Exploring the Mortality Rate of the Trans-Atlantic Slave Trade Using Logistic Regression

Thomas Jefferson High School for Science and Technology

06/08/22

Rationale:

Starting in the 1500s, slave traders began using the Middle Passage to transport enslaved Africans to the Americas as part of the Triangular Trade. As mercantile fervor increased, so did the rates at which captives were forced to become enslaved peoples in the Americas. It is estimated that by the 1700s, over two million Africans had been forcibly removed from their homes and brought to the Amerias where they would be forced to work on plantations (Mustakeem, 2008). In order to do so, slave traders would pack high densities of slaves aboard slave ships, which had deplorable conditions. For starters, the tightly packed confines of the slave ships became a breeding ground for bacteria, and it poses no surprise that disease ran rampant amongst the enslaved peoples. Some of the most common diseases that affected enslaved peoples included scurvy, dysentery (known then as "the flux"), small pox, and malaria. Besides the close proximity, the food rations given out to the slaves lacked nutritional values which also contributed to the slaves' susceptibility to disease (Steckel & Jensen, 1986).

The meals given to the slaves was largely based on which foods could be preserved and stored on ships for long periods of time. Because of high temperatures and changing weather conditions, it was difficult to transport perishable foods like fruits and vegetables, so their diets consisted primarily of carbohydrates and proteins. However, without the consumption of fruits and vegetables, the captives did not have the nutrients that they needed to fight off illnesses. In fact, their meals themselves were often the cause of illnesses for the captives. For instance, in order to cure the meat, large quantities of salt were added to the meat. However, this increase in salt intake caused severe dehydration, which when coupled with the intense conditions of the slave ships, resulted in death for many people. Even their dry provisions, like oatmeal, peas, and flour, were corrupted by maggots and mold. Often undetectable, these bad provisions caused slaves to be even more susceptible to sickness and illness. In addition, a high daily intake of carbohydrates caused dangerous toxins to remain in the body, also resulting in a higher susceptibility to illnesses (Mustakeem, 2008).

Interestingly, different studies have made different conclusions as to which variables affected slave ships the most. In his analysis, Miller concluded that overcrowding aboard slave ships resulted in high death rates and that the levels of death varied by year, indicating that land based pandemics caused many deaths aboard slave ships. Meanwhile, Klein et al. suggested the opposite, indicating that "tight" versus "loose packing" had little effect on the mortality of slaves, because there were still more slaves than free sailors aboard the ships. However, unanimously, all of the studies suggested that the mortality rate for slaves was U-shaped, rather than linear. In general, the slaves died at higher rates on shorter voyages (20 to 40 days long) than they did on intermediate voyages (40 to 50 days long), and it wasn't until long voyages (50+ days) that the hardships of extended time at sea became evident. As a result, the mortality rate turned upwards whenever the voyage length lasted 50 percent longer than the average voyage length (Miller 1981).

Therefore, the purpose of this study is to analyze data from a database containing information about various slave voyages in order to ascertain the effect of several different variables, including the length of the voyage, the number of slaves on the voyage, and the year of the voyage, in order to determine which variables had the most significant affect on slave mortality rates. In doing so, we hope to establish correlation between the variables and thoroughly analyze the slave trade in order to better understand the ordeal that the slaves underwent. Our goal is to determine how factors such as the initial number of slaves, the year that the voyage was made, and the nation-state that sponsored the voyage impacted the mortality rate of the slaves on the vessel.

Some questions we hope to answer include:

- What is the relationship between the mortality rate and various factors including the year the voyage was made, the initial number of slaves on the ship, the length of the voyage, and the nation-state that sponsored the voyage?
- What shape will the mortality rate take on given this combination of variables?

Design of the Study:

The data we will be analyzing has been provided to us by the SlaveVoyages database and contains detailed information concerning each voyage. The data, as listed on the database, is sorted by each vessel's voyage ID, which is based on the year each voyage occurred. Therefore, after downloading the data as an excel sheet, we plan to randomize the order that the vessels are listed in, and then use a random number generator to select two hundred vessels to serve as our training sample and another two hundred to be our holdout sample. Though the database provides us an assortment of information about each vessel, the data that we plan to utilize includes the length of the voyage, the number of slaves on that particular vessel, the year the vessel arrived at the slave landing (as calculated by their algorithm), the flag of the ship, and the slave mortality rate. After compiling the data of all the participating slave vessels, we will perform a logistic regression on the mortality rate. We had ideas about how we expected each of the factors to effect the mortality rate. We hypothesized that higher slave populations on slave ships would lead to higher mortality rates due to overcrowding and faster spread of disease. We also hypothesized that later voyages would have a lower mortality rate as new laws were enacted to 'protect' the slaves, and that ships sailing under the flags of known slave trading empires would have higher mortality rates.

In our analysis of the slave voyage data, we started by developing a multiple logistic regression model for Mortality Rate (MortRate) based on the year the voyage was made (Year), the initial number of slaves on the ship (Total Embarked), the length of the voyage (Length), and the nation-state that sponsored the voyage (Flag). We selected these explanatory variables from the lot because we reasoned that they would have an important impact on the MortRate. We also used interaction terms between Flag and all 3 quantitative explanatory variables individually to account for some nations joining the slave trade later, some nations making shorter voyages, and some nations transporting fewer slaves per ship. We had an interaction term between Year and Total Embarked to encapsulate any changes in slaves transported per ship as the slave trade matured as well as a quadratic term for Year to account for laws that might have mitigated slave deaths aboard ships in the late years of the slave trade.

After the initial regression, we attempted to improve the R^2-value and fix any slight normality and variance issues with transformations on Logit(MortRate) = ln(MortRate / (1-MortRate)). We landed on a square-root transformation of Logit(MortRate). Due to Logit(MortRate) outputting negative values, the transformation we applied wasn't exactly a square-root, but a square-root on -1 * Logit(MortRate). For simplicity's sake, we will still refer to it as the square-root transformation as the result is essentially the same. Once we obtained this full model, we ran the regression again, but using backward elimination with $\alpha = 0.05$ to remove several predictors with very large p-values. The next step was to run a nested-F test on the reduced model to determine if those extra variables were actually important in the full model. Then we took the reduced model and tested it on a holdout sample and retrieved the cross-validation correlation and the shrinkage. Lastly, we decided to look at a multiple linear regression with the same predictors as the reduced model in order to see if the logistic regression was even necessary or if a straight line could obtain superior results at this scale.

Results:

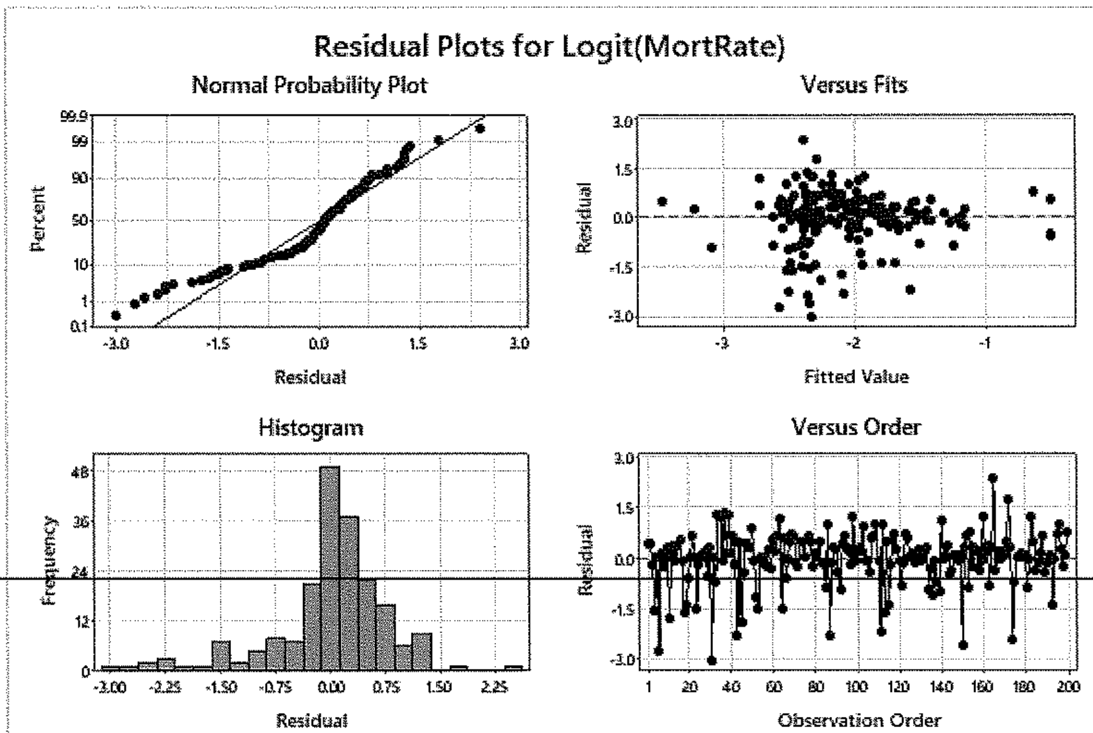
- Part 1: Logistic Regression with all predictors

Regression Equation

Flag (Imputed)	
France	Logit(MortRate) = 67 - 0.079 Year + 0.00216 Length + 0.0574 Total Embarked + 0.000022 Year*Year - 0.000032 Year*Total Embarked
Great Britain	Logit(MortRate) = 67 - 0.079 Year + 0.001135 Length + 0.0582 Total Embarked + 0.000022 Year*Year - 0.000032 Year*Total Embarked
Netherlands	Logit(MortRate) = 67 - 0.078 Year + 0.00168 Length + 0.0550 Total Embarked + 0.000022 Year*Year - 0.000032 Year*Total Embarked
Portugal / Brazil	Logit(MortRate) = 51 - 0.070 Year - 0.000119 Length + 0.0586 Total Embarked + 0.000022 Year*Year - 0.000032 Year*Total Embarked
Spain / Uruguay	Logit(MortRate) = 115 - 0.101 Year - 0.0277 Length + 0.0563 Total Embarked + 0.000022 Year*Year - 0.000032 Year*Total Embarked
U.S.A.	Logit(MortRate) = 69 - 0.080 Year + 0.00601 Length + 0.0589 Total Embarked + 0.000022 Year*Year - 0.000032 Year*Total Embarked

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	25	38.559	1.5424	2.07	0.004
Year	1	0.278	0.2783	0.37	0.542
Length	1	1.378	1.3782	1.85	0.176
Total Embarked	1	2.713	2.7131	3.64	0.058
Flag (imputed)	5	5.714	1.1428	1.53	0.182
Year*Year	1	0.279	0.2790	0.37	0.542
Year*Total Embarked	1	2.641	2.6411	3.54	0.062
Year*Flag (imputed)	5	5.894	1.1787	1.58	0.168
Length*Flag (imputed)	5	8.114	1.6228	2.18	0.059
Total Embarked*Flag (imputed)	5	1.046	0.2091	0.28	0.923
Error	174	129.749	0.7457		
Total	199	168.308			



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.863529	22.91%	11.83%	0.00%

- Part 2: Log-transformed Logistic Regression with all predictors

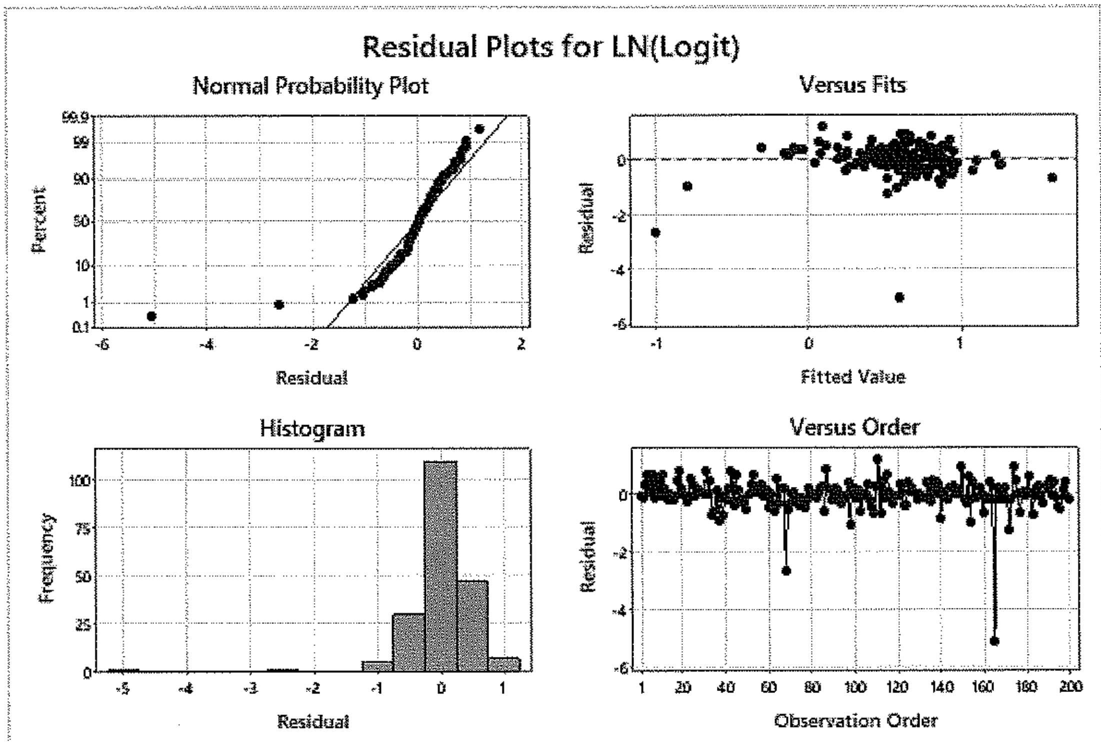
Regression Equation

Flag (imputed)

