Cox Regression Model: Survival Analysis

Survival Analysis
Dr. (to be) Matthew Austin

Orange Team 2
Qing Feng, Andrew Moolenaar, Carlos Chávez, Julie Huang, Bill Jenista

Executive Summary

Tasked with analyzing motor and surge failures for pumps during Hurricane Katrina, our team found that the model that best fit the data was a Cox Regression model which models the pump's failure rate rather than the failure time. When looking specifically at how many consecutive hours the pumps ran during the hurricane, we found that pumps that had been running for at least 12 consecutive hours were more likely to have a motor or surge failure. Statistically, we can say that the motor/surge failure rate for pumps that ran for at least 12 consecutive hours were 1.07 times more than for pumps that did not run 12 consecutive hours.

Analysis & Results

The dataset provided by the Steering Committee of the Center for Risk Management contains information on 770 pump stations on the Gulf Coast during Hurricane Katrina. Within the critical 48-hour storm period, 59% of the 770 pump stations underwent different types of failures. Among all the pump stations with failure, 25% of them were due to the motor or surge with a median survival time of 45 hours and a mean of 41 hours. To understand the possible relationship between different factors and the failure of the pump stations due to the motor and surge, we treated motor and surge failures as the event and all other causes as censored. Additionally, we removed 3 pumps from our dataset due to missing values and data errors before performing the following analysis. Further clarification on the errors will be provided later in this report. The seven factors provided in the dataset are shown below (Table 1).

Factor Name	**Description**
Backup	*Whether a backup pump is present to protect the station from flooding when the main pump is not operating*
Bridgecrane	*Whether a crane is present to allow vertical access to equipment and protect materials*
Servo	*Whether a servo is present to provide control of the desired operation*
Trashrack	*Whether a cleaner is present to provide hydraulic structures*
Elevation	*Elevation of the pump station*
Slope	*Ravine slope surrounding the pump station*
Age	*How long the pump has been installed*

Table 1. Description of Factors in the Dataset

We fit different distributions of failure time, including Weibull, Exponential, Lognormal, and Log-logistic distribution, but none of them fit the data well (Appendix). For this reason, the Accelerated Failure Time (AFT) model is not appropriate as it assumes failure time has a particular structure and distribution. Instead, we decided to model the pump hazard using the Cox Regression model, since this model makes no assumption about the dataset's distribution.

The coefficient estimates (Table 1) from the Cox Regression model represent how variables would increase the risk of failure (higher hazard) or decrease the risk of failure (lower hazard)

that were caused by motor or surge. A positive coefficient estimate implies a higher hazard and vice versa. Trashrack has the largest coefficient estimate in absolute value, -1.05654. Exponentiated, this means for pump stations with the trashrack cleaner, the failure rate is 0.35 times that of those without. Age has a coefficient estimate of 0.11852, which implies an increased risk of failure, and an exponentiated estimate of 1.12583, which indicates the hazard is about 1.13 times higher for every additional degree increase. We think the coefficient estimates and standard errors of Backup, Bridgecrane, Servo, Elevation, and Age indicate that these variables had little effect on pump hazard.[1]

Variable	Coefficient Estimate	Exponentiated	Standard Error
Backup	-0.09	0.91	0.14
Bridgecrane	-0.04	0.96	0.20
Servo	0.29	1.34	0.16
Trashrack	-1.06	0.35	0.15
Elevation	-0.07	0.93	0.08
Slope	-0.10	0.91	0.03
Age	0.12	1.13	0.07

Table 2. Coefficient Estimates of Variables

By stratifying each categorical variable at a time, we plotted the hazards to see if they are proportional (Figure 1-4). Although not perfect parallel lines, Backup, Servo, and Trashrack could indicate some level of parallelism, meaning a comparatively constant effect over time. On the contrary, Bridgecrane's hazard lines are not parallel at all until the very end. This indicates a non-constant effect of bridgecrane.

[1] The standard error can be used to calculate the 95% confidence interval which in the case of the five variables mentioned, means the intervals include zero.

