Lecture Notes: Linear Regression & Correlation Analysis

PSTAT 5A

July 30, 2025

Contents

1 Understanding Relationships Between Variables

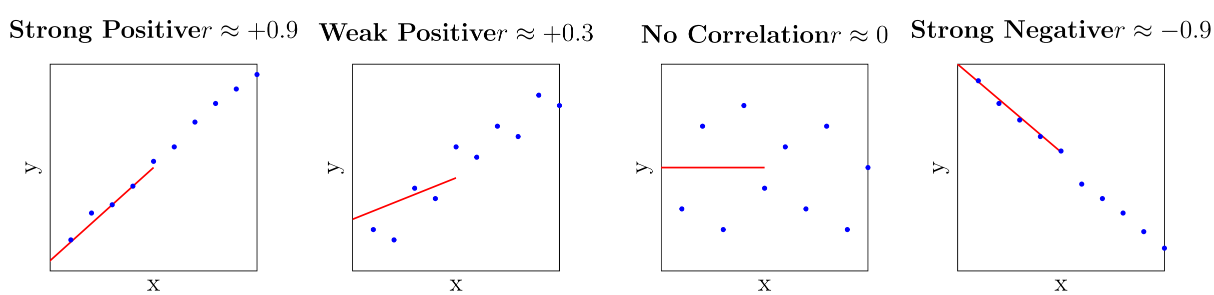
In statistics, we often want to understand how two quantitative variables are related to each other. For example:

- How does study time relate to exam scores?
- Is there a relationship between height and weight?
- Can we predict house prices based on square footage?

Key Concepts

- **Explanatory Variable (x):** The variable we use to explain or predict (independent variable)
- **Response Variable (y):** The variable we want to predict or explain (dependent variable)
- **Correlation:** Measures the strength and direction of a linear relationship
- **Regression:** Uses one variable to predict another variable

1.1 Types of Relationships



2 Correlation Coefficient

The correlation coefficient r (also called Pearson's correlation) measures the strength and direction of a linear relationship between two variables.

Definition 2.1 (Correlation Coefficient)**.** *The sample correlation coefficient is calculated as:*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Alternative computational formula:

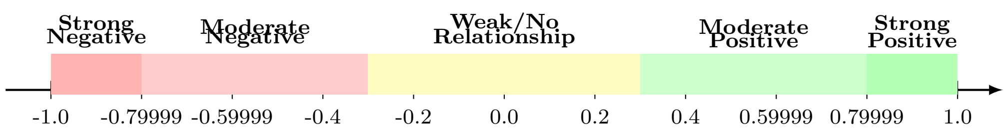
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

2.1 Properties of Correlation

Key Properties of r

1. **Range:** $-1 \leq r \leq +1$
2. **Direction:**
   - $r > 0$: Positive linear relationship
   - $r < 0$: Negative linear relationship
   - $r = 0$: No linear relationship
3. **Strength:**
   - $|r| \geq 0.8$: Strong relationship
   - $0.3 \leq |r| < 0.8$: Moderate relationship
   - $|r| < 0.3$: Weak relationship
4. **Units:** Correlation is unitless (no measurement units)
5. **Symmetry:** $r_{xy} = r_{yx}$

2.2 Correlation Interpretation Guide



Example 2.1 (Calculating Correlation). *Calculate the correlation between study hours (x) and exam scores (y):*

Student	***Hours (x)***	***Score (y)***	x^2	y^2	xy
1	2	65	4	4225	130
2	4	70	16	4900	280
3	6	80	36	6400	480
4	8	85	64	7225	680
5	10	90	100	8100	900
Sum	30	390	220	30850	2470

Using the computational formula:

$$
\begin{aligned}
r &= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} && (1)\\
&= \frac{5(2470) - (30)(390)}{\sqrt{[5(220) - (30)^2][5(30850) - (390)^2]}} && (2)\\
&= \frac{12350 - 11700}{\sqrt{[1100 - 900][154250 - 152100]}} && (3)\\
&= \frac{650}{\sqrt{(200)(2150)}} && (4)\\
&= \frac{650}{\sqrt{430000}} = \frac{650}{655.74} \approx 0.991 && (5)
\end{aligned}
$$

This indicates a very strong positive correlation between study hours and exam scores.

3 Simple Linear Regression

Linear regression allows us to model the relationship between two variables using a straight line, and make predictions.

