Associations Between Tumour Characteristics and Tumour Size in Early Breast Cancer Patients

A Regression Approach

Brian Xue, Sami Ahmed, Nzube Ogbu, Matthew Inkim, Cyrus Chung

STA302: Methods of Data Analysis I
Larry Dong
August 15, 2025

Department of Statistical Sciences
University of Toronto

Contents

1 Contributions

1.1 Brian Xue

- R code for exploratory data analysis, and the model(s).
- R code for diagnostic charts, and coloured scale-factor plots.
- Checked assumptions using the R plots, assisted in applying necessary mitigations.
- Worked on Section 2 and Section 4.

1.2 Sami Ahmed

- Worked on finding the necessary literature to help with our modelling questions.
- R code for backtransformation, tumour size vs Ki67, tumour size vs number of metastases plots.
- Checked assumptions using the R plots.
- Helped format the LaTeXdocument. Assisted with sections 2 and 4 of the report. Worked on Section 3.

1.3 Nzube Ogbu

- Worked on prelimary findings of the report.
- Worked on outliers, cook's distance, and DFBETAs
- Assisted with Section 3 and 4.

1.4 Matthew Inkim

- Worked on section 2 and section 5.
- Proofread and edited sections of the report.

1.5 Cyrus Chung

- R code for exploratory data analysis, and the model(s).
- Proofread and edited sections of the report.

2 Introduction

Breast cancer is a leading cause of cancer-related mortality worldwide, and understanding factors associated with tumour size is crucial for prognosis, treatment planning, and insights into disease biology. Tumour size can be influenced by tumour-intrinsic features, such as growth rate and molecular characteristics, and patient-specific factors, including age and overall health.

This study examines whether tumour size can be described using patient age, number of lymph node metastases, Ki67 index, Estrogen Receptor (ER)/Progesterone Receptor (PR)/Human Epidermal Growth Factor Receptor 2 (HER2) status, tumour type, and molecular subtype in Early Breast Cancer (EBC) patients with clinically negative Axillary Lymph Nodes (ALN).

Hypothesis: Ki67 is a biomarker used to assess the rate of cell division in cancer cells. We assume that a higher Ki67 value results in larger tumours.

Previous research supports the plausibility of these relationships. [3] linked cancer cell size to systematic shifts in protein abundance, suggesting size-related molecular signatures. [2] showed that tumour size influences the concentration of tumour-derived proteins in plasma in a nonlinear manner, indicating that measurable tumour characteristics reflect underlying biological changes. [4] demonstrated that specific circulating proteins increase consistently as tumours grow, highlighting the role of measurable biological markers in tracking tumour progression.

Given that tumour size and the selected predictors are quantitative or categorical variables, linear regression is an appropriate tool to evaluate these relationships. This approach allows for clear interpretation of how each factor is associated with tumour size while adjusting for others, with emphasis on interpretability and description rather than pure prediction.