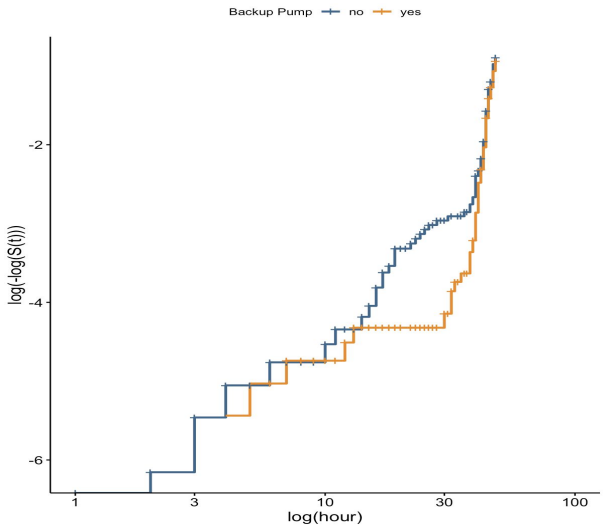


Figure 1

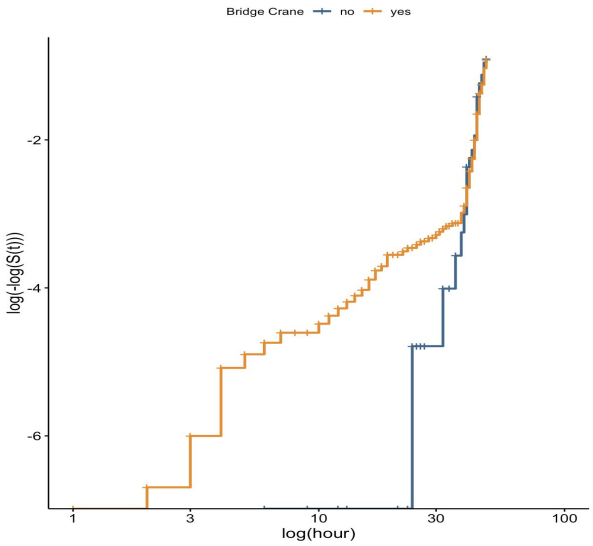


Figure 2

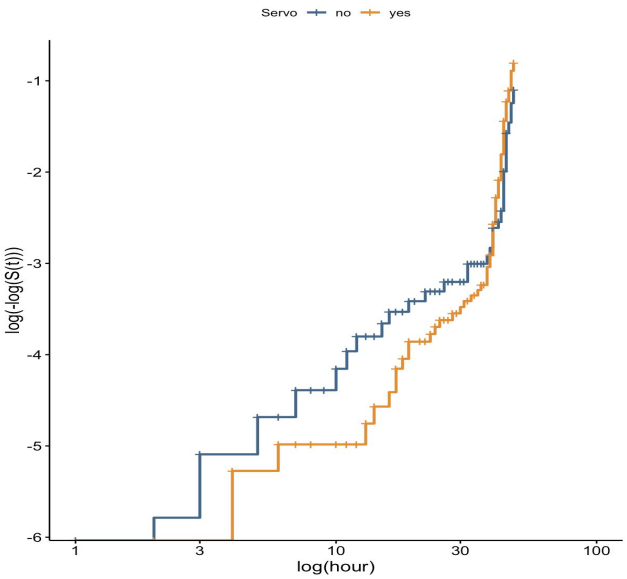


Figure 3

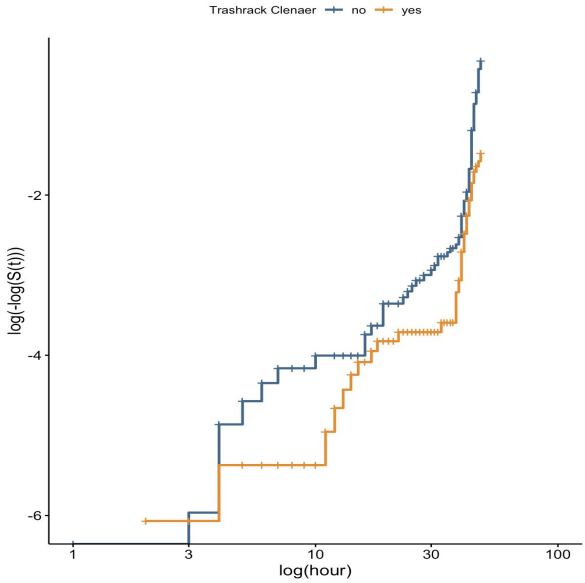


Figure 4

Stratified Hazard Plots

To meet the requests of the Army Corps of Engineers we built an additional model to see if pumps that ran for 12 consecutive hours or more were more likely to fail than others. This involved including a new indicator variable (Twelve) in our dataset to identify these pumps as well as a start and stop hour. Due to missing values, we removed three pumps from the dataset because we could not determine if the pumps were running or not. Based on the coefficient estimate of the variable Twelve from Table 3, it does appear that pumps are more likely to fail if