Flag (imputed)		
France	LN(Logit) =	-97.8 + 0.1211 Year - 0.00066 Length - 0.0729 Total Embarked - 0.000037 Year*Year + 0.000041 Year*Total Embarked
Great Britain	LN(Logit) =	-106.6 + 0.1263 Year - 0.000593 Length - 0.0742 Total Embarked - 0.000037 Year*Year + 0.000041 Year*Total Embarked
Netherlands	LN(Logit) =	-101.2 + 0.1227 Year - 0.00081 Length - 0.0705 Total Embarked - 0.000037 Year*Year + 0.000041 Year*Total Embarked
Portugal / Brazil	LN(Logit) =	-84.8 + 0.1142 Year + 0.000228 Length - 0.0757 Total Embarked - 0.000037 Year*Year + 0.000041 Year*Total Embarked
Spain / Uruguay	LN(Logit) =	-116.7 + 0.1298 Year + 0.01433 Length - 0.0732 Total Embarked - 0.000037 Year*Year + 0.000041 Year*Total Embarked
U.S.A.	LN(Logit) =	-107.8 + 0.1269 Year - 0.00247 Length - 0.0748 Total Embarked - 0.000037 Year*Year + 0.000041 Year*Total Embarked

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	25	19.5327	0.7813	2.18	0.002
Year	1	0.6555	0.6555	1.83	0.178
Length	1	0.1277	0.1277	0.36	0.551
Total Embarked	1	4.3813	4.3813	12.24	0.001
Flag (imputed)	5	5.7320	1.1464	3.20	0.009
Year*Year	1	0.7639	0.7639	2.13	0.146
Year*Total Embarked	1	4.3396	4.3396	12.13	0.001
Year*Flag (imputed)	5	5.7750	1.1550	3.23	0.008
Length*Flag (imputed)	5	1.9312	0.3862	1.08	0.374
Total Embarked*Flag (imputed)	5	2.4457	0.4891	1.37	0.239
Error	174	62.2706	0.3579		
Total	199	81.8033			



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.598228	23.88%	12.94%	0.00%

- Part 3: Square-Root-transformed Logistic Regression with all predictors

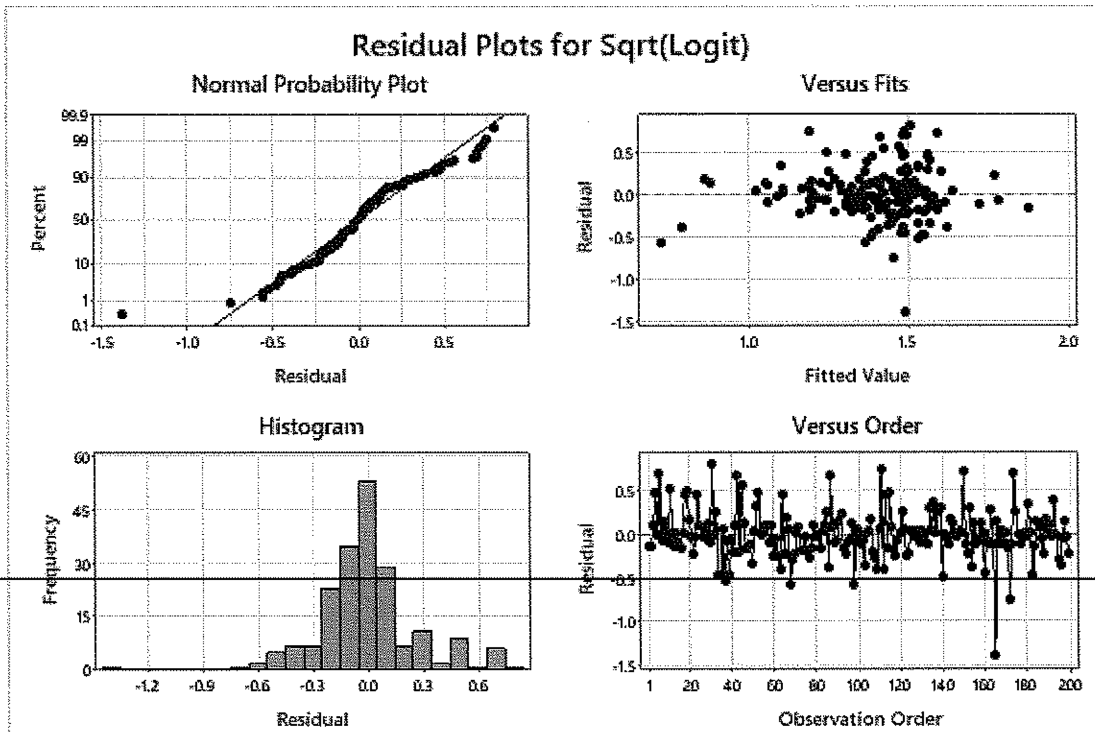
Regression Equation

Flag (imputed)

