Final Report: Breast Cancer Survival Prediction

Sitian Zhou (sz3213), Shuchen Dong (sd3731), Mengxiao Luan (ml5018), Pei Tian (pt2632)

Abstract

Breast cancer is one of the most common cancers affecting millions of women worldwide, necessitating ongoing research and advancements in detection, treatment, and prevention. In our project, we apply survival analysis to a comprehensive breast cancer dataset, aiming to uncover key factors influencing patient outcomes and survival rates. Employing statistical techniques such as correlation analysis and Cox Proportional Hazard model, our study rigorously explores significant risk factors and their interactions, leading to the development of a robust model for predicting survival probabilities. With thorough cross-validation, we have established the reliability of our model while commenting on its performance across different racial groups. Overall, our project offers critical insights into the survival probability analysis for patients affected by breast cancer, enhancing our understanding of this complex disease.

Introduction

Breast cancer in females ranks as the most frequently identified cancer and the fifth primary cause of death worldwide. While breast cancer can also affect men, such cases are usually uncommon. It occurs as healthy cells in the breast undergo alterations, proliferating uncontrollably and typically resulting in the formation of a mass known as a tumor.

Various survival statistics serve as important tools for physicians to assess a breast cancer patient's likelihood of recovery. Survival rate, for instance, is employed to forecast the impact of cancer on life expectancy. It varies based on diverse factors, such as cancer stage, the patient's age and overall health condition, and the effectiveness of the treatment plan [1].

Methods

Correlation

High correlations between covariates can lead to multicollinearity issues, resulting in less reliable statistical inferences. Thus, we first examined the pairwise correlation between variables.

To calculate the Pearson correlation coefficient (PCC) between variables, we first convert variables with binary outcomes (*a_stage*, *estrogen_status*, and *reginol_node_positive*) to dummy variables, as PCC only applies to numeric variables. Then we calculate the pairwise correlation between these variables. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, the Pearson correlation coefficient *r* is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cramer's V is used to evaluate the association between categorical variables. Compared to the Chi-square test which is commonly used to test for independence between categorical variables, Cramer's V includes additional information on how strong the association is.

The formula for Cramer's V coefficient is as follows:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where *N* is the total sample size, *k* is the smallest number of rows or columns, and χ^2 is the computed chi-square.

To explore the association between numeric and categorical variables with more than two levels, we performed ANOVA for each pair and calculated the corresponding eta squared value. Eta squared (η^2) measures the proportion of variance that can be explained by a given variable in the model, which is computed by dividing $SS_{between}$ by SS_{total}. An eta squared of 0.06 indicates a medium effect size between variables and 0.14 or higher indicates a large effect size [2].

Kaplan-Meier Estimation

The Kaplan-Meier estimator is a non-parametric statistical method utilized to estimate the survival function using lifetime data [3]. It breaks down the survival function into sequential probabilities and

accommodates instances where observations are censored at certain time intervals. It's important to note that the K-M estimator of survival might not always integrate to 1 due to censoring.

The K-M estimator $\widehat{S}(t)$ can be produced through the formula below:

$$\widehat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i})$$

where t_i refers to a time point when at least one event takes place, d_i denotes the number of events happening at t_i, and n_i stands for the number of survivors at t_i.

The log-rank test is a non-parametric statistical hypothesis test comparing the survival patterns among multiple sample groups. Its null hypothesis assumes that the distributions of these groups are identical [4]. This test is constructed by calculating the observed and expected event counts within a group at each recorded event time and then aggregating these figures to provide an overall summary across all event time points [5].

Suppose we have *i* groups and *j* distinct times when events take place in any group. Let *N* denote the number of survivors and *O* denote the number of events in a certain group at a certain time point. Then we have:

$$\text{Expectation } E_{i,j} = O_j \frac{N_{i,j}}{N_j} \text{ and variance } V_{i,j} = E_{i,j} \frac{N_j - O_j}{N_j} \frac{N_j - N_{i,j}}{N_j - 1}$$

The test statistic, defined as follows converges to a standard normal distribution as sample size increases.

$$Z_i = \frac{\sum_{j=1}^{J} (O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}}$$

Cox Proportional Hazard Model

The Cox proportional hazard model is a semi-parametric approach used to establish the link between subjects' hazard and a group of predictors. It assumes that the hazard function depends on a set of parameters, known as the regression coefficients of the model [6].

Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ denote the realized values of the *p* covariates for subject *i*. Then the hazard function for the Cox proportional hazards model takes the form below:

$$\lambda(t|X_i) = \lambda_0(t)exp(\beta_1 X_{i1} + \ldots + \beta_p X_{ip}) = \lambda_0(t)exp(X_i\beta)$$

where $\lambda_0(t)$ is the baseline hazard identical for all subjects.

The comparison between Cox models can be accomplished by assessing a Brier score derived from test data. This score represents the mean squared error of the predictions and serves as a measure to gauge the accuracy of the model's forecasts.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

where f_t and o_t denote the forecast probability and the actual outcome of the event at instance *t* respectively, while *N* is the number of forecasting instances [7].

Results

Summary Statistics

The dataset comprises 4,024 breast cancer patients and 16 variables in (Table 1) [8], with demographic distribution as follows: the median age of 54 years, with a majority being White (85%), Black (7.2%), and other races (8%). Marital status is predominantly married (66%), with divorced (12%) and separated (1.1%) also represented as depicted in (Table 2). Clinical features range from tumor stage to hormone receptor status. The outcome variable of interest, patient survival status, is recorded alongside the duration of follow-up, captured in *survival months.*

Transformation

The histograms (Figure 1) generated for the *tumor_size* variable from the dataset suggest that a log transformation effectively normalizes the distribution, potentially improving the model's predictive performance for survival analysis.

Correlation

The correlation plots (Figure 2) suggest a perfect correlation between variables *grade* and *differentiate*. The plots further indicate several strongly correlated variables: *n_stage, t_stage, grade, progesterone_status, a_stage, and reginol_node_positive*, which are excluded from the model fitting.

