Correlation and Regression

Correlation, Scatter Plots, Linear Correlation Coefficient

Regression, Measure of Variation in Regression, Coefficient of Determination

Three Excel Files

Instructor: Qasim Ali

Correlation

- **Correlation:** Relationship between two variables.
- A relation is called linear if a straight line can represent the trend in the data.

Excel Example

Correlation vs Covariance

Calculate the correlation of two datasets

Covariance formula is $\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$

- Multiply the differences from the mean for each value pair and find the sum
- Divide that total by the number of data pairs

Correlation formula is $\frac{\Sigma(x-\bar{x})(y-\bar{y}}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}$

- Top term is the same
- Divided by the term shown

Interpreting the Correlation Values

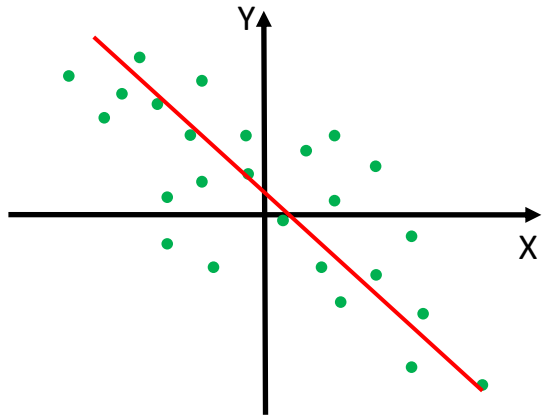
- **Data** that is negatively correlated: $-1 \leq r < 0$
- **Data** that is completely uncorrelated: 0
- **Data** that is positively correlated: $0 < r \leq 1$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

Interpreting the Correlation Values

Negatively Correlated Data

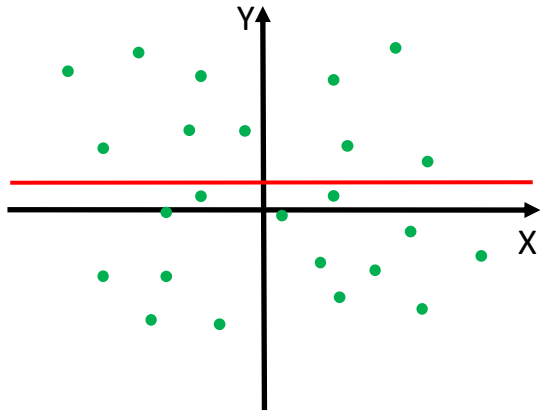
- **Data** that is negatively correlated: $-1 \leq x < 0$



Negative correlation

Interpreting the Correlation Values

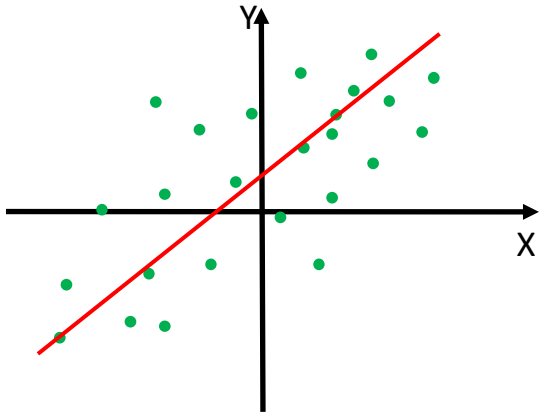
- **Data** that is completely uncorrelated: 0



No correlation

Interpreting the Correlation Values

- **Data** that is positively correlated: $0 < x \leq 1$



Positive correlation

Is my Correlation Significant

- **Depends on few factors**
  - **Number of Measurements**
  - **Whether the value can be positive or negative**
    - Values are only positive or only negative: one-tailed
    - Values are either positive or negative: two-tailed
  - **Look at the correlation value up in a table**
    - Correlation value decreases as sample size increases

Correlation Lookup Table (Two-Tailed)

