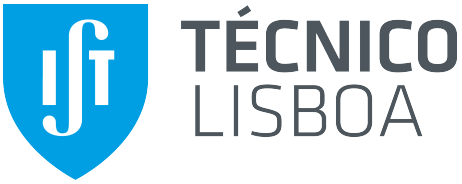

Log-Normal Regression for Survival Analysis in Primary Biliary Cirrhosis

Thomas Gaehtgens (ist186809)
José Pedro Guedes (ist1105774)

MSc in Applied Mathematics and Computation
May 2023

Abstract: This study investigates the impact of various risk factors on the survival time of patients diagnosed with primary biliary cirrhosis (PBC). The analysis employs log-normal regression to assess the statistical significance of risk factors. The results indicate that the treatment with D-penicillamine (DBCA) did not exhibit statistical significance, suggesting no significant difference in survival time between individuals receiving DBCA and those receiving the placebo. Furthermore, the presence of ascites was strongly associated with shorter survival times.

key-words: primary biliary cirrhosis, Survival Analysis, log-normal, regression models, Kaplan-Meier

1 Introduction

Primary biliary cirrhosis (PBC), also referred to as primary biliary cholangitis, is a chronic liver disease characterized by the progressive destruction of small bile ducts in the liver. The exact cause of PBC remains unknown; however, it is hypothesized to involve a combination of genetic, autoimmune, and environmental factors [5]. In this study, our objective is to investigate the association between survival time and a set of observed risk factors in PBC patients.

1.1 Dataset Overview

We analyzed a specific dataset obtained from the Mayo Clinic, which was designed to study the relationship between survival time and risk factors (Table 1) in patients. The dataset encompasses the period from January 1974 to May 1984 and includes information on 424 eligible patients. However, our analysis focused on the 312 patients who actively participated in the trials. The Mayo Clinic conducted a double-blinded randomized trial in primary biliary cirrhosis, comparing the drug D-penicillamine (DPCA) with a placebo [2]. The trial takes into account the patients age, the presence of asictes, a buildup of fluid in the abdomen of the patient as a result of a cirrhosis and the presence of hepatomegaly, an enlargement of the liver.

The dataset contained right-censored data, as not all patients experienced death during the study period. Consequently, we conducted separate analyses on two distinct subsets of the data: one comprising all patients who died and another encompassing the entire cohort. We compared the results obtained from the survival analysis model fitted to the dataset of deceased patients with those obtained from the model applied to the complete dataset. This approach allowed us to investigate the specific factors influencing survival in both subsets and examine the compatibility of the observed survival patterns with the overall population.

Table 1: Variable Descriptions

Variable Name	Description
ID	Case number
TIME	Number of days between registration and the earlier of death, liver transplantation, or study analysis time in July 1986
CEN	Status: 0=censored, 1=death
DRUG	Drug: 1=D-penicillamine, 2=placebo
AGE	Age in days
ASI	Presence of ascites: 0=no, 1=yes
HEP	Presence of hepatomegaly: 0=no, 1=yes
SERUM	Cholesterol levels in the blood serum of individuals in mg/dl

2 Exploratory Data Analysis

The exploratory data analysis began with the computation of summary statistics (Tables 2 and 3) for the complete dataset, followed by the calculation of Pearson correlation coefficients among the continuous variables in both datasets (Figure 1).

Table 2: Non-Binary Features of dataset with all patients.

Feature	Min	Max	Mean	St. Dev.
TIME	41	4556	2006	1123.28
AGE	9598	28650	18269	3864.81
SERUM	0.30	28.00	3.26	4.53

Table 3: Binary Features statistics of dataset with all patients

Feature	# 0	# 1	# 2
CEN	187	125	–
DRUG	–	158	254
ASI	288	24	–
HEP	152	160	–

We conducted an exploratory data analysis using visual representations to gain insights into the dataset. Histograms and bar charts were utilized to visualize the distribution of variables. Histograms provided information about the distributional characteristics of continuous variables,

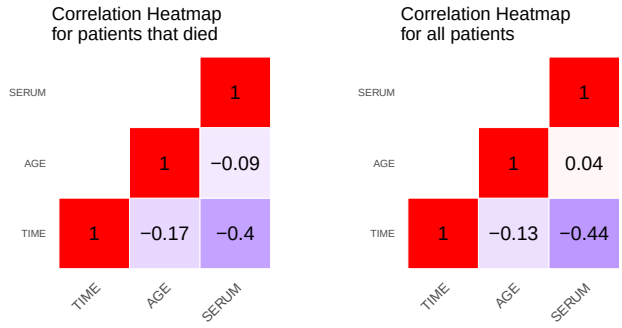

Figure 1: Heatmaps illustrating the Pearson correlation between continuos variables. The left heatmap displays the correlation in the dataset of patients who died, while the right heatmap showcases the correlation among all patients.

while bar charts displayed the distribution of binary variables.