Kaplan–Meier Estimation

First, we create a K-M plot to show the estimated trend of survival probability over time (Figure 3), which falls from 1.00 to approximately 0.75 as survival months approach 100, with some subjects remaining alive at the end of observation.

Then we use some K-M plots based on each categorical variable to explore the underlying difference between the survival functions of different strata (Figure 4). All of the predictors possess a significant difference between groups, with their p-values less than 0.0001. The results of log-rank tests conform to this finding, indicating a potential correlation between these variables and our outcome (Table 3).

Cox Proportional Hazard Model

Variable Selection

Using survival months and status together as a response, we still need to filter the remaining variables to figure out a smaller subset as predictors.

We can carry out the variable selection based on the correlation results we gain above, filtering those with a Pearson correlation coefficient larger than 0.40.

Besides correlation, we should take the clinical implication into consideration as well when selecting predictors. *Differentiate* and *grade* are assigned in an exactly reverse order, resulting in the perfect negative correlation between the two. The three stages, *T-*, *M-,* and *A-stage* are components of the TNM staging system, which assigns the stage of cancer mainly according to the size and extent of the tumor. And the information contained in these columns is integrated into another variable, *6th stage*. The detected correlations between two types of markers and between the numbers of lymph nodes examined in total and as positive might lie in the pathway and progress of cancer.

Combining the two approaches for selection, the final subset of predictors to be used to construct the model include: *age*, *race*, *marital status*, *6th stage*, *differentiate*, *tumor size*, *estrogen status,* and *regional node examined*.

The linear survival model without interaction terms shows a relatively good fitting effect, with a concordance of 0.732 (se = 0.011) and nearly all the predictors significant at a significant level of 0.05.

Interaction

In this module, we delve into the complex relationship between various variables and their impact on survival outcomes. Insisting on the parsimony principle, we designed methodology as the following: starting with the calculation of $\binom{14}{2}$ possible 2-way interaction terms, then employing ANOVA analysis to select significant terms based on a 0.05 significance level (Table 4). Then, with the stepwise addition of these terms to the model, we ensure they enhance the model's performance without absorbing the significance of the main predictor. The culmination of this meticulous process results in our final model: main terms selected in the variable selection section process with interaction terms (*log_tumor_size* * *estrogen_status* + *age* * *differentiate*) (Table 5). The concordance score indicates a slight improvement in model performance with the concordance rising from 0.732 to 0.736. Especially, the inclusion of the interaction term *log_tumor_size* * *estrogen_status* in our model has revealed a noteworthy insight: the impact of *log_tumor_size* on the survival model becomes significant after adding interaction *log_tumor_size* * *estrogen_status*. It underscores the intricate interplay between *tumor_size* and *estrogen_status*, highlighting that the effect of *tumor_size* on survival outcomes is influenced by the various states of *estrogen_status*, thereby verifying the importance of interaction analysis.

Diagnostics

The established model needs to be checked for violation of assumptions as introduced previously.

- Independence: We assume the subjects are drawn independently from the underlying population.
- Collinearity: This is avoided by removing highly correlated variables before model construction.
- Linearity: Distributions of Martingale residuals suggest a linear relationship (Figure 5).
- Proportional: *estrogen_status* has a time-varying coefficient (Figure 6).

To fix this problem, we add a time-varying function on this variable as well as all the significant interactions including it.

Cross-validation

To evaluate the performance of the survival model, we use the Brier score to evaluate the difference between expected survival probability and observed survival probability with 10-fold cross-validation.

According to the boxplot result, the Brier score is always less than 0.25 with a median of around 0.12 for all batches, which indicates the good performance of the survival model.

Race-specific Comparison

In our analysis, we also focus on evaluating the performance of our model across different racial groups, uncovering a notable variance in effectiveness. From the boxplot of the Brier score (Figure 7), the performance ranking emerged as "Other" is the most accurately predicted, followed by "White", and then "Black".

Conclusion & Discussion

Considering all covariates in the data, the estimated survival probability of breast cancer patients after 107 months is 75%. K-M plots with each categorical variable stratified and the corresponding log-rank tests suggest that not all survival curves for strata in each category are the same. Ordinal factors like cancer stages exhibit declining survival with advanced stages, while nominal variables such as race and marital status show lower survival for the Black race and separated marital status.

In our final model, we found significant associations between mortality risk and demographic factors, age, race, and marital status, as well as clinical indicators: 6th stage, estrogen status, regional node count, tumor size, and differentiation grade. Notably, two interaction terms, estrogen status * log(tumor size) and age * differentiation, also contribute to mortality risk among these patients. Among these factors, the 6th stage IIIC notably increases hazard. The associated hazard ratio of 5.47 suggests that patients in the IIIC stage face a risk of death 5.47 times higher than those in the IIA stage while controlling for other factors. Additionally, non-black or white racial categories display the most pronounced effect in reducing hazard. Specifically, the risk of death for non-Black or White individuals is 0.48 times that of Black patients, holding other factors constant.

Regarding the result of cross-validation, the model performance on the race group "Black" is not as good as that of the majority group "White". To enhance the model's accuracy and fairness, we are considering a stratified strategy, which involves training separate models for each racial group. This approach aims to

tailor the predictive capabilities of our models more closely to the unique characteristics and patterns present within each group, ensuring better performance and a more equitable representation and prediction accuracy across diverse racial demographics.

Contribution

Sitian Zhou: Data analysis - correlation; Report writing - Methods, Conclusion & Discussion

Shuchen Dong: Data analysis - exploratory data analysis; Report writing - Methods, Results

Mengxiao Luan: Data analysis - K-M plot, diagnostic; Report writing - Introduction, Methods, Results

Pei Tian: Data analysis - interaction; model construction, validation; Report writing - Abstract, Results

Reference

[1] Lin, R.-H., Lin, C.-S., Chuang, C.-L., Kujabi, B. K., & Chen, Y.-C. (2022). Breast Cancer Survival Analysis Model. *Applied Sciences*, *12*(4), 1971. https://doi.org/10.3390/app12041971

[2] Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00863

[3] Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.2307/2281868

[4] Peto, R., & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. Journal of the Royal Statistical Society. Series A (General), 135(2), 185–207. https://doi.org/10.2307/2344317

[5] Mantel N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer chemotherapy reports, 50(3), 163–170.

[6] Cox, D. R. 1972. “Regression Models and Life-Tables.” Journal of the Royal Statistical Society. Series B (Methodological) 34 (2): 187–220.http://www.jstor.org/stable/2985181.

[7] Brier, G.W. (1950). VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Monthly Weather Review, 78, 1-3.

[8] JING TENG, January 18, 2019, "SEER Breast Cancer Data", IEEE Dataport, doi: https://dx.doi.org/10.21227/a9qy-ph35.

Appendix - A

Table 1. Understanding the variables

Variable	Information contained
Age	age of the patient in years at diagnosis for this cancer
Race	race of the patient
Marital status	the patient's marital status at the time of diagnosis
T-stage	part of the TNM staging system, referring to the size and extent of the primary tumor
N-stage	part of the TNM staging system, referring to the involvement of nearby lymph nodes
6th-stage	assigned stage of the cancer via combination of T, N and A classifications, tumor grade and ER/PR and HER2 test results
Grade	grade of differentiate
Differentiate	the differentiation grade, which refers to how much cancer cells resemble normal, healthy cells in terms of their structure and function
A-stage	metastasis, describes whether the cancer has spread to other parts of the body
Tumor size	indicates exact size in millimeters
Estrogen status	created by combining information from Tumor marker 1 with information from CS site-specific factor 1
Progesterone status	created by combining information from Tumor marker 2 with information from CS site-specific factor 2
Regional nodes examined	the total number of regional lymph nodes that were removed and examined by the pathologist
Regional nodes positive	the exact number of regional lymph nodes examined by the pathologist that were found to contain metastases
Survival months	number of survival months created from complete dates
Status	status of the patient on the follow-up cut-off date

Table 2. Basic attributes of breast cancer patients

Characteristic N = 4,024[1]			
age	54 (47, 61)	**differentiate**	
race		Moderately differentiated	2,351 (58%)
Black	291 (7.2%)	Poorly differentiated	1,111 (28%)
Other	320 (8.0%)	Undifferentiated	19 (0.5%)
White	3,413 (85%)	Well differentiated	543 (13%)
marital_status		**grade**	
Divorced	486 (12%)	1	543 (13%)
Married	2,643 (66%)	2	2,351 (58%)
Separated	45 (1.1%)	3	1,111 (28%)
Single	615 (15%)	4	19 (0.5%)
Widowed	235 (5.8%)	**a_stage**	
t_stage		Distant	92 (2.3%)
T1	1,603 (40%)	Regional	3,932 (98%)
T2	1,786 (44%)	**tumor_size**	25 (16, 38)
T3	533 (13%)	**estrogen_status**	
T4	102 (2.5%)	Negative	269 (6.7%)
n_stage		Positive	3,755 (93%)
N1	2,732 (68%)	**progesterone_status**	
N2	820 (20%)	Negative	698 (17%)
N3	472 (12%)	Positive	3,326 (83%)
x6th_stage		**regional_node_examined**	14 (9, 19)
IIA	1,305 (32%)	**reginol_node_positive**	2.0 (1.0, 5.0)
IIB	1,130 (28%)	**survival_months**	73 (56, 90)
IIIA	1,050 (26%)	**status**	
IIIB	67 (1.7%)	Alive	3,408 (85%)
IIIC	472 (12%)	Dead	616 (15%)

[1] Median (IQR); n (%)

Table 3. Log-rank test results

Variable	P-value
race	1.87134638638362e-07
marital status	2.37184611081271e-06
t-stage	3.26621281266811e-23
n-stage	1.27018858438536e-65
6th stage	7.30061855312582e-66
differentiate	3.53979496372722e-24
grade	3.53979496372729e-24
a-stage	5.16100435908578e-12
estrogen status	2.42400478142036e-39
progesterone status	4.54371568439955e-31

Table 4. Significant interaction terms and ANOVA p-value

Terms	P-value
age * differentiate	0.0475318374093626
age * estrogen_status	0.0195941081308159
age * log_tumor_size	0.0352493444883123
race * marital_status	0.00819441452852169
x6th_stage * regional_node_examined	0.0498067725942598
estrogen_status * log_tumor_size	9.67230122683426e-05

Table 5. Significant level of terms in the final model

Terms	Hazard ratio (=exp(coef))	P-value
age	1.0323818	0.0000053
raceOther	0.4804407	0.0005814
raceWhite	0.6853341	0.0032812
marital_statusMarried	0.8311297	0.1220838
marital_statusSeparated	1.8578236	0.0297719
marital_statusSingle	1.0594297	0.6935638
marital_statusWidowed	0.9375419	0.7213585
x6th_stageIIB	1.5851506	0.0017389
x6th_stageIIIA	2.4025912	0
x6th_stageIIIB	3.3343304	0.000006
x6th_stageIIIC	5.467602	0
differentiatePoorly differentiated	4.8198803	0.0037228
differentiateUndifferentiated	0.5351413	0.7284586
differentiateWell differentiated	0.4270495	0.4941451
estrogen_statusPositive	4.1961241	0.0318832
regional_node_examined	0.9803784	0.0004515
log_tumor_size	1.9975758	0.0000399
estrogen_statusPositive:log_tumor_size	0.5045214	0.0001604
age:differentiatePoorly differentiated	0.9782545	0.0239274
age:differentiateUndifferentiated	1.030656	0.3483772
age:differentiateWell differentiated	1.0064259	0.7643473

