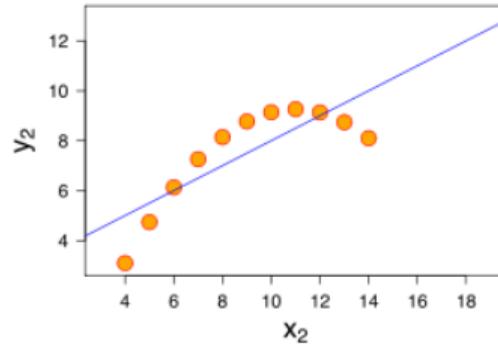
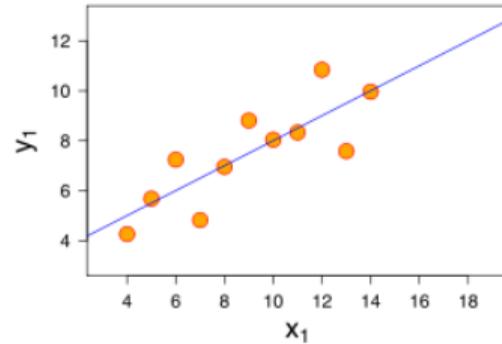
$R = 0.31$



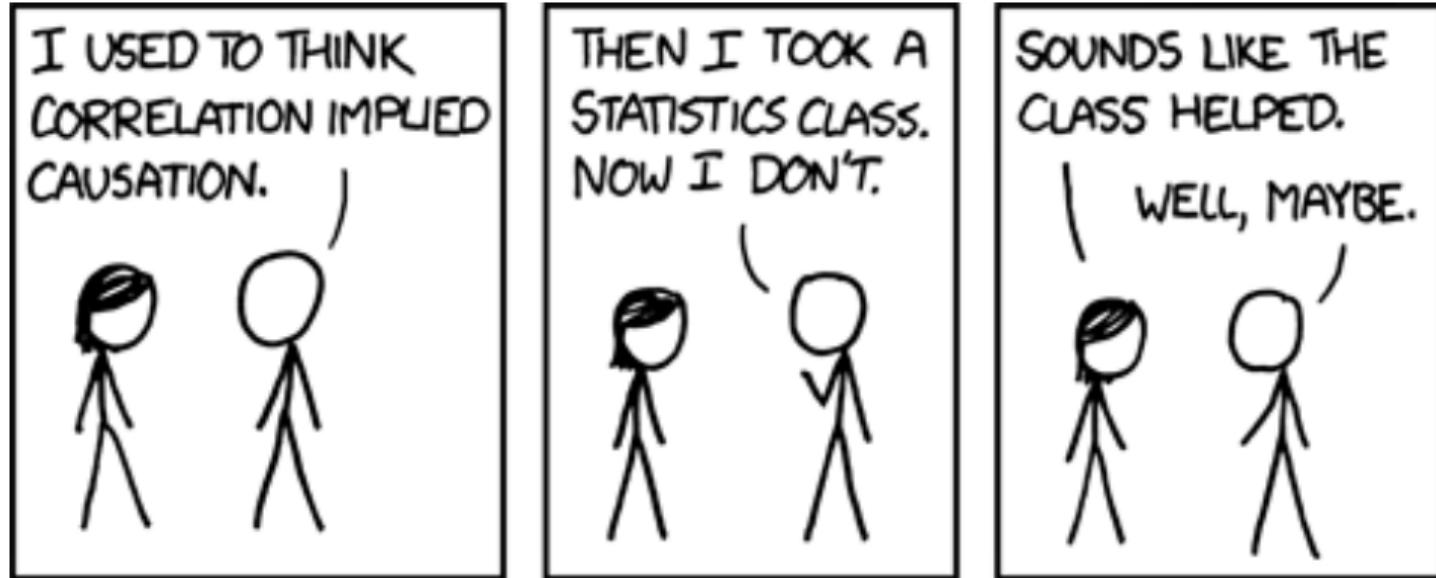
$R = 0.50$

Correlation

Anscombe's Quartet



Correlation does not imply causation



<https://xkcd.com/552/>

Hypothesis (model)

