UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

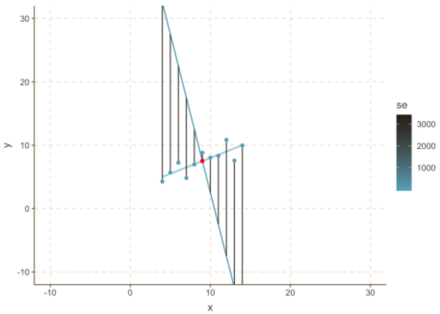
CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Regression and Classification – Part I

Linear and logistic regression

Regression – simple linear regression



$y = \beta_0 + \beta_1 x$

Varying the values β_0 and β_1 obtain different distances of the observations to the curve

A univariate linear regression model is to find a line in the two-dimensional space created by the single independent and dependent variable, which is closest to all data points. A straight line in two dimensions $y = \beta_0 + \beta_1 x$.

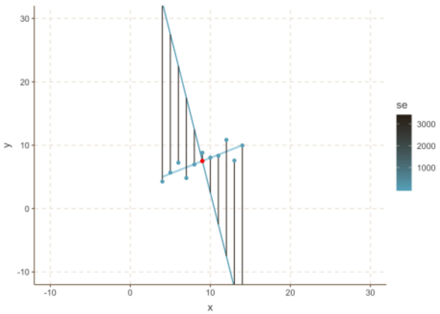
The equation that defines the simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

- The structural part of the model $y = \beta_0 + \beta_1 x$
- The error part of the model ϵ, captures the unexplained behaviour that is not attributed to the independent variable.

The goal of fitting a linear regression model is to estimate the two parameters, the slope β_1 and the intercept β_0, that minimises the overall distance.

Regression – simple linear regression



- Since we estimate $\beta_0$and β_1, we write the estimated regression line

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$$

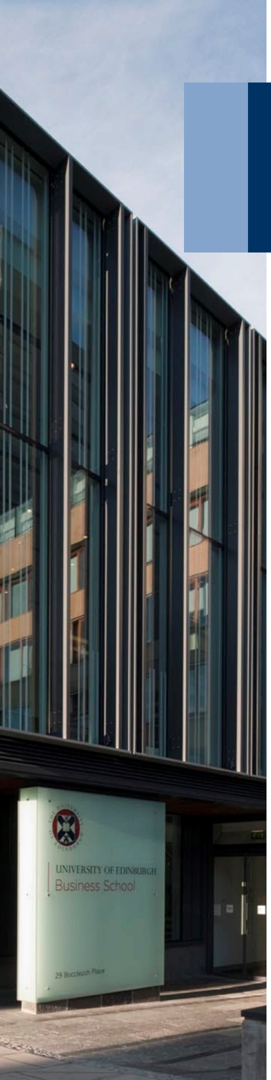
The values $\hat{y}$ of the response variable returned by the estimated regression line are called the predicted (or fitted) values.

In general, $\widehat{y_i}$ differ from the actual observed values for the response (y_i) and the difference between the two are the so called residuals

$$e_i = y_i - \widehat{y_i}$$

We identify as the (estimated) linear relationship between y and x the particular line that fits the data *at best*. The approach usually adopted to find the "best-fit" line is called **the least squares method**.
The idea of least squares: among all possible lines that pass through the points in the scatterplot, the best one is the line that minimizes the sum of squared residuals (which represents a measure of the overall prediction error):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_i\,)\right)^2$$



Regression – simple linear regression

- Example:

If we the relation between supply (s) and demand (d) is

$$d = 10 - 2s$$

And we find in linear regression $\widehat{\beta_0} = 10$ and $\widehat{\beta_1} = 2$, we can predict that a supply of 5 results in a demand of 0. The slope indicates that for every 1 unit increase in supply, the demand will decrease with 2.

- The main: the relationship between the independent and dependent variable is linear.
- One may include higher order relationships.
- In practice, sometimes transformations such as **log transformations** can make the data suitable for finding a linear relationship.

