**From Smoking to Staging:**

**Understanding Lung Cancer Mortality**

**Through Logistic Models**

Thuy Dong, An Nguyen

Advisor: Dr. Ruvini Jayamaha

Applied Regression Analysis – Spring 2025

---------------------------------------------------------------------------------------------------------------------

# **I. Abstract**

According to the American Cancer Society, lung cancer kills more people every year than breast, colon, and prostate cancers combined. However, survival rates are not fixed; they can vary significantly depending on several key factors.

This project aims to explore the factors that influence survival outcomes among lung cancer patients. Utilizing a synthetic yet realistic dataset from Kaggle containing 1000 observations, we applied **logistic regression** models to access how smoking status, cancer stage, and treatment type affect the likelihood of survival. The analysis shows that smoking status is a statistically significant predictor of survival. Although cancer stage and treatment type were not statistically significant in our analysis, observed trends suggest that survival outcomes may depend on cancer stage at diagnosis.

Overall, this study contributes to a better understanding of the key factors associated with lung cancer survival and highlights the importance of early prevention efforts, such as quitting smoking. In the future, this model can be integrated into the medical system to assist doctors in making more informed decisions.

# **II. Introduction**

**i. Dataset**

* **Sources:** Kraggle – ***Lung Cancer Mortality Dataset v2***. (<https://www.kaggle.com/datasets/masterdatasan/lung-cancer-mortality-datasets-v2?select=lung_cancer_mortality_data_test_v2.csv>)
* **Description:** CSV file with 1,000 patient records and 18 variables
* **Variables:** check ***Table I.1***

Key variables used for research includes:

* **Smoking Status** (never smoker, passive smoker, current smoker, former smoker)
* **Cancer Stage** (Stage I, II, III, IV)
* **Treatment Type** (surgery, chemotherapy, radiation, or combined treatments)
* **Survival Outcome** (yes or no)

**ii. Research Question:**

This project aims to answer two primary research questions:

1. **Does smoking status significantly impact lung cancer survival outcomes?**
2. **How do cancer stage and treatment type affect lung cancer survival outcomes?**

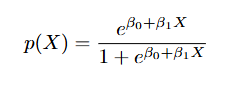
Although no formal literature review is included, the project builds upon existing medical knowledge suggesting that smoking cessation and early diagnosis are highly beneficial for improving lung cancer prognosis.

# **III. Model**

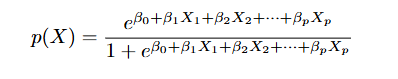
Given the binary nature of the dependent variable – whether the patient survived or not – we selected **logistic regression** as our modeling approach. Logistic regression is appropriate when the outcome variable is categorical (survived: yes or no) and allows for the estimation of odds ratios associated with different predictors such as cancer stage, smoking status, and treatment type.

We developed two logistic regression models:

* **Model 1**: Examines the relationship between survival and smoking status.



* **Model 2:** Examines the relationship between survival, cancer stage, and treatment type.



Both models were fitted using glm function in R, specifying the binomial family to reflect the binary nature of the response variable.

# **IV. Analysis**

**i. Assumption Check**

Here are the assumption checks that we’ve done for Logistic Regression model

1. **Strong outliers**

According to the ***Figure 5***, There is no strong outliers that might influence the final results.

1. **Linearity – Residuals vs Fitted Value**

**Model 1:** according to the ***Residuals vs Fitted*** plot from ***Figure 2.1***, most of the points scattered evenly near the horizontal line bisecting the graph overall. There are no clear systematic grouping or patterns, suggesting that the linearity assumption is met.

**Model 2:** according to the ***Residuals vs Fitted*** plot from ***Figure 2.2***, most of the points scattered evenly near the line bisecting the graph overall. There are no clear systematic grouping or patterns, suggesting that the linearity assumption is met.

1. **Independence – Residuals vs Fitted Value**

**Model 1:** according to the ***Residuals vs Fitted*** plot from ***Figure 2.1***, most of the points scattered evenly near the horizontal line bisecting the graph overall. There are no clear systematic clustering or patterns, suggesting that the independence assumption is met

**Model 2:** according to the ***Residuals vs Fitted*** plot from ***Figure 2.2***, most of the points scattered evenly near the horizontal line bisecting the graph overall. There are no clear systematic clustering or patterns, suggesting that the independence assumption is met

1. **Multicollinearity – VIF**

**Model 1:** There is only 1 predictor so no need for multicollinearity check.

**Model 2:** according to ***Figure 4***, the VIF values of both predictors (cancer\_state, treatement\_type) are approximately equal to 1, confirming that there is no multicollinearity in the model.

1. **Normality – QQ plot**

The normality assumption can’t be applied to this case due to the nature of logistic regression model. Logistic regression yields a binary outcome; additionally, the predictors used in this model are all categorical variables, which are not applicable for normality.

1. **Homoscedasticity - Spread-Location plot**

The homoscedasticity assumption can’t be applied to this case due to the nature of logistic regression model. Logistic regression yields a binary outcome, meaning that the variance of residuals is not constant. The ***Scale-Location*** plot from ***Figure 3.1*** & ***3.2*** are for reference only, not to validate this assumption check.

