A Study on the Need for Using Special Treatment for Colon Cancer

How different indicators reflect the value of Levamisole+5-FU

STAT 362 Group Project Report
April 16, 2023
Instructor: Brian Ling

Group Member:

Name: Qiyue Zhang
Student ID: 20199697
Email: 19qz20@queensu.ca
Contributions: Introduction, Description of the dataset, References

Name: Sizhe Yang
Student ID: 20199745
Email: 19sy24@queensu.ca
Contribution: Method & Code of Kaplan-Meier plot, Cox Regression & Logistic regression model

Name: Weiran Xu
Student ID: 20200716
Email: 19wx11@queensu.ca
Contribution: Interpretation of Kaplan-Meier plot and Analysis in Logistic Regression; Method & Codes of barplots

Name: Jialiang Tao
Student ID: 20201886
Email: 19jt23@queensu.ca
Contribution: Method & Codes of Decision Tree, Box and Bar plots, Tables and Interpretation of Cox Proportional Hazard Regression

Name: Haoyang Xu
Student ID: 20200039
Email: 19hx11@queensu.ca
Contribution: Conclusion

1. Introduction

Our group of five students, who are interested in the field of life science and biostatistics, have embarked on a study of colon cancer. The impetus for our research came from a startling prediction we found on Science Daily, which suggested that colorectal cancer could become the leading cause of cancer death in individuals aged 20-49 in the U.S. by the year 2030 (ScienceDaily, 2023). This prediction has inspired us to delve into the realm of colon cancer, which is one of the most prevalent types of cancer.

Our research revealed that colon cancer has an extremely high mortality rate, with over 17,000 deaths from rectal cancer and approximately 30,000 new cases each year in the UK alone, making it the second leading cause of cancer death after lung cancer (A Leslie, 2002). This information piqued our interest in studying colon cancer and the available therapies to determine whether they can alleviate or cure colon cancer patients.

While there are already thousands of studies on colon cancer, each approaching the issue from a different angle, our focus is on examining the effectiveness of existing therapies. Our goal is to compare and validate different treatments to determine which ones can prolong the survival time of patients and provide effective treatment.

The objective of our study is to address the problem of colon cancer mortality and investigate the impact of certain factors on its treatment. Specifically, we are interested in understanding how a particular factor affects the mortality rate of colon cancer patients. To achieve this, we will be analyzing the colon dataset in R, which provides us with a wealth of information about the disease and its treatments. Our research aims to contribute to the larger body of knowledge in this field, ultimately leading to better outcomes for colon cancer patients. By focusing on the treatment of colon cancer and nodes, we hope to provide valuable insights that can be used to improve the lives of those affected by this disease.

2. Description of dataset

Our study uses a commonly used dataset in survival analysis, which includes information on patients with advanced colorectal cancer. The dataset, which can be accessed directly in R, was initially collected in 1989 and updated until 1994.

The colon cancer dataset contains 1858 observations and 16 variables, including time from surgery to death or end of the study, patient status, age, sex, obstructing tumor, perforated tumor, tumor adherence, and the number of cancerous lymph nodes. These variables include both binary and numeric data, with binary variables such as sex, obstruct, perfor, and adhere indicating the presence or absence of certain conditions. Numeric variables such as time, age, and nodes provide quantitative data. The "status" variable is a binary variable indicating whether or not the patient died during the study. (See Table 1) According to the status column in the colon dataset, there were 920 patients recorded as deceased by the end of the study, marked with a status code of 1. This translates to an event rate of 49.515%, which represents the percentage of patients who passed away during the study period.

Notably, the colon cancer dataset has a relatively low number of missing values, with only 82 observations marked as NA. However, it's worth mentioning that all missing values occur exclusively in the "nodes" and "differ" columns. While this may have a minor impact on our data analysis, we will carefully consider the potential effects of these missing values on our overall findings.(figure1)

3. Method

To address the research questions and test the null hypotheses, we employed a combination of statistical analysis and machine learning methods. The chosen methods include logistic regression, Cox proportional hazard regression model, Kaplan-Meier plot. These methods were selected due to their suitability for the type of data and research questions, as well as their ability to handle binary outcomes and time-to-event data.

3.1 Hypotheses

According to the our objectives, we formulated two null hypotheses. The first null hypothesis is that patients who receive combination chemotherapy do not have a longer time until recurrence and death compared to those who receive observation or Levamisole alone. The second null hypothesis is that the number of nodes is not a significant predictor of time until recurrence or death.

3.2 Statistical Analysis

To investigate the relationship between treatment groups and survival outcomes, we first visualize the survival probability in different groups using the Kaplan-Meier plot. This non-parametric method estimates the survival probability in a population and is useful for accommodating censored data in survival analysis (Bollschweiler, 2003). The plot allows us to compare the survival probability between treatment groups and identify any differences in dataset. Next, we want to investigate the effect of each predictor variable on the outcome of interest, which is binary for logistic regression and time-to-event for Cox proportional hazards regression. We use logistic regression for modeling binary outcomes and Cox proportional hazards regression for modeling time-to-event data. Both methods have distinct assumptions about the data, such as the linearity of the relationship between predictor variables and outcomes and the constancy of hazard ratios over time.

