Regression

Week 5 & 6

Spring 2025

Rida Maryam

What is Regression

- Regression is a statistical method used to model the relationship between a dependent (target) variable and one or more independent (predictor) variables.
- Regression is a statistical method that help us to understand and predict the relationship between a dependent variable and one or more independent variables.
- It represents the best-fit line that predicts the dependent variable based on the independent variable.

Application of Regression

Regression plays a crucial role in various computer science fields and other fields.

- Commonly used for **prediction, forecasting**, and determining the strength of relationships between variables.
- Its applications continue to grow with advancements in AI and data science, Such as
  Finance and Economics
- Predicting stock prices based on historical trends.
- Estimating economic growth using GDP indicators.

Application of Regression

Healthcare

- Predicting disease progression based on patient data.
- Estimating the effectiveness of treatments using medical history.

Marketing and Sales

- Forecasting sales based on advertising spend and consumer behavior.

Real-world Applications of Regression in Software Engineering

1. **Predictive Modeling**: Regression is used to predict outcomes such as software defects, project completion time, or system performance.
2. **Data Analysis**: It helps in analyzing trends and patterns in large datasets, such as user behavior or system logs.
3. **Resource Estimation**: Regression models can estimate the resources (time, cost, effort) required for software development projects.
4. **Quality Assurance**: Predicting the likelihood of bugs or failures in software systems.

Difference Between Regression and Classification

Regression: Predicts continuous values (e.g., house prices, temperature, sales).

Classification: Predicts discrete labels or categories (e.g., spam/not spam, yes/no, high/medium/low).

Example: Predicting the price of a house is a regression problem, while predicting whether a house will sell or not is a classification problem.

Identify which problem requires classification and which requires regression.

1. A bank wants to predict whether a loan applicant will default on their loan or not.
2. A real estate agency wants to estimate the price of a house based on its size, location, and number of rooms.
3. An e-commerce website wants to predict whether a customer will buy a product based on their browsing history.
4. A weather forecasting system needs to predict the amount of rainfall in millimeters for the next day.

Types Of Regression

1. **Liner Regression**
2. **Multiple regression**
3. **Polynomial regression**
4. Ridge Regression (L2 Regularization)
5. Lasso Regression (L1 Regularization)
6. Logistic Regression (For Classification)
7. Stepwise Regression
8. Support Vector Regression (SVR)
9. Decision Tree Regression
10. Random Forest Regression

LINEAR REGRESSION

Linear Regression

Models the relationship between the dependent and independent variables as a straight line.

Equation: $y = mx + c$

where:

1. y = dependent variable,
2. x = independent variable,
3. m = slope of the line, (how much y changes for a unit change in x)
4. c = y-intercept.(value of y, when x is 0)

Example: Predicting house prices based on square footage.

Linear Regression

- **Example**: Predicting Monthly Sales Based on Advertising Budget

Given Data: Advertising Budget vs Sales

Advertising Budget (x) (in \$1000s)	Monthly Sales (y) (in \$1000s)
1	20
2	25
3	30
4	38
5	45
6	50
7	55
8	60

Linear Regression

Formula: y=mx+c

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$c = \bar{y} - m\bar{x}$$

Step 1: Calculate Mean Values

$$\bar{x} = \frac{1+2+3+4+5+6+7+8}{8} = \frac{36}{8} = 4.5$$

$$\bar{y} = \frac{20+25+30+38+45+50+55+60}{8} = \frac{323}{8} = 40.375$$

Step 2: Compute m (Slope)

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	20	-3.5	-20.375	71.3125	12.25
2	25	-2.5	-15.375	38.4375	6.25
3	30	-1.5	-10.375	15.5625	2.25
4	38	-0.5	-2.375	1.1875	0.25
5	45	0.5	4.625	2.3125	0.25
6	50	1.5	9.625	14.4375	2.25
7	55	2.5	14.625	36.5625	6.25
8	60	3.5	19.625	68.6875	12.25

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 248.5$$

$$\sum(x_i - \bar{x})^2 = 40$$

$$m = \frac{248.5}{40} = 6.2125$$

Step 3: Compute c (Intercept)

$$c = \bar{y} - m\bar{x}$$

$$c = 40.375 - (6.2125 \times 4.5)$$

$$c = 40.375 - 27.95625 = 12.41875$$

Final Equation

$$y = 12.42 + 6.21x$$

Step 4: Prediction

- If the company spends **$5,000** on advertising (x=5):
    - $y=12.42+6.21(5)$
    - $y=12.42+31.05$
    - $y=43.47$