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Cost function

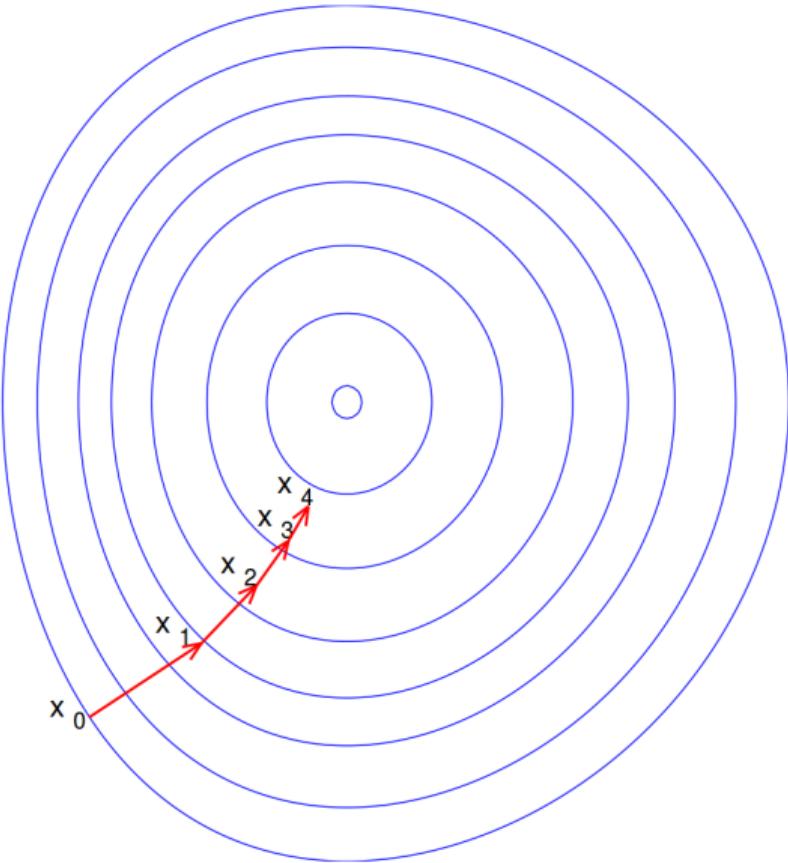
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x_i) - y_i)^2$$

Gradient descent

$$\begin{cases} \theta_0 \leftarrow \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) \\ \theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) \end{cases}$$

Gradient descent

$$\begin{cases} \theta_0 & \leftarrow \theta_0 - \frac{\alpha}{m} \sum_{i=1}^m (h_\theta(x_i) - y_i) \\ \theta_1 & \leftarrow \theta_1 - \frac{\alpha}{m} \sum_{i=1}^m (h_\theta(x_i) - y_i) \end{cases}$$



Hypothesis again

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1$$

$$= \theta_0 + \sum_{i=1}^1 \theta_i x_i$$

$$= [\theta_0, \theta_1] \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$$

$$= \theta^T x$$

Hypothesis (multiple regression)

$$h_{\theta}(x) = \theta_0 + \sum_{i=1}^n \theta_i x_i$$

$$= [\theta_0, \dots, \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \theta^T x$$

Hypothesis (multiple regression)

$$h_{\theta}(x) = \theta^T x$$
$$= \theta^T x^{(1)}$$

Hypothesis (multiple regression)

$$X = \begin{bmatrix} | & | & \cdots & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} x_0^{(1)} & x_0^{(2)} & \cdots & x_0^{(m)} \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \end{bmatrix}$$

Hypothesis (multiple regression)

$$\begin{aligned} h_{\theta}(X) &= \theta^T X \\ &= [h_0(x^{(1)}), h_0(x^{(2)}), \dots, h_0(x^{(m)})] \\ &= \theta^T X \end{aligned}$$

Hypothesis (multiple regression)

or $X\theta$ if row vectors...

