Statistics

Regression (الانحدار)

Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

Agenda

- 🔎 Real-World Stories & Applications of Regression
  - A Data-Driven Advisor
  - Optimizing a Web Application
  - Predicting Heart Disease Risk
  - Predicting Height from a Femur Bone
- What is Regression?
- Types of Regression
- Interpreting the Slope (m)
- Interpreting the Intercept (b)
- Best Fit Line
- Residual
- Least Squares Errors
- Examples
- Q & A

Review

- Scatterplot
  - How to explore the relationship between two quantitative variables.
- Correlation coefficient
  - Numerically describes relationship strength if the relationship has a straight-line
- Analyze the data further by **finding an equation** for the straight line that **best describes** that pattern

A Data-Driven Advisor

📖 Story Setup:

- Imagine you're a **university advisor** trying to help students improve their performance. You’ve collected data from past students, and you're trying to understand:

- What factors best predict a student's final exam score?

A Data-Driven Advisor

📖 Story Setup

- You're a university advisor. You've noticed students struggle or succeed in different ways. You wonder: What factors really predict a student's final exam score?

📊 Data Collected:

Feature	Description
Study Hours per Week	How much the student studies
Attendance (%)	Class attendance rate
Sleep Hours	Average hours of sleep per night
Social Media (hrs)	Time spent on social media
Caffeine Intake	Cups of coffee per day

A Data-Driven Advisor

- Imagine you're a **university advisor** trying to help students improve their performance. You’ve collected data from past students, and you're trying to understand:
- **You want to answer questions like**:

  - If a student studies **one extra hour per week**, how much will their grade improve?
  - Is attendance more important than sleep?
  - Do social media hours really matter that much?

Regression Model

$$\text{Score} = 5 + 1.5 \times StudyHours + 0.8 \times Attendance - 0.2 \times SocialMedia + \cdots$$

🔍 Interpretation of Weights:

- 📈 Study Hours (+1.5): Every extra hour adds 1.5 points.
- 📈 Attendance (+0.8): Attending class boosts scores.
- 📉 Social Media (–0.2): Each hour scrolling drops scores a bit.

Regression Model

🧠 Takeaway:

- Some factors have a **bigger impact**, some have **little or negative impact**, and some may be **statistically insignificant** — that’s how we discover which features are truly important.
- Regression helps us identify what really matters. Now, as an advisor, you’re not just guessing — you're guiding students with data!

Optimizing a Web Application

📖 Story Setup:

- You're a software engineer on a high-traffic e-commerce website. Users are complaining that the site loads too slowly.

- What factors are causing the delays?

Optimizing a Web Application

📖 Story Setup

- You start logging data on thousands of page loads. For each request, you record?

📊 Data Collected:

Feature	Description
Number of Images	How many product images on the page
JavaScript File Size	Size of all JS files loaded (in KB)
Database Query Time	Total query time in seconds
Active Users	Concurrent users on the site
Network Latency	Time delay in the network (ms)

Regression Model

$$\text{LoadTime} = 0.5 + 0.1 \times \text{NumImages} + 0.3 \times \text{JSSize} + 0.7 \times \text{DBTime} + 0.05 \times \text{Users}$$

🔍 Interpretation of Weights:

- DBTime (+0.7): Most influential — optimize your queries!
- JSSize (+0.3): Reducing JS size helps.
- NumImages (+0.1): Images have minor effect.
- Users (+0.05): Slight effect from traffic..

Regression Model

🧠 Takeaway:

- Instead of guessing, you're now using data to focus engineering time where it matters most — improving database performance.

Saving Lives with Data

🏥 Story Setup: Predicting Heart Disease Risk

You work with a hospital's cardiology team. They want to predict which patients are at high risk of heart disease — and prevent it before it's too late.

Saving Lives with Data

📊 Data Collected:

📊 Feature	💡 Description
👵 Age	Patient's age in years
🩸 Cholesterol	Blood cholesterol level (mg/dL)
📊 Blood Pressure	Resting systolic BP (mm Hg)
🏋 BMI	Body Mass Index
🍕 Cigarettes/day	Number of cigarettes smoked daily
⚽ Exercise Frequency	Days per week of moderate physical activity
👥 Family History	Yes/No indicator

Regression Model

$$RiskScore = 10 + 0.8 \times Age + 1.2 \times Cholesterol + 0.9 \times Cigarettes - 0.6 \times Exercise$$

Interpretation of Coefficients:

- Cholesterol (+1.2): The most impactful — high levels sharply raise risk.
- Cigarettes (+0.9): Smoking adds serious risk.
- Exercise (-0.6): More activity reduces risk.
- Age (+0.8): Natural risk increase with age

Regression Model

🧠 Takeaway:

- With regression, doctors move from intuition to insight — enabling personalized care and smarter decisions to save lives.

