Linear Regression

Regression

- Regression focuses on the relationship between an outcome and its input variables.
  - Provides an estimate of the outcome based on the input values.
  - Models how changes in the input variables affect the outcome.
- The outcome can be continuous or discrete.
- Possible use cases:
  - Estimate the lifetime value (LTV) of a customer and understand what influences LTV.
  - Estimate the probability that a loan will default and understand what leads to default.
- **approaches: linear regression and logistic regression**

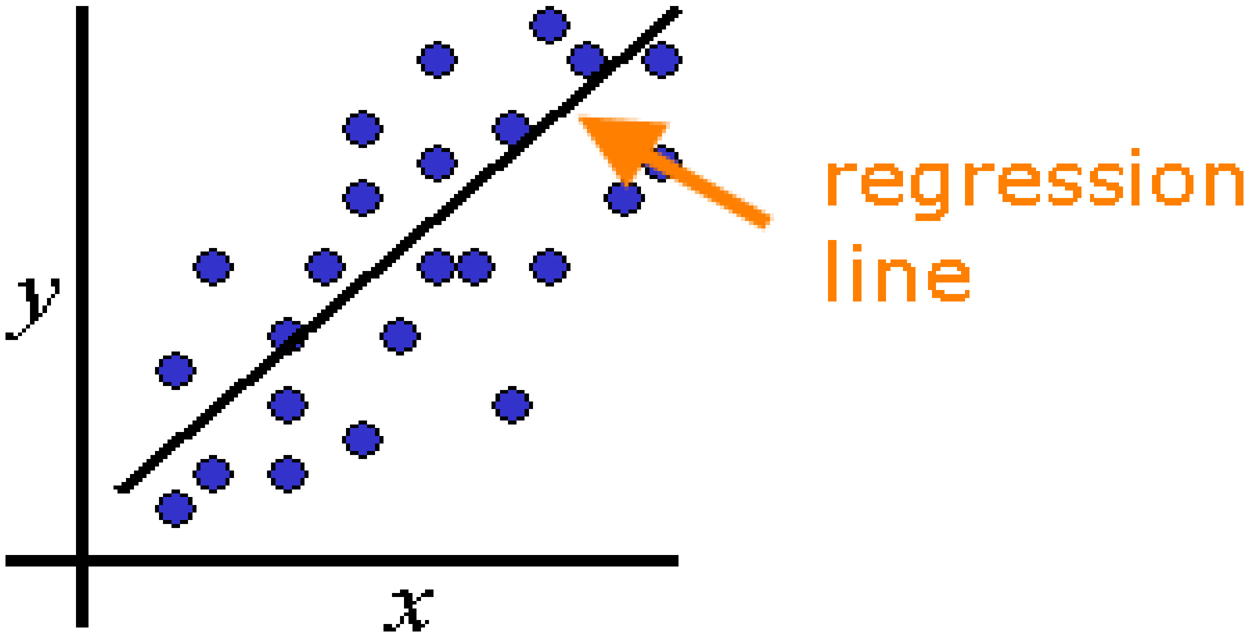
Linear Regression

- Used to estimate a continuous value as a linear function of other variables
  - Income as a function of years of education, age, and gender
  - House sales price as function of area, number of bedrooms/bathrooms, and lot size
- Outcome variable is continuous.
- Input variables can be continuous or discrete.
- Model Output:
  - A set of estimated **coefficients** that indicate the relative impact of each input variable on the outcome
  - A linear expression for estimating the outcome as a function of input variables