In a Cox proportional hazards regression model, we use the exponentiated coefficients (exp(coef)) to represent the hazard ratio for each predictor variable. A hazard ratio greater than 1 indicates that the group with x has a higher hazard rate (i.e., worse survival) than the group with a value of 1, while a hazard ratio less than 1 indicates that the group with x has a lower hazard rate (i.e., better survival) than the group with a value of 1 (Abd ElHafeez, 2021).

To assess the proportional hazards assumption in the Cox model, we employ the Schoenfeld test. The Schoenfeld test is a diagnostic tool used to check whether the hazard ratio for each predictor variable is constant over time (Abeysekera, 2009). A significant test result indicates

a violation of the proportional hazards assumption, which may require adjustments or alternative modeling approaches.

3.3 Potential Problems

Our dataset may contain issues such as missing data, selection bias, and violations of model assumptions. One potential problem with our analysis is the presence of missing data. This can lead to biased estimates and reduced power in our analyses. To handle missing data, we propose using multiple imputation techniques. Another potential problem is selection bias, which can occur when there are systematic differences between the groups being compared. In our study, patients who received combination chemotherapy may have been different from those who received observation or Levamisole alone in terms of factors that were not measured, such as socioeconomic status. This can affect the generalizability of our findings, which means it may not represent the broader population of patients with colon cancer. To minimize selection bias, we can employ propensity score matching or stratification techniques to balance the groups being compared. Furthermore, we will carefully assess the model assumptions (for example, the Cox model assumes that the hazard ratio is constant over time, which may not hold in some cases) and address any violations that may lead to biased estimates or incorrect conclusions.

3.4 Evaluation and Validation

To assess the performance of our logistic regression model and evaluate our results, we will use AUC (Area Under the ROC Curve) as a performance metric. AUC measures the model's ability to distinguish between the two classes (positive and negative), with a value of 1.0 indicating a perfect classifier and a value of 0.5 indicating a random classifier (Stephen Allwright, 2022). ACC(accuracy) measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. While AUC is a measure of how well the model can distinguish between positive and negative classes, ACC is a measure of how often the model makes correct predictions overall. AUC is generally considered to be a better metric than ACC when dealing with imbalanced datasets because it takes into account both true positives and false positives (Stephen Allwright, 2022).
For the Cox proportional hazards regression model, we will use the concordance index (C-index) to evaluate the model's performance. The C-index measures the model's ability to correctly rank the order of survival times for pairs of individuals, one of whom has an event and the other is censored. “The C-index ranges from 0 to 1, with 0.5 indicating no better than random chance and 1.0 indicating perfect prediction” (Zeng, 2021)

4. Result

Kaplan-Meier Plot and Risk Table between Survival Probability and Treatments: The plot reflects the effect of different treatments on survival time of patients. Each colored line represents specific therapy. In the plot, Lev+5FU is the most apparent therapy which extends the life expectancies of humans compared to others. The number of patients in the specific time slot could also prove that. On the contrary, Lev therapy does not work on patients and

even reduces the survival probability. In all, we could find the positive association between the survival time and the Lev+5FU treatment on patients.(figure9)

Kaplan-Meier Plot and Risk Table between Survival Probability and Appearance of More than 4 Positive Lymph Nodes: The plot indicates the survival probability based on whether patients have more than 4 positive lymph nodes or not. Obviously, patients who have less than 4 nodes have much higher survival probability in all time stages, which is proved in the tremendous difference of patients in the risk table through all time periods. In conclusion, there is a positive association between the truth of more than 4 lymph nodes and hazard rate. (figure10)
Logistic Regression: The estimated coefficients of intercept, Lev, Lev+5FU and nodes are 0.47234, 0.13067 ,0.50326 and -0.19658. This shows a positive association between the treatment Lev and lev+5FU. However, the positive effect is not apparent since the coefficients are small. Also, all factors are significant with P value less than 0.05 except Lev treatment and the number of nodes is more significant than treatments on the logistic regression. This reflects that if we set a null hypothesis that Lev treatment does not work on patients, then we cannot reject the null hypothesis.(figure2)

Logistic Regression Residual Plot: The plot shows the relationship between residuals, which is the difference between observed values and predicted values of survival time, and the predicted values of the logistic model based on the predictors rx and nodes. The points show apparent patterns and extreme outliers, indicating the model does not fit the data well and capture the relationship. More predictors need to be considered for improving the accuracy. (figure4)
Logistic Regression QQ Plot: The QQ plot compares the probability distribution by plotting the quantiles of residuals of difference between observed and predicted values, and quantiles of predicted values of the logistic model based on two predictors. Lines in the plot deviate from the straight reference line with heavy tails and outliers, representing the trait of nonnormality. From Shpiro-Wilk test, P value is 2.2*10^-6 which is much less than 0.05 indicating the apparent normality. We will use data transformation methodology to improve the accuracy of the model as we can next.(figure5)

Logistic Regression Scale Plot: For all values in the y-axis, we replace all of them with their root values. However, the values still present non-random scattering and it indicates variance of residuals is not constant across predictors we use in the logistic regression.(figure6)

For accuracy (ACC), a value of 0.636824 indicates that the model correctly predicted the outcome for approximately 64% of the cases. While this may be better than chance, it may not be sufficient for many practical applications.