3 Methods

3.1 Methodology Flowchart

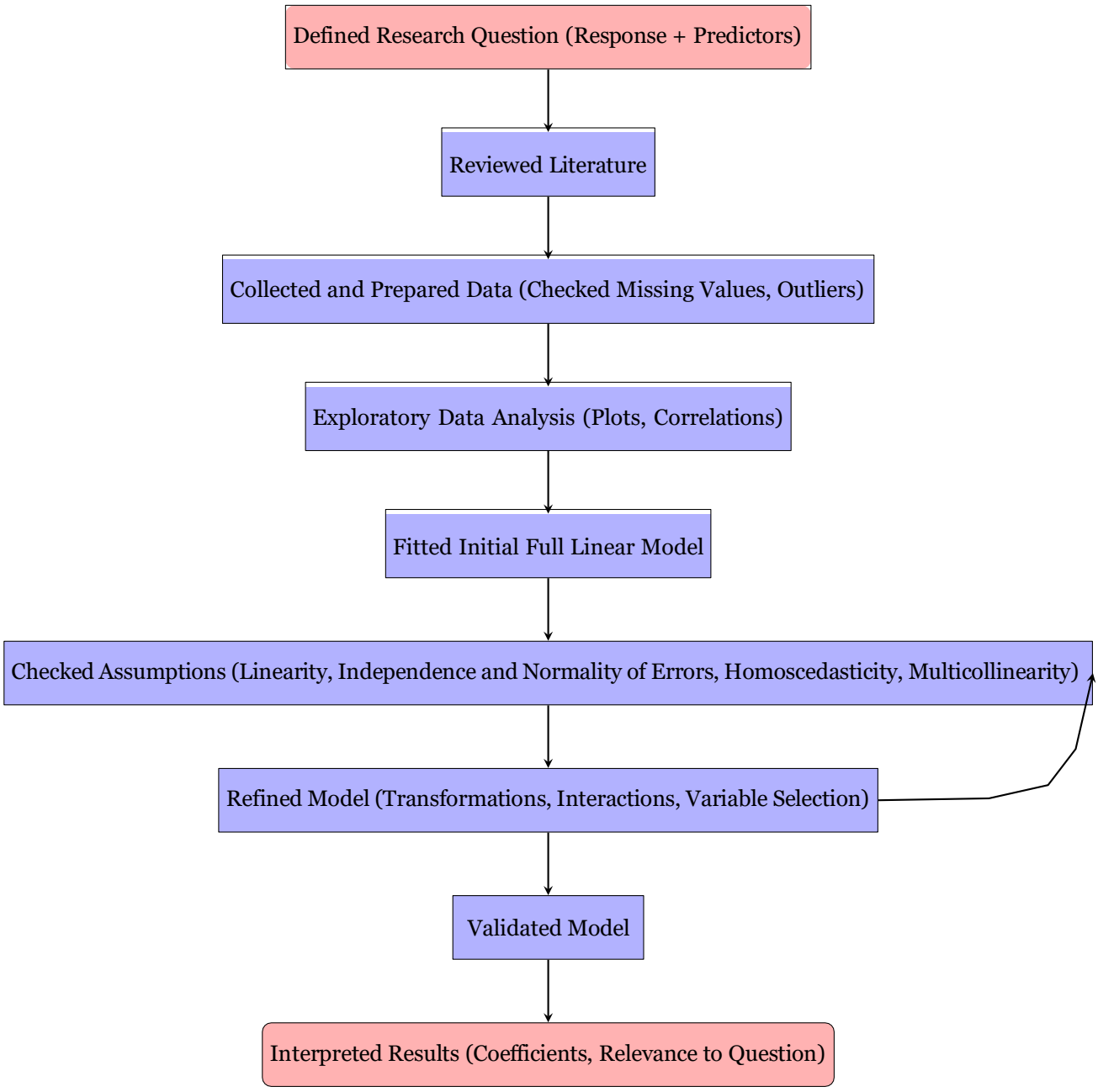

Figure 1: Flowchart of study methodology

3.2 Data Preparation and Exploratory Analysis

Data was obtained from the "Early Breast Cancer Core-Needle Biopsy WSI Dataset" [9] used in [8], which developed a deep learning-based primary tumour biopsy signature to predict ALN metastases in EBC patients with clinically negative ALN. After exclusions, 1058 patients were analyzed.

We treated tumour size (cm) as a continuous response variable, while predictors included both continuous (age, Ki67, number of metastases (META)) and categorical variables (ER/PR/HER2/, tumour type (TT), molecular subtype (MST)).

Continuous predictors were assessed for missing values, distributional characteristics, and outliers. Summary statistics and visualisations (scatterplots for continuous variables, and boxplots for categorical predictors) were used to assess initial associations with tumour size. Pearson correlations and one-way ANOVAs were used to preliminarily explore relationships between predictors and response.

Skewness was examined for continuous variables to detect deviations from normality, and correlations between predictors was assessed to anticipate multicollinearity in the model.

3.3 Initial Model Specification

An initial multiple linear regression (MLR) model fit and included all the predictors. This served as the baseline for diagnostics and refinement. The focus was on interpretability to determine the independent contribution of each predictor, particularly Ki67, to tumour size. Giving:

$$Y_{\sqrt{\text{tumour size}}} = \beta_{\text{META}}x_1+\beta_{\text{age}}x_2+\beta_{\text{Ki67}}x_3+\beta_{\text{ER}}x_4+\beta_{\text{PR}}x_5+\beta_{\text{HER2}}x_6+\beta_{\text{TT}}x_6+\beta_{\text{MST}}x_7+\epsilon$$

Dummy variables were used for categorical variables, and reference levels were chosen based on relevance to ensure interpretable coefficients. Continuous predictors were centered to facilitate interpretation of the intercept and reduce potential multicollinearity.

3.4 Assumption Checks

The standard regression assumptions were then evaluated. We looked at residuals versus fitted plots for linearity, scale-location plots for homoscedasticity, Q-Q plots for normality, and Variance Inflation Factor (VIF) for multicollinearity. If assumptions were violated, appropriate mitigation strategies were implemented, including polynomial terms for non-linear trends, and centering or standardizing predictors.