Furthermore, box plots were employed to examine the distribution of time across different values of the binary variables. This facilitated initial observations regarding the influence of these binary variables on the variable of interest.

The histograms and bar plots show similar distributions across both datasets, indicating comparable distributions of the variables of interest. However, a notable difference is observed in the proportion of patients with hepatomegaly, which is significantly larger in the dataset with patients who died, suggesting it might be associated with patients' death.

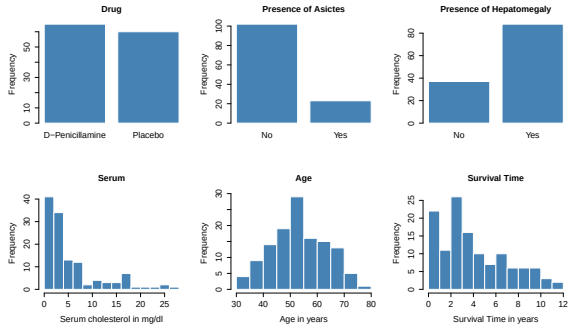

Figure 2: Distribution of variables regarding patients that died.

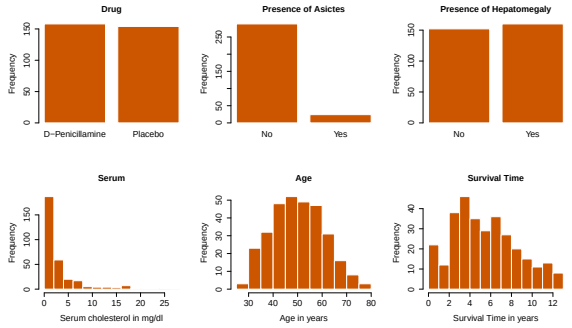

Figure 3: Distribution of variables regarding all patients.

The correlation between continuous variables appears to be consistent between the two datasets, suggesting a similar relationship between these variables. Moreover, the independent variables show little to no correlation with each other, suggesting their independence. Among the variables, SERUM exhibits the strongest correlation with survival time. The negative correlation suggests that higher levels of serum or older age are associated with shorter survival times.

When considering the survival time associated with the binary variables, the distributions are again similar in both datasets. It appears that the administration of DPCA does not have a significant impact on the survival time of patients. However, the presence of ascites and hepatomegaly seems to be associated with shorter survival times. Patients with ascites show the shortest survival time.

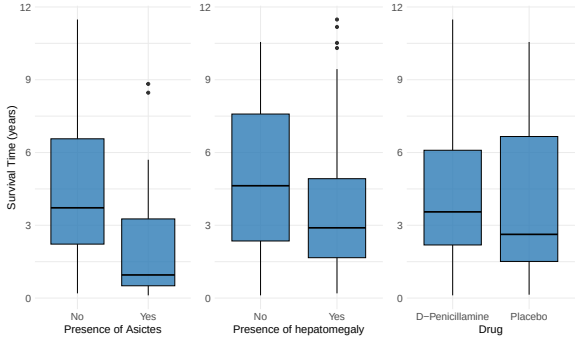

Figure 4: Boxplot of the variables ASI, HEP and DRUG regarding patients that died.

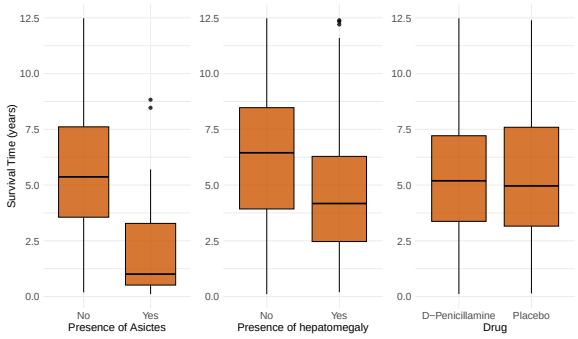

Figure 5: Boxplot of the variables ASI, HEP and DRUG regarding all patients.