Cost function (multiple regression)

$$\begin{aligned} J(\theta) &= \frac{1}{2m} \sum_{i=1}^m \left(h_{\theta}(x^{(i)}) - y^{(i)} \right)^2 \\ &= \frac{1}{2m} (X\theta - Y)^T (X\theta - Y) \end{aligned}$$

Gradient descent (multiple regression)

$$\theta_j \leftarrow \theta_j - \frac{\alpha}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

for $j = 1, \dots, n$

Gradient descent (multiple regression)

$$\theta \leftarrow \theta - \nabla J(\theta)$$

where $\nabla = \begin{bmatrix} \frac{\partial}{\partial \theta_0} \\ \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_n} \end{bmatrix}$

A photograph of a narrow, sunlit street in a European city, likely Italy, based on the architecture and laundry style. Laundry is hung from multiple clotheslines strung between tall, multi-story buildings. The laundry includes various items like red shirts, striped tunics, and dark pants. The buildings have traditional features like arched windows and small balconies with potted plants. The sky is clear and blue.

Questions?

Linear regression

- Continuous output
- Normal residues
- Predict \hat{y} for x given $\{(x_i, y_i)\}$

Logistic regression

- Binary output
- Classification

Logistic regression

- Have: continuous and discrete inputs
- Want: class (0 or 1)

Probabilistic inspiration

$h_\theta(x) = .75 \iff$ event has 75% of being true

Probabilistic inspiration

$$h_{\theta}(x) = \Pr(y = 1 \mid x; \theta) = 0.75$$

Probabilistic inspiration

So this must be true:

$$\Pr(y = 0 \mid x; \theta) + \Pr(y = 1 \mid x; \theta) = 1$$

Probabilistic inspiration

Set $y = 1 \iff h_\theta(x) = \Pr(y = 1 \mid x; \theta) > \frac{1}{2}$

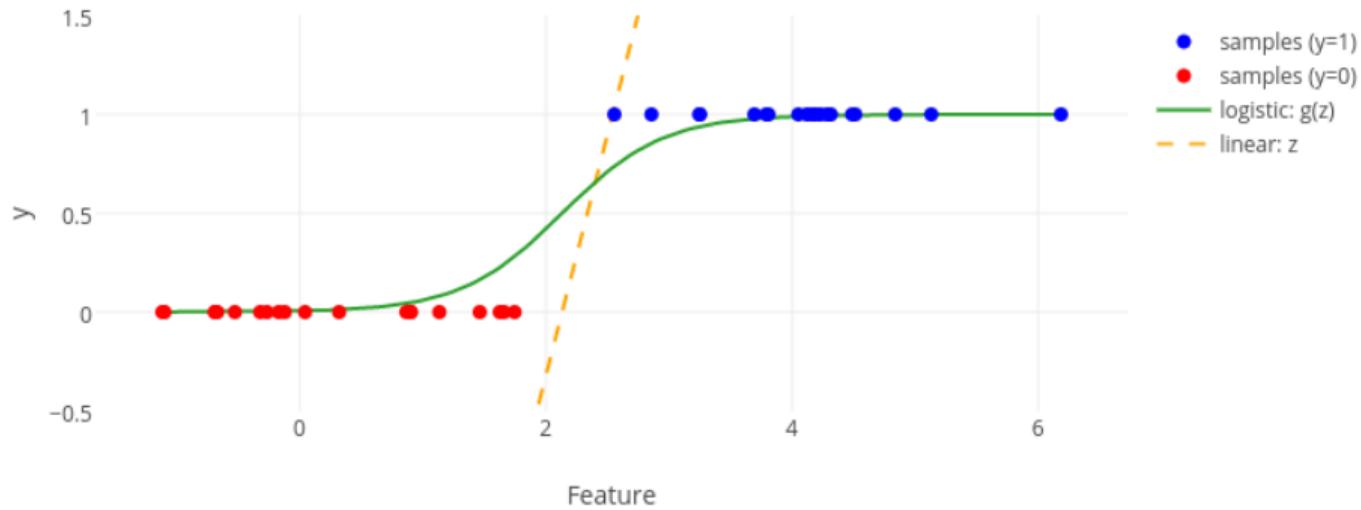
Probabilistic inspiration

Math review:

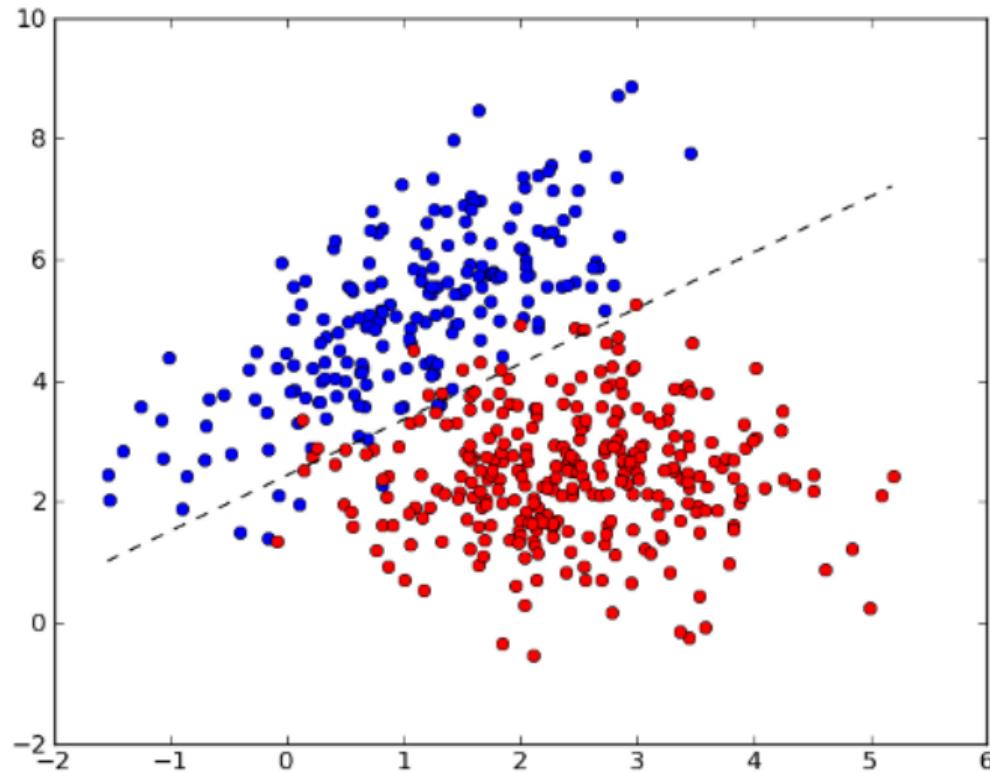
- $z = (\theta^T x)$
- $\theta^T x \geq 0 \iff h_\theta \geq 0.5$
- $\theta^T x \geq 0 \iff \text{predict } y = 1$

Logistic Regression

Logistic Regression: 1 Feature



Logistic Regression



Logistic (sigmoid, logit) function

$$g(z) = \frac{1}{1 + e^{-z}}$$

Logistic (sigmoid, logit) function

$$g(z) = \frac{1}{1 + e^{-z}}$$

Exercise: plot this

Cost function in logistic regression

In linear regression, we had

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

Cost function in logistic regression

In linear regression, we had

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x) - y)^2$$

Cost function in logistic regression

In linear regression, we had

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m \text{Cost}(h_\theta(x), y)$$

Cost function in logistic regression

Here's a convex cost function:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Cost function in logistic regression

Here's a convex cost function:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Exercise: Plot this (cost vs y).

