AdaptVarLM: A Linear Regression Model for Covariate-Dependent Non-Constant Error Variance

by

Wanmeng Wang
B.Sc., University of Manitoba, 2022

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

AdaptVarLM: A Linear Regression Model for Covariate-Dependent Non-Constant Error Variance

by

Wanmeng Wang
B.Sc., University of Manitoba, 2022

Supervisory Committee

Dr. Xuekui Zhang, Supervisor
(Department of Mathematics and Statistics,University of Victoria)

Dr. Li Xing, Member
(Department of Mathematics and Statistics,University of Victoria)

Dr. Xiaojian Shao, Member
(Department of Mathematics and Statistics,University of Victoria)

ABSTRACT

In biological research, traditional multiple regression models assume homoscedasticity — constant variance of error terms — an assumption that is difficult to maintain in complex biological data. This thesis introduces AdaptVarLM, a novel linear regression model specialized in dealing with non-constant error variance dependent on one covariate. AdaptVarLM integrates an auxiliary linear relationship between the logarithmic variance of the error term and a specific explanatory variable, and uses maximum likelihood estimation (MLE) in the iterative updating process to improve the parameter estimation accuracy. By modelling non-constant error variance, AdaptVarLM outperforms the traditional regression model in capturing the complex variability inherent in biological data. Applying to the study of Alzheimer's disease, AdaptVarLM detects genetically linked genes associated with the disease and error variance. The results of analyzing both bulk and single-cell data validate the effectiveness of AdaptVarLM in detecting significant genes.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Xuekui Zhang, for his invaluable guidance and support throughout my research. His essential guidance has been instrumental in my academic journey and has inspired me to grow as a researcher. I would also like to express my deep appreciation to my committee members, Dr. Li Xing, Dr. Xiaojian Shao, and Dr. Ke Xu for their insightful feedback that greatly assisted my research. Special thanks also go to my family and friends, whose encouragement has been a source of strength and motivation. AI tools are licensed to assist with refining the language. Their support made me enjoy this challenging journey even more.

Chapter 1

Introduction

The homoscedasticity assumption-that error terms have constant variance-is a foundational assumption of some statistical models such as the traditional multiple regression model. Although this assumption helps simplify the analysis process, it ignores the inherent complexity of biological data. The multiple regression model assumes that all error terms in the data have constant variance, thus failing to capture small changes in error variance of the data [5] [8] [12]. This simplification step may result in unsatisfactory parameter estimation and prediction performance of the multiple regression model in some data scenarios. Non-constant error variances that depend on covariates occurs in the field of biological gene expression. For example, the variability in gene expression data, particularly when influenced by diseases or environmental factors, can introduce significant bias in statistical estimates, thereby challenging the assumption of constant variance and affecting the reliability of genomic analyses. In summary, the inherent complexity in biological data challenges the homoscedasticity assumption in the multiple regression model. Therefore, how to capture the non-constant error variance in the data and include the covariance-dependent characteristics in the model has become an important content in the field of statistical modelling.

Over recent years, the problems caused by non-constant error variance in regression analysis have received considerable attention. Weighted Least Squares (WLS) and Robust Standard Errors are the two most common methodologies for correcting non-constant error variance. In the WLS approach, each data point is assigned a different weight, which is generally the inverse of its square of residual [7]. In this way, the weight ensures that observations with larger uncertainty have less impact on the model estimates [7]. The Robust Standard Errors approach eliminates the effects

of non-constant error variance on the model by adjusting the covariance matrix of parameter estimates [14][17]. Specifically, the Robust Standard Errors method provides a corrected standard error for parameter estimates, which makes statistical inference valid even if the error variance is not constant. These two traditional methods, WLS and Robust Standard Errors, use a transformation mechanism to eliminate the impact of non-constant error variance in linear analysis as much as possible. A new Bayesian approach distributes error variance by leveraging the Dirichlet process mixture prior [16]. This Bayesian method is shown to provide more accurate parameter estimates and predictions, especially in the case of non-normal error terms[16]. Although these three approaches have proven effective in solving the non-constant error variance problem, they don't adequately account for the error variance dependence of covariates in the model.

To make full use of the inherent dependence of non-constant error variance on covariates in the data, rather than eliminating its impact in the regression analysis, this thesis proposes a novel approach, AdaptVarLM, based on the multiple regression model. AdaptVarLM aims to improve parameter estimation accuracy in the linear regression model by removing the restriction of the homoscedasticity assumption and exploiting the dependence information between error variance and covariates. AdaptVarLM introduces an auxiliary linear relationship between the logarithmic error variance and one specific explanatory variable. In the process of parameter estimating, AdaptVarLM uses a method similar to the Expectation-Conditional-Maximization (ECM) [10][15]. In detail, maximum likelihood estimation (MLE) is iteratively applied to update the parameter estimates in two linear relationships alternately: the linear relationship between the output variable and the explanatory variables, and the auxiliary relationship between the log error variance and a specific covariate. In this way, AdaptVarLM progressively improves parameter estimation accuracy in the linear regression model and is not restricted by the homoscedasticity assumption. Additionally, while the regression model is limited to identifying factors that significantly influence the mean of the explained variable, AdaptVarLM can detect significant factors associated with the error variance of the explained variable.

This thesis is structured to provide a detailed explanation and study of the AdaptVarLM methodology. Section 2 introduces the notation and mathematical model representation of AdaptVarLM, followed by a description of the model fitting and parameter estimation algorithms. Section 3 provides a simulation study, including model parameter settings (Section 3.1), simulation procedure (Section 3.2), and sim-

ulation results (Section 3.3). Section 4 contains the application of AdaptVarLM to the Religious Groups Study and Memory and Aging Project (ROSMAP) on Alzheimer's disease [1]. Real data analysis includes two datasets: Bulk RNA-Seq data analysis (Section 4.1) and Single-Cell RNA-Seq data analysis (Section 4.2). This thesis concludes with the discussion in Section 5, where we summarize conclusions about AdaptVarLM, and reflect on the limitations and future research directions in AdaptVarLM.

Chapter 2

Methods

2.1 Notation and Statistical Model

In our research, we consider a continuous outcome variable Y and its possible association with k explanatory variables $X_1, X_2, \ldots, X_k$. We hypothesize that these variables may influence Y but do not prescribe any particular distribution for the explanatory variables X. The vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ comprises the observed values of Y from n samples. The $n \times k$ matrix $\mathbf{x}$ contains observed data of explanatory variables, where each row $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$ encapsulates the values of the k explanatory variables for the i-th observation, for $i = 1, \ldots, n$.

Our model is tailored to accommodate situations where certain factors, like treatment or disease status, may influence both the mean and the variance of the outcome Y. This leads to a variance that is not constant but varies in response to a specific predictor. We enhance the multiple linear regression framework as follows:

$$\begin{aligned} y_i &= a_0 + a_1 x_{1i} + a_2 x_{2i} + \ldots + a_k x_{ki} + \epsilon_i, \\ \epsilon_i &\sim N(0, \sigma_i^2), \\ \log(\sigma_i^2) &= b_0 + b_1 x_{1i}, \quad i \in \{1, \ldots, n\}. \end{aligned} \tag{2.1}$$

In this formulation, $\mathbf{a} = (a_0, a_1, \ldots, a_k)$ denotes the regression coefficients, similar to those in a traditional multiple linear regression, reflecting the influence of the predictors on the mean of Y. Notably, X_1 is identified as a key predictor that impacts the variance of the outcome. The coefficients $\mathbf{b} = (b_0, b_1)$ are unique to our model, illustrating the effect of X_1 on the variance of Y. This aspect distinguishes our

approach from standard multiple linear regression, where the variance is considered constant.