2.1 Survival Function

In addition to the descriptive visualizations, the non-parametric Kaplan-Meier estimator [1] was employed to analyze survival time through the survival curve. This plot is useful in providing a visual representation of the survival of the study participants and allows the identification of potential differences in survival between the different groups.

The similarity observed in the Kaplan-Meier curves between the DPCA and placebo groups in both datasets, along with the L-shaped pattern in the subgroup with complete data, provides insights into the effectiveness of DPCA and the influence of different data sets on survival analysis.

The comparable curves between DPCA and placebo (Figures 6 and 7) indicate that DPCA did not yield a significant improvement in survival outcomes compared to the placebo, as previously

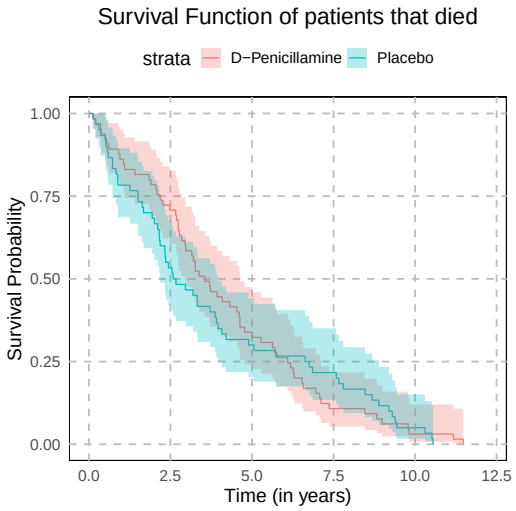

Figure 6: Survival function of the patients that died (complete data) using the Kaplan-Meier estimator.

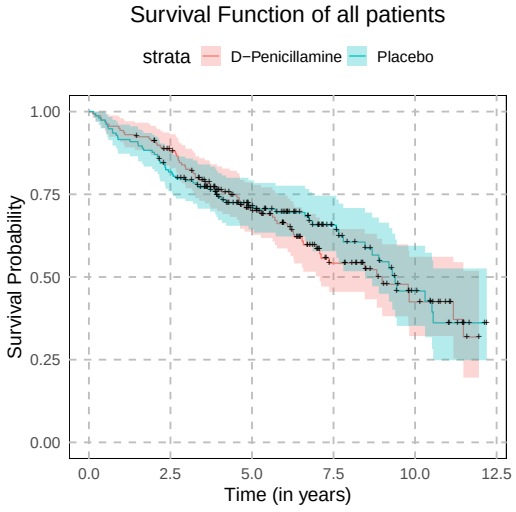

Figure 7: Survival function of all patients using the Kaplan-Meier estimator. + signs signal censored data.

analyzed. This consistency across the censored data and complete data sets suggests that the presence of censored data did not substantially impact the overall survival trends.

However, the L-shaped pattern observed in the subgroup with complete data suggests that the initial decrease in survival probabilities is more pronounced, followed by a stabilization over time, suggesting the presence of underlying factors or characteristics within the subgroup that may have a distinct influence on survival outcomes compared to the overall population.

Significant differences are observed in the survival curves between patients with ascites and hepatomegaly (Figures 11 and 10). In both datasets, patients with ascites exhibit shorter survival times. However, the impact of hepatomegaly appears to be less pronounced in the dataset of patients who died, when compared to the dataset of all patients.

We present a mortality table (Table 4) regarding the age of PBC patients. The table provides information on the occurrence of death and individuals at risk at different time periods. Mortality rates seem to vary across age groups, with an increasing trend as age advances. For instance, the [20, 30) age group records no deaths, while mortality rates rise from [30, 40) to [70, 80). These findings suggest a risk of mortality with increasing age.

3 Methods

3.1 Log-normal Distribution

Following survival analysis methods [6], we propose a parametric log-normal model where the logarithm of the survival times are distributed according to a Normal distribution $\log(T) \sim \mathrm{N}(\mu, \sigma^2)$ or $T \sim \text{log-normal}(\mu, \sigma^2)$. The probability density function of log-normal distribution [8] is defined as:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$$

And the survival and hazard functions are defined as:

$$S_T(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$$

$$\lambda_T(t) = \frac{1}{\sigma t} \times \frac{\phi(\frac{\ln t-\mu}{\sigma})}{1 - \Phi(\frac{\ln t-\mu}{\sigma})}$$

Table 4: Mortality Table

Age Group	Deaths	Exposures	Mortality Rate	Survival Rate
[20,30)	0	2	0.00	1.00
[30,40)	12	42	0.22	0.78
[40,50)	30	63	0.32	0.68
[50,60)	46	54	0.46	0.54
[60,70)	28	21	0.57	0.43
[70,80)	9	5	0.64	0.36

Where Φ and ϕ are the cumulative and probability density function of the standard Gaussian.

