DegreeLive

BBA | Semester 6 | UK6DSCMGT303

Statistics for Business Decisions

Course Code: UK6DSCMGT303 • Module 4 Notes

degreelive.in

Module 4: Regression Analysis

Linear Equations, Least Squares, Slopes, Mathematical Theorems, Standard Error, and Applications

Chapter 1: The Epistemological Nature and Purpose of Regression Analysis

1.1 Historical Origin and Definitional Framework

While correlation analysis quantifies the strength and directional sympathy of association between two variables, it does not provide an operational mathematical mechanism for predicting the unknown numerical value of one variable from the known value of another. In corporate management, executives do not merely ask whether advertising correlates with sales; they demand a precise predictive formula: "If we allocate an additional Rs. 25 lakhs to digital advertising next quarter, exactly how many crores of sales revenue can we expect to generate?" The statistical methodology that models functional dependencies and generates empirical predictions is **Regression Analysis**.

Historically, the term **Regression** (literally meaning "stepping back" or "returning to an average state") was introduced in 1877 by Sir Francis Galton in his biological studies on heredity. Galton observed that although tall parents tended to have tall children, the average height of children born to exceptionally tall parents was shorter than their parents' height, trending backward or "regressing" toward the average height of the broader population. Similarly, exceptionally short parents tended to have children whose average height regressed upward toward the population norm. In modern statistical science, regression has transcended this hereditary context to become the universal mathematical discipline of functional modeling, predictive estimation, and causal econometric inference.

M. M. Blair defines regression concisely: "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data." Morris Hamburg states: "The term regression analysis refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."

1.2 Operational Components of the Regression Relationship

In bivariate linear regression, the two variables occupy distinct mathematical roles:

- **Dependent Variable (Criterion / Regressand / Response Variable):** Conventionally denoted by the symbol Y, this is the primary variable whose values are to be predicted, forecasted, or explained by the model (e.g., corporate sales revenue, crop yield, consumer demand).
- **Independent Variable (Predictor / Regressor / Explanatory Variable):** Conventionally denoted by the symbol X, this is the variable whose values are known, preset, or controlled by

the decision-maker, deployed to estimate the unknown dependent variable (e.g., advertising expenditure, chemical fertilizer dosage, unit selling price).

Chapter 2: The Critical Distinction between Correlation and Regression

Although correlation and regression are intimately interconnected branches of bivariate statistical theory, they possess profound philosophical, structural, and operational differences:

Analytical Dimension	Correlation Analysis	Regression Analysis
Core Analytical Objective	Measures the **degree, strength, and direction** of association between variables.	Estimates the **average functional value** of one variable from the known value of another.
Nature of Variables	Both variables are treated symmetrically as random co-varying entities; no cause-and-effect relationship is assumed.	Variables are asymmetrical: one is established as the **Independent Variable (X)** and the other as the **Dependent Variable (Y)**.
Dimensional Units	A pure, dimensionless relative number bounded strictly between **-1.0 and +1.0**.	Expressed in the **concrete physical units of measurement** of the dependent variable per unit of the independent variable (e.g., Rs. of sales per rupee of advertising).
Symmetry of Computation	Completely symmetrical: the correlation between X and Y is identical to that between Y and X ($r_{xy} = r_{yx}$).	Asymmetrical: the regression line of Y on X is mathematically distinct from the regression line of X on Y ($b_{yx} \neq b_{xy}$).
Impact of Scale and Origin	Completely independent of changes in both origin and scale.	Independent of change of origin, but **directly responsive and sensitive to change of scale**.
Number of Equations	Produces a single statistical coefficient (r).	Generates two distinct predictive regression equations (Y on X, and X on Y).

Chapter 3: The Principle of Least Squares and the Two Lines of Regression

3.1 Why There Are Exactly Two Regression Lines

In bivariate analysis, students frequently question why two distinct lines of regression exist for a single dataset. In Cartesian geometry, a single straight line connects two variables. However, in statistical regression, the two lines fulfill two fundamentally distinct predictive functions:

1. **Regression Line of Y on X:** Deployed exclusively when management seeks to **estimate or predict the value of Y** corresponding to a given known value of X. It treats X as independent and minimizes the sum of squared **vertical deviations** (errors in Y).
2. **Regression Line of X on Y:** Deployed exclusively when management seeks to **estimate or predict the value of X** corresponding to a given known value of Y. It treats Y as independent and minimizes the sum of squared **horizontal deviations** (errors in X).

