COMP517 – Data Analysis

Lab 9 – Introduction to One-way ANOVA and Simple Linear Regression

Introduction

In this lab, we introduce the fundamental concepts of hypothesis testing using one-way analysis of variance (ANOVA) and simple linear regression models. We will use ANOVA statistical techniques to analyze exam scores across various streams of students. Then we perform Tukey's Honestly Significant Difference (HSD) post hoc test after conducting an analysis of variance (ANOVA) to compare test scores among different streams. In the latter part of the lab, we will employ the identical dataset for conducting a linear regression analysis. Simple linear regression serves various purposes, but in this lab, we concentrate on specific objectives. It enables the prediction of the dependent variable by leveraging the independent variable, facilitating the estimation, or forecasting of outcomes for new or unobserved data points. In addition, it provides a valuable means of understanding relationships between variables. The slope coefficient (β1) within the regression model quantifies how variations in the independent variable influence changes in the dependent variable while holding all other factors constant, thus offering insights into the cause-and-effect dynamics between the variables of interest.

A **preliminary** version of this lab is made available under the name "Lab 09_Preload.ipynb". This Python file comes with explanatory comments and some existing codes designed to guide you through the tasks. Download the file and save it your lab 09 directory.

Lab Objectives

- ✓ **One-Way ANOVA:** One-way analysis of variance (ANOVA) is a statistical method to compare the means of three or more groups. We will extend our analysis to multiple streams and assess whether there is a significant variance in exam scores across these streams.
- ✓ **Tukey's Honestly Significant Difference (HSD) Test:** Conduct Tukey's HSD post hoc test following an analysis of variance (ANOVA) to assess and compare test scores across various stream categories.
- ✓ **Simple Linear Regression:** One of the objectives of simple linear regression is to establish and understand the relationship between two variables: one is the dependent variable (the variable you want to predict or explain), and the other is the independent variable (the variable used to make predictions or provide explanations). This analysis then aims to predict a student's test score based on their performance in the lab test. Following that, we will provide a concise discussion on the findings of the model.

Data Source

Our dataset comprises exam scores from students in different streams of a given course. By applying these hypothesis testing techniques, we aim to uncover potential disparities in academic performance among these streams. Download the 'student_streamscores.csv' and save it under your lab 09 directory.

Assumptions

For this lab, we assume that our data follows a normal distribution, and we also assume that the key assumptions for linear regression validity are met. We will cover these assumptions next week.

Grading of the Lab

To receive credit for this lab, you must complete all assigned tasks and answer all questions. Submit your responses to the questions as a PDF file via Canvas.

1. One-way Analysis of Variance (ANOVA)

ANOVA is a statistical technique used to analyze and compare the means of two or more groups or populations to determine if there are statistically significant differences among them. It is a parametric test that extends the principles of the t-test to situations involving more than two groups.

1. **Null Hypothesis (H_0)**: The null hypothesis in ANOVA states that there are no significant differences among the group means. In other words, all group means are equal.
2. **Alternative Hypothesis (H_a)**: The alternative hypothesis suggests that at least one group mean is different from the others. It implies that there is a significant difference among the groups.
3. **Test Statistic (F-Statistic)**: The test statistic in ANOVA is the F-statistic, which is a ratio of two variances: the variance between groups and the variance within groups. It measures how much the group means differ relative to the variability within each group.
4. ***P*-Value**: The *p*-value associated with the F-statistic is used to make a decision about the null hypothesis. If the p-value is less than a chosen significance level (often denoted as alpha, typically 0.05), the null hypothesis is rejected in favour of the alternative hypothesis.

Example: Let's conduct a one-way ANOVA analysis where we have multiple streams (more than two), and we want to determine if there is a statistically significant difference in the mean test scores among these streams.

Research Question: Is there a statistically significant difference in the mean test scores among different streams?

Step 1.1: Formulate Hypotheses

Question 1.1: *formulate the hypotheses for the above research question.*

Step 1.2: Collect Data: Read data from the CSV file, load it into a DataFrame, extract the unique values from the "Stream" column, and print these unique stream values to the console. It helps you understand the different categories or groups within the "Stream" column of your dataset.