The survival time in the logarithm side is expressed in terms of linear combination of p covariates and its associated coefficient plus a measurement error, as:

$$\log(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. The estimation of parameter values β_0, β_1, ... , β_p is computed through maximum likelihood estimation. If only considering the complete data, the likelihood function is given by:

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{x_i \sigma_i \sqrt{2\pi}} e^{-\frac{(\log x_i - \mu_i)^2}{2\sigma_i^2}}$$

We want to find the parameters that maximize the likelihood function, which is equivalent to minimizing the negative log-likelihood.

$$\hat{\beta} = \arg\max_{\beta} \{L(\beta)\} = \arg\min_{\beta} \{-\log L(\beta)\}$$

3.2 Censored Data

It is crucial to consider the presence of censored data when analyzing the dataset containing all patients, as previously mentioned. The inclusion of censored observations requires a correction to the likelihood that accounts for this phenomenon. The approach to calculate the likelihood depends on the specific type of censoring present in the data. In this case, we have right-censoring and, therefore, some observations for which the exact time of death is unknown, but we know that it has occurred after a certain time C_i.

The likelihood function for right-censored data can be calculated using the general formula for the likelihood with right-censored data:

$$L(\theta|\underline{t}, \underline{\gamma}) = \prod_{i=1}^{n} f(t_i|\theta)^{\gamma_i} \; S(t_i|\theta)^{1-\gamma_i}$$

where $f(t_i)^{\gamma_i} S(t_i)^{1-\gamma_i} = \lambda(t_i)^{\gamma_i} S(t_i)$ dependent on an array of parameters θ, and

$$\gamma_i = \begin{cases} 1, & X_i \leq C_i \\ 0, & X_i > C_i \end{cases}$$

3.3 Log-Normal Model Fitting

The log-normal model was fitted on each of the datasets using a stepwise iterative selection process [7] [3]. In each step of the stepwise iterative selection process, a covariate is removed based on the statistical significance of its coefficient through Wald's test. The process continues until no further significant variables are found to be removed.

Three models were fitted to both datasets. The saturated model (i.e. including all covariates), which we refer as F, is given by:

$$\log(T) = \beta_0 + \beta_1 X_{DRUG} + \beta_2 X_{AGE} +$$

$$+ \beta_3 X_{HEP} + \beta_4 X_{ASI} + \beta_5 X_{SERUM}$$

For the reduced models, we chose the covariates that remained after applying the stepwise iterative selection process on each of the datasets.

For the complete data dataset, R_{com}, we have:

$$\log(T) = \beta_0 + \beta_1 X_{HEP} + \beta_2 X_{ASI} + \beta_3 X_{SERUM}$$

For the dataset regarding all patients:

$$\log(T) = \beta_0 + \beta_1 X_{AGE} +$$

$$+ \beta_2 X_{HEP} + \beta_3 X_{ASI} + \beta_4 X_{SERUM}$$

3.4 Goodness of fit

The goodness of the fit was measured by the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), with k being the number of parameters in the model, n the size of the sample and L the likelihood [3]:

$$AIC = 2k - 2\log L(\hat{\beta})$$

$$BIC = k\log n - 2\log L(\hat{\beta})$$

These statistical measures are used to assess the quality of models, regarding model selection. Both criteria aim to balance the trade-off between model fit and complexity. When comparing multiple models, the model with the lowest AIC and BIC is preferred as it indicates a better balance between goodness of fit and model complexity.

Q-Q (quantile-quantile) plots were used to assess the distributional assumptions of the residuals. The Q-Q plot compares the observed quantiles of the residuals against the expected quantiles of a chosen theoretical distribution, in this case the normal distribution. This graphical technique helps to identify departures from the assumed distribution and provides insights into the goodness-of-fit of the model. If the residuals follow a perfect normal distribution, the points on the plot will form a straight line.