2.2 Model Fitting and Parameter Estimation

To fit the model with data $(\mathbf{y}, \mathbf{x})$ and estimate parameters $(\mathbf{a}, \mathbf{b})$ from model (2.1), we maximize the log-likelihood:

$$l(\mathbf{a}, \mathbf{b}; \mathbf{y}, \mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a_0 - \sum_{j=1}^{k} a_j x_{ji})^2}{\sigma_i^2}, \tag{2.2}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(b_0 + b_1 x_{1i}) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - a_0 - \sum_{j=1}^{k} a_j x_{ji})^2}{e^{b_0 + b_1 x_{1i}}}. \tag{2.3}$$

Given the complexity of the likelihood (2.3), there is no analytic solution to maximize it with respect to parameters $(\mathbf{a}, \mathbf{b})$. We develop an algorithm that mimics the Expectation-Conditional-Maximization (ECM) approach. This algorithm simplifies the maximization process by conditioning on certain information. We focus on three elements: σ_i^2, $\mathbf{a}$, and $\mathbf{b}$. By conditioning on one element, maximizing the likelihood with respect to another becomes more manageable.

First, we treat σ_i^2 as a latent variable, analogous to the E-step in the ECM algorithm. Given the values of $\mathbf{b}$, the values of σ_i^2 can be directly calculated as per model (2.1).

Second, with σ_i^2 known, maximizing the log-likelihood (2.2) with respect to $\mathbf{a}$ reduces to a weighted least squares (WLS) problem with a closed-form solution:

$$\mathbf{a} = (\mathbf{x}^\top \mathbf{w}\mathbf{x})^{-1}\mathbf{x}^\top \mathbf{w}\mathbf{y}, \tag{2.4}$$

where the diagonal weight matrix $\mathbf{w}$ is defined as

$$w_{ij} = \begin{cases} 0 & i \neq j, \\ 1/\sigma_i^2 & i = j, \end{cases} \quad i, j = 1, \ldots, n. \tag{2.5}$$

Third, given the values of $\mathbf{a}$, the maximization of the likelihood with respect to $\mathbf{b}$ demands a numerical approach. For this purpose, we employ the Nelder-Mead method [11], a well-established technique for finding solutions to problems where analytical

methods are infeasible, to solve the equation:

$$\frac{\partial}{\partial \mathbf{b}} l(\mathbf{a}, \mathbf{b}; \mathbf{y}, \mathbf{x}) = \mathbf{0}. \tag{2.6}$$

To implement this in R, we utilize the *optim* function with the 'method=Nelder-Mead' option. This iterative process, involving calculating σ_i^2, updating $\mathbf{a}$, and updating $\mathbf{b}$, mirrors the E-step and two CM-steps in the ECM algorithm. We iterate these steps until convergence is achieved for the final solution of $(\mathbf{a}, \mathbf{b})$. The procedure requires an initial value, for which we use the traditional multiple regression model to initialize $\mathbf{a}$. To ensure practical convergence, we set a threshold $\varepsilon = 10^{-6}$ and define convergence as the condition where the change in every element of both $\mathbf{a}$ and $\mathbf{b}$ is less than this threshold in a single iteration. Like the EM algorithm, each update monotonically increases the log-likelihood, leading to convergence. Typically, significant changes do not occur after several iterations. When computational efficiency is crucial, we may impose a maximum number of iterations (for example, no more than 100) to balance precision and computing speed. The complete method is outlined in Algorithm 1.

Algorithm 1 ECM-like Algorithm for Parameter Estimation in Model (2.1)

Input: Data $(\mathbf{y}, \mathbf{x})$, Convergence threshold ε, Log-likelihood $l(\mathbf{a}, \mathbf{b}; \mathbf{y}, \mathbf{x})$
Initialization:
Fit linear regression on $(\mathbf{y}, \mathbf{x})$ to obtain initial $\mathbf{a} = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}$
Iterative Update Loop:
while not converged **do**
    Update b: Solve $\mathbf{b}$ from $\frac{\partial}{\partial \mathbf{b}} l(\mathbf{a}, \mathbf{b}; \mathbf{y}, \mathbf{x}) = \mathbf{0}$, numerical solution (e.g., Nelder-Mead method)
    Calculate σ_i^2 and diagonal weight matrix w: $w_{ii} = 1/\sigma_i^2 = 1/e^{b_0 + b_1 x_{1i}}$
    Update a: Solve $\mathbf{a}$ from $\frac{\partial}{\partial \mathbf{a}} l(\mathbf{a}, \mathbf{b}; \mathbf{y}, \mathbf{x}) = \mathbf{0}$, solution $\mathbf{a} = (\mathbf{x}^\top \mathbf{w} \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{w} \mathbf{y}$
    Check for convergence by comparing the change in $\mathbf{a}$ and $\mathbf{b}$ with threshold ε
end while
Denote $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ as the estimation of parameters in the last iteration.
Estimate SE of point estimation :
  Compute the observed Fisher Information matrix $I_{\text{obs}}(\hat{\mathbf{a}}, \hat{\mathbf{b}})$
  Invert the Fisher Information matrix to obtain the covariance matrix $\text{Cov}(\hat{\mathbf{a}}, \hat{\mathbf{b}})$
  Calculate SEs as the square roots of the diagonal elements of the covariance matrix
Output: The final point estimation $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$; SE of point estimation.

To calculate the Standard Error (SE) of the parameter estimations from Algorithm 1 output, we employ the Observed Fisher Information [6], which evaluates the second

derivative at the maximum likelihood estimate (MLE) $\hat{\theta} = (\hat{\mathbf{a}}, \hat{\mathbf{b}})$. The covariance matrix of the parameter estimates is the inverse of the Observed Fisher information matrix. The standard errors are the square roots of the diagonal elements of this covariance matrix. More detailed formulas for each element in the Observed Fisher Information matrix are shown in the Appendix. The relevant formulas are:

$$I_{\text{obs}}(\hat{\theta}) = - \left. \frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \tag{2.7}$$

$$\text{Cov}(\hat{\theta}) = I_{\text{obs}}(\hat{\theta})^{-1} \tag{2.8}$$

Chapter 3

Simulation Studies

3.1 Simulation Settings

In our simulation study, we investigate various scenarios for the true values of **a** and **b**, with a focus on both moderate ($n = 100$) and large ($n = 1000$) sample sizes. We configure our experiments with $k = 2$, using $a_2 = 0$ or $a_2 \neq 0$ as a means to explore scenarios both with and without additional covariates beyond the key variable X_1. We set $a_0 = 0$, as it merely introduces a consistent shift in the outcome, thus not influencing the assessment results. Likewise, b_0 is fixed at 0, establishing a baseline variance of $\sigma_i^2 = 1$ for the regression error. We vary b_1 to examine how the predictor X_1 influences the variance of Y. The values of a_1 are selected to represent different signal-to-noise ratio levels, considering the fixed value of b_0 [3]. Similarly, the settings for a_2 correspond to the strength of association with additional covariates.

In summary, as displayed in Table 3.1, the above simulation plan involves $150 = 2 \times 5 \times 5 \times 3$ settings as combinations of values in 4 parameters, and set $a_0 = b_0 = 0$ without loss of generality in model evaluation.

Table 3.1: The Table of Simulation Parameters and level Settings

Variable	**Possible Values**
Sample Size (n)	100, 1000
Effect of X_1 on Outcome Mean (a_1)	0, 0.1, 1, 10, 100
Effect of Covariate (a_2)	0 (No Covariate), 0.1, 1, 10, 100
Effect of X_1 on Outcome Variance (b_1)	0 (Standard Linear Model), 1, 3

3.2 Simulation and Analysis Procedure

For each setting described above, the simulation and analysis procedure includes the following steps:

- **GENERATE DATASETS**

1. Generate explanatory variables from a standard normal distribution:

$$x_{1i}, x_{2i} \sim \mathrm{N}(0,1), \quad \text{for} \quad i = 1, \ldots, n.$$

2. Calculate σ_i^2 using $\sigma_i^2 = e^{b_0 + b_1 x_{1i}}$

3. Generate regression errors ϵ_i from $\epsilon_i \sim \mathrm{N}(0, \sigma_i^2)$

4. Calculate the response variable y_i using the model:

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \epsilon_i$$

5. Repeat Steps 1-4 to generate a data set $(\mathbf{x}, \mathbf{y})$.

- **EVALUATION**

6. On the data set, fit AdaptVarLM and other compared models to obtain estimated parameters $\hat{\theta} = (\hat{\mathbf{a}}, \hat{\mathbf{b}})$, Observed Fisher Information Matrix using $(\mathbf{x}, \mathbf{y})$ and $\hat{\theta}$, and calculate parameter estimation standard errors (SEs).

