Course: Mathematical statistics

Week 12: The least square method

Lecturer: Nagulama Moses

Kumi University

May 29, 2025

Outline

Intended learning outcomes

- Define the least squares method and explain its purpose in regression analysis.
- Distinguish between dependent and independent variables in the context of regression modeling.
- Formulate the equation of the regression line using the least squares approach.
- Calculate the least squares estimates for the intercept α and slope β of the regression line.

Introduction to Regression

Regression refers to the statistical concept of "stepping back" or returning toward an average.

It was introduced by the British statistician **Sir Francis Galton** (1822–1911), who observed that children's heights tend to regress towards the average height of the population.

Definition:
Regression analysis is a statistical technique used to examine the average or expected relationship between two or more variables, expressed in terms of the original units of the data.

Purpose of Regression

Regression is a powerful tool used to:

- Understand the nature of the relationship between variables.
- Predict the value of one variable based on the known values of others.
- Quantify how changes in an independent variable affect a dependent variable.

Types of Variables in Regression

1. Dependent Variable (Y):
The variable whose value we are trying to predict or explain. It depends on the independent variable(s).

2. Independent Variable (X):
The variable that provides the basis for prediction. It is assumed to influence or determine the value of the dependent variable.

Example:
In studying the effect of fertilizer (X) on crop yield (Y),
Fertilizer is the *independent variable*, and
Crop yield is the *dependent variable*.

Terminology in Regression

Independent Variable:

Also known as the regressor, predictor, or explanatory variable. It is denoted by X and is used to predict or explain the response variable.

Dependent Variable:

Also referred to as the *response*, regressed, or explained variable. It is denoted by Y and depends on the values of X.

The Core Problem in Regression

The main objective in regression is to determine the conditional expectation of the dependent variable Y given a specific value of the independent variable X.

Conditional Expectation:

$$\mu_{Y|X=x} = E(Y|X = x)$$

This represents the expected value of Y when X takes the value x.

Linear Regression Model

For a linear relationship between X and Y, the conditional expectation is modeled as:

$$\mu_{Y|X=x} = \alpha + \beta x$$

where:

- α is the **intercept**,
- β is the **slope or regression coefficient**,
- α and β are constants estimated from sample data.

This is known as the **regression equation of Y on X**.

Line of Best Fit

The **regression line** (line of best fit) is determined from observed data points (x_i, y_i) using the method of least squares.

It provides the best linear approximation of the conditional expectation $\mu_{Y|X=x}$.

Equation of the regression line:

$$\mu_{Y|X=x} = \alpha + \beta x$$

This line minimizes the sum of the squared vertical deviations between the observed y_i values and the predicted values.

Introduction to Least Squares Method

The **Least Squares Method** is used to find the line of best fit for a set of data points (x_i, y_i).

Objective: Minimize the sum of squared differences between the observed values y_i and the predicted values $\hat{y}_i$ from the model.

Regression Line Equation:

$$y = \alpha + \beta x$$

In practice, we estimate it as:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

Model Assumptions

The simple linear regression model assumes:

$$y_i = \alpha + \beta x_i + e_i$$

where:

- α and β are the true regression coefficients,
- e_i is the error term for observation i,
- $E(e_i) = 0$ (mean of error terms is zero),
- Error terms e_i are independently and identically distributed (i.i.d.).

Estimating Parameters

The least squares estimates minimize:

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

The formulas for the least squares estimators are:

$$\hat{\beta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

So the regression line becomes:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Key Point on the Regression Line

The point $(\bar{x}, \bar{y})$ always lies on the least squares regression line. That is:

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

Interpretation: The regression line passes through the mean of the data.

Regression of X on Y

Similarly, to predict X from Y, we reverse the roles:

$$x_i = a + by_i + \epsilon_i$$

Estimates:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - (\sum y_i)^2} = \frac{S_{xy}}{S_{yy}}$$

$$a = \bar{x} - b\bar{y}$$

So, the regression line of X on Y is:

$$\hat{x} = a + by$$

Example: Regression of Blood Pressure on Age

The data below shows the age and systolic blood pressure (in mmHg) of 6 randomly selected subjects:

Age (X)	43	48	56	61	67	70
Pressure (Y)	128	120	135	143	141	152

Find the regression line of Y on X, then use it to estimate the blood pressure of a subject aged 65.

