Statistical Modelling of Prostate Cancer Biomarkers: A Comparative Regression Analysis

Swagato Das
M.Stat - 1st Year
Roll No: MB2534
Indian Statistical Institute, Kolkata
Instructor : **Dr. Abhik Ghosh**

November 26, 2025

Abstract

This report presents a rigorous statistical analysis of prostate cancer data, aiming to model the Log Prostate Specific Antigen (`lpsa`) using various clinical covariates. Utilizing a dataset of 97 patients, we employ Multiple Linear Regression, Robust Regression, Lasso Regularization, and Generalized Additive Models (GAM). Our findings suggest that despite the presence of influential observations, linear structures dominate the data generating process, with Lasso Regression providing the optimal balance between bias and variance (Test RMSE: 0.7037). The study highlights the critical role of tumor volume and capsular penetration in predicting PSA levels.

Reproducible implementations of the tests, models, and plotting scripts utilized in this analysis are available at the following GitHub repository: Statistical Modelling of Prostate Cancer Biomarkers.

1 Introduction

Prostate cancer is emerging as an increasingly serious public health threat due to a rise in advanced-stage diagnoses and an aging global population, which counteracts improved survival rates for early-detected cases [James et al., 2024]. It stands as the most common cancer in men in 112 countries, accounting for roughly 15% of all male cancers. Recent projections by the Lancet Commission suggest a surge in annual new cases from 1.4 million in 2020 to 2.9 million by 2040 [James et al., 2024]. This anticipated rise cannot be mitigated by lifestyle changes or public health interventions alone, necessitating robust governmental strategies.

While incidence rates for some localized prostate cancers have declined in the United States, there has been a concerning trend where advanced, metastatic prostate cancer cases have increased [WebMD, Desai et al., 2024]. This shift is partially attributed to changes in screening guidelines and the complexity of distinguishing between indolent and aggressive disease [Keck Medicine of USC, 2024, hea, 2024]. Consequently, the ability to accurately model biomarkers such as Prostate Specific Antigen (PSA) using non-invasive or pre-surgical clinical data is of paramount importance. High precision in statistical modelling of these biomarkers can aid in risk stratification, potentially reducing the need for invasive biopsies which often carry side effects [Harvard Health Publishing, 2023], and supporting active surveillance strategies for low-risk patients [National Cancer Institute, 2023].

2 Data Description

The dataset consists of clinical data from 97 men who were about to undergo a radical prostatectomy. The variables capture various physiological and histological attributes of the prostate and the tumor. The transformations (logarithms) applied to several variables are intended to stabilize variance and linearize relationships, a common practice in biological statistics.

Table 1: Description of Variables in the Prostate Cancer Dataset

Variable	Description & Interpretation
`lcavol`	Logarithm of cancer volume. A direct measure of tumor burden; expected to be positively correlated with PSA.
`lweight`	Logarithm of prostate weight. Indicates the size of the organ; larger prostates produce more antigen.
`age`	Age of the patient. Prostate cancer risk and PSA levels typically increase with age.
`lbph`	Logarithm of benign prostatic hyperplasia amount. Non-cancerous enlargement, which can confound PSA readings.
`svi`	Seminal vesicle invasion (Binary: 0/1). Indicator of metastasis; implies a more advanced disease state.
`lcp`	Logarithm of capsular penetration. Measures extra-prostatic extension, a sign of aggressiveness.
`gleason`	Gleason score. Pathological grading system (6-9) indicating tumor architecture and aggressiveness.
`pgg45`	Percentage of Gleason scores 4 or 5. A granular measure of high-grade disease presence.
`lpsa`	**Response Variable**. Log of Prostate Specific Antigen. The primary clinical biomarker for screening and monitoring.

3 Research Question

The primary objective of this study is to model the **Log Prostate Specific Antigen (`lpsa`)** as a function of the available clinical and pathological covariates.

