Coursework 4: Statistic for AI and Data Science Part 1

Q1 Answer) The study explores the correlation between stork populations and European human birth rates, yielding a statistically significant p-value of 0.008. A frequent misconception is to equate correlation with causation, leading to the erroneous conclusion that storks deliver babies.

The correct interpretation of the p-value is rooted in its role as a measure of evidence against the null hypothesis, which, in this case, asserts no correlation between storks and human births. Despite the statistical significance, it is crucial to recognise that a low p-value does not imply causation. As aptly put by the author, the association between stork populations and human births is 'clearly ludicrous.'

In this context, the most plausible explanation for the observed correlation is the presence of a confounding variable, such as land area, which may influence both birth rates and the number of breeding pairs of storks. Understanding the distinction between correlation and causation is paramount in statistical analysis to avoid drawing unwarranted conclusions.

Q2 Answer) The correlation coefficient measures the strength and direction of the linear relationship between two variables. Specifically, the correlation coefficient quantifies how well changes in one variable predict changes in the other, with a high absolute value indicating a robust linear relationship. At the same time, a low correlation coefficient suggests a weak or non-linear relationship. A linear relationship between two variables means that a change in one variable is consistently associated with a proportional change in the other, forming a straight line when plotted on a graph. However, the correlation coefficient alone does not provide information about the statistical significance of the relationship.

The p-value, on the other hand, measures the statistical significance of the relationship between the dependent and independent variables (s). A low p-value indicates that the observed relationship is statistically significant, meaning it is unlikely to be due to chance. This suggests that the regression model provides meaningful insights into the data, constituting what we consider a 'good fit.'

Therefore, both the correlation coefficient and the p-value are essential measures of the goodness of fit of a regression model, but they have different interpretations. The correlation coefficient assesses the strength and direction of the linear relationship, while the p-value measures the statistical significance of the relationship. Together, they offer a more complete picture of how well the regression model fits the data, considering the nature of the relationship and its statistical reliability.

Q3 Answer) In terms of correlation, the statistical association between the stork population and the human birth rate is $r = 0.62$. A high correlation coefficient might indicate a strong positive relationship, implying that regions with more storks tend to have higher birth rates. However, correlation alone does not imply causation; it merely points to a consistent pattern or trend.

Causation, in contrast, suggests a cause-and-effect relationship. In the stork example, claiming causation would mean asserting that the presence of storks directly causes an increase in human births.

The possible relationships highlighted in this example include the notion that a statistically significant correlation does not automatically imply causation. The observed correlation between the stork population and the human birth rate might be coincidental or influenced by a confounding variable (an extraneous factor that is not the main focus of a study but can affect the relationship between the variables under investigation), such as land area, leading to a misinterpretation of their association.

This example is a comical yet insightful illustration of the fallacy of inferring causation from correlation. It underscores the importance of considering alternative explanations and caution when interpreting statistical relationships.

Q4 Answer) A confounding variable is an extraneous factor that is not the main focus of a study but can affect the relationship between the variables under investigation. Regarding the relationship between storks and births, a possible confounding variable is land area. Countries with larger land areas may have more storks and higher birth rates, but this does not necessarily mean that storks cause births. Another possible confounding variable is population density. Countries with higher population densities may have fewer storks and lower birth rates, but this does not necessarily mean that storks prevent births.

Other possible confounding variables include socioeconomic status, climate, and cultural factors. For example, countries with higher levels of poverty may have higher birth rates and more storks due to a lack of access to birth control and family planning services. Similarly, countries with colder climates may have fewer storks and lower birth rates due to environmental factors. Cultural factors, such as beliefs about fertility and childbirth, may also affect the relationship between storks and births.

Identifying and controlling for confounding variables in statistical analyses is essential to avoid drawing incorrect conclusions about the relationship between two variables. For storks and births, controlling for confounding variables would be necessary to determine the proper relationship between the two variables.

Diagram

The below diagram shows a causal map between our 4 variables. Area relates to both stock and human population. This is self-explanatory as the geographical area can influence both stork and human populations, creating a direct link between these variables in the causal map. Humans relate to birth rate as they are the contributors to the birth rate, with the number of births directly influenced by the human population. This connection underscores the direct impact of human population size on the overall birth rate within the causal map.