France	Sqrt(Logit) =	-32.6 + 0.0406 Year - 0.000634 Length - 0.0279 Total Embarked - 0.000012 Year*Year + 0.000016 Year*Total Embarked
Great Britain	Sqrt(Logit) =	-34.4 + 0.0416 Year - 0.000361 Length - 0.0283 Total Embarked - 0.000012 Year*Year + 0.000016 Year*Total Embarked
Netherlands	Sqrt(Logit) =	-33.8 + 0.0411 Year - 0.00056 Length - 0.02687 Total Embarked - 0.000012 Year*Year + 0.000016 Year*Total Embarked
Portugal / Brazil	Sqrt(Logit) =	-26.5 + 0.0372 Year + 0.000075 Length - 0.0287 Total Embarked - 0.000012 Year*Year + 0.000016 Year*Total Embarked
Spain / Uruguay	Sqrt(Logit) =	-49.0 + 0.0482 Year + 0.00973 Length - 0.0275 Total Embarked - 0.000012 Year*Year + 0.000016 Year*Total Embarked
U.S.A.	Sqrt(Logit) =	-34.3 + 0.0416 Year - 0.00190 Length - 0.0286 Total Embarked - 0.000012 Year*Year + 0.000016 Year*Total Embarked

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	25	5.3859	0.21544	2.44	0.000
Year	1	0.0737	0.07367	0.84	0.362
Length	1	0.1185	0.11852	1.34	0.248
Total Embarked	1	0.6399	0.63988	7.26	0.008
Flag (imputed)	5	1.0207	0.20413	2.32	0.046
Year*Year	1	0.0806	0.08059	0.91	0.340
Year*Total Embarked	1	0.6284	0.62839	7.13	0.008
Year*Flag (imputed)	5	1.0386	0.20771	2.36	0.042
Length*Flag (imputed)	5	0.9258	0.18516	2.10	0.067
Total Embarked*Flag (imputed)	5	0.2726	0.05452	0.62	0.686
Error	174	15.3326	0.08812		
Total	199	20.7185			



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.296847	26.00%	15.36%	0.00%

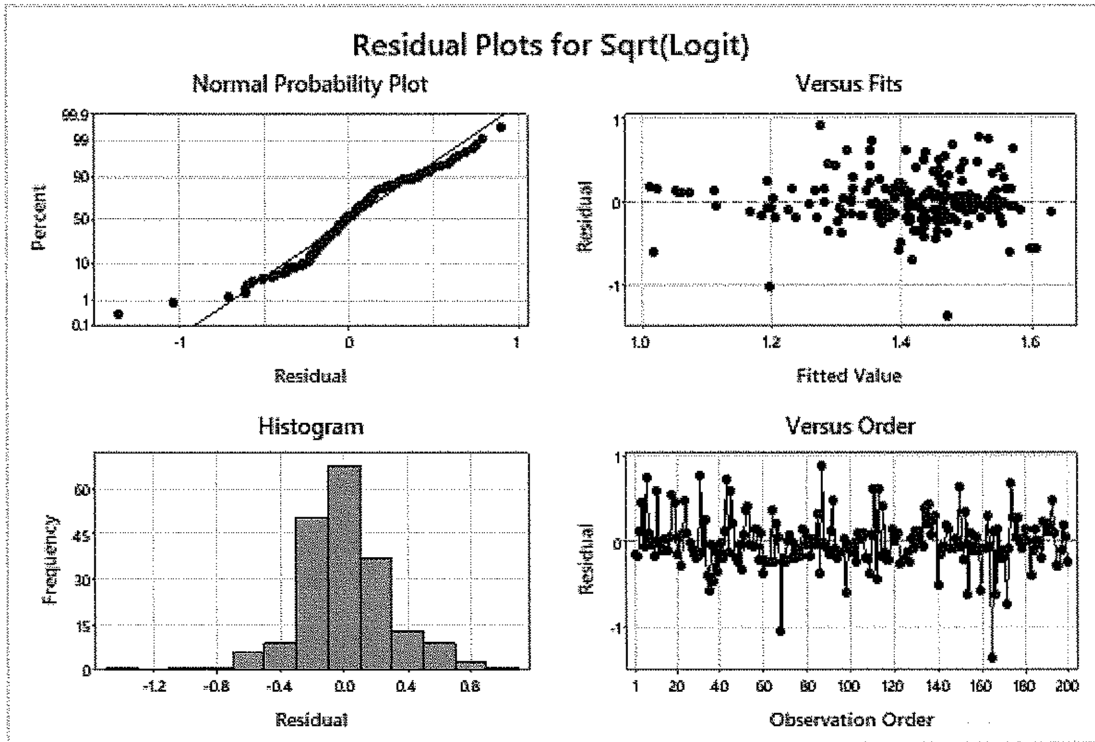
- Part 4: Square-Root-transformed Logistic Regression with Backward Elimination

Regression Equation

Flag (imputed)	
France	Sqrt(Logit) = -2.466 + 0.002287 Year - 0.000369 Total Embarked
Great Britain	Sqrt(Logit) = -2.488 + 0.002287 Year - 0.000369 Total Embarked
Netherlands	Sqrt(Logit) = -2.501 + 0.002287 Year - 0.000369 Total Embarked
Portugal / Brazil	Sqrt(Logit) = -2.505 + 0.002287 Year - 0.000369 Total Embarked
Spain / Uruguay	Sqrt(Logit) = -2.522 + 0.002287 Year - 0.000369 Total Embarked
U.S.A.	Sqrt(Logit) = -2.793 + 0.002287 Year - 0.000369 Total Embarked