7. Calculate **Criteria 1: Estimation bias** $\delta_\theta = \hat{\theta} - \theta$, which is defined as the difference between estimations and true parameter values used in simulation setting.

8. Record the Coverage Indicator $I_{coverage}$, which indicates whether the 95% CI of parameter estimation covers the real value: $\hat{\theta} - 1.96 \times SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + 1.96 \times SE(\hat{\theta})$

9. Record the Detect Indicator I_{detect}, which indicates whether the null hypothesis is rejected (e.g., parameter estimate is significantly different from zero).

- **REPLICATES**

10. Repeat 100 times of Steps 5 to generate and analyze 100 data sets.

11. Repeat the above Steps 6-9 100 times to assess the model's performance. Specifically, use 100 Coverage Indicators to calculate the **Criteria 2: Type I Error** and **Criteria 3: Power** in the test. Finally, use 100 Coverage Indicators to evaluate the 95% confidence interval (CI) rejection rate to determine the proportion of times the null hypothesis is rejected when the true parameter values fall within this interval.

3.3 Simulation Results

To compare the performance of AdaptVarLM against existing models, we selected the multiple regression model, WLS, and RSE for comparison. For each parameter, these steps above produce $15,000 = 150 \times 100$ estimating values for the three criteria mentioned above for both AdaptVarLM and the other models being compared. We analyzed these criteria to explore the impact of different settings on the parameter estimation of AdaptVarLM, and compare the performance of AdaptVarLM across various parameter settings against other models.

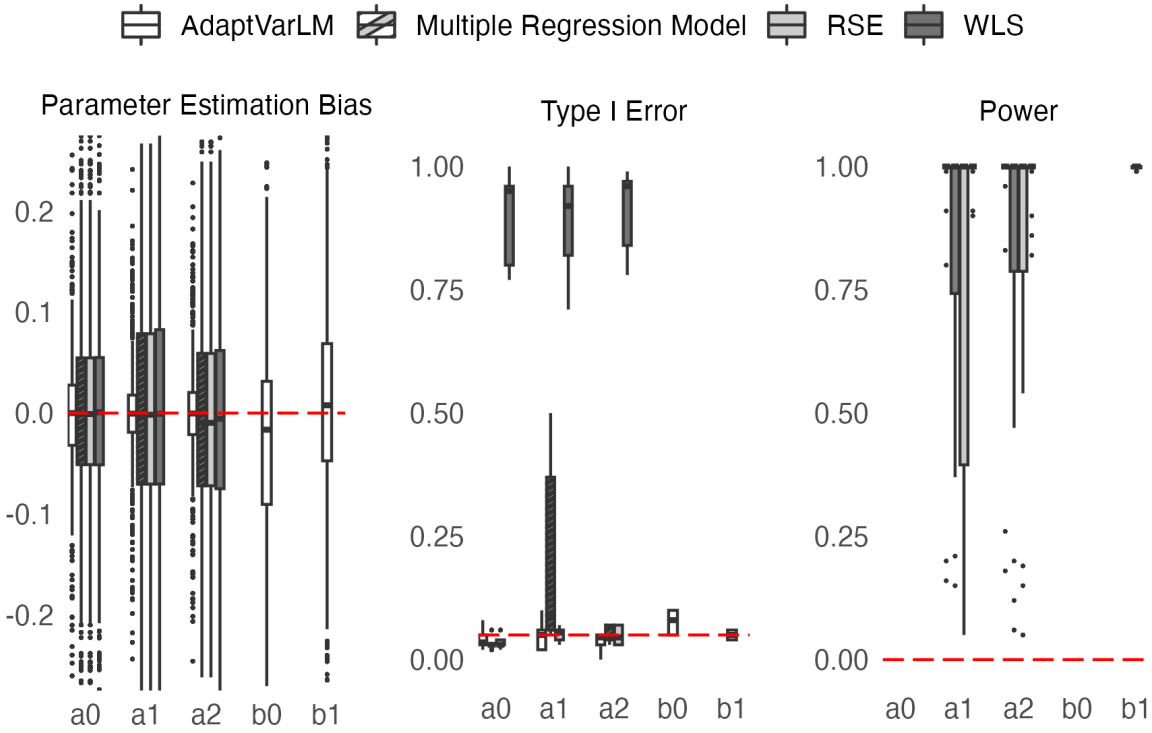


Figure 3.1: The Boxplots of Estimation Bias, Type I Error, and Power in Four Models

Figure 3.1 presents boxplots that compare the performance of the AdaptVarLM with other models in terms of Parameter Estimation Bias, Type I Error, and Power across all 150 simulation settings. The left subfigure shows the boxplots of Parameter Estimation Bias, and the red horizontal line at zero bias serves as a reference, indicating where the model's estimates are equal to the true parameter values. The subfigure show that AdaptVarLM produces parameter estimates with less overall variability, suggesting greater accuracy in parameter estimation. In contrast, the other models have more spread in their biases. The middle subfigure shows the boxplots of Type I Error for AdaptVarLM compared to other models across different parameters a_0, a_1, a_2, b_0, and b_1. The red vertical line at 0.05 represents the nominal significance level ($\alpha = 0.05$). The boxplots for AdaptVarLM are closer to the nominal significance level, with less variability in Type I Error for most parameters. This indicates that AdaptVarLM performs more reliably in maintaining the error rate. In contrast, the multiple regression model shows greater variability in Type I Error, especially for the a_1 parameter, suggesting difficulties in controlling the error rate when the error variance is not constant. The WLS model increases power at the cost of significantly increased Type I Error for parameters a_0, a_1, and a_2, which highlights a trade-off by sacrificing control over false positives to enhance its ability to detect true effects. The right subfigure displays boxplots of Power for parameters a_1, a_2, and b_1, as power values are not applicable for a_0 and b_0 since these parameters are consistently set to zero. The white boxplots representing AdaptVarLM are very short and appear almost as lines close to 1, which suggests its strong ability to correctly detect true effects of covariates on error variance with minimal variability in power across different simulation settings. In contrast, the WLS and RSE models have a much wider spread in their power values, which suggests their ability to detect true effects is less consistent in some cases, potentially leading to unstable performance across different conditions. Overall, AdaptVarLM shows greater consistency and effectiveness in all three criteria, making it a more dependable model for scenarios that require accurate detection of true effects of the covariate on error variance.

In our simulation study, we conducted 150 simulation settings, as detailed in Table 3.1, to investigate the performance of AdaptVarLM and other models. To accurately compare the performance of the AdaptVarLM and other models across the criteria mentioned above on a pairwise basis, we define a metric for each criteria

as follows:

$$\text{Absolute Bias Difference (ABD)}(\theta) = |\hat{\theta}^{\text{AdaptVarLM}} - \theta^{\text{true}}| - |\hat{\theta}^{\text{Other}} - \theta^{\text{true}}|,$$
$$\text{Absolute Type I Error Difference (ATD)}(\theta) = |\hat{\alpha}_{\theta}^{\text{AdaptVarLM}} - 0.05| - |\hat{\alpha}_{\theta}^{\text{Other}} - 0.05|,$$
$$\text{Power Difference (PD)}(\theta) = (1 - \hat{\beta}_{\theta}^{\text{AdaptVarLM}}) - (1 - \hat{\beta}_{\theta}^{\text{Other}}).$$

These metrics provide a direct comparison of AdaptVarLM with other models: the Absolute Bias Difference(ABD) assesses which model's parameter estimates are closer to the true value, the Absolute Type I Error Difference(ATD) evaluates which model's Type I Errors are closer to the nominal level of 0.05, and the Power Difference(PD) measures which model has greater power.