4 Results

Tables 5 and 6 present the results of the stepwise iterative selection process, including the corresponding p-values and the covariates that were removed at each step.

Contrary to the initial expectations, the analysis of the dataset of patients who died revealed that *HEP* was the first covariate to be removed, rather than *DRUG*. This suggests that the association between *HEP* and the outcome variable may be weaker or that its inclusion in the model contributes less to improving the overall fit compared to other covariates.

Furthermore, the remaining covariates in the dataset with censored data have the largest p-value of 0.00078 for *HEP*. In the dataset with complete data, the largest p-value among the remaining covariates is 0.000026 for *AGE*. Both these values are smaller than the standard p-value of 0.05, suggesting all the remaining covariates are statistically significant in light of Wald's test.

Table 5: Variables removed in stepwise iterative selection process on **dataset with all patients**.

Step	Removed covariate	Hypothesis $\beta_j = 0$		
		deviance	d.f.	p-value
1	-	-	-	-
2	DRUG	0.01	1	0.92246

Table 6: Variables removed in stepwise iterative selection process on **dataset with patients that died**.

Step	Removed covariate	Hypothesis $\beta_j = 0$		
		deviance	d.f.	p-value
1	-	-	-	-
2	HEP	0.28	1	0.59488
3	DRUG	0.83	1	0.36064

4.1 Model Diagnostic & Selection

The goodness of fit results are presented in Tables 7 and 8. As anticipated, the models that demonstrate the best fit for each dataset, indicated by lower BIC and AIC values, are the models derived from the stepwise iterative selection process specific to each dataset.

Table 7: Goodness of fit measures - complete data

Model	AIC	BIC
F	577.81	597.61
R_{com}	**574.93**	**589.07**
R_{cen}	576.62	593.59

Table 8: Goodness of fit measures - all data

Model	AIC	BIC
F	775.31	801.52
R_{com}	782.91	801.63
R_{cen}	**773.32**	**795.78**

In comparing the coefficients between the complete data and all data groups in Tables 10 and 11, there are variations worth considering.

The intercept term exhibits a higher value in the all data group, indicating an increased baseline risk associated with censored data. This suggests that the presence of censored observations has an impact on the overall risk estimation.

While the coefficient for *AGE* remains consistent across both datasets, the coefficients for *ASI* and *SERUM* show significant variations. These differences suggest that the presence of ascites (*ASI*) and the level of serum (*SERUM*) have differing effects on survival time between the two groups. The larger magnitude of the coefficients in the all data group implies a stronger association between these variables and shorter survival time, which was also observed on the Kaplan-Meier curves (Figures 10 and 11).

These observed variations in coefficient magnitudes indicate the importance of considering the specific dataset when interpreting the significance of each covariate. It highlights the need to account for the presence of censored data and its potential impact on the estimated risks and associations with survival time.

Table 9: Coefficients of Model F

Variable	complete data		all data	
	β	st. error	β	st. error
(intercept)	2.635	0.459	4.436	0.420
DRUG	-0.137	0.152	-0.014	0.145
AGE	-0.019	0.008	-0.030	0.007
HEP	-0.090	0.169	-0.513	0.153
ASI	-0.802	0.208	-1.098	0.244
SERUM	-0.057	0.013	-0.110	0.015

Table 10: Coefficients of Model R_{com}

Variable	complete data		all data	
	β	st. error	β	st. error
(intercept)	2.470	0.427	4.205	0.397
AGE	-0.018	0.008	-0.030	0.007
ASI	-0.795	0.209	-1.142	0.248
SERUM	-0.060	0.013	-0.124	0.015

Table 11: Coefficients of Model R_{cen}

Variable	complete data		all data	
	β	st. error	β	st. error
(intercept)	2.552	0.452	4.426	0.407
AGE	-0.019	0.008	-0.030	0.007
HEP	-0.094	0.170	-0.514	0.153
ASI	-0.790	0.209	-1.097	0.244
SERUM	-0.058	0.013	-0.110	0.015

4.2 Residual Analysis

The Q-Q plot for the dataset containing only complete data and the Q-Q plot for the dataset with all the patient's data (Figures 9 and 8 respectively) both exhibit left skewness. The Q-Q plot for the dataset with censored data displays a more pronounced left skewness compared to the plot for the complete data.