For AUC, a value of 0.665086 suggests that the model's ability to distinguish between the two classes (positive and negative) is not very good. An AUC of 0.5 indicates a random classifier, while an AUC of 1.0 indicates a perfect classifier. An AUC value between 0.6 and 0.7 is generally considered to be a poor classifier.

Boxplot of survival time and treatments: The boxplot combination shows the distribution of survival time of patients based on three therapies. Obviously, Lev+5FU therapy has the highest median and Q3 value compared to others, indicating the improvement of survival time in mid and early periods. Besides, the variability of other specific values in the plot is small which means no treatment works in the final stage of patients and Lev therapy does not work effectively.(figure7)

Cox-Proportional Hazard Regression: In the first regression (Nodes), the study investigates the relationship between the number of positive lymph nodes (Nodes4 = 1) and the hazard rate in patients with colon cancer. The positive coefficient indicates that patients with more than 4 positive lymph nodes have a significantly higher hazard rate than those with 4 or fewer nodes. Specifically, the hazard rate is 2.474 times higher in patients with more than 4 positive lymph nodes. The likelihood ratio, Wald, and score tests confirm the significance of this relationship, indicating that the model with the Nodes4 variable is a better fit than the null model without any predictor. The concordance index suggests moderate predictive accuracy for this model.(See figure12)

In the second regression (rx, the treatment), the study examines the effect of two treatments, Lev and Lev+5FU, on the hazard rate compared to the reference group (no treatment). The negative coefficients indicate that both treatments have a reduction in the hazard rate, but only Lev+5FU shows a significant effect. Specifically, the hazard rate is 35.66% lower for patients receiving Lev+5FU compared to the reference group, while the reduction in hazard rate for Lev is not statistically significant. The likelihood ratio, Wald, and score tests confirm the significance of the relationship between the treatment variable (rx) and hazard rate, indicating that the model with the rx variable is a better fit than the null model without any predictor. The concordance index suggests modest predictive accuracy for this model, which also indicates the regression might be overfitted. (See figure11)

In the Schenofeld test, the likelihood ratio test for the rx predictor variable has a chi-squared statistic of 0.648, with 2 degrees of freedom and a p-value of 0.72. This indicates that there is no significant evidence of an association between rx and the outcome (survival status).

There is a significant association between the node4 predictor variable and the outcome (survival status), with a chi-squared statistic of 11.2, 1 degree of freedom, and a p-value of 0.00082. This suggests that node4 is a significant predictor of survival.

5. Conclusion & Discussion

In conclusion, based on our hypotheses and the analysis using the methods, it can be concluded that the combination chemotherapy (Levamisole+5-FU) has a significant effect on time until recurrence or death, and the number of nodes is a significant predictor of time until

recurrence or death. Patients receiving the combination chemotherapy have a higher survival probability compared to those receiving Levamisole alone or observation, and patients with fewer nodes have a higher survival probability compared to those with more nodes. Also (Levamisole+5-FU) can effectively limit the metastasis of cancer cells.

There are some limitations of our project, for example: we cannot obtain median survival time and data is small. While our hypotheses and machine learning methods can provide valuable insights into the factors affecting time until recurrence or death in colon cancer patients, they have certain limitations that should be taken into account when interpreting the results. It is important to consider other factors that may influence the outcome and to use multiple methods to validate the findings.

The summary can be useful in guiding researchers or analysts who are interested in investigating the impact of combination chemotherapy and the number of nodes on the time until recurrence or death in patients with colon cancer.

6. References

Abd ElHafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021, November 30). Methods to analyze time-to-event data: The Cox Regression Analysis. *Oxidative Medicine and Cellular Longevity*. Retrieved April 10, 2023, from https://www.hindawi.com/journals/omcl/2021/1302811/

Abeysekera, W., & Sooriyarachchi, M. (2009). Use of Schoenfeld's global test to test the proportional hazards assumption in the Cox proportional hazards model: an application to a clinical study. *Journal of the National Science Foundation of Sri Lanka*, *37*(1), 41–51. DOI: http://doi.org/10.4038/jnsfsr.v37i1.456

A Leslie, R J C Steele. (2002, August 1). Management of colorectal cancer. *Postgraduate Medical Journal.* Volume 78, Issue 922, August 2002, Pages 473–478. Retrieved April 8, 2023 from https://doi.org/10.1136/pmj.78.922.473

Bollschweiler E. (2003). Benefits and limitations of Kaplan-Meier calculations of survival chance in cancer surgery. *Langenbeck's archives of surgery*, 388(4), 239–244. https://doi.org/10.1007/s00423-003-0410-6

Dana-Farber Cancer Institute. (2023, March 16). Researchers chart a course for understanding, preventing, and treating young-onset colorectal cancer. *ScienceDaily*. Retrieved April 8, 2023 from www.sciencedaily.com/releases/2023/03/230316140939.htm

J. Salazar-Roa. (2022). Rewiring Cancer Metabolism. *Harvard Medical School News*. Retrieved April 8, 2023 from https://hms.harvard.edu/news/rewiring-cancer-metabolism

Stephen Allwright. (2022, August 23). AUC vs accuracy, which is the best metric? *Stephen Allwright*. Retrieved April 19, 2023, from https://stephenallwright.com/auc-vs-accuracy/