Appendix - B

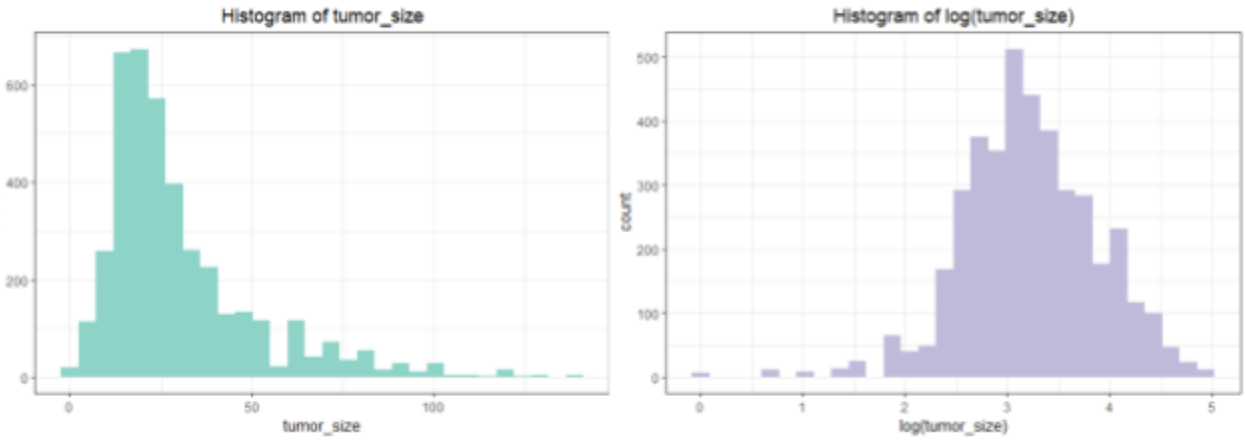


Figure 1. Transformation of *tumer_size*

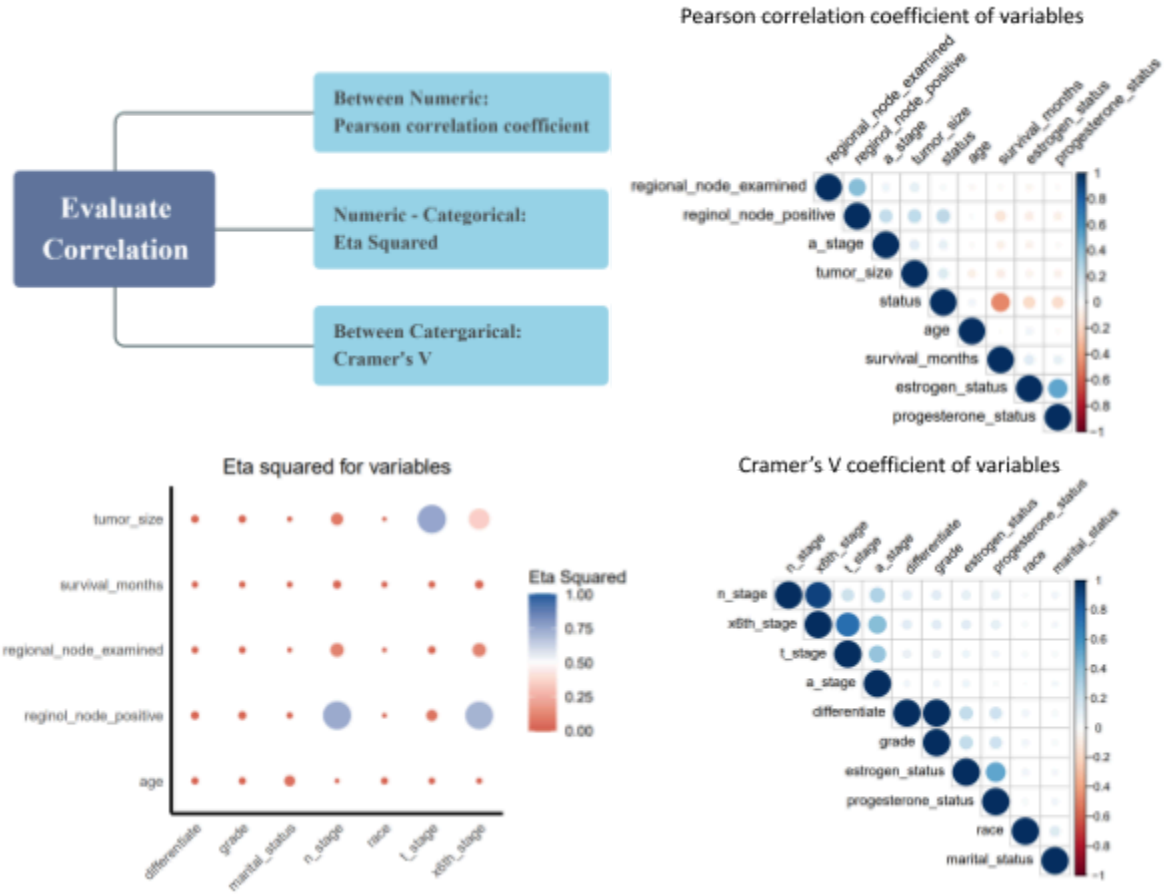