N	0.1	0.05	0.02	0.01	0.001
5	0.80	0.88	0.93	0.96	0.99
6	0.73	0.81	0.88	0.92	0.97
7	0.67	0.75	0.83	0.87	0.95
8	0.62	0.71	0.79	0.83	0.93
9	0.58	0.67	0.75	0.80	0.90
10	0.55	0.63	0.71	0.77	0.87
15	0.44	0.51	0.59	0.64	0.76
20	0.38	0.44	0.52	0.56	0.68
30	0.31	0.36	0.42	0.46	0.57

Correlation
Practice Problems

Example (Samara Lighting sales and advertising): Samara Office Lighting Store knows the importance of proper illumination. During COVID-19 pandemic, a lot of people are working from home and very picky when it comes to interior lighting. **The store** has been among the few economic success stories.

Julia Kim, Samara Lighting's owner, understands that her company's sales depend on many factors including economic conditions, interest rates, fashion trends, etc. Most of the factors are beyond her control, so she decides to focus on one factor: advertising. Julia collects the annual advertising and sales figures for the last 14 years and wants to analyze how closely annual sales (in $ millions) are related to annual advertising expenses (in $ thousands).

Data

Year	Annual Advertising Expenditure	Annual Sales		Year	Annual Advertising Expenditure	Annual Sales
2007	44	12.1		**2014**	92	27.2
2008	81	19.8		**2015**	104	26.6
2009	99	21.9		**2016**	103	24.1
2010	102	24.0		**2017**	121	30.2
2011	56	15.4		**2018**	101	27.9
2012	77	21.7		**2019**	115	31.3
2013	80	23.3		**2020**	132	34.5

Excel Example

Correlation

In the following examples, please identify the type of data.

- Correlation and Causation are the same

Answer: FALSE

Causation manipulate independent variable and observe effect on dependent variable

- How the values look like in one-tailed test?

Answer: Either only positive or only negative

- Correlation tells you if there is an association between *x* and *y*

Answer: TRUE

- Correlation describe the relationship or allow you to predict one variable from the other.

