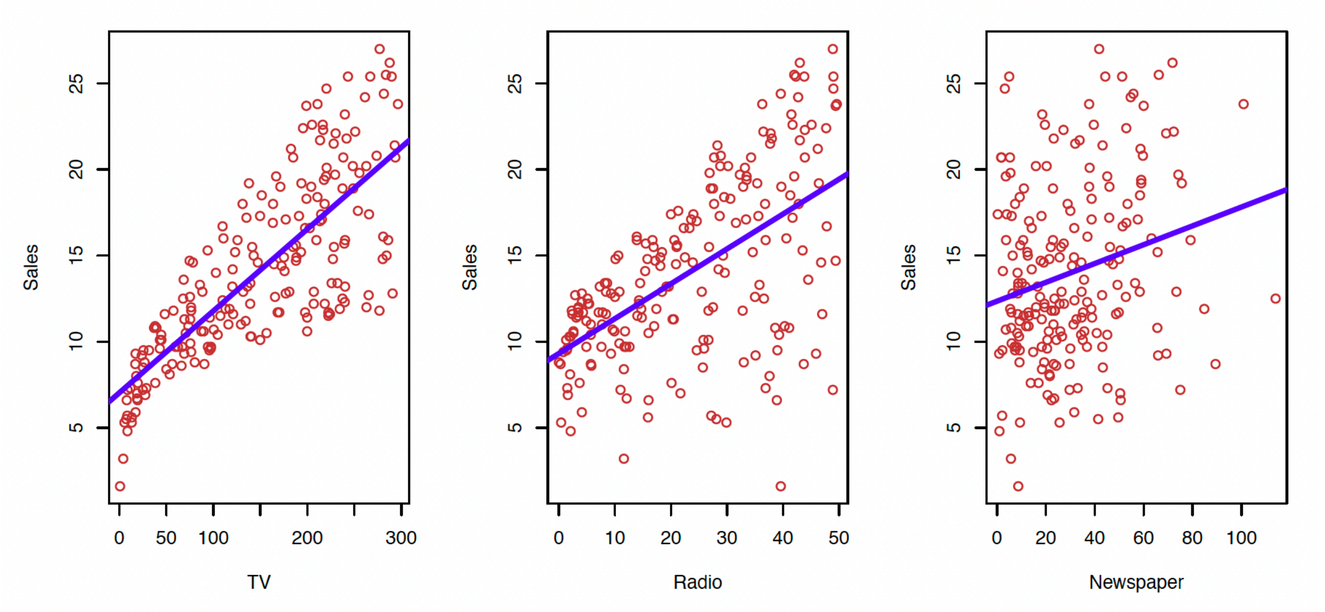
What is statistical learning?

@manuel__pita

Dr. Manuel Pita

UNIVERSIDADE
LUSÓFONA



X-axis: advertising budget in thousands of Dollars
Y-axis: units sold x 1000

TV, Radio and Newspaper are shown separately, along with the sales.

We want to predict sales based on the data we have.

$$\texttt{sales} \approx f(\texttt{Radio, Newspaper, TV})$$

Maybe considering the three types of budget together is a better idea.

Some terminology and notation

Sales is our **response, target, output or dependent** variable,
while Radio, Newspaper and TV are the **independent, feature, input or predictor** variables

We can refer to the input variable X as a vector:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Now we can write our model as:

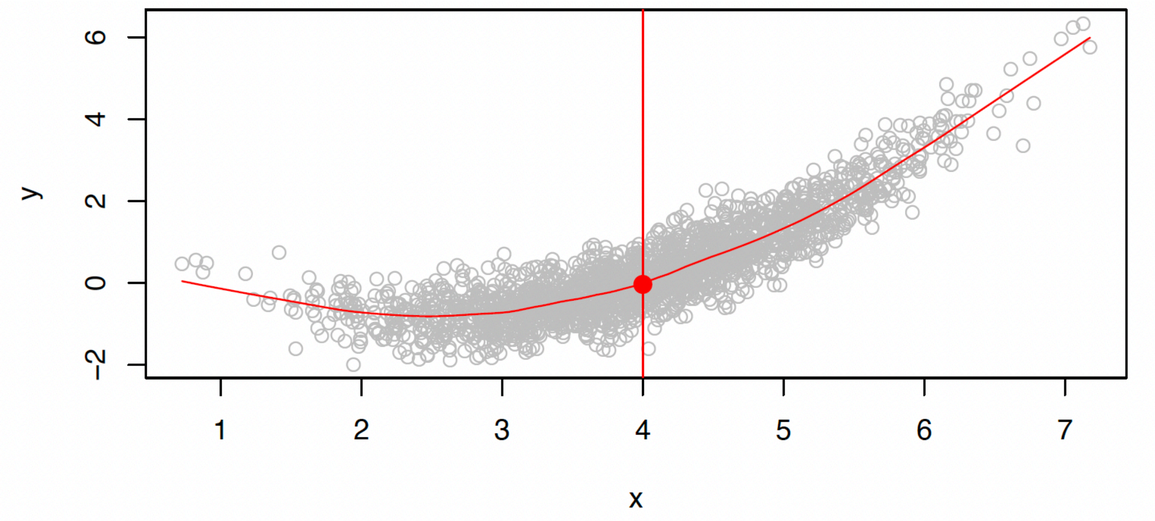
$$Y = f(X) + \epsilon$$

- ϵ is a random error we always assume to exist because the only perfect model of a system is the system itself: **all models have errors!**

What do we want a statistical model for?

- A good statistical model f can help us **predict** system Y at new points X
- We can understand what predictor variables are most important
- For example, the "**years of education**" variable has a big impact on **income**
- But "**marital status**" does not
- Often, the complexity of f is critical in separating the effects of predictor vars.

What is the best model?



What would be a good value of f(X) when X = 4?

We can use the **expected value**, given X = 4:

$$f(4) = E(Y|X = 4)$$

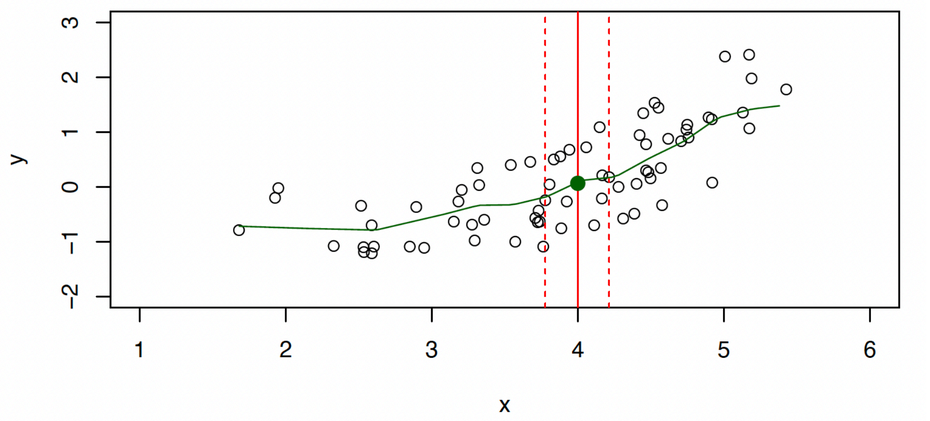
In this case, the expected value can be the average, extrapolating to the whole model:

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

And we call this the **regression function**.

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

How to estimate $f(x)$?