Definition 3.1 (Simple Linear Regression Model)**.** *The simple linear regression model is:*

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y = *response variable*
- x = *explanatory variable*
- β_0 = *y-intercept (population parameter)*
- β_1 = *slope (population parameter)*
- ϵ = *random error term*

3.1 Sample Regression Line

Since we don't know the true population parameters, we estimate them from sample data:

Sample Regression Equation

$$\hat{y} = b_0 + b_1 x$$

where:

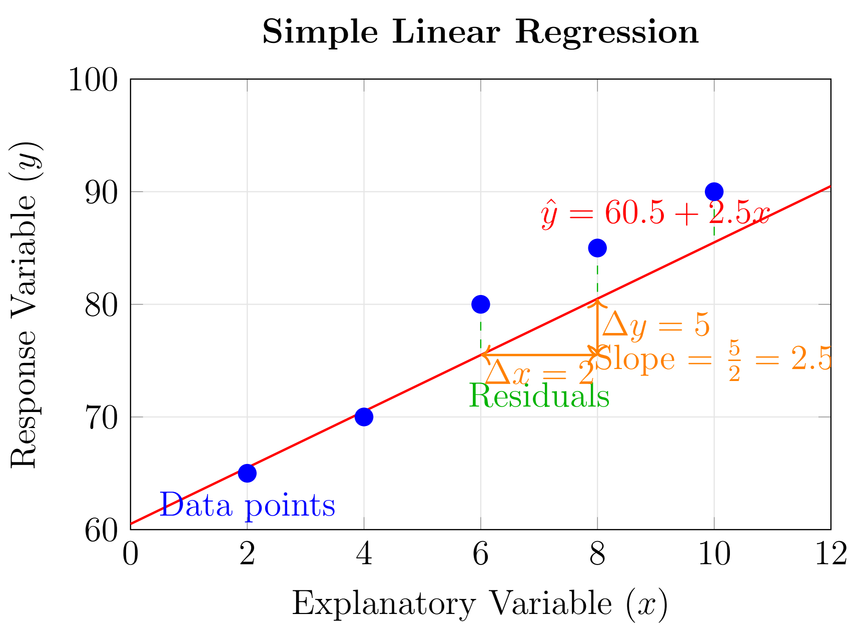
- $\hat{y}$ = predicted value of y
- b_0 = sample y-intercept
- b_1 = sample slope

Formulas:

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{6}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{7}$$

3.2 Visual Representation of Regression



3.3 Interpretation of Regression Components

Interpreting Regression Components

Slope (b_1):

- Represents the change in y for each one-unit increase in x
- Units: (units of y) per (unit of x)
- Example: "For each additional hour of study, exam score increases by 2.5 points on average"

Y-intercept (b_0):

- The predicted value of y when $x = 0$
- May or may not have practical meaning depending on context
- Example: "A student who studies 0 hours is predicted to score 60.5 points"

Example 3.1 (Finding the Regression Line)**.** *Using our study hours and exam scores data from earlier:*

Step 1: Calculate the slope

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \tag{8}$$
$$= \frac{2470 - 5(6)(78)}{220 - 5(6)^2} \tag{9}$$
$$= \frac{2470 - 2340}{220 - 180} \tag{10}$$
$$= \frac{130}{40} = 3.25 \tag{11}$$

Step 2: Calculate the y-intercept

$$b_0 = \bar{y} - b_1\bar{x} \tag{12}$$
$$= 78 - 3.25(6) \tag{13}$$
$$= 78 - 19.5 = 58.5 \tag{14}$$

Step 3: Write the regression equation

$$\hat{y} = 58.5 + 3.25x$$

Interpretation: *For each additional hour of study, exam score increases by 3.25 points on average.*

4 Making Predictions and Understanding Residuals

4.1 Predictions

Once we have the regression equation, we can make predictions for new values of x.

Making Predictions

Steps for prediction:

1. Substitute the x-value into the regression equation
2. Calculate $\hat{y} = b_0 + b_1 x$
3. Interpret the result in context

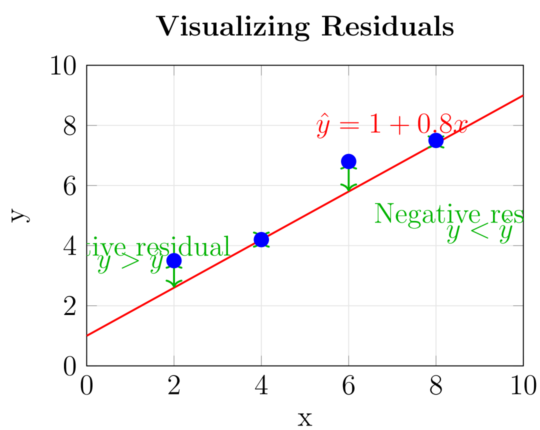
Important considerations:

- Only predict within the range of observed x-values (avoid extrapolation)
- Predictions are estimates with uncertainty
- The relationship may not hold outside the observed range

4.2 Residuals and Model Fit

Definition 4.1 (Residual)**.** *A residual is the difference between the observed value and the predicted value:*

$$\textit{Residual} = y - \hat{y} = \textit{Observed} - \textit{Predicted}$$



4.3 Properties of Residuals

Key Properties of Residuals

1. The sum of residuals equals zero: $\sum(y_i - \hat{y}_i) = 0$
2. Small residuals indicate good fit
3. Large residuals suggest outliers or poor model fit
4. Residual plots help check regression assumptions

5 Coefficient of Determination (R^2)

The coefficient of determination measures how much of the variation in y is explained by the regression line.