Step 1.3: Create and visualise histogram plots with Kernel Density Estimation (KDE): This step is essential in the context of a one-way analysis of variance (ANOVA) because it serves multiple crucial purposes. Firstly, it allows for the examination of the distribution of test scores within each group, helping to assess whether the assumption of normality is met. Additionally, it aids in identifying potential outliers and understanding variations in score distributions among different groups, all of which are fundamental considerations in ANOVA. Moreover, these visualizations facilitate data exploration and can reveal any violations of ANOVA assumptions, guiding decisions on appropriate statistical tests or data transformations. Lastly, they provide an intuitive means of communicating the findings to both technical and non-technical audiences, enhancing the interpretability of the results.

Question 1.2: *Analyse your findings and provide a brief and definitive summary regarding the distributional attributes of both the overall test scores and the test scores for each individual stream.*

Step 1.4: Perform a One-way ANOVA test: using **stats.f_oneway** to analyze whether there are significant differences in the means of test scores among the different streams. To do so we took below steps:

- Grouping Data:

  This code creates a list called grouped_data that contains the test scores for each stream. It uses a list comprehension to iterate through each unique stream value in stream_values and filters the data DataFrame to select only the test scores corresponding to that stream.

- Performing ANOVA:

  The stats.f_oneway function from a library like SciPy is used to perform the one-way ANOVA. It takes the individual test score arrays from grouped_data as arguments and calculates the F-statistic and associated p-value.

Step 1.5: Provide the results to make a decision based on significance level (alpha=0.05): The code in this step prints out the results of the one-way ANOVA. It displays the F-statistic, which measures the ratio of variation between group means to variation within groups, and the p-value, which indicates the significance level of the ANOVA test. We set the significance level (alpha) to 0.05.

Compare the *p*-value to the chosen significance level (α).

Question 1.3: *Fill in the table below with the results of your code.*

Table 1.1: Results of one-way ANOVA

Dataset statistical output	
	Sample Size
Stream 50	
Stream 51	
Stream 52	
Stream 53	
one-way ANOVA Statical outputs	
f-statistic	
p-Value	

Step 1.6: Draw a Conclusion:

Question 1.4: *Based on your results, draw a conclusion about whether the exam scores of different streams are significantly different or not.*

Step 1.7: Use of Critical F-value to interpret the results and then draw a conclusion.

In this step, we will be performing a statistical hypothesis test using an F-test to determine if there is a significant difference in test scores among different streams. The code provided in this section is part of a hypothesis test commonly used in ANOVA to compare the means of multiple groups (streams in this case). The F-test helps determine if the variance between group means is statistically significant or if it can be explained by random chance.

1. Calculate the degrees of freedom for the numerator/between (dfn) and denominator/within (dfd) in the F-distribution where **k** is the number of groups or streams (in this case, 4 streams) and **N** is the total number of observations, which is the sum of the lengths of scores from each stream.
    - **df_between** is the degrees of freedom between groups. It represents the degrees of freedom for the numerator and is calculated as **k – 1** where K is number of unique streams.
        - This value is associated with the variability between group means. It quantifies the **number of group means** that are free to vary independently when comparing them in the ANOVA.
    - **df_within** is the degrees of freedom within groups. It represents the degrees of freedom for the denominator and is calculated as **N – k** where K is number of unique streams and N is the total number of dataset rows.
        - This value is associated with the variability within each group. It quantifies the **number of data points** within the sample that are free to vary independently when calculating the within-group variability.
    - "free to vary independently" implies that a parameter or value is not constrained or restricted by other factors or parameters in a way that limits its range of possible values. It can vary or change on its own without being overly influenced by external factors, and this independence is a fundamental concept in statistical analyses and hypothesis testing.
2. Calculate the critical F-value for a given significance level (alpha) and degree of freedom:
    - **stats.f.ppf(1 - alpha, df_between, df_within)** uses the **ppf** (percent point function) method from **scipy.stats** to calculate the critical F-value. The critical F-value represents the value from the F-distribution that corresponds to a specific probability (1 - alpha) and the degrees of freedom for the numerator (**df_between**) and denominator (**df_within**).
3. Print the critical F-value rounded to two decimal places.
4. Test the hypothesis: **f_statistic** represents the computed F-statistic from your data, which is used to test the null hypothesis.
    - If **f_statistic** is greater than the critical F-value, it means that the test statistic is in the tail of the F-distribution, and you reject the null hypothesis.
    - If **f_statistic** is less than or equal to the critical F-value, you fail to reject the null hypothesis.

