Predictive Power and Efficient Sample Size in Linear Regression Models

by

Yifan Ma

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

by

May 2023

APPROVED:

Professor Zheyang Wu, Major Thesis Advisor

Abstract

This work explores the relationship between prediction accuracy, the impact of additional predictors, and sample size in the context of multiple linear regression models. The objective is to facilitate sample size calculations for study designs that directly target predictive power (i.e., prediction accuracy) in various applicational studies. To achieve this goal, we analyze the functional relationship between prediction mean square error (PMSE) and factors such as the number, effect sizes, and correlations among predictors, as well as sample size.

Building on this analysis, we introduce a metric referred to as the percentage of PMSE reduction (pPMSEr) to quantify the improvement in prediction accuracy when sample size is increased and/or new important predictors are added to a model. Given a set of predictors, we can compute an efficient sample size, defined as the smallest sample size that achieves, for example, 90% of the prediction accuracy ever achievable at an infinite sample size. Beyond this efficient sample size, increasing the sample size does not significantly improve prediction accuracy unless more important predictors are incorporated into the model.

We validate these calculations through computations and simulations based on a pain study, demonstrating a practical application and interpretation of the proposed measures in planning prediction-related studies.

Acknowledgements

I would like to express my sincerest gratitude to my thesis advisor, Zheyang Wu, for his unwavering support, encouragement, and guidance throughout my research journey. Their valuable insights, constructive feedback, and expertise have been instrumental in shaping the direction and scope of this thesis.

I am also thankful to the faculty members of the Mathematical Science Department for providing me with a stimulating academic environment and for sharing their knowledge and expertise, which have been invaluable in enriching my research.

I would like to thank my family for their constant love, encouragement, and support, which have been a source of strength and motivation throughout my academic journey.

Thank you all for your invaluable contributions to this thesis.

Contents

List of Tables

Chapter 1

Introduction

Predictive modeling is playing an increasingly important role in many fields, including finance, healthcare, and marketing. However, achieving high predictive accuracy is a complex task that depends on multiple factors, including the number of predictors and their effect size, as well as the sample size used to train and test the model.

The focus of this thesis is to investigate the relationship between prediction accuracy, the number of predictors, their effect sizes, and the sample size. Our goal is to establish recommendations for designing studies that aim to enhance a model's predictive power by identifying new predictors. To accomplish this, we will measure the impact of various factors that contribute to the PMSE.

The importance of sample size in predictive modeling has been recognized for several decades. In 1974, Narula (1974)explored the relationship between predictive mean square error and stochastic regressor variables. Since then, numerous studies have been conducted to determine the optimal sample size for different types of outcomes and model structures. For instance, Riley et al. (2019a), Riley et al. (2019b)published the minimum sample size required for developing multivariable prediction models for continuous, binary, and time-to-event outcomes. van der Ploeg et al. (2014) conducted a simulation study to determine the sample size requirements for predicting dichotomous endpoints, and van Smeden et al. (2019) discussed the limitations of the events per variable criteria for sample size determination in binary logistic prediction models. These studies have contributed to our understanding of the importance of sample size in predictive modeling and provided valuable insights into determining the optimal sample size based on the outcome of interest and model structure.

While a larger sample size can lead to more accurate predictions, there is a limit to the effect of sample size on model performance. One reason for this limit is that beyond a certain point, adding more samples may not provide much additional information. For example, if a dataset already contains representative samples of the population, collecting more samples may not significantly improve the model's predictive ability. Another reason is that the relationship between the input variables and the target variable may not be linear or may have a limited range. In such cases, adding more data may not improve model performance beyond a certain point. Additionally, the complexity of the model and the quality of the features used in the model can also impact its predictive ability. It is essential to acknowledge that the true predictive model remains unknown and that achieving 100% prediction accuracy may not be possible, even with knowledge of the effect sizes of the predictors. The uncertainty of the error term can still have an impact on the model's accuracy, which is why it's essential to consider other factors in addition to sample size to ensure high prediction accuracy.

Therefore, while the sample size is an important consideration in predictive modeling, its effect on model performance is limited, and other factors such as model complexity and feature quality also play a crucial role. On the other hand, new predictors can provide additional information about the outcome variable, potentially capturing previously unknown or unmeasured factors that are related to the outcome. This can help to increase the model's predictive power and provide a more comprehensive understanding of the factors that contribute to the outcome. In practical medical reports, the effect size of newly added predictors usually only considers this new variable and controls the effect size of other existing variables.

This thesis aims to investigate the relationship between prediction accuracy and newly developed predictors considering sample size. Through analyzing the impact of various factors on PMSE, we will develop guidelines for study designs that can improve the accuracy and reliability of research results. We will not only investigate the relationship between sample size, number of predictors, and effect sizes on the PMSE but also summarize the degree of influence of new predictors on the model's prediction accuracy. To quantify this influence, we will use the reduced prediction mean square error percentage $rPMSEp$ or Correlation between true value and prediction, which corresponds to the efficient sample size required to achieve a certain prediction accuracy. We will provide calculations based on the pain study to support our findings.

Chapter 2

Literature Review

Multiple regression analysis is a commonly used method of predictive modeling, which involves using several predictor variables to make predictions about a dependent variable. However, when the predictor variables are stochastic or random, the accuracy of the predictions can be compromised. This chapter reviews three seminal studies that investigate the impact of stochastic predictor variables on the accuracy of predictions made using multiple regression equations.

Kerridge (1967)focuses on examining the predictive errors of multiple regression equations when the predictor variables are treated as random variables drawn from a multivariate normal population. The conventional treatment of multiple regression assumes independent variables as constants, but in many practical applications, it is more reasonable to consider them as random variables. The paper suggests that despite the limitations of using regression or stochastic predictor variables, it can still be useful if the limitations are well-understood, particularly when dealing with a large number of predictor variables. Kerridge investigates the prediction error that can occur when using multiple regression with stochastic predictor variables. Kerridge shows that the prediction error can be approximated distributionally by the product of a standard normal variable and the square root of an independent Beta-variate. The mean square can be approximated by

$$PMSE = \sigma^2 \left(1 + \frac{1}{n}\right) \left(\frac{n-2}{n-k-2}\right), \tag{2.1}$$

where σ^2 is the variance of the error term in the regression, n is the sample size, k is the number of predictors. The paper provided a widely used approximation of $PMSE$.

Narula (1974) discusses the problem of variable selection in regression analysis. Although this topic has been thoroughly researched in the literature, most previous studies have focused on selecting subsets of predictor variables that are treated as fixed. Narula proposes a decision rule for selecting a subset of predictor variables that are stochastic, which leads to a smaller prediction mean squared error (PMSE) compared to the conventional approach. Specifically, Narula shows that the PMSE can be decomposed into the variance of the regression coefficients and the variance of the random errors in the case of stochastic predictor variables. By quantifying these two sources of variance, Narula provides a method for assessing the accuracy of predictions made using stochastic predictor variables. The article presents a proof for the derived subset PMSE formula:

$$PMSE_{subset} = \sigma_p^2 \left(1 + \frac{1}{n}\right) \left(\frac{n-2}{n-k-2}\right), \tag{2.2}$$

where σ_p^2 represents the error term for the subset regression, p is the number of predictors in the subset. The proof is provided in the Appendix, as Narula (1974) only presents the conclusion without the supporting details. Overall, Narula's approach is a valuable contribution to variable selection in regression analysis, especially when working with stochastic predictor variables, as it leads to more accurate predictions.

Sawyer (1982) focuses on cases where one is interested in specifying prediction accuracy before collecting sample data, predictors are not subject to experimental control, and one is interested in specifying prediction accuracy averaged with respect to the distribution of the predictors.Sawyer (1982) explores the accurate approximate distribution of mean absolute error (MAE) as a combination of normal distribution and derivative of the normal distribution as

$$Prob(\hat{y} - y \leq t) = \Phi\left(\frac{t}{\sigma'}\right) + \frac{p}{4(n-2)(n-p-4)}\Phi^{(4)}\left(\frac{t}{\sigma'}\right), \qquad (2.3)$$

where $\hat{y}$ is the prediction of response, y is the corresponding true value of response,Φ is the standard normal distribution function, $\Phi^{(4)}$ is its fourth derivative, $\sigma' = \sqrt{MSE}$, . The approximation has been significantly proven to be better than a simple normal approximation. In the discussion section, Sawyer argues that as the number of predictor variables increases, the sample size required to achieve a given level of accuracy also increases. The approximate inflation in MAE due to the regression coefficients is a simple function of the base sample size and the number of predictors. The advantage of using MAE is that it has a simple definition, making it easily understandable to people who use prediction equations but have little statistical training.

The three seminal studies investigate the impact of stochastic predictor variables on the accuracy of predictions made using multiple regression equations. These studies propose different approaches for handling stochastic predictor variables in regression analysis, including approximations for prediction, mean squared error, and mean absolute error. These approaches are useful for improving the accuracy of predictions in practical applications, particularly when dealing with a large number of predictor variables. Additionally, the studies provide valuable insights into variable selection in regression analysis and predicting accuracy before collecting sample data.

Chapter 3

Analytical Result

3.1 Linear Model and Properties

We consider a sample size of n independent individuals for k predictors. For each individual $i = 1, ..., n$, we have a vector of all predictors $\boldsymbol{z_i'} = (z_{i1}, ..., z_{ik})$. We define the design matrix as $\boldsymbol{Z} = (\boldsymbol{z_1'}, ..., \boldsymbol{z_n'})'$.

Among the total k predictors, p of them are seen as basic predictors, corresponding to $\boldsymbol{z_{1i}'} = (z_{i1}, ..., z_{ip})$. The remaining $k - p$ predictors are called non-basic predictors (i.e., factors to be discovered in a newly proposed study), corresponding to $\boldsymbol{z_{2i}'} = (z_{i,(p+1)}, ..., z_{ik})$. Therefore, the partition can be written as $\boldsymbol{z_i'} = (\boldsymbol{z_{1i}'}, \boldsymbol{z_{2i}'})$.

