Linear regression

Part I

@manuel__pita

Dr. Manuel Pita

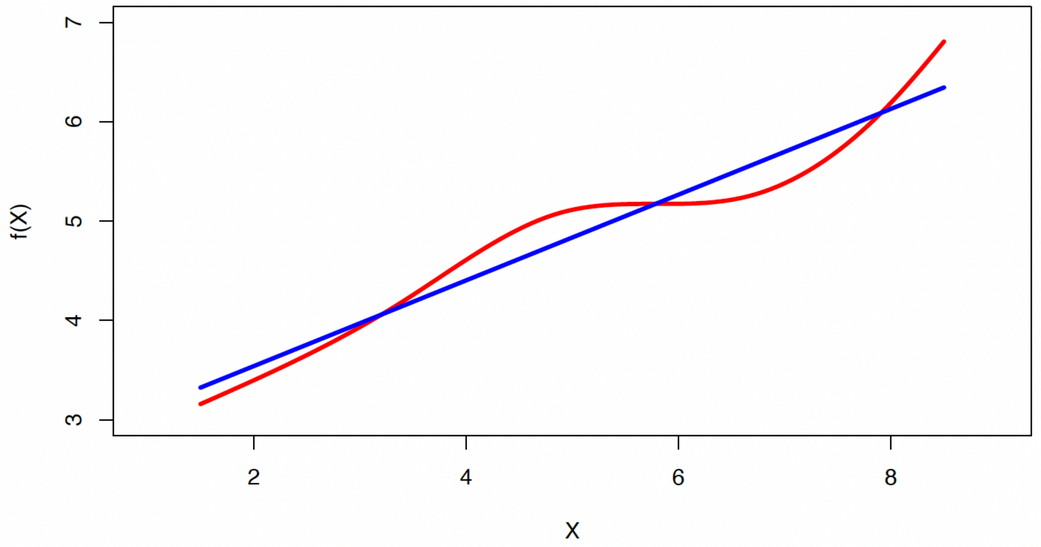
UNIVERSIDADE
LUSÓFONA

From last week…

- **Goal**: predict a dependent variable from one or more independent variables
- Dependent and independent variables go by several names
- We introduced the definition of a **statistical model**
- We learnt that the **only perfect model of a system Y is the system Y itself**
- Therefore, **all models are imperfect** (have some errors)
- **Some errors can be reduced, if we have good data**. Some errors are not reducible
- **Not all predictor variables have the same power** in making predictions
- We **model by fitting a function on the expected value of Y**, given the data
- We use some **training data** to learn the model and keep some **testing data** to see if the model generalises well
- We mentioned that **some models overfit**: they do great on training data but do not generalise (more on that will come)
- **Linear models are unrealistic, but the best starting point**

Linear regression

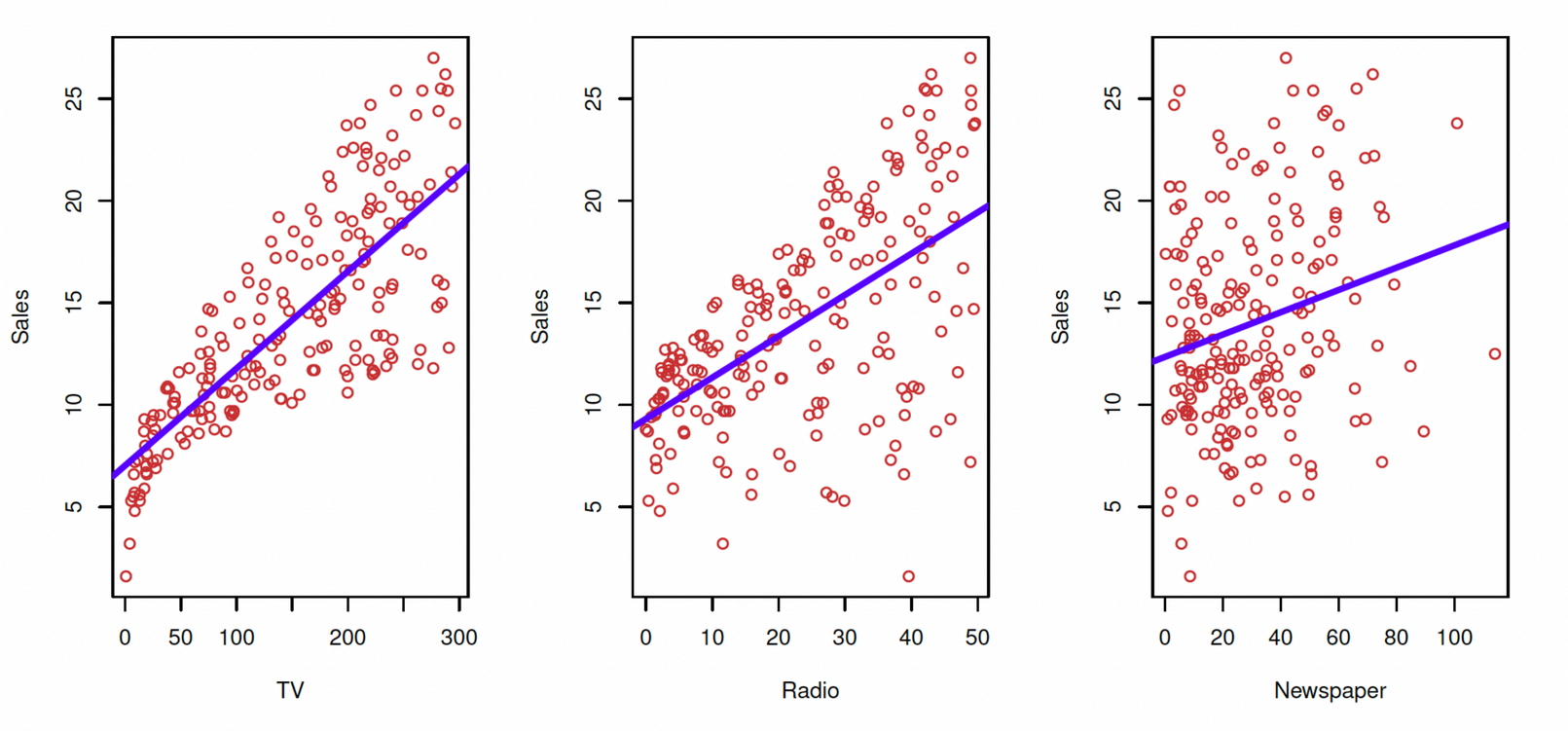
- Linear regression is a simple approach to supervised learning
- It assumes that the dependence of Y on $X_1, X_2, \ldots, X_p$ is linear
- Real systems are rarely linear
- They are also extremely useful in practice



The advertising dataset

Consider the advertising data we introduced last week. Some questions:

- *Is there a relationship between advertising budget and sales?*
- *How strong is the relationship between advertising budget and sales?*
- *Which media contribute to sales?*
- *How accurately can we predict future sales?*
- *Is the relationship linear?*
- *Is there synergy among the advertising media?*


Sales
5
10
15
20
25
0
50
100
200
300
TV
Sales
5
10
15
20
25
0
10
20
30
40
50
Radio
Sales
5
10
15
20
25
0
20
40
60
80
100
Newspaper

Simple linear regression using a single predictor X

- Assume a model of the form $Y = \beta_0 + \beta_1 X + \epsilon$
- β_0 represents the *intercept* and β_1 the *slope.*
- They are also referred to as the *coefficients* or *parameters* of the model
- Given estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ we predict Y (sales) as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- where $\hat{y}$ indicates a prediction of Y based on $X = x$
- The hat symbol denotes an estimated value

Estimating parameters

The least squares method

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i^{th} value of X
- Then $e_i = y_i - \hat{y}_i$ represents the i^{th} $residual$
- We can then define the residual sum of squares (RSS) as,

$$\text{RRS} = e_1^2 + e_2^2 + \ldots + e_n^2$$

Or equivalently as:

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

What we want is to find the parameters that minimise RSS.

Estimating parameters

The least squares method

- We can show, mathematically that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