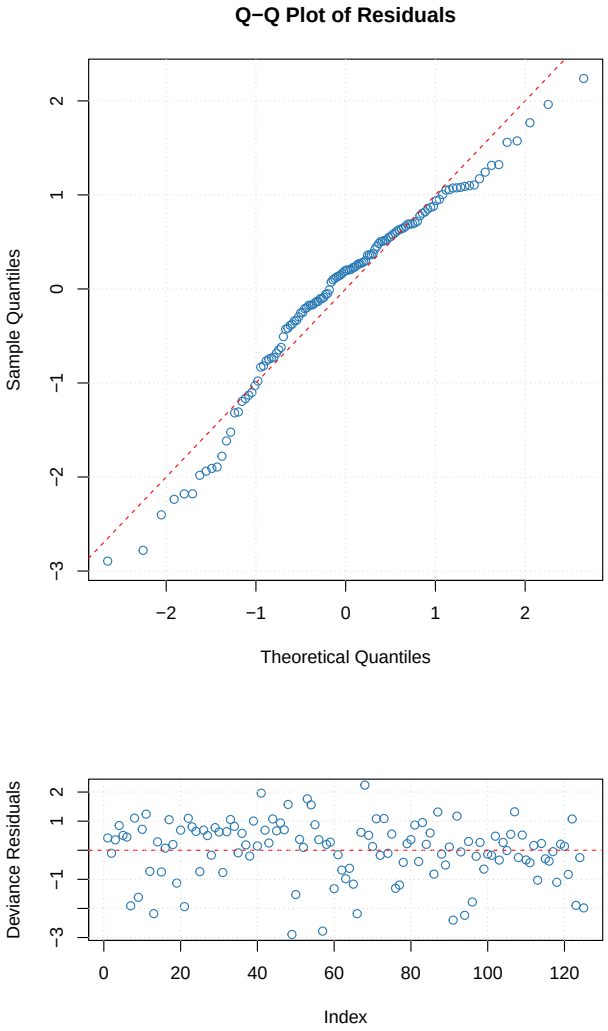

Figure 8: Q-Q plot of residuals and deviance residuals regarding all data.

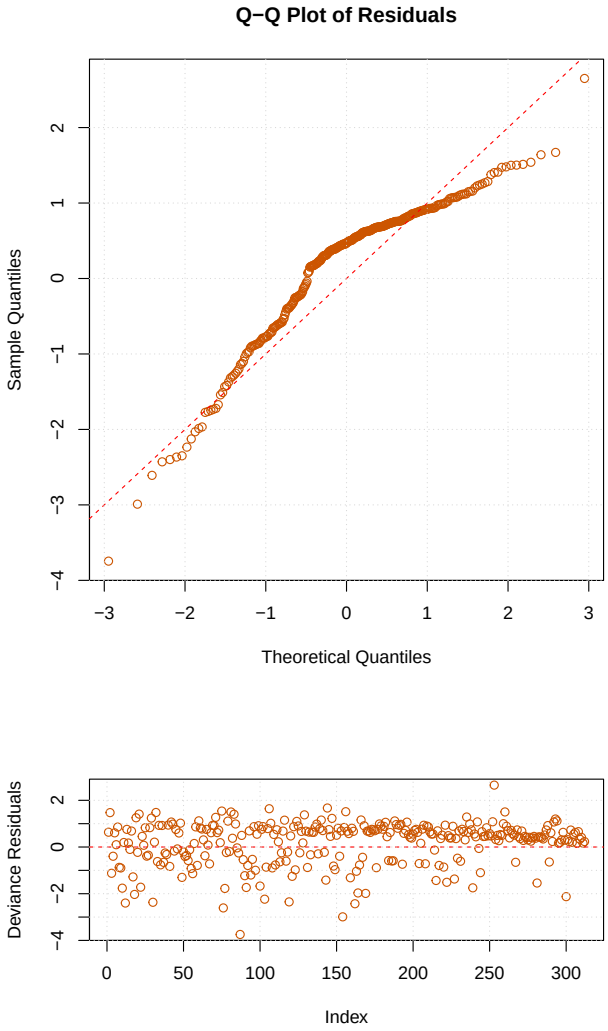

Figure 9: Q-Q plot of residuals and deviance residuals regarding patients that died.

This increased skewness in the Q-Q plot for the censored data suggests that the presence of right-censored observations contributes to the deviation from a normal distribution assumption. This result is to somewhat expected as right-censored data can introduce uncertainty in estimating extreme values on the upper end of the distribution, leading to a more pronounced left skew in the residuals plot.

