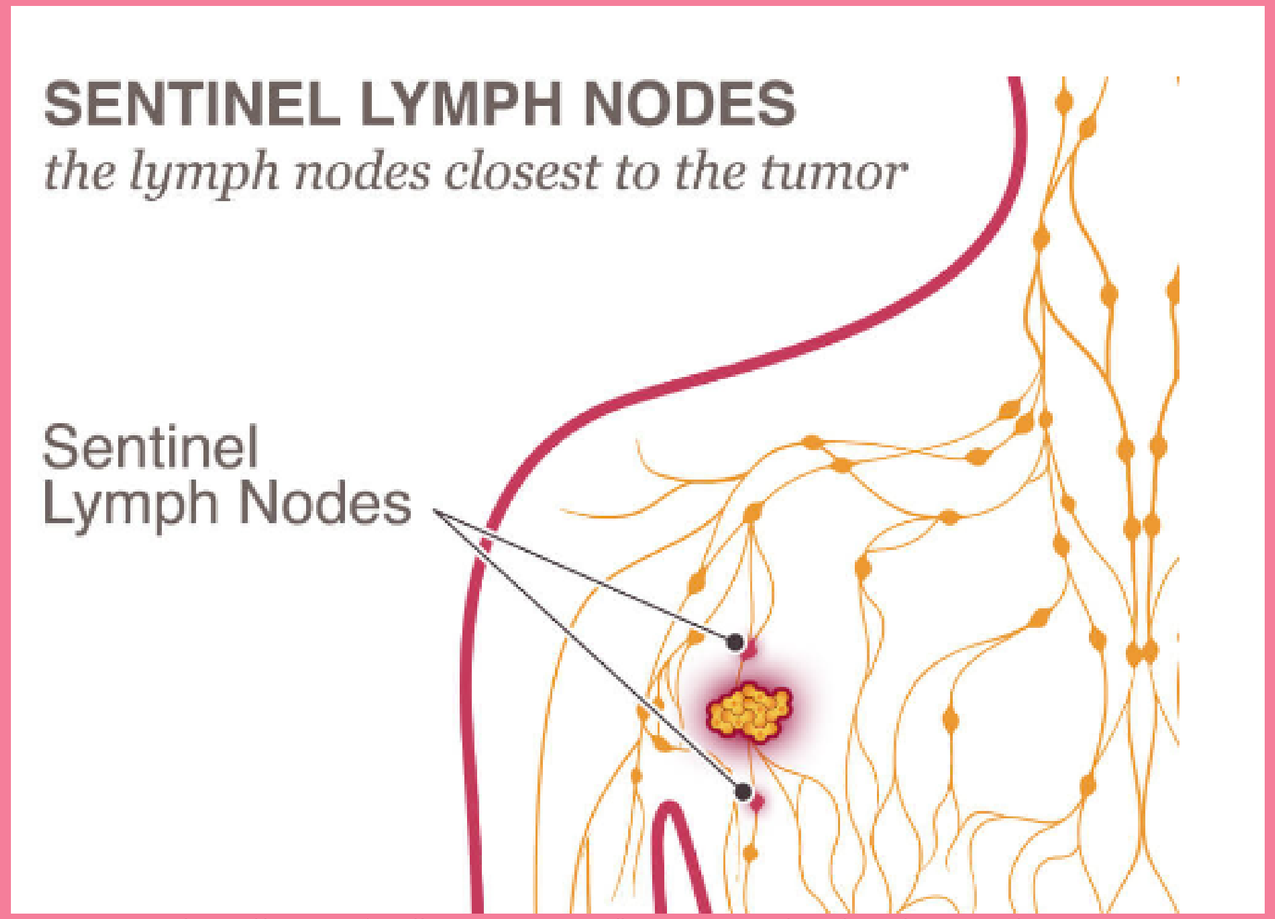
SHURAYA CHOUDHURY

CANCER SURVIVAL ANALYSIS

Background

- A small bean-shaped structure that is part of the body’s immune system.
- Lymph nodes filter substances that travel through the lymphatic fluid
- They are connected to one another by lymph vessels.
- Clusters of lymph nodes are found in the neck, axilla (underarm), chest, abdomen, and groin.

https://www.cancer.gov/publications/dictionaries/cancer-terms/def/lymph-node



https://www.nationalbreastcancer.org/breast-cancer-lymph-node-removal

Objective

Analyze the survival of patients who have undergone breast cancer surgery to predict whether a patient will survive five years or longer after the surgery.

Data Description

Data Set Information:

The dataset contains cases from a study that was conducted between
1958 and 1970 at the University of Chicago's Billings Hospital. There are 306 observations.