1. Often, we don't have enough points for $X = 4$ exactly
2. This means we cannot compute $E(Y|X = x)$
3. So, we relax the definition of expected value: $\hat{f}(x) = \texttt{Average}(Y|X \in \mathcal{N}(x))$
4. Where $\mathcal{N}(x)$ is a neighbourhood of points around x

If we have many data points, and about four or less predictor variables (p) the nearest neighbours method usually works fine. In other cases, specially large p

This is known as the **curse of dimensionality**: in high-dimensional spaces data points tend to be very far from each other

Parametric and structure models

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

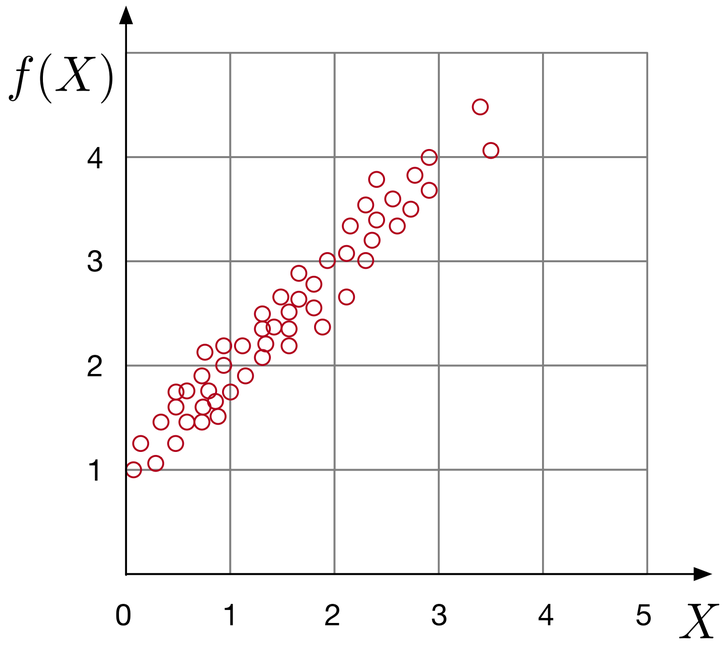
$[\beta_0, \beta_1, \beta_2, ..., \beta_p]$ are the model parameters

Linear models are rarely good models of the real world, but they are usually our starting point.

Parametric and structure models

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

What linear model would be a good fit for these data points?