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	3.0261	0.43230	4.69	0.000
Year	1	2.1181	2.11809	22.99	0.000
Total Embarked	1	0.4444	0.44443	4.82	0.029
Flag (imputed)	5	1.1809	0.23618	2.56	0.029
Error	192	17.6924	0.09215		
Total	199	20.7185			



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.303559	14.61%	11.49%	7.34%

- Part 5: Nested-F Tests and Cross-Validation Correlation on Final Model

Nested-F Test of Full Model against Year, Total Embarked, Flag - Model

Hypotheses:

H_0: $\beta_i = 0$ for all predictors in the reduced model

H_a: $\beta_i \neq 0$ for at least 1 predictor in the reduced model

Conditions:

Random: The data was randomly selected from the overall dataset

Independent: The factors of one sailing voyage do not have an impact on the others.

Normality: As seen from the normal probability plots of the full and reduced square root models above, the distributions are roughly normal.

F-statistic & P-value:

F = (SSM_{full} - $SSM_{reduced}$ / #predictors) / (SSE_{full} / n-k-1) = 8.626

dfNUM = 3, dfDEN = 175

P-value = 0.00002266

Conclusion:

Since the p-value was less than alpha = 0.05, we can reject the null hypothesis that the predictors in the reduced model are not significant. We have sufficient evidence to conclude that the reduced model is still significant compared to the full model.

Cross-Validation & Shrinkage:

When using the model to predict the MortRate in the holdout set, we obtained a **cross-validation correlation of 0.354**.

Pairwise Pearson Correlations

Sample 1	Sample 2	N	Correlation	95% CI for ρ	P-Value
Sqrt(Logit)	Prediction	200	0.354	(0.226, 0.469)	0.000

The shrinkage we calculated was: $R^2_{training} - R^2_{holdout} = 0.1461 - 0.1253 = 0.0208$

The shrinkage is 2.08%, which is quite reasonable.

- Part 6: Comparing Logistic and Linear Models

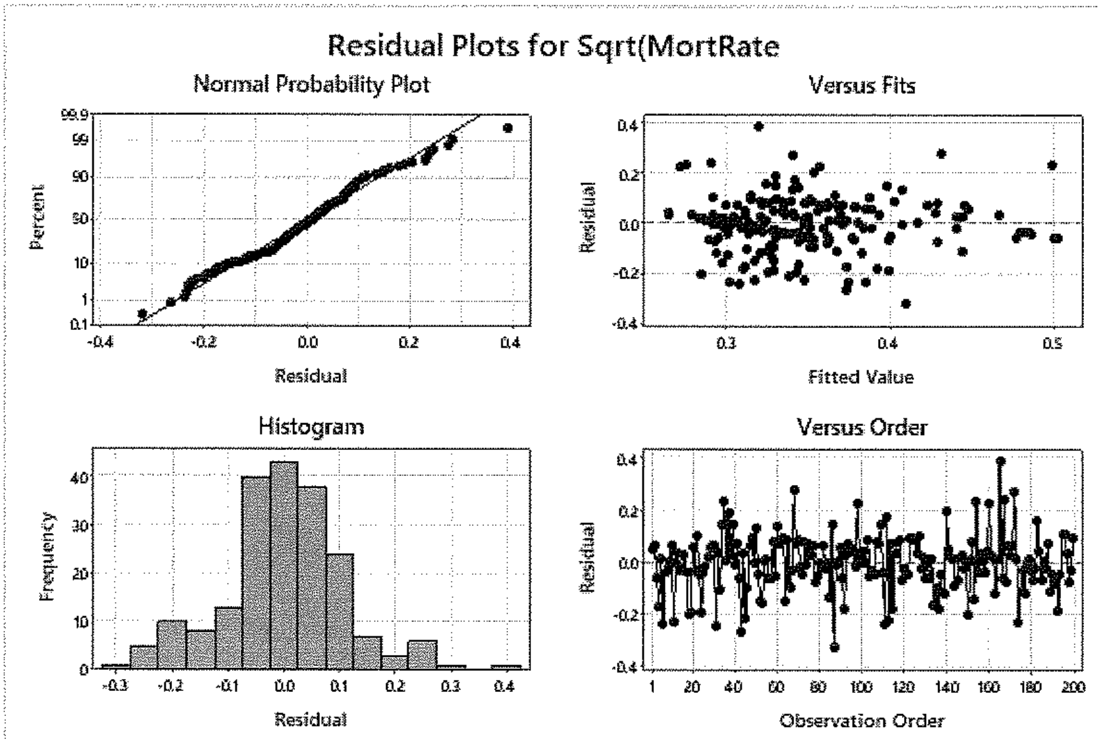
Linear:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.881	0.304	6.19	0.000	
Year	-0.000899	0.000174	-5.16	0.000	1.13
Total Embarked	0.000128	0.000061	2.08	0.039	1.27
Flag (imputed)					
Great Britain	0.0063	0.0228	0.27	0.784	2.10
Netherlands	0.0128	0.0463	0.28	0.782	1.18
Portugal / Brazil	0.0111	0.0258	0.43	0.667	1.84
Spain / Uruguay	0.0212	0.0423	0.50	0.617	1.25
U.S.A.	0.1313	0.0367	3.57	0.000	1.53

