Estimating Secchi Depth using Tobit Regression

Group No. 4

Aditya Kishore 21011, Akshat Pandey 21024
Aryan Jain 21055 & Siddhi Lipare 21262

Introduction

A. Background

The Secchi disk is a simple device used to measure water transparency or clarity in bodies of water, such as lakes or oceans. The Secchi disk is a black-and-white circular disk that is lowered into the water until it is no longer visible, and then the depth at which it disappears is measured. This measurement, known as the Secchi depth, is an indicator of water transparency and is influenced by factors such as the concentration of suspended particles and algae in the water.

There are datasets that record Secchi depth measurements along with other water quality parameters, collected at various locations and times. These datasets are valuable for understanding the health and characteristics of aquatic ecosystems.

B. Objective

The primary objective of this report is to investigate and quantify the relationship between Secchi depth, a crucial indicator of water transparency, and various water characteristics/chemistry. By employing a Tobit regression model, specifically tailored for handling right-censored data inherent in the Secchi depth measurements, our aim is to derive consistent and meaningful estimates of the impact of independent variables, such as alkalinity, chloride, carbon organic content, nitrogen, and sulphate, on the observed Secchi depth. Additionally, this study seeks to assess the significance of each variable in influencing water transparency and to provide insights into the overall model fit. Through these analyses, we aim to contribute to a deeper understanding of the factors influencing Secchi depth, thereby enhancing our ability to assess and manage water quality effectively.

Data Description

The dataset utilized in this analysis has been sourced from Kaggle, a prominent platform for data science and machine learning enthusiasts. This dataset encompasses comprehensive information from more than 135 geographical locations, each associated with Secchi depth measurements. These measurements, indicative of water transparency, serve as the dependent variable in our analysis.

- The dataset consists of 404 observations, each representing a distinct geographical location.
- Each observation is associated with 30 features, providing a diverse set of water characteristics and chemistry information.
- The 30 features encapsulate various parameters related to water quality, including but not limited to alkalinity, chloride levels, carbon organic content, nitrogen content, and sulphate concentration.
- Georgian Bay Secchi Dataset

Methodology

Tobit Regression

Tobit regression is a statistical model designed for handling censored data, where the dependent variable is only partially observed due to detection limits or other constraints. Unlike standard Ordinary Least Squares (OLS) regression, Tobit regression accounts for both observed and censored values by employing a maximum likelihood estimation approach. This makes Tobit regression particularly suitable for situations where the dependent variable is either left- or right-censored. In the presence of censored observations, OLS regression can lead to biased and inconsistent parameter estimates, as it treats censored values as if they were fully observed, disregarding the inherent limitations in the data. OLS assumes a complete and unrestricted range for the dependent variable, making it ill-suited for scenarios where observations are intentionally constrained, such as in environmental studies or economic surveys. Tobit regression, by addressing these limitations, provides more accurate and reliable estimates, making it a preferred choice when dealing with censored data.

Choice of Independent Variables

The choice of independent variables is based on the correlation matrix analysis. In this approach, features with high correlation are identified, and one of them is selected as the independent variable to avoid multi-collinearity issues. High correlation between variables can lead to inflated standard errors and difficulties in interpreting individual variable contributions.