3.5 Outliers and Influential Observations

Outlier and influence diagnostics were conducted using Cook's Distance, leverge and DFBETAs. Potential outliers were flagged but not immediately removed, instead

kept for further review. The criteria for being flagged were *leverage* $> \frac{2p}{n}$, cooks distance > 1, and DFBETA > 1. Graphical representations and influence pl ots were used to visualise these points.

3.6 Model Refinement

As stated above, transformations of the response and/or predictors were aplied when assumption violations were detected. Interaction terms(Ki67 x MST, ER x PR etc.), and polynomial terms were explored when suggested by residual patterns. Model refinement was guided by statistical criteria and clinical rationale; nested models were compared using ANOVA F-tests, and non-nested models were evaluated with AIC and BIC to identify the ”best” model. This process was iterative, with assumption checks repeated after each modification to ensure the final model satisfied linear regression requirements.

3.7 Validation

The final model was assessed on whether all assumptions were met, the amount of high-influence points was reasonable, $VIF \leq 5$, interpretability, and metrics such as R^2 and standardised residuals.

3.8 Interpretation

Final model coefficients and confidence intervals were interpreted to determine the independent effects of each predictor on tumour size. Particular attention was given to Ki67 as a biomarker of cell proliferation. The clinical relevance of the coefficients was considered, and potential limitations, including sample size, measurement error, and generalizability to other populations, were discussed when drawing conclusions.

4 Results

The response variable in this model is the diameter of tumours. Initially, the tumour size exhibited a mean of 2.235 cm, a standard deviation of 0.863 cm, and a minimum of 0.5 cm, suggesting a right-skewed distribution.

Since MLR assumes that the residuals of the response variable are normally distributed, this skewness raises concerns about the model. A square root transformation (SRTS) was applied to the tumour size , resulting in a more symmetrical distribution with a reduced mean of 1.467 and a standard deviation of 0.290. SRTS is continuous, unbounded, and approximately normal after transformation. Hence, suitable for MLR.

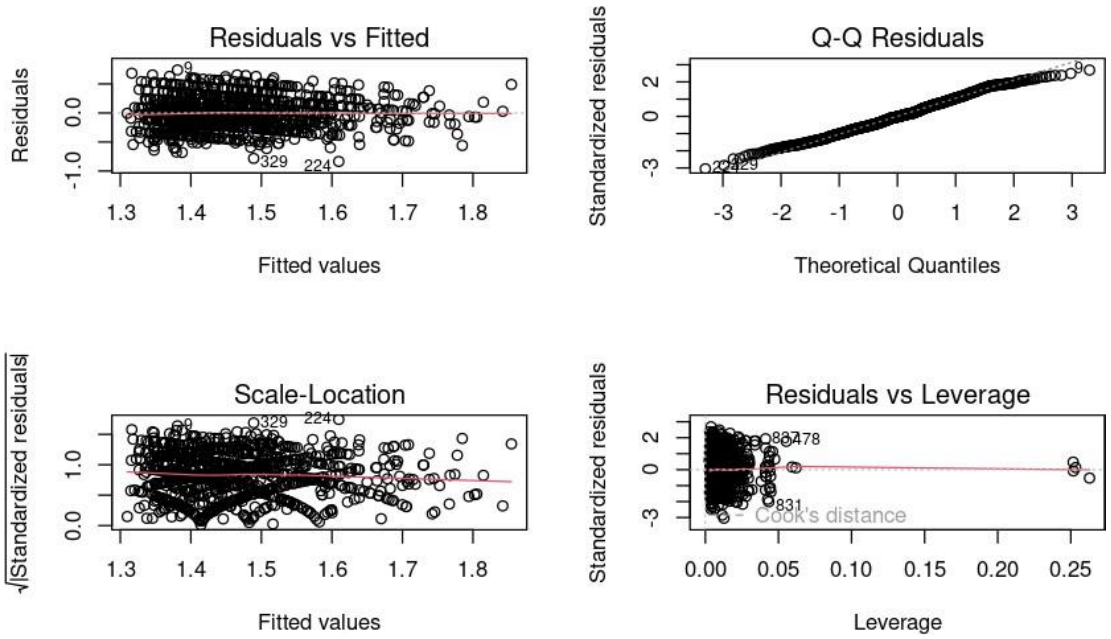

Figure 2: (a) Residuals versus fitted (b) Q-Q Residuals (c) Scale-Location (d) Residuals versus leverage - The diagnostic plots after the transformation

The initial residuals vs. fitted plot showed a downward trend, indicating missing non-linear effects. Applying SRTS produced a more random scatter around zero. Although a faint trend remains, it reflects dataset characteristics (see section 5.1), so the linearity assumption is reasonably satisfied.