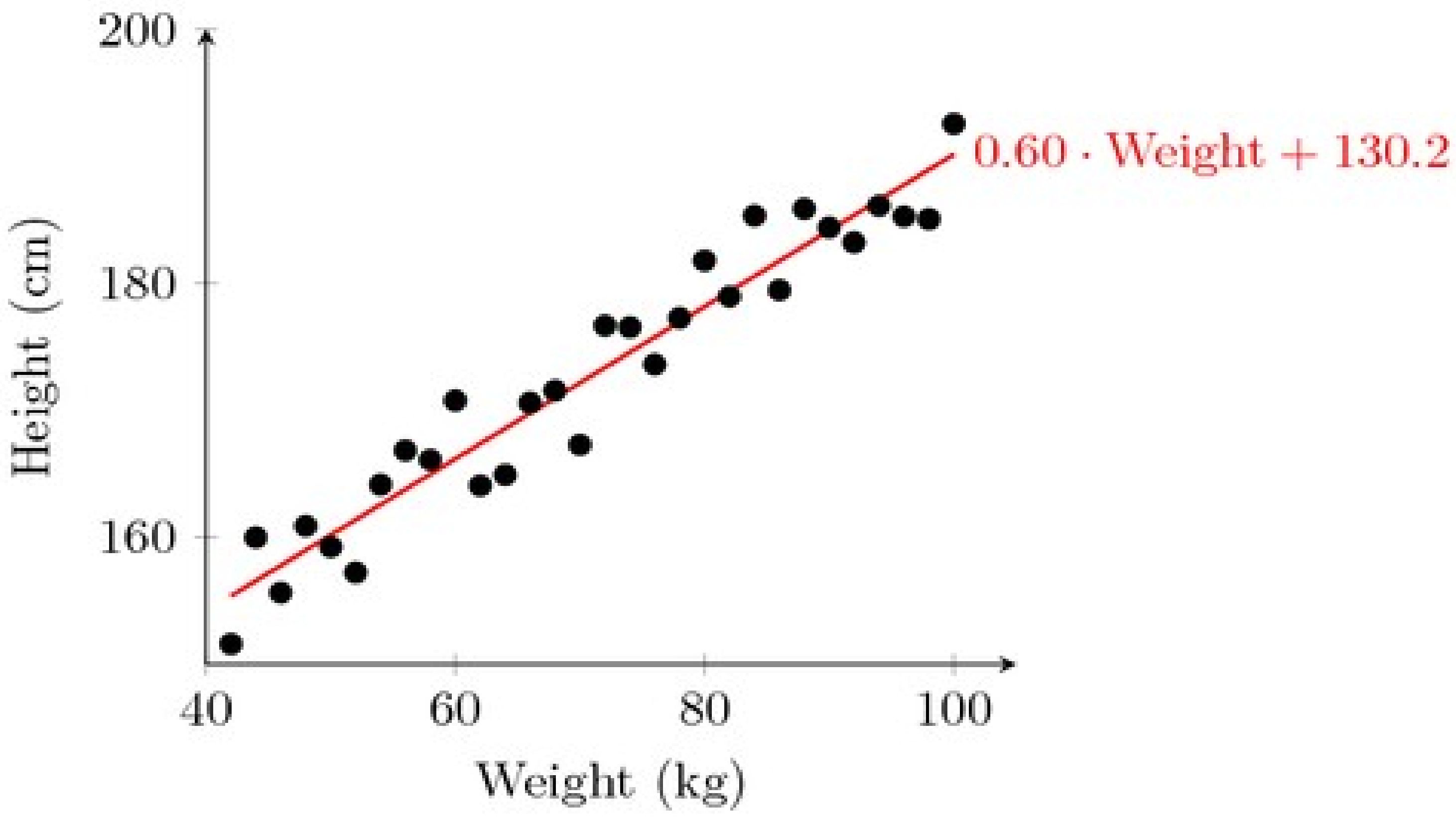
Linear Regression other examples

- Banks assess the risk of home-loan Applicants based on their:
  - age,
  - income,
  - expenses,
  - occupation,
  - number of dependents,
  - Personal status
  - total credit limit, etc.

Regression Line



Regression Line, example ….Height Vs. Weight



Use Cases

- Linear regression is often used in business, government and other scenarios
- Some common practical application
  - **Real Estate:** A simple linear regression analysis can be used to model residential home prices as a function of the home's living area. Such a model helps set or evaluate the list price of home on the market.
  - **Demand forecasting**: Business and government can use linear regression models to predict demand for goods and services. For instance, a chain restaurant can predict the quantity of food that customer ask for based on the weather, the day of the week, the time, etc.
  - **Medical:** A linear regression model can be used to analyze the effect of a proposed radiation treatment on tumor size. Input variables might include duration of a single radiation treatment, frequency of radiation treatment, and patient attributes such as age and weight.