Definition 5.1 (Coefficient of Determination)**.**

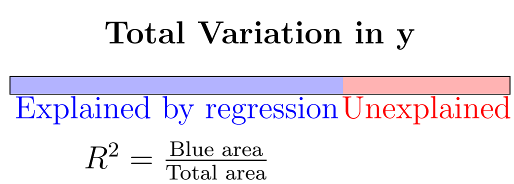
$$R^2 = r^2 = \frac{\textit{Variation explained by regression}}{\textit{Total variation in } y}$$

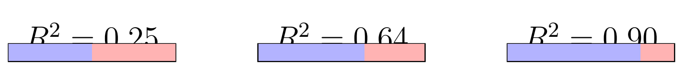
Alternative formulas:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{15}$$

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \tag{16}$$

5.1 Interpreting R^2





Interpreting R^2 Values

- **Range:** $0 \leq R^2 \leq 1$ (often expressed as percentage)
- $R^2 = 0$**:** Regression line explains 0% of variation (no linear relationship)
- $R^2 = 1$**:** Regression line explains 100% of variation (perfect linear relationship)
- $R^2 = 0.64$**:** "64% of the variation in y is explained by the linear relationship with x"

Rule of thumb:

- $R^2 \geq 0.70$: Strong predictive relationship
- $0.30 \leq R^2 < 0.70$: Moderate predictive relationship
- $R^2 < 0.30$: Weak predictive relationship

6 Conditions for Linear Regression

Before using linear regression, we must check that certain conditions are met.

CONDITIONS: LINE

Linear relationship between x and y
Independent observations
Normal distribution of residuals
Equal variance (homoscedasticity)

6.1 Checking Conditions

6.1.1 1. Linear Relationship

- Check scatterplot for linear pattern
- Look for curved or nonlinear patterns
- Consider transformations if relationship is not linear

6.1.2 2. Independence

- Observations should not be related to each other
- Random sampling helps ensure independence
- Be careful with time series data or clustered data

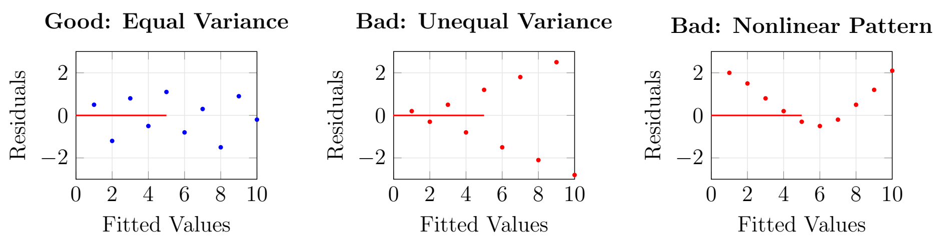
6.1.3 3. Normal Residuals

- Check histogram or normal probability plot of residuals
- Residuals should be approximately normally distributed
- Small departures from normality are often acceptable

6.1.4 4. Equal Variance

- Plot residuals vs. fitted values
- Look for constant spread (no fan shape)
- Residual spread should be similar across all x-values

6.2 Diagnostic Plots



7 Hypothesis Testing for Regression Slope

We can test whether there is a significant linear relationship between x and y by testing the slope.

Hypothesis Test for Slope

Hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{(no linear relationship)} \tag{17}$$
$$H_a : \beta_1 \neq 0 \quad \text{(linear relationship exists)} \tag{18}$$

Test statistic:

$$t = \frac{b_1 - 0}{SE_{b_1}} = \frac{b_1}{SE_{b_1}}$$

where SE_{b_1} is the standard error of the slope.
Distribution: t with $df = n - 2$

7.1 Standard Error of the Slope

$$SE_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

where s is the residual standard error:

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

Example 7.1 (Testing Regression Slope). *For our study hours and exam scores example, suppose we find:*

- $b_1 = 3.25$ *(slope)*
- $SE_{b_1} = 0.45$ *(standard error of slope)*
- $n = 5$ *students*

Test at $\alpha = 0.05$ whether there is a significant relationship.
Solution:

1. ***Hypotheses:*** $H_0 : \beta_1 = 0$ *vs.* $H_a : \beta_1 \neq 0$
2. ***Test statistic:***
$$t = \frac{3.25}{0.45} = 7.22$$
3. ***Degrees of freedom:*** $df = 5 - 2 = 3$
4. ***Critical value:*** $t_{0.025,3} = 3.182$
5. ***Decision:*** *Since* $|7.22| > 3.182$, *reject* H_0
6. ***Conclusion:*** *There is significant evidence of a linear relationship between study hours and exam scores.*

8 Complete Worked Example

Let's work through a comprehensive regression analysis.