Wayne, L. M. (n.d.). Cox Proportional Hazards Regression Analysis. Cox proportional hazards regression analysis. Retrieved April 10, 2023, from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html

Zeng, Q., Li, J., Tan, F. et al. Development and Validation of a Nomogram Prognostic Model for Resected Limited-Stage Small Cell Lung Cancer Patients. *Ann Surg Oncol* 28, 4893–4904 (2021). https://doi.org/10.1245/s10434-020-09552-w

* Graphs and Plots

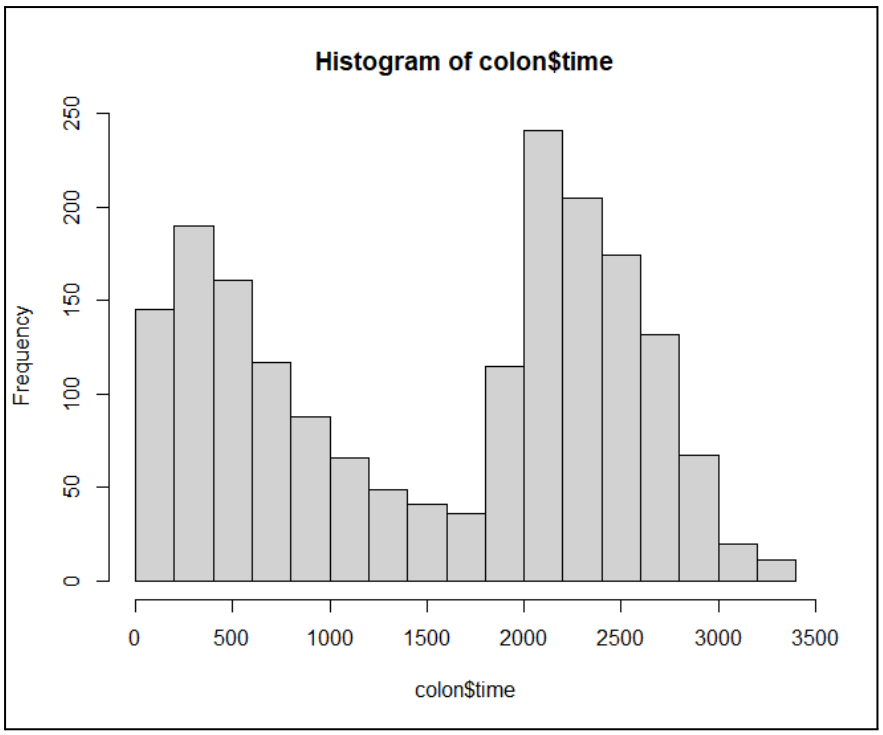
	Overall (N=1858)
Study Identifier	
Mean (SD)	1.00 (0)
Median [Min, Max]	1.00 [1.00, 1.00]
Treatment	
Lev	620 (33.4%)
Lev+5FU	608 (32.7%)
Obs	630 (33.9%)
Gender	
Mean (SD)	0.521 (0.500)
Median [Min, Max]	1.00 [0, 1.00]
Age (years)	
Mean (SD)	59.8 (11.9)
Median [Min, Max]	61.0 [18.0, 85.0]
Obstruction	
Mean (SD)	0.194 (0.395)
Median [Min, Max]	0 [0, 1.00]
Perforation	
Mean (SD)	0.0291 (0.168)
Median [Min, Max]	0 [0, 1.00]
Adherence	
Mean (SD)	0.145 (0.353)
Median [Min, Max]	0 [0, 1.00]
Number of Lymph Nodes	
Mean (SD)	3.66 (3.57)
Median [Min, Max]	2.00 [0, 33.0]
Missing	36 (1.9%)
Days Until Event/Censoring (days)	
Mean (SD)	1540 (947)
Median [Min, Max]	1860 [8.00, 3330]
Censoring Status	
Mean (SD)	0.495 (0.500)
Median [Min, Max]	0 [0, 1.00]
Tumor Differentiation	
Mean (SD)	2.06 (0.514)
Median [Min, Max]	2.00 [1.00, 3.00]
Missing	46 (2.5%)
Extent of Local Spread	
Mean (SD)	2.89 (0.488)
Median [Min, Max]	3.00 [1.00, 4.00]
Time from Surgery to Registration	
Mean (SD)	0.266 (0.442)
Median [Min, Max]	0 [0, 1.00]
Presence of More than 4 Positive Lymph Nodes	
Mean (SD)	0.274 (0.446)
Median [Min, Max]	0 [0, 1.00]
Event Type	
Mean (SD)	1.50 (0.500)
Median [Min, Max]	1.50 [1.00, 2.00]

(**figure1)** (Table 1: This table presents the number and percentage of observations in each category for categorical variables, and the mean, standard deviation, minimum, and maximum values for continuous variables in "colon" dataset. It also indicates whether any missing values are present in the dataset.)