We model `lpsa` because PSA levels spans several orders of magnitude, often following a log-normal distribution in the population. Modeling the logarithm stabilizes the error variance (homoscedasticity) and renders the treatment effects multiplicative rather than additive, which is biologically more plausible. The covariates considered—specifically `lcavol`, `gleason`, and `svi`—are established prognostic factors in oncology. By establishing a statistical relationship between these physical/pathological traits and the biomarker `lpsa`, we aim to understand the extent to which tumor burden and aggressiveness explain the variance in serum antigen levels. This investigation finds its place in the broader context of *nomogram construction*, where clinicians predict pathological outcomes based on pre-operative variables to guide decision-making regarding radical prostatectomy versus radiation or active surveillance.

4 Methods

To analyze the relationship between `lpsa` and the predictors, we employed a multi-stage statistical framework.

4.1 Multiple Linear Regression & Stepwise Selection

We began with the Ordinary Least Squares (OLS) estimator, $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$. To induce parsimony, we utilized stepwise regression minimizing the Akaike Information Criterion (AIC) [Akaike, 1974]. AIC penalizes model complexity to prevent overfitting.

4.2 Diagnostics & Influence Analysis

We assessed the OLS assumptions using the Shapiro-Wilk test for normality of residuals and the Breusch-Pagan test for heteroscedasticity. Influential observations were identified using Cook's Distance (D_i). Points exceeding the threshold $4/n$ were flagged for further investigation. We also utilized Component-Residual plots to visually assess the linearity assumption for each predictor variable.

4.3 Robust Regression (M-Estimation)

To address potential outliers detected by Cook's Distance, we employed Robust Linear Models (RLM) using Huber M-estimation [Huber, 1964], which minimizes a function increasing less rapidly than the square for large residuals.

4.4 Lasso Regularization

Recognizing the potential for multicollinearity (e.g., between `gleason` and `pgg45`), we applied the Least Absolute Shrinkage and Selection Operator (Lasso) [Tibshirani, 1996]. Standardization of predictors is performed internally to ensure the penalty λ applies uniformly.

4.5 Generalized Additive Models (GAM)

To test for non-linear relationships without assuming specific parametric forms, we utilized Generalized Additive Models [Hastie and Tibshirani, 1990], utilizing smooth spline functions for continuous predictors.

5 Results

5.1 Exploratory Data Analysis

The correlation matrix (Figure 1) reveals strong positive correlations between the response `lpsa` and predictors `lcavol` (0.73) and `svi` (0.56).

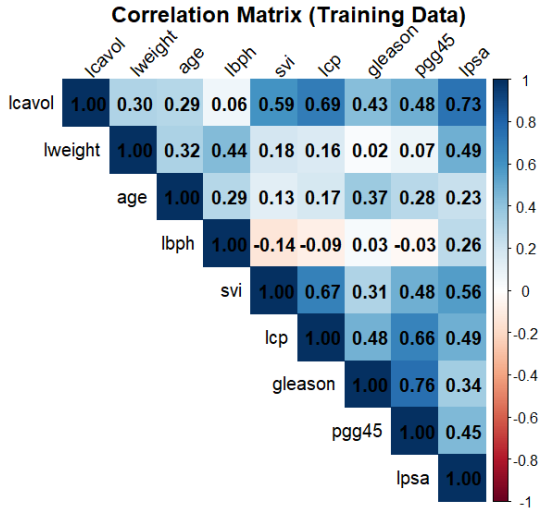

Figure 1: Correlation Matrix of Training Data

Scatterplots (Figure 2) confirm a largely linear trend between `lcavol` and `lpsa`, while variables like `age` show a weaker, more diffuse relationship.

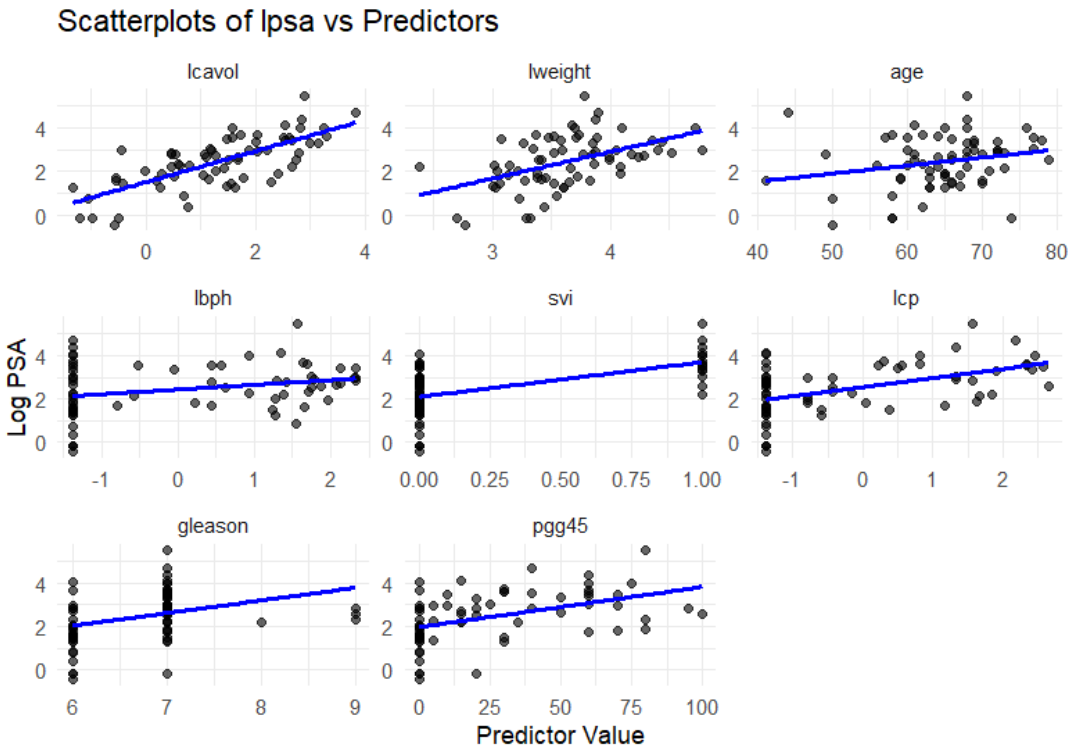

Figure 2: Scatterplots of Covariates against Log PSA

5.2 Model Fitting and Diagnostics

The Stepwise AIC procedure selected a model retaining `lcavol`, `lweight`, `age`, `lbph`, `svi`, `lcp`, and `pgg45`. Diagnostic tests were favorable (Shapiro-Wilk $p = 0.3178$; Breusch-Pagan $p = 0.7302$). Figure 3 displays the standard regression diagnostic plots.

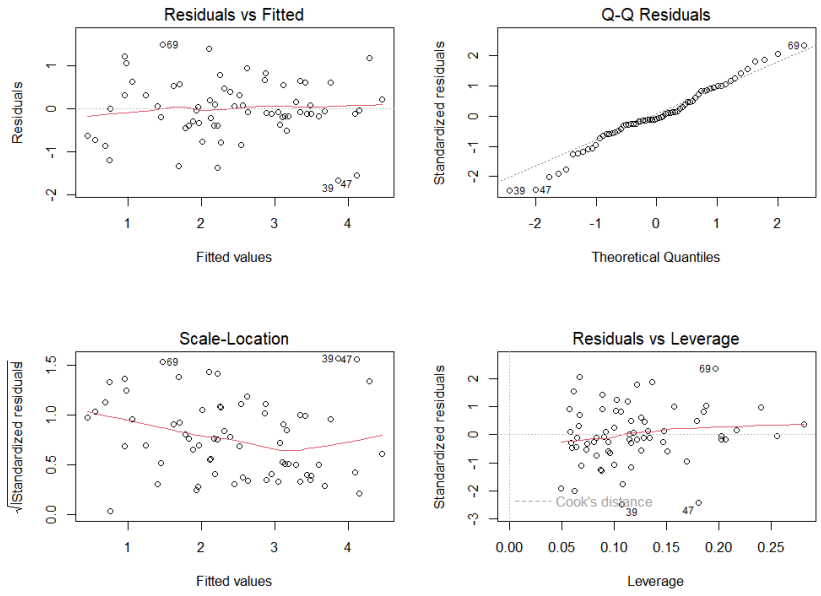

Figure 3: Regression Diagnostics: Residuals vs Fitted, Q-Q Plot, Scale-Location, and Leverage.

