How to Calculate the Correlation Between Continuous Input Features and Ordinal Output

When working with data, one of the primary goals is to understand the relationships between input features and the output variable. But what if your input features are **continuous** (e.g., age, salary, temperature) and your output variable is **ordinal** (e.g., satisfaction level: "low," "medium," "high")? This scenario presents a unique challenge because you're dealing with variables that have different types: a continuous input and an ordinal output.

In this blog post, we'll explore the methods for calculating the correlation between a continuous input feature and an ordinal output variable. We'll walk through several statistical techniques, provide practical examples, and discuss how to interpret the results.

What Are Continuous and Ordinal Variables?

Before we jump into the methods, let's define these types of variables:

- **Continuous input**: A continuous variable can take any value within a range and is typically measured on a scale with consistent intervals. Examples include height, weight, or income. These values can be any real number, with no inherent breaks between the values.

- **Ordinal output**: An ordinal variable consists of categories that have a meaningful order or ranking, but the intervals between the categories are not uniform. For example, a satisfaction rating might be categorized as "low," "medium," or "high," but the difference in satisfaction between "low" and "medium" is not necessarily the same as the difference between "medium" and "high."

The challenge in calculating correlation between these types of variables lies in properly measuring how the continuous input influences the ordinal output. Unlike traditional Pearson correlation, which assumes a linear relationship between continuous variables, ordinal variables require a different approach.

Let's now dive into the methods you can use to assess the relationship between continuous input features and ordinal output variables.

---

1. Spearman's Rank Correlation

Spearman's rank correlation is one of the most popular methods for assessing the relationship between continuous and ordinal variables. This method is particularly useful when you have a continuous input and an ordinal output, as it doesn't assume a linear relationship but instead measures how well the ranks of the input correspond to the ranks of the output.

How It Works:

1. **Rank the continuous input**: If your continuous input variable is not already in rank form, you'll first rank the values. For example, for a set of salaries, you would assign the lowest salary a rank of 1, the second-lowest rank 2, and so on.

2. **Rank the ordinal output**: The ordinal output already has a natural ranking. For example, if your output is a satisfaction level ("low," "medium," "high"), you would assign ranks like this: "low" = 1, "medium" = 2, and "high" = 3.

3. **Calculate the Spearman rank correlation**: After ranking the continuous and ordinal variables, you can calculate the Spearman rank correlation coefficient, denoted as **ρ**. The formula for **ρ** is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where:

- did_i is the difference in ranks for each pair of data points,
- nn is the number of data points.

Why Use It:

Spearman's rank correlation is ideal when you're dealing with non-linear relationships between continuous and ordinal variables. It simply looks for a monotonic relationship — that is, a consistent increase or decrease between the ranks of the input and output variables.

Example:

Let's say you're analyzing the relationship between **income** (continuous) and **customer satisfaction** (ordinal: "low," "medium," "high"). After ranking both income and satisfaction, Spearman's rank correlation will give you a coefficient between -1 and +1:

- **+1** indicates a perfect positive monotonic relationship (as income increases, satisfaction increases),
- **-1** indicates a perfect negative monotonic relationship (as income increases, satisfaction decreases),
- **0** indicates no relationship.

---

2. Ordinal Logistic Regression

If you're interested in predicting the **ordinal output** based on the **continuous input**, **ordinal logistic regression** is a powerful technique to consider. Unlike Spearman's rank, which simply quantifies the strength of the relationship, ordinal logistic regression models the probability that the ordinal output will fall into a particular category based on the continuous input.

How It Works:

1. **Model the ordinal output**: Ordinal logistic regression treats the ordinal output as a categorical variable with ordered levels (for example, "low," "medium," and "high").

2. **Continuous input as predictor**: The continuous variable is used as the predictor in the model, and the goal is to predict the probability that the output falls into each category of the ordinal output (low, medium, or high).

3. **Calculate odds ratios**: The model calculates **odds ratios** that describe how the likelihood of the ordinal output changing from one level to another is influenced by changes in the continuous input.

Why Use It:

Ordinal logistic regression is particularly useful if you want to **predict** the ordinal outcome based on continuous input and assess how the input influences the probability of different output categories.

Example:

If you want to predict the **customer satisfaction level** (ordinal: low, medium, high) based on **age** (continuous), ordinal logistic regression can model how the likelihood of each satisfaction level changes as age increases. This gives you not just a correlation but a probabilistic framework for understanding how age impacts satisfaction.

---

3. Kendall's Tau-b

Kendall's Tau-b is another rank-based correlation method that is similar to Spearman's rank correlation but has certain advantages when dealing with ties in the data. It's particularly useful when there are many tied ranks in the continuous variable, making it a robust choice for smaller datasets or datasets with a lot of equal values.

How It Works:

- **Rank the variables**: As with Spearman's correlation, you rank both the continuous input and the ordinal output.

- **Calculate the Tau-b coefficient**: Kendall's Tau-b measures the strength of the monotonic relationship between the variables by comparing the number of concordant and discordant pairs. A concordant pair is where both variables move in the same direction, while a discordant pair moves in opposite directions.

Why Use It:

Kendall's Tau-b is a more robust choice than Spearman's correlation when there are ties in the data, making it especially useful for datasets with many repeated values.

Example:

If you are analyzing the relationship between **employee age** (continuous) and their **job satisfaction** level (ordinal: low, medium, high), Kendall's Tau-b can be used to calculate how strongly these two variables are related while accounting for tied ranks.

---

4. Mann-Whitney U Test (for Two Categories)

If your ordinal output has only two categories, you could also use the **Mann-Whitney U test** (also known as the **Wilcoxon rank-sum test**). This test compares the distributions of the continuous input across the two categories of the ordinal output to assess whether there is a significant difference between the groups.

How It Works:

- **Group the data** based on the two categories of the ordinal output (e.g., "low" vs. "high").
- **Rank the continuous input** across all groups.
- **Compare the ranks** between the two groups using the U statistic.

Why Use It:

The Mann-Whitney U test is particularly useful when you have a **binary ordinal output** (e.g., "low" vs. "high" satisfaction) and want to compare the distributions of a continuous input across the two groups.

Example:

If you are analyzing whether **salary** (continuous) differs significantly between employees who report "low" satisfaction and those who report "high" satisfaction, the Mann-Whitney U test will assess whether the salary distributions for these two satisfaction categories are different.

Summary of Methods

Method	Best Used For	Strength
Spearman's Rank	Measuring monotonic relationships between continuous input and ordinal output	Simple and effective for monotonic relationships, works well with rankings
Ordinal Logistic Regression	Predicting ordinal outcomes based on continuous input	Provides probabilistic predictions, useful for modeling relationships
Kendall's Tau-b	Handling ties and small datasets when assessing monotonic relationships	Robust to ties, especially in small datasets
Mann-Whitney U Test	Comparing continuous input between two ordinal categories	Great for comparing distributions with binary ordinal outcomes

---

Conclusion

When dealing with a **continuous input** and an **ordinal output**, several methods can help you calculate the correlation and assess the relationship between the two variables. The best method for your analysis will depend on the nature of your data and your specific goals:

- **Spearman's Rank Correlation** is great for measuring monotonic relationships.
- **Ordinal Logistic Regression** is ideal for predictive modeling.
- **Kendall's Tau-b** is useful when handling ties in the data.
- **Mann-Whitney U Test** is perfect for comparing two groups when your ordinal output has only two categories.

By using these techniques, you can uncover the relationship between continuous input features and ordinal output, providing you with the insights necessary to make informed decisions in your analysis.