



COS10022 – DATA SCIENCE PRINCIPLES

Dr Pei-Wei Tsai (Lecturer, Unit Convenor)
ptsai@swin.edu.au, EN508d

SWIN
BUR
NE

SWINBURNE
UNIVERSITY OF
TECHNOLOGY

.
.

Data Science Principles

COS10022

2023 Hawthorn HS1

Presented By Dr Pei-Wei Tsai

7th March 2023



. . .

. . .

.

.

- • • • •
- • • • •

Acknowledgement of Country

We respectfully acknowledge the Wurundjeri People of the Kulin Nation, who are the Traditional Owners of the land on which Swinburne's Australian campuses are located in Melbourne's east and outer-east, and pay our respect to their Elders past, present and emerging.

We are honoured to recognise our connection to Wurundjeri Country, history, culture, and spirituality through these locations, and strive to ensure that we operate in a manner that respects and honours the Elders and Ancestors of these lands.

We also respectfully acknowledge Swinburne's Aboriginal and Torres Strait Islander staff, students, alumni, partners and visitors.

We also acknowledge and respect the Traditional Owners of lands across Australia, their Elders, Ancestors, cultures, and heritage, and recognise the continuing sovereignties of all Aboriginal and Torres Strait Islander Nations.

- •
- •

- • • • • • • • • • • • • •
- • • • • • • • • • • • • •

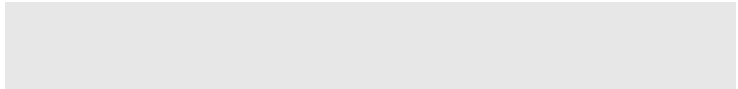


.
.
.

Linear Regression

Case Study

.
.
.
.
.
.
.



Revisit

Linear Regression Model

Source: <http://onlinestatbook.com/2/regression/intro.html>

How did we calculate the previous Linear Regression equation in the first place?

Five statistics are required:


mean of X: μ_x

mean of Y: μ_y

standard deviation of X: s_x

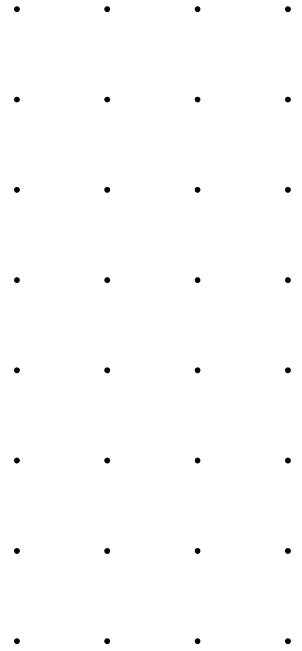
standard deviation of Y: s_y

Pearson's correlation coefficient: r_{xy}


$$\mu_x = \frac{\sum_{i=1}^N x_i}{N}, \text{ where}$$

x_i : an input value

N : the total number of values in a given input variable x



Revisit

Linear Regression Model

Source: <http://onlinestatbook.com/2/regression/intro.html>

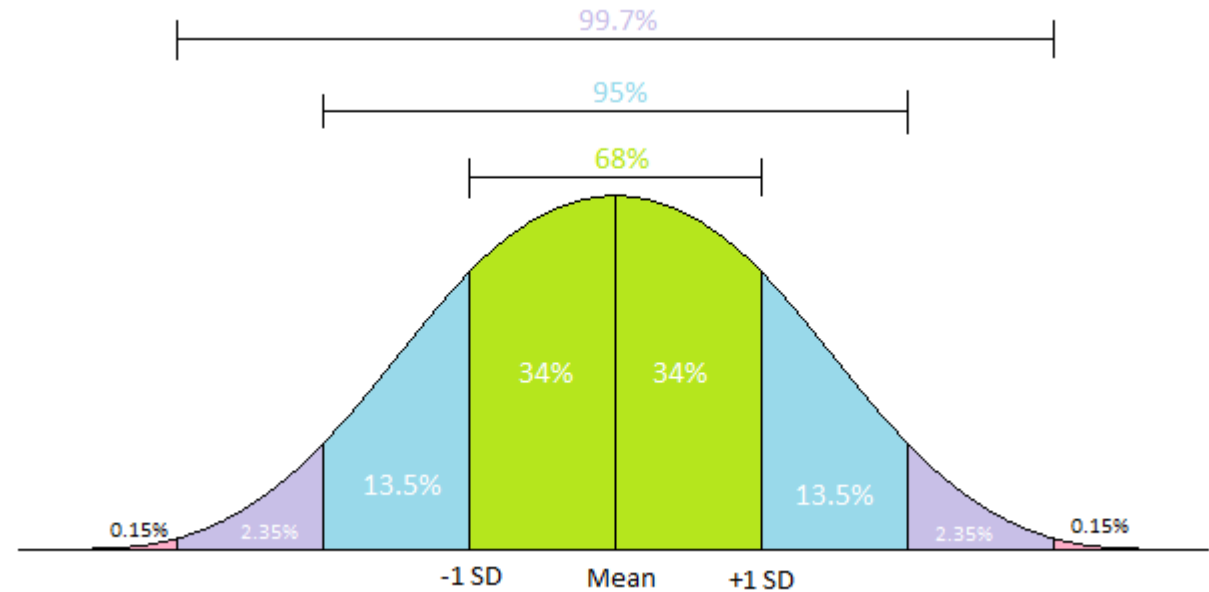
Standard deviation.