Predicting Height from a Femur Bone

- Anthropologists can reconstruct information using partial human remains at burial sites (مواقع الدفن)
- After finding a femur (thighbone 🦴) عظمة الفخذ they can predict how tall an individual was
- Sadly, this particular prediction equation had to be applied to bones found in **mass graves** (مقابر جماعية) **in Kosovo**, to help identify Albanians who had been executed by Serbians in 1998

Predicting Height from a Femur Bone

$$\hat{y} = 61.4 + 2.4x$$

Example

- A femur found at a particular site has a length of 50 cm. What is the predicted height of the person who had that femur?

- $\hat{y} = 61.4 + 2.4(50) = 181.4\ cm$

Conclusion: The estimated height of the individual is 181.4 cm

What is Regression?

- Regression is a fundamental concept in statistics and machine learning used to model the relationship between a dependent variable (often called the output or target) and one or more independent variables (inputs or features).

What is Regression?

🧠 In Simple Terms

- Regression is like asking:

  If I know X, can I predict Y?

- Example:

🏠 Predicting house price based on size and location
👩‍🎓 Predicting student score from study hours

Types of Regression

- **Linear Regression**: Assumes a straight-line relationship between input(s) and the output.
  - $price = a * size + b$

- **Multiple Linear Regression**: Like linear regression, but with multiple inputs.
  - $price = a1 * size + a2 * bedrooms + a3 * location + b$

Types of Regression

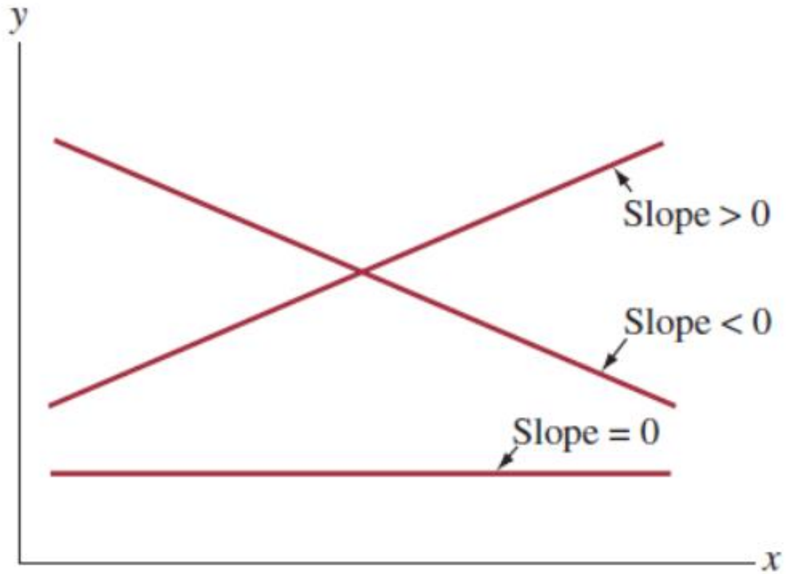
- **Polynomial Regression**: Models non-linear relationships by using powers of the input variables.
- **Logistic Regression**: Despite its name, it's used for classification problems (e.g., yes/no, spam/ham).
- **Ridge, Lasso, and Elastic Net Regression**: Regularized versions of linear regression to prevent overfitting.
- **Non-linear Regression**: When the relationship between variables is not linear or polynomial.

Simple Linear Regression

$$y = \boldsymbol{m}x + \boldsymbol{b}$$

Where:

- y: Predicted value
- x: Input feature
- m: Slope or weight
- b: Intercept (starting point)



Interpreting the Slope (m)

- The slope tells us **how much** the dependent variable y is expected to change when x increases by one unit, assuming all else is constant.
- In a model predicting salary from years of experience:
  - $Salary = \mathbf{2000} \times Experience + 30{,}000$
  - Slope = 2000: Each additional year of experience is associated with an increase of $2000 in salary.
- Positive vs. Negative Slope:
  - Positive m: y increases as x increases.
  - Negative m: y decreases as x increases.