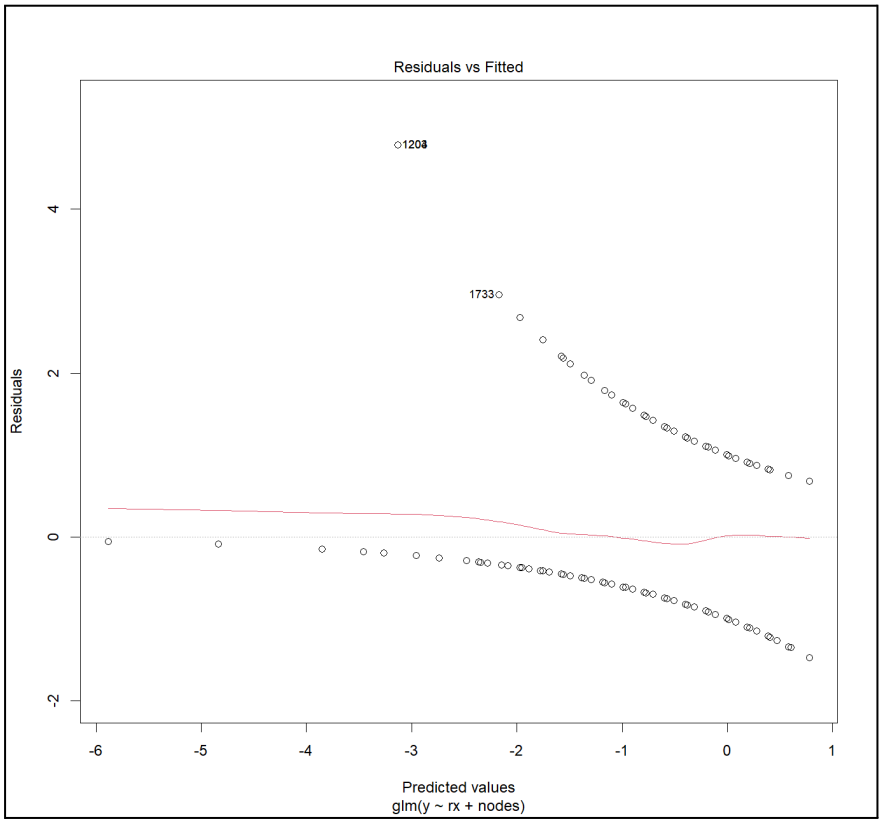
Coefficients

	Estimate	Std. Error	z valur	Pr(>|z|)
(Intercept)	0.47234	0.1032	4.577	4.72E-06
rxlev	0.13067	0.12056	1.084	0.278
rxlev+5FU	0.50326	0.12192	4.128	3.66E-05
nodes	-0.19658	0.01863	-10.551	< 2e-16

