BUS 310 – Lessons 10 & 11 Notes

Correlation and Simple Linear Regression

1. Correlation

Definition

Correlation measures the **degree and direction of linear association** between two **quantitative variables**.

Correlation Coefficient (r)

Key Properties

- Unit-free measure
- Range:

  $-1 \leq r \leq 1$

  Interpretation:

  - $r = 1$: Perfect positive linear relationship
  - $r = -1$: Perfect negative linear relationship
  - $r = 0$: No linear relationship
  - Closer to ± 1 → stronger relationship
  - Closer to 0 → weaker relationship

Correlation Coefficient Formula

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{(\sum(X - \bar{X})^2)(\sum(Y - \bar{Y})^2)}}$$

Scatterplots

- Graphical display of two variables (X and Y)
- Used to:
    - Detect linear relationships
    - Identify positive, negative, or no relationship
- A **shapeless swarm of points** indicates **no relationship**

2. Simple Linear Regression

Purpose of Regression

Regression analysis is used to:

- **Predict** the value of a dependent (response) variable
- **Explain** the effect of an independent (explanatory) variable on the dependent variable

Regression Model

Population Regression Model

$Y = \beta_0 + \beta_1 X + \varepsilon$

Sample Regression Equation

$$\boldsymbol{\hat{Y} = b_0 + b_1 X}$$

Where:

- $\boldsymbol{\hat{Y}}$: predicted value of Y
- $\boldsymbol{b_0}$: sample intercept
- $\boldsymbol{b_1}$: sample slope
- **X** : independent variable

Interpretation

- **Intercept (b_0)**: Average value of Y when X=0
- **Slope (b_1)**: Average change in Y for a one-unit increase in X

Least Squares Method

- The regression line is chosen to **minimize the sum of squared residuals**:

$$\sum e^2 = \sum \left(Y - \hat{Y}\right)^2$$

- Residual:

$$e = Y - \hat{Y}$$

3. Prediction

Using the Regression Equation

$$\boldsymbol{\hat{Y} = b_0 + b_1 X}$$

Predictions should only be made **within the observed range of X values**.

4. Measures of Fit and Precision

Coefficient of Determination

$$r^2$$

- Measures the **proportion of variation in Y explained by X**
- Example:
  r^2 = 0.942⇒ 94.2% of variation in Y is explained by X

Coefficient of Alienation

$1 - r^2$

- Measures the **unexplained variation**

Standard Error of Estimate

$$s_{yx}$$

- Measures the average prediction error around the regression line
- Smaller values indicate better predictive accuracy

5. Inference About the Slope

Hypotheses

H0: β1=0 (no linear relationship)

H1: β1≠0 (linear relationship exists)

Test Statistic for Slope

$$\boldsymbol{t = \frac{b_1}{s_{b_1}}}$$

Where:

- s_{b_1} = standard error of the slope
- Degrees of freedom:

df = n−2

Decision Rule

- Reject H0 if:
  - $|t| > t_{\alpha/2, n-2}$

    or

- p-value ≤ α

6. Confidence Interval for the Slope

$$b_1 \pm t_{\alpha/2, n-2} \cdot s_{b_1}$$

Interpretation:
We are (1−α)×100% confident that the true slope β1 lies within this interval.

7. Key Takeaways for Exams

- Correlation measures **association**, not causation
- Regression provides:
  - Prediction
  - Explanation of relationships
- r shows **strength and direction**
- r^2 shows **explanatory power**
- Slope inference uses a **t-test with df = n−2**