Linear Regression and Correlation

Introduction

Many real-world situations involve understanding relationships between two or more numerical variables includ:

- Education & Income: Does a higher education level lead to a higher salary?
- Exam Performance: Does a stuckentis grade on a midterm predict their final exam score?
- Repair costs: How does the total cost of a repair depend on hourly labor fees?

The data in these examples are called bivariate data because they involve two variables. More complex analysis use multivariate data (three or more variables). This focuses on the simplest type of regression: linear regression with one independent variable (X). We also study correlation, which measures how strong the relationship is between two variables

Key concepts

Measures the strength and direction of the relationship between X and Y

The correlation coefficient (r) ranges from (-1) to (+1).

$r = +1$ → Perfect positive relationship (as X increases, Y increases)

$r = -1$ → perfect negative relationship (as X increases, Y decreases)

$r = 0$ → no correlation

3 Outlier

Outliers are data points that do not follow the pattern of the rest of the data

They can affect regression results and correlation calculations

Example: A student who scores low on midterms but high on the final exam may be an outlier

Linear Equations in Regression

Linear regression is based on the equation of a straight line, written as: $y = a + bx$

Where

x = Independent variable, y = dependent variable, a = Y-Intercept (value of y when $x = 0$) b = slope (rate of change)

Date / /
Page No.

Examples of linear Equations

1. $y = 3 + 2x$
   When $x = 0$, $y = 3$
   for every increase of 1 in x, y increases by 2
2. $y = -0.01 + 1.2x$
   When $x = 0$, $y = -0.01$
   for every increase of 1 in x, increases by 1.2

Key fact: The graph of any linear equation is a straight line (unless it is vertical)

Real-life Example: Duron's word processing service (AWPS)
AWPS charges: $32 per hour
: One time setup fee of $31.50

Step 1: Define variable
- Let x = number of hours needed for the job
- let y = total cost

Step 2: Write the Equation
- Fixed cost = $31.50
- Hourly rate = $32 per hour

$y = 31.50 + 32x$

This equation models total cost based on the number of hours worked.

Example calculation:
If the job takes 5 hours, the total cost is
$y = 31.50 + (32 \times 5) = 31.50 + 160 = 191.50$
so the total charge for 5 hours is $191.50

Why are linear Equation Useful?
predicts costs & values based on given inputs
Analyze trends in business, economics, psychology & science
Create models for decision making

Interpreting the slope (b)
The slope b determines the direction and steepness of the line
if $b > 0$: The line slopes upward (positive relationship)
if $b = 0$ The line is horizontal (no change)
if $b < 0$: The line slopes downward (negative relationship

Understanding the correlation coefficient (r) and coefficient of Determination (r^2)
The correlation coefficient (r) measures the strength and direction of a linear relationship between two variables.

Interpreting the correlation coefficient (r)
Range : $-1 \leq r \leq 1$ $r = 1 \rightarrow$ perfect positive correlation
$r = -1 \rightarrow$ perfect negative correlation
$r = 0 \rightarrow$ No linear correlation
Strength of relationship : $|r| \approx 1 \rightarrow$ strong correlation
$|r| \approx 0.5 \rightarrow$ Moderate correlation
$|r| \approx 0 \rightarrow$ weak or no correlation

Coefficient of Determination (r^2)
What does r^2 tell us?
r^2 is the square of r and is expressed as a percentage
Represents the percentage of variation in y explained by x

Example

- given $r = 0.6637$, then:

$$r^2 = (0.6637)^2 = 0.4397 \approx 44\%$$

Interpretation

- 44% of the variation in final exam scores is explained by third exam scores
- The remaining 56% is due to other factors (eg study habits, sleep, stress)

Summary of Partitioning the sums of squares

1. Understanding the Sum of Squares (SSY)

The total variation in Y is measured using sum of squares Y (SSY), which represent the sum of squared deviation of each Y value from the mean of Y:

$$SSY = \Sigma (Y - \bar{Y})^2$$

Eg Dataset

Mean of $Y = 2.06$

$SSY = 4.397$ (sum of squared deviation from the mean)

2. Converting Raw scores to deviation scores

- Instead of using raw scored y it is useful to compute deviation scores, represented as $y = Y - MY$
- Conventionally small letters (y) denote deviations from the mean

3. Computing predicted scores from the Regression Equation

The predicted value Y' is obtained using the regression equation $Y' = 0.425X + 0.785$

Eg. for $X = 1$ $\quad Y' = 0.425(1) + 0.785 = 1.210$

Date / /
Page No.