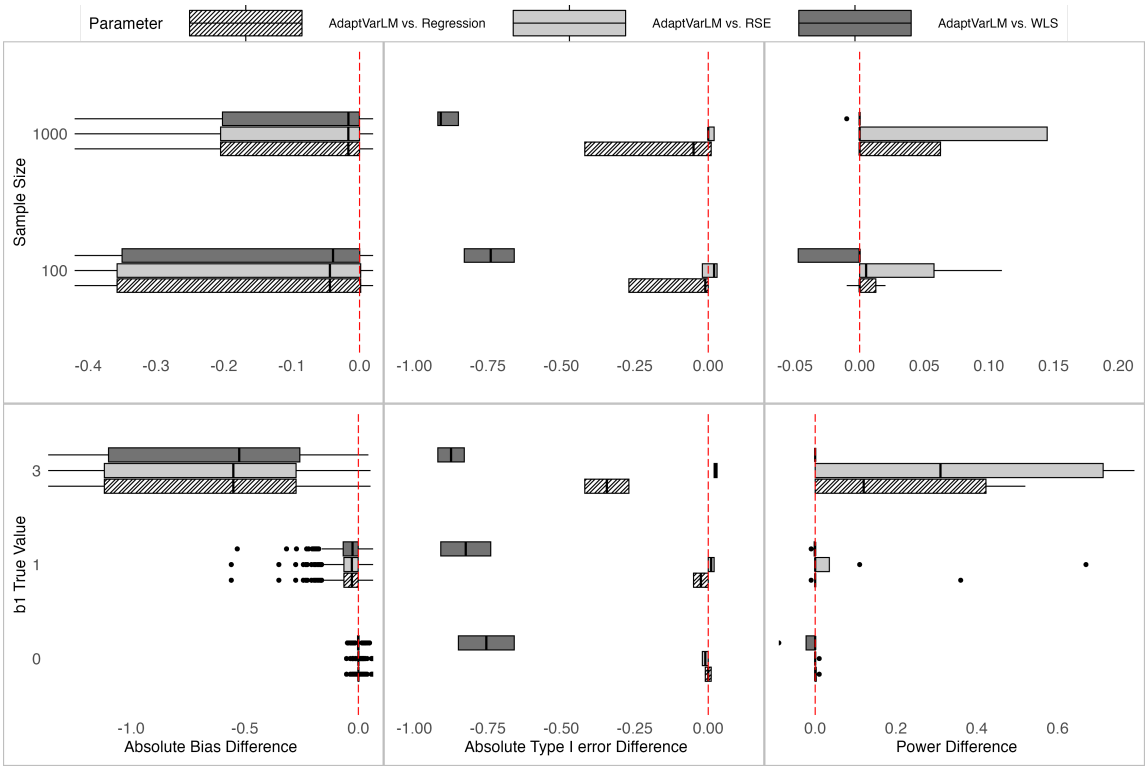


Figure 3.2: The Boxplots of Three Criteria for Parameter $a1$ across Sample Sizes and $b1$ True Values

We focus on analyzing the three criteria for parameter a_1 across different simulation settings since a_1 is the coefficient of the explanatory variable that also affects error variance. The results were grouped based on different simulation settings, and boxplots were created to visualize the outcomes. Figure 3.2 displays the boxplots of ABDs, ATDs, and PDs across different sample sizes and b_1 true values. The top

row of boxplots shows the performance of AdaptVarLM compared to other models in terms of ABD, ATD, and PD for the parameter a_1 across various sample sizes. The negative ABDs suggest that AdaptVarLM produces more accurate parameter estimates with less bias compared to other models. The nearly negative ATDs indicate that AdaptVarLM maintains Type I Error closer to the nominal significance level. The PDs are generally positive when comparing AdaptVarLM to the multiple regression and RSE model, showing that AdaptVarLM tends to achieve higher power than these two models. It should be noted that WLS produces higher power than AdaptVarLM, but this outperformance is achieved by significantly raising the Type I Error. The bottom row of boxplots shows the performance of AdaptVarLM compared to other models in terms of ABD, ATD, and PD for the parameter a_1 across various b_1 true values. The performance of AdaptVarLM across various b_1 true values shows a similar pattern to its performance across various sample sizes. As the true value of b_1 increases, reflecting a stronger effect of X_1 on error variance, all three metrics show a greater deviation from the red line and increased variability. This suggests that AdaptVarLM's performance advantage becomes more evident when the covariate has a significant impact on error variance. Overall, AdaptVarLM shows a better performance compared to the other three models across various sample sizes and b_1 true value settings, providing more accurate parameter estimates and a better balance between Type I Error control and power.

Chapter 4

APPLICATION

We analyze data derived from the Religious Orders Study and Memory and Aging Project (ROSMAP), generated by the Rush Alzheimer's Disease (AD) Center [1]. The ROSMAP dataset aims to investigate the underlying mechanisms and risk factors of Alzheimer's disease by collecting detailed clinical, neuropathological, and genetic data from participants [1]. This dataset helps to understand the progression of Alzheimer's disease and identify treatment options and prevention strategies [1]. The dataset includes bulk RNA-seq and single-cell RNA-seq (snRNA-seq) data from 637 postmortem samples across 8 individuals. We demonstrate AdaptVarLM by analyzing bulk RNA-seq data, which does not include or require cell type information. Paired single-cell RNA-seq data, serving as a known reference, validates our model's effectiveness in capturing variance changes resulting from shifts in cell abundance.

4.1 Bulk Data Analysis

Several data cleaning and pre-processing steps were implemented to ensure the comparability of the data. In detail, we removed sample records containing missing values, excluded genes with zero expression levels in more than 90% of the bulk samples, and standardized the expression levels for every gene to ensure comparability across samples.

To explore the association between gene expression levels and common clinical characteristics in Alzheimer's disease, we applied the AdaptVarLM and multiple regression model to bulk data. The response variable, Y, represents the expression level of one specific gene. Our fitted model includes four important clinical predictors

and two Principal Component Analysis (PCA) predictors. The clinical predictors are: Clinical Cognitive Diagnosis Summary (CCDS) [2] [4], using integers from 1 to 6 to reflect the severity of Alzheimer's disease from mild to severe (X_1); age at the last visit (X_2); sex (X_3); and APOE gene type (X_4), a known Alzheimer's risk factor [13]. To capture major sources of variation in all gene expression, we incorporated the first two principal components from PCA on all genes as predictors, with PCA1 (X_5) explaining 18.20% of the variance and PCA2 (X_6) explaining 9.82% of the variance. Additionally, we modelled the logarithm of the error variance as a linear function of X_1 to address covariance between Y and X_1.

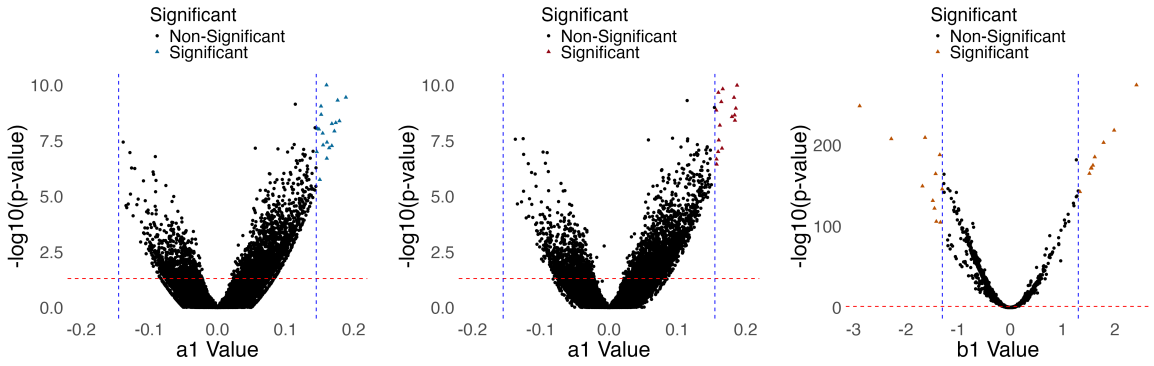
We used volcano plots to visualize the association between gene expression levels and CCDS, focusing on the association coefficient values and their p-values. We identify significant genes based on two criteria:

1. The p-value of the association coefficient is smaller than 0.05, represented by the red line in the volcano plot.

2. Among the genes that satisfy the above criterion, the top 20 genes with the largest absolute association coefficient values are selected as significant.