Figure 2. Correlation (or association) between variables

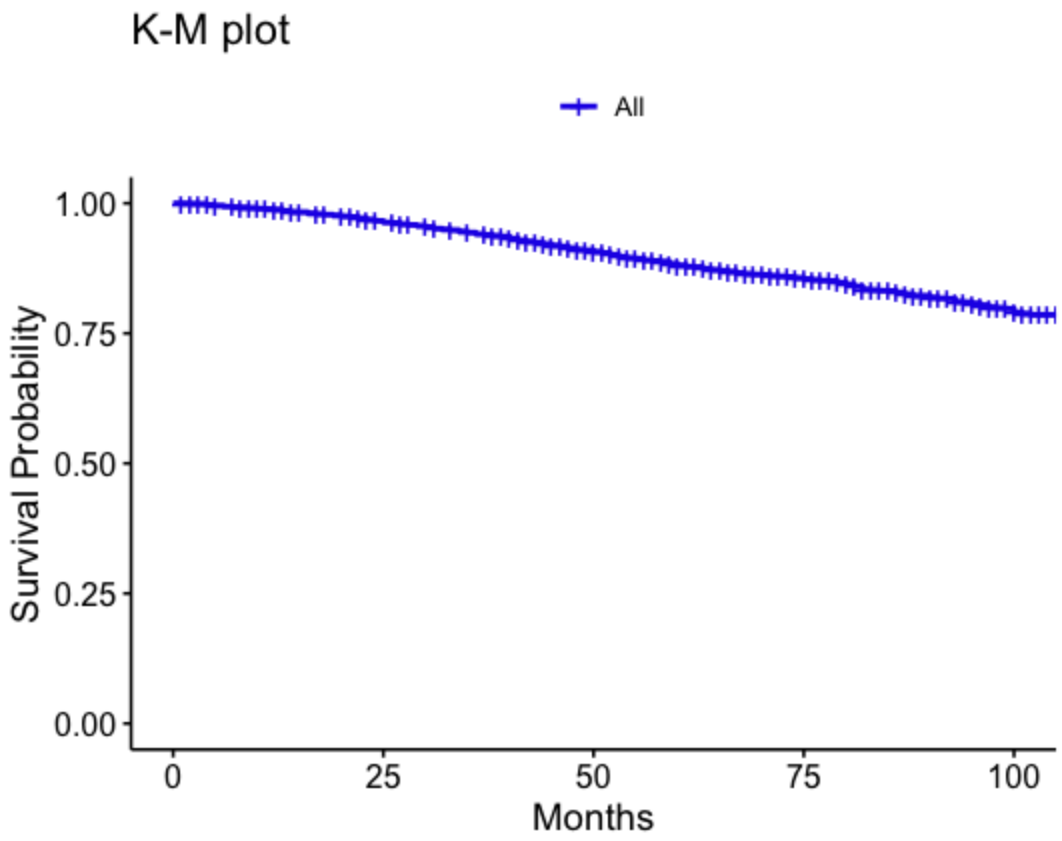
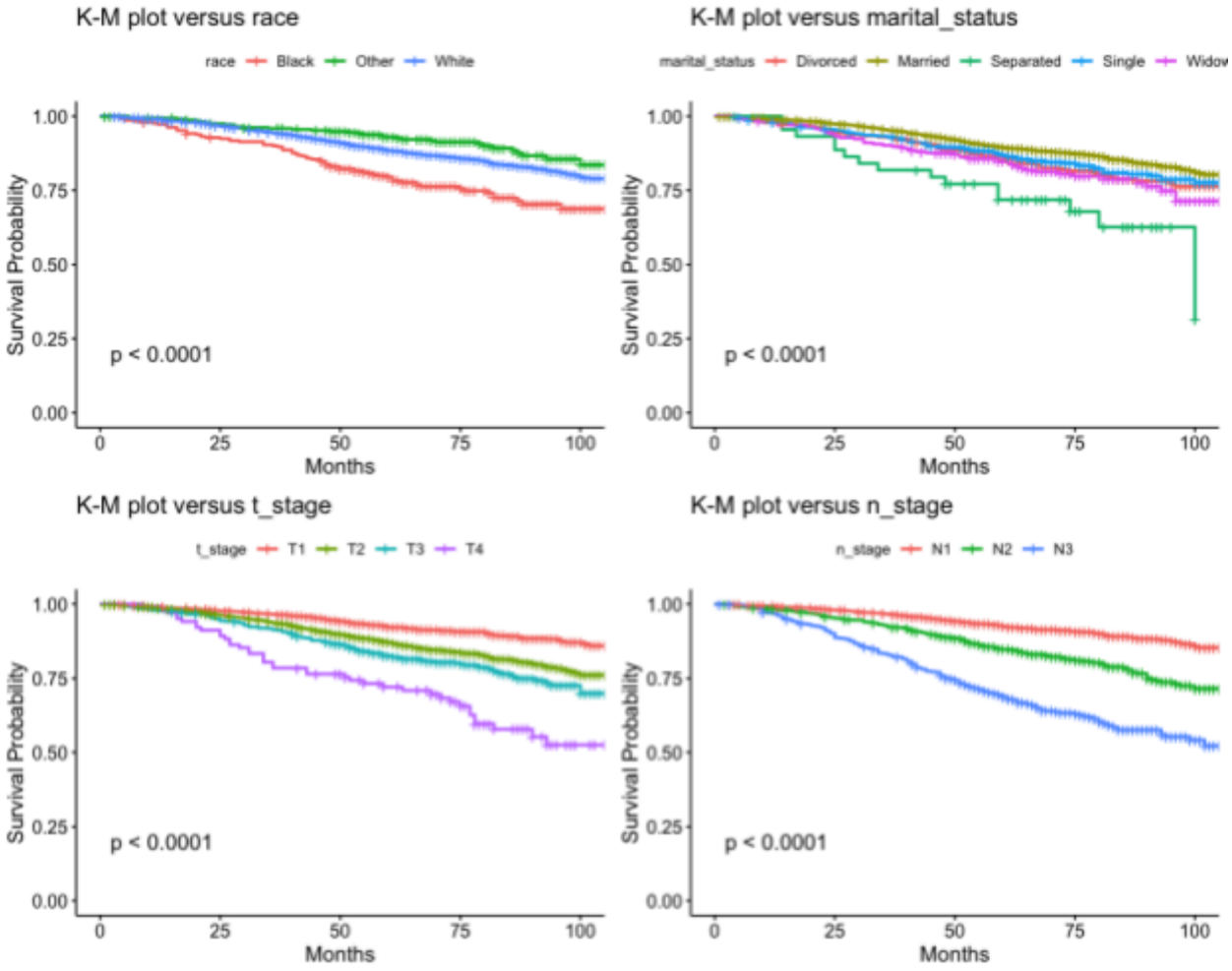