$$s_x = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu_x)^2}{N-1}}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^N (y_i - \mu_y)^2}{N-1}}$$

this part of the equation is called the 'sample variance'.

Standard deviation measures how far a set of random numbers are spread out from their average value (mean).



Source: <https://www.biologyforlife.com/standard-deviation.html>

Revisit Linear Regression Model

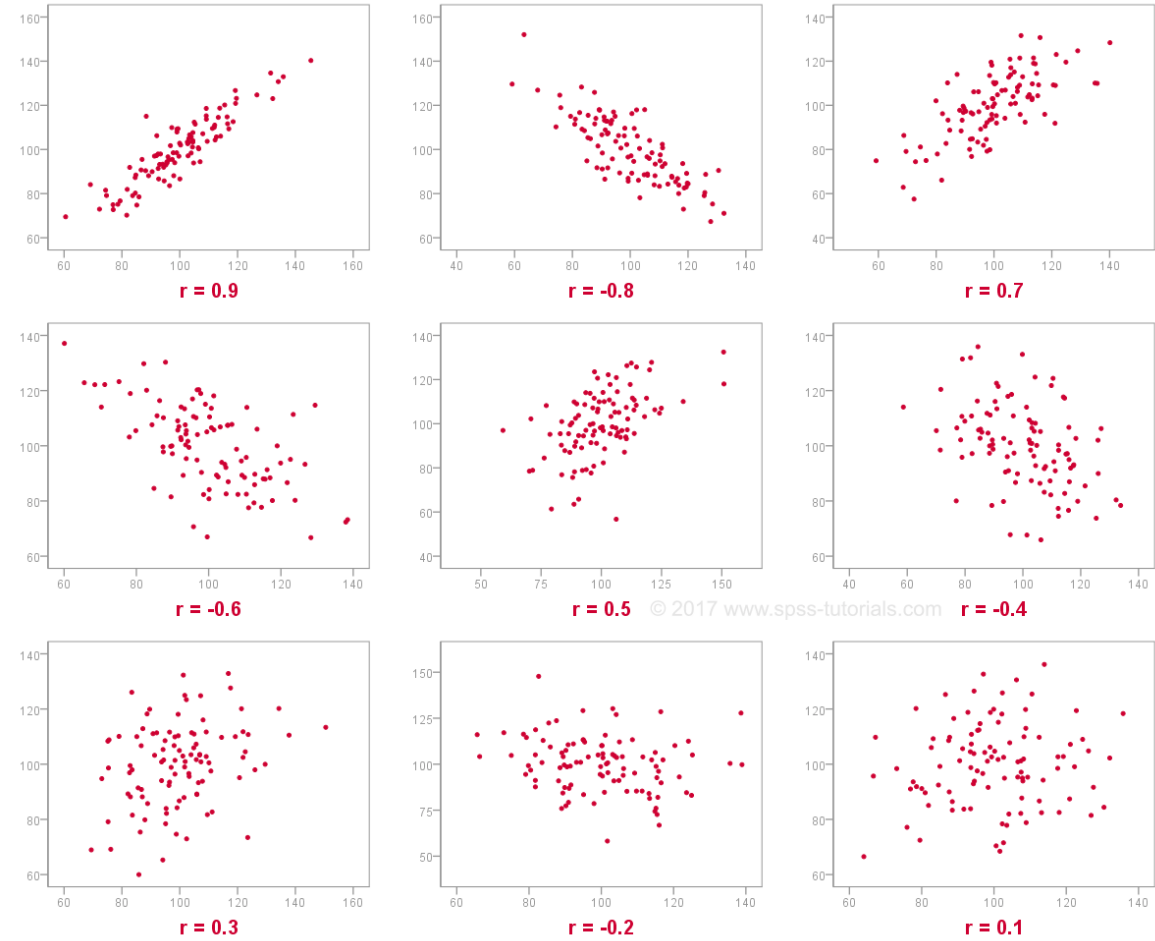
Source: <http://onlinestatbook.com/2/regression/intro.html>

Pearson's correlation coefficient.