Figure 4.1 presents the volcano plots illustrating the strength of association between gene expression levels and CCDS by fitting both the multiple regression model and the AdaptVarLM. In the multiple regression model, we considered only the association coefficient $a1$ between the mean of Y and X_1. The volcano plot of $a1$ values and the corresponding p-values is presented in Subfigure (a). In the AdaptVarLM model, we considered both $a1$ and the association coefficient $b1$, which represents the association between the error variance of Y and X_1. The volcano plots of $a1$ and $b1$ are presented in Subfigure (b) and Subfigure (c), respectively. In each volcano plot, genes above the red line (p-value less than 0.05) and outside the blue lines (absolute coefficient values greater than the threshold) are considered significant.

Table 4.1 presents the significant genes detected by the multiple regression model and the AdaptVarLM model. The regression model identifies genes whose "mean" gene expression levels are significantly associated to disease status (CCDS), the AdaptVarLM extends this analysis by detecting genes whose error variance also associated with disease status. These significant genes provide insights into genes that may be related to Alzheimer's Disease. It is noteworthy that the genes in the third column of Table 4.1 are supported by the AdaptVarLM model, which shows that CCDS significantly affects the error variance of expression levels in these genes.



(a) $a1$ in regression model (b) $a1$ in AdaptVarLM (c) $b1$ in AdaptVarLM

Figure 4.1: The Volcano Plots of Significant Genes Associated with Clinical Cognitive Diagnosis Summary(CCDS)

4.2 Single Cell Data Analysis

To verify the effectiveness of AdaptVarLM and the significant genes detected by the AdaptVarLM in the third column of Table 4.1, we performed analysis on single-cell RNA-Seq data derived from 24 bulk samples from patients with varying degrees of CCDS. Each bulk sample comprises 9 distinct cell types, and for each cell type, the expression levels for 28,781 genes were measured to offer detailed information at the single-cell level [1].

To investigate which cell type proportions are affected by CCDS, we combined the single-cell data with clinical CCDS information to determine the CCDS for each bulk sample. A proportion matrix of size 9 x 6 was created to store the cell counts for each cell type across the 6 CCDS categories. For each bulk sample, cell counts of 9 cell types were added to the corresponding CCDS column in the proportion matrix. The proportions of each cell type per CCDS were calculated. The proportion matrix, displayed in Table 4.2, shows the proportions of each cell type for the various CCDS categories, providing a clear representation of how cell type proportions vary between CCDS1 and CCDS5. From Table 4.2, the cell types Exc and Oligodendrocytes show the most significant changes in proportion between CCDS values of 1 and 5.

Focusing on the differences in gene expression density between Exc and Oligodendrocytes cell types, we only consider bulk samples with CCDS values of 1 or 5 because the cell type proportions varied the most in these two CCDS values, resulting in 13 bulk samples. For each gene, the single-cell gene expression levels in 13 bulk samples were extracted separately for Exc and Oligodendrocytes, and subjected to a

Table 4.1: Significant Genes Detected by Regression Model and AdaptVarLM

Genes by $a1$ in Regression	**Genes by $a1$ in AdaptVarLM**	**Genes by $b1$ in AdaptVarLM**
SLC6A9	SLC6A9	LIPF
AQP6	SLC4A11	ENSG00000231656
PRELP	PRELP	PGA3
CCDC69	CCDC69	CTRB2
SLC4A11	AQP6	PGA4
NRIP2	APLN	TXNDC2
APLN	NRIP2	ICAM4
SYTL1	HSPB2	FABP4
HSPB2	SYTL1	CPB1
OLMALINC	SLC6A12	EGR2
ZC3H12A-DT	ZC3H12A-DT	ENSG00000224413
TMCC2	ARRDC2	LAMB3
GAREM2	TMCC2	IGHD
CSRP1	CSRP1	SAA2
SLC6A12	GAREM2	AQP10
ADAMTS2	CLIP2	FAT2
CDC42BPG	OLMALINC	SLC30A2
ARRDC2	CDC42BPG	MTND4P12
ISYNA1	SLC25A48	SEMG2
TMEM184B	ISYNA1	SEMG1

Kullback-Leibler Divergence (KL_Divergence) calculation to quantify the difference in gene expression distributions between the two cell types [9]. To implement this in R, we utilize the *KL* function with the 'unit=log2' option. This analysis produced 17,578 KL_Divergence values for all genes.

The b_1 values from model fitting results in bulk data set were extracted and merged with KL_Divergence values by gene names. To study the KL_Divergence values in b_1 significant genes and non-significant genes, we selected genes with the biggest positive b_1 values as significant gene group and genes with the smallest positive b_1 values as non-significant gene group. Figure 4.2 shows the gene expression level density plots in Exc and Oligodendrocytes of the top 1 significant and non-significant genes and the KL_Divergence value, which indicates the degree of divergence between the distributions. Significant genes ICAM4, display bigger KL_Divergence values compared to non-significant genes IZUMO1. This figure visually shows that the gene expression densities between Exc and Oligodendrocytes have more significant differences in the b_1 significant gene set, compared with the b_1 non-significant gene set. Figure 4.3 shows the box plot of KL_Divergence values for different selected sizes of b_1 significant and non-significant gene groups. The p-value evaluates the statistical significance of the difference in KL_Divergence between the two groups. This plot highlights that significant genes generally have higher KL_Divergence values, and there are greater differences in gene expression distributions between Exc and

Table 4.2: The Table of Cell Type Proportions Across CCDS Categorie 1 and 5

Cell Type	CCDS 1	CCDS 5	Proportion Change
astrocytes	0.16	0.18	0.02
endothelials	0.01	0.01	0.00
Exc	0.42	0.38	-0.04
Inh	0.14	0.12	-0.02
microglia	0.01	0.03	0.02
oligodendrocytes	0.17	0.23	0.06
opcs	0.04	0.05	0.01
pericytes	0.01	0.01	0.00
VLMC	0.03	0.00	-0.03

Oligodendrocytes in b_1 significant gene group.

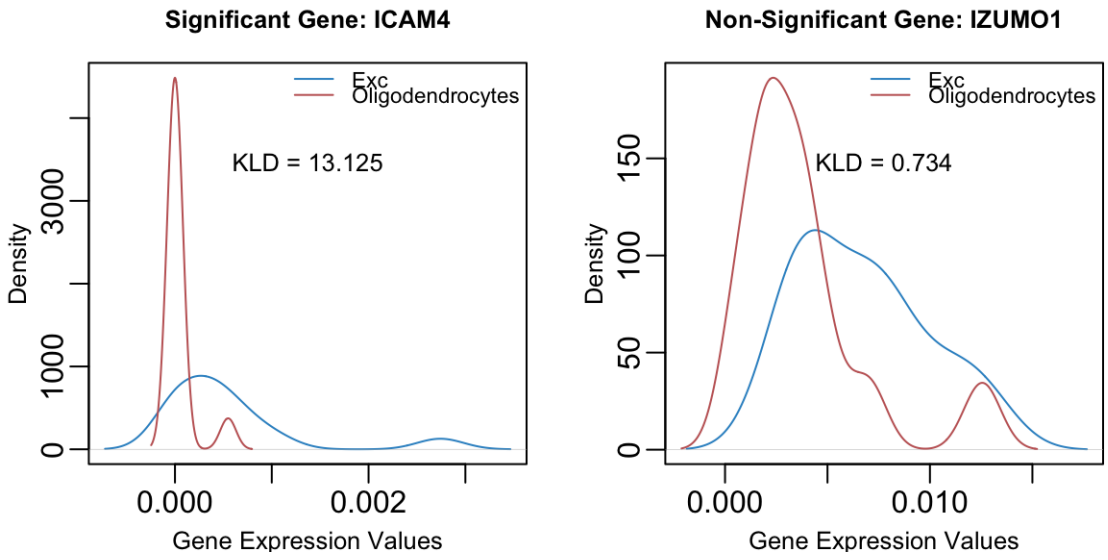


Figure 4.2: The Density Plots of Gene Expression Levels and KL_Divergence in Exc and Oligodendrocytes

The single-cell analysis provides evidence that the severity of Alzheimer's Disease significantly affects the cell type proportions of Exc and Oligodendrocytes, which in turn influences the distributions of gene expression levels. The results from the bulk and single-cell data analysis validate the robustness of AdaptVarLM in detecting significant genes associated with the severity of Alzheimer's disease.