When examining the deviance residuals, we observe a different pattern between the two datasets. The dataset with complete data shows a distribution of residuals that appears more consistent with a normal distribution. On the other hand, the dataset with censored data displays residuals that seem to follow a trend as the index increases. This concentration of residuals indicates potential departures from normality and suggests that the model's estimates are less accurate for certain observations in the censored dataset.

There results are indications of potential departures from the normal distribution assumption, suggesting limitations to the log-normal regression model to the data, especially when considering censored data.

5 Conclusion & Future Work

We conclude that treatment with DBCA was not statistically significant, implying no significant difference in survival time between individuals receiving DBCA and those receiving the placebo.

Furthermore, we observed variations in coefficient magnitudes between the patients that died and all the patients, suggesting potential differences in the impact of covariates on survival time. We found that Presence of hepatomegaly wasn't statistically significant for survival time of patients that died, however it was a significant risk factor when considering all patients survival time.

Both datasets exhibited left skewness in the QQ plots, indicating deviations from the assumption of normality for the residuals of the log-normal regression model.

To validate and enhance the reliability of our findings, we suggest the assessment of the predictive performance of the models using new data. Additionally, future research should consider alternative statistical methods, such as Weibull or gamma regressions, to potentially improve model fit and address potential issues associated with residual skewness.

References

[1] T. Fleming and D. Harrington. Counting processes and survival analysis. wiley. *Biometrical Journal - BIOM J*, 34:674–674, 01 1992.

[2] Patricia M. Grambsch, E. Rolland Dickson, Russell H. Wiesner, and Alice Langworthy. Application of the mayo primary biliary cirrhosis survival model to mayo liver transplant patients. *Mayo Clinic Proceedings*, 64(6):699–704, 1989.

[3] G. Loiola da Silva. Notas de bioestatística. *Dep. Matemática - Instituto Superior Técnico*, 05 2023.

[4] Carlo Selmi, Christopher Bowlus, M Gershwin, and Ross Coppel. Primary biliary cirrhosis. *Lancet*, 377:1600–9, 05 2011.

[5] Jayant Talwalkar and Keith Lindor. Primary biliary cirrhosis. *Lancet*, 362:53–61, 08 2003.

[6] Terry M Therneau. *A Package for Survival Analysis in R*, 2023. R package version 3.5-5.

[7] Maria A Amaral Turkman and Giovani Loiola Silva. Modelos lineares generalizados-da teoria à prática. *Sociedade Portuguesa de Estatística, Lisboa*, 2000.

[8] Shuyi Wang and Wenhao Gui. Corrected maximum likelihood estimations of the log-normal distribution parameters. *Symmetry*, 12(6), 2020.