We assume that the response and predictor variables follow a multivariate normal distribution, i.e.,

$$(y_i, \boldsymbol{z_i'})' \sim MVN(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),$$

where $\boldsymbol{\mu} = (\mu_0, \boldsymbol{\mu}')'$ and $\boldsymbol{\Sigma}^*$ is the covariance matrix is defined by

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \sigma_{00} & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{pmatrix},$$

where $\boldsymbol{\sigma}$ is the covariance vector, $\boldsymbol{\Sigma}$ is the covariance matrix among predictor variables, and σ_{00} is the variance of response variable $\boldsymbol{y}$. We further partition the covariance vector as $\boldsymbol{\sigma} = Cov(y_i, \boldsymbol{z_i}) = (\boldsymbol{\sigma_1'}, \boldsymbol{\sigma_2'})'$, and the variance matrix as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Based on the distribution of the random variables $(y_i, \boldsymbol{z_i'})'$, we introduce the full regression model as follows:

$$y_i = \alpha + \boldsymbol{z_i'}\boldsymbol{\beta} + \epsilon_i.$$

The error term ϵ_i follows a normal distribution $N(0, \sigma_k^2)$,

$$\sigma_k^2 = \sigma_{00} - \boldsymbol{\sigma'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}, \tag{3.1}$$

which is independent of $\boldsymbol{z_i}$. Additionally, $\boldsymbol{\beta}$ denotes the full-model effects and is given by:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}. \tag{3.2}$$

where $\boldsymbol{\sigma}$ and $\boldsymbol{\Sigma}$ are the covariance vector and covariance matrix, respectively, of the random variables $(y_i, \boldsymbol{z_i'})'$. In particular, $\boldsymbol{\sigma}$ is the vector of covariances between y_i and each element of $\boldsymbol{z_i}$, while $\boldsymbol{\Sigma}$ contains the covariances between the elements of $\boldsymbol{z_i}$.

Similarly, we can partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ denote the effects of the predictors in the first and second subsets of $\boldsymbol{z_i}$, respectively. Using matrix algebra, we obtain the following expressions for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

$$\begin{aligned}\boldsymbol{\beta}_1 &= (\boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}})^{-1}[\boldsymbol{\sigma}_1 - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\sigma}_2];\\ \boldsymbol{\beta}_2 &= (\boldsymbol{\Sigma_{22}} - \boldsymbol{\Sigma_{21}}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\Sigma_{12}})^{-1}[\boldsymbol{\sigma}_2 - \boldsymbol{\Sigma_{21}}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma}_1].\end{aligned} \tag{3.3}$$

Furthermore, based on the distribution of the random variables $(y_i, \boldsymbol{z'_{1i}})'$, we consider the reduced regression model:

$$y_i = \alpha + \boldsymbol{z'_{1i}}\boldsymbol{\beta}_1^{\sharp} + \epsilon_i^{\sharp}, \tag{3.4}$$

where $\epsilon_i^{\sharp}$ follows a normal distribution $N(0, \sigma_p^2)$,

$$\sigma_p^2 = \sigma_{00} - \boldsymbol{\sigma'_1}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma_1}, \tag{3.5}$$

which is independent of $\boldsymbol{z_{1i}}$. In this case, the effects of the predictors in the first subset of $\boldsymbol{z_i}$ are denoted by $\boldsymbol{\beta}_1^{\sharp}$, and are given by:

$$\boldsymbol{\beta}_1^{\sharp} = \boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma_1}, \tag{3.6}$$

note that $\epsilon_i^{\sharp}$ is not independent of $\boldsymbol{z_{2i}}$ and they are multivariate normal.

We can obtain the joint effects based on the marginal effects. Specifically, consider the marginal model regarding the jth predictor, $j = 1, ..., k$,

$$y_i = \alpha + z_{ij}\beta_j^* + \epsilon_i^*. \tag{3.7}$$

We have $\sigma_j = Cov(y_i, z_{ij}) = \Sigma_{jj}\beta_j^*$. Denote the vector of the marginal coefficients/effects $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_k^*)'$. We have

$$\boldsymbol{\sigma} = (\Sigma_{11}\beta_1^*, ..., \Sigma_{kk}\beta_k^*)' = \text{diag}(\boldsymbol{\Sigma})\boldsymbol{\beta}^*. \tag{3.8}$$

Following Equation(3.8), the coefficients/effects in joint models Equation(3.2) and Equation(3.6) can be obtained.

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\text{diag}(\boldsymbol{\Sigma})\boldsymbol{\beta}^*; \tag{3.9}$$

$$\boldsymbol{\beta}_1^\sharp = \boldsymbol{\Sigma}_{11}^{-1}\text{diag}(\boldsymbol{\Sigma}_{11})\boldsymbol{\beta}_1^*. \tag{3.10}$$

In practice, β_j^* and $\boldsymbol{\Sigma}$ may come from literature, prior studies or be estimated by data.

3.2 Percentage of PMSE Reduction

The predictive mean squared error (PMSE) based on the least squares estimator (LSE) of the full regression can be calculated using the formula presented by Kerridge in Kerridge (1967). This formula is as follows:

$$PMSE = E(y_0 - \hat{y}_0)^2 = \sigma_k^2 \frac{(n+1)(n-2)}{n(n-k-2)}. \tag{3.11}$$

Here, y_0 is the true response value, $\hat{y}_0$ is the estimated response value, σ_k^2 is the error variance, and n and k are the sample size and the number of predictors, respectively. Similarly, the PMSE based on the LSE of the reduced regression can be calculated using the following formula:

$$PMSE_1 = E(y_0 - \tilde{y}_0)^2 = \sigma_p^2 \frac{(n+1)(n-2)}{n(n-p-2)}. \tag{3.12}$$

Here, $\tilde{y}_0$ is the estimated response value based on the reduced regression, and σ_k^2 is the error variance of the reduced regression. The present study includes an investigation into the approximation distribution for PMSE, which, although not utilized in the current analysis, is nevertheless explicated comprehensively with supporting evidence, proof, and simulation results, all of which are provided in the Appendix.

To measure the improvement in the prediction that occurs by adding predictors, we use the percentage of PMSE reduction measure, which is given by the following equation:

$$pPMSEr = \left(\frac{PMSE_1 - PMSE}{PMSE_1} \right) \times 100\% = \left(1 - \frac{\sigma_k^2}{\sigma_p^2} \cdot \frac{n-p-2}{n-k-2} \right) \times 100\%. \tag{3.13}$$

which calculates the percentage difference between the PMSE of the reduced regression and the PMSE of the full regression. The percentage of PMSE reduction can also be expressed in terms of the error variance ratio (EVR), which is defined as the ratio of the error variances of the full and reduced regressions:

$$EVR = \frac{\sigma_k^2}{\sigma_p^2}.$$

Additionally, an inflation factor called $\lambda(n;p,p_2)$ can be defined as follows:

$$\lambda(n;p,p_2) = \frac{n-p-2}{n-k-2} = \frac{1}{1-\frac{k-p}{n-p-2}} = \frac{1}{1-\frac{p_2}{n-p-2}},$$

where p_2 represents the number of non-basic predictors added to the reduced model to form the full model. The inflation factor $\lambda(n;p,p_2)$ is related to estimation error and uncertainty (analog to the inflation factor $K = \frac{(n+1)(n-2)}{n(n-k-2)}$ in Sawyer (1982)). It should be noted that the percentage reduction measure can be negative, i.e.,

$E(y_0 - \hat{y}_0)^2 > E(y_0 - \tilde{y}_0)^2$, in certain situations, such as when the new and basic predictors are negatively correlated.

3.3 Efficient Sample Size

The reduction of prediction mean squared error (PMSE) in Equation(3.13) is determined by several factors, including the sample size n, the number of known predictors p, and the number of new predictors $k - p$. A larger sample size n tends to make the inflation factor $\lambda(n; p, p_2)$ approach 1, particularly while k and p remain fixed. This reflects the fact that the sample size has a limited impact on PMSE in the absence of estimation error or uncertainty.

When considering the influence of the number of new predictors p_2 on PMSE, it is important to note that while increasing p_2 tends to increase the inflation factor, it may also decrease the full model variance σ_k^2. It is therefore desirable to have the reduction in variance outweigh the increase in inflation, assuming the effect sizes of the new predictors are not too small. However, if the new predictors have no effect (i.e., are false), the inflation increases while σ_k^2 remains constant, resulting in an increase in PMSE. From a sample size perspective, a larger n is needed to effectively control inflation when a small or moderate number of false predictors are present. In other words, at any given p_2, a sufficiently large sample size can always ensure that the inflation factor is equal to 1. Assuming that the addition of new predictors always results in a decrease in σ_k^2 relative to σ_p^2, a large enough sample size is needed to ensure that the addition of new predictors does not worsen prediction accuracy.

In this study, we observe that the pPMSEr is increasing as the sample size n increases. To define an efficient sample size, denoted as n^*, we seek the smallest sample size such that the ratio of pPMSEr at n^* to pPMSEr at infinity is greater than or equal to a specified efficiency level of $1-\alpha$, where α represents the complement of the efficiency level (e.g., 90% of the largest pPMSEr at $n=\infty$). We can determine the efficient sample size by using the equations:

$$\frac{pPMSEr(n^*)}{pPMSEr(\infty)} \geq 1-\alpha,$$

By

$$\frac{1-EVR\cdot\lambda^*}{1-EVR} = 1-\alpha \text{ and } \lambda^* = \frac{1}{1-\frac{p_2}{n^*-p-2}},$$

Solving for λ^* and substituting into the first equation yields:

$$n^* = p+2+p_2\cdot\frac{\lambda^*}{\lambda^*-1}, \text{ where } \lambda^* = 1+\alpha(\frac{1}{EVR}-1). \tag{3.14}$$

Equation(3.14) gives us the efficient sample size n^*, which we can use to obtain accurate predictions while minimizing sample size and associated costs.

3.4 Effect sizes

The reduction of PMSE in Equation(3.13) is determined by the variances σ_k^2 and σ_p^2, which are relavent to the effect sizes and covariances of the predictors. There are various measures and interpretations regarding effect sizes. One representative measure is Cohen's f^2, which is based on the proportion of the variation explained by the predictors. Let the proportion of the response's variance accounted for by all k predictors defined by

$$R^2 = \frac{\sigma_{00}-\sigma_k^2}{\sigma_{00}} = \frac{\boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}}{\sigma_{00}},$$

and the proportion of the response's variance accounted for by p basic predictors

$$R_1^2 = \frac{\sigma_{00} - \sigma_p^2}{\sigma_{00}} = \frac{\boldsymbol{\sigma}'_{\mathbf{1}}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}\mathbf{1}}{\sigma 00}.$$

Cohen's f^2 for the effects of all predictors is

$$f^2 = \frac{R^2}{1 - R^2},$$

while Cohen's f^2 for the effects of new predictors conditional on the basic predictors is given by

$$f_2^2 = \frac{R^2 - R_1^2}{1 - R^2} = \frac{\sigma_p^2 - \sigma_k^2}{\sigma_k^2} = \frac{1 - \sigma_k^2/\sigma_p^2}{\sigma_k^2/\sigma_p^2}. \tag{3.15}$$

which yields

$$\frac{\sigma_k^2}{\sigma_p^2} = \frac{1}{f_2^2 + 1}. \tag{3.16}$$

Another meaningful measure is regression coefficients, which provide a practical interpretation based on the original data scale (unstandardized measures). Using the joint or marginal regression coefficients, we can calculate f_2^2 from Equation(3.15) and Equation(3.17) as follows:

$$\frac{\sigma_k^2}{\sigma_p^2} = \frac{\sigma_{00} - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}}{\sigma_{00} - \boldsymbol{\sigma}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_1} = \frac{\sigma_{00} - \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{\sigma_{00} - \boldsymbol{\beta}_1^{\sharp\prime}\boldsymbol{\Sigma}_{11}\boldsymbol{\beta}_1^{\sharp}} = \frac{\sigma_{00} - \boldsymbol{\beta}^{*\prime}\mathrm{diag}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathrm{diag}(\boldsymbol{\Sigma})\boldsymbol{\beta}^{*}}{\sigma_{00} - \boldsymbol{\beta}_1^{*\prime}\mathrm{diag}(\boldsymbol{\Sigma}_{11})\boldsymbol{\Sigma}_{11}^{-1}\mathrm{diag}(\boldsymbol{\Sigma}_{11})\boldsymbol{\beta}_1^{*}}. \tag{3.17}$$

Chapter 4

Example

4.1 Data

The study utilizes the analytical result outlined in the previous section. The dataset utilized in this study was based on the findings reported by pain study Baker TA (2008). However, as the paper only provided a correlation matrix, a standard deviation of $\mathbf{SD} = \mathbf{1}$ was used to derive the covariance matrix.

In the calculation, the full regression model included $k = 12$ predictors, while the reduced regression model included the first $p = 3$ predictors. The basic predictors included in the full and reduced regression models were Age, Education, and Income. The remaining $p_2 = k - p = 9$ predictors were classified as non-basic and included Comorbidities, Pain locations, Medications, Physical functioning, Depressive symptoms, Life satisfaction, LOC-chance, LOC-powerful, and LOC-internal.

Given the correlation matrix provided by Baker TA (2008), and assuming the covariates were standardized, the covariance matrix can be obtained using the formula: $\mathbf{\Sigma}^* = diag(\boldsymbol{SD}) \times \mathbf{Cor} \times diag(\boldsymbol{SD})$, where $\mathbf{Cor}$ is the correlation matrix given and $\mathbf{SD}$ is standard deviation vector. The response variable, Pain intensity, and the

$k = 12$ predictors were assumed to follow a multivariate normal distribution with mean $\boldsymbol{\mu}^* = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma^*}$, which is presented below in Equation(4.1).

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} 1 & -0.24 & 0.00 & -0.03 & 0.45 & 0.33 & 0.26 & 0.39 & -0.21 & -0.05 & 0.10 & 0.16 & 0.34 \\ & 1 & -0.21 & -0.05 & -0.27 & -0.21 & -0.09 & 0.00 & 0.27 & 0.09 & 0.34 & 0.00 & -0.05 \\ & & 1 & 0.46 & -0.19 & 0.00 & -0.19 & -0.14 & 0.12 & -0.24 & -0.29 & -0.02 & -0.13 \\ & & & 1 & -0.30 & -0.04 & -0.18 & -0.16 & 0.16 & -0.05 & -0.02 & 0.07 & -0.10 \\ & & & & 1 & 0.20 & 0.64 & 0.34 & -0.14 & 0.20 & 0.00 & 0.03 & 0.14 \\ & & & & & 1 & 0.33 & 0.34 & -0.07 & 0.02 & -0.1 & -0.07 & 0.11 \\ & & & & & & 1 & 0.46 & -0.25 & 0.15 & -0.05 & -0.07 & 0.18 \\ & & & & & & & 1 & -0.17 & 0.13 & 0.13 & -0.03 & 0.26 \\ & & & & & & & & 1 & -0.03 & 0.08 & 0.03 & -0.57 \\ & & & & & & & & & 1 & 0.68 & 0.19 & 0.10 \\ & & & & & & & & & & 1 & 0.20 & 0.09 \\ & & & & & & & & & & & 1 & 0.05 \\ & & & & & & & & & & & & 1 \end{bmatrix}. \tag{4.1}$$

4.2 Calculation

The relationship between $pPMSEr$, sample size, and Cohen's f^2 was examined. According to Baker TA (2008), the predictors were categorized into demographic, health, and psychological factors. The reduced regression model only included demographic factors, while the full regression model considered both health and psychological factors while controlling for the reduced predictors.

The variances of the error terms in the full and reduced regression models were calculated as follows:

$$\sigma_k^2 = \sigma_{00} - \boldsymbol{\sigma'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} = 0.4687399, \sigma_p^2 = \sigma_{00} - \boldsymbol{\sigma_1'}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma_1} = 0.9393167.$$

Here, σ_{00} represents the variance of the response, which was standardized to 1.

4.2.1 Predictor Effects

The full model effects were obtained using Equation(3.2), and the calculated result is as shown in Equation(4.2).

$$\begin{aligned}\boldsymbol{\beta} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} \\ &= (-0.13, 0.07, 0.10, 0.50, 0.23, -0.15, 0.18, -0.05, -0.47, 0.43, 0.14, 0.21).\end{aligned} \tag{4.2}$$

Notably, the coefficients calculated in this study differ from Equation(4.3) presented in Table 2 of Baker TA (2008) as Equation(4.3),

$$\boldsymbol{\beta}^{*} = (-0.20, -0.03, -0.02, -0.04, 0.12, 0.18, 0.26, 0.25, -0.01, 0.08, -0.26, 0.21), \tag{4.3}$$

as the Equation(4.2) considers all predictors in the full model, whereas Equation(4.3) was computed with added-up predictors controlling for prior sets of predictors.

On the other hand, the reduced-model effects, given in Equation(4.4), were calculated using only the three demographic basic predictors. The effects for demographic factors in the reduced-model $\boldsymbol{\beta}_1^{\sharp}$ were found to be closer to Equation(4.3) in the coefficients presented in Baker TA (2008) than the calculated effects Equation(4.2) in the full model effects, as adding all predictors in the model would change the effect of previous predictors, especially when the predictors are not significant.

$$\begin{aligned}\boldsymbol{\beta}_1^{\sharp} &= \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_1 \\ &= (-0.25, -0.04, -0.02).\end{aligned} \tag{4.4}$$

4.2.2 pPMSEr

The prediction mean square error is used to measure the prediction accuracy and the calculated PMSE is as shown in Table 4.1. The table provides the calculated

PMSE by Equation(3.11) and Equation(3.12) by adding predictors sequentially for each sample size from 30 to 600.

Sample Size	Basic Model	Comorbidities	Pain Locations	Medications	Physical Functioning	Depressive Symptoms	Life Satisfaction	LOC chance	LOC powerful	LOC internal
30	1.0871	0.9322	0.9073	0.9366	0.8996	0.9121	0.9280	0.8318	0.8427	0.8476
60	1.0071	0.8444	0.8025	0.8075	0.7549	0.7436	0.7333	0.6357	0.6212	0.6009
90	0.9833	0.8191	0.7732	0.7727	0.7172	0.7013	0.6863	0.5903	0.5721	0.5488
120	0.9719	0.8071	0.7595	0.7565	0.6999	0.6820	0.6652	0.5701	0.5506	0.5262
150	0.9651	0.8001	0.7515	0.7472	0.6899	0.6710	0.6531	0.5587	0.5384	0.5135
180	0.9607	0.7954	0.7462	0.7411	0.6834	0.6638	0.6454	0.5514	0.5307	0.5054
210	0.9576	0.7922	0.7425	0.7368	0.6789	0.6588	0.6400	0.5462	0.5252	0.4998
240	0.9553	0.7898	0.7398	0.7336	0.6755	0.6552	0.6360	0.5425	0.5213	0.4957
270	0.9535	0.7879	0.7377	0.7311	0.6729	0.6523	0.6329	0.5396	0.5182	0.4925
300	0.9520	0.7864	0.7360	0.7292	0.6709	0.6501	0.6304	0.5373	0.5158	0.4900
330	0.9509	0.7852	0.7346	0.7276	0.6692	0.6482	0.6285	0.5354	0.5138	0.4880
360	0.9499	0.7842	0.7335	0.7263	0.6678	0.6467	0.6268	0.5339	0.5122	0.4863
390	0.9491	0.7833	0.7326	0.7252	0.6666	0.6454	0.6254	0.5326	0.5109	0.4849
420	0.9484	0.7826	0.7317	0.7242	0.6656	0.6443	0.6243	0.5315	0.5097	0.4837
450	0.9478	0.7820	0.7310	0.7234	0.6648	0.6434	0.6232	0.5305	0.5087	0.4827
480	0.9472	0.7814	0.7304	0.7227	0.6640	0.6426	0.6224	0.5297	0.5078	0.4818
510	0.9467	0.7809	0.7299	0.7221	0.6633	0.6419	0.6216	0.5289	0.5071	0.4810
540	0.9463	0.7805	0.7294	0.7215	0.6628	0.6412	0.6209	0.5283	0.5064	0.4803
570	0.9460	0.7801	0.7289	0.7210	0.6622	0.6406	0.6203	0.5277	0.5058	0.4797
600	0.9456	0.7798	0.7286	0.7205	0.6618	0.6401	0.6197	0.5272	0.5052	0.4791

Table 4.1: Prediction Mean Square Error from the calculation by sequentially added predictors over the basic 3 predictors

The "improvement" of prediction by adding the new $k - p = 9$ health and psychological predictors with sample size $n = 181$, as Baker TA (2008), can be measured by the "percentage of PMSE reduction":

$$\begin{aligned} pPM\hat{S}Er &= \left(\frac{PMSE_1 - PMSE}{PMSE_1}\right) \times 100\% \\ &= \left(1 - \frac{\sigma_k^2}{\sigma_p^2} \cdot \frac{n-p-2}{n-k-2}\right) \times 100\% = 47.40\%. \end{aligned}$$

The prediction mean square error measures the expected squared distance between the prediction for a specific value and what the true value is whereas $pPMSEr$ is used to demonstrate how much accuracy the added $k - p = 9$ predictors bring to the model. In this example, the introduction of psychological predictors into the model increased the model's prediction accuracy by 47.40%.

4.2.3 Efficient Sample Size

Generally speaking, PMSE will decrease as the sample size increases, so pPMSEr will increase as the sample size increases, that is, the larger the sample size, the better the prediction effect. However, the positive impact of sample size increase on prediction accuracy is limited. It shows that the prediction accuracy is stable when the sample size equals or exceeds a threshold. After the threshold, the increase in sample size is not cost-efficient to increase prediction accuracy.