Answer: FALSE

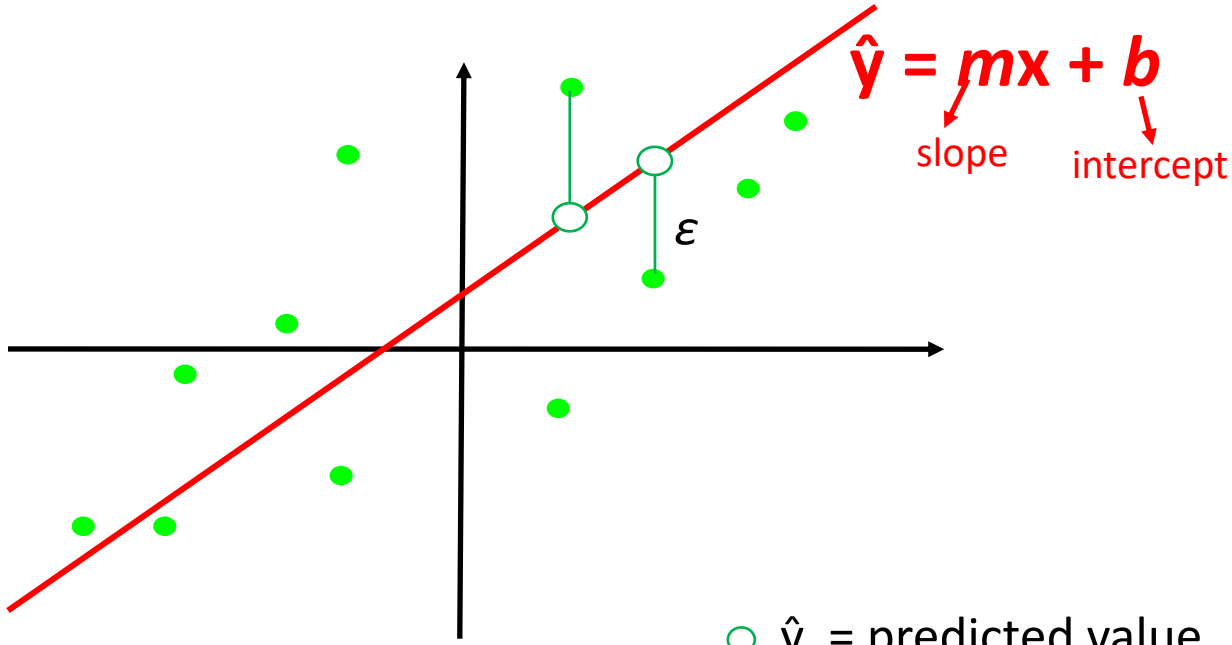
Regression

Best Fit

- Correlation tells you if there is an association between x and y but it doesn't describe the relationship or allow you to predict one variable from the other.

- Aim of linear regression is to fit a straight line, $\hat{y} = mx + b$, to data that gives best prediction of y for any value of x.

- This will be the line that minimizes distance between data and fitted line, i.e. the residuals



Regression

- To find the best line we must minimize the sum of the squares of the residuals (the vertical distances from the data points to our line)

  Model line: $\hat{y} = mx + b$ m = slope, b = intercept

  Residual (ε) = $y - \hat{y}$

  Sum of squares of residuals = $\Sigma (y - \hat{y})^2$

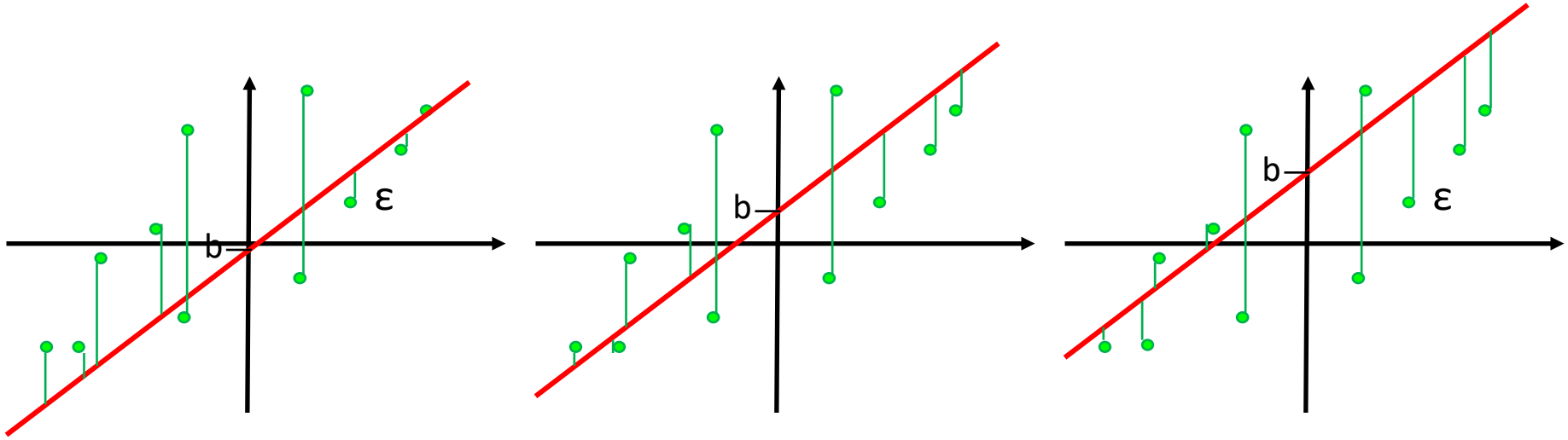
- We must find values of m and b that minimize

  $\Sigma (y - \hat{y})^2$

Regression

How to find b?