3.2 The Mathematical Criterion of Least Squares

The **Principle of Least Squares**, formulated independently by Adrien-Marie Legendre (1805) and Carl Friedrich Gauss (1809), provides the mathematical foundation for fitting regression lines. The best-fitting line is mathematically defined as that line which satisfies two simultaneous conditions:

1. The algebraic sum of the deviations of actual observed values from their corresponding estimated points on the regression line is equal to zero:

$$\sum(Y - \hat{Y}) = 0 \quad \text{and} \quad \sum(X - \hat{X}) = 0$$

2. The sum of the squared deviations of actual values from the fitted line is at a minimum:

$$\sum(Y - \hat{Y})^2 = \text{Minimum} \quad \text{and} \quad \sum(X - \hat{X})^2 = \text{Minimum}$$

3.3 Relative Spatial Orientation of the Two Regression Lines

The geometric relationship between the two regression lines reveals the underlying correlation between the variables:

- **Intersection Coordinates:** Both regression lines intersect at the point of their arithmetic means: $(\bar{X}, \bar{Y})$\$.
- **Perfect Correlation ($r = \pm 1.0$):** When correlation is perfect, both regression lines coincide into a single identical line. There is zero scatter.
- **Zero Correlation ($r = 0$):** When variables are completely uncorrelated, the two regression lines are perpendicular to each other (90°\$). The line of Y on X runs horizontally parallel to the X-axis (Y = \bar{Y}), and the line of X on Y runs vertically parallel to the Y-axis (X = \bar{X}).
- **Moderate Correlation:** The acute angle (θ) between the two regression lines reflects the strength of association: the higher the correlation, the narrower the angle between the lines. The angle is formulated as:

$$\tan\theta = \left(\frac{1-r^2}{|r|}\right)\left(\frac{\sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

Chapter 4: Mathematical Theory and Properties of Regression Coefficients

4.1 Definitional Architecture

The slope of a linear regression line is quantified by its **Regression Coefficient**. Bivariate analysis features two distinct regression coefficients:

1. Regression Coefficient of Y on X (b_{yx})

Measures the expected average change in the dependent variable Y resulting from a one-unit change in the independent variable X. It represents the slope of the regression line of Y on X:

$$b_{yx} = r\frac{\sigma Y}{\sigma X} = \frac{\text{Cov}(X,Y)}{\sigma^2 X} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\sum xy}{\sum x^2}$$

2. Regression Coefficient of X on Y (b_{xy})

Measures the expected average change in variable X resulting from a one-unit change in variable Y. It represents the slope of the regression line of X on Y:

$$b_{xy} = r\frac{\sigma X}{\sigma Y} = \frac{\text{Cov}(X,Y)}{\sigma^2 Y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2} = \frac{\sum xy}{\sum y^2}$$

4.2 The Five Core Mathematical Theorems of Regression Coefficients

1. **Geometric Mean Property:** The correlation coefficient is the geometric mean of the two regression coefficients:

   $$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

   Significance: If $b_{yx} = 0.8$ and $b_{xy} = 0.45$, then $r = \sqrt{0.8 \times 0.45} = \sqrt{0.36} = +0.60$.

2. **Congruence of Algebraic Signs:** Both regression coefficients (b_{yx} and b_{xy}) and the correlation coefficient (r) must always possess the **identical algebraic sign**. If one is positive, all three must be positive; if one is negative, all three must be negative. It is mathematically impossible for b_{yx} to be positive while b_{xy} is negative.

3. **Reciprocal Magnitude Constraint:** If one of the regression coefficients is numerically greater than unity (> 1.0), the other regression coefficient must be strictly less than unity (< 1.0):

   $$b_{yx} \times b_{xy} = r^2 \leq 1.0$$

   Because r^2 cannot exceed 1.0, both regression coefficients cannot simultaneously be greater than 1.0.