The threshold as efficient sample size with specific efficiency $1 - \alpha = 0.9$(e.g., 90% of the largest pPMSEr at $n = \infty$).

$$n^* = p + 2 + (k - p)\left(\frac{EVR}{\alpha(1 - EVR)} + 1\right) = 103.6 \approx 104,$$

where $EVR = \frac{\sigma_k^2}{\sigma_p^2} = 0.499$. The actual used sample size in the paper is 181, which means the $rPMSEp$ should be greater than 0.1. On the flip side, with 181 sample size, the efficiency $1 - \alpha = 0.953$. The pPMSEr we obtained with sample size of 181 could reach 95.3% of the largest pPMSEr at $n = \infty$.

4.2.4 Cohen's f^2

The R^2 for full and reduced regression models are

$$R^2 = \frac{\sigma_{00} - \sigma_k^2}{\sigma_{00}} = \frac{\boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}}{\sigma_{00}} = 0.53126, R_1^2 = \frac{\sigma_{00} - \sigma_p^2}{\sigma_{00}} = \frac{\boldsymbol{\sigma_1'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_1}{\sigma_{00}} = 0.0606,$$

corresponding to R^2 given in Baker TA (2008) written below:

$$R^2 = 0.44, R_1^2 = 0.06.$$

It is noteworthy that the calculated R^2 deviates from the R^2 reported in the paper. This inconsistency can be attributed to the fact that the paper's full model was obtained by controlling for previous predictors, which may have amplified the effects of insignificant predictors. By the definition of the squared multiple correlations R^2 and Equation(3.16), Cohen's f^2 can be calculated $f_2^2 = 0.3328571$, that is, the new $k - p = 9$ predictors have large effect size since $f_2^2 \geq 0.15$ for Cohen's f_2^2 interpretation.

4.3 Simulation

The simulation aims to investigate the impact of sample size and the inclusion of non-basic predictors on prediction accuracy. Our hypothesis is that as more variables are added, the efficient sample size will decrease, resulting in a stable prediction accuracy that does not significantly improve with the addition of a larger sample size. To generate our data, we used the covariate matrix $\mathbf{\Sigma}$ from Equation(4.1). The simulation result using covariance generated response is also consistent with the result using the linear model with generated predictors data and error term with variance σ_k^2. The comparison between the two ways to generate the response variable is provided in the Appendix.

The basic model contained three demographic predictors, while the non-basic predictors were sequentially added to the model. For each iteration, a new response is generated to be considered as future value. We simulated the prediction mean square error (PMSE) and the correlation between the future value and prediction by taking the mean of 5000 iterations. The resulting PMSE and correlation values are presented in Table 4.2 and Table 4.3, respectively.

Table 4.2 presents the PMSE values for each sample size by adding non-basic

Sample Size	Basic Model	Comorbidities	Pain Locations	Medications	Physical Functioning	Depressive Symptoms	Life Satisfaction	LOC chance	LOC powerful	LOC internal
30	1.0947	0.9452	0.9356	0.9639	0.9406	0.9549	0.9846	0.8684	0.8744	0.8674
60	1.0487	0.8859	0.8599	0.8771	0.8370	0.8407	0.8531	0.7485	0.7460	0.7313
90	1.0319	0.8712	0.8343	0.8469	0.8017	0.7988	0.8000	0.6956	0.6879	0.6700
120	1.0115	0.8483	0.8102	0.8193	0.7716	0.7662	0.7625	0.6604	0.6504	0.6323
150	1.0016	0.8376	0.7978	0.8036	0.7544	0.7478	0.7410	0.6415	0.6301	0.6099
180	0.9947	0.8291	0.7867	0.7907	0.7402	0.7322	0.7237	0.6252	0.6135	0.5926
210	0.9894	0.8265	0.7828	0.7854	0.7337	0.7238	0.7137	0.6160	0.6024	0.5805
240	0.9840	0.8228	0.7782	0.7792	0.7272	0.7157	0.7044	0.6071	0.5919	0.5701
270	0.9840	0.8206	0.7753	0.7757	0.7236	0.7105	0.6988	0.6007	0.5844	0.5617
300	0.9807	0.8162	0.7694	0.7690	0.7164	0.7025	0.6904	0.5945	0.5776	0.5544
330	0.9762	0.8119	0.7644	0.7634	0.7099	0.6951	0.6830	0.5877	0.5707	0.5475
360	0.9731	0.8087	0.7606	0.7592	0.7057	0.6905	0.6777	0.5835	0.5659	0.5423
390	0.9695	0.8061	0.7576	0.7553	0.7021	0.6867	0.6737	0.5792	0.5618	0.5380
420	0.9673	0.8045	0.7558	0.7532	0.7002	0.6842	0.6702	0.5756	0.5583	0.5346
450	0.9657	0.8027	0.7533	0.7504	0.6971	0.6810	0.6667	0.5720	0.5541	0.5302
480	0.9649	0.8019	0.7522	0.7492	0.6959	0.6797	0.6650	0.5694	0.5514	0.5276
510	0.9659	0.8019	0.7522	0.7490	0.6954	0.6787	0.6633	0.5670	0.5486	0.5245
540	0.9648	0.8003	0.7510	0.7476	0.6935	0.6766	0.6611	0.5657	0.5473	0.5229
570	0.9640	0.7997	0.7503	0.7469	0.6920	0.6749	0.6589	0.5635	0.5449	0.5204
600	0.9619	0.7968	0.7469	0.7433	0.6886	0.6713	0.6550	0.5602	0.5417	0.5171

Table 4.2: Prediction Mean Square Error from simulation by sequentially added predictors over the basic 3 predictors

predictors sequentially. As the sample size increased, the PMSE decreased for all regression models. However, as some of the non-basic predictors were added, such as Medications, Depressive Symptoms, the PMSE increased, indicating that these predictors may not be improving the prediction accuracy. We also observed that adding more non-basic predictors led to a decreasing marginal gain in prediction accuracy with an increasing sample size.

The results presented in Table 4.2 provide empirical validation for the theoretical findings reported in Table 4.1. The simulation results demonstrate consistent patterns across each added predictor. The absolute value of difference for each parameter setting is observed to be less than 0.05, confirming that the estimation formulas for PMSE for full and reduced models are adequate. Specifically, the calculated $pPMSEr$ for a sample size of $n = 181$ is 47.40%. However, the simulated result yields an empirical $pP\hat{M}SEr$ of 40.42% for the same sample size, indicating notably lower prediction accuracy than the calculation expected.

$$
\begin{aligned}
pPMSEr &= \left(\frac{PMSE_1 - PMSE}{PMSE_1}\right) \times 100\% \\
&= \frac{0.9947 - 0.5926}{0.9947} = 40.42\%,
\end{aligned}
$$

This discrepancy may be attributed to the potential introduction of random error due to the inclusion of insignificant predictors, which can diminish the amount of variance explained and increase the effective sample size n^* needed. Since we considered the correlation in the definition of the error term in Equation(3.1), the phenomenon may potentially trace back to a low correlation between response and predictors. Thus, it may be necessary to increase the sample size to offset the negative impact of adding insignificant predictors on prediction accuracy. Since 7 out of the 12 predictors we are using are insignificant, only an efficiently large sample size, such as $n = 600$, can offset the impact with simulated $pPMSEr = 46.24\%$, which is closer to the theoretical result for $n = 600$, $pPMSEr = 49.33\%$.

Sample Size	Basic Model	Comorbidities	Pain Locations	Medications	Physical Functioning	Depressive Symptoms	Life Satisfaction	LOC chance	LOC powerful	LOC internal
30	0.1210	0.3252	0.3684	0.3584	0.4012	0.4009	0.4074	0.4729	0.4753	0.4752
60	0.1464	0.3657	0.4115	0.4052	0.4516	0.4561	0.4659	0.5370	0.5436	0.5498
90	0.1594	0.3868	0.4328	0.4288	0.4762	0.4841	0.4957	0.5673	0.5763	0.5858
120	0.1701	0.4007	0.4472	0.4450	0.4933	0.5026	0.5149	0.5876	0.5978	0.6096
150	0.1789	0.4108	0.4573	0.4563	0.5054	0.5161	0.5286	0.6019	0.6129	0.6257
180	0.1850	0.4180	0.4648	0.4646	0.5143	0.5259	0.5388	0.6124	0.6239	0.6377
210	0.1900	0.4233	0.4705	0.4712	0.5214	0.5337	0.5469	0.6204	0.6326	0.6469
240	0.1943	0.4279	0.4756	0.4769	0.5272	0.5400	0.5535	0.6268	0.6394	0.6542
270	0.1977	0.4316	0.4795	0.4814	0.5319	0.5452	0.5589	0.6322	0.6450	0.6600
300	0.2010	0.4348	0.4829	0.4853	0.5361	0.5497	0.5637	0.6368	0.6498	0.6652
330	0.2037	0.4376	0.4860	0.4887	0.5396	0.5536	0.5678	0.6406	0.6538	0.6695
360	0.2058	0.4398	0.4884	0.4915	0.5426	0.5568	0.5712	0.6438	0.6573	0.6731
390	0.2079	0.4418	0.4906	0.4940	0.5451	0.5597	0.5741	0.6467	0.6603	0.6763
420	0.2096	0.4436	0.4926	0.4961	0.5474	0.5622	0.5768	0.6492	0.6630	0.6791
450	0.2115	0.4452	0.4943	0.4981	0.5494	0.5644	0.5791	0.6515	0.6654	0.6817
480	0.2131	0.4468	0.4959	0.4999	0.5513	0.5665	0.5813	0.6536	0.6677	0.6840
510	0.2144	0.4481	0.4973	0.5015	0.5529	0.5683	0.5832	0.6554	0.6695	0.6860
540	0.2157	0.4491	0.4985	0.5029	0.5544	0.5698	0.5849	0.6570	0.6712	0.6878
570	0.2168	0.4502	0.4996	0.5042	0.5557	0.5713	0.5865	0.6585	0.6728	0.6895
600	0.2178	0.4511	0.5007	0.5054	0.5570	0.5727	0.5879	0.6599	0.6742	0.6910

Table 4.3: Correlation between the true value and prediction changes by sequentially added predictors over the basic 3 predictors

Basic Predictors	Comorbidities	Pain Locations	Medications	Physical Functioning	Depressive Symptoms	Life Satisfaction	LOC -Chance	LOC -Powerful
103.6487	137.1353	143.9351	129.5519	141.2487	129.9053	113.9951	206.2313	191.7463

Table 4.4: Efficient sample size n^* by sequentially added predictors over the basic 3 predictors

Table 4.3 presents the correlation values for each sample size by adding predictors sequentially. The result is simulated using 5000 iterations using 5-fold cross-validation. We observed a similar pattern as with $PMSE$: the correlation increased with increasing sample size but decreased as some insignificant non-basic predictors were added. Additionally, we observed a similar diminishing marginal gain in prediction accuracy with the inclusion of more non-basic predictors.

