Applied Statistics for Data Scientists with R

Class 18: Correlation and Regression

Types of Relationships

1. Categorical vs. Categorical
2. Categorical vs. Numeric
3. Numeric vs. Numeric

Categorical vs. Numeric

Common Techniques:

- **T-test / ANOVA:** Compares means across categories.

Visualizations: Boxplot, histogram.

Example:

Do different job roles have different average salaries?

Smoking Status vs. Average Lung Capacity

Study Mode (Online/In-Person) vs. Final Exam Score

Numeric vs. Numeric

Common techniques:

- Correlation (Pearson, Spearman, Kendall).
- Simple Linear Regression: Predicts Y from X.

Visualizations: Scatterplot.

Example:

Is higher BMI associated with increased blood pressure?

Do students who attend more classes perform better?

Categorical vs Categorical

Common Techniques:

- **Chi-Square Test of Independence:** Checks if two categorical variables are associated.
- **Cramer's V:** Measures the strength of association between categorical variables.

Visualization: Percent stacked bar plots.

Example:

Does gender influence product preference?

Blood Type vs. Disease Susceptibility

Smoking Status (Smoker/Non-Smoker) vs. Disease Type (Cancer, Diabetes, etc.)

Correlation Analysis

- It measure the **strength** and **direction** of the relationship between two numerical variables.
  - Strength: How strong the relationship is (e.g., weak, moderate, strong).
  - Direction: Positive (both increase together) or Negative (one increases while the other decreases).

- Types of Correlation:
  - Pearson's Correlation (r) – Measures linear relationships.
  - Spearman's Rank Correlation (ρ) – Measures monotonic relationships (not necessarily linear).
  - Kendall's Tau (τ) – Measures ordinal relationships (for ranking data).

Pearson’s Correlation

For valid results, these assumptions should be met

1. Linearity: The relationship between the two variables should be linear (check using scatterplot).
2. Normality: Both variables should be approximately normally distributed (not necessary for Spearman/Kendall).
3. No Outliers: Extreme values can distort the correlation coefficient.

Pearson's Correlation: Interpretation

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

Correlation Coefficient (r)	Interpretation
$r = 1$	Perfect positive correlation
$0.7 \leq r < 1$	Strong positive correlation
$0.5 \leq r < 0.7$	Moderate positive correlation
$0.3 \leq r < 0.5$	Weak positive correlation
$r = 0$	No correlation
$-0.3 \leq r < 0$	Weak negative correlation
$-0.5 \leq r < -0.3$	Moderate negative correlation
$-0.7 \leq r < -0.5$	Strong negative correlation
$r = -1$	Perfect negative correlation

Regression Analysis

- To model the relationship between a dependent variable (target) and one or more independent variables (predictors).

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Regression: Example

$$\text{House Price} = \beta_0 + \beta_1 \times \text{Size (sq. ft.)} + \beta_2 \times \text{Bedrooms} + \beta_3 \times \text{Location Score} + \epsilon$$

$$\text{House Price} = 50,000 + 250 \times \text{Size} + 15,000 \times \text{Bedrooms} + 20,000 \times \text{Location Score}$$

- Intercept ($\beta_0 = 50{,}000$): When Size = 0, Bedrooms = 0, and Location Score = 0, the baseline house price is $50,000
- Size ($\beta_1 = 250$) → For every additional 1 sq. ft., the house price increases **on average** by $250.
- Bedrooms ($\beta_2 = 15{,}000$) → Each additional bedroom increases house price by $15,000.
- Location Score (β_3) → Every 1-point increase in neighborhood quality score increases the price by $20,000.