Regression analysis

- Linear Regression is a predictive analytical technique used to model the relationship between several input variables (X) and a continuous outcome variable (Y)
  - The input variable are known as independent variables and the outcome variable is known as a dependent variable
- The aim is to establish a linear relationship (a mathematical formula) between the independent/predictor variables (X) and the outcome/dependent variable (Y), so that, we can use this formula to estimate the value of the response Y, when only the predictors (Xs) values are known.

Example #1

- Suppose that we got **java1** and **java2** scores from the previous term for five students:
- Let's see if we can build a model to predict, what score that a new student can score based on her mark in java1?

	Java1	Java2
Noura	95	85
Asma	85	95
Fatima	80	70
Salma	70	65
Hind	60	70

Example #1

	Java1 (predictor)	Java2 (response)
Noura	95	85
Asma	85	95
Fatima	80	70
Salma	70	65
Hind	60	70

- We want to build a linear model such that:
  - Y = c0 + c1*x1
    - Y represents the predicted mark for Java2
    - x1: Java1
    - c0 and c1 are coefficients

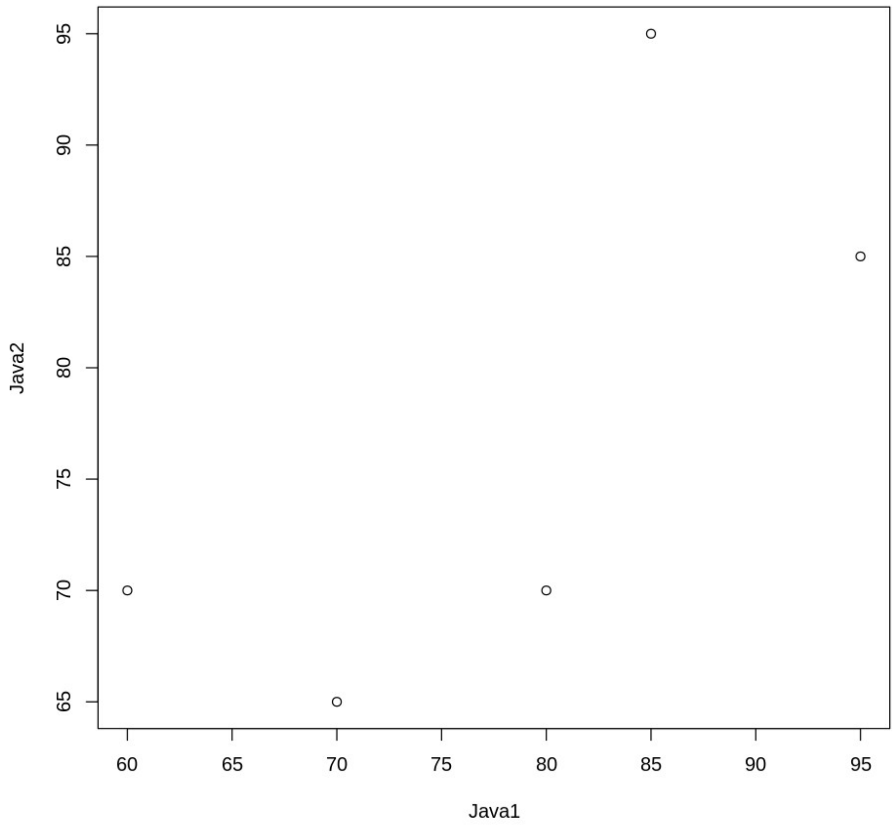
Example #1

Lets use R to Plot these values:

```
# Regression

Java1 <- c(95, 85, 80, 70, 60)
Java2 <- c(85, 95, 70, 65, 70)

plot(Java2, Java1)
```



Example #1

we use **lm()** to build a linear regression model in R

```
fit <- lm(Java2 ~ Java1)
fit
```

Example #1

Call: lm(formula = Java2 ~ Java1)
Coefficients:

Intercept: 26.7808 (This is c0)
Java1: 0.6438 (This is c1)

According to the equation:

$$Y = c0 + c1*x1$$

$$Java2 = 26.7808 + 0.6438*Java1$$

Example #1

$$Java2 = 26.7808 + 0.6438*Java1$$

If Mariam scored 80 in Java1, what probably she can get in Java2 ?

