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

In the world of data science, the relationship between input features and output variables is the foundation of building predictive models and drawing meaningful insights. However, when the input features are **ordinal** (i.e., variables with a meaningful order but no consistent difference between categories), and the output is **continuous** (i.e., a variable that can take any value within a range, such as age, price, or temperature), the correlation methods become a bit more specialized.

This blog will walk you through how to calculate the correlation between ordinal input features and continuous output variables, focusing on the best techniques and the key things to consider when analyzing this type of data.

Understanding Ordinal Input Features and Continuous Output

Before diving into the calculation methods, let's first define the two types of variables involved:

- **Ordinal Input**: An ordinal variable consists of categories that have a meaningful order or ranking, but the intervals between the categories are not necessarily consistent. For example, consider a scale of "poor," "average," and "excellent" for customer satisfaction. These categories have a logical order, but we can't say that the difference between "poor" and "average" is the same as the difference between "average" and "excellent."

- **Continuous Output**: A continuous variable can take any value within a certain range. Examples include variables like height, temperature, weight, or price. These variables have an infinite number of possible values and can be measured at fine intervals.

When dealing with ordinal inputs and continuous outputs, the goal is to determine if there's a meaningful relationship between the ordered categories of the input variable and the numerical values of the output variable. Here are the most commonly used methods to calculate such a correlation:

1. Spearman's Rank Correlation

Spearman’s rank correlation is one of the most commonly used methods for assessing the relationship between an ordinal input and a continuous output. Unlike Pearson’s correlation, which assumes a linear relationship between the variables, Spearman’s rank correlation works

by ranking the values of the ordinal input and then calculating the correlation between these ranks and the continuous output.

How It Works:

- **Step 1**: Convert the ordinal input categories into numerical ranks. For example, if you have the categories "low," "medium," and "high," you would assign them ranks like 1, 2, and 3, respectively.

- **Step 2**: Rank the continuous output values (if necessary). This step involves converting the continuous values into ranks based on their magnitude. For example, if the continuous output is income, the lowest value would get the rank of 1, the second-lowest would get rank 2, and so on.

- **Step 3**: Compute the correlation between the ranks of the ordinal input and the ranks of the continuous output. Spearman's rank correlation gives a coefficient (ranging from -1 to +1), where:

    - **+1** means a perfect positive monotonic relationship,
    - **-1** means a perfect negative monotonic relationship,
    - **0** means no monotonic relationship.

Why Use It:

Spearman's rank correlation is particularly useful because it doesn't assume the relationship between the variables is linear. It simply looks for a consistent increase or decrease in the variables' ranks, making it ideal for situations where the data is skewed or where there are outliers.

Example:

Let's say you are analyzing the relationship between customer satisfaction (ordinal: "low," "medium," "high") and product price (continuous). You would rank the customer satisfaction categories and then rank the prices of the products. Spearman's rank correlation would then quantify how well the ranks of satisfaction are associated with the ranks of price.

---

2. Kendall's Tau

Kendall's Tau is another rank-based correlation measure, similar to Spearman's rank correlation, but it has some advantages when dealing with smaller datasets or when there are many tied ranks (i.e., multiple values that are equal).

How It Works:

- **Step 1**: Convert the ordinal input to ranks.
- **Step 2**: Convert the continuous output to ranks.
- **Step 3**: For each pair of data points, determine if the ranks are in the same direction (concordant), opposite directions (discordant), or tied.

Kendall's Tau is calculated based on the number of concordant and discordant pairs, and it also ranges from -1 to +1. A positive value indicates that the ranks of the ordinal input and continuous output tend to increase together, while a negative value suggests they move in opposite directions.

Why Use It:

Kendall's Tau is preferred when you have small datasets or a lot of ties in the ranks, as it tends to be more robust than Spearman's correlation in these situations.

Example:

If you're studying the relationship between employee satisfaction (ordinal: "low," "medium," "high") and salary (continuous), Kendall's Tau would provide a measure of the strength of association between satisfaction levels and salary, adjusting for ties or small sample sizes.

---

3. Ordinal Logistic Regression

Ordinal logistic regression (also called **proportional odds regression**) is a more advanced method that models the relationship between an ordinal input and a continuous output, accounting for the fact that the ordinal variable has an inherent order.

How It Works:

- Ordinal logistic regression treats the ordinal input variable as an ordered factor and models the continuous output variable based on the ordered levels of the input.
- This method calculates the probability that the continuous output variable falls into a certain range, given the ordered levels of the ordinal input.
- The result of ordinal logistic regression is a set of odds ratios, which describe the likelihood of the continuous output being higher or lower as the ordinal input changes.

Why Use It:

Ordinal logistic regression is particularly useful if you want to predict or model a continuous outcome based on an ordinal input. It's especially helpful when the relationship between the ordinal input and the continuous output is more complex than a simple monotonic relationship.

Example:

If you're trying to predict the price of a product (continuous) based on its quality rating (ordinal: "low," "medium," "high"), ordinal logistic regression can help you understand how each level of quality influences the price, while accounting for the ordered nature of the quality variable.

---

4. ANOVA (Analysis of Variance)

If you have an ordinal input with relatively few categories (e.g., "low," "medium," "high") and a continuous output, you might consider using **ANOVA** to assess whether the means of the continuous output differ significantly across the levels of the ordinal input.

How It Works:

- **Step 1**: Group the continuous output data based on the categories of the ordinal input.
- **Step 2**: Perform ANOVA to compare the means of the continuous output across the different ordinal input levels.

Why Use It:

ANOVA is useful if you are interested in understanding whether there are significant differences in the continuous output between the levels of the ordinal input. However, ANOVA assumes that the data are normally distributed, so it may not be appropriate for highly skewed data.

Example:

If you are studying how customer satisfaction (ordinal: "low," "medium," "high") affects spending (continuous), ANOVA can help you determine if the mean spending differs between customers in different satisfaction categories.

Conclusion

When you have an **ordinal input** and a **continuous output**, calculating the correlation between them involves considering both the ordinal nature of the input and the continuous nature of the output. The methods outlined in this blog—**Spearman's Rank Correlation**, **Kendall's Tau**, **Ordinal Logistic Regression**, and **ANOVA**—are all powerful techniques to help you understand how these types of variables are related.

- **Spearman's rank** and **Kendall's Tau** are excellent for measuring monotonic relationships.
- **Ordinal logistic regression** is useful for more advanced modeling when you want to predict a continuous outcome based on an ordinal input.
- **ANOVA** is great when you want to test for significant differences in the continuous output across different levels of the ordinal input.

Each method has its strengths and is suited to different types of analyses. Choosing the right one will depend on your data's specific characteristics and the goals of your analysis. By understanding these methods, you can more accurately model and interpret the relationship between ordinal inputs and continuous outputs.