Before transformation, the Q-Q plot showed a faint S-shape with tail deviations, indicating skewed residuals. Applying SRTS stabilized variance and reduced skewness, and aligned residuals closer to the 45° line, thereby satisfying the normality assumption and supporting valid inference in the linear regression model.

The scale-location plot of the initial model (see appendix Figure 5) shows even variance of standardized residuals across fitted values, although obvious parabolic patterns are visible. These curves suggest that certain subsets of observations share

similar fitted values and residual spreads, which may occur when categorical predictors produce grouped patterns in the data. To investigate, scale-location plots were generated with points colored by each categorical predictor (HER2, ER, PR, Tumour Type, and Molecular Subtype; Figure 3).

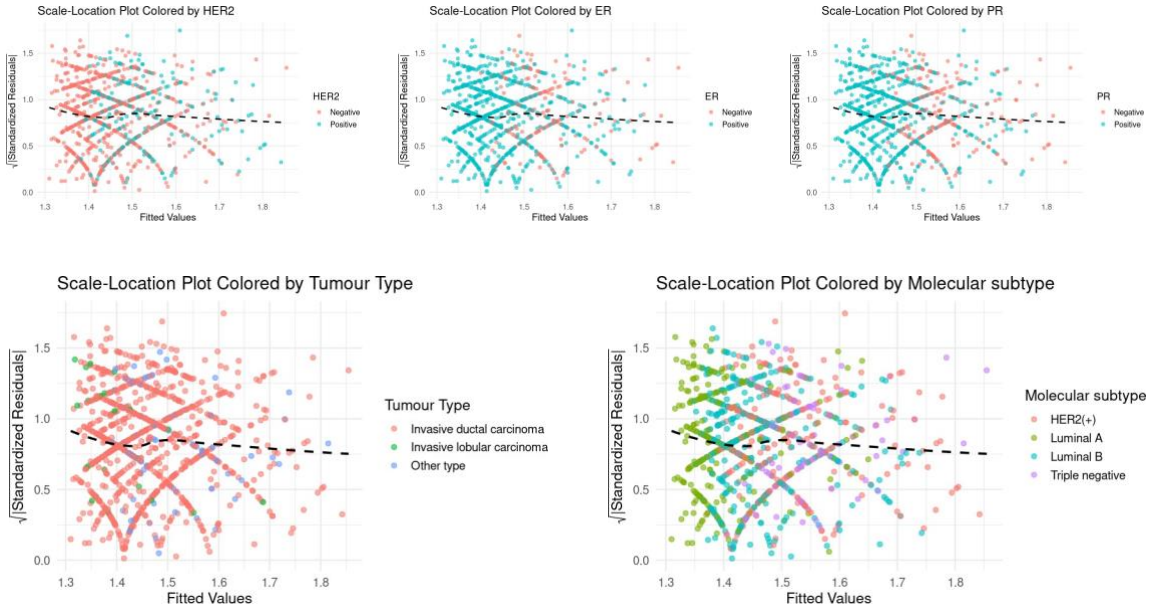

Figure 3: Colour Coded Scale-Location Plots, filtered by HER2, ER, PR, TT, and MST

In the coloured plots, no category displayed a systematic deviation from the horizontal trend line, and the variance appeared comparable across levels. The observed parabolas were roughly symmetric in the sense that if the plots were flipped horizontally, the residual distribution would appear similar, indicating no directional heteroscedasticity related to fitted values.

Given that neither the original nor the color-coded plots reveal substantial issues, the homoscedasticity assumption is considered to be reasonably met. The structured patterns are likely due to hidden predictors or some interaction terms that are not available in the dataset, rather than violations caused by included variables.

There are 74 observations with high leverage. However, there are 0 observations with a DFBETA greater than the threshold of 1, nor are there any observations with a Cook's Distance exceeding 1. This would suggest that any existing leverage point is a "good leverage point", not likely to influence the model, negatively or otherwise. Thus, we can safely assume an absence of outliers, per the decided upon criteria. (see appendix figure 6)