Attribute Information:

1. Age of patient at time of operation
2. Patient's year of operation
3. Number of positive axillary nodes detected
4. Survival status

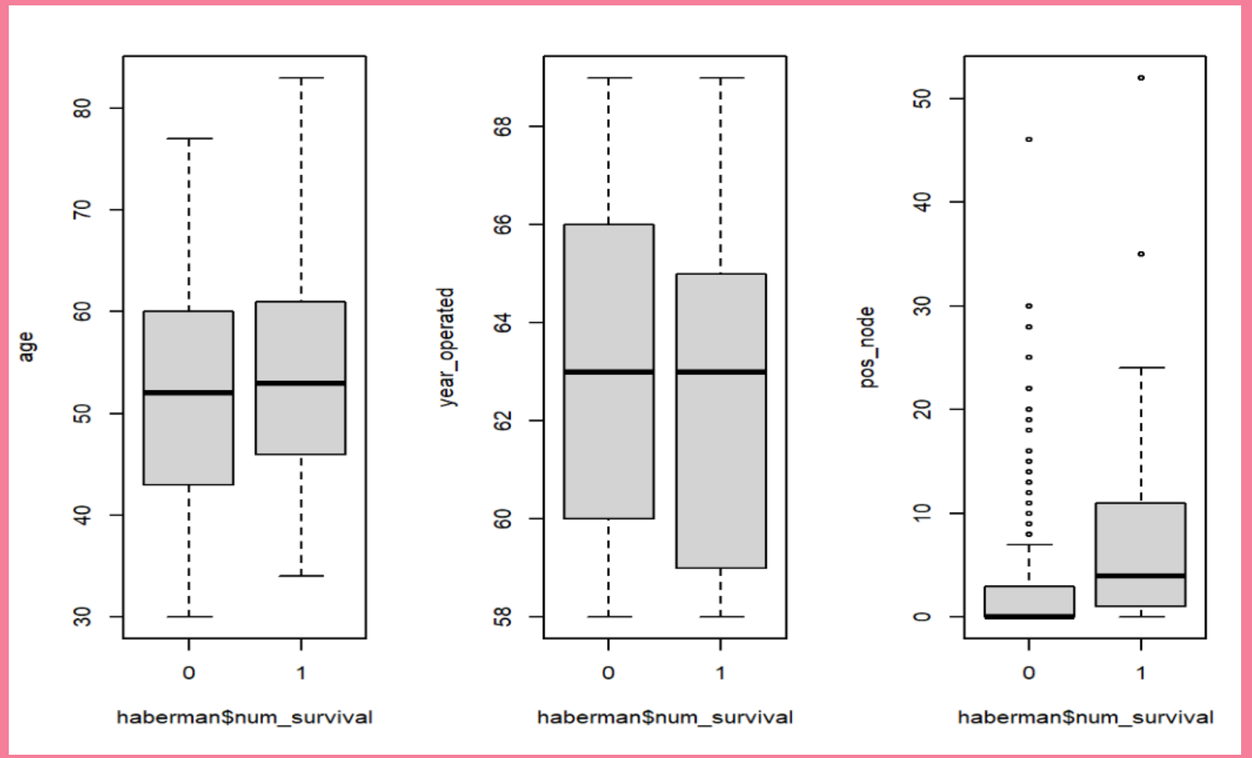
-- 0 = the patient survived 5 years or longer
-- 1 = the patient died within 5 year