Predicted sales: $43,470

Linear Regression

- Example: Predicting Exam Scores Based on Study Hours

Study Hours (X)	Exam Score (Y)
2	50
4	70
6	80
8	90

Formula to Find m and c

$$m = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$c = \frac{\sum Y - m(\sum X)}{n}$$

Using the example dataset:

Study Hours (X)	Exam Score (Y)	X^2	XY
2	50	4	100
4	70	16	280
6	80	36	480
8	90	64	720
Sum	**290**	**120**	**1580**

Calculate m and c:

$$m = \frac{4(1580) - (20)(290)}{4(120) - (20)^2} = \frac{6320 - 5800}{480 - 400} = \frac{520}{80} = 6.5$$

$$c = \frac{290 - 6.5(20)}{4} = \frac{290 - 130}{4} = \frac{160}{4} = 40$$

So, the regression equation is:

$$Y = 6.5X + 40$$

Now put X=5 to predict the Exam score.

***Y*=6.5(5)+40**

y=32.5+40

y=72.5

If a student studies for 5 hours, he can score 72.5.

Your Task?

Write A computer program in any language to implement the following program. (Take a data set of 10 rows)

1. Predicting house prices based on square footage.
2. Predicting Exam Scores Based on Study Hours

MULTIPLE REGRESSION

Multiple Regression?

Multiple regression is a statistical technique used to model the relationship between a **dependent variable (Y)** and **two or more independent variables (X_1, X_2, ..., X_n)**. It extends simple linear regression, which involves only one independent variable, to cases where multiple factors influence the outcome.

Multiple Linear Regression

Extension of linear regression with multiple independent variables.

Formula: $$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon$$

- **Y**: Dependent variable (the outcome you want to predict).
- **X_1, X_2, …, X**□: Independent variables (predictors).
- **β_0**: Intercept (value of **Y** when all **X** are 0).
- **β_1, β_2, …, β**□: Coefficients (represent the change in **Y** for a unit change in the corresponding **X**, holding other variables constant).
- **ϵ**: Error term (accounts for variability in **Y** not explained by the independent variables).

Example: Predicting Scores based on study hours, and Attendance.

Predicting Scores based on study hours, and Attendance.

Here's the sample dataset again:

Student	Study Hours (X_1)	Attendance % (X_2)	Exam Score (Y)
1	2	80	50
2	4	90	70
3	6	95	80
4	8	100	90

The regression equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Where:

- Y: Exam score (dependent variable, range: 0 to 100).
- X_1: Study hours.
- X_2: Attendance percentage (range: 0 to 100).
- β_0: Intercept.
- β_1, β_2: Coefficients for X_1 and X_2.
- ϵ: Error term.

The normal equations for two variables are:

1. $\sum Y = n\beta_0 + \beta_1 \sum X_1 + \beta_2 \sum X_2$
2. $\sum X_1 Y = \beta_0 \sum X_1 + \beta_1 \sum X_1^2 + \beta_2 \sum X_1 X_2$
3. $\sum X_2 Y = \beta_0 \sum X_2 + \beta_1 \sum X_1 X_2 + \beta_2 \sum X_2^2$

Student	X_1	X_2	Y	X_1^2	X_2^2	X_1Y	X_2Y	X_1X_2
1	2	80	50	4	6400	100	4000	160
2	4	90	70	16	8100	280	6300	360
3	6	95	80	36	9025	480	7600	570
4	8	100	90	64	10000	720	9000	800

Using the sample data, we calculate the sums:

Sum	Value
$\sum Y$	290
$\sum X_1$	20
$\sum X_2$	365
$\sum X_1^2$	120
$\sum X_2^2$	33525
$\sum X_1X_2$	1890
$\sum X_1Y$	1580
$\sum X_2Y$	26900

The normal equations for two variables are:

1. $\sum Y = n\beta_0 + \beta_1 \sum X_1 + \beta_2 \sum X_2$
2. $\sum X_1 Y = \beta_0 \sum X_1 + \beta_1 \sum X_1^2 + \beta_2 \sum X_1 X_2$
3. $\sum X_2 Y = \beta_0 \sum X_2 + \beta_1 \sum X_1 X_2 + \beta_2 \sum X_2^2$

Using the sample data, we calculate the sums:

Sum	Value
$\sum Y$	290
$\sum X_1$	20
$\sum X_2$	365
$\sum X_1^2$	120
$\sum X_2^2$	33525
$\sum X_1 X_2$	1890
$\sum X_1 Y$	1580
$\sum X_2 Y$	26900

Substitute the sums into the normal equations:

1. $290 = 4\beta_0 + 20\beta_1 + 365\beta_2$
2. $1580 = 20\beta_0 + 120\beta_1 + 1890\beta_2$
3. $26900 = 365\beta_0 + 1890\beta_1 + 33525\beta_2$

$$\begin{bmatrix} 4 & 20 & 365 \\ 20 & 120 & 1890 \\ 365 & 1890 & 33525 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 290 \\ 1580 \\ 26900 \end{bmatrix}$$

$\beta_0 = 10$: The expected exam score when study hours and attendance are both 0.

$\beta_1 = 5$: For every additional hour of study, the exam score increases by 5 points, holding attendance constant.

$\beta_2 = 0.2$: For every 1% increase in attendance, the exam score increases by 0.2 points, holding study hours constant.

Equation:

$$Y=10+5X1+0.2X2$$

For example, if a student:

Studies **5 hours**, and has **85% attendance**,

The predicted exam score is:

Y=10+5(5)+0.2(85)=

Y=10+25+17=

Y=**52**

Polynomial Regression

Polynomial regression is a type of regression analysis that models the relationship between a dependent variable y and an independent variable x as an nth-degree polynomial

Formula: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \epsilon$

- **y** → Dependent variable (output or response).
- **x** → Independent variable (input or predictor).
- **n** → Degree of the polynomial
- **$\beta_0, \beta_1, \beta_2, \ldots, \beta_n$** → Regression coefficients (weights assigned to each power of x).

- **Example:** Predicting population growth over time.

Polynomial Regression

•B0 is the **intercept**, the value of y when x=0.
•B1 x is the **linear term**, representing a straight-line effect.
•B2 x2 is the **quadratic term**, capturing curvature.

A higher-degree polynomial allows the model to fit more complex relationships, but too high a degree can lead to **overfitting**.

Evaluating Regression Models: R-squared & Error Analysis

- Once a regression model is built, it is essential to evaluate its performance.
- Common evaluation metrics include:
  1. **R-squared**
  2. **Mean Absolute Error (MAE)**
  3. **Mean Squared Error (MSE)**

R Squared

•R-squared shows how well the model predicts the actual values, ranging from 0 (no fit) to 1 (perfect fit).
•A higher R-squared means the model is more accurate in explaining the data.

The formula for R^2 (coefficient of determination) is:

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$$

Where:

- $SS_{\text{total}} = \sum(y_{\text{actual}} - \bar{y})^2$ (Total sum of squares)
- $SS_{\text{residual}} = \sum(y_{\text{actual}} - y_{\text{predicted}})^2$ (Residual sum of squares)
- $\bar{y}$ is the mean of actual y-values
- y_{actual} are the actual values
- $y_{\text{predicted}}$ are the predicted values from the model

We have the **linear regression equation**:

$y=12.42+6.21x$

The **given data**:

Advertising Budget ($1000s) (x)	Monthly Sales ($1000s) (y)
1	20
2	25
3	30
4	38
5	45
6	50
7	55
8	60

R-squared

R-squared

Step 1: Compute Mean of y

$$\bar{y} = \frac{20 + 25 + 30 + 38 + 45 + 50 + 55 + 60}{8}$$

$$\bar{y} = \frac{323}{8} = 40.375$$

Step 2: Compute Predicted Values $y_{\text{predicted}}$

Using the equation $y = 12.42 + 6.21x$, we compute $y_{\text{predicted}}$:

x	y_{actual}	$y_{\text{predicted}} = 12.42 + 6.21x$
1	20	$12.42 + (6.21 \times 1) = 18.63$
2	25	$12.42 + (6.21 \times 2) = 24.84$
3	30	$12.42 + (6.21 \times 3) = 31.05$
4	38	$12.42 + (6.21 \times 4) = 37.26$
5	45	$12.42 + (6.21 \times 5) = 43.47$
6	50	$12.42 + (6.21 \times 6) = 49.68$
7	55	$12.42 + (6.21 \times 7) = 55.89$
8	60	$12.42 + (6.21 \times 8) = 62.10$

Step 3: Compute SS_{total} (Total Sum of Squares)