The calculation of the efficient sample size for each model with a significance level of $\alpha = 0.1$ (i.e., the sample size that attains 90% of the largest pPMSEr as n approaches infinity) is provided as a reference in Tables 4.2 and 4.3, as demonstrated in Table 4.4.

Upon attaining the efficient sample size n^*, for each model, it is anticipated that the $rPMSEp$, $PMSE$, and correlations will remain stable, as the sample size continues to increase. This discovery lends support to the definition used to determine efficient sample size n^*. The low correlation between variables and response variables may result in a reduced possibility of obtaining statistically significant results. Furthermore, the presence of non-significant predictors can have an impact on the estimation of the efficient sample size n^*, leading to an underestimation of the efficient sample size necessary to achieve a prediction accuracy level. The incorporation of insignificant predictors, such as Medications, Depressive Symptoms, Life Satisfaction, and LOC-internal, into the model, is expected to result in an increase in the efficient sample size n^*, as opposed to a decrease for significant predictors. The inclusion of such predictors is unlikely to enhance the predictive performance, it is probable to weaken it instead. Moreover, it is unlikely to provide additional information but rather introduce random errors. Consequently, a larger sample size would be required to achieve the same level of prediction accuracy. Conversely, the significant predictors can be readily identified by an increase in $pPMSEr$, a decrease in $PMSE$, or a less efficient sample size n^*.

To summarize, our simulation study demonstrates that adding significant non-basic predictors can decrease the effective sample size and improve prediction accuracy when the sample size is enough. However, including additional insignificant predictors may lead to extra randomness and reduced prediction accuracy instead of contributing to variance explanation. Therefore, careful consideration of the choice of predictors to include in the model is necessary, taking into account both the sample size and the objectives of the analysis. The calculated formulas are proven to be adequate to provide estimations for $PMSE$ when introducing predictors in the linear model.

Chapter 5

Discussion

The present study utilized the analytical result to examine the relationship between $pPMSEr$, sample size n, and $Cohen'sf^2$. Our findings indicate that there is a significant relationship between these variables, with $pPMSEr$ decreasing as sample size and effect size increase. These results are consistent with previous research on statistical power and highlight the importance of adequate sample sizes in hypothesis testing. The analysis also revealed that the non-basic predictors have the ability to significantly contribute to explaining the variability in response. The variance of the response variable, pain intensity, was assumed to be 1 in the simulation as the error variance for the full model is 0.9393, which implies that the model was able to capture a substantial amount of the variance in pain intensity. The effects of each predictor were also examined using the calculated coefficients. The results indicated that Physical functioning, Depressive symptoms, and LOC-internal were the most significant predictors of pain intensity. These findings are consistent with previous research, which has also identified these factors as important contributors to pain intensity.

It is worth noting that our results of predictor effects differ from those reported by Baker TA (2008), who utilized the same correlation matrix but different coefficients. While this discrepancy may be due to methodological differences or variations in sample characteristics, it underscores the importance of replication studies in scientific research.

One limitation of our study is that it only considered a single quantitative outcome variable (pain intensity) and a limited set of predictors, without considering the interactions or random effects. Future research could extend our approach to other kinds of models such as general linear models, additional outcome measures, and a more comprehensive set of predictors, potentially utilizing machine learning or other statistical techniques to identify complex relationships among predictors. Other possible limitation includes the use of a correlation matrix rather than real-world data and the assumption of normality in the distribution of the response variable and covariates. These limitations may affect the accuracy of the results and should be taken into consideration when interpreting the findings.

Future research may benefit from using individual-level data and exploring the relationship between response and other factors obtained in newly published studies. Moreover, additional studies may also investigate the effectiveness of interventions aimed at reducing pain intensity by addressing the significant predictors identified in this study.

Bibliography

Baker TA, C. N., Buchanan NT (2008). Factors influencing chronic pain intensity in older black women: examining depression, locus of control, and physical health. *Womens Health (Larchmt)*.

Kerridge, D. (1967). Errors of prediction in multiple regression with stochastic regressor variables. *Technometrics*, **9** 309–311.

Narula, S. C. (1974). Predictive mean square error and stochastic regressor variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **23** 11–17.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G. and Collins, G. S. (2019a). Minimum sample size for developing a multivariable prediction model: Part i–continuous outcomes. *Statistics in medicine*, **38** 1262–1275.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G. and Collins, G. S. (2019b). Minimum sample size for developing a multivariable prediction model: Part ii-binary and time-to-event outcomes. *Statistics in medicine*, **38** 1276–1296.

Sawyer, R. (1982). Sample size and the accuracy of predictions made from multiple regression equations. *Journal of Educational Statistics*, **7** 91–104.

van der Ploeg, T., Austin, P. C. and Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, **14** 1–13.

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J. and Reitsma, J. B. (2019). Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical methods in medical research*, **28** 2455–2474.

Chapter 6

Appendix

6.1 Proof for Subset PMSE

The purpose of section 6.1 is to provide comprehensive proof of the work of Narula (1974). While Narula (1974) provides a valuable theoretical framework for PMSE, the derivation of the formula is not fully elucidated. As such, this appendix aims to fill in the gaps and provide a clear and rigorous proof of the PMSE formula for stochastic regressor variables. Through this analysis, we hope to deepen our understanding of the PMSE and its relevance in statistical modeling.

The response variable and the predictor variables follow a joint $(k+1)$-variate normal distribution with unknown mean vector $\mu^* = [\mu_0, \boldsymbol{\mu}']'$, and covariance matrix

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} \sigma_{00} & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{bmatrix}.$$

Let $\boldsymbol{z_1, z_2, ..., z_n}$ be n independent (k-component vector) observations on the predictor variables, $\boldsymbol{x_i} = \boldsymbol{z_i} - \bar{\boldsymbol{z}}$. Let $\boldsymbol{S}^* = \begin{bmatrix} s_{00} & \boldsymbol{s}' \\ \boldsymbol{s} & \boldsymbol{S} \end{bmatrix}$ be sample covariance matrix, where

$$s_{00} = \sum \frac{(y_i - \bar{y})^2}{n-1}, \boldsymbol{s} = \sum \frac{(y_i - \bar{y})\boldsymbol{x_i}}{n-1}, \boldsymbol{S} = \sum \frac{\boldsymbol{x_i}\boldsymbol{x_i}'}{n-1}.$$

We assume the correct model as follows,

$$\boldsymbol{y = \alpha + \beta_1 z_1 + \beta_2 z_2 + ... + \beta_k z_k + \epsilon},$$

Meanwhile, the LSE prediction equation

$$\boldsymbol{\hat{y}} = \bar{y} + \boldsymbol{\hat{\beta}_1}(\boldsymbol{z_1} - \bar{z_1}) + \boldsymbol{\hat{\beta}_2}(\boldsymbol{z_2} - \bar{z_2}) + ... + \boldsymbol{\hat{\beta}_k}(\boldsymbol{z_k} - \bar{z_k}) = \bar{y} + \boldsymbol{X'\hat{\beta}},$$

$$\hat{y_i} = \bar{y} + \boldsymbol{x_i'\hat{\beta}}.$$

For any given $\boldsymbol{z_i}$,

$$\begin{aligned} E(y_i|\boldsymbol{z_i}) &= \alpha + \boldsymbol{\beta z_i} \\ &= \mu_0 - \boldsymbol{\sigma' \Sigma^{-1} \mu} + \boldsymbol{\sigma' \Sigma^{-1} z_i} \\ &= \mu_0 + \boldsymbol{\sigma' \Sigma^{-1} (z_i - \mu)}, \end{aligned}$$

where $\alpha = \mu_0 - \boldsymbol{\sigma' \Sigma^{-1} \mu}, \boldsymbol{\beta = \sigma' \Sigma^{-1}}$. Thus $\hat{\alpha} = \bar{y} - \boldsymbol{s' S^{-1} \bar{x}}, \boldsymbol{\hat{\beta}} = \boldsymbol{S^{-1} s}$.

The conditional predictive mean square error by

$$\begin{aligned}
E[(y_0 - \hat{y_0})^2|\boldsymbol{z_0}] &= E[(\alpha + \boldsymbol{(z_0 - \mu)'\beta} + \epsilon_0 - \bar{y} - \boldsymbol{x_0'}\hat{\boldsymbol{\beta}}|\boldsymbol{z_0})^2] \\
&= E[(\alpha + \boldsymbol{(z_0 - \mu)'\beta + \epsilon_0} - \alpha - \boldsymbol{(\bar{z} - \mu)'\beta - \bar{\epsilon}} - \boldsymbol{x_0'}\beta|\boldsymbol{z_0})^2] \\
&= E[(\boldsymbol{x_0'\beta} - \boldsymbol{\bar{x}'\beta} + \epsilon_0 - \bar{\epsilon} - \boldsymbol{x_0'}\hat{\boldsymbol{\beta}}|\boldsymbol{z_0})^2] \\
&= E[(\boldsymbol{x_0'\beta} + (\epsilon_0 - \bar{\epsilon}) - \boldsymbol{x_0'}\hat{\boldsymbol{\beta}}|\boldsymbol{z_0})^2] \\
&= E[(\boldsymbol{x_0'\beta} + (\epsilon_0 - \bar{\epsilon}))^2 + (\boldsymbol{x_0'}\hat{\boldsymbol{\beta}})^2 - 2(\boldsymbol{x_0'\beta} + (\epsilon_0 - \bar{\epsilon}))\boldsymbol{x_0'}\hat{\boldsymbol{\beta}}|\boldsymbol{z_0}] \\
&= E[(\boldsymbol{x_0'\beta})^2 + (\epsilon_0 - \bar{\epsilon})^2 + (\boldsymbol{x_0'}\hat{\boldsymbol{\beta}})^2 - 2\boldsymbol{x_0'\beta x_0'}\hat{\boldsymbol{\beta}}|\boldsymbol{z_0}] \\
&= \boldsymbol{\beta'}E(\boldsymbol{x_0x_0'}|\boldsymbol{z_0})\boldsymbol{\beta} + E[(\epsilon_0 - \bar{\epsilon})^2|\boldsymbol{z_0}] + E[(\boldsymbol{x_0'}\hat{\boldsymbol{\beta}})^2|\boldsymbol{z_0}] - 2\boldsymbol{\beta'}E[\boldsymbol{x_0x_0'}\hat{\boldsymbol{\beta}}|\boldsymbol{z_0}].
\end{aligned}$$