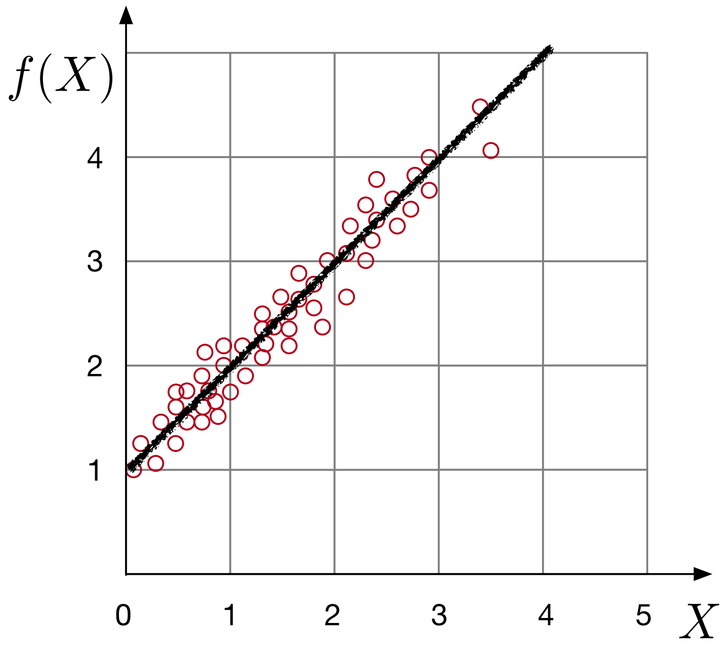


Parametric and structure models

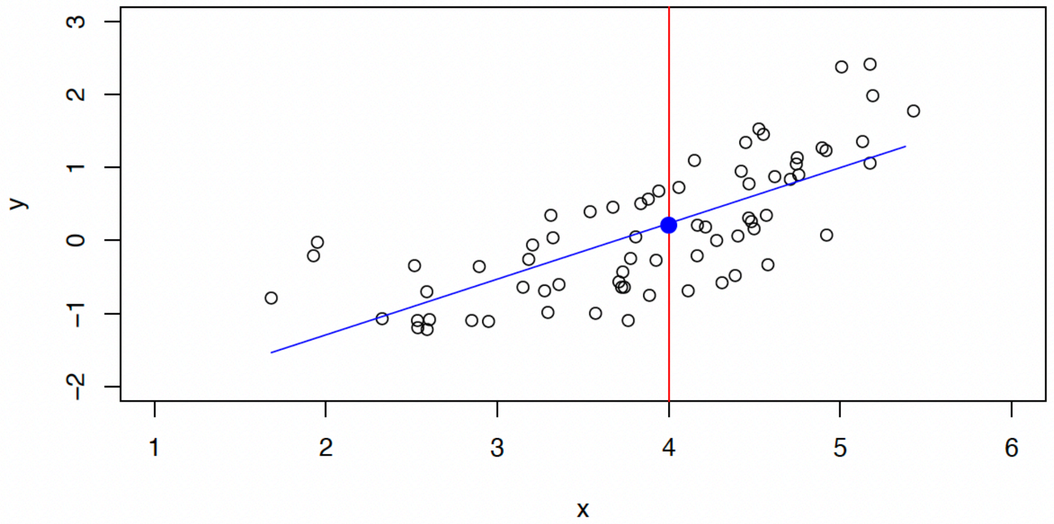
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

What linear model would be a good fit for these data points?