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^N (x_i - \mu_x)^2 \cdot \sum_{i=1}^N (y_i - \mu_y)^2}}, \text{ where}$$

N : the total number of data points

Person's correlation coefficient measures the strength of association between two variables.



Source: <https://www.spss-tutorials.com/pearson-correlation-coefficient/>

Revisit

Linear Regression Model – Example

Source: <http://onlinestatbook.com/2/regression/intro.html>

Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

The resulting statistics:

Linear Regression formula.

$$\mu_x = 3.00$$

$$\mu_y = 2.06$$

$$s_x = 1.581$$

$$s_y = 1.072$$

$$r_{xy} = 0.627$$

$$y = 0.785 + 0.425 x$$

$$w_x = r_{xy} \cdot \frac{s_y}{s_x} = 0.425$$

$$w_0 = \mu_y - w_x \mu_x = 2.06 - (0.425)(3)$$



Revisit

Linear Regression Model – Example

Source: <http://onlinestatbook.com/2/regression/intro.html>

Given the following data:

Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

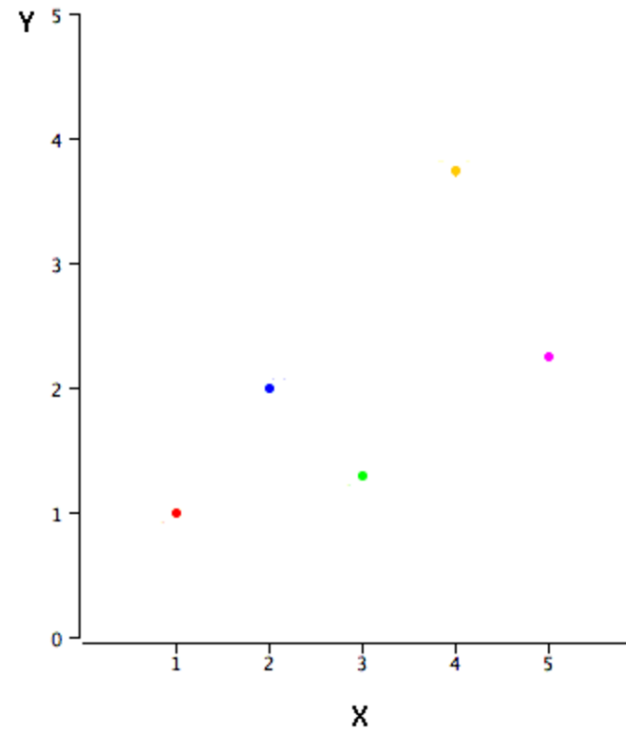


Figure 1. A scatter plot of the example data.