Data Analysis

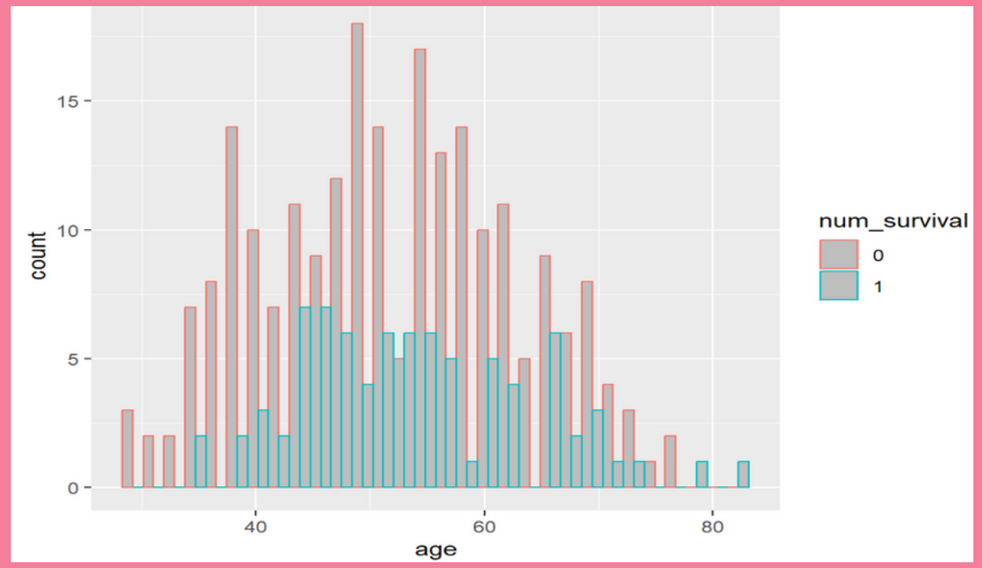
Data Preprocessing

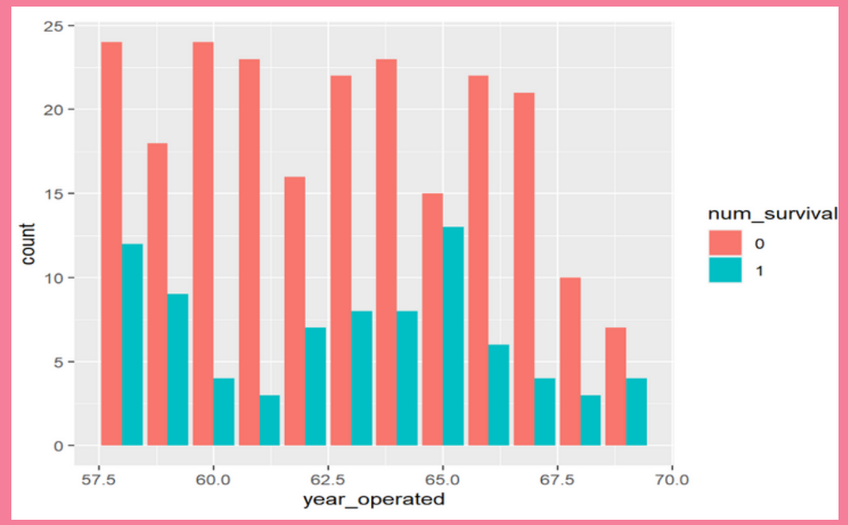
- The original .data file was converted to a CSV file.
- The columns were named after their attribute.
- Survival status was changed to 0 and 1. Patient survived 5 years or longer was represented by 0 and patient dieing within five years was represented by 1.
- There was no NA values.

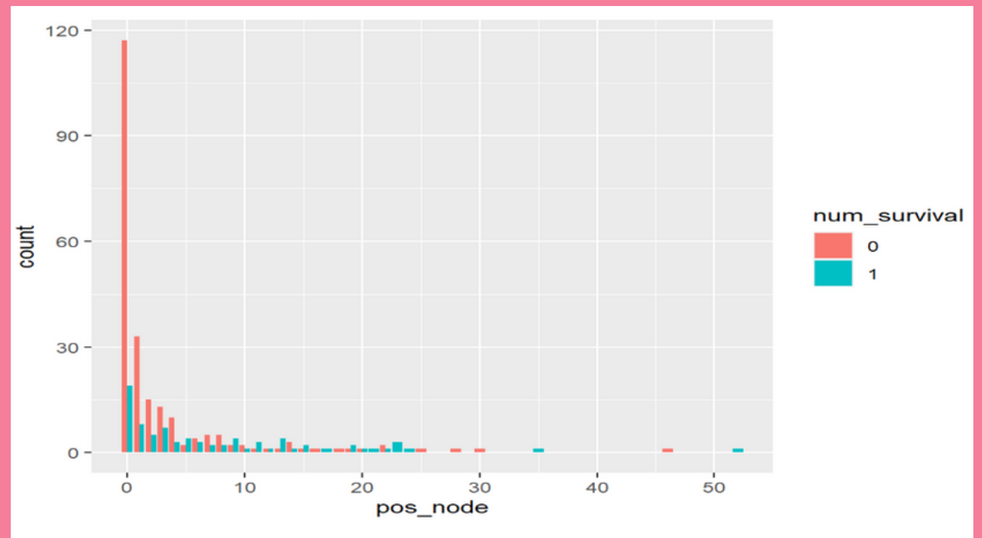
Boxplot



Descriptive Statistics







Methodologies

$\eta = \beta 0 + \beta 1(age) + \beta 2\ (positive\ nodes) + \beta 2\ (operation\ year)$

Logistic Model

$\eta = \log(p/(1-p))$

Probit Model

$\eta = \Phi^{-1}(p)$
where Φ^{-1} is the inverse normal CDF

Cauchit Model

$\eta = \arctan(\pi(p - 1/2))$

Logistic Model

$$\eta = -1.862 + 0.020 * \text{age} + 0.088 * \text{positive node} + -0.010 * \text{operation year}$$

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.861625	2.675197	-0.696	0.487	
age	0.019899	0.012735	1.563	0.118	
year_operated	-0.009784	0.042013	-0.233	0.816	
pos_node	0.088442	0.019849	4.456	8.36e-06	***

Logistic Model

Test

H0 : θ = (β1,⋯ ,βq) = 0 against
H1 : θ != 0, D0 −D ~ χ2q

D0 −D = 353.69 −328.26= 25.43

Degrees of Freedom = 3

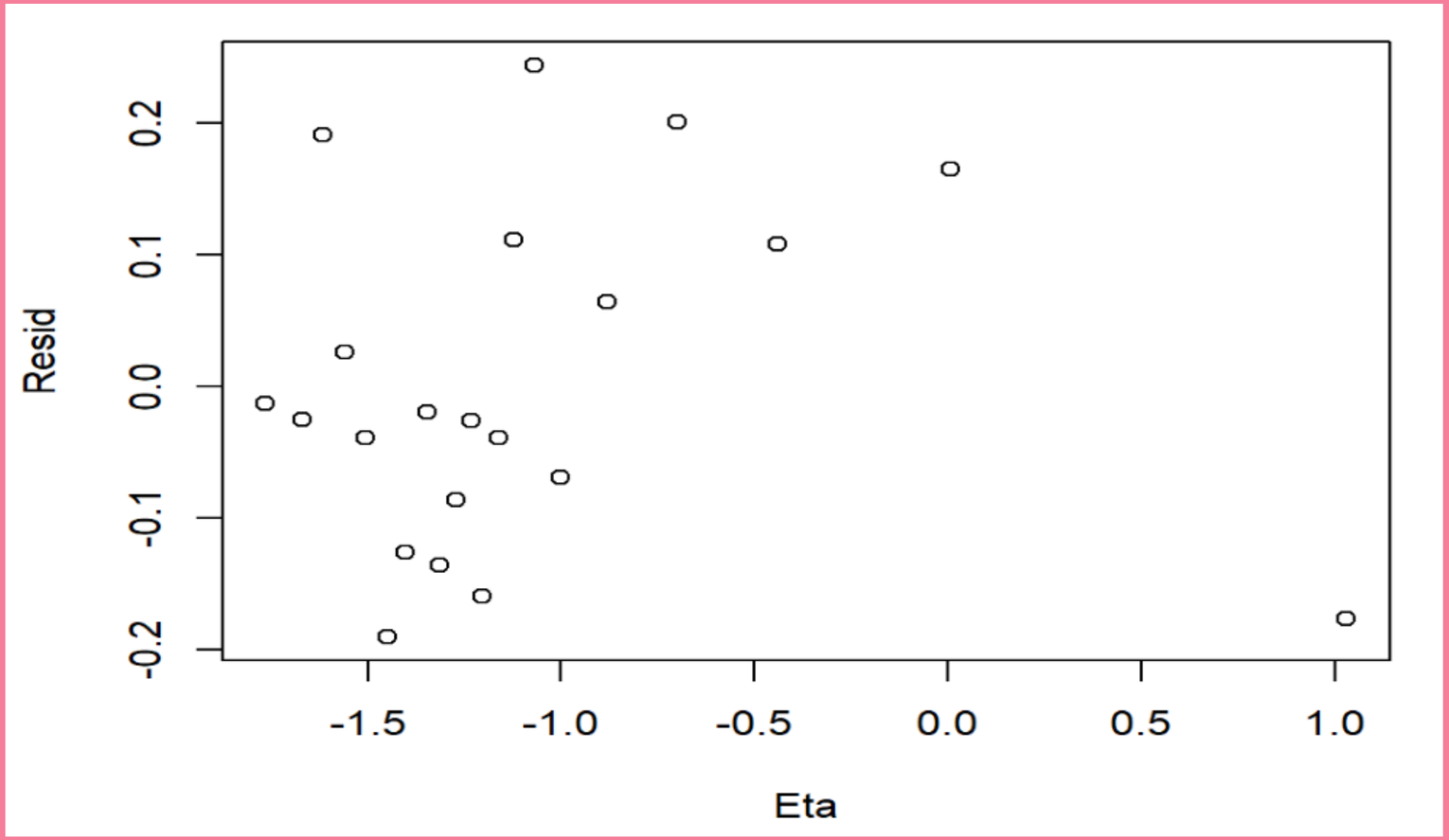
p-value = 1.255242e-05

Confidence Interval

	2.5 %	97.5 %
(Intercept)	-7.137314371	3.37989869
age	-0.004979245	0.04510314
year_operated	-0.092569105	0.07260941
pos_node	0.051293040	0.12925848

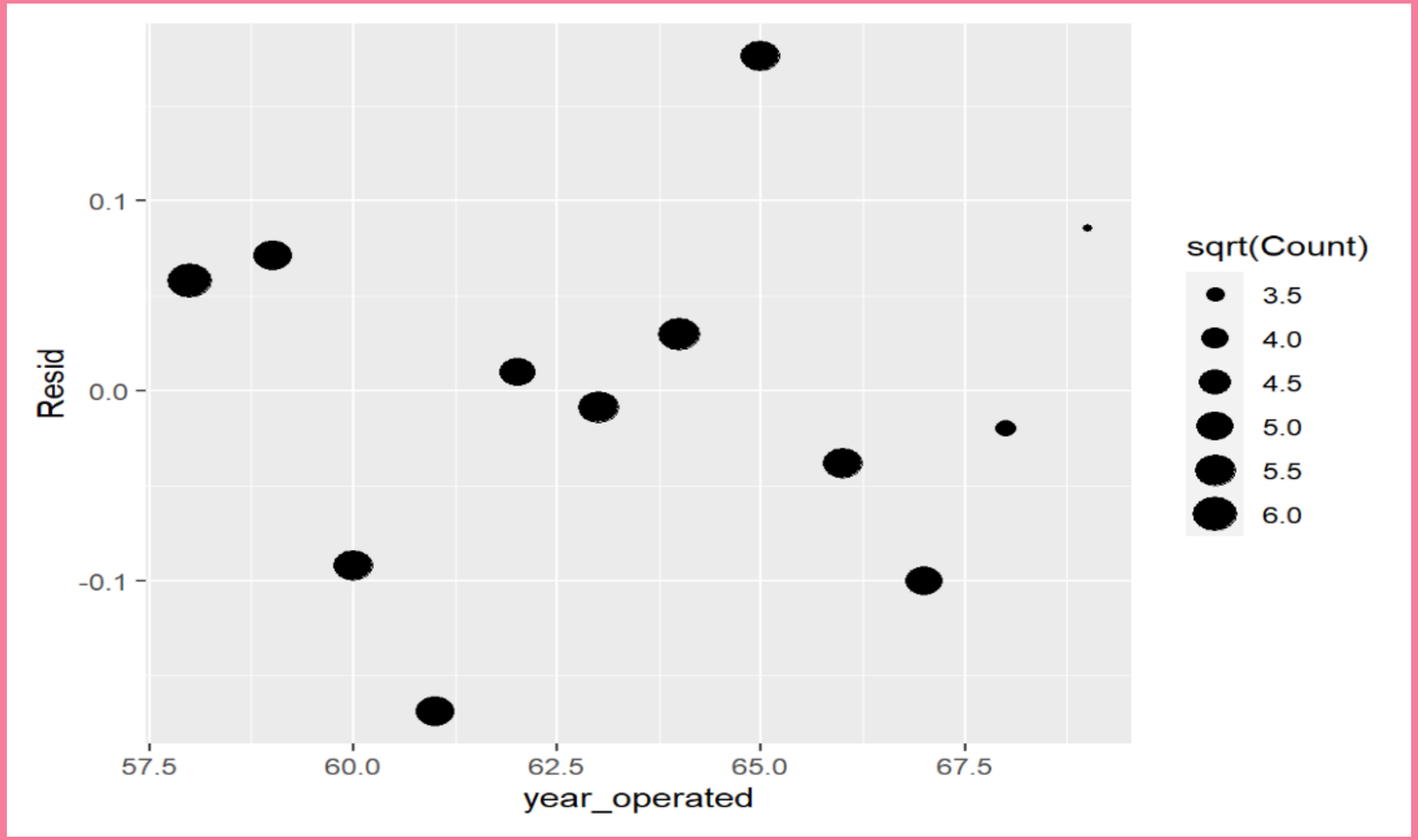
Diagnostics

Binned Residuals by η



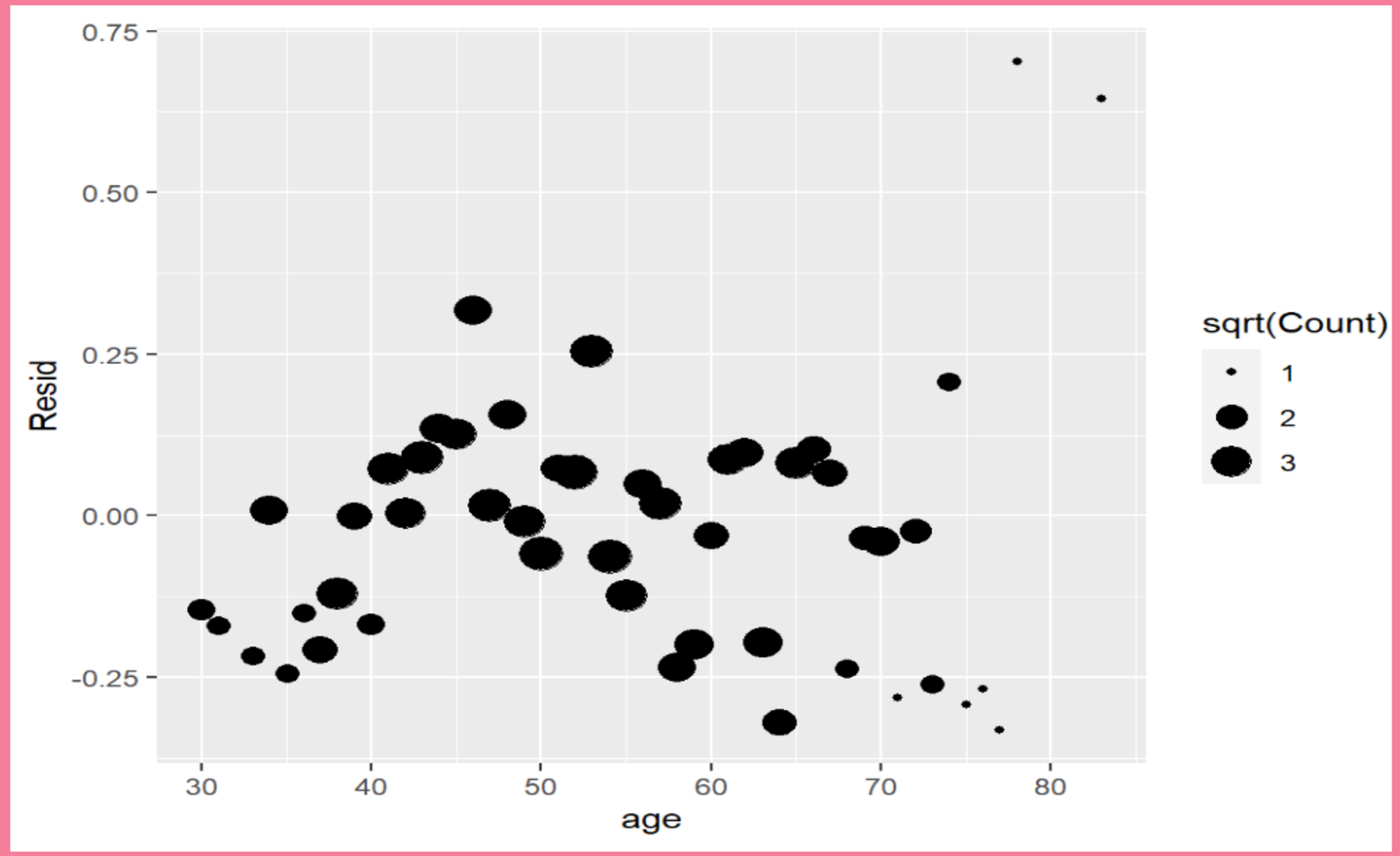
Diagnostics

Binned Residuals by Operation year



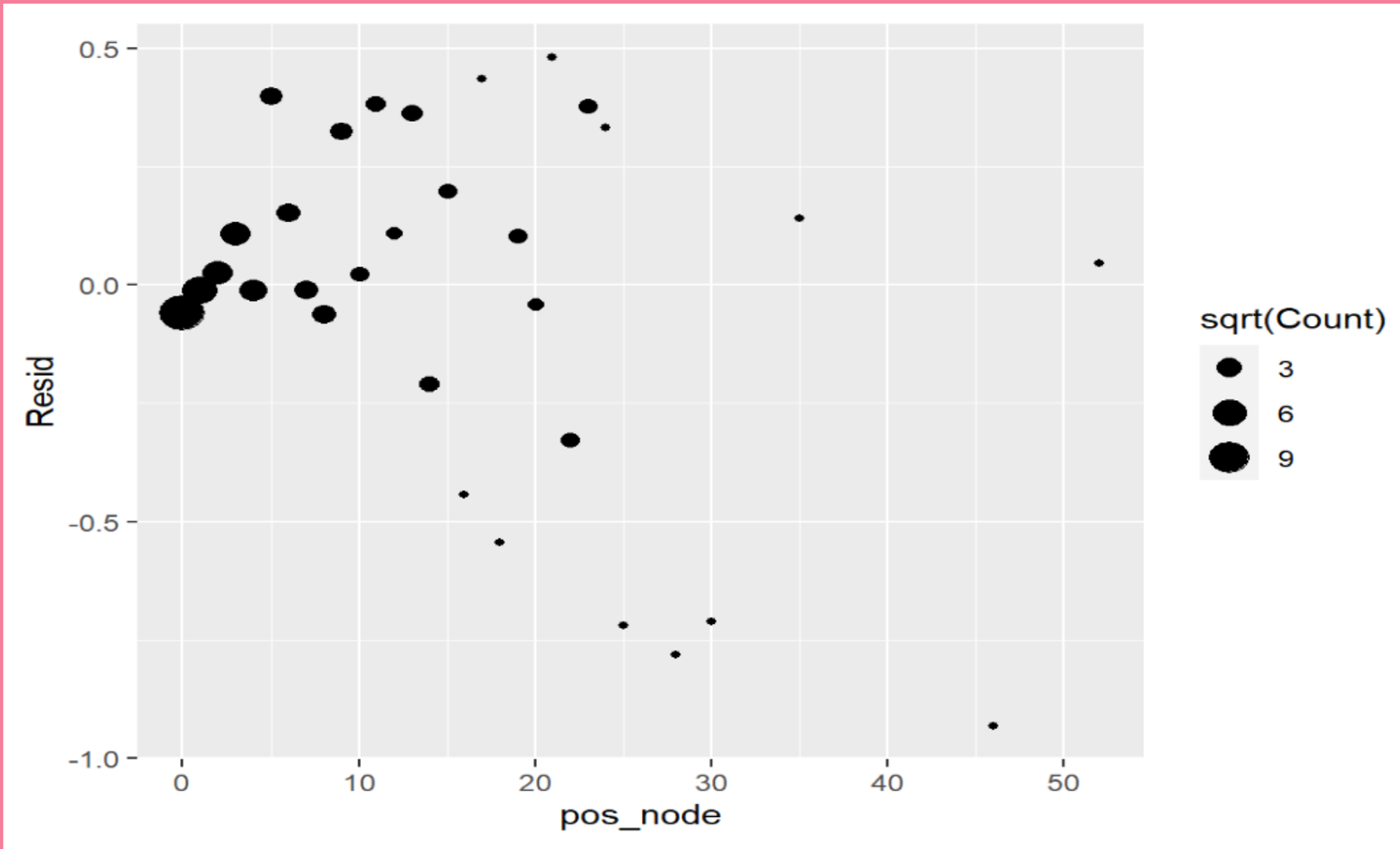
Diagnostics

Binned Residuals by age



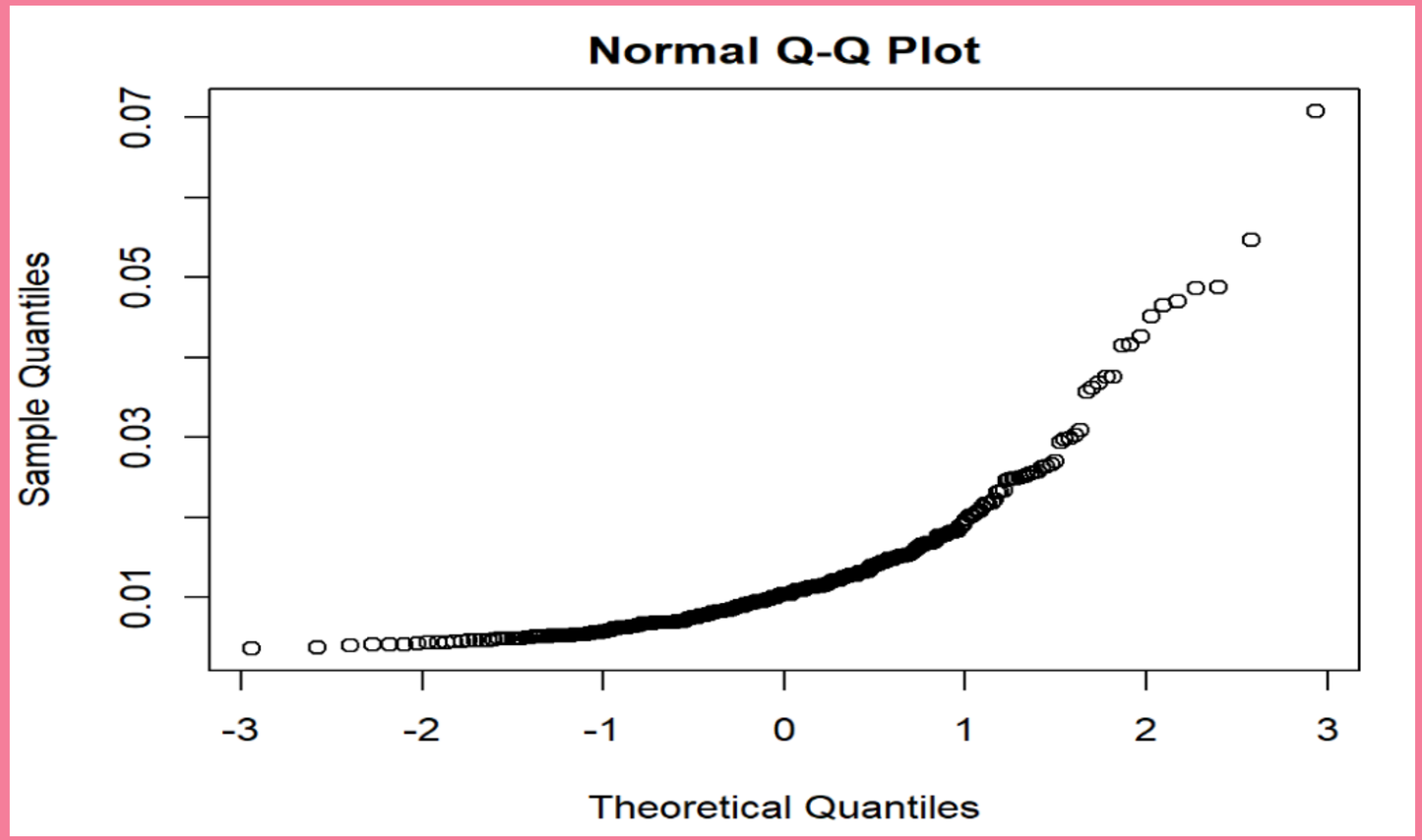
