**Coursework 4: Statistic for AI and Data Science**

2 Part 1: Review of the paper ‘Storks Deliver Babies’

**Question 1: The paper explains that the p-value can be misunderstood. Explain the misunderstanding described in the paper and give the correct interpretation of the p-value in the context of the analysis in the paper.**

The paper points out a common confusion regarding p-values, which is the mistaken belief that a small p-value, like the one in the stork-baby study (p = 0.008), directly suggests a high probability of the investigated effect being true. This misinterpretation can lead to incorrect conclusions.

The correct understanding of a p-value is that it indicates the likelihood of observing data as extreme as the data actually observed, under the assumption that there is no association (the null hypothesis is true). In this context, it means there is a 0.8% chance of seeing such a strong correlation between storks and birth rates by random chance when there is no real correlation.

This p-value does not mean there is a 99.2% chance that storks deliver babies. It simply signifies that the observed correlation is statistically unlikely to have occurred if storks and births were truly uncorrelated.

The misunderstanding occurs when the p-value is wrongly taken as direct proof of the alternative hypothesis. It's essential to interpret statistical significance in the context of plausibility and real-world evidence, not as an isolated measure. Despite the low p-value, the implausible nature of storks delivering babies reminds us that statistical significance should not be conflated with actual significance or causation.

**Question 2: Explain how the correlation coefficient and the p-value relate to the question ‘how good is my regression model?’, making clear the difference between them**

In assessing the quality of a regression model, the correlation coefficient and the p-value provide different insights:

1. Correlation Coefficient: Indicates the strength and direction of the linear relationship between two variables. Ranges from -1 to 1, with values close to 1 or -1 signifying a strong linear relationship. Informs about the extent to which the model captures the linear relationship between variables, but does not address causality, influence of other variables, or prediction accuracy.

2. P-value: Tests the null hypothesis that a particular coefficient (like the slope) is zero, meaning no effect of the independent variable on the dependent variable. A low p-value (typically < 0.05) suggests a statistically significant effect of the independent variable on the dependent variable. Indicates the statistical significance of the relationship but doesn’t directly reveal the strength or practical significance of this relationship.

Key Differences:

The correlation coefficient measures the strength of a linear relationship but doesn't imply causation or test significance. The p-value assesses the statistical significance of the relationship but doesn’t comment on its strength or practical importance.

Conclusion:

To gauge "how good" a regression model is, consider both the correlation coefficient and the p-value. However, a thorough evaluation also involves examining model fit (like R-squared), predictive accuracy, and other factors like confounding variables and model assumptions. A high correlation and significant p-value suggest statistical soundness but don’t alone confirm the model's practical utility or appropriateness for causal inference.

**Question 3: Using the example from the paper, explain the difference between causation and correlation, covering possible relationships between them.**

In the paper "Storks Deliver Babies," the difference between correlation and causation is illustrated using the example of storks and birth rates:

Correlation: The paper demonstrates a statistical correlation between stork populations and human birth rates, with a moderate positive correlation coefficient. This means areas with more storks tend to have more births, but this correlation does not imply that storks cause more births.

Causation: Causation implies a direct cause-and-effect relationship. The paper humorously uses the old folklore of storks delivering babies to highlight that, despite the correlation, there is no causal relationship between the number of storks and the number of births. The idea is biologically and logically implausible

Relationship Between Correlation and Causation: The paper exemplifies several key points:

• Correlation Without Causation: Two variables can be correlated due to coincidence, third factors, or other unknown reasons, without one causing the other.

• Causation with Correlation: True causal relationships often result in correlation.

• Causation Without Correlation: Causation can exist without a clear correlation, especially in complex systems.

• Spurious Correlation: Apparent correlations may be due to a third factor influencing both variables, leading to misleading interpretations.

Conclusion: The paper emphasizes that correlation alone does not imply causation. While two variables can be correlated, careful analysis is necessary to establish if one truly causes the other. The stork-baby example serves to underline the importance of critical thinking in interpreting statistical data.

**Question 4: Explain what is meant by a confounding variable and suggest possible confounders for any relationship between storks and births. Draft and describe a diagram of causes that you believe is most likely to explain the relationship between the 4 variables Area, Storks, Humans and BirthRate used in the paper**

A confounding variable is one that influences both the dependent and independent variables in a study, potentially leading to false conclusions about the relationship between these variables. In the context of the stork-birth rate relationship, potential confounders could include:

**Area of the Country:** Larger areas might have more storks due to ample habitat and also larger human populations, thus more births.

**Rural vs. Urban Distribution:** Rural areas may support more storks and also have different birth rates compared to urban areas.

**Economic Factors:** Wealthier regions may support better conditions for both storks and human populations.

**Climate and Environmental Conditions:** Climates favorable to storks could also correlate with factors influencing human birth rates.

Causal Diagram for Area, Storks, Humans, and BirthRate:

Area → Storks: More area means more habitat for storks.

Area → Humans: Larger areas support larger human populations.

Humans → BirthRate: More people generally lead to more births.

Area → BirthRate: Indirect impact through various socio-economic factors.

Storks ↔ BirthRate: Observed correlation, not causation, likely influenced by Area and Humans.

Area(A) Storks(S)

Spurious Correlation

Humans(H) Birth Rate (B)

In this diagram, arrows indicate influence direction. The correlation between storks and birth rate is shown without a direct causal link, suggesting that the observed association is influenced by other factors like area size and human population, rather than a direct causative relationship between storks and births.

# References

Matthews, R. (2000). Storks Deliver Babies (p= 0.008). *Teaching Statistics, 22*(2), 36-38. Retrieved 11 21, 2023, from https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9639.00013

(Matthews, 2000)