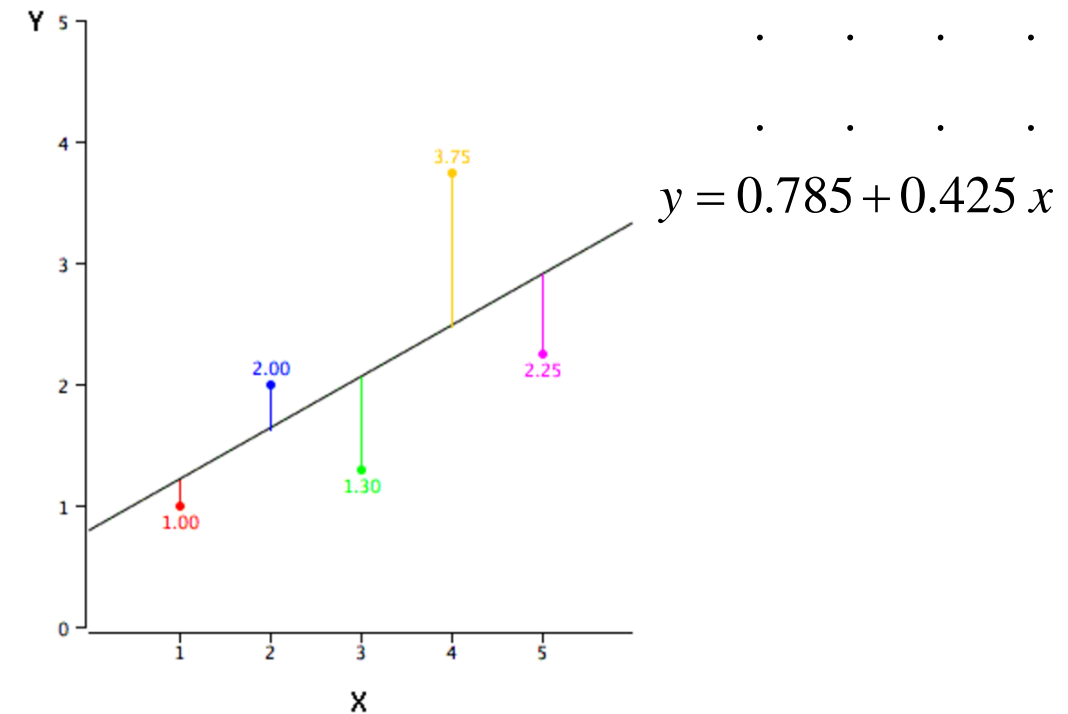
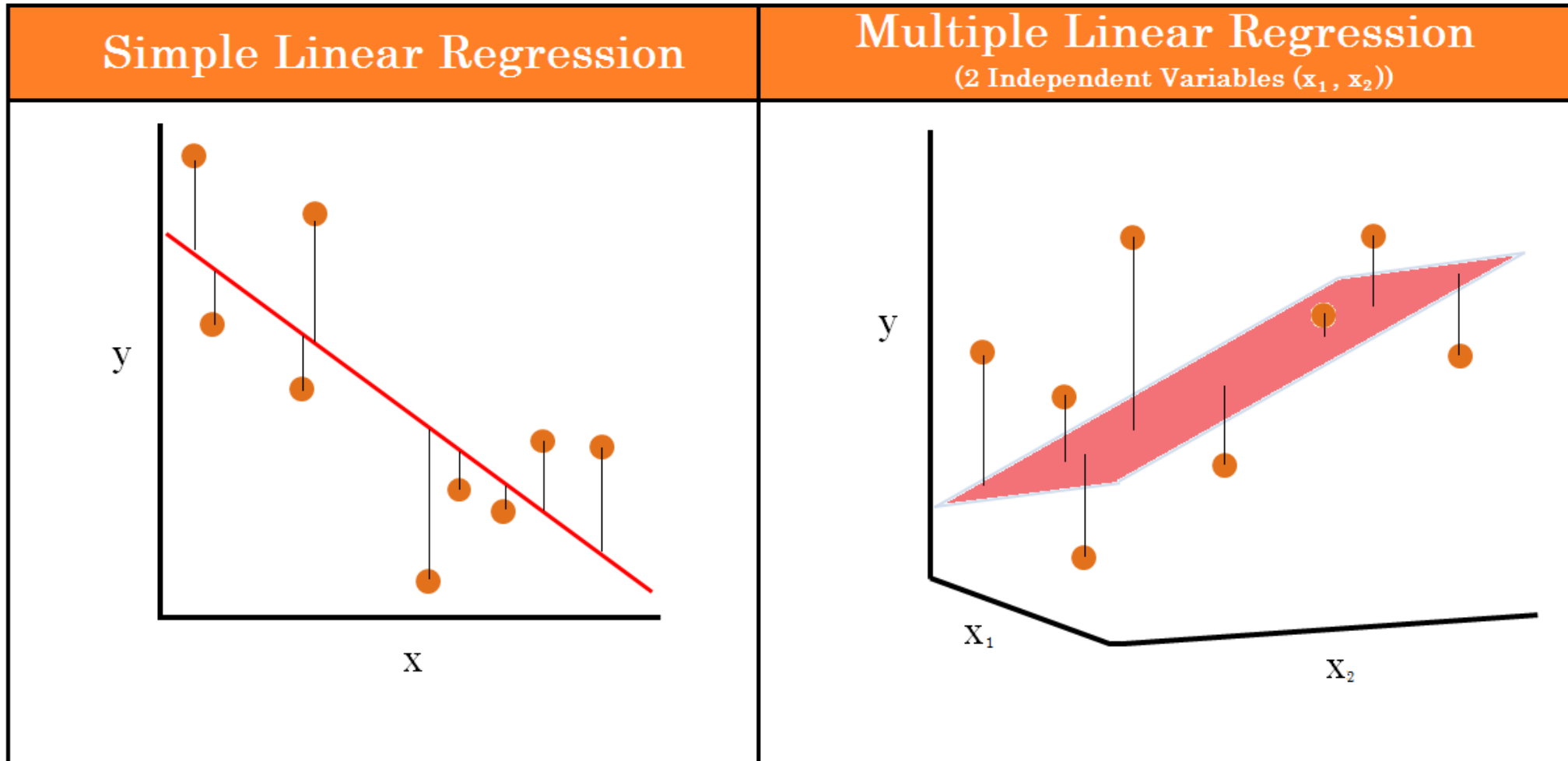


Figure 2. A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

The Change from "Line" to "Plane"



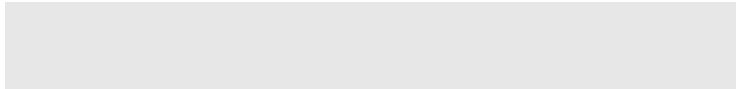
Source: https://www.linkedin.com/pulse/magic-linear-regression-model-bhagyashree-ghosh/?trk=public_post_main-feed-card_feed-article-content

.
.
.