**ii. Results: Interpret the results**

**Model 1: Smoking Status**

Model 1 explores how the smoking\_status variable affects the lung cancer survival outcome. According to ***Table M1***, using the Current Smoker as the reference group, the Former Smoker is statistically significantly associated with the likelihood to survive lung cancer, with the p-value = 0.0187 (less than 0.05) and log-odds increase of 0.5132. The model’s AIC is 1034.4.

Although the coefficients for Never Smoked (0.1789) and Passive Smoker (0.4160) are positive, they are not statistically significant at the 0.05 level (p = 0.4353 and p = 0.0642, respectively).

This suggests that quitting smoking can make a big difference, even for lung cancer patients.

**Model 2: Cancer Stage and Treatment Type**

Model 2 explores how cancer\_stage and treatment\_type jointly affect the lung cancer survival outcome. According to ***Table M2***, using Stage I and No Treatment as the reference group, the coefficients for Radiation (0.231) and Surgery (0.0087) treatment types are positive, and those cancer stage II (-0.047), III (-0.144), IV (-0.068) are negative. However, there is no statistically significant variable as all p-value are larger than 0.05. The model’s AIC is 1044.7, suggesting minimal improvement.

Based on figure, the model suggests that there is no single best cancer\_stage or treatment\_type that reliably predicts the chance of survival, but it has to depend on specific case.

# **V. Conclusion**

From our model results we conclude that individuals who quit smoking have the highest chance of lung cancer survival – the clearest and most consistent pattern observed. While cancer stage and treatment type did not reach statistical significance, the findings suggest there is no single “best” treatment; instead, treatment efficacy appears to depend on the stage at diagnosis.

**Limitations and Future Works:**

There are various limitations with our research which may have influenced our final result:

* **Dataset:** This study is limited by the representativeness of the Kaggle dataset and the exclusion of potentially influential factors such as genetic data, detailed medical history, environmental exposures (like pollution or occupational risks), and how closely patient follow their treatment plans. Our Kraggle dataset is also a synthetic data which might not capture all the complexity and variability present in actual clinical data.   
  This limitations can help explain why the people who Never Smoked has lower survival rate compared to people who quit smoking (Former Smoker) in Model 1. It might be because there are less records of Never Smoked than Former Smoker in the dataset.
* **Features Selection:** The predictors (smoking\_status, treatment\_type, cancer\_stage) are intuitively handpicked as we purposely wanted to explore the relationship between survival outcomes with those variables. However, without including other variables from the dataset, we might have missed other important predictors that can affect the survival outcome.
* **Imbalance data:** The data was imbalanced (the survived variable is not equally distributed as the majority is “not survived” patients).

In future work, we plan to incorporate these additional variables and validate our models on larger diverse cohorts to improve accuracy and support more personalized treatment strategies. We plan to use the machine learning method such as Lasso or PCA to help with feature selection instead of intuitively handpick variables. We will also obtain a more realistic dataset that can closely represent our real-world lung cancer survival pattern. Additionally, our data should be carefully clean and balanced using machine learning method like undersampling or oversampling.

# **VI. References**

American Cancer Society. (2024). *Key statistics for lung cancer*. American Cancer Society. <https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html>

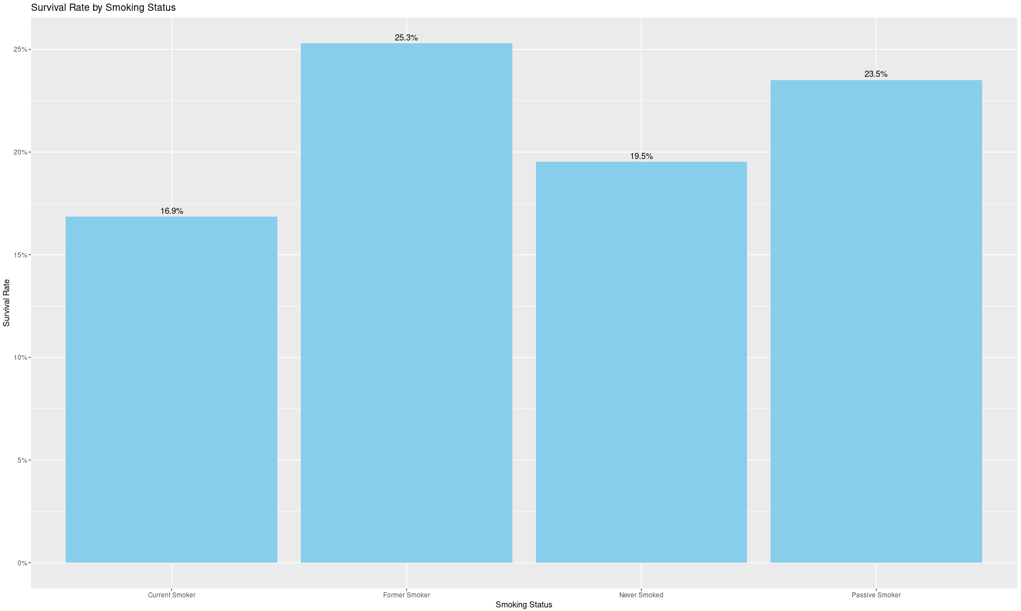
MasterDataSan. (n.d). Lung Cancer Mortality Datasets v2 [Data set]. Kaggle. Retrieved April 7, 2025, from [https://www.kaggle.com/datasets/masterdatasan/lung-cancer-mortality-datasets-v2?](https://www.kaggle.com/datasets/masterdatasan/lung-cancer-mortality-datasets-v2?select=lung_cancer_mortality_data_test_v2.csv) s elect=lung\_cancer\_mortality\_data\_test\_v2.csv