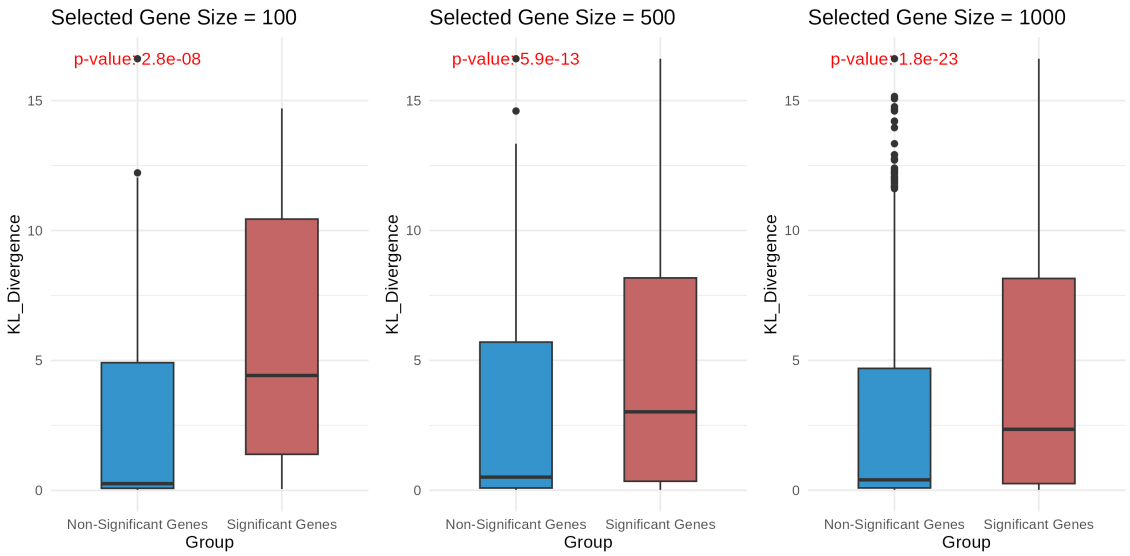


Figure 4.3: The Box Plots of KL Divergence for Significant and Non-significant Genes

Chapter 5

DISCUSSION

This thesis addresses a critical limitation of traditional multiple regression models in biological research: the assumption of homoscedasticity. By introducing AdaptVarLM, we provide a framework for analyzing data with covariate-dependent error variance, overcoming the limitations of traditional models.

This thesis addresses a critical limitation of traditional multiple regression models in biological research: the assumption of homoscedasticity. By introducing AdaptVarLM, we provide a framework for analyzing data with covariate-dependent error variance, overcoming the limitations of traditional models.

AdaptVarLM solves the problem of variability in error terms by establishing an auxiliary linear relationship between the logarithmic error variance and a specific explanatory variable. This allows for more accurate modelling in scenarios where error variance is not constant but varies with one certain covariate. The iterative maximum likelihood estimation (MLE) process is employed in AdaptVarLM to improve parameter estimation accuracy. We apply AdaptVarLM to the ROSMAP dataset to identify significant genes associated with Alzheimer's disease, particularly those highlighted in gene expression distribution variance.

AdaptVarLM assumes that the error variance has a linear relationship with only one covariate. This may limit its applicability in situations where multiple covariates influence variance. To address these limitations and expand the effectiveness of AdaptVarLM, future research could extend this approach to adapt multiple sources of variability, which leads to greater flexibility to a wider range of datasets.

Chapter 6

APPENDIX

6.1 Formulas for the Observed Fisher Information Matrix $I_{\mathbf{obs}}(\hat{\theta})$

Each element in $I_{\text{obs}}(\hat{\theta})$ is derived based on the partial second derivatives of the likelihood function with respect to the model parameters. First, we define the notations used in the formulas of $I_{\text{obs}}(\hat{\theta})$:

- n represents the number of data points.
- (X_i, Y_i) represents the i-th data point, where $i \in \{1, \ldots, n\}$.
- k represents the number of predictor variables X.
- $\hat{e_i}$ represents the i-th residual of (X_i, Y_i) , calculated as:

$$e_i = y_i - a_0 - \sum_{j=1}^{k} a_j x_{ji}$$

- Q_i represents the exponential transformation between error variance σ_i^2 and x_{1i}:

$$Q_i = e^{b_0 + b_1 x_{1i}}.$$

- l is the log-likelihood function of AdaptVarLM with respect to the model parameters.

For each element in the m-th row and n-th column of $I_{\text{obs}}(\hat{\theta})$, where $1 \leq m \leq k+1$ and $1 \leq n \leq k+1$, the formula for $I_{m,n}$ is:

When $1 \leq m \leq k+1$, $1 \leq n \leq k+1$:

$$I_{(m,n)} = I(a_{m-1}, a_{n-1}) = -\frac{\partial^2 l}{\partial a_{m-1} \partial a_{n-1}} = \sum_{i=1}^{n} \frac{(x_{(m-1)i})(x_{(n-1)i})}{Q_i}$$

When $1 \leq m \leq k+1$, $n = k+2$:

$$I_{(m,n)} = I(a_{m-1}, b_0) = -\frac{\partial^2 l}{\partial a_{m-1} \partial b_0} = \sum_{i=1}^{n} \frac{(e_i)(x_{(m-1)i})}{Q_i}$$

When $1 \leq m \leq k+1$, $n = k+3$:

$$I_{(m,n)} = I(a_{m-1}, b_1) = -\frac{\partial^2 l}{\partial a_{m-1} \partial b_1} = \sum_{i=1}^{n} \frac{(e_i)(x_{1i})(x_{(m-1)i})}{Q_i}$$

When $m = k+2$, $n = k+2$:

$$I_{(m,n)} = I(b_0, b_0) = -\frac{\partial^2 l}{\partial b_0 \partial b_0} = \frac{1}{2} \sum_{i=1}^{n} \frac{e_i^2}{Q_i}$$

When $m = k+2$, $n = k+3$:

$$I_{(m,n)} = I(b_0, b_1) = -\frac{\partial^2 l}{\partial b_0 \partial b_1} = \frac{1}{2} \sum_{i=1}^{n} \frac{(e_i^2)(x_{1i})}{Q_i}$$

When $m = k+3$, $n = k+3$:

$$I_{(m,n)} = I(b_1, b_1) = -\frac{\partial^2 l}{\partial b_1 \partial b_1} = \frac{1}{2} \sum_{i=1}^{n} \frac{(e_i^2)(x_{1i}^2)}{Q_i}$$

Notes: Given that $I_{\text{obs}}(\hat{\theta})$ is symmetric, we present the formulas for the elements in the upper triangular portion of I_{obs}. The elements $I_{n,m}$ in the lower triangular portion of I_{obs} can be calculated as $I_{(n,m)} = I_{(m,n)}$, where $m \leq n$.