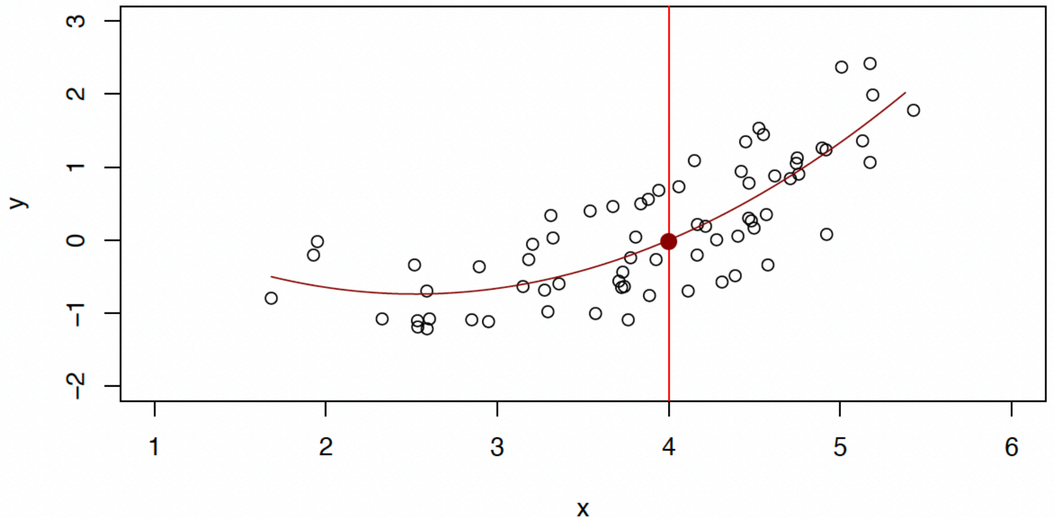
$$f(X) = X + 1$$

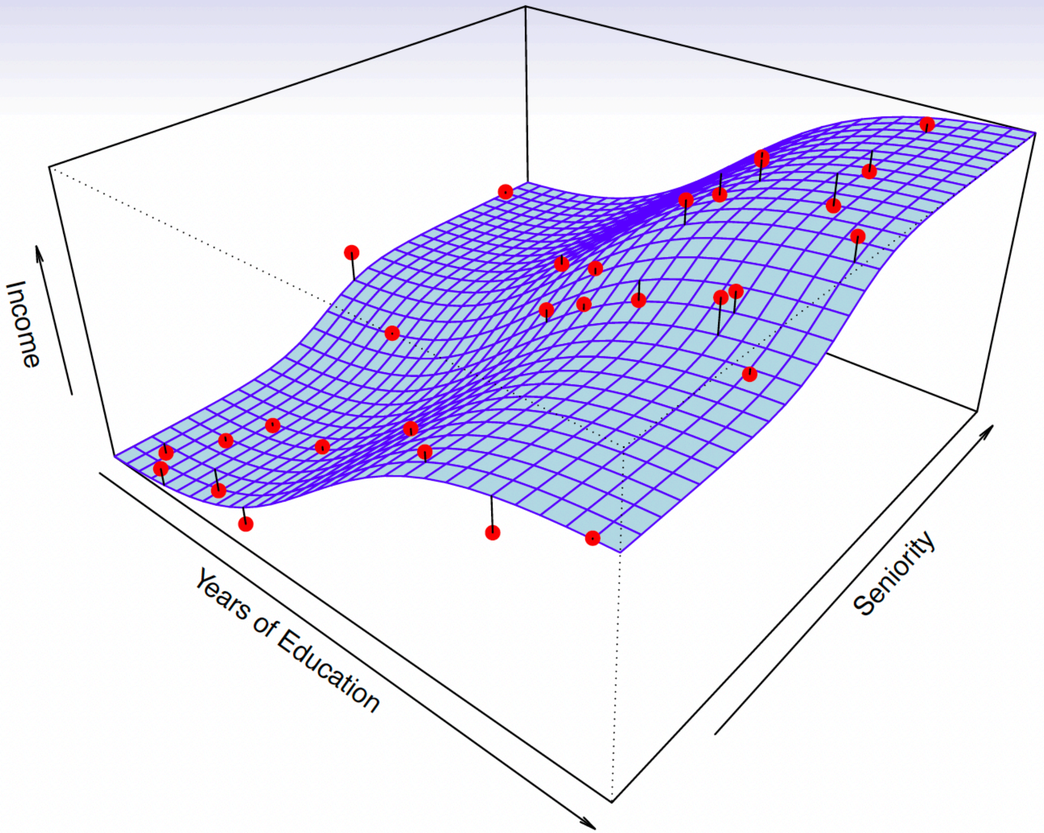


A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A quadratic model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better.