Predictor	**Sum Sq**	**Df**	F **value**	$\Pr(>F)$
Type II ANOVA				
Age (years)	0.292	1	3.8386	$0.0504^{\cdot}$
Ki67	0.825	1	10.8372	0.0010^{**}
ER	0.003	1	0.0401	0.8413
PR	0.361	1	4.7414	0.0297^{*}
HER2	0.000	1	0.0003	0.9869
Tumour Type	0.304	2	1.9952	0.1365
Number of lymph node metastases	4.986	1	65.5002	$1.64 \times 10^{-15***}$
Molecular subtype	0.025	3	0.1107	0.9539
Residuals	78.019	1025		
Type III ANOVA				
(Intercept)	6.484	1	85.1907	$< 2.2 \times 10^{-16***}$
Age (years)	0.292	1	3.8386	$0.0504^{\cdot}$
Ki67	0.825	1	10.8372	0.0010^{**}
ER	0.003	1	0.0401	0.8413
PR	0.361	1	4.7414	0.0297^{*}
HER2	0.000	1	0.0003	0.9869
Tumour Type	0.304	2	1.9952	0.1365
Number of lymph node metastases	4.986	1	65.5002	$1.64 \times 10^{-15***}$
Molecular subtype	0.025	3	0.1107	0.9539
Residuals	78.019	1025		

$^{\cdot}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

Table 1: ANOVA results for the model predicting $\sqrt{\text{Tumour Size (cm)}}$

The ANOVA (Table 1) and VIF (Table 2) results support a clear understanding of the predictors of tumour size. Ki67, representing the percentage of actively dividing tumour cells, is highly significant ($F = 10.84$, $p = 0.001$) and exhibits low multi-collinearity ($VIF \approx 1.30$), confirming its independent contribution to tumour size, consistent with the hypothesis that higher proliferation rates produce larger tumours.

Lymph node metastases, indicating cancer spread to axillary nodes, are the strongest predictor ($F = 65.50$, $p < 0.001$) and show minimal multicollinearity ($VIF \approx 1.01$), aligning with literature demonstrating their correlation with tumour size [5]. (see appendix figure 7 for comparison between the two)

Age and PR are modestly associated ($F \approx 3.84$ and 4.74, $p \approx 0.05$ and 0.03) and show low VIFs (< 2), indicating reliable estimation. ER shows no significant effect ($F = 0.04$, $p = 0.84$), while HER2 and molecular subtype have high VIFs (≈ 7.16 and 2.35), suggesting moderate-to-high collinearity, and tumour type is non- significant ($F = 1.99$, $p = 0.14$).

Overall, the model supports that Ki67, and lymph node metastases are key continuous predictors of tumour size, while categorical biomarkers require cautious interpretation due to potential collinearity.

Predictor	GVIF	Df	$GVIF^{1/(2\times Df)}$
Age (years)	1.047	1	1.023
Ki67	1.686	1	1.299
ER	4.047	1	2.012
PR	3.017	1	1.737
HER2	51.214	1	7.156
Tumour Type	1.047	2	1.012
Number of lymph node metastases	1.020	1	1.010
Molecular subtype	166.991	3	2.347

Table 2: Variance Inflation Factor (VIF) for predictors in the tumour size model

Figure 4 illustrates the predicted tumour size across the observed range of Ki67, while holding all other predictors at typical values. The highlighted region represents ±10% increase or decrease in Ki67 and how that translates to a 0.05cm increase or decrease in tumour size respectively.

Tumour size was back-transformed from the square-root scale used in the model to allow straightforward interpretation. The graph addresses the research question by illustrating how increases in Ki67 correspond to larger tumours, supporting the hypothesis that higher proliferation rates (higher Ki67) drive tumour growth.

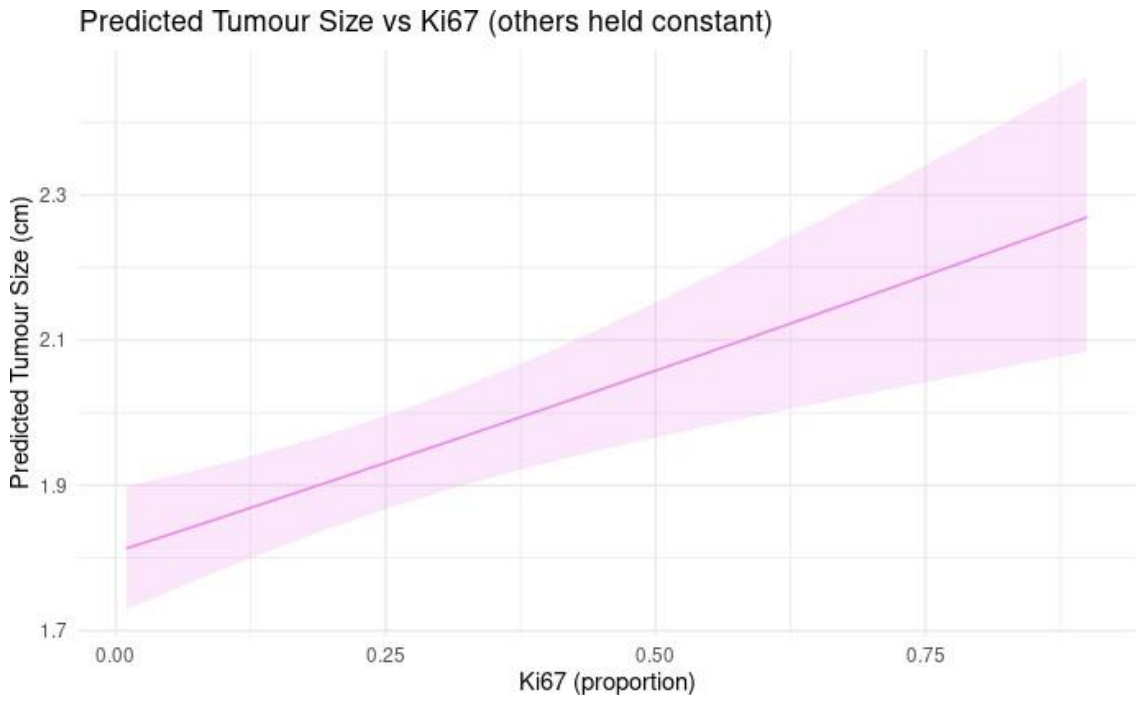

Figure 4: Predicted Tumour Size vs Ki67, other predictors held constant

5 Conclusion and Limitations

5.1 Interpretation of Final Model

The final linear model demonstrates a significant positive association between Ki67 and predicted tumour size, indicating that higher proliferation rates correspond to larger tumours. Predicted tumour size was back-transformed from the square-root scale to maintain interpretability in centimetres, ensuring that model predictions align with observed units. While ER, PR, and HER2 were included as predictors, their effects were less pronounced, consistent with prior work showing correlations between receptor status and tumour biology [1]. Including these markers enhances interpretability without dominating the model, and captures variations linked to tumour type and size. [6]

5.2 Research Question Discussion

Interaction effects were not considered, which may obscure combined effects of predictors. Discordance in receptor status [7] introduces variability, contributing to residual trends. Extreme values and underrepresentation of some tumour subtypes reduce generalizability. The model assumes linear additive effects and cannot capture non-linear or complex interactions in tumour biology.

Our model directly addresses the research question by quantifying how Ki67, age, lymph node metastases, tumour type, molecular subtype, and receptor status collectively relate to tumour size. The strong effect of lymph node metastases is consistent with literature highlighting their role in tumour burden. Ki67 emerges as a key predictor, supporting its use as a proliferation biomarker. The faint parabolic trend in the scale-location plot may reflect partial receptor discordance between primary tumours and metastatic sites [7], indicating inherent biological variability rather than model mis-specification.

5.3 Limitations of Analysis

Several limitations affect the robustness and generalizability of the model. Interaction effects were not considered, which may obscure combined effects of predictors. Omitted variable bias may arise if other biological or genetic factors influencing tumour size are missing. The dataset may not represent the broader population, being limited by geography, age, or clinical characteristics. Multicollinearity among predictors can inflate standard errors and destabilize coefficients. Our sample size is relatively small (just above a 1000), and tumour type is skewed toward invasive ductal carcinomas (90.45%), underrepresenting invasive lobular (2.41%) and other tumours (7.14%). Linear regression does not account for temporal dynamics as tumour size and predictors are measured at different points.

6 Appendix

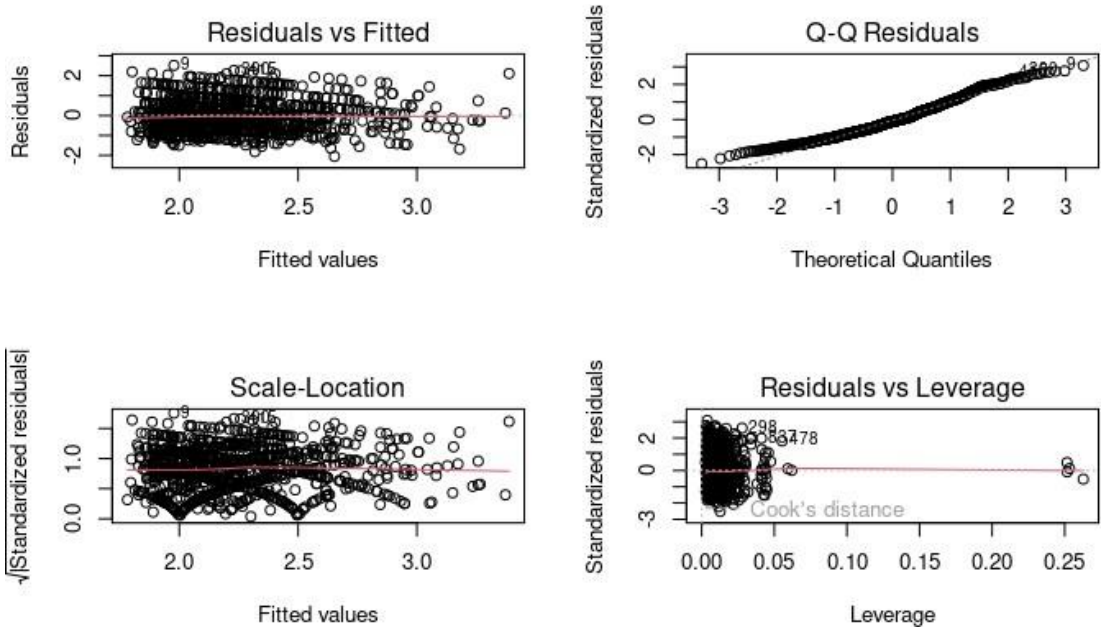

Figure 5: (a) Residuals versus fitted (b) Q-Q Residuals (c) Scale-Location (d) Residuals vs leverage - Diagnostic plots before the transformation