Linear Regression

Case Study

.
.
.
.
.
.
.



.
.
.

Example 1 – Business Revenue Estimation

.
.
.
.
.
.
.

Linear Regression Real Life Example #1

Businesses often use linear regression to understand the relationship between advertising spending and revenue.

The business might fit a simple linear regression model using *advertising spending* as the predictor variable and *revenue* as the response variable.

The regression model would take the following form:

$$R = \beta_0 + \beta_1 \times A$$

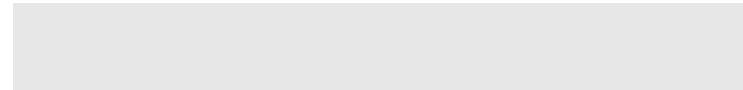
where R stands for the revenue, A is the advertisement spending, β_0 and β_1 are predefined constants.

- β_0 implies the total expected revenue when the advertisement spending is zero.
- β_1 represents the average change in total revenue when the advertisement spending is increased by one unit, e.g., one dollar.
 - If $\beta_1 < 0$, the more you spend on the ad, the less revenue you will get.
 - If $\beta_1 \approx 0$, the advertisement spending has little effect on the revenue.
 - If $\beta_1 > 0$, the more advertisement spending is associated with more revenue.
- Depending on the value of β_1 , a company may decide to adjust the investment made in the advertisement.

.
.
.

Example 2 - Understanding the Relationship between Drug Dosage and Treatment Effects

.
.
.
.
.



Linear Regression Real Life Example #2

Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients.

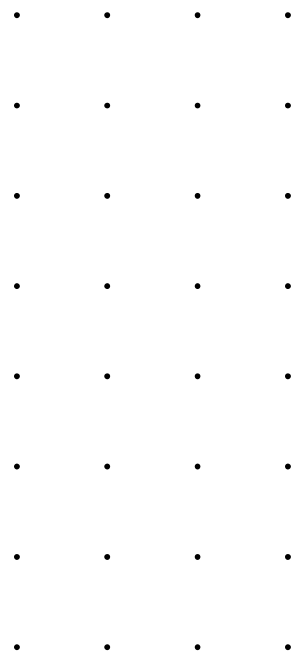
The researchers might administer various dosages of a certain drug to patients and observe how their blood pressure responds. They might fit a simple linear regression model using dosage as the predictor variable and blood pressure as the response variable.

The regression model would take the following form:

$$P = \beta_0 + \beta_1 \times D$$

where P stands for the blood pressure, D is the dosage, β_0 and β_1 are predefined constants.

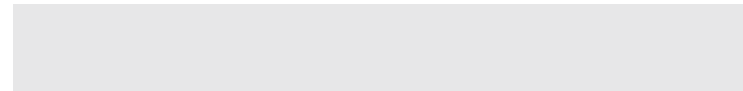
- β_0 stands for the basic blood pressure without any dosage.
- β_1 represents the average change in blood pressure when the dosage is increased by one unit.
 - If $\beta_1 < 0$, the dosage is associated with a decrease in blood pressure.
 - If $\beta_1 \approx 0$, the drug is more like a placebo.
 - If $\beta_1 > 0$, the dosage is associated with an increase in blood pressure.
- Depending on the value of β_1 , researchers may decide to change the dosage given to a patient.



.
.
.

Example 3 - Estimating How Fertilizer and Water Affect on Crop Yields

.
.
.
.
.
.
.



Linear Regression Real Life Example #3

Agricultural scientists often use linear regression to measure the effect of fertilizer and water on crop yields.

The scientists might use different amounts of fertilizer and water on different fields and see how it affects crop yield. They might fit a multiple linear regression model using fertilizer and water as the predictor variables and crop yield as the response variable.

The regression model would take the following form:

$$C = \beta_0 + \beta_1 \times F + \beta_2 \times W$$

where C means the crop yield, F is the amount of fertilizer, W stands for the amount of water, β_0 , β_1 and β_2 are predefined constants.

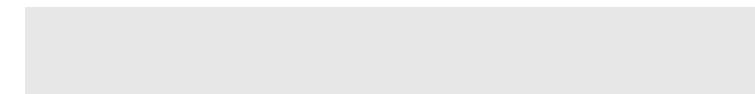
- β_0 stands for the expected crop yield with no fertilizer or water.
- β_1 is the average change in crop yield when fertilizer is increased by one unit, assuming the amount of water remains unchanged.
- β_2 represents the average change in crop yield when water is increased by one unit, assuming the amount of fertilizer remains unchanged.

.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

.
.
.

Example 4 - Estimating the Effect that Different Training Regimens Have on Player Performance

.
.
.
.
.
.