6.2 Other Simulation Result Figures

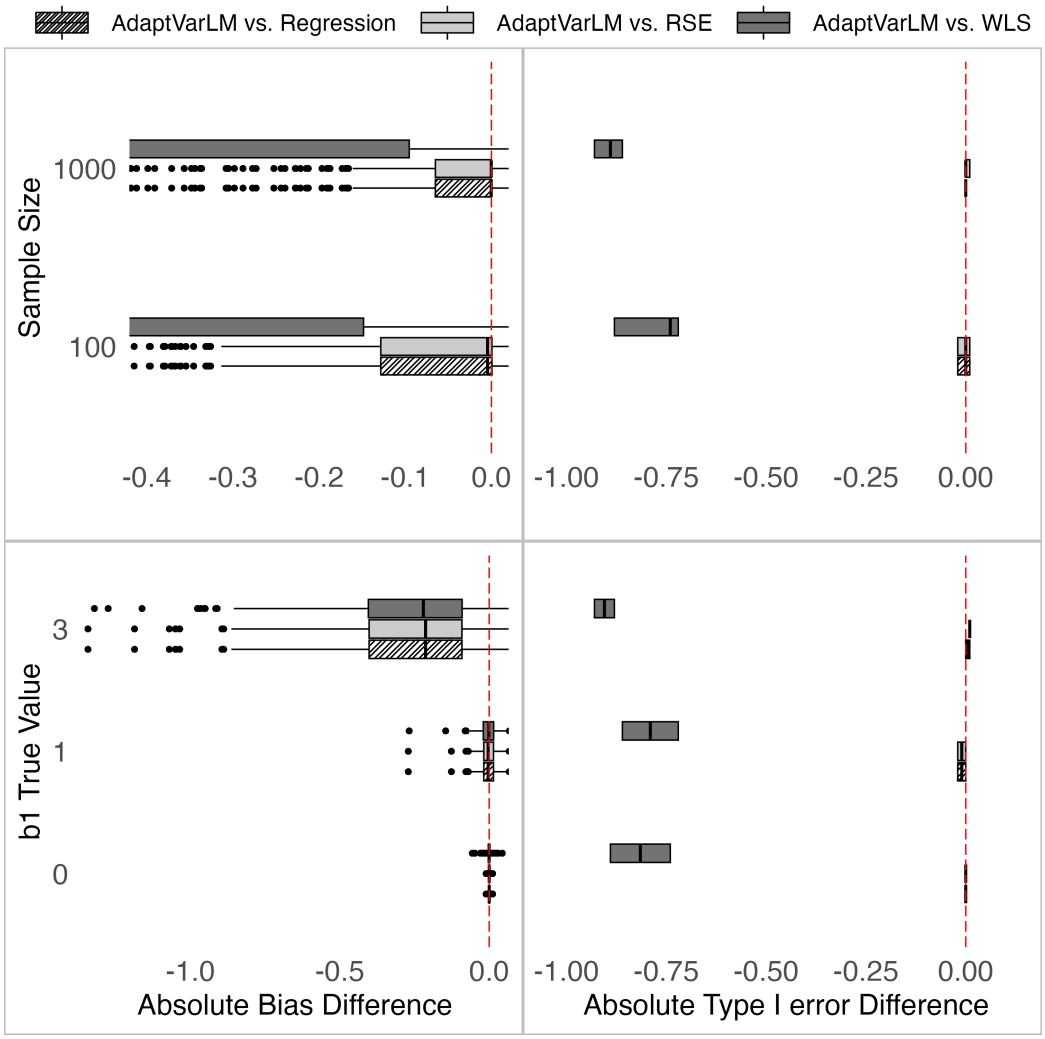


Figure 6.1: The Boxplots of Three Criteria for Parameter $a0$ across Sample Sizes and $b1$ True Values

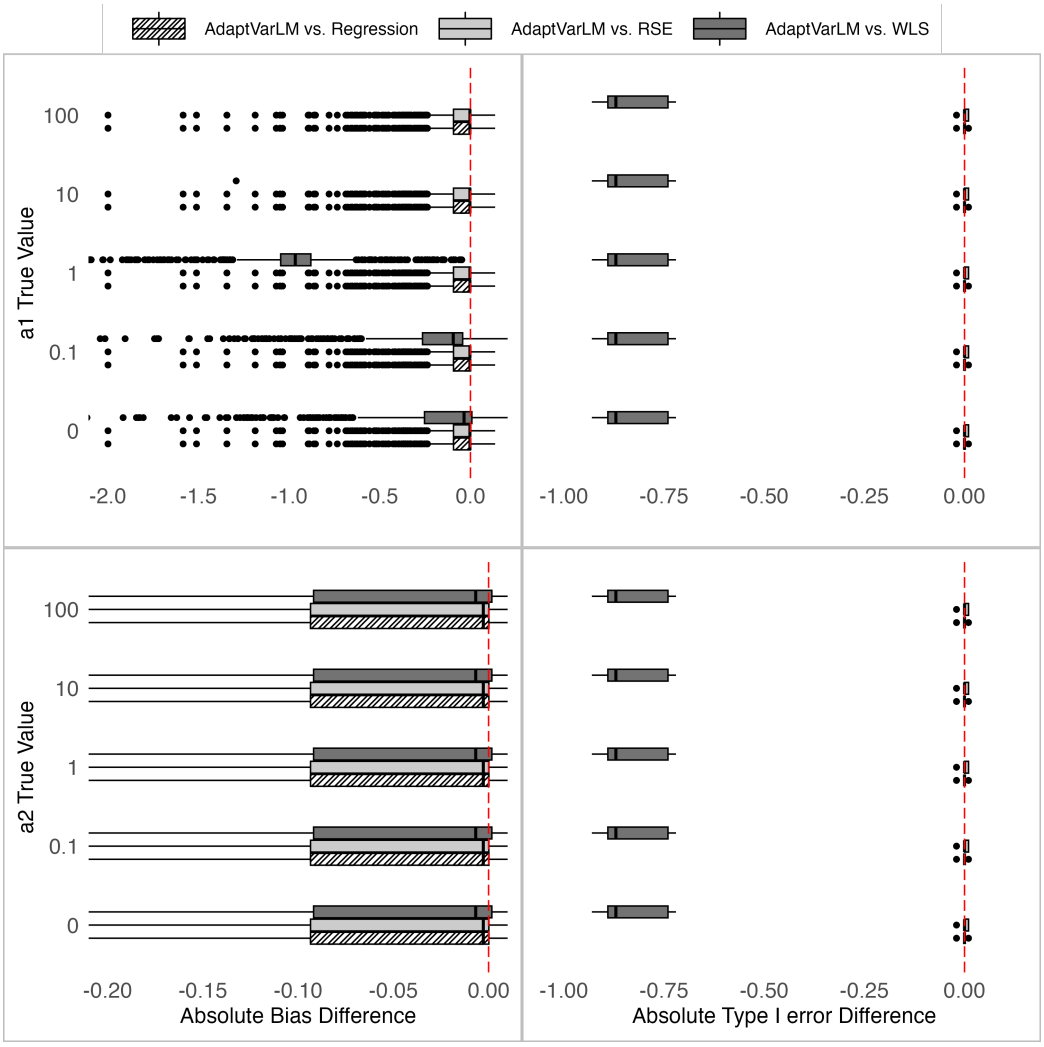


Figure 6.2: The Boxplots of Three Criteria for Parameter $a0$ across $a1$ True Values and $a2$ True Values

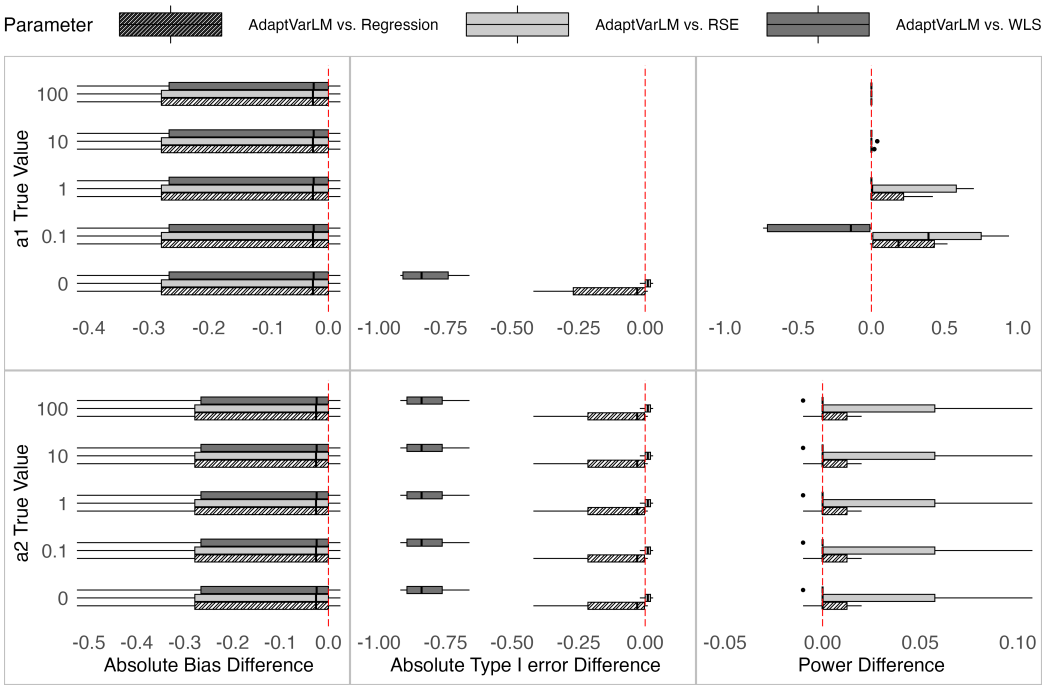


Figure 6.3: The Boxplots of Three Criteria for Parameter $a1$ across $a1$ True Values and $a2$ True Values