# **Appendix**

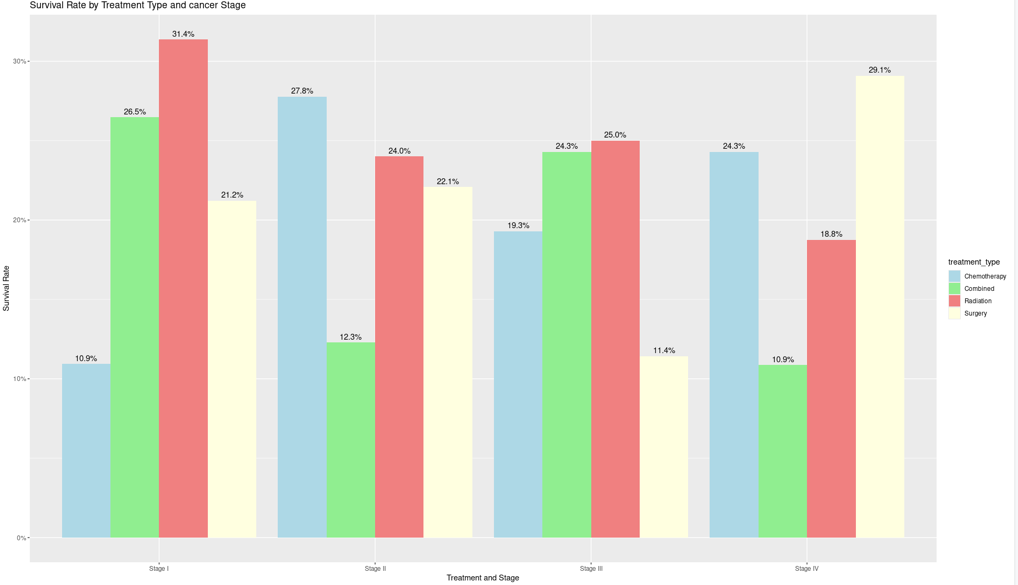
**Table I.1:**

|  |  |  |  |
| --- | --- | --- | --- |
| No. | **Variable Name** | **Type** | **Description** |
| 1 | id | numeric | A unique identifier for each patient in the dataset |
| 2 | age | numeric | The age of the patient at the time of diagnosis. |
| 3 | gender | categorical | The gender of the patient (e.g., male, female). |
| 4 | country | categorical | The country or region where the patient resides. |
| 5 | diagnosis\_date | numeric | The date on which the patient was diagnosed with lung cancer. |
| 6 | cancer\_stage | categorical | The stage of lung cancer at the time of diagnosis (e.g., Stage I, Stage II, Stage III, Stage IV). |
| 7 | family\_history | categorical | Indicates whether there is a family history of cancer (e.g., yes, no). |
| 8 | smoking\_status | categorical | The smoking status of the patient (e.g., current smoker, former smoker, never smoked, passive smoker). |
| 9 | bmi | numeric | The Body Mass Index of the patient at the time of diagnosis. |
| 10 | cholesterol\_level | numeric | The cholesterol level of the patient (value). |
| 11 | hypertension | categorical | Indicates whether the patient has hypertension (high blood pressure) (e.g., yes, no). |
| 12 | asthma | categorical | Indicates whether the patient has asthma (e.g., yes, no). |
| 13 | cirrhosis | categorical | Indicates whether the patient has cirrhosis of the liver (e.g., yes, no). |
| 14 | other\_cancer | categorical | Indicates whether the patient has had any other type of cancer in addition to the primary diagnosis (e.g., yes, no). |
| 15 | treatment\_type | categorical | The type of treatment the patient received (e.g., surgery, chemotherapy, radiation, combined). |
| 16 | beginning\_off\_treatment\_date | numeric | The date on which the patient started their cancer treatment. |
| 17 | end\_treatment\_date | numeric | The date on which the patient completed their cancer treatment or died. |
| 18 | survived | categorical | Indicates whether the patient survived (e.g., yes, no). |

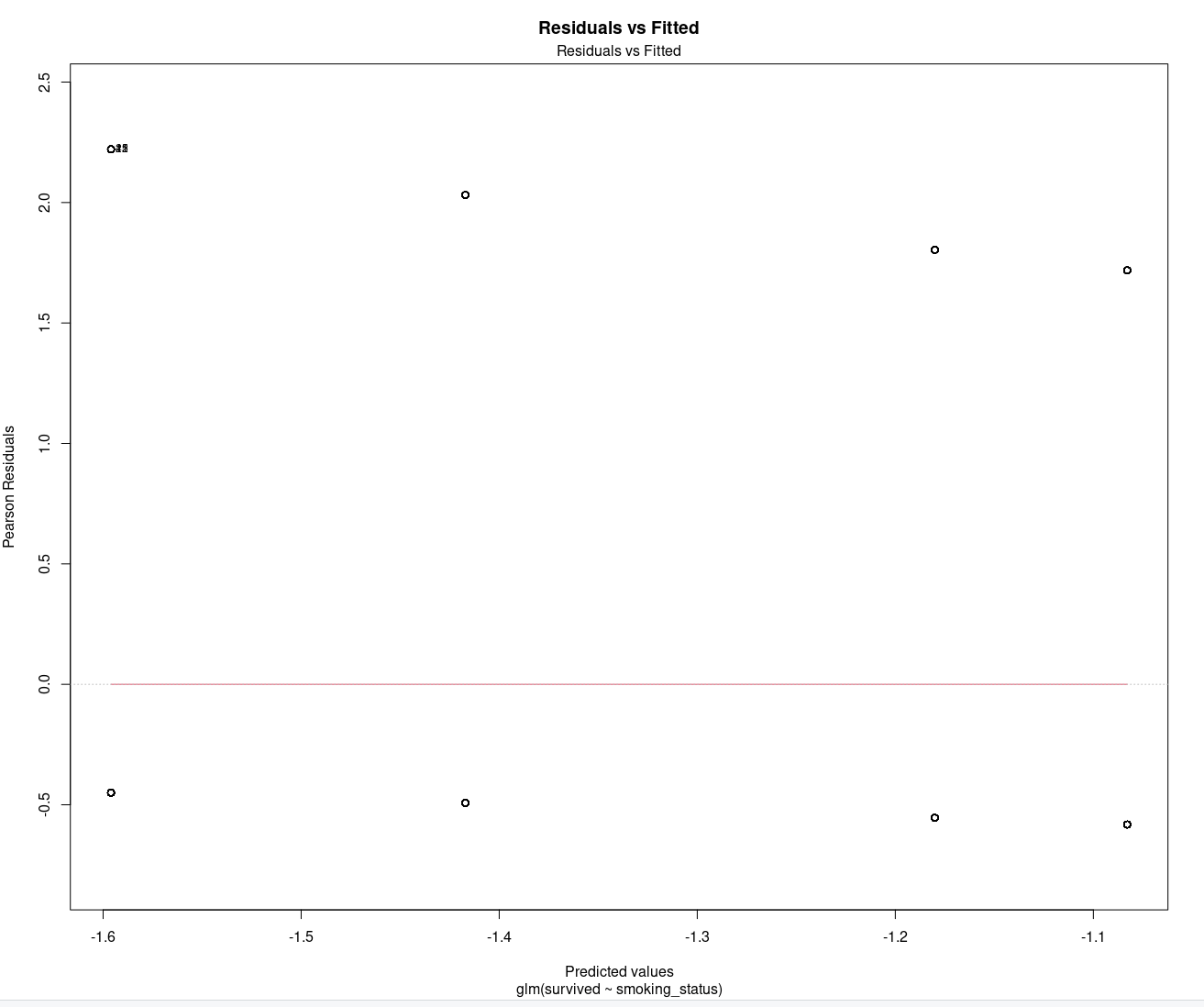
**Figure 1.1**: Full Logistic Regression Output for Model 1 (Smoking Status)



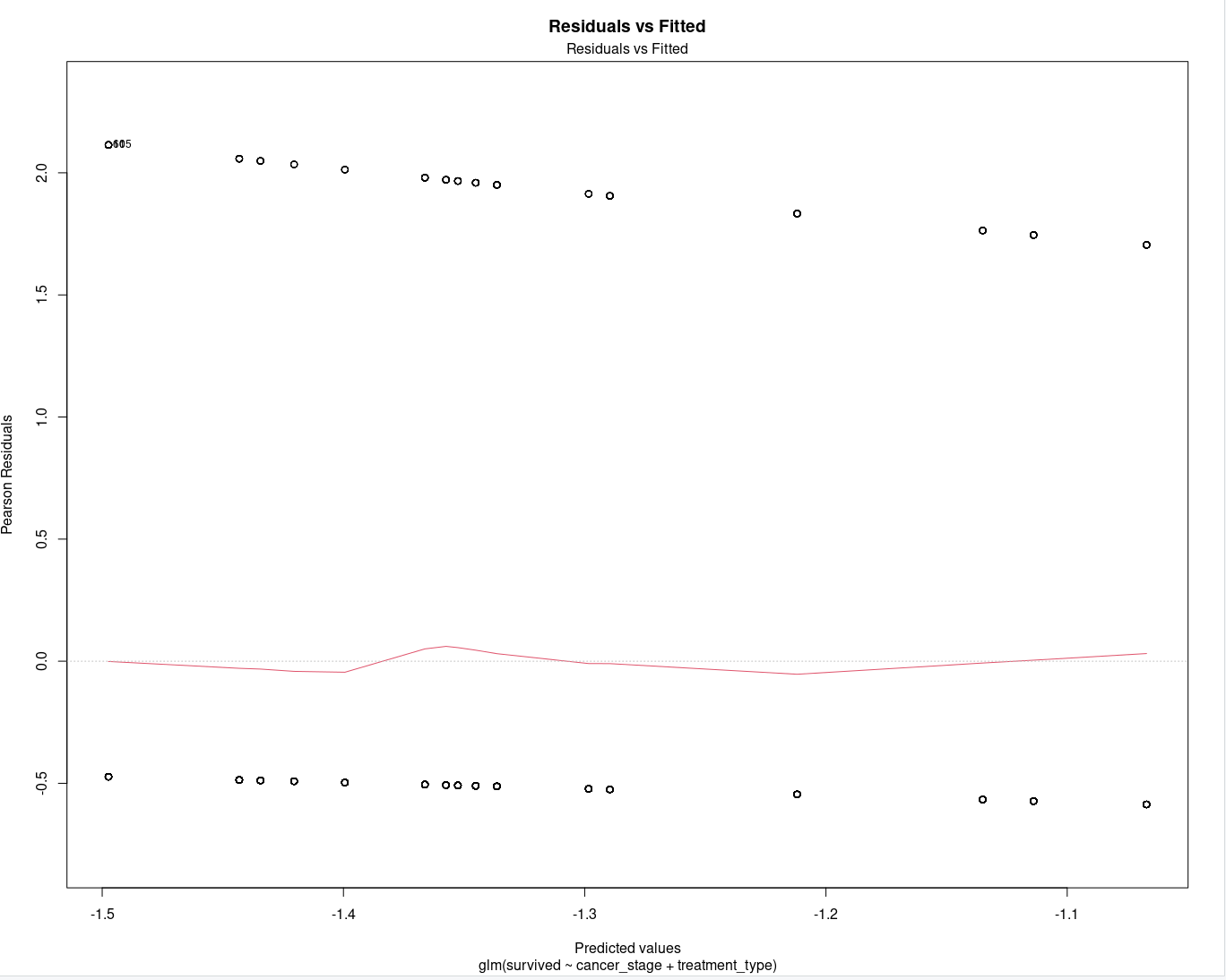
**Figure 1.2**: Full Logistic Regression Output for Model 2 (Cancer Stage and Treatment Type)



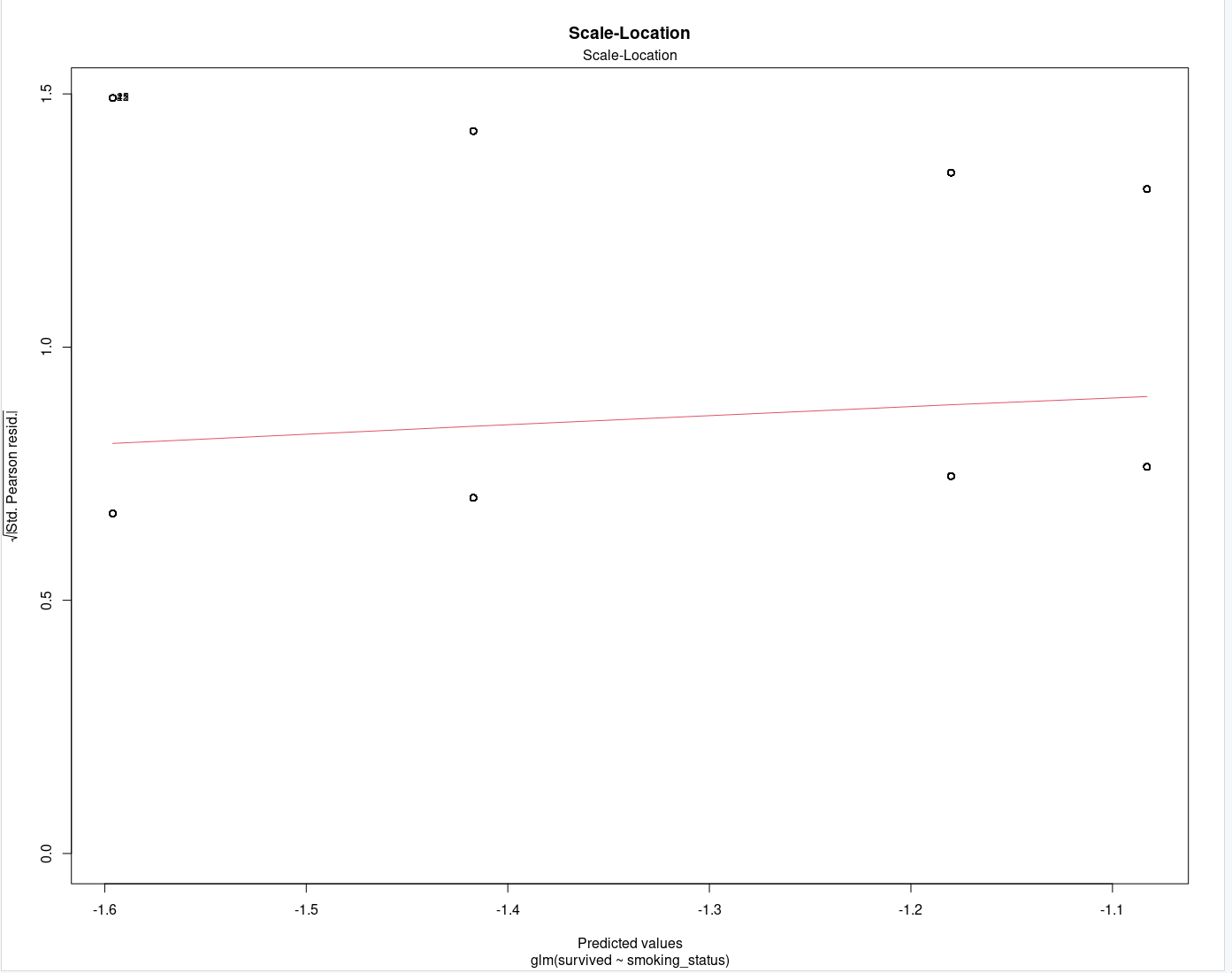
**Figure 2.1**: Linearity & Independence Check (Model 1)