By following equations, corresponding Lemma A1, Lemma A3, Lemma A3.1, Lemma A7, Lemma A9 proven by Narula (1974),

$$E(\tilde{\boldsymbol{\beta}}_{\boldsymbol{1}}|\boldsymbol{X_1}) = \boldsymbol{\beta_1} + \boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\beta_2} = \boldsymbol{\Phi_1}, \tag{6.1}$$

$$E(\boldsymbol{x_{01}x_{01}'}|\boldsymbol{z_0}) = (\boldsymbol{z_{01}} - \boldsymbol{\mu_1})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})' + \frac{\boldsymbol{\Sigma_{11}}}{n}, \tag{6.2}$$

$$E(\boldsymbol{x_0x_0'}|\boldsymbol{z_0}) = \boldsymbol{(z_0 - \mu)}(\boldsymbol{z_0} - \boldsymbol{\mu})' + \frac{\boldsymbol{\Sigma}}{n}, \tag{6.3}$$

$$\begin{aligned}
E[(\boldsymbol{x_{01}'}\tilde{\boldsymbol{\beta}}_{\boldsymbol{1}})^2|\boldsymbol{z_0}] =& \sigma_p^2 \left[\boldsymbol{(z_{01} - \mu_1)'\Sigma_{11}^{-1}(z_{01} - \mu_1)} + \frac{p}{n}\right] \frac{1}{n-p-2} \\
&+ \boldsymbol{\Phi_1'\Sigma_{11}\Phi_1}\frac{1}{n} + \boldsymbol{\Phi_1'(z_{01} - \mu_1)(z_{01} - \mu_1)'\Phi_1},
\end{aligned} \tag{6.4}$$

$$\boldsymbol{\sigma'}\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\boldsymbol{\Sigma_{11}} \\ \boldsymbol{\Sigma_{21}}\end{bmatrix} = \sigma_1', \tag{6.5}$$

The conditional PMSE can be written as

$$\begin{aligned}
E[(y_0-\hat{y_0})^2|\boldsymbol{z_0}] &= \boldsymbol{\beta'}E(\boldsymbol{x_0x_0'}|\boldsymbol{z_0})\boldsymbol{\beta} + E[(\epsilon_0-\bar{\epsilon})^2|\boldsymbol{z_0}] + E[(\boldsymbol{x_0'}\hat{\boldsymbol{\beta}})^2|\boldsymbol{z_0}] - 2\boldsymbol{\beta'}E[\boldsymbol{x_0x_0'}\hat{\boldsymbol{\beta}}|\boldsymbol{z_0}] \\
&= \boldsymbol{\beta'}\left[\boldsymbol{(z_0-\mu)(z_0-\mu)'}+\frac{\boldsymbol{\Sigma}}{n}\right]\boldsymbol{\beta}+\sigma_k^2\left(1+\frac{1}{n}\right) \\
&\quad +\sigma_k^2\left[\boldsymbol{(z_0-\mu)'\Sigma^{-1}(z_0-\mu)}+\frac{k}{n}\right]\frac{1}{n-k-2} \\
&\quad +\frac{1}{n}\boldsymbol{\Phi'\Sigma\Phi}+\boldsymbol{\Phi'(z_0-\mu)(z_0-\mu)'\Phi}-2\boldsymbol{\beta'}\left[\boldsymbol{(z_0-\mu)(z_0-\mu)'}+\frac{\boldsymbol{\Sigma}}{n}\right]\boldsymbol{\beta} \\
&= \sigma_k^2\left(1+\frac{1}{n}\right)+\sigma_k^2\left[\boldsymbol{(z_0-\mu)'\Sigma^{-1}(z_0-\mu)}+\frac{k}{n}\right]\frac{1}{n-k-2}.
\end{aligned}$$

Since $\boldsymbol{\Phi}$ is the notation of expectation of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{1}}$, when we are using all predictors, the LSE is unbiased, which means $\boldsymbol{\Phi}=\boldsymbol{\beta}$.

The unconditional PMSE

$$\begin{aligned}
E[(y_0-\hat{y_0})^2] &= E\{E[(y_0-\hat{y_0})^2|\boldsymbol{z_0}]\} \\
&= E\left[\sigma_k^2\left(1+\frac{1}{n}\right)+\sigma_k^2\left[\boldsymbol{(z_0-\mu)'\Sigma^{-1}(z_0-\mu)}+\frac{k}{n}\right]\frac{1}{n-k-2}\right] \\
&= \sigma_k^2\left(1+\frac{1}{n}\right)+\sigma_k^2E\left[\boldsymbol{(z_0-\mu)'\Sigma^{-1}(z_0-\mu)}+\frac{k}{n}\right]\frac{1}{n-k-2} \\
&= \sigma_k^2\left(1+\frac{1}{n}\right)+\sigma_k^2\left(k+\frac{k}{n}\right)\frac{1}{n-k-2} \\
&= \sigma_k^2\left(1+\frac{1}{n}\right)\left(1+\frac{1}{n-k-2}\right) \\
&= \sigma_k^2\left(1+\frac{1}{n}\right)\left(\frac{n-2}{n-k-2}\right).
\end{aligned}$$

For subset, we partition the k-component vector of predictor variables into two parts, $\boldsymbol{Z}=[\boldsymbol{Z_1},\boldsymbol{Z_2}], \boldsymbol{X}=[\boldsymbol{X_1},\boldsymbol{X_2}], \boldsymbol{x_1'}=[\boldsymbol{x_{i1}'},\boldsymbol{x_{i2}'}], \boldsymbol{\mu'}=[\boldsymbol{\mu_1'},\boldsymbol{\mu_2'}], \boldsymbol{\sigma'}=[\boldsymbol{\sigma_1'},\boldsymbol{\sigma_2'}], \boldsymbol{s'}=[\boldsymbol{s_1'},\boldsymbol{s_2'}]$, $\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}}\\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}}\end{bmatrix}, \boldsymbol{S}=\begin{bmatrix}\boldsymbol{S_{11}} & \boldsymbol{S_{12}}\\ \boldsymbol{S_{21}} & \boldsymbol{S_{22}}\end{bmatrix}$, so the subset prediction equation is given by

$$\tilde{y_i} = \bar{y} + \boldsymbol{x'_{i1}}\boldsymbol{\tilde{\beta_1}},$$

where $\boldsymbol{\tilde{\beta_1} = S_{11}^{-1}s_1}$. By LemmaA1, LemmaA3, LemmaA7, $\boldsymbol{\Phi_1 = \beta_1 + \Sigma_{11}^{-1}\Sigma_{12}\beta_2}$, the conditional PMSE at z_0 is given by

$$\begin{aligned}
E[(y_0 - \tilde{y_0})^2|\boldsymbol{z_0}] &= E[(\boldsymbol{x'_0\beta} + \epsilon_0 - \bar{\epsilon} - \boldsymbol{x'_{01}\tilde{\beta_1}}|\boldsymbol{z_0})^2] \\
&= E[(\boldsymbol{x'_0\beta})^2 + (\boldsymbol{x'_{01}\tilde{\beta_1}})^2 - 2\boldsymbol{x'_0\beta x'_{01}\tilde{\beta_1}} + (\epsilon_0 - \bar{\epsilon})^2 + 2(\boldsymbol{x'_0\beta - x'_{01}\tilde{\beta_1}})(\epsilon_0 - \bar{\epsilon})|\boldsymbol{z_0}] \\
&= E[(\boldsymbol{x'_0\beta})^2|\boldsymbol{z_0}] + E[\boldsymbol{x'_{01}\tilde{\beta_1}})^2|\boldsymbol{z_0}] - 2E[\boldsymbol{x'_0\beta x'_{01}\tilde{\beta_1}}|\boldsymbol{z_0}] + E[(\epsilon_0 - \bar{\epsilon})^2|\boldsymbol{z_0}] \\
&\quad + 2E[(\boldsymbol{x'_0\beta - x'_{01}\tilde{\beta_1}})(\epsilon_0 - \epsilon)|\boldsymbol{z_0}],
\end{aligned}$$

where

$$E[(\boldsymbol{x'_0\beta})^2|\boldsymbol{z_0}] = \boldsymbol{\beta}\left[\boldsymbol{(z_0 - \mu)(z_0 - \mu)'} + \frac{\boldsymbol{\Sigma}}{n}\right]\boldsymbol{\beta},$$

$$\begin{aligned}
E[(\boldsymbol{x'_{01}\tilde{\beta_1}})^2|\boldsymbol{z_0}] =& \sigma_p^2\left[\boldsymbol{(z_{01} - \mu_1)'\Sigma_{11}^{-1}(z_{01} - \mu_1)} + \frac{p}{n}\right]\frac{1}{n-p-2} \\
&+ \boldsymbol{\Phi'_1\Sigma_{11}\Phi_1}\frac{1}{n} + \boldsymbol{\Phi'_1(z_{01} - \mu_1)(z_{01} - \mu_1)'\Phi_1},
\end{aligned}$$

$$E[(\epsilon_0 - \bar{\epsilon})^2|\boldsymbol{z_0}] = \sigma_k^2 + \frac{1}{n}\sigma_k^2,$$

$$E[(\boldsymbol{x'_0\beta - x'_{01}\tilde{\beta_1}})(\epsilon_0 - \epsilon)|\boldsymbol{z_0}] = E[(\boldsymbol{x'_0\beta - x'_{01}\tilde{\beta_1}})\epsilon_0|\boldsymbol{z_0}] - E[(\boldsymbol{x'_0\beta - x'_{01}\tilde{\beta_1}})\bar{\epsilon}|\boldsymbol{z_0}] = 0.$$

(but $\bar{\epsilon} = 0$ is not necessary)