Linear Regression Real Life Example #4

Data scientists for professional sports teams often use linear regression to measure the effect that different training regimens have on player performance.

The data scientists in the NBA team might analyse how different amounts of weekly yoga sessions and weightlifting sessions affect the number of points a player scores.

The regression model would take the following form:

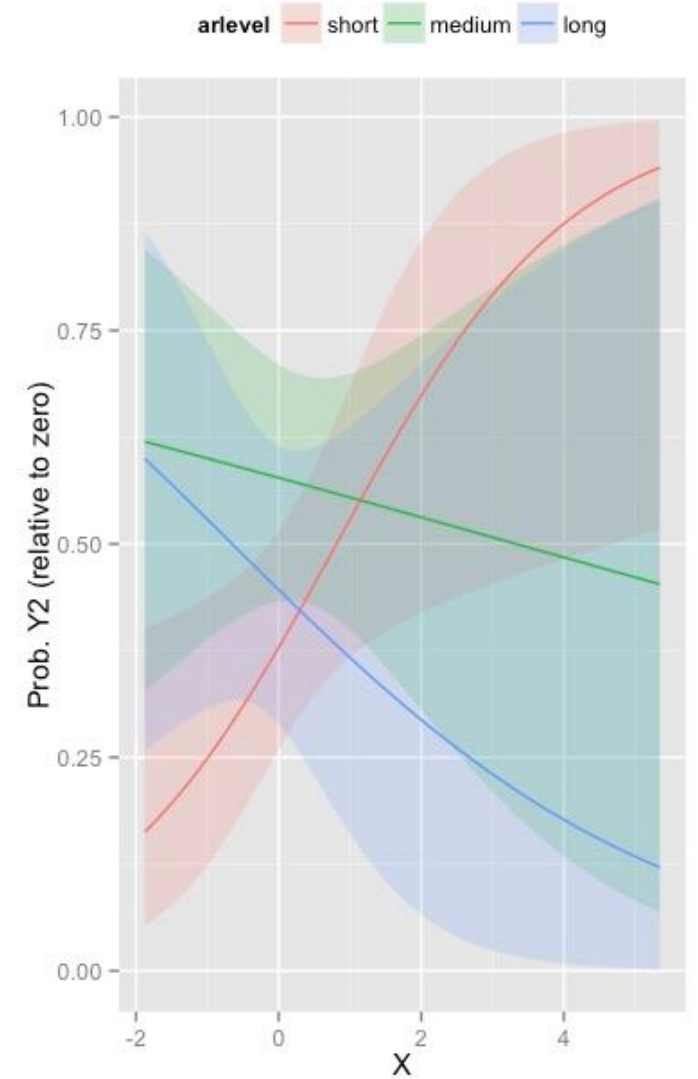
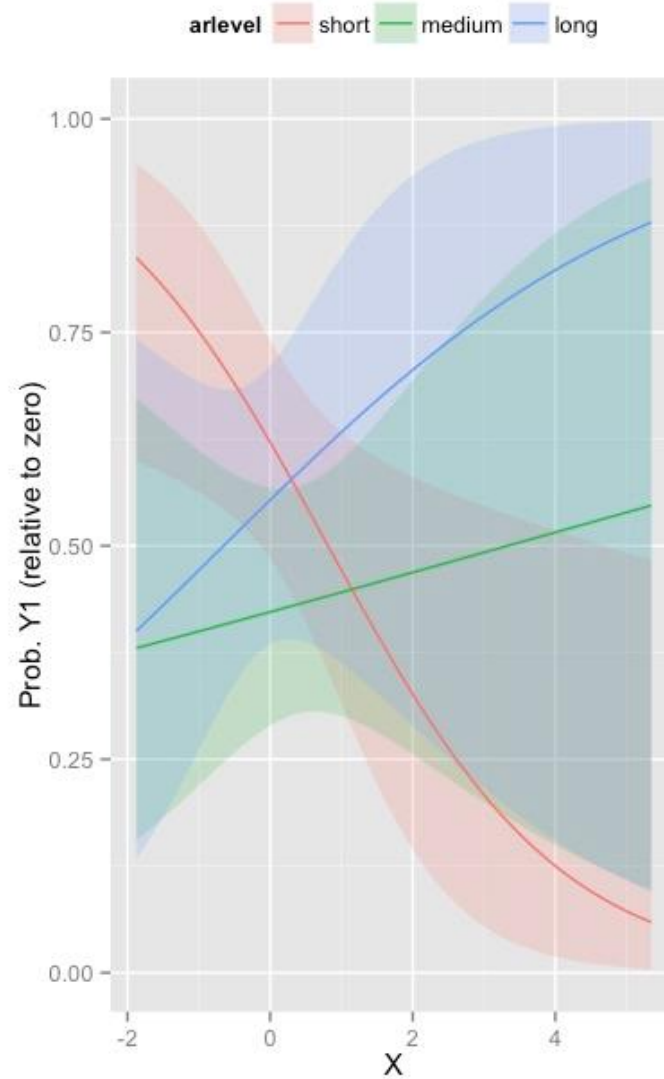
$$S = \beta_0 + \beta_1 \times Y + \beta_2 \times E$$