4. **Arithmetic Mean Superiority Theorem:** The arithmetic mean of the two regression coefficients is always greater than or equal to the correlation coefficient:

$$\frac{b_{yx}+b_{xy}}{2} \geq r$$

5. **Effect of Change of Origin and Scale:** Regression coefficients are completely independent of changes of origin, but are directly responsive to changes of scale. If $u = \frac{X-A}{c1}$ and $v = \frac{Y-B}{c2}$:

$$b_{yx} = \left(\frac{c2}{c1}\right) b_{vu} \quad \text{and} \quad b_{xy} = \left(\frac{c1}{c2}\right) b_{uv}$$

Chapter 5: Computation of Linear Regression Equations

5.1 The Deviation Form (Standard Operational Method)

When arithmetic means $\bar{X}$ and $\bar{Y}$ are known integers, regression equations are formulated directly around the means:

1. Regression Equation of Y on X:

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y = \bar{Y} - b_{yx}\bar{X} + b_{yx}X \implies Y = a + b_{yx}X$$

Where: a = \bar{Y} - b_{yx}\bar{X represents the Y-intercept.

2. Regression Equation of X on Y:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X = \bar{X} - b_{xy}\bar{Y} + b_{xy}Y \implies X = c + b_{xy}Y$$

Where: c = \bar{X} - b_{xy}\bar{Y represents the X-intercept.

5.2 The Raw Data Normal Equations Form

When computing directly from raw uncentered data via the classical method of least squares, the parameters of the line $Y = a + bX$ are obtained by solving two simultaneous **Normal Equations**:

$$\text{Equation 1: } \sum Y = Na + b\sum X$$

$$\text{Equation 2: } \sum XY = a\sum X + b\sum X^2$$

Similarly, for the line of X on Y ($X = c + dY$):

> Equation 1: $\sum X = Nc + d\sum Y$
>
> Equation 2: $\sum XY = c\sum Y + d\sum Y^2$

Chapter 6: The Standard Error of Estimate and Forecasting Confidence Bands

6.1 Concept and Mathematical Formulation

Even the most meticulously fitted regression line does not guarantee that every future empirical observation will land precisely upon the line. Residual errors occur due to unobserved environmental shocks. The **Standard Error of Estimate** (S_{yx} or S_e\$) measures the dispersion or scatter of actual observed points around the fitted regression line, functioning as the standard deviation of regression residuals:

> $S_{yx} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{N-2}} = \sigma_Y\sqrt{1-r^2}\cdot\sqrt{\frac{N}{N-2}}$
>
> Standard Large-Sample Formula: $S_{yx} = \sigma_Y\sqrt{1-r^2}$
>
> Standard Error of Estimate of X on Y: $S_{xy} = \sigma_X\sqrt{1-r^2}$

6.2 Managerial Interpretation and Forecasting Confidence Bands

- If $r = \pm 1.0$, then $S_{yx} = 0$, indicating perfect prediction with zero forecasting error.
- If $r = 0$, then $S_{yx} = \sigma_Y$, indicating that the regression model provides zero predictive improvement over simply guessing the population mean $\bar{Y}$.
- Under the assumption of normally distributed residuals, management constructs **Forecasting Confidence Bands** around any point prediction $\hat{Y}$\$:
  - $\hat{Y} \pm 1.0 \cdot S_{yx}$ provides approximately **68.27% confidence**.
  - $\hat{Y} \pm 1.96 \cdot S_{yx}$ provides exactly **95.0% confidence**.
  - $\hat{Y} \pm 2.58 \cdot S_{yx}$ provides exactly **99.0% confidence**.

Chapter 7: Comprehensive Worked Numerical Problems in Regression Analysis

7.1 Worked Problem 1: End-to-End Computation of Both Regression Equations

A consumer goods enterprise analyzed the monthly promotional advertising expenditure (X in lakhs of rupees) and corresponding sales revenue (Y in lakhs of rupees) across five retail regions:

Region	Advertising X	Sales Y	x = X - 6	y = Y - 10	x^2	y^2	x · y
1	2	4	-4	-6	16	36	+24
2	4	7	-2	-3	4	9	+6
3	6	10	0	0	0	0	0
4	8	13	+2	+3	4	9	+6
5	10	16	+4	+6	16	36	+24
Total	**Sum X = 30**	**Sum Y = 50**	**Sum x = 0**	**Sum y = 0**	**Sum x^2 = 40**	**Sum y^2 = 90**	**Sum xy = 60**