Linear regression

- Please watch the video through this [link](https://media.ed.ac.uk/media/Linear+Regression/1_v9yj8lq9/117185481):

https://media.ed.ac.uk/media/Linear+Regression/1_v9yj8lq9/117185481

Regression – Logistic Regression

- Logistic regression is used for binary classification, meaning its output predictions $y \in \{0,1\}$. It models the probability of the occurrence of an event.
- Assume the chance of an event occurring is p, consider the ratio between the event occurring over the event not occurring: $\frac{p}{1-p}$, also called the odds.
- The **logit transformation** of p is the logarithm of the odds for the event

$$log(\frac{p}{1-p})$$

- **The logistic regression model** models the log odds of the event occurs (y=1) as a linear function of the predictors

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

typically the natural log, ln, is used.

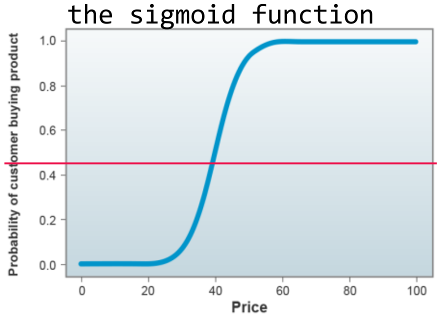
Regression – Logistic Regression

- Consider calculate whether someone is going to buy a product given a particular price x. We would calculate the log odds of it being bought $log(p/(1-p)) = \beta_0 + \beta_1 x$
- To calculate the actual probability of a customer buying the product

$$\frac{p}{1-p} = \ln(e^{\beta_0 + \beta_1 x})$$

- So

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



➔ the **sigmoid** function. It moves between 0 and 1.

- A decision rule is required: e.g. >0.5 ➔ the event happens
- Finding the best parameters β_0 *and* β_1 means maximising/optimising the log odds of the function.

  If $\beta_1 > 0$ ➔ **the log odds** increase with an increase of x; then the probability increases with x (as $e^{-\beta_1 x}$ is larger than 1 and subtracts of the denominator).

Now please study the file:
18 - modelling_linear_and_logistic_regression.ipynb

Activity: Linear and logistic regression

Quiz

Q1. Which two of the following statements are true?

- A. Linear regression requires numeric inputs and continuous outputs.
- B. Linear regression requires numeric inputs and discrete outputs.
- C. Logistic regression requires numeric inputs and continuous outputs.
- D. Logistic regression requires numeric inputs and continuous outputs.

Quiz

Q2. What makes a logistic regression a classification algorithm?

- A. Using probabilities
- B. Using a decision function
- C. Using a sigmoid function
- D. None of the above

Quiz

Q3. Which of the following statements are correct?

- A. The steeper the slope, the stronger the impact of the independent variable on the dependent
- B. The flatter the slope, the stronger the impact of the independent variable on the dependent
- C. The direction of the relationship is determined by the intercept.
- D. The distance between observations is measured by the intercept.

Quiz

Q4. Why do we use the sigmoid function?

- A. Because it is the best way to model the log transform of odds
- B. Because it has a 0/1 outcome
- C. Because its shape forces outcomes to the edges of its range
- D. None of the above