Therefore, the conditional PMSE could be written as

$$\begin{aligned}
&= \sigma_p^2 \left[(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Sigma_{11}^{-1}}(\boldsymbol{z_{01}} - \boldsymbol{\mu_1}) + \frac{p}{n} \right] \frac{1}{n-p-2} + \boldsymbol{\Phi_1'}\boldsymbol{\Sigma_{11}}\boldsymbol{\Phi_1}\frac{1}{n} + \boldsymbol{\Phi_1'}(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Phi_1} \\
&\quad + \boldsymbol{\beta}(\boldsymbol{z_0} - \boldsymbol{\mu})(\boldsymbol{z_0} - \boldsymbol{\mu})'\boldsymbol{\beta} + \boldsymbol{\beta'\Sigma\beta}\frac{1}{n} - 2\boldsymbol{\beta'}E(\boldsymbol{x_0x_{01}'}|\boldsymbol{z_0})\boldsymbol{\Phi_1} + \sigma_k^2 + \frac{1}{n}\sigma_k^2 \\
&= \sigma_k^2 + \frac{1}{n}(\sigma_p^2 + \boldsymbol{\sigma_1'}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma_1} - \boldsymbol{\sigma'}\boldsymbol{\Sigma^{-1}}\boldsymbol{\sigma}) + [(\boldsymbol{z_0} - \boldsymbol{\mu})'\boldsymbol{\beta}]^2 + [(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Phi_1}]^2 \\
&\quad + \sigma_p^2 \left[(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Sigma_{11}^{-1}}(\boldsymbol{z_{01}} - \boldsymbol{\mu_1}) + \frac{p}{n} \right] \frac{1}{n-p-2} \\
&\quad + \boldsymbol{\beta'\Sigma\beta}\frac{1}{n} + \boldsymbol{\Phi_1'}\boldsymbol{\Sigma_{11}}\boldsymbol{\Phi_1}\frac{1}{n} - 2\boldsymbol{\beta'}E(\boldsymbol{x_0x_{01}'}|\boldsymbol{z_0})\boldsymbol{\Phi_1} \\
&= \sigma_k^2 + \frac{1}{n}\sigma_p^2 + \sigma_p^2 \left[(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Sigma_{11}^{-1}}(\boldsymbol{z_{01}} - \boldsymbol{\mu_1}) + \frac{p}{n} \right] \frac{1}{n-p-2} + [(\boldsymbol{z_0} - \boldsymbol{\mu})'\boldsymbol{\beta}]^2 + [(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Phi_1}]^2 \\
&\quad + \frac{1}{n}\boldsymbol{\beta'\Sigma\beta} + \frac{1}{n}\boldsymbol{\Phi_1'}\boldsymbol{\Sigma_{11}}\boldsymbol{\Phi_1} + \frac{1}{n}\boldsymbol{\sigma_1'}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma_1} - \frac{1}{n}\boldsymbol{\sigma'}\boldsymbol{\Sigma^{-1}}\boldsymbol{\sigma} - 2\boldsymbol{\beta'}E(\boldsymbol{x_0x_{01}'}|\boldsymbol{z_0})\boldsymbol{\Phi_1},
\end{aligned}$$

where

$$E(\boldsymbol{x_0x_{01}'}|\boldsymbol{z_0}) = E\left[\begin{bmatrix} \boldsymbol{x_{01}x_{01}'} \\ \boldsymbol{x_{02}x_{01}'} \end{bmatrix} |\boldsymbol{z_0} \right] = \begin{bmatrix} (\boldsymbol{z_{01}} - \boldsymbol{\mu_1})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})' + \boldsymbol{\Sigma_{11}/n} \\ (\boldsymbol{z_{02}} - \boldsymbol{\mu_2})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})' + \boldsymbol{\Sigma_{21}/n} \end{bmatrix},$$

$$\boldsymbol{\sigma'}\boldsymbol{\Sigma^{-1}}\boldsymbol{\sigma} = \boldsymbol{\beta'\Sigma\beta}, \boldsymbol{\sigma_1'}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma_1} = \boldsymbol{\beta_1'}\boldsymbol{\Sigma_{11}}\boldsymbol{\beta_1},$$

so that

$$\begin{aligned}
\boldsymbol{\beta'}E(\boldsymbol{x_0x_{01}'}|\boldsymbol{z_0})\boldsymbol{\Phi_1} =& [\boldsymbol{\beta_1'}(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})' + \boldsymbol{\beta_1'}\boldsymbol{\Sigma_{11}}/n \\
&+ \boldsymbol{\beta_2'}(\boldsymbol{z_{02}} - \boldsymbol{\mu_2})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})' + \boldsymbol{\beta_2'}\boldsymbol{\Sigma_{21}}/n]\boldsymbol{\Phi_1} \\
=& \boldsymbol{\beta_1'}(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Phi_1} + \boldsymbol{\beta_2'}(\boldsymbol{z_{02}} - \boldsymbol{\mu_2})(\boldsymbol{z_{01}} - \boldsymbol{\mu_1})'\boldsymbol{\Phi_1} + \boldsymbol{\sigma_1'}\boldsymbol{\Phi_1}/n.
\end{aligned}$$

The other terms in condition PMSE would be equal to

$$\begin{aligned}
&\frac{1}{n}\boldsymbol{\beta'\Sigma\beta}+\frac{1}{n}\boldsymbol{\Phi_1'\Sigma_{11}\Phi_1}+\frac{1}{n}\boldsymbol{\sigma_1'\Sigma_{11}^{-1}\sigma_1}-\frac{1}{n}\boldsymbol{\sigma'\Sigma^{-1}\sigma}-\frac{2}{n}\boldsymbol{\sigma_1'\Phi_1}\\
&=\frac{1}{n}\boldsymbol{\Phi_1'\Sigma_{11}\Phi_1}+\frac{1}{n}\boldsymbol{\sigma_1'\Sigma_{11}^{-1}\sigma_1}-\frac{2}{n}\boldsymbol{\sigma_1'\Phi_1}\\
&=\frac{1}{n}\boldsymbol{\sigma_1'\Sigma_{11}^{-1}\sigma_1}+\frac{1}{n}\boldsymbol{\sigma_1'\Sigma_{11}^{-1}\sigma_1}-\frac{2}{n}\boldsymbol{\sigma_1'\Sigma_{11}^{-1}\sigma_1}=0,
\end{aligned}$$

in which $\Phi_1=\boldsymbol{\Sigma_{11}^{-1}\sigma_1}$.

The conditional PMSE is equal to

$$\begin{aligned}
E[(y_0-\tilde{y_0})^2|\boldsymbol{z_0}]&=\sigma_k^2+\frac{\sigma_p^2}{n}+\sigma_p^2\left[\boldsymbol{(z_{01}-\mu_1)'\Sigma_{11}^{-1}(z_{01}-\mu_1)'}+\frac{p}{n}\right]\frac{1}{n-p-2}\\
&+[\boldsymbol{(z_0-\mu)'\beta-(z_{01}-\mu_1)'\Phi_1}]^2\\
&=\sigma_k^2+\frac{\sigma_p^2}{n}+\sigma_p^2\left[\boldsymbol{(z_{01}-\mu_1)'\Sigma_{11}^{-1}(z_{01}-\mu_1)'}+\frac{p}{n}\right]\frac{1}{n-p-2}\\
&+[\boldsymbol{(z_{01}-\mu_1)'\beta_1+(z_{02}-\mu_2)'-(z_{01}-\mu_1)'(\beta_1+\Sigma_{11}^{-1}\Sigma_{12}\beta_2)}]^2\\
&=\sigma_k^2+\frac{\sigma_p^2}{n}+\sigma_p^2\left[\boldsymbol{(z_{01}-\mu_1)'\Sigma_{11}^{-1}(z_{01}-\mu_1)'}+\frac{p}{n}\right]\frac{1}{n-p-2}\\
&+[\boldsymbol{(z_{02}-\mu_2)'-(z_{01}-\mu_1)'\Sigma_{11}^{-1}\Sigma_{12}\beta_2}]^2,
\end{aligned}$$

take expectation

$$\begin{aligned}
E[(y_0-\tilde{y_0})^2]&=E[E(y_0-\tilde{y_0})^2|\boldsymbol{z_0}]\\
&=E\{\sigma_k^2+\frac{\sigma_p^2}{n}+\sigma_p^2\left[\boldsymbol{(z_{01}-\mu_1)'\Sigma_{11}^{-1}(z_{01}-\mu_1)'}+\frac{p}{n}\right]\frac{1}{n-p-2}\\
&+[\boldsymbol{(z_0-\mu)'\beta-(z_{01}-\mu_1)'\Phi_1}]^2\} \qquad .\\
&=\sigma_p^2+\frac{\sigma_p^2}{n}+\boldsymbol{\sigma_1'\Sigma_{11}^{-1}\sigma_1-\sigma'\Sigma^{-1}\sigma}+\sigma_p^2\left(p+\frac{p}{n}\right)\frac{1}{n-p-2}\\
&+E[\boldsymbol{\beta'(z_0-\mu)(z_0-\mu)'\beta+\Phi_1'(z_{01}-\mu_1)(z_{01}-\mu_1)'\Phi_1}\\
&-2\boldsymbol{\beta'(z_0-\mu)(z_{01}-\mu_1)'\Phi_1}]
\end{aligned}$$

The expectation term $E[\boldsymbol{\beta'(z_0 - \mu)(z_0 - \mu)'\beta}$ is equal to

$$\boldsymbol{\beta'\Sigma\beta + \Phi_1'\Sigma_{11}\Phi_1 - 2\beta'} \begin{bmatrix} \boldsymbol{\Sigma_{11}} \\ \boldsymbol{\Sigma_{21}} \end{bmatrix} \boldsymbol{\Phi_1}.$$

The unconditional PMSE

$$\begin{aligned} E[(y_0 - \tilde{y_0})^2] &= \sigma_p^2(1+\frac{1}{n})(n-2)/(n-p-2) \\ &\quad + \boldsymbol{\sigma_1'\Sigma_{11}^{-1}\sigma_1 - \sigma'\Sigma^{-1}\sigma + \beta'\Sigma\beta + \Phi_1'\Sigma_{11}\Phi_1 - 2\beta'} \begin{bmatrix} \boldsymbol{\Sigma_{11}} \\ \boldsymbol{\Sigma_{21}} \end{bmatrix} \boldsymbol{\Phi_1} \\ &= \sigma_p^2(1+\frac{1}{n})(n-2)/(n-p-2) \\ &\quad + \boldsymbol{\Phi_1'\Sigma_{11}\Phi_1 - \beta'\Sigma\beta + \beta'\Sigma\beta + \Phi_1'\Sigma_{11}\Phi_1 - 2\beta'} \begin{bmatrix} \boldsymbol{\Sigma_{11}} \\ \boldsymbol{\Sigma_{21}} \end{bmatrix} \boldsymbol{\Phi_1} \\ &= \sigma_p^2(1+\frac{1}{n})(n-2)/(n-p-2). \end{aligned}$$

Thus, the unconditional PMSE $= \sigma_p^2(1+\frac{1}{n})(n-2)/(n-p-2)$

6.2 Distribution Approximation of PMSE

6.2.1 Analytical Result

To investigate the unconditional prediction square error distribution of $(\hat{y} - y^*)^2$, we start with moments of $(\hat{y} - y^*)$, proven by Sawyer (1982): Let M be a positive integer. Then

$$\mathbb{E}\left[(\hat{y} - y^*)^{2M}\right] = \frac{\sigma^{2M}\frac{(2M)!}{M!}\left(\frac{n+1}{2n}\right)^M \prod_{j=1}^{M}(n-2j)}{\prod_{j=1}^{M}(n-p-2j)},$$

when $2M \leq n-p-1$.