Required:

1. Calculate arithmetic means $\bar{X}$ and $\bar{Y}$.
2. Calculate regression coefficients b_{yx} and b_{xy}.
3. Derive the Regression Equation of Y on X.
4. Derive the Regression Equation of X on Y.
5. Estimate sales revenue (Y) if advertising spending is expanded to Rs. 15 lakhs.
6. Estimate required advertising budget (X) to achieve a target sales revenue of Rs. 25 lakhs.
7. Calculate the correlation coefficient r using the regression coefficients.

Step-by-Step Mathematical Resolution:

1. Arithmetic Means:

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6.0 \text{ lakhs}$$
$$\bar{Y} = \frac{\sum Y}{N} = \frac{50}{5} = 10.0 \text{ lakhs}$$

2. Regression Coefficients:

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{60}{40} = 1.50$$
$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{60}{90} = 0.6667 \approx 0.67$$

DegreeLive

> **Interpretation:** $b_{yx} = 1.50$ signifies that for every additional Rs. 1.00 lakh allocated to advertising, sales revenue expands on average by Rs. 1.50 lakhs.

3. Regression Line of Y on X:

$$Y - \bar{Y} = b_{yx}(X - \bar{X}) \implies Y - 10 = 1.50(X - 6)$$
$$Y - 10 = 1.50X - 9.00 \implies Y = 1.00 + 1.50X$$

4. Regression Line of X on Y:

$$X - \bar{X} = b_{xy}(Y - \bar{Y}) \implies X - 6 = 0.6667(Y - 10)$$
$$X - 6 = 0.6667Y - 6.6667 \implies X = -0.67 + 0.6667Y$$

5. Estimation of Sales Y for Advertising X = 15 Lakhs:

$$\hat{Y} = 1.00 + 1.50(15) = 1.00 + 22.50 = 23.50 \text{ lakhs of rupees}$$

6. Estimation of Required Advertising X for Target Sales Y = 25 Lakhs:

$$\hat{X} = -0.6667 + 0.6667(25) = -0.6667 + 16.6667 = 16.00 \text{ lakhs of rupees}$$

7. Calculation of Correlation Coefficient:

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{1.50 \times 0.6667} = \sqrt{1.00} = +1.00$$

Both regression slopes are positive, confirming perfect positive linear association ($r = +1.0$).

7.2 Worked Problem 2: Reconstructing Statistical Metrics from Two Regression Equations

An econometric research firm investigating industrial employment and output provides two estimated regression equations:

> Equation 1: $8X - 10Y + 66 = 0$
>
> Equation 2: $40X - 18Y = 214$
>
> Given: Variance of X (σ_X^2) = 9.

Required:

1. Determine arithmetic means $\bar{X}$ and $\bar{Y}$.
2. Identify which equation is Y on X and which is X on Y, and find regression coefficients b_{yx} and b_{xy}.
3. Calculate the correlation coefficient r.
4. Calculate the standard deviation of Y (σ_Y).

Step-by-Step Mathematical Resolution:

1. Solve for Means by Intersecting Equations:

$$8\bar{X} - 10\bar{Y} = -66 \quad \text{— (1)}$$

$$40\bar{X} - 18\bar{Y} = 214 \quad \text{— (2)}$$

$$\text{Multiply (1) by 5: } 40\bar{X} - 50\bar{Y} = -330 \quad \text{— (3)}$$

$$\text{Subtract (3) from (2): } [-18\bar{Y}] - [-50\bar{Y}] = 214 - (-330)$$

$$32\bar{Y} = 544 \implies \bar{Y} = \frac{544}{32} = 17.0$$

$$\text{Substitute into (1): } 8\bar{X} - 10(17) = -66 \implies 8\bar{X} - 170 = -66$$

$$8\bar{X} = 104 \implies \bar{X} = \frac{104}{8} = 13.0$$

Arithmetic means are $\bar{X} = 13.0$ and $\bar{Y} = 17.0$.