Interpreting the Intercept (b)

- The intercept is the predicted value of y when x=0.

- Example:
  - Intercept = 30,000: When experience is 0 years, the expected salary is $30,000.

What is the Best Fit Line? - خط الانحدار الأفضل

- The **best fit line** (also called the **regression line**) is a straight line that best represents the data on a scatter plot.
- It shows the overall trend (الاتجاه العام) between an independent variable (x) and a dependent variable (y).

📐 Purpose: To minimize the difference (errors) between the actual data points and the predicted points on the line. This difference is called a **residual**.

Residual

- **Residuals** measure the size of the prediction errors, the vertical distance between the point and the regression line.
- Each observation has a residual.

$$\boldsymbol{y - \hat{y}}$$

  - A large residual indicates an unusual observation.
  - The smaller the absolute value of a residual, the closer the predicted value is to the actual value, so the better is the prediction.

Error

- Minimize the difference between actual and predicted values — typically using Least Squares Error

$$\text{Error} = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

Best Fit Line Formulas:

- Given n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$:

- Slope (m)

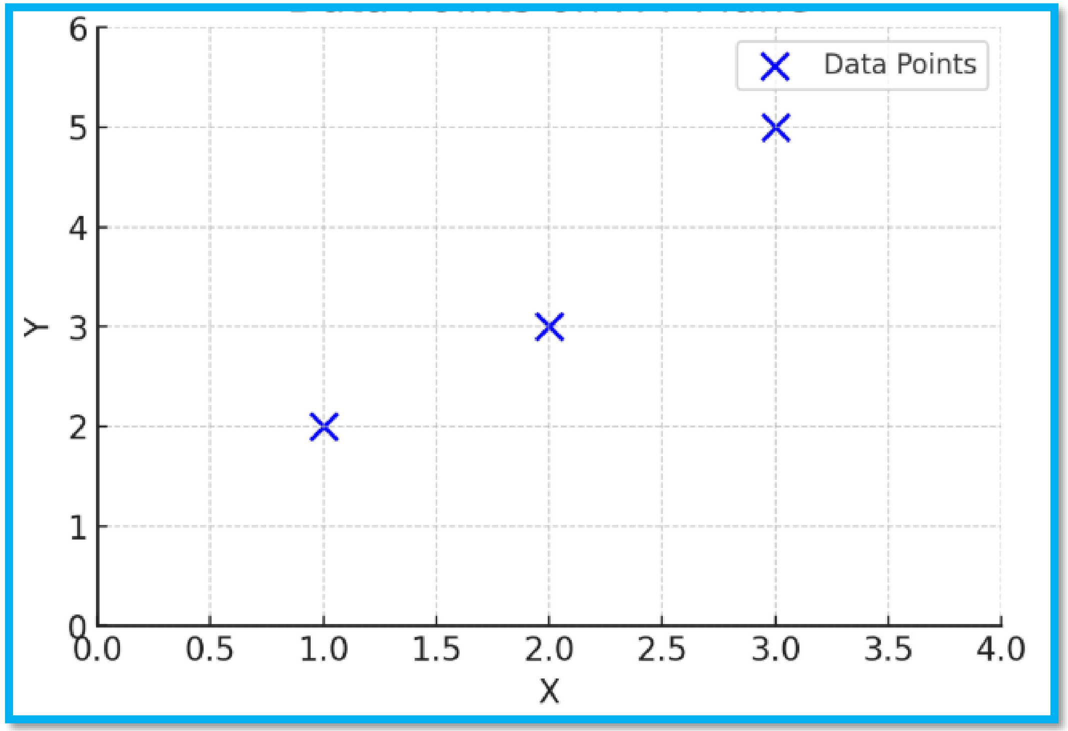
$$m = \frac{n\sum(xy) - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

- Intercept (b):

$$b = \frac{\sum y - m\sum x}{n}$$

Example:

- Suppose we have points: (1, 2), (2, 3), (3, 5)



Example:

- Suppose we have points: (1, 2), (2, 3), (3, 5)
- Step 1: Compute sums
  - $\sum x = 1 + 2 + 3 = 6$
  - $\sum y = 2 + 3 + 5 = 10$
  - $\sum xy = 1 \times 2 + 2 \times 3 + 3 \times 5 = 2 + 6 + 15 = 23$
  - $\sum x^2 = 12 + 22 + 32 = 1 + 4 + 9 = 14$
  - $n=3$

Example:

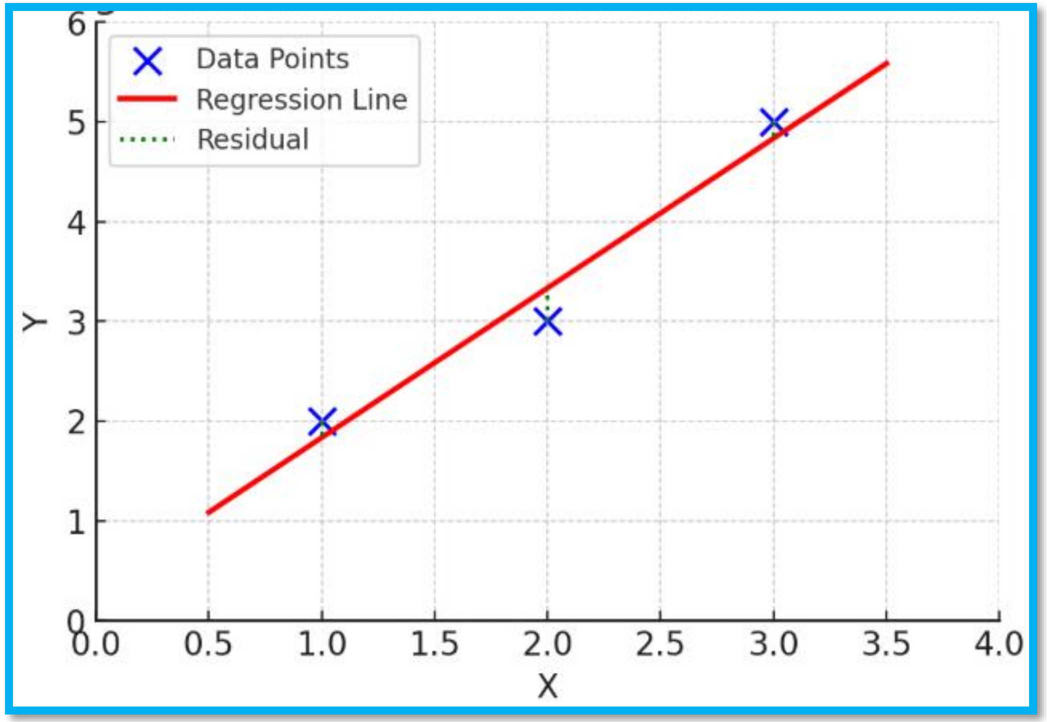
- Step 2: Calculate m and b

$$m = \frac{3\times23-6\times10}{3\times14-6^2} = \frac{69-60}{42-36} = \frac{9}{6} = 1.5$$

$$b = \frac{10-1.5\times6}{3} = \frac{10-9}{3} = \frac{1}{3} \approx 0.33$$

✅ **Final Equation:**

$$\boxed{y = 1.5x + 0.33}$$



Example 02

Observation	x (Study Hours)	y (Exam Score)	x^2	$x \cdot y$
1	1	2	1	2
2	2	4	4	8
3	3	5	9	15
4	4	4	16	16
5	5	5	25	25
Total	**15**	**20**	**55**	**66**

Example 02

- Step 1: Compute the slope (m)

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

- Plug in the values:
  - $n = 5$
  - $\sum x = 15$
  - $\sum y = 20$
  - $\sum xy = 66$
  - $\sum x^2 = 55$

$$m = \frac{5(66) - (15)(20)}{5(55) - (15)^2} = \frac{330 - 300}{275 - 225} = \frac{30}{50} = 0.6$$

Example 02

- Step 2: Compute the intercept (b)

$$b = \frac{\sum y - m \sum x}{n} = \frac{20 - 0.6(15)}{5} = \frac{20 - 9}{5} = \frac{11}{5} = 2.2$$

- The best fit line is: $\boxed{y = 0.6x + 2.2}$

- Interpretation:
  - Slope (0.6): For each additional hour studied, the expected exam score increases by 0.6 points.
  - Intercept (2.2): If a student studies 0 hours, their expected score is 2.2 (though this might not have a practical meaning — it's just the model's baseline).