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	0.47760	0.06823	5.56	0.000
Year	1	0.32735	0.32735	26.67	0.000
Total Embarked	1	0.05319	0.05319	4.33	0.039
Flag (imputed)	5	0.19591	0.03918	3.19	0.009
Error	192	2.35623	0.01227		
Total	199	2.83384			



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.110779	16.85%	13.82%	9.51%

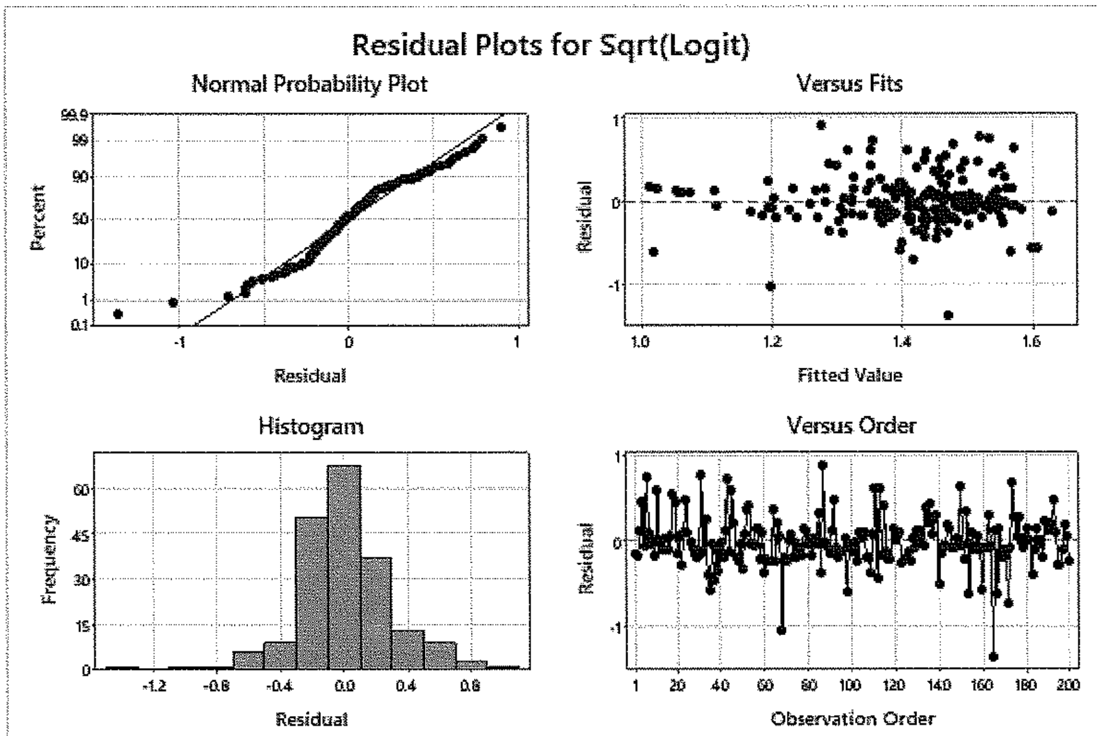
Logistic:

Regression Equation

Flag (imputed)	
France	Sqrt(Logit) = -2.466 + 0.002287 Year - 0.000369 Total Embarked
Great Britain	Sqrt(Logit) = -2.488 + 0.002287 Year - 0.000369 Total Embarked
Netherlands	Sqrt(Logit) = -2.501 + 0.002287 Year - 0.000369 Total Embarked
Portugal / Brazil	Sqrt(Logit) = -2.505 + 0.002287 Year - 0.000369 Total Embarked
Spain / Uruguay	Sqrt(Logit) = -2.522 + 0.002287 Year - 0.000369 Total Embarked
U.S.A.	Sqrt(Logit) = -2.793 + 0.002287 Year - 0.000369 Total Embarked

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	3.0261	0.43230	4.69	0.000
Year	1	2.1181	2.11809	22.99	0.000
Total Embarked	1	0.4444	0.44443	4.82	0.029
Flag (imputed)	5	1.1809	0.23618	2.56	0.029
Error	192	17.6924	0.09215		
Total	199	20.7185			



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.303559	14.61%	11.49%	7.34%

Both models come from the same data, just transformed in different ways, so they are both random and independent, as stated above. Both models are also roughly linear on the normal probability plot, so the normality condition is cleared. However, the R^2 values for the linear model are slightly larger than that of the logistic model. The linear model's normal probability plot is also slightly more linear and histogram seems more bell-shaped compared to the same plots of the logistic model.