2. Identify Regression Lines and Compute Slopes:

- Assume Equation 1 is Y on X:

$$10Y = 8X + 66 \implies Y = 0.80X + 6.6 \implies b_{yx} = 0.80$$

- Then Equation 2 is X on Y:

$$40X = 18Y + 214 \implies X = \frac{18}{40}Y + \frac{214}{40} \implies b_{xy} = 0.45$$

- Check product condition:

$$b_{yx} \times b_{xy} = 0.80 \times 0.45 = 0.36 \leq 1.0$$

Product is less than 1.0! Our assumption is validated and correct.

3. Correlation Coefficient:

$$r = +\sqrt{b_{yx} \times b_{xy}} = \sqrt{0.36} = +0.60$$

4. Standard Deviation of Y:

$$\sigma X = \sqrt{9} = 3.0$$
$$b_{yx} = r\frac{\sigma Y}{\sigma X} \implies 0.80 = 0.60\left(\frac{\sigma Y}{3.0}\right)$$
$$0.80 = 0.20 \cdot \sigma Y \implies \sigma Y = \frac{0.80}{0.20} = 4.0$$

The standard deviation of Y is exactly **4.0** (and Variance of Y = 16).

7.3 Worked Problem 3: Standard Error of Estimate and 95% Forecasting Confidence Band

An industrial power distribution utility has established the following empirical regression parameters linking monthly electricity demand (Y in Megawatt-hours) to ambient industrial temperature (X in degrees Celsius):

- Regression Equation of Y on X: $\hat{Y} = 120.0 + 4.50X$
- Correlation Coefficient: $r = +0.80$
- Standard Deviation of Electricity Demand: $\sigma Y = 15.0$ MWh
- Sample size of historical operating months: $N = 66$

Required:

1. Calculate the Standard Error of Estimate (S_{yx}) for the regression model.
2. Forecast expected electricity demand when next month's temperature is projected at 30 degrees Celsius.
3. Construct the 95% statistical confidence interval for this forecast.

Step-by-Step Mathematical Resolution:

1. Standard Error of Estimate:

$$S_{yx} = \sigma Y \sqrt{1 - r^2} = 15.0 \times \sqrt{1 - (0.80)^2} = 15.0 \times \sqrt{1 - 0.64}$$
$$S_{yx} = 15.0 \times \sqrt{0.36} = 15.0 \times 0.60 = 9.00 \text{ MWh}$$

Interpretation: The standard deviation of actual observed demand values around the fitted regression line is 9.00 MWh.

2. Point Forecast at X = 30 Degrees Celsius:

$$\hat{Y} = 120.0 + 4.50(30) = 120.0 + 135.0 = 255.00 \text{ MWh}$$

3. Construction of 95% Confidence Interval:

$$95\% \text{ Confidence Limits} = \hat{Y} \pm 1.96(S_{yx}) = 255.00 \pm 1.96(9.00)$$

$$1.96 \times 9.00 = 17.64 \text{ MWh}$$

$$\text{Lower Limit} = 255.00 - 17.64 = 237.36 \text{ MWh}$$

$$\text{Upper Limit} = 255.00 + 17.64 = 272.64 \text{ MWh}$$

Executive Risk Insight: Management can be 95% confident that actual grid electricity demand during the upcoming month will fall comfortably within the band between **237.36 MWh and 272.64 MWh**, providing clear boundaries for fuel procurement and generator scheduling.

Chapter 8: Managerial Decision Framework for Choosing Between Correlation and Regression

8.1 Decision Protocols in Enterprise Analytics

Corporate analysts select between correlation and regression based on five strategic decision criteria:

1. **Exploratory Association vs. Actionable Intervention:** If the goal is purely diagnostic (e.g., assessing whether employee satisfaction correlates with customer retention), correlation is sufficient. If leadership plans to intervene (e.g., setting compensation budgets to hit customer targets), regression is required.
2. **Direction of Predictive Power:** Regression demands identifying a clear dependent outcome. When variables mutually reinforce each other without clear direction, correlation represents the more scientifically honest metric.
3. **Model Generalizability and Forecasting Horizon:** Regression models assume that underlying structural relationships remain stable over time. In volatile market disruptions, historical regression slopes may decay rapidly, requiring continuous econometric recalibration.

DegreeLive