To further validate the linearity assumption, Component-Residual plots were examined (Figure 4). The relatively straight trend lines (magenta) compared to the fitted lines (blue) suggest that the logarithmic transformations applied to the predictors were sufficient to capture the linear relationships.

Figure 4: Component + Residual Plots for the Stepwise Model. These plots isolate the relationship between each predictor and the response residuals, confirming linearity.

The Influence Plot (Figure 5) identified several points (indices 38, 39, 47, 69) exceeding the Cook's distance threshold.

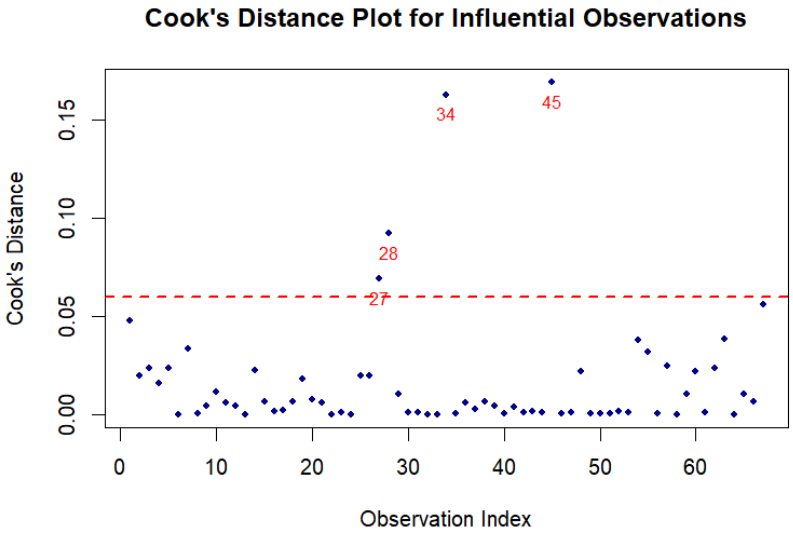

Figure 5: Cook's Distance Plot identifying influential observations.

5.3 Comparative Performance

The models were evaluated on a held-out test set ($n = 30$). Figure 6 illustrates the predicted vs actual values for all considered models.