The null hypothesis in this context likely states that there is no significant difference in test scores among the different streams. If you reject the null hypothesis, it suggests that there is a significant difference in test scores among at least some of the streams.

Question 1.5: *What was the calculated critical f-value? Based on your results and the provided general rule, draw a conclusion about whether the exam scores of these streams are significantly different.*

Step 1.8: Let's identify which specific stream pairs had significant differences in means after conducting an ANOVA. To do so, we can use Tukey's Honestly Significant Difference (HSD) post hoc test to determine where those differences lie when you have three or more groups. This is particularly useful when ANOVA indicates that there are significant overall differences among groups but doesn't specify which groups are different from each other. Below explains the steps we took to perform the test:

- First, we create a **MultiComparison** object called **multicomp**. This object is used to perform multiple comparison tests, such as Tukey's HSD, to assess the differences between groups after an ANOVA.
    - data['Test_Score'] represents the dependent variable (test scores), and
    - data['Stream'] represents the independent variable (the stream or group labels) that the Tukey's HSD test will compare.
- The multicomp.tukeyhsd() method is called on the multicomp object to compute Tukey's HSD test results. This test compares the means of all pairs of groups (streams) and determines if there are statistically significant differences between them.
- Displaying the Results: This output typically includes a table that provides information about the mean differences between groups, confidence intervals, and whether those differences are statistically significant.
    - **"meandiff"** indicates the mean difference between the two groups being compared.
    - **"p-adj"** represents the adjusted p-value, which is corrected for multiple comparisons.
    - **"lower"** and **"upper"** refer to the lower and upper bounds of the confidence interval for the mean difference.
    - **"reject"** is a binary indicator that tells us whether to reject the null hypothesis of no significant difference between the groups.

Example on how to interpret the results: Comparing "st50" and "st51" groups, the mean difference is 0.86, but the *p*-value (*p*-adj) is 0.9, indicating that there's no statistically significant difference between these two groups (reject = False).

Question 1.6: *Provide the results generated by Tukey's Honestly Significant Difference (HSD) test and interpret the result using above example.*

2. Simple Linear Regression

Simple linear regression is a statistical technique used to model and quantify the linear relationship between two variables: one is considered the dependent variable, which is the outcome of interest, and the other is the independent variable, which is used to predict or explain changes in the dependent variable. It seeks to fit a straight line to the data that best represents the association between the two variables, typically by estimating two coefficients: the slope (indicating the direction and steepness of the relationship) and the intercept (representing the value of the dependent variable when the independent variable is zero). Simple linear regression is widely used for prediction, understanding cause-and-effect relationships, and assessing the strength and significance of the linear association between the variables.

Research Question:

Is there a statistically significant relationship between students' performance in their lab tasks (Lab_Grade) and their final test scores (Test_Score)? Specifically, do students who excel in their lab tasks tend to achieve higher scores on their final tests?

Objective:

The objective is to investigate and quantify the relationship between students' performance in their lab tasks (independent variable) and their test scores (dependent variable). By performing a simple linear regression analysis, the goal is to determine whether there exists a statistically significant linear association between these two variables. Additionally, the objective is to assess the direction and strength of this relationship, allowing us to understand if students who perform well in their labs tend to score higher on their final tests or vice versa.