they have been running for 12 consecutive hours prior. Using Table 3, we can say that pumps that ran for 12 consecutive hours are 1.07 times more likely to fail than pumps that did not run for 12 consecutive hours. While this may seem like a somewhat small increase, it is still statistically significant nonetheless.

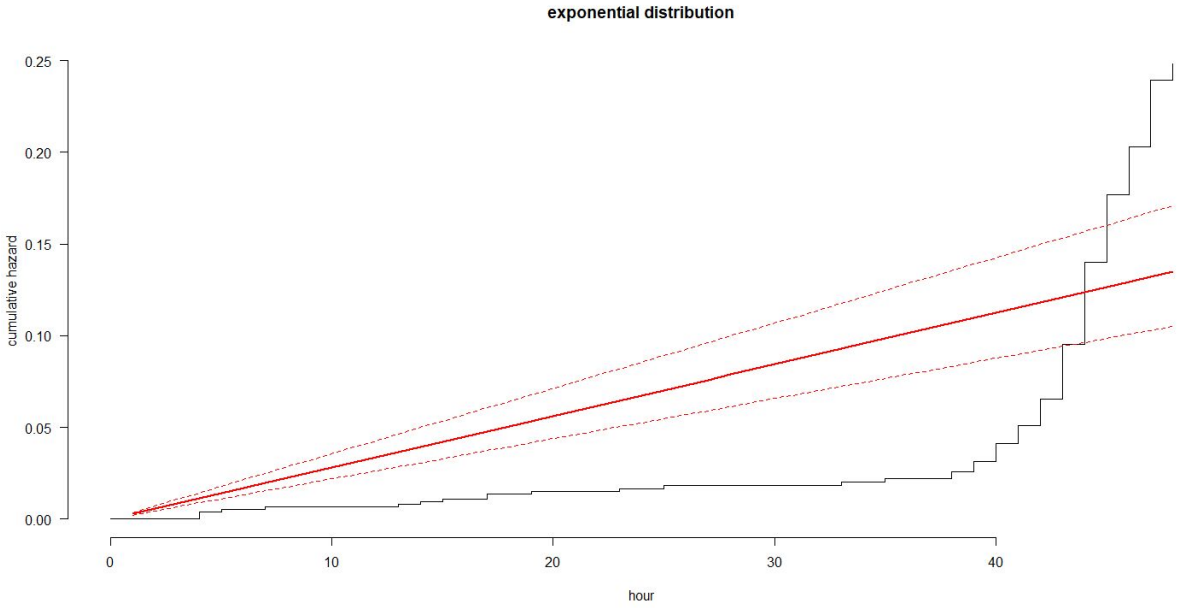
Variable	Coefficient Estimate	Exponentiated	Standard Error
Twelve	0.07	1.07	0.03
Backup	0.05	1.05	0.02
Bridgecrane	-0.10	0.91	0.03
Servo	0.33	1.38	0.03
Trashrack	-1.21	0.30	0.03
Elevation	0.02	1.02	0.01
Slope	-0.18	0.83	0.01
Age	0.15	1.16	0.01

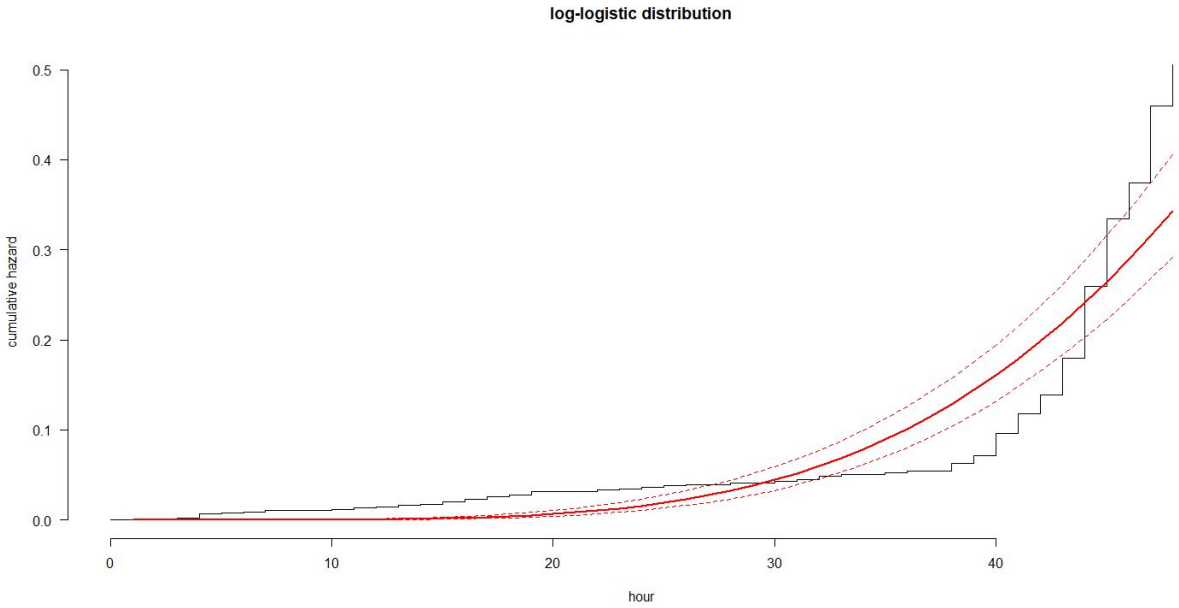
Table 3. Coefficient Estimates of Variables