Example #1

Predict what score a student can get in Java2 if she got 80 in Java1

We can do it in R:

```
(java2_ <- fit$coefficients[[1]] + fit$coefficients[[2]]*80)
```

Is it 78.28 ?

Example #2

Marks in Java1 : {95, 85, 80, 70, 60, 80, 75, 90, 88}

Marks in Java2 : {85, 95, 70, 65, 70, 78, 72, 88, 91}

Marks in swEng: {82, 89, 70, 72, 75, 77, 75, 89, 89}

Activity

- Use linear regression to build a model to predict score in Software Engineering (swEng)
- Plot the marks in 3D
- Very good source for 3D plot:
- https://www.r-bloggers.com/getting-fancy-with-3-d-scatterplots/

Example #2

```
# 1. read marks
Java1 <- c(95, 85, 80, 70, 60, 80, 75, 90, 88)
Java2 <- c(85, 95, 70, 65, 70, 78, 72, 88, 91)
swEng <- c(82, 89, 70, 72, 75, 77, 75, 89, 89)

#2. plot
library(scatterplot3d)
 scatterplot3d(Java1,   # x axis
                         Java2,   # y axis
                         swEng,   # z axis
                         main="3-D Scatterplot Example 1")
```

Example #2

```
# 3. predict the mark in software engineering based on the marks in java1 and java2
fit2 <- lm(swEng ~ Java1 + Java2)
fit2


# 4. visualize the model
library(visreg)
visreg(fit2)
```

Example #3

- Linear regression is demonstrated below with function **lm()** on the Australian CPI **(Consumer Price Index)** data, which are quarterly CPIs from 2008 to 2010
- At first, the data is created and plotted.
- In the code below, an x-axis is added manually with function **axis()**, where **las=3** makes text vertical.

*A **consumer price index** (CPI) measures changes in the **price** level of a market basket of **consumer** goods and services purchased by households.*

Example #3

```
# each year 4 quarters
# rep() function replicates values
year <- rep(2008:2010, each=4)
quarter <- rep(1:4, 3) # 3 years / 4 Q

cpi <- c(162.2, 164.6, 166.5, 166.0,
        + 166.2, 167.0, 168.6, 169.5,
        + 171.0, 172.1, 173.3, 174.0)
```

Example #3

- We then check the correlation between CPI and the other variables, year and quarter.

```
cor(year,cpi)
[1] 0.9096316

cor(quarter,cpi)
[1] 0.3738028
```

Example #3

- Then a linear regression model is built with function **lm()** on the above data:
  - Predictors: year and quarter
  - Response: CPI

```
fit <- lm(cpi ~ year + quarter)
fit
```

Example #3

- With the generated linear model in the previous slide, CPI is calculated as:

$$cpi = c0 + c1*year + c2*quarter$$

- where c0, c1 and c2 are coefficients from model fit. Therefore, the CPIs in 2011 can be get as follows.

```
(cpi2011 <- fit$coefficients[[1]] + fit$coefficients[[2]]*2011 + fit$coefficients[[3]]*(1:4))
```

Predict new inputs

```
data2011 <- data.frame(year=2011, quarter=1:4)
cpi2011 <- predict(fit, newdata=data2011)
```

Example #4

- Returning to the *Income* example, in addition to the variables age and education, the person’s gender, female or male, is considered an input variable. The following code reads a comma-separated-value (CSV) file of 400 people’s incomes, ages, years of education, and gender. The first 10 rows are displayed:

- income_input = read.csv("c:/R Codes/Ch6/income.csv")
- income_input[1:10,]

Example #4

- Each person in the sample has been assigned an identification number, *ID*. *Income* is expressed in thousands of dollars. (For example, 113 denotes $113,000.)
- As described earlier, *Age* and *Education* are expressed in years. For *Gender*, a 0 denotes male and a 1 denotes female.
- A summary of the imported data
- summary(income_input)

Example #4

- Using the linear model function, lm(), in R, the income model can be applied to the data as follows:
- results <- lm(Income~Age + Education + Gender, income_input)
- summary(results)
- The following R code provides the modified model results:
- results2 <- lm(Income ~ Age + Education, income_input)
- summary(results2)