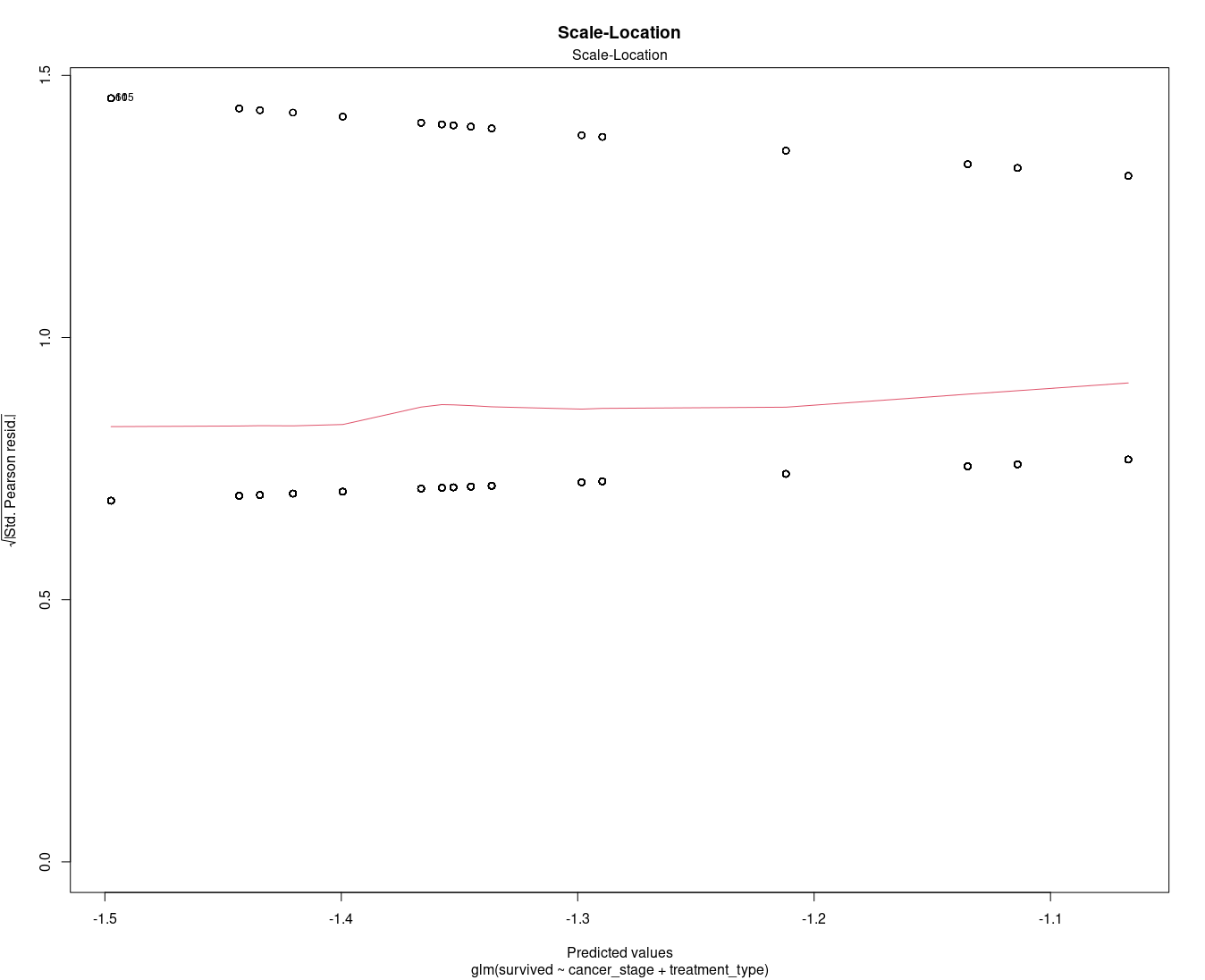
**Figure 2.2**: Linearity & Independence Check (Model 2)