Formula:

$$SS_{\text{total}} = \sum (y_i - \bar{y})^2$$

y_{actual}	$y_{\text{actual}} - \bar{y}$	$(y_{\text{actual}} - \bar{y})^2$
20	$20 - 40.375 = -20.375$	$(-20.375)^2 = 415.13$
25	$25 - 40.375 = -15.375$	$(-15.375)^2 = 236.44$
30	$30 - 40.375 = -10.375$	$(-10.375)^2 = 107.69$
38	$38 - 40.375 = -2.375$	$(-2.375)^2 = 5.64$
45	$45 - 40.375 = 4.625$	$(4.625)^2 = 21.39$
50	$50 - 40.375 = 9.625$	$(9.625)^2 = 92.62$
55	$55 - 40.375 = 14.625$	$(14.625)^2 = 213.94$
60	$60 - 40.375 = 19.625$	$(19.625)^2 = 385.06$

Step 4: Compute SS_{residual} (Residual Sum of Squares)

Formula:

$$SS_{\text{residual}} = \sum (y_{\text{actual}} - y_{\text{predicted}})^2$$

y_{actual}	$y_{\text{predicted}}$	$y_{\text{actual}} - y_{\text{predicted}}$	$(y_{\text{actual}} - y_{\text{predicted}})^2$
20	18.63	$20 - 18.63 = 1.37$	$(1.37)^2 = 1.88$
25	24.84	$25 - 24.84 = 0.16$	$(0.16)^2 = 0.03$
30	31.05	$30 - 31.05 = -1.05$	$(-1.05)^2 = 1.10$
38	37.26	$38 - 37.26 = 0.74$	$(0.74)^2 = 0.55$
45	43.47	$45 - 43.47 = 1.53$	$(1.53)^2 = 2.34$
50	49.68	$50 - 49.68 = 0.32$	$(0.32)^2 = 0.10$
55	55.89	$55 - 55.89 = -0.89$	$(-0.89)^2 = 0.79$
60	62.10	$60 - 62.10 = -2.10$	$(-2.10)^2 = 4.41$

$$SS_{\text{residual}} = 11.20$$

Step 5: Compute R^2

Formula:

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$$

$$R^2 = 1 - \frac{11.20}{1477.91}$$

$$R^2 = 1 - 0.0076$$

$$R^2 = 0.9924$$

Final Interpretation

- $R^2 = 0.9924$ **means 99.24% of the variance** in monthly sales is explained by the advertising budget.
- Since R^2 is very close to **1**, the model fits the data **very well**.

Conclusion

- **Regression Equation:** $y = 12.42 + 6.21x$
- $R^2 = 0.9924$**: Model is highly accurate.**

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) in Regression

Mean Absolute Error (MAE) is a metric used to evaluate the performance of a regression model. It measures the average absolute difference between the predicted values and the actual values. A lower MAE indicates better model accuracy.

Formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- y_i = Actual value
- $\hat{y}_i$ = Predicted value
- n = Number of observations

MAE Example

Example: Predicting Pizza Delivery Time

A pizza shop predicts delivery times (in minutes). Here's real vs. predicted data for 5 orders

Order	Actual Time	Predicted Time	Absolute Error
1	20 min	18 min	$\|20 - 18\|$ = **2 min**
2	30 min	35 min	$\|30 - 35\|$ = **5 min**
3	25 min	23 min	$\|25 - 23\|$ = **2 min**
4	40 min	45 min	$\|40 - 45\|$ = **5 min**
5	15 min	10 min	$\|15 - 10\|$ = **5 min**

Step 1: Find absolute errors (ignore +/- signs).
Step 2: Sum the errors:

2+5+2+5+5=19

- ➢Divide by number of orders (5): 19/5 =3.8 min

Conclusion: On average, predictions are **off by 3.8 minutes**.

Your Task?

Find MAE

Student	Actual Score	Predicted Score
1	85	82
2	60	65
3	90	88
4	75	70
5	95	98

Mean Squared Error (MSE)

Mean Squared Error (MSE)

- **MSE** is another popular metric for regression. It calculates the average of the **squared differences** between actual and predicted values.

◆ **Formula:**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- y_i = actual value
- $\hat{y}_i$ = predicted value
- n = number of data points

Example

House	Actual Price (y)	Predicted Price	Error	Squared Error
1	200	180	-20	400
2	150	160	10	100
3	300	310	10	100
4	250	240	-10	100

MSE= (400+100+100+100)/4
= 700/4
= 175