table_of_coefficients**(figure2)**

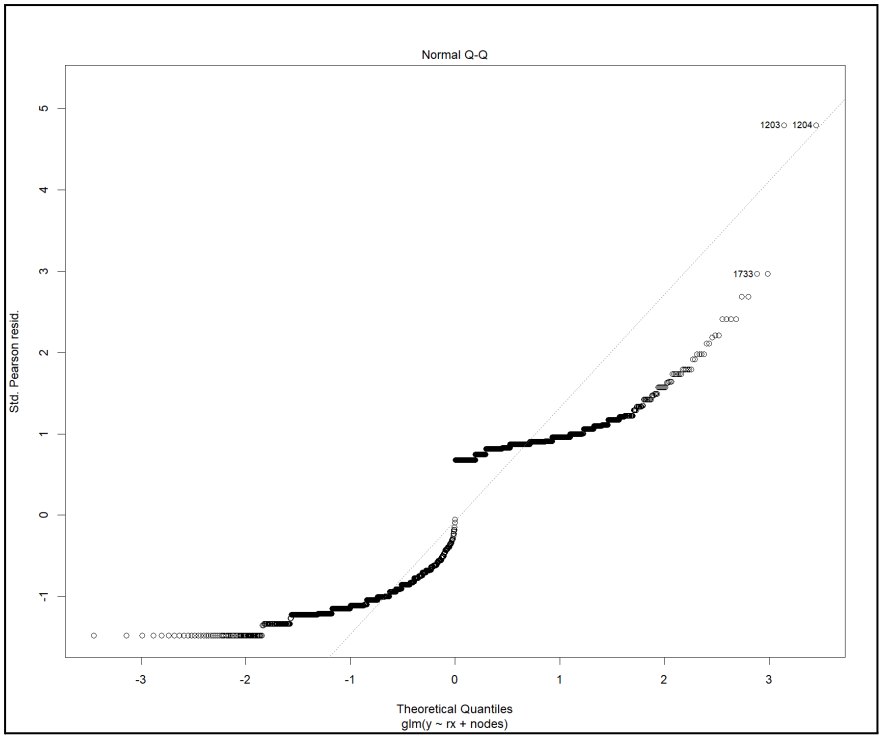


survival_histogram**(figure3)**

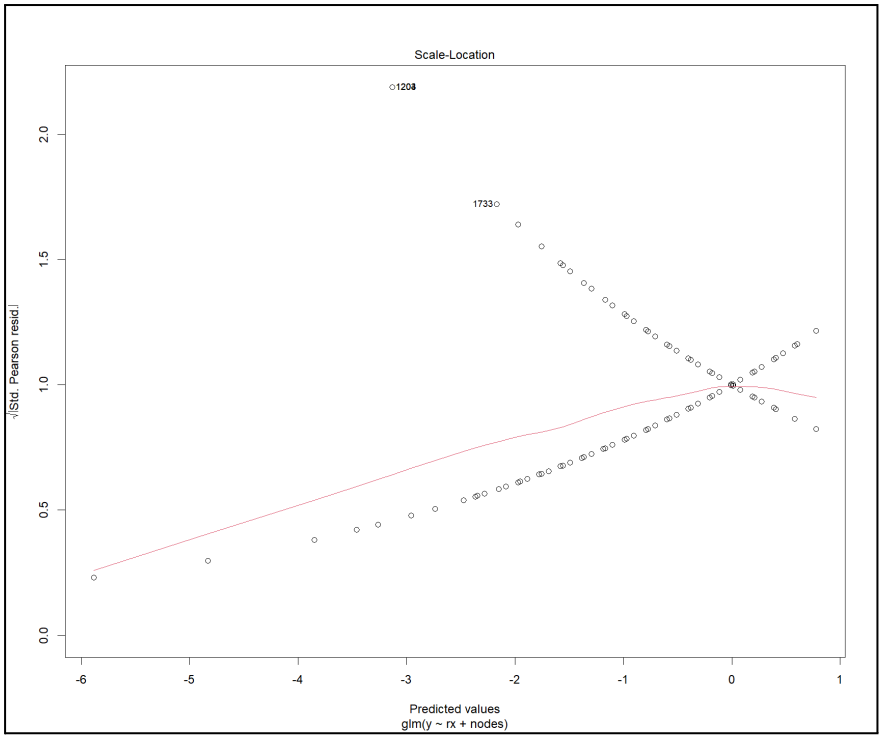


residual_fitted**(figure4)**

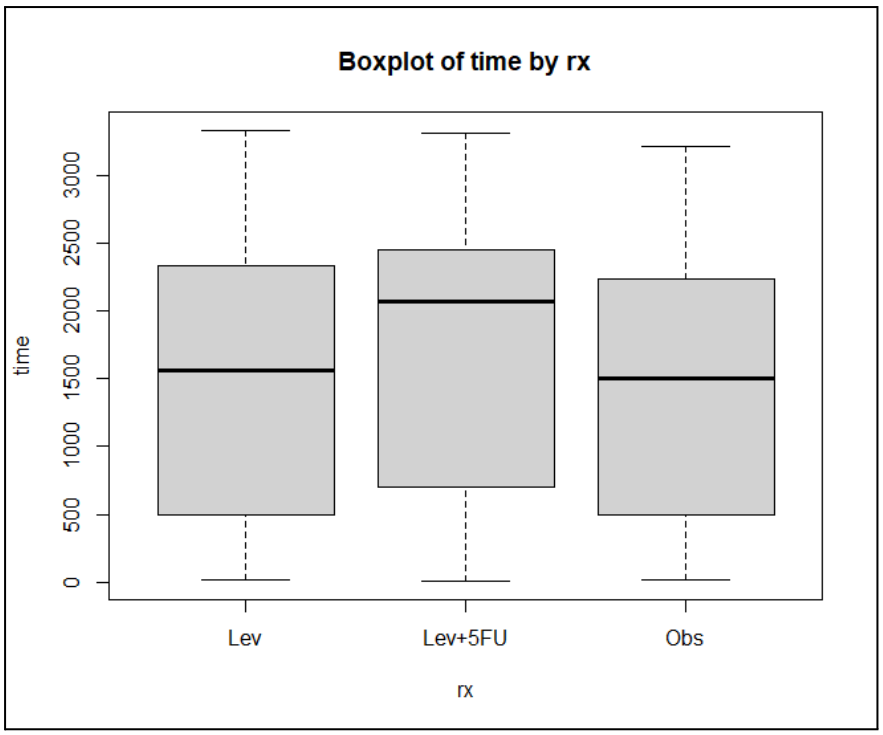
x:Predicted y:Residuals



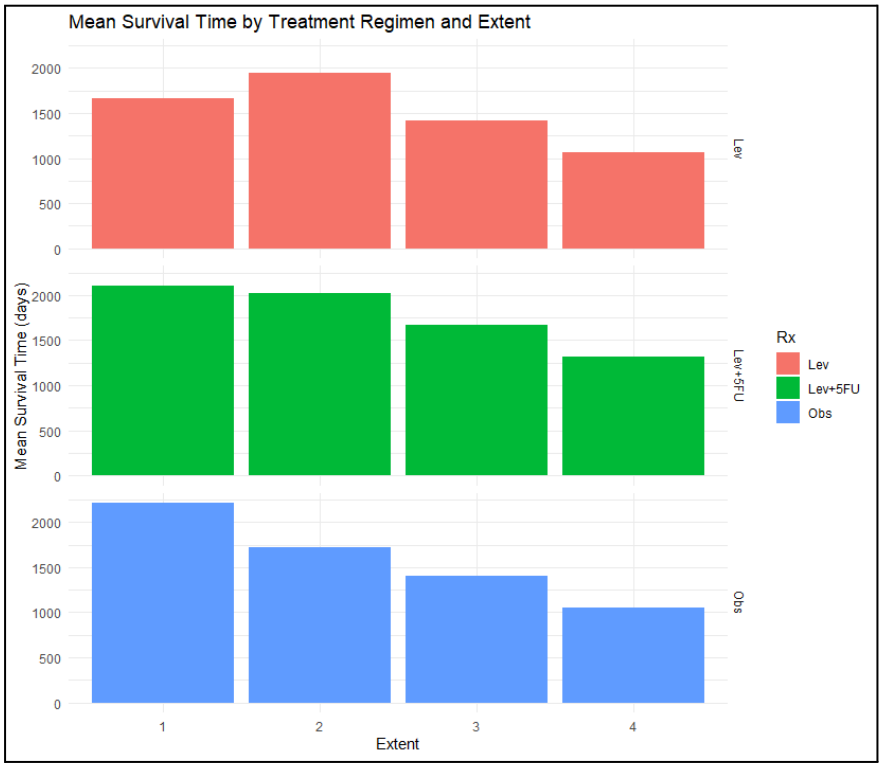
normal_Q-Q**(figure5)** x:Theoretical y:Std. Pearson resid.