Key Terms You Should Know

- 🔢 Independent Variable: The input(s) you control or observe
- 🎯 Dependent Variable: The output you're trying to predict
- 📈 Best Fit Line: Line that minimizes the prediction error
- 🧮 Residual: The difference between actual and predicted values
- 🧠 Model: The function or rule the algorithm finds to relate X and Y

Why is Regression Useful?

- 📊 Make predictions (e.g., sales forecasting)
- 🩺 Understand relationships (e.g., health risk factors)
- 🛠️ Optimize systems (e.g., web performance tuning)
- 💡 Gain insights from data

📚 Part 1: Conceptual Questions (Theory)

❓ Q1: What is the purpose of a regression line?

📚 Part 1: Conceptual Questions (Theory)

❓ Q1: What is the purpose of a regression line?

✅ A: To model the relationship between the dependent and independent variable and make predictions based on that relationship.

📚 Part 1: Conceptual Questions (Theory)

❓ Q2: What does the slope in a regression line tell us?

📚 Part 1: Conceptual Questions (Theory)

❓ Q2: What does the slope in a regression line tell us?

✅ A: It tells us how much the dependent variable changes for a one-unit change in the independent variable.

📚 Part 1: Conceptual Questions (Theory)

❓ Q3: What is a residual?

📚 Part 1: Conceptual Questions (Theory)

❓ Q3: What is a residual?

✅ A: It is the difference between the actual value and the predicted value from the regression line.

🔢 Part 2: Computationnel Exercices

🧮 Q4: Given the data points (1, 2), (2, 4), and (3, 5), calculate the regression line.

🔢 Part 2: Computationnel Exercices

🧮 Q4: Given the data points (1, 2), (2, 4), and (3, 5), calculate the regression line.

✅ A:

- $\sum x = 6, \sum y = 11, \sum xy = 25, \sum x^2 = 14, n = 3$
- $m = \frac{3(25)-6(11)}{3(14)-6^2} = \frac{75-66}{42-36} = \frac{9}{6} = 1.5$
- $b = \frac{11-1.5(6)}{3} = \frac{11-9}{3} = 0.67$
- Final regression equation:

$$y = 1.5x + 0.67$$

🔢 Part 2: Computationnel Exercices

🧮 Q5: For the regression line $y = 2x + 1$, what is the predicted value of y when x=4?

🔢 Part 2: Computationnel Exercices

🧮 Q5: For the regression line $y = 2x + 1$, what is the predicted value of y when x=4?

✅ A:

$$y = 2(4) + 1 = 8 + 1 = 9$$

🔢 Part 2: Computationnel Exercices

🧮 Q6: A regression model gives the equation $y = -0.5x + 10$. Interpret the slope.

🔢 Part 2: Computationnel Exercices

🧮 Q6: A regression model gives the equation $y = -0.5x + 10$. Interpret the slope.

✅ A:

The slope -0.5 means that for every 1 unit increase in x, y **decreases by 0.5** units.

🔢 Part 2: Computationnel Exercices

💡 Q9: Given the model y=0.7x+2, a student observes the following:

x	Actual y
2	3.5
4	5.2
6	6.5

Calculate the predicted y and residuals.

🔢 Part 2: Computationnel Exercices

💡 Q9: Given the model y=0.7x+2, a student observes the following:

✅ A:

x	Actual y	Predicted y = 0.7x + 2	Residual (Actual - Predicted)
2	3.5	3.4	0.1
4	5.2	4.8	0.4
6	6.5	6.2	0.3

References

- Montgomery, D.C. and Runger, G.C., 2010. Applied statistics and probability for engineers. John wiley & sons.
- Devore, J.L., 2011. Probability and Statistics for Engineering and the Sciences. Cengage learning.
- Gubner, J.A., 2006. Probability and random processes for electrical and computer engineers. Cambridge University Press.
- Agresti, A. and Franklin, C., 2007. The art and science of learning from data. Upper Saddle River, New Jersey, 88.
- Navidi, W.C., 2008. Statistics for engineers and scientists (Vol. 2, No. 4). New York, NY, USA: McGraw-Hill Higher Education.