- First we find the value of *b* that gives the minimum sum of squares

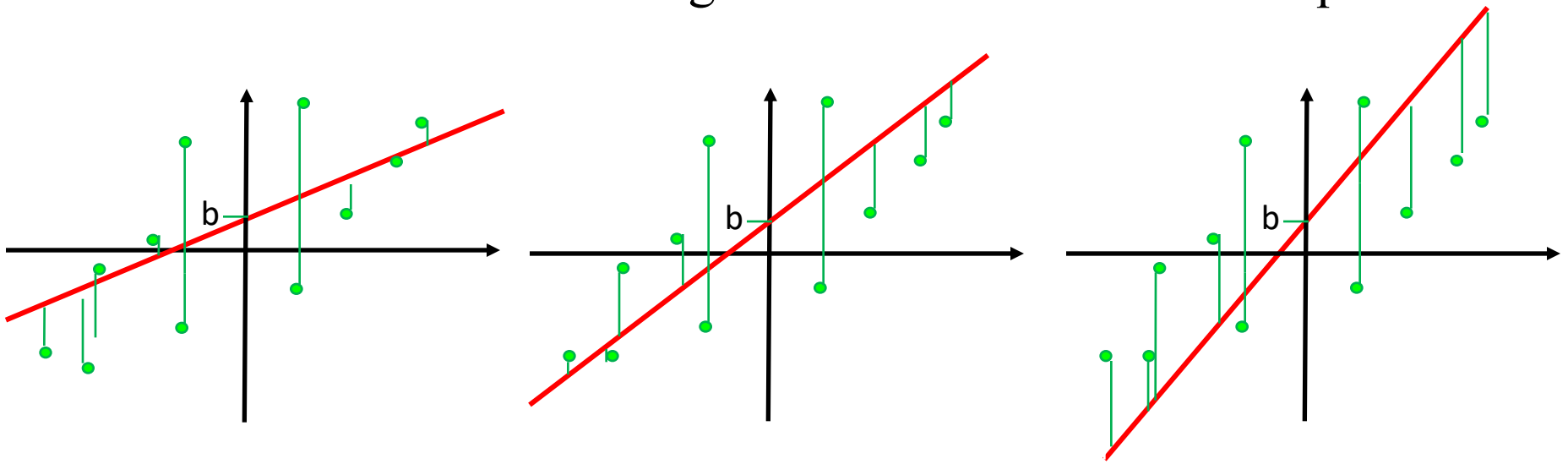


- Trying different values of *b* is equivalent to shifting the line up and down the scatter plot.

Regression

How to find *m*?

- Now we find the value of *m* that gives the minimum sum of squares



- Trying out different values of *m* is equivalent to changing the slope of the line, while *b* stays constant.

Regression

How to find m?

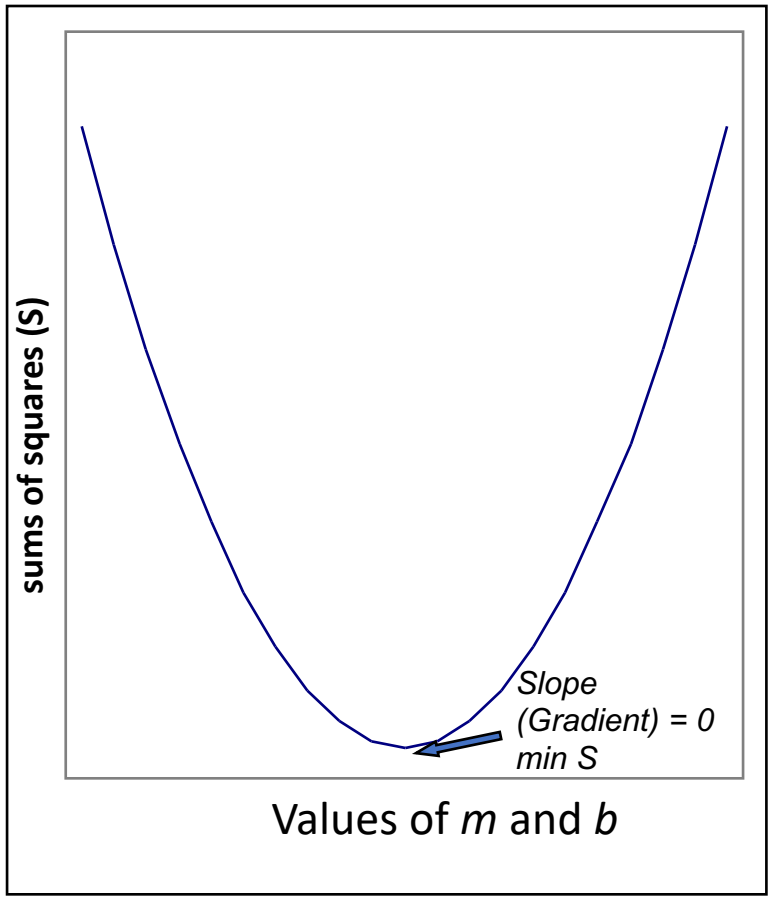
- Need to minimize $\Sigma(y-\hat{y})^2$