4 Partitioning SSY into SSY' and SSE

SSY (Total variation) can be divided into

SSY' (Sum of Squares Predicted): Variation of Predicted Y' values from the mean

SSE (Sum of Squares Error): Validation of actual Y values from Y' (unexplained variation)

Formula : $SSY = SSY' + SSE$

Eg calculation $4.597 = 1.806 + 2.79$

5 Interpretation of r^2

r^2 represents the proportion of explained variation

$$r^2 = \frac{SSY'}{SSY}$$

Eg: If $r = 0.4$, then $r^2 = 0.16$, meaning 16% of variation is explained by the predictor variable

Re proportion of unexplained variation : $\frac{SSE}{SSY}$

Summary of standard error of the estimate

1. Understanding the Standard error of the estimate (Sest)
- Re standard error of the estimate (Sest) measures the accuracy of regression predictions
- A smaller standard error means the predictions are more accurate (points closer to the regression line), while a larger error indicates more spread (less accuracy)

2) Formula for standard error of the estimate

The standard error is calculated using

$$Sest = \sqrt{\frac{\Sigma(Y - Y')^2}{N}}$$

Date / /
Page No.

3 Example Calculation: (Population Case, N=5)
from the dataset

X	Y	y'	$Y-y'$	$(Y-y')^2$
1.0	1.00	1.210	-0.210	0.044
2.0	2.00	1.635	0.365	0.133
3.0	1.30	2.060	-0.760	0.578
4.0	3.75	2.485	1.265	1.600
5.0	2.25	2.910	-0.660	0.436
Sum	15.00	10.30	0.00	2.791

4 Computing standard Error wrt Pearson's correlation (p)
An alternative formula relates σ_{est} to Pearson's Correlation

$$\sigma_{est} = \sqrt{(1-p^2)\frac{SSY}{N}}$$

where; $p = 0.6768$ (Pearson correlation)
$SSY = 4.597$ (Total sum of squares)
$N = 5$
substituting values = $\sigma_{est} = 0.747$

5 Standard error of the estimate for a sample
When estimating from a sample, the denominator changes from N to N-2

Why use N-2?
Two parameters (slope and intercept) are estimated before computing SSY

Inferential statistics for regression slope (b) and correlation (ρ)

1) Assumptions for inferential statistics in Regression

Inferential statistics for regression rely on several assumptions about the population (not just the sample):

i. Linearity - The relationship between X (predictor) and Y (criterion) is linear

ii. Homoscedasticity - The variance of residuals (errors) is constant across all values of X
- If the spread of residuals changes (e.g. wider spread at one end), the assumption is violated (Heteroscedasticity)
- Example (Figure 3): Predictions are more accurate for students with higher GPAs, but less accurate for lower GPAs → Heteroscedasticity

iii. Normality of errors: The residuals (deviations from the regression lines) follow a normal distribution

Note: This does not mean X or Y is normally distributed - only the errors should be

2 Significance Test for the slope (b)

We test whether the regression slope (b) is significantly different from zero. (i.e. whether X significantly predicts Y)

formula for t-test on the slope (b) $t = \frac{b - 0}{s_b}$

Standard error of the slope (S-b)

$$s_b = \frac{s_{est}}{\sqrt{SS_X}}$$

Date / /
Page No.

3 Confidence Interval for the Slope (b)

A 95% confidence interval for the slope is $b \pm (t_{0.05} \cdot S_b)$

4 Significance Test for pearson's correlation (r)

To test whether the correlation (r) is significantly different from zero, we use

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$