Diagnostics

Binned Residuals by Positive Nodes



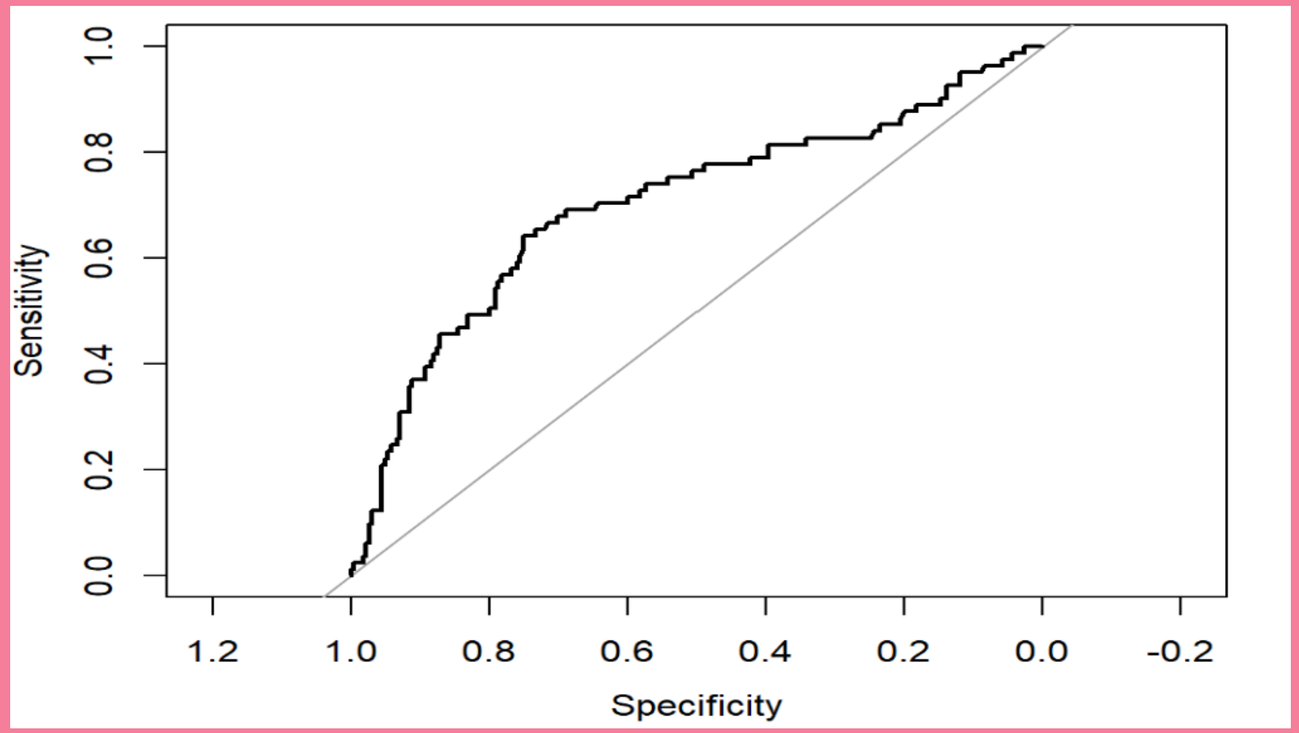
Diagnostics

Plot of hat values



Goodness of Fit

ROC Curve



The area under the curve is 0.70. The area under the curve (AUC) quantifies the model classification accuracy.

Model Comparison

Goodness of Fit

	Area Under the Curve (AUC)
Logit Model	0.7065
Probit Model	0.7037
Cauchit Model	0.6908

Conclusion

Analysis for patients who had undergone breast cancer surgery was conducted using a logistic model. For goodness of fit comparison, link function: logit, probit and cauchit were utilized to see the AUC.

Future Research Suggestion:

- There are only 306 observations so collecting more data will be useful.
- For possible future improvemnts, one can use other classification methods like KNN to test for a better model.

Thank you!