The objective of performing a simple linear regression analysis, in the context of above research question, is to:

a) **Model the Relationship:** Build a statistical model that represents the linear relationship between the independent variable (Lab_Grade) and the dependent variable (Test_Score).
b) **Quantify the Relationship:** Determine the strength and direction of the relationship by estimating the regression coefficients (slope and intercept).
c) **Assess Significance:** Evaluate whether the observed relationship is statistically significant, which helps us determine if the relationship is unlikely to have occurred by random chance.
d) **Predict Outcomes:** Utilize the regression model to predict or explain how changes in the Lab_Grade are associated with changes in the Test_Score.
e) **Provide Insights:** Interpret the findings to answer the research question and draw conclusions about the extent to which students' lab performance relates to their final test scores.

Step 2.1: Define the independent variable X (grade score) and the dependent variable y (test score) based on the data in the 'Lab_Grade' and 'Test_Score' columns of the dataset, respectively.

Step 2.2: Add a constant term (intercept) to the independent variable X. This is necessary for estimating the intercept of the linear regression model.

Step 2.3: Create a linear regression model using the Ordinary Least Squares (OLS) method from StatsModels. It fits the model to the data, estimating the coefficients (slope and intercept) that best describe the linear relationship between the independent and dependent variables.

Step 2.4: Print a summary of the linear regression model. The summary includes several important statistics and information. The regression model summary helps interpret the relationship between the grade score and test score, including the strength and significance of the association and other diagnostic statistics for assessing the goodness of fit of the model. We will go through the detail of each output in future lab. See Appendix A for more detail.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Test_Score   R-squared:                       0.725
Model:                            OLS   Adj. R-squared:                  0.724
Method:                 Least Squares   F-statistic:                     522.6
Date:                Tue, 26 Sep 2023   Prob (F-statistic):           1.92e-57
Time:                        20:32:23   Log-Likelihood:                -655.88
No. Observations:                 200   AIC:                             1316.
Df Residuals:                     198   BIC:                             1323.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.3232      2.749      4.119      0.000       5.902      16.745
Lab_Grade      0.8102      0.035     22.860      0.000       0.740       0.880
==============================================================================
Omnibus:                       11.787   Durbin-Watson:                   0.817
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               14.154
Skew:                           0.448   Prob(JB):                     0.000844
Kurtosis:                       3.946   Cond. No.                         467.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

❖ Interpretation of few metrics provided in the above model summary:

a) **F-statistic:** The F-statistic (522.6) is a measure of the overall significance of the model. A high F-statistic suggests that at least one independent variable significantly contributes to explaining the dependent variable. In this case, the F-statistic is high, indicating the model's overall significance.

b) **Prob (F-statistic):** The *p*-value associated with the F-statistic (1.92e-57) is extremely low, effectively zero. This indicates strong evidence against the null hypothesis that all coefficients are zero. In other words, the model as a whole is highly significant.

c) **Coefficients:** The coefficients provide information about the intercept (const) and the coefficient of the independent variable (Lab_Grade).

i. The intercept (const) is estimated to be 11.3232, suggesting that when Lab_Grade is zero, the predicted Test_Score is approximately 11.32.

d) ***P*-values for Coefficients:** The *p*-values associated with the coefficients are both very low (close to zero). This indicates that both the intercept and Lab_Grade coefficient are statistically significant. In this context, Lab_Grade significantly contributes to predicting Test_Score.

Question 2.1: *Provide the coefficient associated with Lab_Grade and provide an explanation of its meaning and significance.*

Appendix A: Simple Linear Regression Model Summary

- Coefficients: It provides information about the estimated coefficients (slope and intercept).
- R-squared: The coefficient of determination, which measures the proportion of variance in the dependent variable explained by the independent variable.
- *p*-values: Indicates the significance of the coefficients. Low *p*-values suggest that the coefficients are statistically significant.
- Confidence intervals: Provides a range within which the true population coefficients are likely to fall.
- Standard errors: Estimates of the variability of the coefficients.
- F-statistic and *p*-value: Assess the overall significance of the model.
- Residuals: Information about the residuals (differences between observed and predicted values).
- Durbin-Watson statistic: Assesses the presence of autocorrelation in the residuals.