These details seem to indicate that the linear model would actually be a bit better for the data than the logistic model that we used. However, we stand by the logistic model because at the extreme ends of the range of percentages (close to 0 or close to 1), the logistic model works better for percentages than the linear model would. Even though our data doesn't tend to reach 0 or 1, that may not always be the case if we were to test against more and more data points.

NOTE: The Linear model was square-root transformed because a square-root transform was applied to the Logistic model as well and we wanted to keep the comparison between the two strategies as balanced as possible.

Discussion:

Our results indicate that the model with the most simplicity yet effectiveness is the Square-Root Logit model with backward elimination. From the ANOVA tables for Square-Root Logits, with and without backward elimination, both are statistically significant at $\alpha = 0.05$. From the Nested-F of full vs. reduced, we determined that the model with backward elimination was still statistically significant. This model shows us that the relationship between Sqrt(Logit(MortRate)) and Year, Total Embarked, and Flag, is a linear relationship. In other words, MortRate and our predictors have a roughly S-shaped relationship. The coefficient of Year was positive, which indicates that over the years, mortality aboard ships decreased, while the coefficient on Total Embarked was negative, indicating that the more people there were onboard, the higher mortality rates tended to be (to see why the signs of the coefficients seem to be swapped, please refer to the final paragraph of the "Design of the Study" section). The model also indicates that ships sponsored by the USA tended to have much higher mortality rates than the other 5 main nation-states (Great Britain, Netherlands, France, and Spain, and Portugal). Our findings of an S-shaped mortality rate differed from the U-shaped mortality rate previously found by other studies. However, we believe that this difference can be attributed to the fact that we looked at several variables, while the studies we looked at focused primarily on the effect of the distance on the mortality rate. All of the variables we looked at were linearly associated with ln(x / 1-x), resulting in our S-shaped mortality rate. In the future, we would be interested in continuing our analysis by including variables like weather and disease that the database did not include. Several of our studies mentioned noticing a correlation between mortality rates and the year a ship left a certain port, which they believe indicates that seasonal diseases affected the mortality rates aboard slave ships. If we had information about those factors, we may have been able to develop a more powerful model, considering that ours had a smaller R^2 value of only 14.61%. We may have also been able to corroborate the U-shaped distribution that many authors predicted. A more powerful model would be able to more accurately reflect the true nature of the Trans-Atlantic Slave Trade and the ordeal these people went through.

Reflection:

Overall, we thoroughly enjoyed this project as it challenged us to pair our writing skills with our recently acquired RS3 knowledge to create a comprehensive research paper. Over the course of the last few weeks, we utilized a variety of skills, including our ability to find reliable sources to base our project rationale on, our ability to find a database with information in it, and our ability to perform logistic regressions on the data. The skills will continue to benefit us as we continue on with higher education. In college, there are standardized courses that everyone must complete like biology and chemistry, where we will be required to analyze our data and write up lab reports. Being able to thoroughly analyze the data and pinpoint the most significant variables will enhance the quality of our writing and the authenticity of our results.

References

Eltis, D. (1989). Fluctuations in Mortality in the Last Half Century of the Transatlantic Slave Trade. *Social Science History*, *13*(3), 315–340. https://doi.org/10.2307/1171374

Haines, R., & Shlomowitz, R. (2000). Explaining the Mortality Decline in the Eighteenth-Century British Slave Trade. *The Economic History Review*, *53*(2), 262–283. http://www.jstor.org/stable/2598698

Klein, H. S., Engerman, S. L., Haines, R., & Shlomowitz, R. (2001). Transoceanic mortality: The slave trade in comparative perspective. *The William and Mary Quarterly*, *58*(1), 93. https://doi.org/10.2307/2674420

Miller, J. C. (1981). Mortality in the Atlantic Slave Trade: Statistical Evidence on Causality. *The Journal of Interdisciplinary History*, *11*(3), 385–423. https://doi.org/10.2307/203625

Mustakeem, S. (2008). "I Never Have Such a Sickly Ship Before": Diet, Disease, and Mortality in 18th-Century Atlantic Slaving Voyages. *The Journal of African American History*, *93*(4), 474–496. http://www.jstor.org/stable/25610019

Simon J. Hogerzeil, & Richardson, D. (2007). Slave Purchasing Strategies and Shipboard Mortality: Day-to-Day Evidence from the Dutch African Trade, 1751-1797. *The Journal of Economic History*, *67*(1), 160–190. http://www.jstor.org/stable/4501137

Steckel, R. H., & Jensen, R. A. (1986). New Evidence on the Causes of Slave and Crew Mortality in the Atlantic Slave Trade. *The Journal of Economic History*, *46*(1), 57–77. http://www.jstor.org/stable/2121267