Compute Necessary Sums

Let's calculate the needed values:

X	43	48	56	61	67	70
Y	128	120	135	143	141	152
XY	5504	5760	7560	8723	9447	10640
X^2	1849	2304	3136	3721	4489	4900
$\sum$	345	819	47634	20399		

- $n = 6$
- $\sum X = 345$
- $\sum Y = 819$
- $\sum XY = 47634$
- $\sum X^2 = 20399$

Compute Means and Slope

$$\bar{X} = \frac{345}{6} = 57.5, \quad \bar{Y} = \frac{819}{6} = 136.5$$

Slope:

$$\hat{\beta} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{6 \cdot 47634 - 345 \cdot 819}{6 \cdot 20399 - (345)^2}$$

$$\hat{\beta} = \frac{285804 - 282555}{122394 - 119025} = \frac{3249}{3369} \approx 0.964$$

Intercept:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 136.5 - 0.964 \cdot 57.5 \approx 136.5 - 55.43 = 81.07$$

Regression Line and Prediction

Regression Line of Y on X:

$$\hat{Y} = 81.07 + 0.964X$$

Estimate the blood pressure for a 65-year-old:

$$\hat{Y}_{65} = 81.07 + 0.964 \cdot 65 \approx 81.07 + 62.66 = 143.73$$

Interpretation:
A 65-year-old subject is predicted to have a systolic blood pressure of approximately **143.7 mmHg**.

Example:

The scores of students in Mathematics and Physics tests are as follows:

Math Score (X)	83	97	80	95	73	78	91	86
Physics Score (Y)	78	95	83	93	78	72	90	80

Tasks:

(a) Determine the regression line of Y on X, and estimate Y when $X = 85$.

(b) Determine the regression line of X on Y, and estimate X when $Y = 84$.

Compute Required Values

Let us define:

X	83	97	80	95	73	78	91	86
Y	78	95	83	93	78	72	90	80
XY	6474	9215	6640	8835	5694	5616	8190	6880
X^2	6889	9409	6400	9025	5329	6084	8281	7396
Y^2	6084	9025	6889	8649	6084	5184	8100	6400
$\sum$	683	669	58544	59413	54315			

$$n = 8, \quad \sum X = 683, \quad \sum Y = 669, \quad \sum XY = 58544,$$

$$\sum X^2 = 59413, \quad \sum Y^2 = 54315$$

Compute Means and Slope

$$\bar{X} = \frac{683}{8} = 85.375, \quad \bar{Y} = \frac{669}{8} = 83.625$$

Regression of Y on X:

$$\hat{\beta}_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{8(58544) - 683 \cdot 669}{8(59413) - (683)^2}$$

$$\hat{\beta}_{yx} = \frac{468352 - 456927}{475304 - 466489} = \frac{11425}{8815} \approx 1.295$$

$$\hat{\alpha}_{yx} = \bar{Y} - \hat{\beta}_{yx}\bar{X} = 83.625 - 1.295 \cdot 85.375 \approx -27.9$$

Regression line:

$$Y = -27.9 + 1.295X$$

Estimate Y When X = 85

$$Y = -27.9 + 1.295 \cdot 85 = -27.9 + 110.075 = 82.18$$

Prediction:
When a student scores 85 in Math, their predicted Physics score is approximately **82.18**.

Regression of X on Y

$$\hat{\beta}_{xy} = \frac{n\sum XY - \sum X \sum Y}{n\sum Y^2 - (\sum Y)^2} = \frac{468352 - 456927}{434520 - 447561} = \frac{11425}{-13041} \approx -0.876$$

$$\hat{\alpha}_{xy} = \bar{X} - \hat{\beta}_{xy}\bar{Y} = 85.375 - (-0.876)(83.625) \approx 85.375 + 73.27 = 158.64$$

Regression line:

$$X = 158.64 - 0.876Y$$

Estimate X when $Y = 84$:

$$X = 158.64 - 0.876 \cdot 84 \approx 158.64 - 73.58 = 85.06$$

Prediction: A student who scores 84 in Physics is predicted to score **85.06** in Math.

Regression Line Property

Theorem: The regression lines of Y on X and X on Y both pass through the point:

$$(\bar{x}, \bar{y}) \quad \text{where} \quad \bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i$$

Regression Line Property

Theorem: The regression lines of Y on X and X on Y both pass through the point:

$$(\bar{x}, \bar{y}) \quad \text{where} \quad \bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i$$

This point is the **centroid** (mean point) of the dataset.