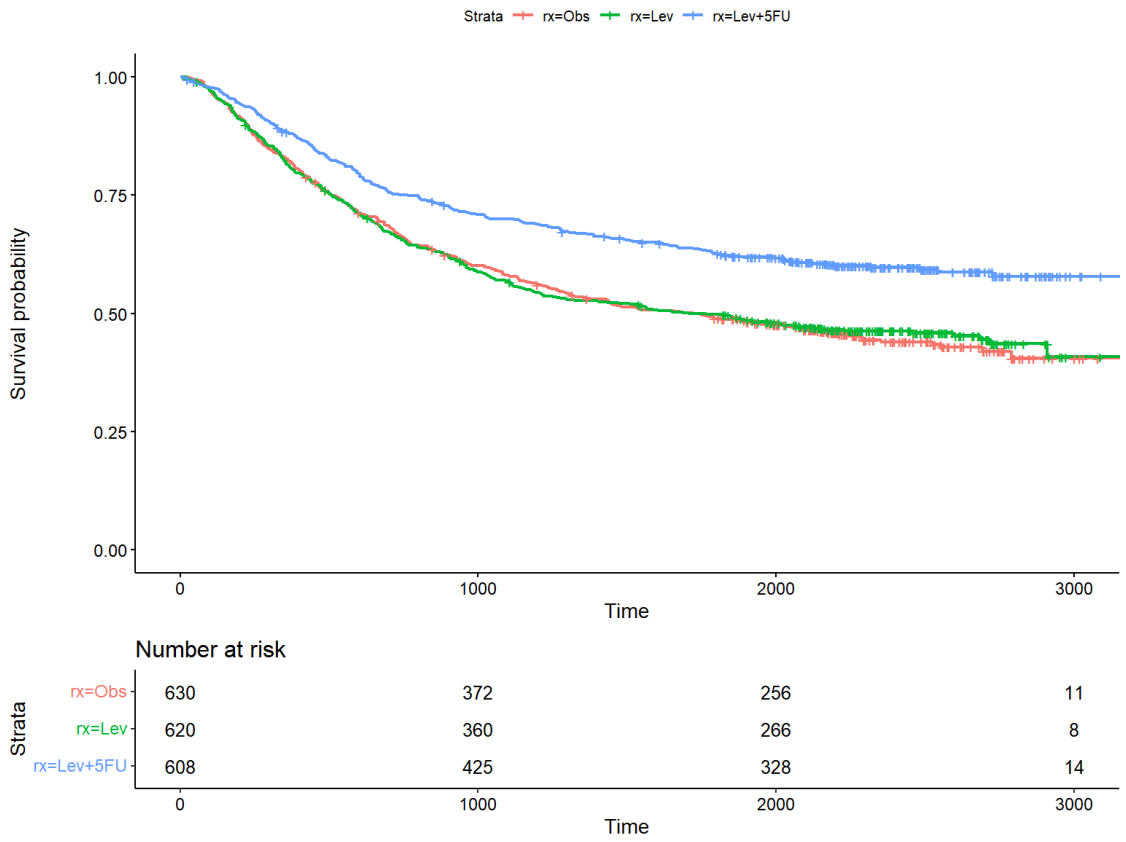
scale_location**(figure6)** x:Predicted values y:Std. Pearson resid