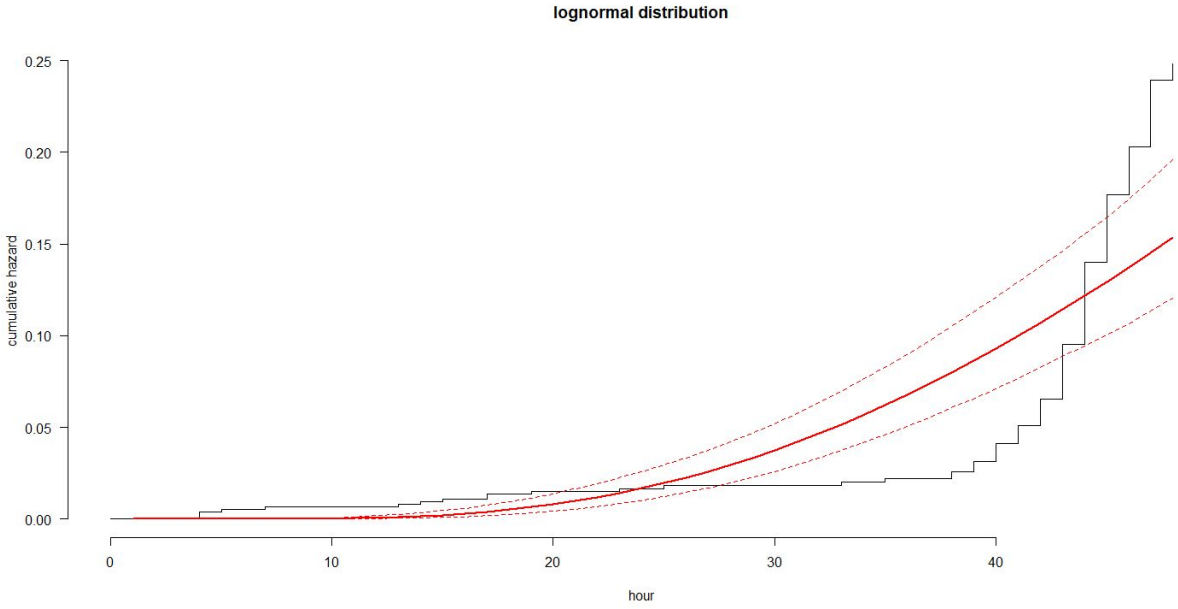
Conclusion

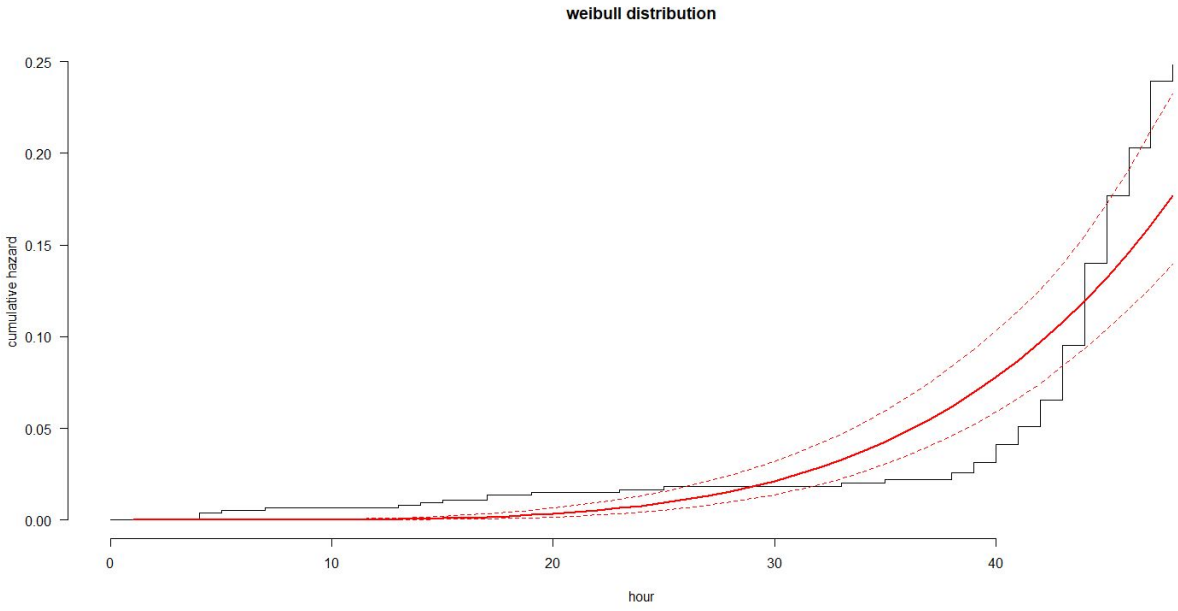
Through our analysis of the association between the motor and surge failure rates and different factors relating to the pump stations, we found that the most significant factor on these failures is whether the pump has a trashrack. For those pumps with a trashrack, they were 0.35 times less likely to have motor or surge failure than those without a trashrack. Additionally, we checked the Army Corps assumptions that pumps running for 12 consecutive hours were more likely to have motor or surge failure. We found this to be statistically true, as pumps running for 12 consecutive hours were 1.07 times more likely to have these failures compared those which did not run for 12 consecutive hours. With these findings, we would recommend identifying areas and/or pumps that are more susceptible to motor and surge failures and consider adding trashracks to the most frequently running pumps and investigate what features affect pump run time.

Appendix


exponential distribution
cumulative hazard
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
hour


log-logistic distribution
cumulative hazard
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
hour


lognormal distribution
cumulative hazard
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
hour


weibull distribution
cumulative hazard
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
hour