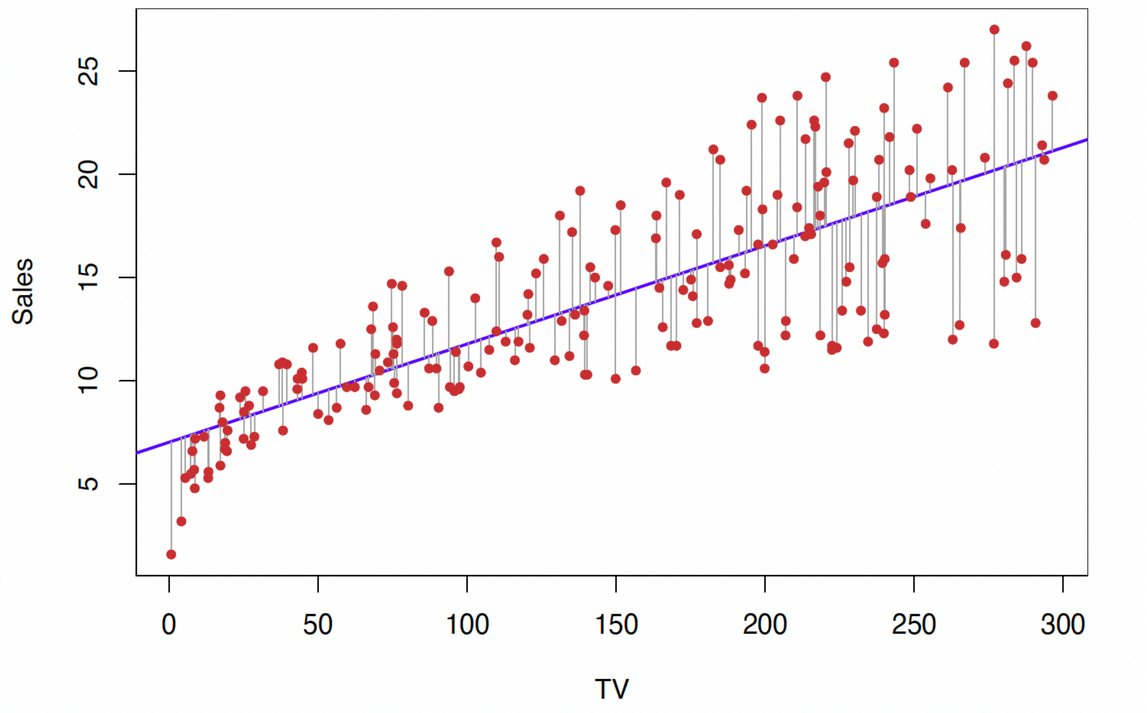
And

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

Where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the *sample means* of our data points)

The advertising data example

Sales predicted by *TV*



The least squares t for the regression of *sales* onto *TV*. In this case, a linear fit captures the essence of the relationship, although it is somewhat deficient on the left of the plot.

Accuracy of the coefficient estimates

The *standard error* of an estimator reflects how it varies under repeated sampling. We have

$$\text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Where $\sigma^2 = Var(\epsilon)$. However, remember, we do not know ϵ.

We will estimate it from the data as the *Residual Standard Error*:

$$\sigma \approx \text{RSE} = \sqrt{\text{RSS}/(n-2)}$$

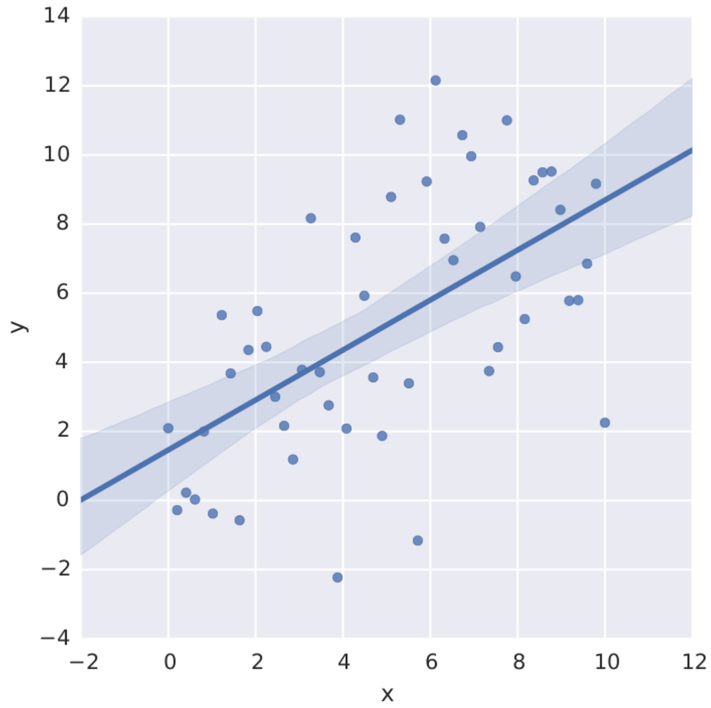
The *standard errors* above allow us to compute confidence intervals

Accuracy of the coefficient estimates

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form,

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

For the advertising data, the 95% confidence interval for β_1 is [0.042, 0.053]



Dummy visual representation of what a confidence interval looks like

Hypothesis testing

Hypothesis testing

Some intuitions

- Standard errors can be used to perform *hypothesis tests* on the coefficients.
- The most common hypothesis test involves testing the *null hypothesis* of

H_0 : There is no relationship between X and Y versus the *alternative hypothesis*

H_A: There is some relationship between X and Y

If H_0 was true, then $\beta_1 = 0$, conversely if H_A was true, then $\beta_1 \neq 0$

Notice that if $\beta_1 = 0$ then the model becomes $Y = \beta_0 + \epsilon$ ...

Hypothesis testing

Some intuitions

- To test the null hypothesis, we compute a t-statistic given by

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

- This will have a t-distribution with $n - 2$ df, assuming $\beta_1 = 0$
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the $p\text{-}value$.

Hypothesis testing

Some intuitions

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Hypothesis testing

Some intuitions

Distribution?

DF?

Test statistic?

Null and alternative hypotheses?

P-value?

😵


IN PRAISE OF ANALYTICS