Figure 3. Overall K-M plot


K-M plot versus race
race Black Other White
Survival Probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0 25 50 75 100
Months
K-M plot versus marital_status
marital_status Divorced Married Separated Single Widow
Survival Probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0 25 50 75 100
Months
K-M plot versus t_stage
t_stage T1 T2 T3 T4
Survival Probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0 25 50 75 100
Months
K-M plot versus n_stage
n_stage N1 N2 N3
Survival Probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0 25 50 75 100
Months

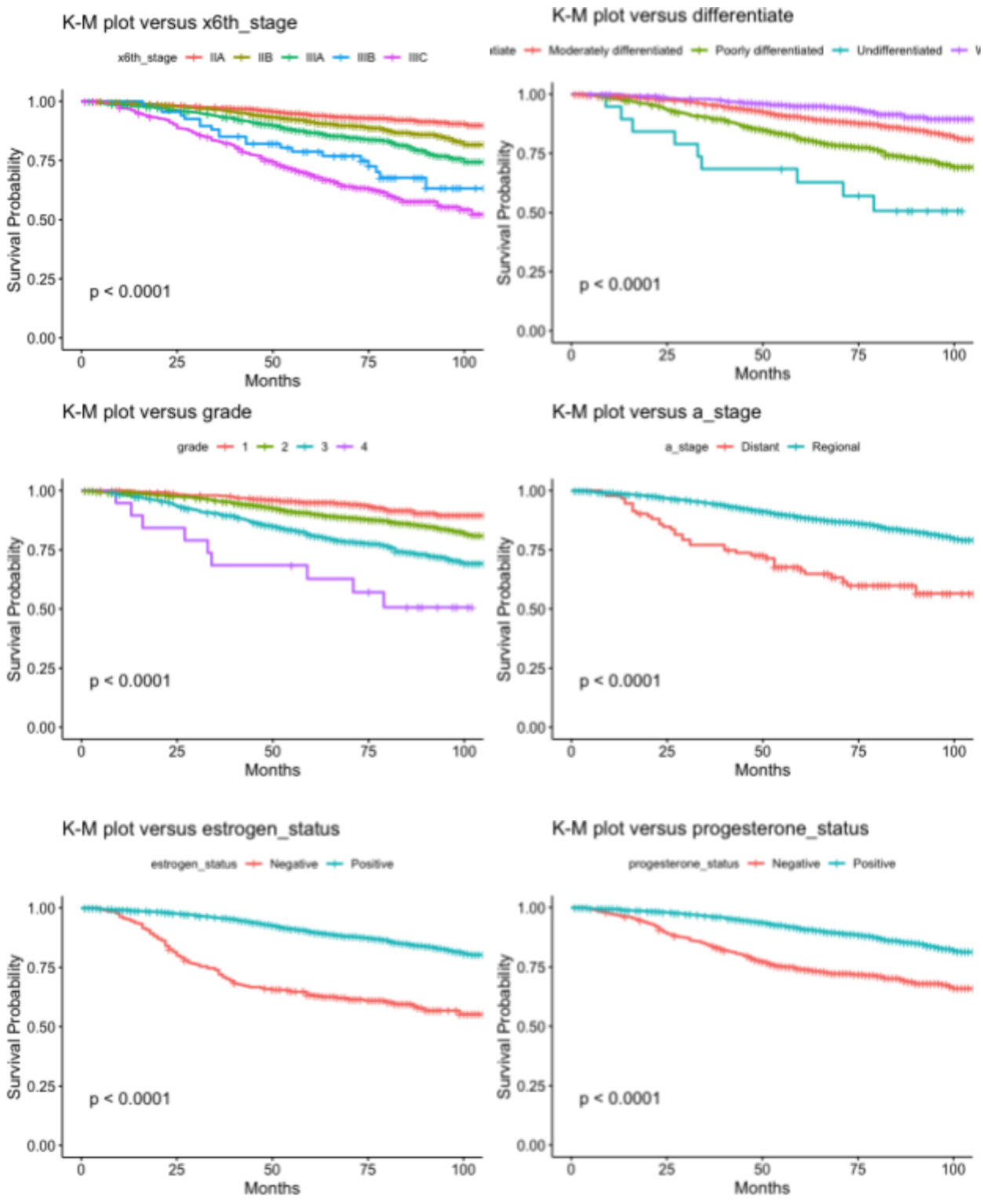


Figure 4. K-M plot for each categorical variable

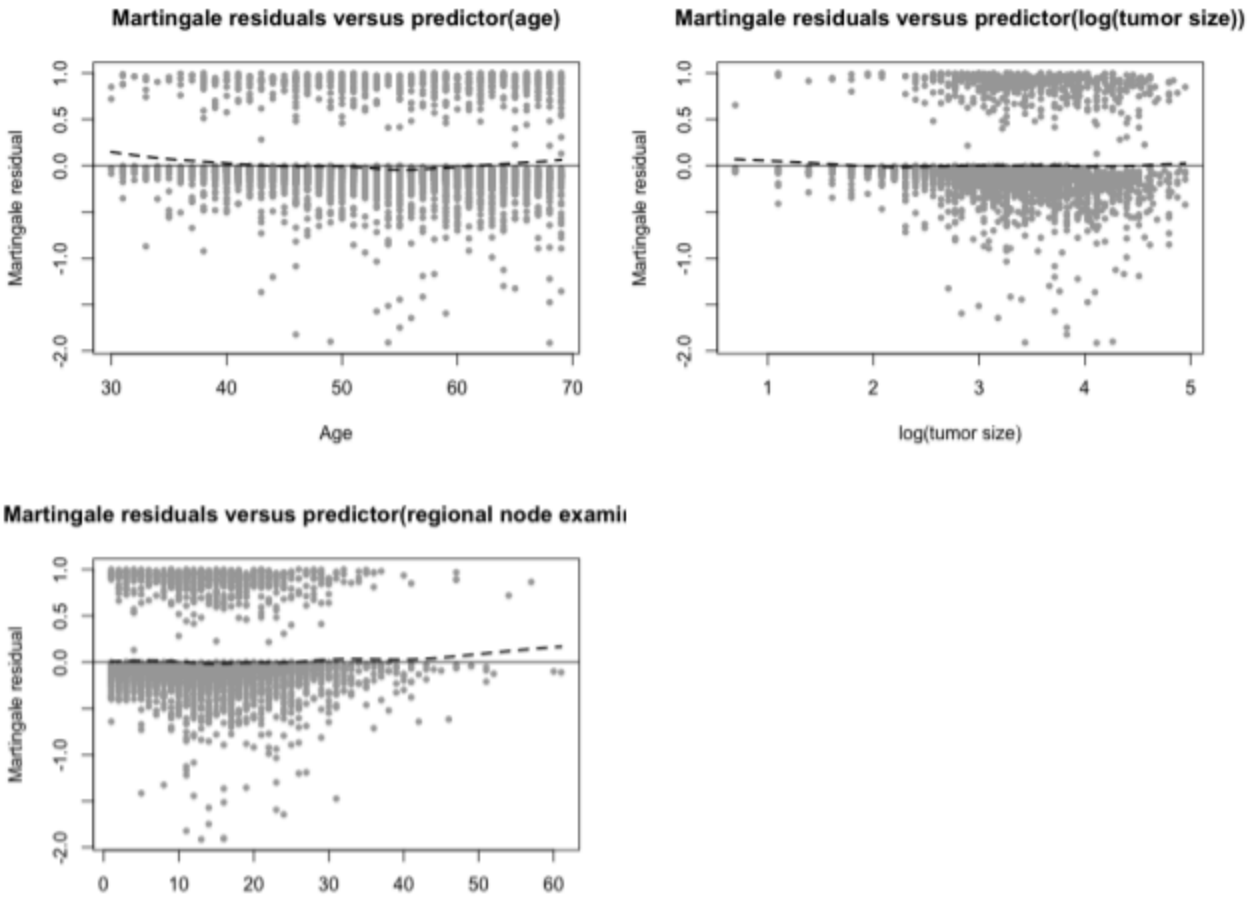
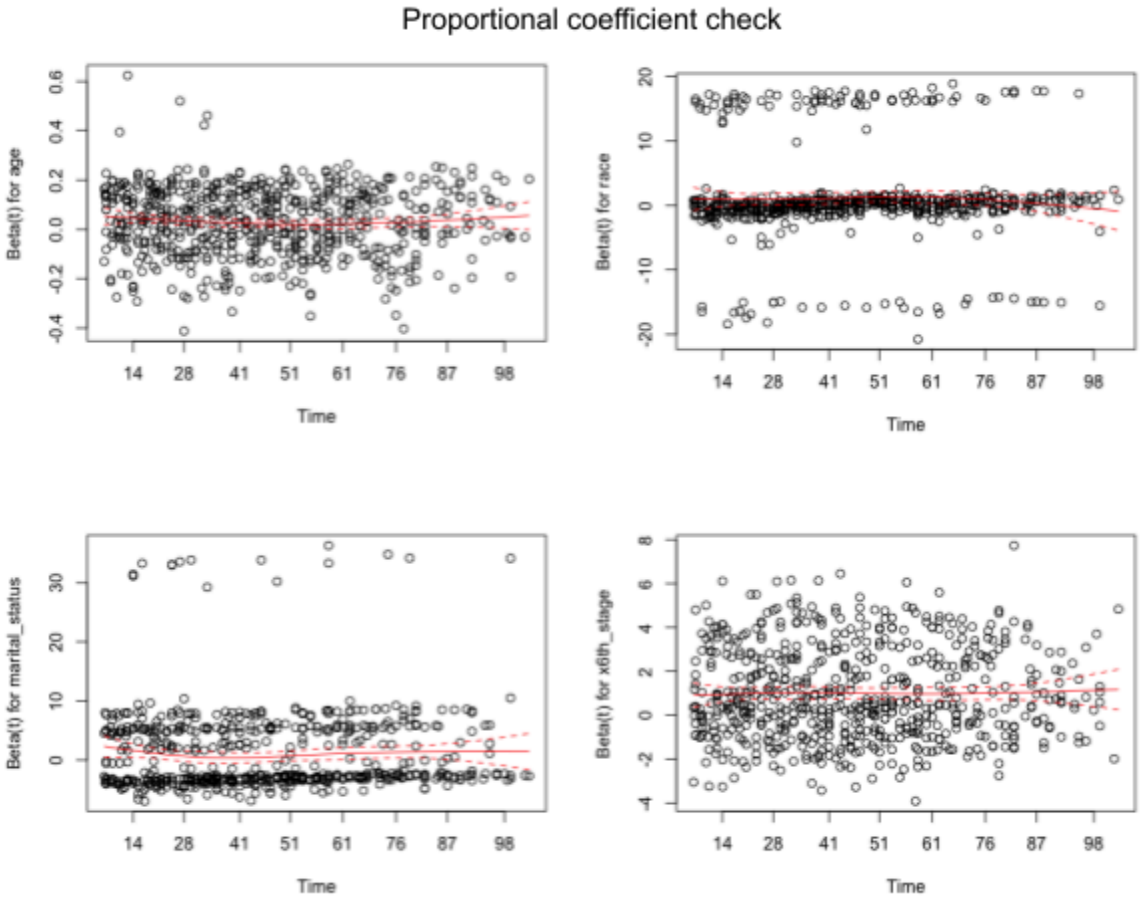