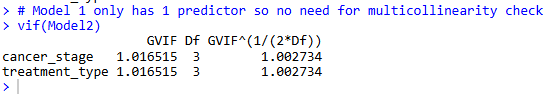
**Figure 3.1**: Homoscedasticity Check (Model 1)



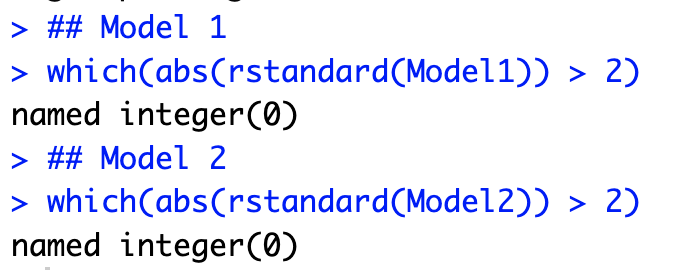
**Figure 3.2**: Homoscedasticity Check (Model 2)



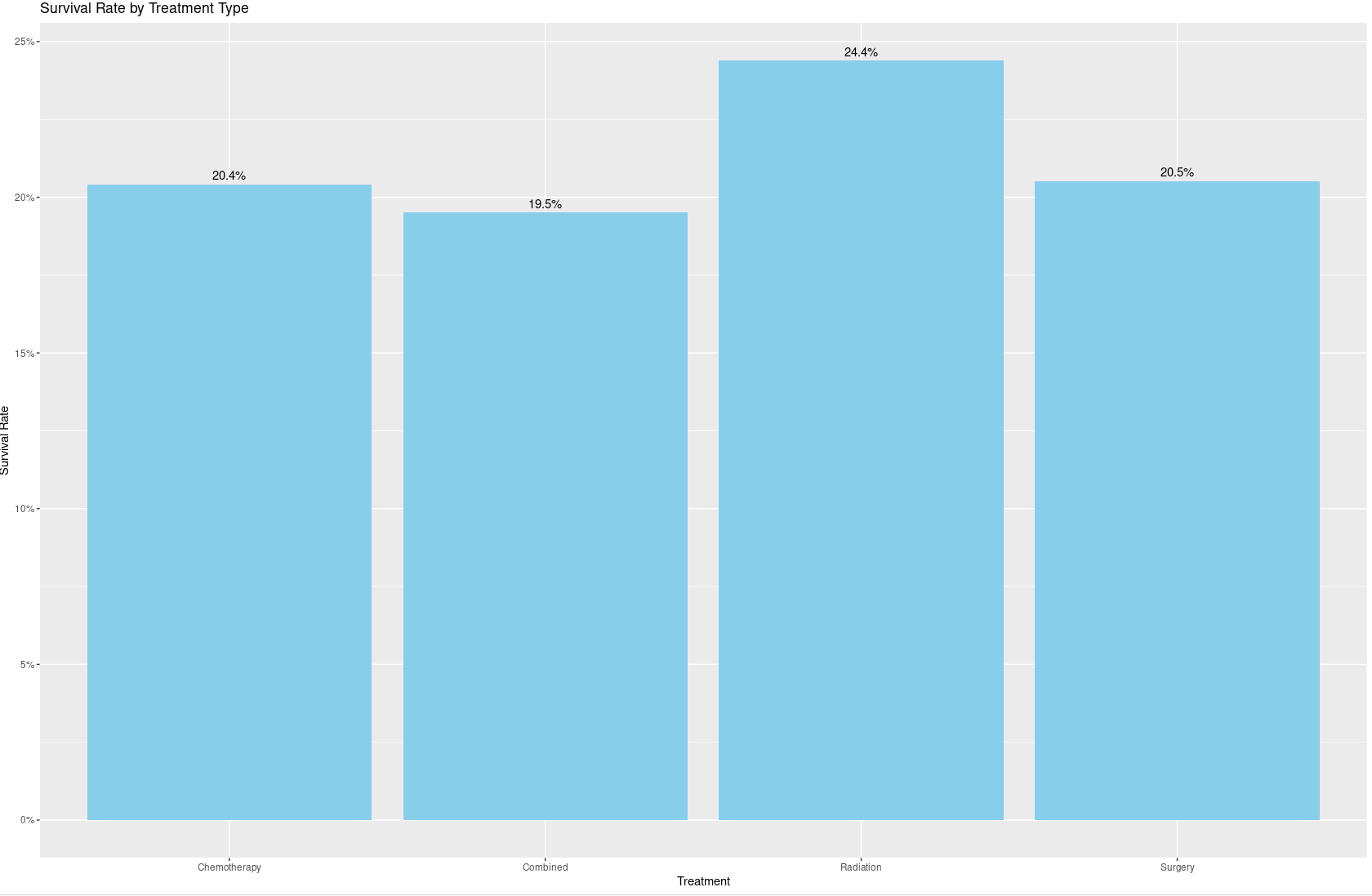
**Figure 4**: Multicollinearity Check (Model 2)