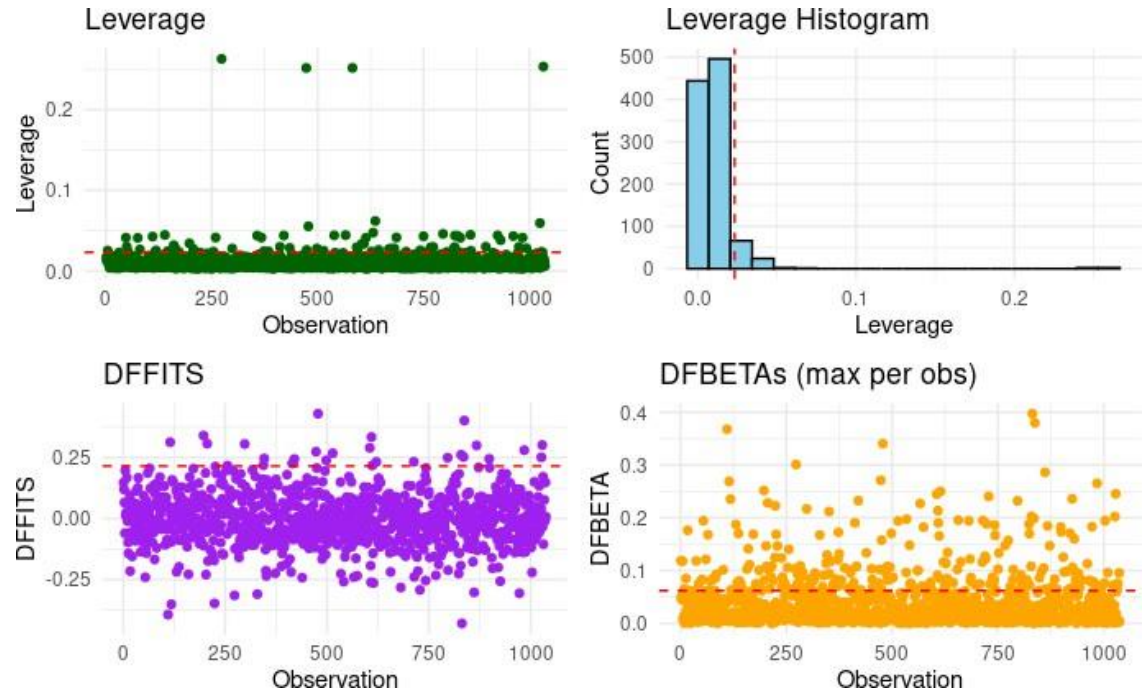

Figure 6: (a and b) Leverage and Leverage Histogram, (c) DFFITS (d)DFBTAs

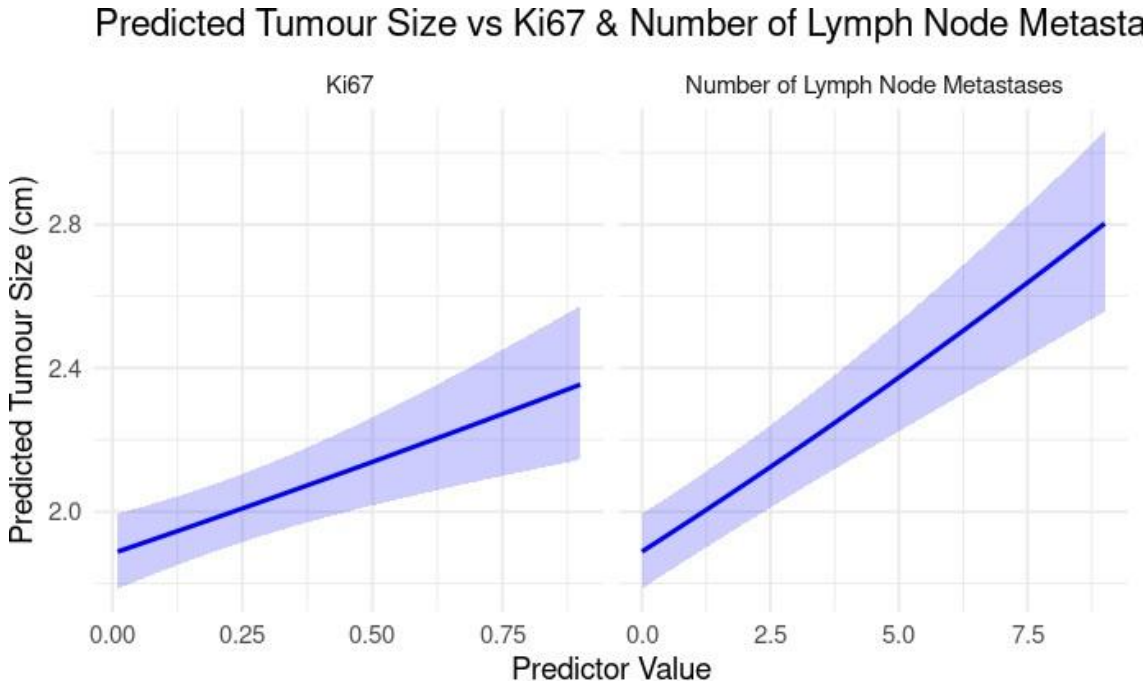

Figure 7: (a) Predicted tumour size vs Ki67 (b) Predicted tumour size vs number of lymph node metastasis ([5])

7 Bibliography

References

[1] Anna M Badowska-Kozakiewicz et al. "Immunohistochemical evaluation of human epidermal growth factor receptor 2 and estrogen and progesterone receptors in invasive breast cancer in women". In: *PubMed* (2013). Validation of IHC methods for receptor status. URL: https://pubmed.ncbi.nlm.nih.gov/23847668/.

[2] Qiaojun Fang et al. "Impact of Protein Stability, Cellular Localization, and Abundance on Proteomic Detection of Tumor-Derived Proteins in Plasma". In: *PLoS ONE* 6.7 (2011), e22691. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3146523/.

[3] Ian Jones et al. "Characterization of proteome-size scaling by integrative omics reveals mech- anisms of proliferation control in cancer". In: *Science Advances* 9.4 (2023), eadd0636. DOI: 10.1126/sciadv.add0636. URL: https://www.science.org/doi/10.1126/sciadv.add0636.

[4] Huiyan Li et al. "Serial Analysis of 38 Proteins during the Progression of Human Breast Tumor in Mice Using an Antibody Colocalization Microarray". In: *PLOS ONE* 10.7 (2015), e0130302. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4390249/.

[5] Seung Ki Min et al. "Relation Between Tumor Size and Lymph Node Metastasis According to Subtypes of Breast Cancer". In: *Journal of Breast Cancer* (2021), pp. 75–84. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7920868/.

[6] Jigisha Thakar and Divyes Mehta. *A Review of an Unfavorable Subset of Breast Cancer: Estrogen Receptor Positive Progesterone Receptor Negative*. Review on PR and ER/PR-positive breast cancers. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3228102/#sec5.

[7] Vincent Walter et al. "Estrogen, progesterone, and human epidermal growth factor receptor 2 discordance between primary and metastatic breast cancer". In: *PubMed* (2020). Evidence of ER/HER2/PR discordance; not studied here. URL: https://pubmed.ncbi.nlm.nih.gov/32613540/.

[8] F. Xu et al. "Predicting Axillary Lymph Node Metastasis in Early Breast Cancer Using Deep Learning on Primary Tumor Biopsy Slides". In: *Frontiers in Oncology* 11 (2021). Paper describ- ing DL-based prediction model. DOI: 10.3389/fonc.2021.75900. URL: https://doi.org/10.3389/fonc.2021.75900.

[9] Chuang Zhu. *Predicting Axillary Lymph Node Metastasis in Early Breast Cancer Using Deep Learning on Primary Tumour Biopsy Slides, BCNB Dataset*. https://github.com/bupt-ai-cz/BALNMP?tab=readme-ov-file. Dataset repository. Oct. 2021.