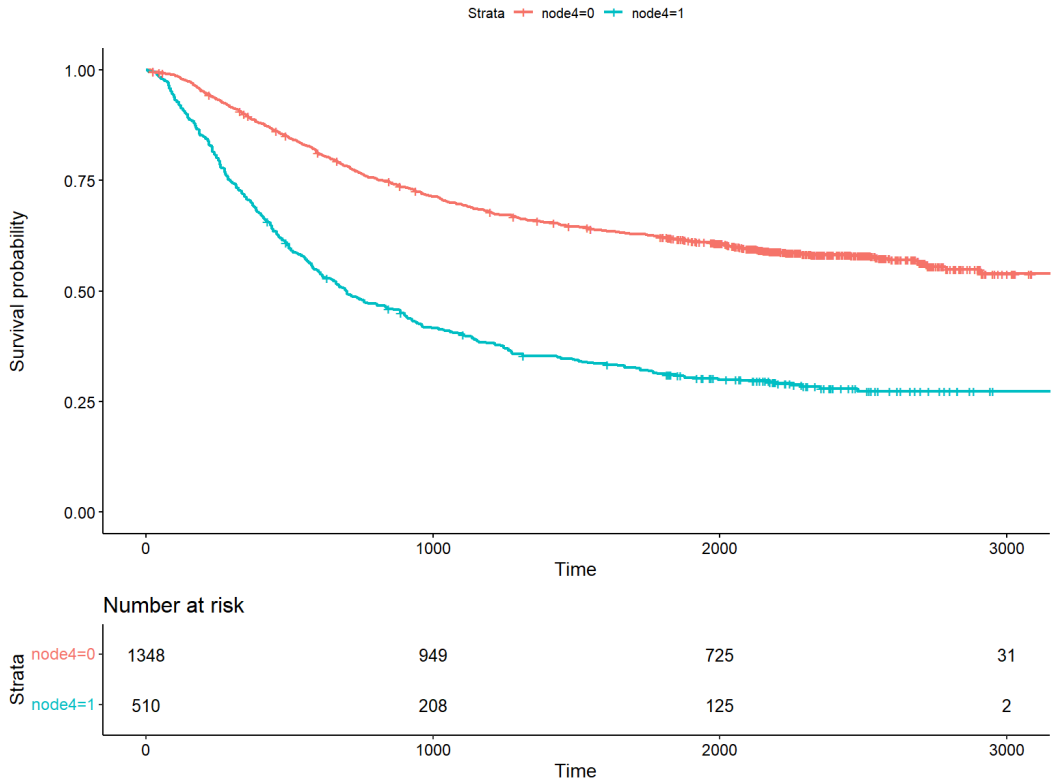
rx_time**(figure7)**



time_rx_extent**(figure8)** x:Extent y:Mean survival Time(days)



Kaplan-Meier plot of rx **(figure9)** x:Time y:Survival probability



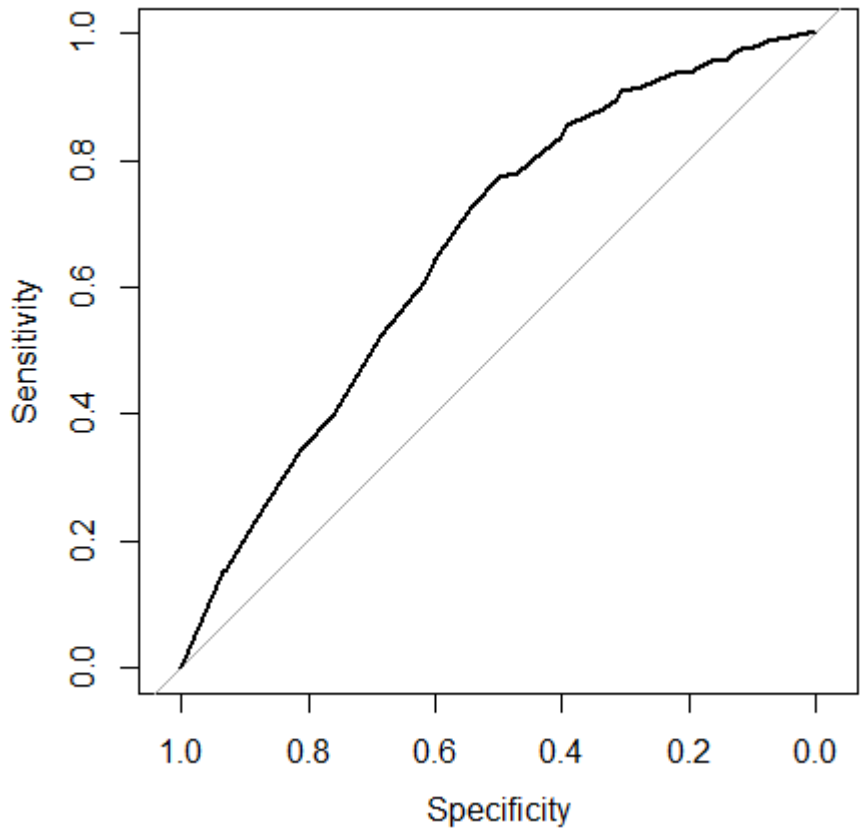
Kaplan-Meier plot of node4 **(figure10)** x:Time y:Survival Probability

	coef	exp(-coef)	se(coef)	z	Pr(>|z|)
rxLev	-0.021	0.98	0.077	-0.272	0.786
rxLev+5FU	-0.441	0.643	0.084	-5.256	1.47E-07
Concordance = 0.545 (se= 0.009)					
			df	p	
Likelihood ratio test			35.23	2	2E-08
wald test			33.11	2	6E-08
Score (logrank) test			33.63	2	5E-08

time~rx **(figure11)**

	coef	exp(-coef)	se(coef)	z	Pr(>|z|)
node4	0.91	2.47	0.068	13.34	<2e-16
Concordance = 0.6 (se= 0.008)					
			df	p	
Likelihood ratio test			162.2	1	<2e-16
wald test			177.9	1	<2e-16
Score (logrank) test			190.2	1	<2e-16

time~node4 **(figure12)**



AUC (Area Under the ROC Curve) plot **(figure13)**