Research & Discussion

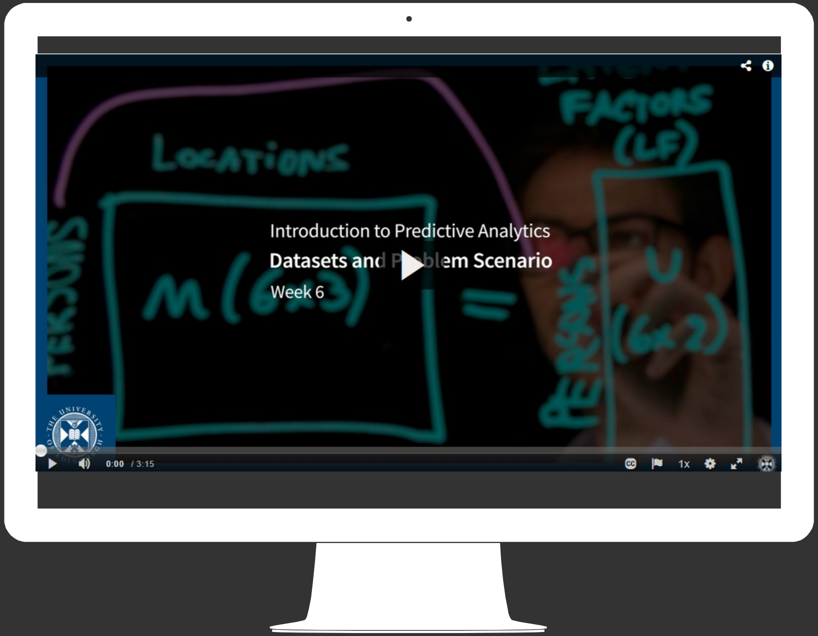
Try to find papers or Real-life applications that discuss linear and logistic regression.

Please summarise your findings into a succinct and descriptive summary which you will share in the discussion area.

Think about the following two questions as you share your thoughts:

- How where the results measured and visualised?
- Did other techniques perform significantly better?

Datasets and problem scenario



- Please watch the video through this link:

https://media.ed.ac.uk/media/Datasets+and+Problem+Scenario/1_1nu58b46/117185481

Activity: Investigate the problem

You will explore two datasets that were covered in the previous video.

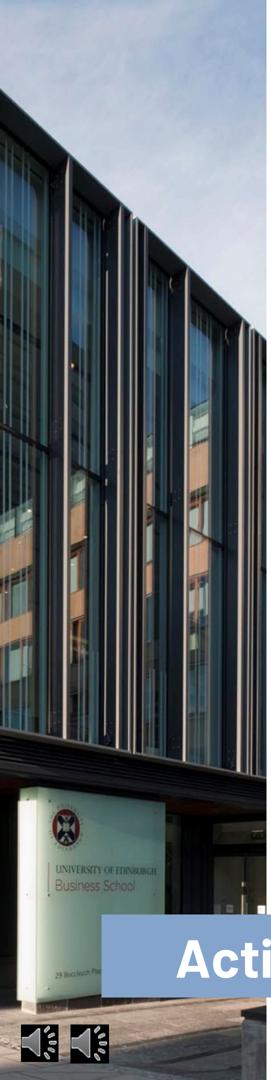
- CS_Purchase_data.csv
- CS_Customer_data.csv

Before start modelling, you should think about the problem description and the following research question: can we obtain a well-performing predictive model that predicts the purchase size of a customer?

Have a first look at the data (in Python, Excel, R, whatever suits you), and check all steps we have performed before:

- pre-processing
- transformation
- normalisation

Brainstorm the various actions that you believe are necessary before modelling can begin. Then, think about what predictive model would be best suited for this.



Homework: Complete the modelling process
You will now implement the complete modelling process in three parts; starting with pre-processing, next transformation and finally modelling.

This activity is all about transforming variables into the correct format, preparing datasets and feeding data into a model to get answers.

1. 19 - A1 - Assessment_pre_proces_complete_the_modelling_process_solution.ipynb + CS_Purchase_data.csv + CS_Customer_data.csv
   ➔ Save your result in a file called 'CS_pre_processed_data.csv', which will be used for the next activity A2
2. 20 - A2 - Assessment_trans_complete_the_modelling_process.ipynb + CS_pre_processed_data.csv (from A1)
   ➔ Save your result in a file called 'CS_transformed_data.csv', which will be used for the next activity A3
3. 21 - A3 - Assessment_model_complete_the_modelling_process.ipynb + CS_transformed_data.csv (from A2)

Activity: More on the datasets

UNIVERSITY OF EDINBURGH
Business School