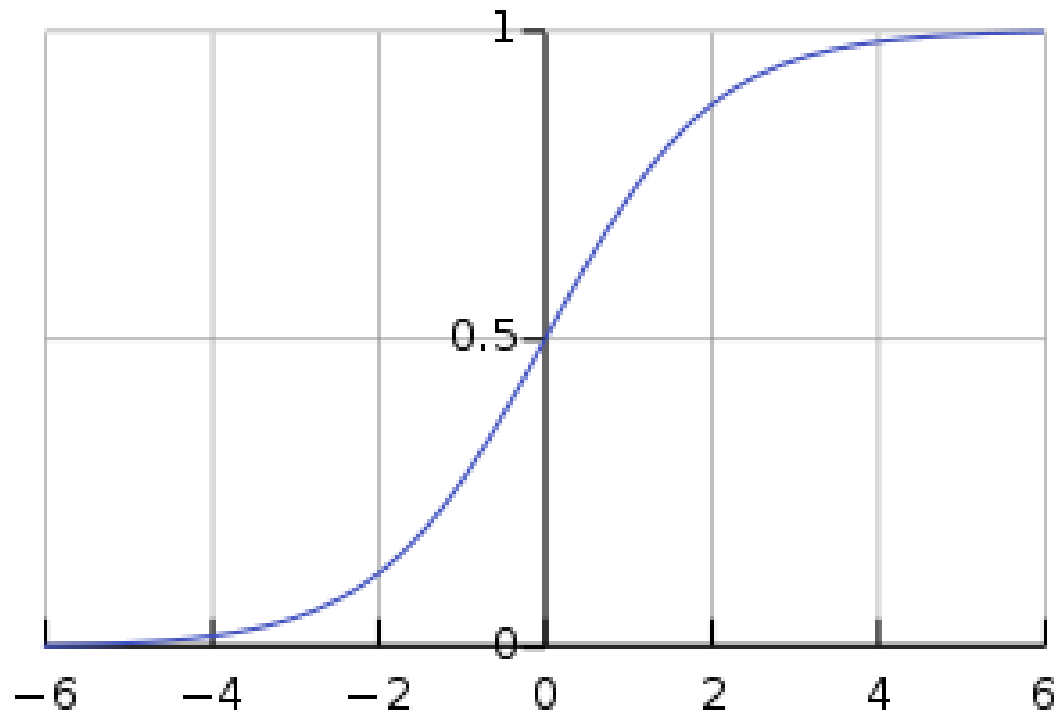
where S is the points scored, Y denotes the yoga sessions, E means the weightlifting sessions, β_0 , β_1 and β_2 are predefined constants.

- β_0 stands for the expected points scored for a player who participates in neither yoga sessions nor weightlifting sessions.
- β_1 is the average change in points scored when weekly yoga sessions is increased by one unit, assuming the number of weekly weightlifting sessions remains unchanged.
- β_2 represents the average change in points scored when weekly weightlifting sessions is increased by one unit, assuming the number of weekly yoga sessions remains the same.
- Depending on the values of β_1 and β_2 , the data scientists may recommend that a player participates in more or less weekly yoga and weightlifting sessions in order to maximize their points scored.

Logistic Regression

- Logistic regression is a special case of regression analysis and is calculated when the dependent variable is nominally or ordinally scaled.





Sigmoid Function

- A sigmoid function only produces output values between 0 and 1.
- $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1 - S(-x)$

Logistic Regression Model

- Let's replace the “x” in the sigmoid function with our linear regression function, then the logistic regression model is obtained.
- A logistic function or logistic curve is a common S-shaped curve (sigmoid curve) with equation

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

where z stands for the linear regression function.

- Thus, a logistic regression function for the model given in the revisit is:

$$f(X) = \frac{1}{1 + e^{-(a_1x_1 + a_2x_2 + \dots + b)}}$$

.
.
.

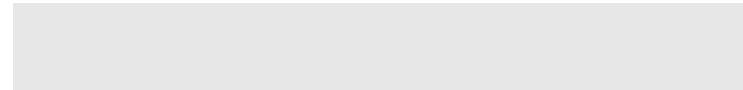
Logistic Regression Case Study

.
.
.
.
.
.
.

.
.
.

Example 1 – Finding Out How Exercise and Weight Impact the Probability of Having a Heart Attack

.
.
.



Logistic Regression Real Life Example #1

Medical researchers want to know how exercise and weight impact the probability of having a heart attack.