6 Appendix

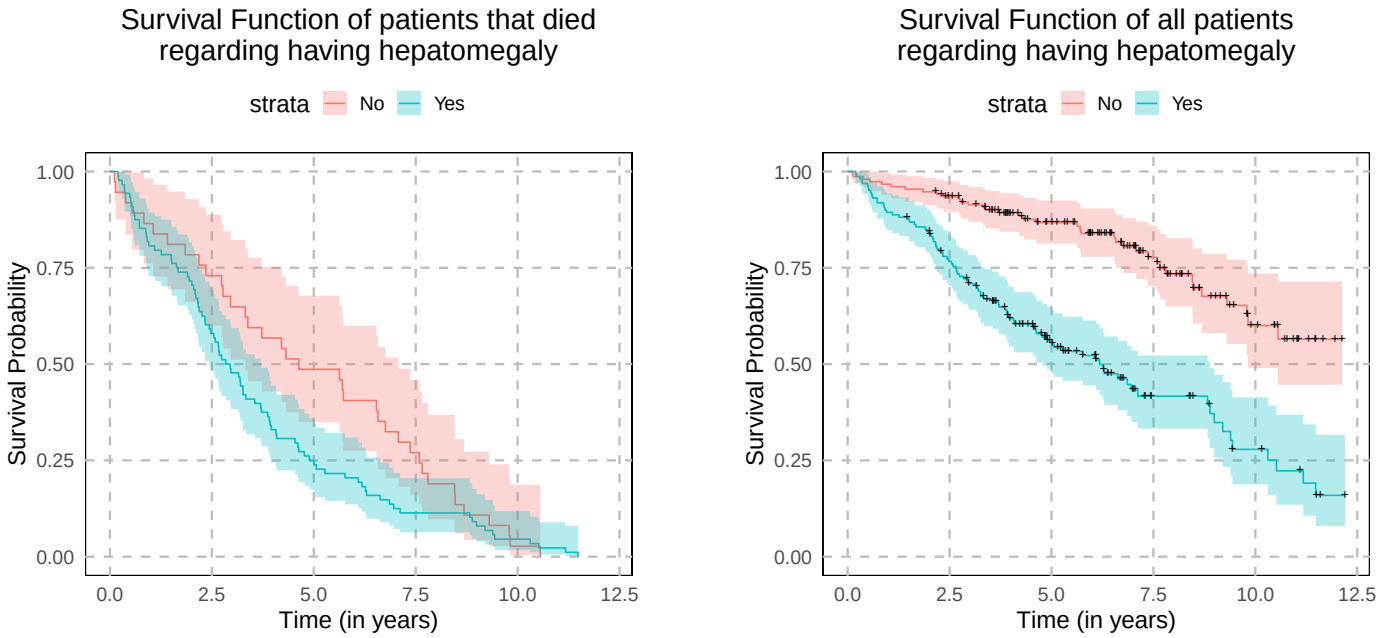

Figure 10: Kaplan-Meier plots for Hepatomegaly data.

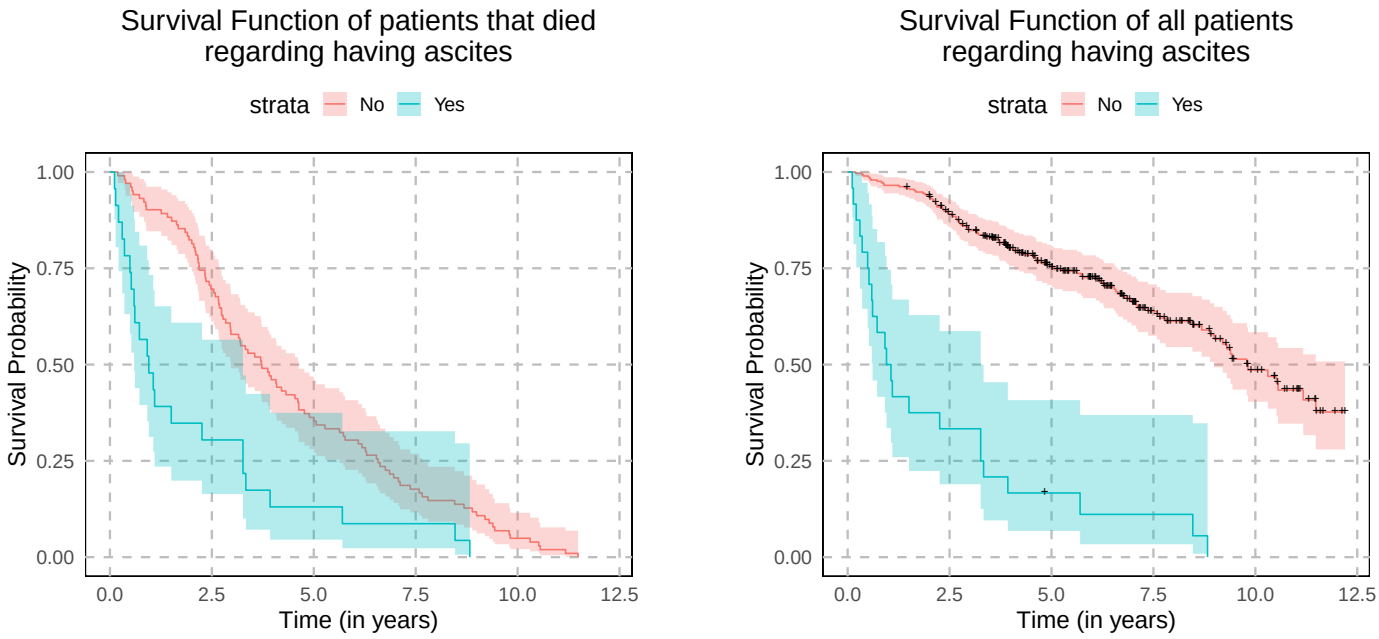

Figure 11: Kaplan-Meier plots for Ascites data.