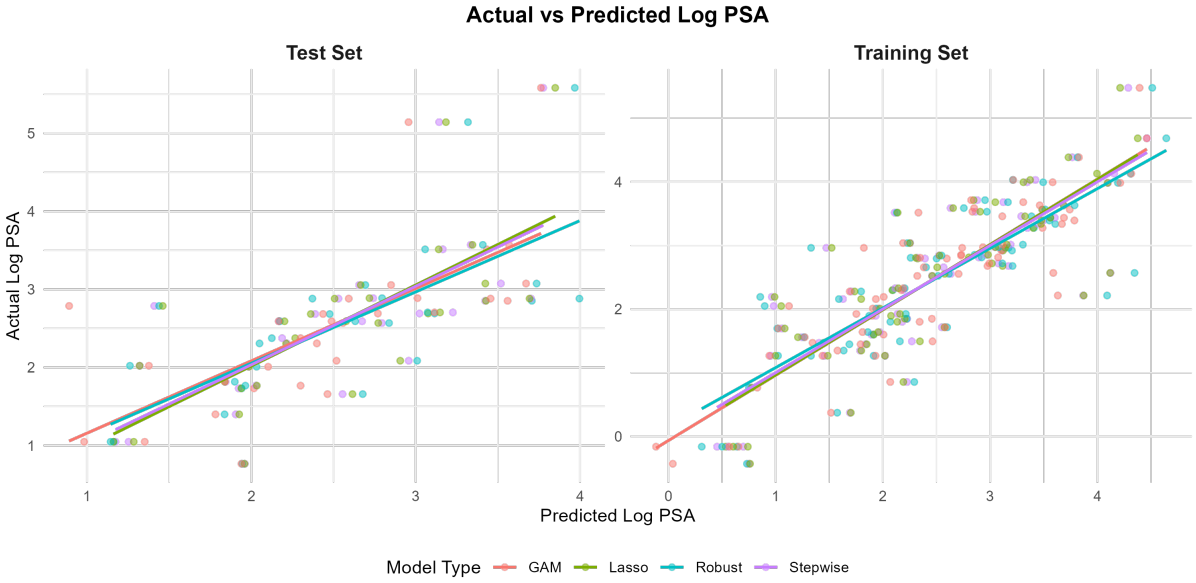

Figure 6: Actual vs Predicted Log PSA for Training and Test Sets across all models. The dashed line represents perfect prediction.

Table 2 summarizes the performance metrics. Notably, Robust Regression performed slightly worse than OLS, implying the "influential" points carried valid signal. The Lasso regression achieved the best generalization (Test RMSE: 0.7037).

Table 2: Model Performance Comparison on Test Data

Model	**Test RMSE**	**Test R^2**
Stepwise OLS	0.7187	0.5079
Robust Regression (RLM)	0.7234	0.5015
Lasso Regression	**0.7037**	**0.5282**
Generalized Additive Model (GAM)	0.7464	0.4693

6 Discussion

6.1 Inference

The analysis confirms that tumor volume (`lcavol`) is the single most significant predictor of PSA levels, followed by prostate weight (`lweight`) and seminal vesicle invasion (`svi`). The linearity of these relationships validates the use of `lpsa` as a surrogate marker for tumor burden.

6.2 Constructive Criticism

From the standpoint of statistical academia, while the linear assumptions held, the reliance on Stepwise AIC is often criticized for post-selection inference validity. The standard errors in the final stepwise model do not account for the uncertainty of the selection process itself, leading to potentially overconfident p-values. Furthermore, the handling of influential points via RLM demonstrated a common bottleneck: distinguishing between "bad" outliers (measurement error) and "good" outliers (extreme biological variation). In this case, treating them as noise was detrimental. A more Bayesian approach could have incorporated prior clinical knowledge to robustly handle these deviations without discarding their signal.

6.3 Shortcomings

A major limitation of this analysis is the small sample size ($n = 97$), which limits the power of non-linear methods like GAM. Additionally, the dataset is observational. We model correlation, not causation; elevated PSA correlates with tumor volume, but other unmeasured factors (inflammation, prostatitis) could confound this. In practice, data collection for such studies faces challenges in standardization—pathological grading (Gleason) can have inter-observer variability among pathologists.

6.4 Reflection

This study underscores the utility of regularization techniques (Lasso) even in small-data regimes, where they outperform complex non-parametric methods. Future work could focus on integrating genomic biomarkers alongside these pathological features to improve predictive accuracy. Studies such as this should be promoted because accurate statistical risk stratification is the cornerstone of reducing over-treatment in prostate cancer, directly addressing the "surge" in cases warned by the Lancet Commission [James et al., 2024].

Acknowledgements

I, **Swagato Das**, Indian Statistical Institute, Kolkata, M.Stat - 1st Year, Roll No. - MB2534, would like to express my sincere gratitude to **Professor Dr. Abhik Ghosh**, ISRU, ISI Kolkata, for his careful instructions throughout the course and his genius insights. It has been an enlightening encounter with him.

References

Screening strategies could reduce prostate cancer mortality, overdiagnosis among Black men. *Healio Hematology/Oncology*, 2024.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Mihir M Desai et al. Trends in incidence of metastatic prostate cancer in the US. *JAMA Network Open*, 2024.

Harvard Health Publishing. Prostate biopsy side effects are common, 2023.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Nicholas D James et al. The Lancet Commission on prostate cancer: planning for the surge in cases. *The Lancet*, 403(10437):1683–1722, 2024.

Keck Medicine of USC. Metastatic prostate cancer on the rise since decrease in cancer screenings, 2024.

National Cancer Institute. Active surveillance for low-risk prostate cancer continues to rise, 2023.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

WebMD. Advanced prostate cancer on the rise. `https://www.webmd.com/prostate-cancer/advanced-prostate-cancer-rise`. Accessed: 2025.