Another example

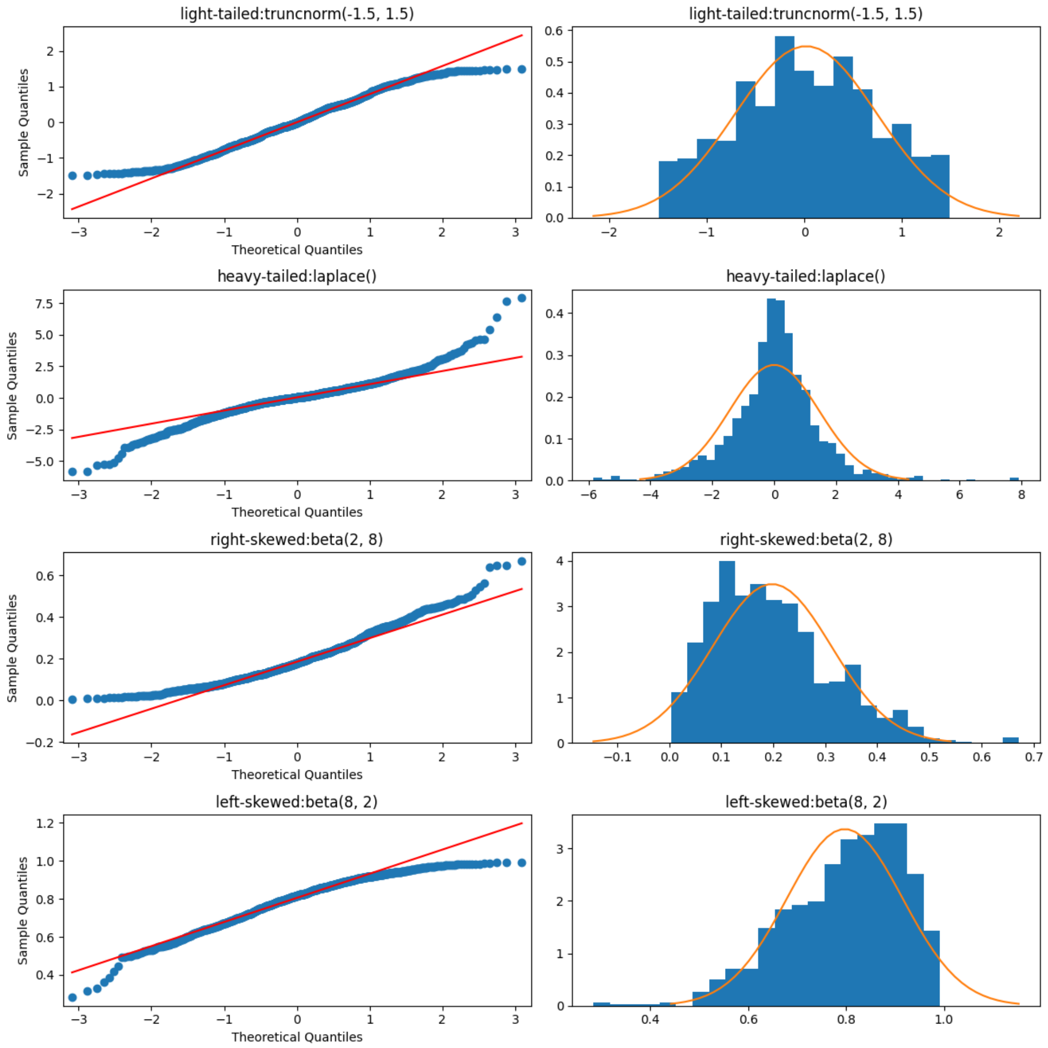
$$\text{House Price} = \beta_0 + \beta_1 \times \text{Size} + \beta_2 \times \text{Bedrooms} + \beta_3 \times \text{Los Angeles} + \beta_4 \times \text{Chicago} + \epsilon$$

- Intercept (β_0) = The base price of a house in New York (reference category).
- Los Angeles (β_3) = Adjusts the price if the house is in Los Angeles (compared to New York).
- Chicago (β_4) = Adjusts the price if the house is in Chicago (compared to New York).

Assumptions

1. The relationship between the independent variable(s) and the dependent variable is linear
   Check: Residual vs Fitted plot should not show any pattern
2. Independent variables should not be highly correlated with each other. (No multicollinearity)
   Check: Variance inflating factors should be ideally less than 5, or at least less than 8
3. Constant variance of errors (Homoscedasticity)
   Check: Residual vs Fitted plot, bp test
4. Residuals should be normally distributed
   Check: QQ plot, histogram, Shapiro wilk test of residuals
5. Residuals should not be correlated
   Check: Dw test

QQ Plot meaning



https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot