ECS7024 Statistics for AI and Data Science - Coursework 4

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

Content

Section 1: Introduction

Section 1.1 Objective and Outline

The objective of this pdf is to answer the following questions about points made in the paper by Robert Matthews.

“Storks Deliver Babies (p = 0.008)”. Teaching Statistics. Volume 22,

Number 2, Summer 2000, p36-8.

Section 2: 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.

Answer

The paper is an exploration into the misconceptions that can occur if the p value is taken as truth for 2 variables where a confounding effect may in fact cause the observed result.

The paper found a p value of 0.008 when comparing number of stork breeding pairs and human birth rate pairs (1000/yr). The P value of 0.08 does not show that the two variables are correlated with any degree of strength. Despite this, many researchers would report it and conclude that the two variables are 'significantly correlated', the paper mentioned more likely a confounding variable is the reason for the correlation. In this case it was land area. The p value can be improperly used to determine the hypothesis is correct.

Section 3: 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.

Answer

When answering the question ‘how good is my regression model?', we rely on two metrics: the correlation coefficient and the p-value. The correlation coefficient, shown with 'r,' tells us about the strength and direction of the linear relationship between variables. A positive 'r' indicates that as one variable increases, the other tends to increase. Meanwhile, the p-value helps us determine the statistical significance of our findings. A low p-value suggests that our model's predictions are statistically meaningful, providing evidence against the null hypothesis. These two metrics together offer a comprehensive judge when answering the question ‘how good is my regression model?', considering both the strength of the observed relationships and their statistical significance, are essential for conclusions in regression modelling.

Section 4: Question 3

ECS7024 Statistics for AI and Data Science - Coursework 4

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

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

Answer

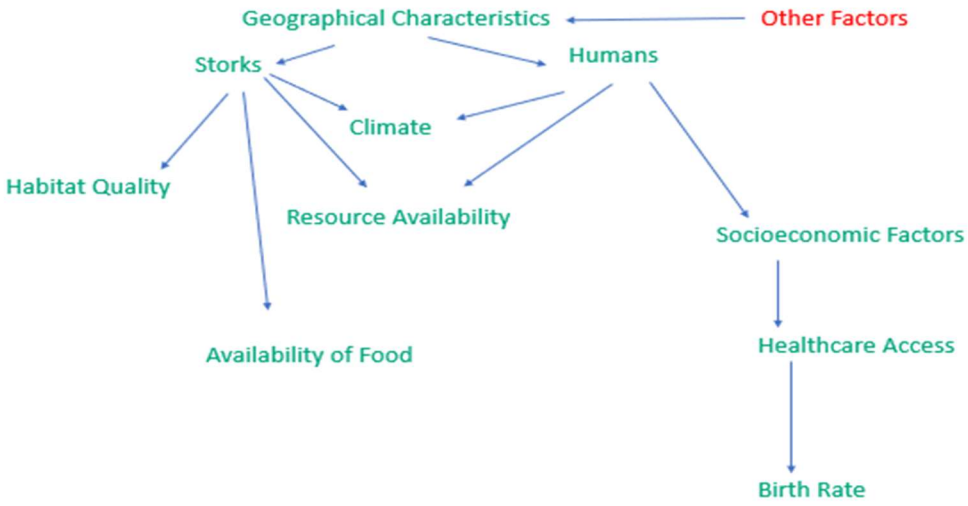
The paper "Storks Deliver Babies" shows the difference between correlation and causation, using the example of the purported relationship between stork populations and human birth rates. While a statistically significant correlation coefficient ($r = 0.62$, $p = 0.008$) suggests a potential association between the two variables, it does not imply causation. The paper emphasises the importance of considering confounding variables, such as land area, which may explain the observed correlation without indicating a direct causal link. The findings highlight the issue of drawing causal inferences only based on correlation and is a calling for researchers to take care when trying to disprove null hypothesis. The misinterpretation of p-values leads to errors drawing conclusions if confounding variables are not accounted for. Despite the statistical significance, the interpretation of the stork-birth correlation serves as an example, urging an understanding of statistical outcomes and emphasising that rejecting the null hypothesis does not validate the hypothesis.

Section 5: 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 (see lecture topic 13) that you believe is most likely to explain the relationship between the 4 variables Area, Storks, Humans and BirthRate used in the paper.

Answer

A confounding variable is an external factor that can distort the true connection. In the relationship between storks and births, it would be a factor affecting the relationship seen between stork populations and human birth rates. Possible confounders include climate conditions influencing both stork habitats and human behaviours, the quality of stork habitats correlating with areas conducive to human settlements, environmental changes impacting both stork populations and human lifestyles, and human interactions through activities like pollution affecting stork habitats and indirectly influencing human birth rates. The interconnected relationship is shown in a diagram below, illustrating the complexity of factors that may confound the observed correlation and emphasising the need to consider multiple variables for a comprehensive understanding of the relationship.



Acknowledgments

The data in this coursework has been provided by Chris Harbard of the Royal Society for the Protection of Birds. Additionally, credit is extended to Robert Matthews for his original paper, "Storks Deliver Babies ($p = 0.008$)," published in Teaching Statistics, Volume 22, Number 2, Summer 2000, pages 36-8. Their contributions form the foundation for this paper.