Example 8.1 (House Prices and Square Footage). *A real estate agent collected data on 8 houses:*

House	***Sq Ft (x)***	***Price ($1000s) (y)***	x^2	y^2	xy
1	*1200*	*150*	*1,440,000*	*22,500*	*180,000*
2	*1500*	*180*	*2,250,000*	*32,400*	*270,000*
3	*1800*	*210*	*3,240,000*	*44,100*	*378,000*
4	*2000*	*240*	*4,000,000*	*57,600*	*480,000*
5	*2200*	*260*	*4,840,000*	*67,600*	*572,000*
6	*2500*	*290*	*6,250,000*	*84,100*	*725,000*
7	*2800*	*320*	*7,840,000*	*102,400*	*896,000*
8	*3000*	*350*	*9,000,000*	*122,500*	*1,050,000*
Sum	*17,000*	*2,000*	*38,860,000*	*533,200*	*4,551,000*

Part 1: Calculate basic statistics

$$\bar{x} = \frac{17,000}{8} = 2,125 \ sq \ ft \tag{19}$$

$$\bar{y} = \frac{2,000}{8} = 250 \ thousand \ dollars \tag{20}$$

Part 2: Calculate correlation

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{21}$$

$$= \frac{8(4,551,000) - (17,000)(2,000)}{\sqrt{[8(38,860,000) - (17,000)^2][8(533,200) - (2,000)^2]}} \tag{22}$$

$$= \frac{36,408,000 - 34,000,000}{\sqrt{[310,880,000 - 289,000,000][4,265,600 - 4,000,000]}} \tag{23}$$

$$= \frac{2,408,000}{\sqrt{(21,880,000)(265,600)}} \tag{24}$$

$$= \frac{2,408,000}{2,411,651} \approx 0.998 \tag{25}$$

Part 3: Find regression line

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \tag{26}$$

$$= \frac{4,551,000 - 8(2,125)(250)}{38,860,000 - 8(2,125)^2} \tag{27}$$

$$= \frac{4,551,000 - 4,250,000}{38,860,000 - 36,125,000} \tag{28}$$

$$= \frac{301,000}{2,735,000} \approx 0.110 \tag{29}$$

$$b_0 = \bar{y} - b_1\bar{x} \tag{30}$$

$$= 250 - 0.110(2,125) \tag{31}$$

$$= 250 - 233.75 = 16.25 \tag{32}$$

Regression equation: $\hat{y} = 16.25 + 0.110x$
Part 4: Interpretation

- ***Slope:*** *For each additional square foot, house price increases by $110 on average*
- ***Y-intercept:*** *A house with 0 square feet would cost $16,250 (not meaningful in context)*
- ***Correlation:*** $r = 0.998$ *indicates a very strong positive linear relationship*
- ***R^2:*** $R^2 = (0.998)^2 = 0.996$, *so 99.6% of price variation is explained by square footage*

Part 5: Make a prediction *Predict the price of a 2,400 square foot house:*

$$\hat{y} = 16.25 + 0.110(2,400) = 16.25 + 264 = 280.25$$

The predicted price is $280,250.

9 Summary and Quick Reference

Key Formulas Summary

Correlation:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Regression Line:

$$\hat{y} = b_0 + b_1 x \quad \text{where} \quad b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}, \quad b_0 = \bar{y} - b_1\bar{x}$$

Coefficient of Determination:

$$R^2 = r^2$$

Hypothesis Test for Slope:

$$t = \frac{b_1}{SE_{b_1}} \quad \text{with } df = n - 2$$

Common Mistakes to Avoid

1. **Correlation vs. Causation:** High correlation doesn't imply causation
2. **Extrapolation:** Don't predict outside the range of observed x-values
3. **Ignoring Conditions:** Always check LINE conditions before using regression
4. **Overinterpreting R^2:** High R^2 doesn't guarantee a good model
5. **Wrong Units:** Pay attention to units in slope interpretation

Regression Analysis Checklist

Before Analysis:

- Create scatterplot to visualize relationship
- Check for outliers and influential points
- Verify conditions (LINE)

During Analysis:

- Calculate correlation coefficient
- Find regression equation
- Interpret slope and y-intercept in context
- Calculate R^2 and interpret

After Analysis:

- Check residual plots for model adequacy
- Test significance of regression slope
- Make predictions within appropriate range
- State conclusions in context