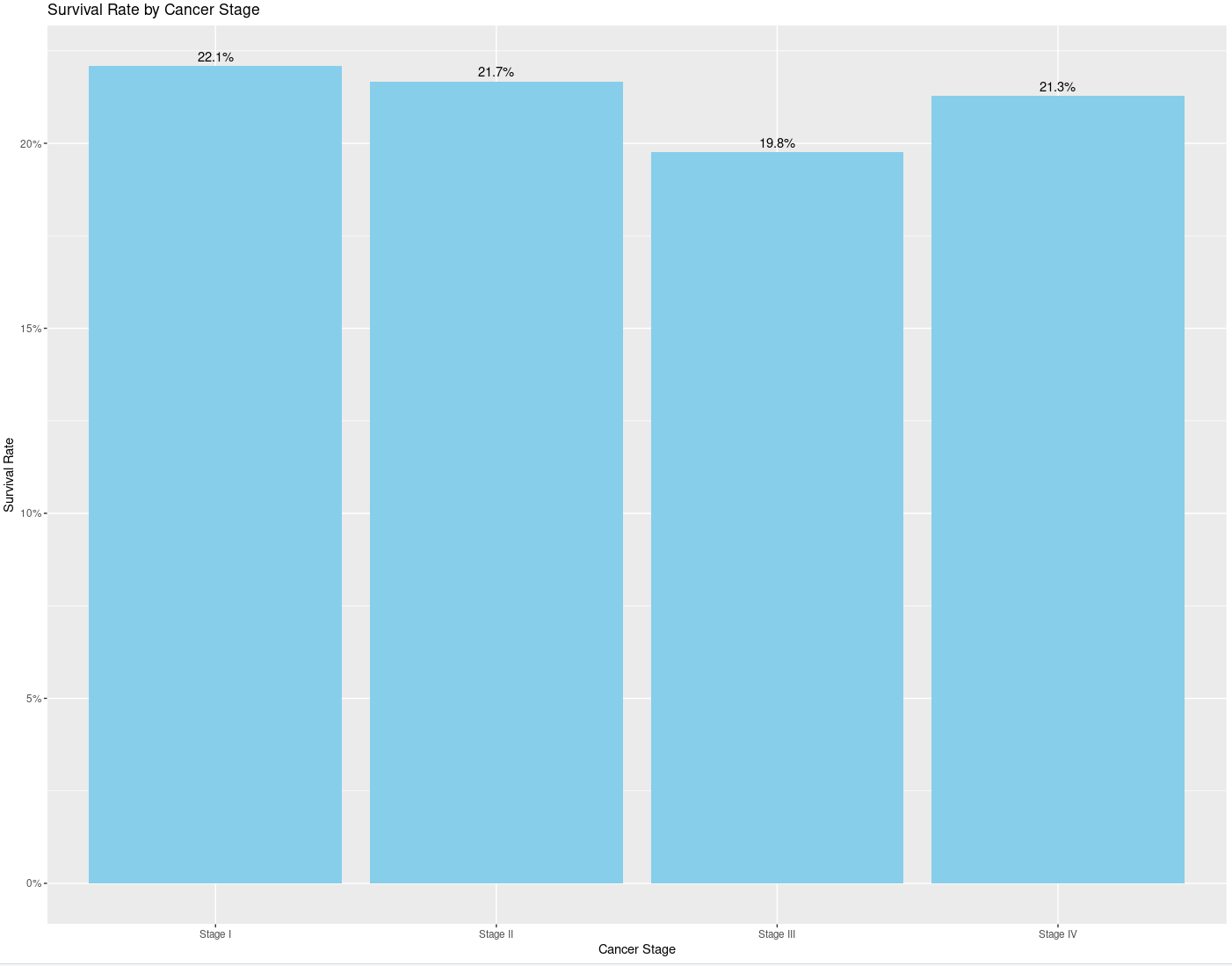
**Figure 5:** Outliers check (Model 1 & 2)

****

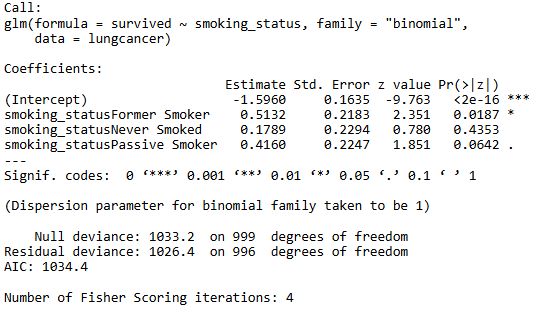
**Figure 6.1**: Survival Rate by Treatment Type



**Table 6.2**: Survival Rate by Cancer Stage



**Table M1**: Bar Chart – Survival Rate by Smoking Status



**Table M2**: Coefficient Plot – Cancer Stage and Treatment Type