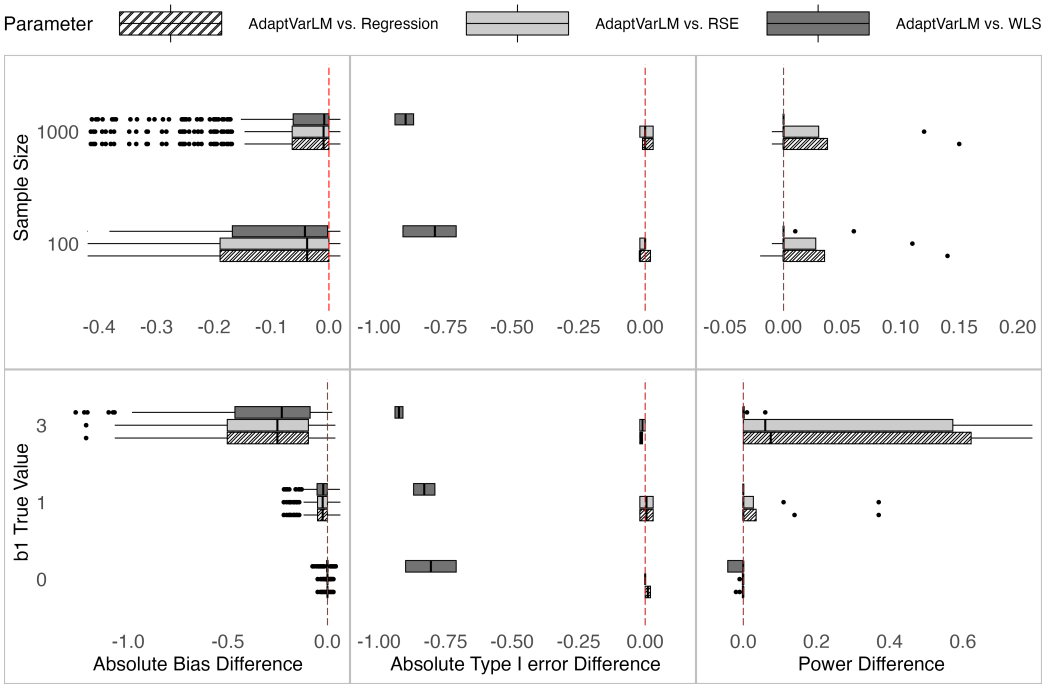


Figure 6.4: The Boxplots of Three Criteria for Parameter $a2$ across Sample Sizes and $b1$ True Values

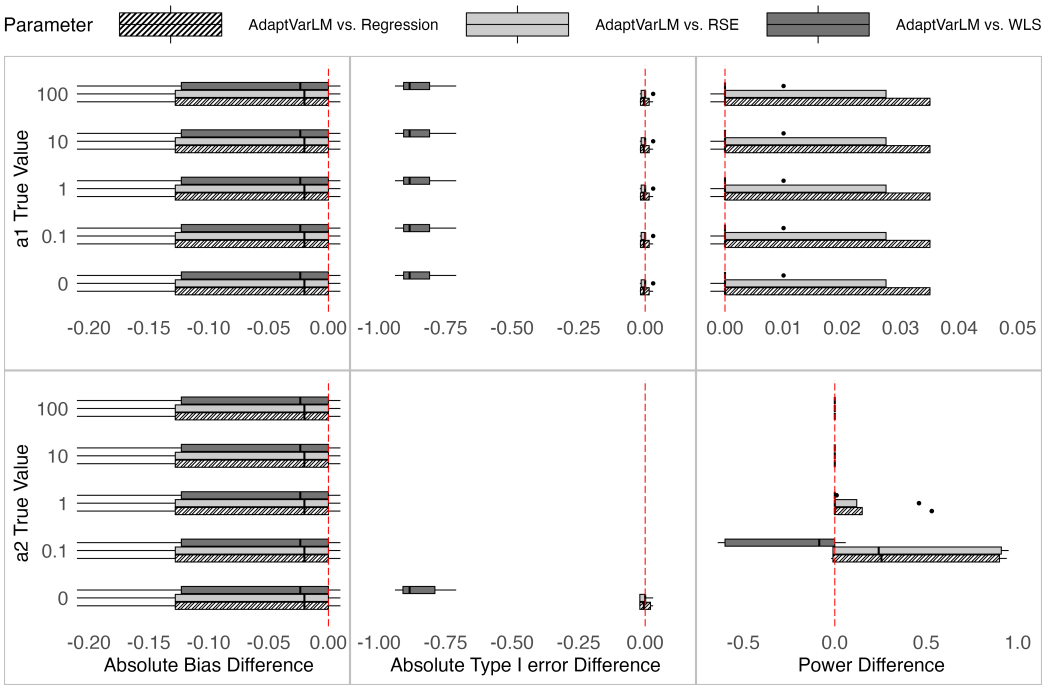


Figure 6.5: The Boxplots of Three Criteria for Parameter $a2$ across $a1$ True Values and $a2$ True Values

Bibliography

[1] David A Bennett, Aron S Buchman, Patricia A Boyle, Lisa L Barnes, Robert S Wilson, and Julie A Schneider. Religious orders study and rush memory and aging project. *Journal of Alzheimer's disease*, 64(s1):S161–S189, 2018.

[2] David A Bennett, Julie A Schneider, Neelum T Aggarwal, Zoe Arvanitakis, Raj C Shah, Jeremiah F Kelly, Jacob H Fox, Elizabeth J Cochran, Danielle Arends, Anna D Treinkman, et al. Decision rules guiding the clinical diagnosis of alzheimer's disease in two community-based cohort studies compared to standard practice in a clinic-based cohort study. *Neuroepidemiology*, 27(3):169–176, 2006.

[3] George E. P. Box. Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, 30(1):1–17, 1988.

[4] Rush Alzheimer's Disease Center. Clinical diagnosis: Dementia - variable detail: dcfdx. `https://www.radc.rush.edu/docs/var/detail.htm;jsessionid=9F662F621EBFE4672368AFEFCB743F2E?category=Clinical+Diagnosis&subcategory=Dementia&variable=dcfdx`, 2024. Accessed: 2024-08-27.

[5] Marie Davidian and Perry D Haaland. Regression and calibration with nonconstant error variance. *Chemometrics and Intelligent Laboratory Systems*, 9(3):231–248, 1990.

[6] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

[7] Herbert Glejser. A new test for heteroskedasticity. *Journal of the American Statistical Association*, 64(325):316–323, 1969.

[8] John A Jacquez, Frances J Mather, and Charles R Crawford. Linear regression with non-constant, unknown error variances: sampling experiments with least squares, weighted least squares and maximum likelihood estimators. *Biometrics*, pages 607–626, 1968.

[9] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[10] Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[11] John A. Nelder and Roger Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[12] Silian Shen and Changlin Mei. Estimation of the variance function in heteroscedastic linear regression models. *Communications in Statistics—Theory and Methods*, 38(7):1098–1112, 2009.

[13] Philip B Verghese, Joseph M Castellano, and David M Holtzman. Apolipoprotein e in alzheimer's disease and other neurological disorders. *The Lancet Neurology*, 10(3):241–252, 2011.

[14] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[15] Yan Xu, Li Xing, Jessica Su, Xuekui Zhang, and Weiliang Qiu. Model-based clustering for identifying disease-associated snps in case-control genome-wide association studies. *Scientific Reports*, 9:1–10, 09 2019.

[16] Tingting Yu, Shangyuan Ye, and Rui Wang. High-dimensional variable selection accounting for heterogeneity in regression coefficients across multiple data sources. *Canadian Journal of Statistics*, 2023.

[17] Achim Zeileis. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004.