To understand the relationship between the predictor variables and the probability of having a heart attack, researchers can perform logistic regression.

The **response variable** in the model will be heart attack and it has two potential outcomes:

- 1) A heart attack occurs.
- 2) A heart attack does not occur.

The results of the model can tell:

- How changes in exercise and weight affect the probability that a given individual has a heart attack.
- The prediction on whether a given individual has a heart attack based on the weight and the time this individual spent exercising.

$$P(H|E, W)$$

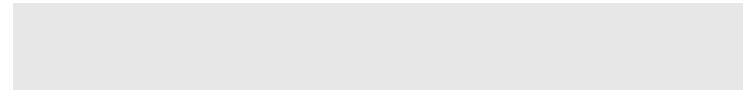


.
.

Example 2 -
Finding Out How GPA, ACT Score
and the Number of AP Classes
Taken Impact the Probability of
Getting into a Particular University

. .
. .
. .
. .

.
.



Logistic Regression Real Life Example #2

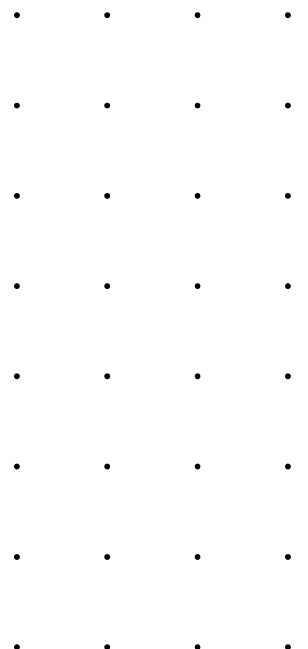
To understand the relationship between the predictor variables and the probability of getting accepted, researchers can perform logistic regression.

- Response variable: *acceptance* with potential outcomes:
 - The application is accepted
 - The application is rejected
- The result of the model tells exactly how changes in GPA, ACT score, and number of AP classes taken affect the probability that a given individual gets accepted into the university.

$$P(U|GPA, ACT, AP)$$

Question:

- Should the model be the same for all universities?



.
.
.

Example 3 – SPAM Detection

.
.
.
.
.
.
.

Logistic Regression Real Life Example #3

A business wants to know whether the word count and the country of origin impact the probability that an email is spam.

To understand the relationship between the predictor variables and the probability of an email being spam, researchers can perform logistic regression.

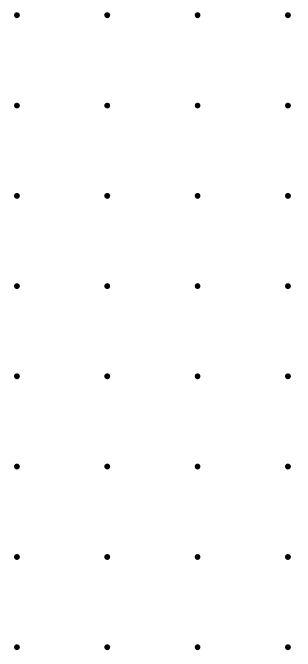
The **response variable** in the model will be “spam” and it has two potential outcomes:

- 1) The email is spam.
- 2) The email is not spam.

The results of the model can tell the business exactly how changes in word count and the country of origin affect the probability of a given email being spam.

The business can also use the fitted logistic regression model to predict the probability that a given email is spam, based on its word count and country of origin.

$$P(S|W, C)$$



.
.
.

Example 4 – Fraudulent Transaction Detection

. . . .
. . . .
.
.
.
.

Logistic Regression Real Life Example #4

The credit card company wants to know whether transaction amount and credit score impact the probability of a given transaction being fraudulent.

To understand the relationship between these two predictor variables and the probability of a transaction being fraudulent, the company can perform logistic regression.

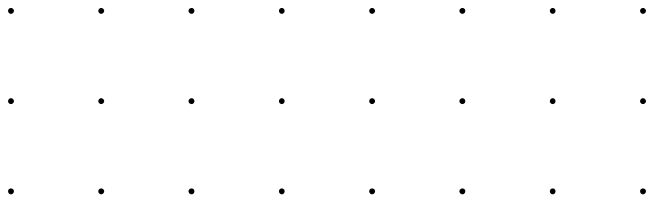
The **response variable** in the model will be “fraudulent” and it has two potential outcomes:

- 1) The transaction is fraudulent.
- 2) The transaction is not fraudulent.

The results of the model will tell the company exactly how changes in transaction amount and credit score affect the probability of a given transaction being fraudulent. The company can also use the fitted logistic regression model to predict the probability that a given transaction is fraudulent, based on the transaction amount and the credit score of the individual who made the transaction.

$$P(T|C,S)$$





Thank you

17th February 2023