Proof for Regression Line of Y on X

The regression equation of Y on X is:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

From the normal equation:

$$\sum y_i = n\hat{\alpha} + \hat{\beta}\sum x_i$$

Divide both sides by n:

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

The point $(\bar{x}, \bar{y})$ lies on the regression line of Y on X.

Proof for Regression Line of X on Y

The regression line of X on Y is:

$$\hat{x} = a + by$$

From the normal equation:

$$\sum x_i = na + b\sum y_i$$

Divide by n:

$$\bar{x} = a + b\bar{y}$$

The point $(\bar{x}, \bar{y})$ also lies on the regression line of X on Y.

Therefore:

- Both regression lines pass through the mean point $(\bar{x}, \bar{y})$.
- This point is common to both the line of best fit for Y on X and X on Y.
- It reflects the overall trend of the data.

$$\boxed{(\bar{x}, \bar{y}) \text{ lies on both regression lines}}$$

Example

A study shows the relationship between advertising spend X (in \$1000s) and sales revenue Y (in \$1000s) for five weeks.

Advertising (X)	1	2	3	4	5
Sales (Y)	3	4	5	7	8

- Find the regression line of Y on X
- Interpret the slope and intercept
- Predict sales when advertising is \$6000

The regression line of Y on X is given by:

$$Y = \hat{\alpha} + \hat{\beta}X$$

Where:

$$\hat{\beta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Given:

$$\sum x_i = 1 + 2 + 3 + 4 + 5 = 15, \quad \sum y_i = 3 + 4 + 5 + 7 + 8 = 27$$

$$\sum x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55,$$

$$\sum x_i y_i = 1(3) + 2(4) + 3(5) + 4(7) + 5(8) = 92$$

$$\bar{x} = \frac{15}{5} = 3, \quad \bar{y} = \frac{27}{5} = 5.4$$

$$\hat{\beta} = \frac{5(92) - (15)(27)}{5(55) - (15)^2} = \frac{460 - 405}{275 - 225} = \frac{55}{50} = 1.1$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 5.4 - 1.1(3) = 5.4 - 3.3 = 2.1$$

$$\text{Regression Line: } \hat{Y} = 2.1 + 1.1X$$

What is the expected sales revenue when advertising spending is \$6000?

$$X = 6 \Rightarrow \hat{Y} = 2.1 + 1.1(6) = 2.1 + 6.6 = 8.7$$

Predicted Sales Revenue: \$8700

Example: Regression Analysis

The following data shows the number of hours students studied (X) and their scores (Y) in a statistics test.

Hours Studied (X)	2	3	5	7	9
Test Score (Y)	50	55	65	70	80

Tasks:

- Find the regression line of Y on X.
- Estimate the score of a student who studied for 6 hours.

Solution: Regression Line of Y on X

We use the regression equation:

$$\hat{Y} = a + bX$$

where:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \quad a = \bar{y} - b\bar{x}$$

Calculate necessary sums

$$\sum X = 2 + 3 + 5 + 7 + 9 = 26, \quad \sum Y = 50 + 55 + 65 + 70 + 80 = 320$$

$$\sum XY = (2)(50) + (3)(55) + (5)(65) + (7)(70) + (9)(80) = 1795$$

$$\sum X^2 = 2^2 + 3^2 + 5^2 + 7^2 + 9^2 = 168$$

Compute slope b and intercept a

$$b = \frac{5(1795) - (26)(320)}{5(168) - (26)^2} = \frac{8975 - 8320}{840 - 676} = \frac{655}{164} \approx 3.9939$$

$$\bar{x} = \frac{26}{5} = 5.2, \quad \bar{y} = \frac{320}{5} = 64$$

$$a = 64 - 3.9939(5.2) \approx 64 - 20.768 \approx 43.232$$

Regression line:

$$\hat{Y} = 43.232 + 3.9939X$$

Prediction

Estimate: What is the expected test score if a student studies for 6 hours? Substitute $X = 6$ into the regression line:

$$\hat{Y} = 43.232 + 3.9939(6) = 43.232 + 23.9634 \approx 67.20$$

Answer: A student who studies for 6 hours is expected to score approximately **67.2**.

References

- Hogg,R;Mckean,J;Craig,A(2012).Introduction to mathematical statistics, 7th edition, pearson Prentice Hall, 2012.

- Hastings K.J,(1997) Probability and statistics, Addison Wesley reading,massachusetts.

Thank You!

Next Lecture: Confidence interval for regression coefficient