Figure 5. Linear regression check for continuous variables


Proportional coefficient check
Beta(t) for age
Beta(t) for race
Beta(t) for marital_status
Beta(t) for x6th_stage
Time
14 28 41 51 61 76 87 98

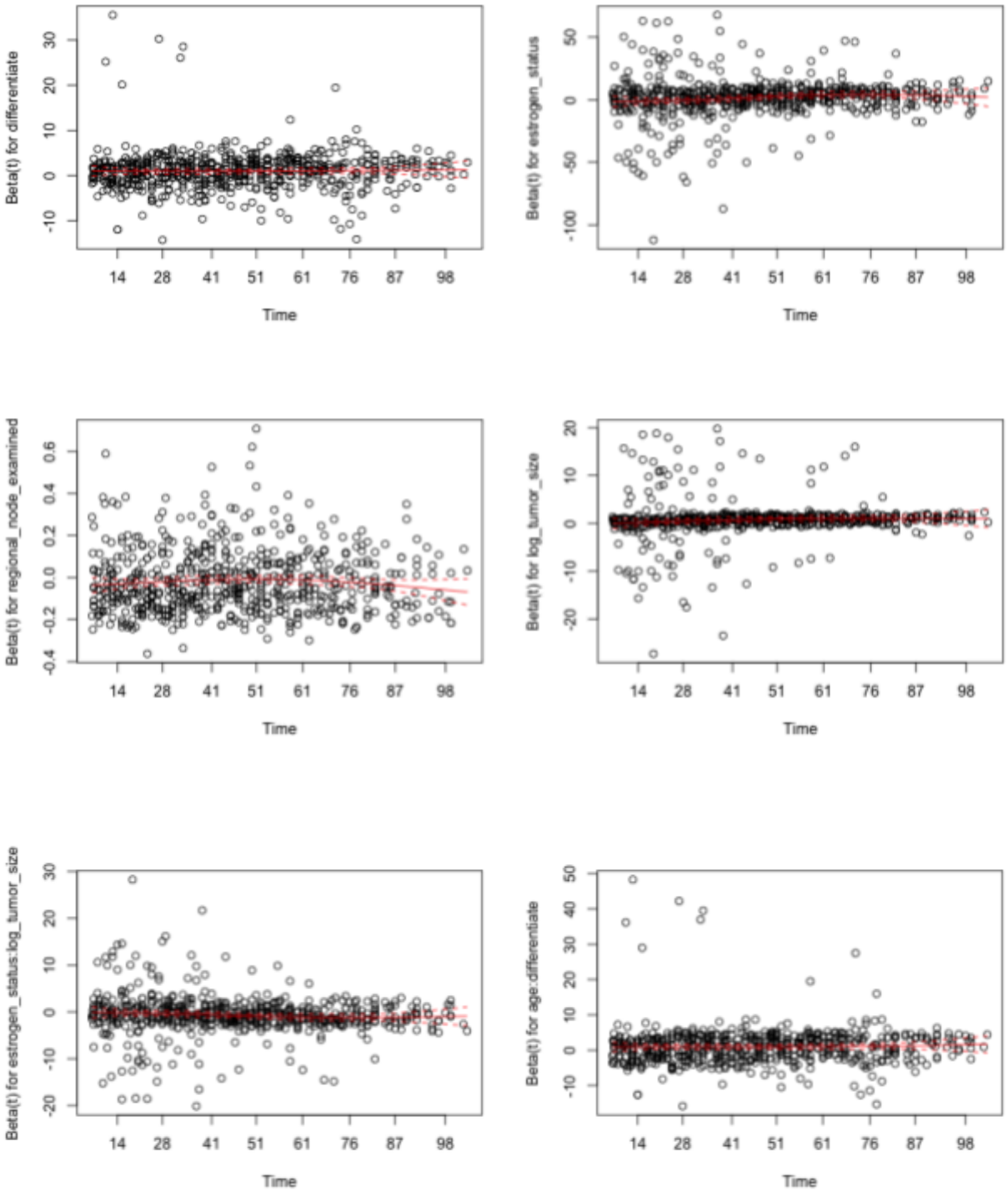


Figure 6. Proportional coefficient check for each term

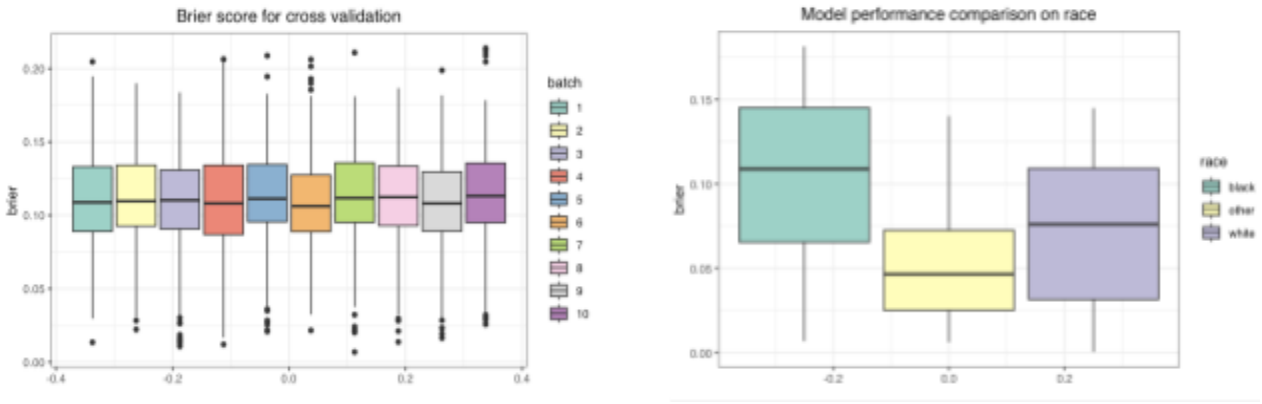


Figure 7. Model validation and evaluation