Cost function in logistic regression

Here's a convex cost function:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m \text{Cost}(h_\theta(x), y)$$

Cost function in logistic regression

Here's a convex cost function:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$J(\theta) = y \cdot \log(h_\theta(x)) + (1 - y) \cdot \log(1 - h_\theta(x))$$

Gradient descent

$$\theta_j \leftarrow \theta_j - \frac{\alpha}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

for $j = 1, \dots, n$

null hypothesis

true positive, true negative

false positive, false negative

type I error

(incorrect rejection of null hypothesis)

type II error

(failure to reject null hypothesis)

sensitivity

100% sensitivity = no false negatives

specificity

100% specificity = no false positives

Precision

$$P = \frac{TP}{TP + FP}$$

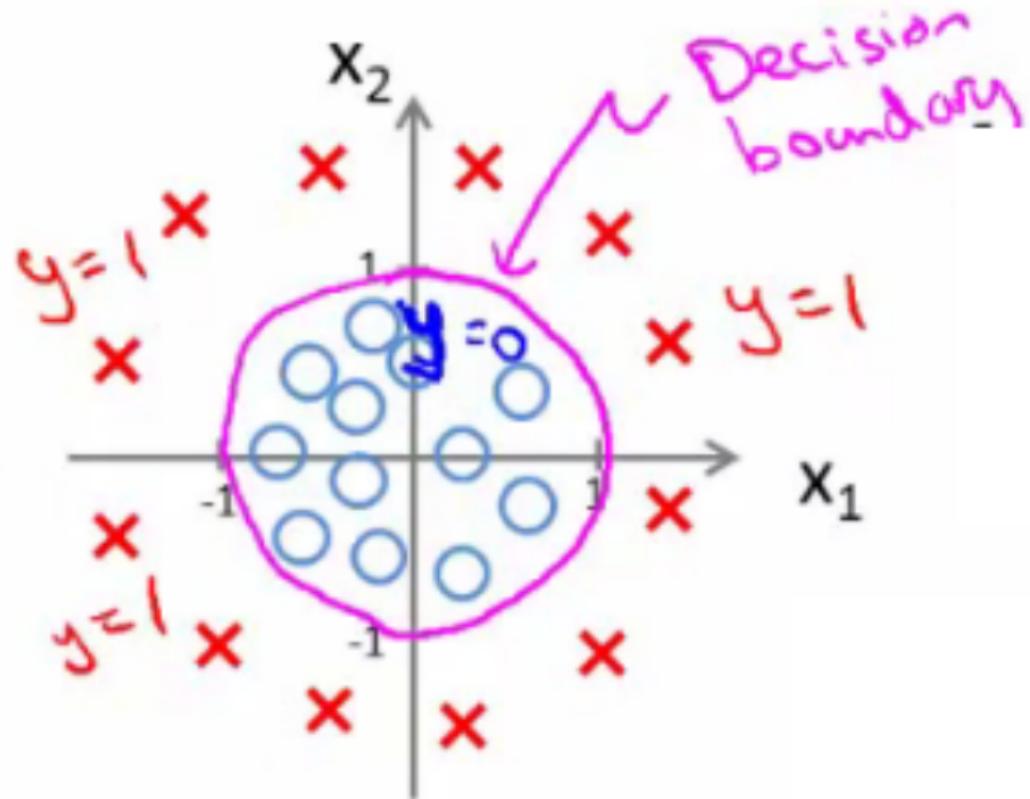
Recall

$$R = \frac{TP}{TP + FN}$$

F1 score

$$F1 = \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Non-linear decision boundaries



Non-linear decision boundaries

OvA = OvR

OvO

Non-linear decision boundaries

One vs All = One vs Rest

One vs One

A brown teddy bear is sitting on a weathered wooden fence. The bear is facing forward, looking slightly upwards and to the right. The background is a clear blue sky. In the upper left quadrant of the image, the word "questions?" is written in a blue, sans-serif font.

questions?