If we consider the distribution $\hat{y} - y^*$ follows an asymptotically normal distribution, then $(\hat{y} - y^*)^2$ is likely to be approximate Gamma distribution, by the proposition above,

$$\mathbb{E}\left[(\hat{y} - y^*)^2\right] = \sigma^2 \frac{(n+1)(n-2)}{n(n-p-2)}, \tag{6.6}$$

$$\mathbb{E}\left[(\hat{y} - y^*)^4\right] = 3\sigma^4 \frac{(n+1)^2(n-2)(n-4)}{n^2(n-p-2)(n-p-4)}, \tag{6.7}$$

$$\mathrm{Var}\left[(\hat{y} - y^*)^2\right] = \sigma^4 \frac{(n+1)^2(n-2)}{n^2(n-p-2)} \left(\frac{3n-12}{n-p-4} - \frac{n-2}{n-p-2} \right). \tag{6.8}$$

Applying method of moments, for $Gamma(\alpha, \beta)$ with shape-scale parameters

$$\mathbb{E}(X^n) = \frac{\beta^n (n+\alpha-1)!}{(\alpha-1)!},$$

$$\mathbb{E}X = \alpha\beta, \mathbb{E}X^2 = \alpha(\alpha+1)\beta^2,$$

let

$$\begin{cases} \mathbb{E}\left[(\hat{y} - y^*)^2\right] = \alpha\beta \\ \mathbb{E}\left[(\hat{y} - y^*)^2\right] = \alpha(\alpha+1)\beta^2 \end{cases} \Rightarrow \begin{cases} \alpha = \frac{(n-2)(n-p-4)}{3(n-4)(n-p-2)-(n-2)(n-p-4)} \\ \beta = \sigma^2 \frac{n+1}{n} \left(\frac{3(n-4)}{n-p-4} - \frac{n-2}{n-p-2} \right) \end{cases}$$

Based on the normal assumption, we approximate the prediction square error as

$$Gamma \left(\frac{(n-2)(n-p-4)}{3(n-4)(n-p-2) - (n-2)(n-p-4)}, \sigma^2 \frac{n+1}{n} \left(\frac{3(n-4)}{n-p-4} - \frac{n-2}{n-p-2} \right) \right) \tag{6.9}$$

For $n - p \geq 5$, the approximation using Gram-Charlier is

$$P(\hat{y} - y^* \leq t) \doteq \Phi\left(\frac{t}{\sigma'}\right) + \frac{p}{4(n-2)(n-p-4)} \Phi^{(4)}\left(\frac{t}{\sigma'}\right), \tag{6.10}$$

where Φ is the standard normal distribution function, $\Phi^{(4)}$ is its fourth derivative,

$$\sigma' = \sqrt{MSE} = \sigma\sqrt{\frac{(n+1)(n-2)}{n(n-p-2)}}.$$

Though adding the second term may not even improve the approximation by Sawyer (1982) Appendix, we examine the approximation for $P((\hat{y}-y^*)^2 \leq t)$ distribution.

Denote $X = \hat{y}-y^*, Y = X^2 = (\hat{y}-y^*)^2$,

$$\begin{aligned}
F_Y(t) &= P(Y \leq t) = P(|\hat{y}-y^*| \leq \sqrt{t}) \\
&= P(\hat{y}-y^* \leq \sqrt{t}) - P(\hat{y}-y^* \leq -\sqrt{t}) \\
&= \Phi(\frac{\sqrt{t}}{\sigma'}) - \Phi(-\frac{\sqrt{t}}{\sigma'}) = 2\Phi(\frac{\sqrt{t}}{\sigma'}) - 1 \quad (first \quad term), \\
or & \\
&= \Phi(\frac{\sqrt{t}}{\sigma'}) + \frac{p}{4(n-2)(n-p-4)}\Phi^{(4)}(\frac{\sqrt{t}}{\sigma'}) \\
&\quad - \Phi(-\frac{\sqrt{t}}{\sigma'}) - \frac{p}{4(n-2)(n-p-4)}\Phi^{(4)}(-\frac{\sqrt{t}}{\sigma'}) \quad (first\&second \quad term),
\end{aligned}$$

where $\Phi^{(4)}(t) = \frac{1}{\sigma^4}(t^4 - 6t^2 + 3)\Phi(t)$ for standard normal distribution by Gram-Charlier definition. Thus, the approximation using first two terms can be written as

$$\begin{aligned}
&2\Phi(\frac{\sqrt{t}}{\sigma'}) - 1 + \frac{p}{4(n-2)(n-p-4)}\left\{\frac{1}{\sigma^4}\left[(\frac{\sqrt{t}}{\sigma'})^4 - 6(\frac{\sqrt{t}}{\sigma'})^2 + 3\right]\left(2\Phi(\frac{\sqrt{t}}{\sigma'}) - 1\right)\right\} \\
&= \left(2\Phi(\frac{\sqrt{t}}{\sigma'}) - 1\right)\left(1 + \frac{p}{4\sigma^4(n-2)(n-p-4)}\left[\frac{t^2}{MSE^2} - 6\frac{t}{MSE} + 3\right]\right).
\end{aligned} \tag{6.11}$$

Plug in that $MSE = \sigma^2\frac{(n+1)(n-2)}{n(n-p-2)}$

The results of the MAE approximation suggest that the normal distribution is a satisfactory approximation for the MSE distribution and that including the second term of the Gram-Charlier series is not necessary. Therefore, it may be reasonable to consider the use of the product of a constant and a Chi-square distribution with one degree of freedom, represented as

$$\sigma^2\chi_1^2, \tag{6.12}$$

as a viable option for approximating the MSE distribution.

6.2.2 Approximation Simulation

In order to maintain consistency with the parameter settings employed in the MAE approximation, we retained the same values of p (1, 2, 3, 5, 8, and 10) and sample sizes n (25, 50, 75, and 100) for our simulation study. To satisfy the condition $n - p \geq 5$, we excluded the sample size of 10 for cases where p was equal to 8 or 10, while also including a sample size of 100. For each parameter setting, we assumed a constant value for the mean of the predictors, denoted as $\boldsymbol{\mu}$, as well as a constant covariance between every pair of predictors, and also set the intercept of the linear model, α, and the error variance, σ^2, as constant.

To simulate the unconditional mean square error $(\hat{y} - y^*)^2$, we generated 10,000 random samples for each combination of n and p. In each simulation, the response variable was generated using a linear relationship with the predictors, and the new data was drawn from a normal distribution with the same mean and variance as the predictors. We then calculated the empirical cumulative distribution function and compared it with the predicted values obtained from the Gram-Charlier approximations, gamma approximation, and chi-square approximation.

The results of the simulation study are presented in Table 6.1. In this table, $\hat{F}$ represents the empirical distribution function of $(\hat{y} - y^*)^2$, which corresponds to the true distribution F of the MSE. $\hat{F}$ is calculated from the 10,000 random samples in each category defined by the values of n and p. F_1 and F_2 denote the computed approximations to F based on the first and second Gram-Charlier approximations, as Equation(6.10) and Equation(6.11) respectively. F_3 and F_4 correspond to the gamma distribution as equation Equation(6.9) and chi-square distribution as Equation(6.12), respectively, as discussed previously. MSE_{est} and $Var(SE)_{est}$ represent the estimated mean and variance, respectively, for the square error using equations Equation(6.6) and Equation(6.8). To account for randomness in the simulation, several random seeds were employed.

Number of Predictors p	Sample Size n	$\max\left\|\hat{F}_t - F_1\right\|$	$\max\left\|\hat{F}_t - F_2\right\|$	$\max\left\|\hat{F}_t - F_3\right\|$	$\max\left\|\hat{F}_t - F_4\right\|$	$\left\|MSE_{true} - MSE_{ep}\right\|$	$\left\|VarSE_{true} - VarSE_{ep}\right\|$
1	10	0.0132	0.0078	0.0140	0.0123	0.0225	0.2350
	25	0.0062	0.0059	0.0062	0.0069	0.0052	0.0163
	50	0.0062	0.0063	0.0065	0.0062	0.0170	0.1008
	75	0.0061	0.0061	0.0060	0.0063	0.0175	0.1722
2	10	0.0063	0.0206	0.0497	0.0057	0.0097	0.0204
	25	0.0107	0.0115	0.0142	0.0102	0.0086	0.0739
	50	0.0075	0.0078	0.0082	0.0064	0.0079	0.0200
	75	0.0107	0.0106	0.0104	0.0084	0.0031	0.1118
3	10	0.0281	0.0132	0.0763	0.0284	0.0229	0.9169
	25	0.0121	0.0101	0.0085	0.0104	0.0268	0.1493
	50	0.0081	0.0085	0.0087	0.0074	0.0300	0.1267
	75	0.0071	0.0070	0.0067	0.0100	0.0037	0.0419
5	10	0.0509	0.1252	0.2775	0.0507	0.0354	7.6245
	25	0.0094	0.0063	0.0095	0.0094	0.0055	0.0873
	50	0.0097	0.0094	0.0097	0.0126	0.0258	0.0533
	75	0.0065	0.0067	0.0073	0.0092	0.0088	0.0014
8	25	0.0060	0.0076	0.0215	0.0069	0.0009	0.0075
	50	0.0047	0.0056	0.0079	0.0058	0.0038	0.0537
	75	0.0076	0.0078	0.0075	0.0076	0.0239	0.1107
	100	0.0111	0.0114	0.0114	0.0105	0.0002	0.0385
10	25	0.0070	0.0120	0.0317	0.0067	0.0358	0.2587
	50	0.0114	0.0129	0.0152	0.0093	0.0188	0.0632
	75	0.0078	0.0084	0.0083	0.0083	0.0066	0.0119
	100	0.0077	0.0078	0.0083	0.0074	0.0166	0.0365

Table 6.1: Absolute Value of Difference Between the Empirical Distribution Function and Approximation Distribution Function

To evaluate the performance of each approximation, we calculated the absolute value of the maximum difference between the estimated distribution function and the empirical distribution function for different random seeds. Although we identified the best-performing approximation for each category, the differences between the four estimates were not statistically significant and were generally below 0.015, with the exception of the case where $(p, n) = (5, 10)$. In this case, the variance of the square error estimation deviated from the empirical square error variance, indicating that a larger inflation factor K may lead to poor accuracy of the approximation. Taking this into account, we determined that the chi-square approximation was the best-performing approximation for the case where $(p, n) = (5, 10)$.