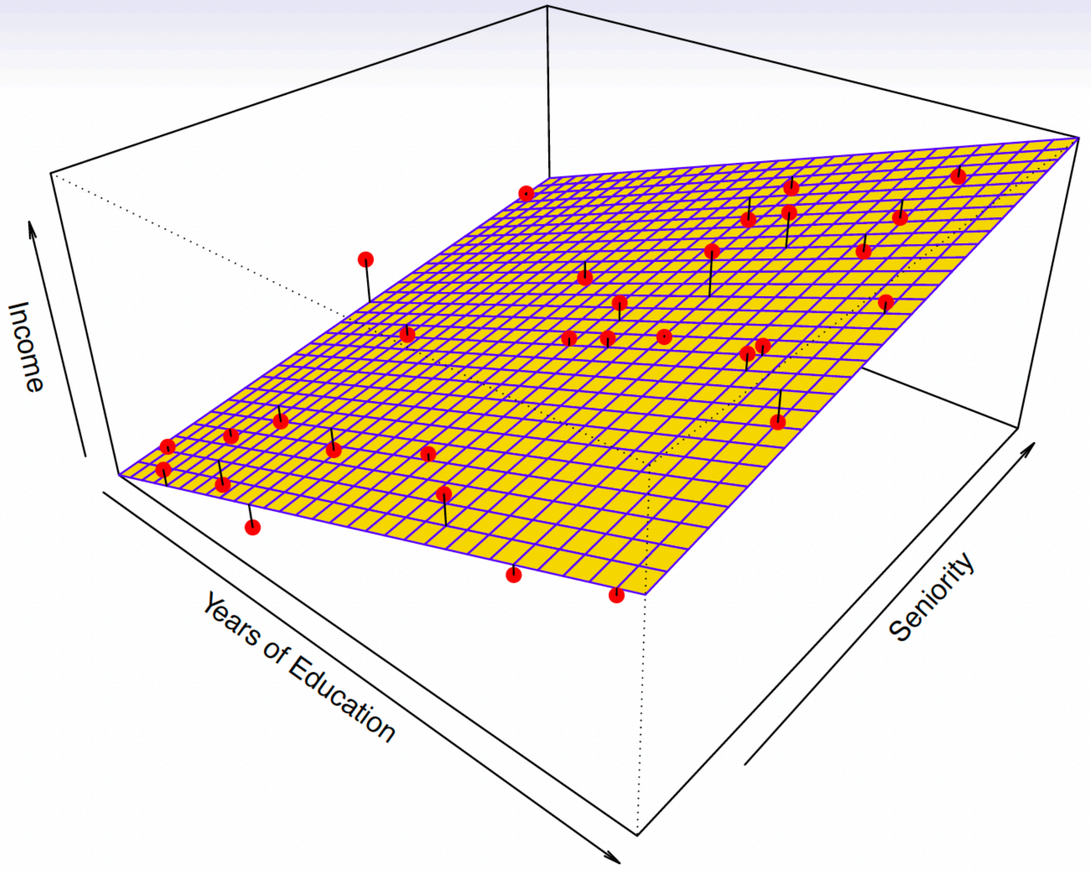




Simulated example. Red points are simulated values for `income` from the model

$$\texttt{income} = f(\texttt{education}, \texttt{seniority}) + \epsilon$$

f is the blue surface.



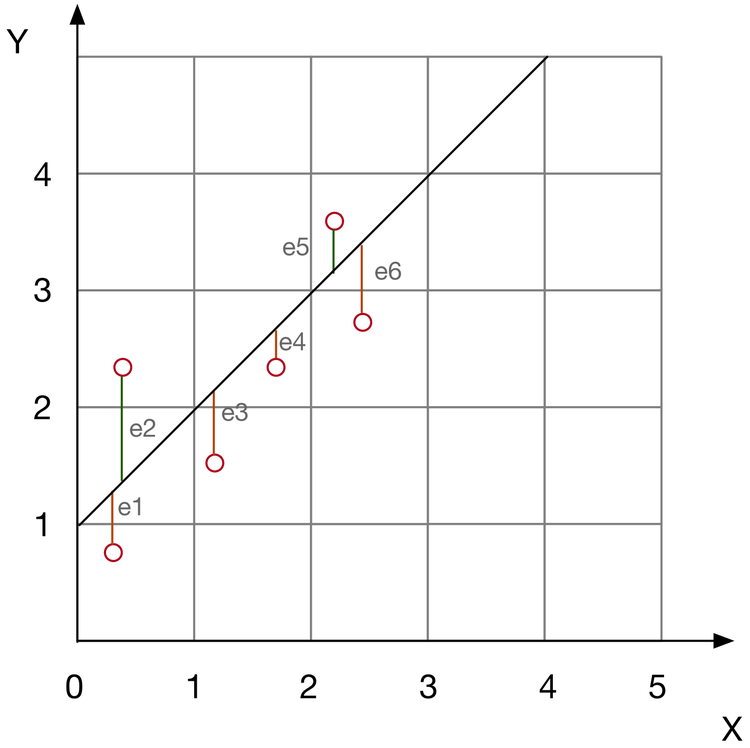
Linear regression model fit to the simulated data.

$$\hat{f}_L(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$$

How to validate our model?

- When we get out data, we split it into **training** and **testing** (usually 70:30)
- Minimise the mean sum of squared residuals to fit the model (training)
- Compute the mean sum of squared residuals to the testing data: what for?
- Answer: to see how the model behaves on previously unseen data.
- If we train and test with the same data, the model may become too fine-tuned to the training data and do poorly on testing data, this is called **overfitting**.

Variance (statistics)



$$\text{SSR} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

This Sum of Squared Residuals is a number that quantifies how dispersed the data points e_i are to the line. As you will see later in the module, this is related to the statistical concept of **variance**.

Self evaluation #1

1. What are structured, semi-structured and unstructured data (with examples)?
2. What is the definition of rectangular data?
3. What is a regression model?
4. What is a residual?

For each answer, add your degree of confidence that you got it right:

Low, medium or high