$$\hat{y} = mx + b$$

- so need to minimise:

$$\Sigma(y - (mx + b))^2$$

- If we plot the sums of squares for all different values of *m* and *b* we get a parabola, because it is a squared term
- So the minimum sum of squares is at the bottom of the curve, where the gradient is zero.



Regression

How to find *m* and *b*?

- The minimum sum of squares is at the bottom of the curve where the gradient = 0
- So we can find m and b that give min sum of squares by taking partial derivatives of $\Sigma(y - mx - b)^2$ with respect to m and b separately
- Then we solve these for 0 to give us the values of m and b that give the min sum of squares

Regression

How to find *m* and *b*?

- Doing this gives the following equations for m and b:

$$m = \frac{r\, s_{xy}}{s_x^2} \qquad b = \bar{y} - \frac{r\, s_{xy}}{s_x^2}\, \bar{x}$$

r = correlation coefficient of x and y
s_{xy} = standard deviation of y
s_x = standard deviation of x

- If we apply the values of s_{xy} and s_x to the above equations, we get:

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \qquad m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Regression

Example

We collected a data that tells **hours of sunshine** vs number of **ice creams** sold at the shop from Monday to Friday:

Hours of Sunshine "x"	Ice Creams Sold "y"
2	4
3	5
5	7
7	10
9	15

Find the best **m** (slope) and **b** (y-intercept) that suits that data represented by $y = mx + b$.

https://www.mathsisfun.com/data/least-squares-regression.html

Regression
Example

Step 1: For each (x,y) calculate x^2 and xy

Step 2: Calculate Slope **m**

Step 3: Calculate Intercept **b**

Step 4: Assemble the equation of a line:

Step 5: Find the (x,y) points and plot line **y = mx + b**

Excel Example

Regression
Example

Step 1: For each (x,y) calculate x^2 and xy

x	y	x2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
26	41	168	263

Regression
Example

Step 2: Calculate slope m

$$m = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

$$= \frac{5 \times 263 - 26 \times 41}{5 \times 168 - 262}$$

$$= \frac{1315 - 1066}{840 - 676}$$

$$= \frac{249}{164}$$

$$= 1.5183 \ldots$$

Regression
Example

Step 3: Calculate Intercept **b**

$$\boldsymbol{b} = \frac{\Sigma y - m\,\Sigma x}{N}$$

$$= \frac{41 - 1.5183 \times 26}{5}$$

$$= 0.3049\ldots$$

Regression
Example

Step 4: Assemble the equation of a line:

$$y = mx + b$$

$$y = 1.5183x + 0.3049$$

Regression
Example

Step 5: Find the (x,y) points and plot line **y = mx + b**

x	y	x2	xy	y = 1.518x + 0.305	ERROR
2	4	4	8	3.3415	-0.6585
3	5	9	15	4.8598	-0.1402
5	7	25	35	7.8963	0.8963
7	10	49	70	10.9329	0.9329
9	15	81	135	13.9695	-1.0305

Regression

Example

Now if the weather prediction says we have 8 hours of sunshine, you can predict that 12.45 ice creams can be sold out.

Summary

- ✓ **Percentiles and Box Plot**
  - ✓ Percentiles
  - ✓ Quartiles
  - ✓ Box Plots
- ✓ **Correlation**
  - ✓ Difference between Covariance and Correlation
  - ✓ How to variables are correlated – positive, negative or none
  - ✓ Correlation vs Regression
- ✓ **Regression**
  - ✓ Linear Regression
  - ✓ Found slope ’m’ and intercept ‘b’
  - ✓ Learning through excel