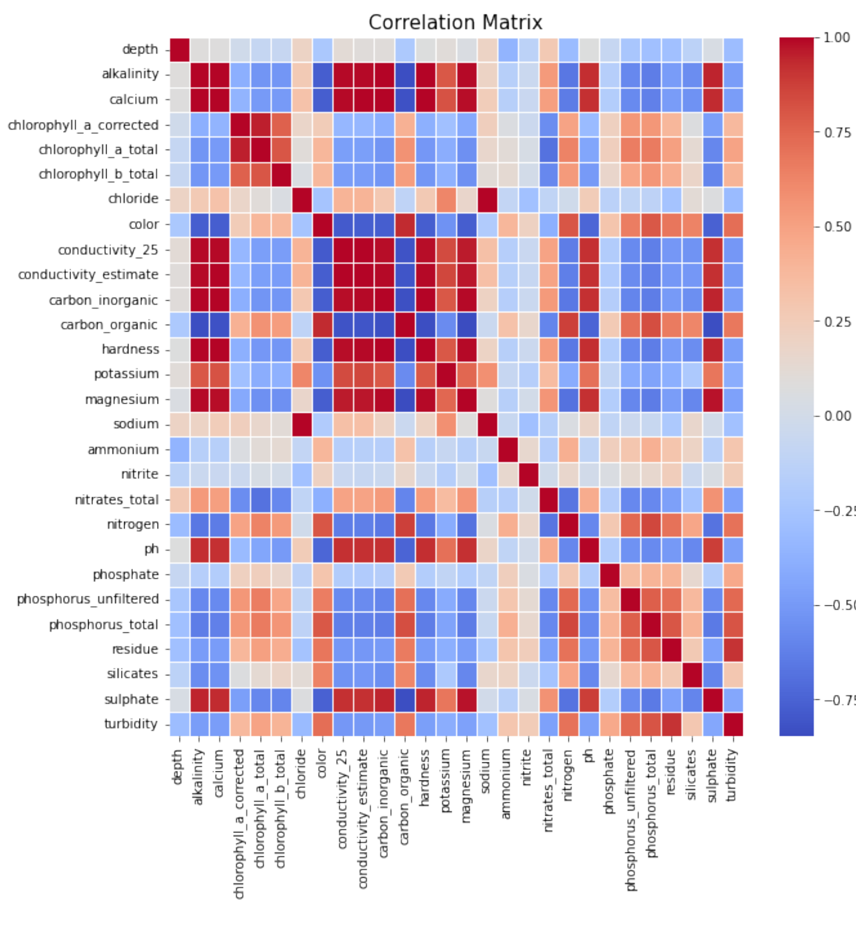

Figure 1: Correlation Matrix depicting correlation between features

Results

A. Tobit Regression Results

The Tobit regression analysis reveals insights into the relationship between Secchi depth and water characteristics/chemistry. The model, as evidenced by a significant likelihood ratio test (LL-Ratio) and a low AIC, provides a reasonable fit to the data. Notably, nitrogen emerges as a substantial predictor, displaying a highly significant negative coefficient, indicating that increased nitrogen levels are associated with a considerable decrease in Secchi depth. Conversely, alkalinity, chloride, carbon organic content, and sulphate exhibit non-significant coefficients, suggesting limited evidence of their impact on Secchi depth. The presence of heteroscedasticity, highlighted by a significant Log(Sigma) term, indicates variability in the error term across observations, warranting attention. Overall, these results contribute valuable insights into the factors influencing Secchi depth, emphasizing the significance of nitrogen as a key determinant.

```
tobit_res = tobit.Tobit(endog, exog, cens).fit()
tobit_res.summary()
```

```
Optimization terminated successfully.
         Current function value: 1.932589
         Iterations: 700
         Function evaluations: 1054
```

Figure 2: Tobit Model Implementation

Tobit Regression Results

Dep. Variable:	secchi_depth	**Pseudo R-squ:**	0.148
Method:	Maximum Likelihood	**Log-Likelihood:**	-780.8
No. Observations:	404	**LL-Null:**	-916.7
No. Uncensored Obs:	359	**LL-Ratio:**	271.8
No. Left-censored Obs:	0	**LLR p-value:**	0.000
No. Right-censored Obs:	45	**AIC:**	1573.5
Df Residuals:	398	**BIC:**	1597.5
Df Model:	5	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	8.1926	1.389	5.899	0.000	5.471	10.914
alkalinity	0.0375	0.023	1.624	0.104	-0.008	0.083
chloride	0.0359	0.060	0.601	0.548	-0.081	0.153
carbon_organic	-0.0364	0.235	-0.155	0.877	-0.497	0.424
nitrogen	-23.0158	3.609	-6.377	0.000	-30.089	-15.942
sulphate	0.0047	0.164	0.029	0.977	-0.316	0.325
Log(Sigma)	0.7515	0.037	20.269	0.000	0.679	0.824

Figure 3: Tobit Regression Results

B. Comparison with OLS

The Ordinary Least Squares (OLS) regression results provide insights into the relationship between Secchi depth and water characteristics/chemistry. The model exhibits a moderate explanatory power with an R-squared of 0.505, suggesting that approximately 50.5% of the variability in Secchi depth is explained by the chosen independent variables. Notably, nitrogen stands out as a significant predictor with a highly negative coefficient, indicating that an increase in nitrogen levels is associated with a substantial decrease in Secchi depth. The intercept is statistically significant, representing the estimated Secchi depth when all independent variables are zero. However, other variables such as alkalinity, chloride, carbon organic content, and sulphate show non-significant coefficients, suggesting limited evidence of their individual impact on Secchi depth. The model's overall significance is confirmed by a significant F-statistic (81.24). The Durbin-Watson statistic (1.523) indicates potential autocorrelation, while the Omnibus and Jarque-Bera tests highlight deviations from normality in the residuals. These results offer valuable insights into the relationships between water characteristics and Secchi depth, though the potential issues indicated by diagnostic tests should be considered in the interpretation.

OLS Regression Results

Dep. Variable:	secchi_depth	**R-squared:**	0.505
Model:	OLS	**Adj. R-squared:**	0.499
Method:	Least Squares	**F-statistic:**	81.24
Date:	Wed, 15 Nov 2023	**Prob (F-statistic):**	1.25e-58
Time:	16:57:06	**Log-Likelihood:**	-879.17
No. Observations:	404	**AIC:**	1770.
Df Residuals:	398	**BIC:**	1794.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
intercept	8.9357	1.374	6.504	0.000	6.235	11.637
alkalinity	0.0390	0.023	1.716	0.087	-0.006	0.084
chloride	0.0260	0.059	0.440	0.660	-0.090	0.142
carbon_organic	0.0505	0.230	0.219	0.827	-0.402	0.503
nitrogen	-25.7125	3.465	-7.422	0.000	-32.524	-18.901
sulphate	-0.0441	0.161	-0.274	0.784	-0.360	0.272

Omnibus:	212.470	**Durbin-Watson:**	1.523
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1737.946
Skew:	2.098	**Prob(JB):**	0.00
Kurtosis:	12.254	**Cond. No.**	1.85e+03

Figure 4: Tobit Regression Results

Discussion

Difference between OLS & Tobit Results

The primary difference between the Ordinary Least Squares (OLS) regression and Tobit regression lies in their suitability for handling data with censored observations. OLS assumes continuous, fully observed data and does not explicitly account for censoring, making it less reliable in scenarios where certain values are only partially observed. In contrast, Tobit regression is specifically designed for censored data, providing consistent estimates by incorporating both observed and censored values in the modeling process. While OLS is appropriate for standard linear relationships, Tobit is tailored for scenarios, such as the analysis of Secchi depth, where accurate estimation requires accounting for the presence of right-censored observations and the associated limitations in the data.

Conclusion

The analysis of Secchi depth and water characteristics/chemistry using both Ordinary Least Squares (OLS) regression and Tobit regression yields valuable insights. In the OLS model, nitrogen emerges as a significant predictor, demonstrating a substantial negative impact on Secchi depth. However, the OLS results may be influenced by the presence of censored data, potentially leading to biased estimates. In contrast, Tobit regression, specifically designed for censored data, confirms the significance of nitrogen while addressing potential biases associated with censoring. The Tobit model, with its likelihood-based statistics and consideration of both observed and censored values, provides a more robust approach for estimating the relationship between water parameters and Secchi depth. While some variables in both models exhibit non-significant coefficients, the Tobit results emphasize the importance of accounting for the censored nature of the data for more accurate and reliable estimates. The report concludes that Tobit regression is the preferred model for this analysis, offering a comprehensive understanding of the factors influencing Secchi depth in the presence of censored observations.