Prediction Accuracy on Adding New Predictors In Linear Regression

Yifan Ma

Worcester Polytechnic Institute

April 5, 2023

Introduction

- While a larger sample size can lead to more accurate predictions, there is a limit to the effect of sample size on model performance.
- New predictors can provide additional information about the outcome variable.
- The effect size of newly discovered predictors usually only considers this new variable and controls the effect size of other existing variables in medical reports.
- Investigating the relationship between prediction accuracy and newly developed predictors considering sample size as a reference point can improve the accuracy and reliability of research results.

Introduction

- This thesis investigates the relationship between sample size, predictors, effect sizes, and Predictive Mean Squared Error (PMSE).
- The impact of various factors that contribute to PMSE will be measured.
- Quantify the influence of new predictors on the model's prediction accuracy using reduced prediction mean squared error percentage (*pPMSEr*) or correlation between true value and prediction.
- Provide calculations based on data to support findings.

Literature Review

- When predictor variables are stochastic or random, the accuracy of predictions can be compromised.
- Kerridge (1967) examines the predictive errors of multiple regression equations when the regressor variables are treated as random variables drawn from a multivariate normal population.

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

- Narula (1974) 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.

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

Literature Review

- Sawyer (1982) explores the accurate approximate distribution of mean absolute error (MAE) as a combination of normal distribution and derivative of the normal distribution.

$$Prob(|\hat{y} - y| \leq t) = \Phi\left(\frac{t}{\sigma'}\right) + \frac{p}{4(n-2)(n-p-4)}\Phi^{(4)}\left(\frac{t}{\sigma'}\right) \quad (3)$$

- These approaches are useful for improving the accuracy of predictions in practical applications, particularly when dealing with a large number of regressor variables.
- Additionally, the studies provide valuable insights into variable selection in regression analysis and predicting accuracy before collecting sample data.

Method

We consider a sample of n independent individuals. 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 $\boldsymbol{Z} = (\boldsymbol{z_1'}, ..., \boldsymbol{z_n'})'$.

Model Assumptions

- The response and predictors follow a multivariate normal distribution.
- The covariance matrix of the response and predictors is given by $\boldsymbol{\Sigma}^*$, which has the following form:

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

  where $\boldsymbol{\sigma}$ is the covariance vector, $\boldsymbol{\Sigma}$ is the variance matrix, and σ_{00} is the variance of response $\boldsymbol{y}$.
- The k predictors can be divided into two subsets: p known as "basic predictors" and $k - p$ as "new predictors".

Method

Full Regression Model

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

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

where ϵ_i follows a normal distribution with mean zero and variance $\sigma_k^2 = \sigma_{00} - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$.

Full Model Effects

$\boldsymbol{\beta}$ denotes the full-model effects and is given by:

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

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.

Method

Reduced Regression Model

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

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

where $\epsilon_i^{\sharp}$ follows a normal distribution with mean zero and variance

$$\sigma_p^2 = \sigma_{00} - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} \tag{4}$$

Reduced Model Effects

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:

$$\boldsymbol{\beta}_1^{\sharp} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_1. \tag{5}$$

Method

Percentage of PMSE Reduction

- PMSE of the full regression: $PMSE = \sigma_k^2 \frac{(n+1)(n-2)}{n(n-k-2)}$
- PMSE of the reduced regression: $PMSE_1 = \sigma_p^2 \frac{(n+1)(n-2)}{n(n-p-2)}$
- Percentage of PMSE reduction:
  $pPMSEr = \left(1 - \frac{\sigma_k^2}{\sigma_p^2} \cdot \frac{n-p-2}{n-k-2}\right) \times 100\%$
- Error variance ratio (EVR): $EVR = \frac{\sigma_k^2}{\sigma_p^2}$
- Inflation factor: $\lambda(n; p, p_2) = \frac{n-p-2}{n-k-2} = \frac{1}{1-\frac{p_2}{n-p-2}}$

Method

Efficient Sample Size

- Sample size: larger sample size n tends to make the inflation factor $\lambda(n; p, p_2)$ approach 1.
- Number of new predictors: increasing p_2 tends to increase the inflation factor.
- 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:

$$n^* = p + 2 + p_2 \cdot \frac{\lambda^*}{\lambda^* - 1}, \text{ where } \lambda^* = 1 + \alpha(\frac{1}{EVR} - 1). \quad (6)$$

Effect Size

- One representative measure is Cohen's f^2, which is based on the proportion of the variation explained by the predictors.
- Let $R^2 = \frac{\sigma_{00} - \sigma_k^2}{\sigma_{00}} = \frac{\boldsymbol{\sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}}{\sigma_{00}}$ be the proportion of the response's variance accounted for by all k predictors, and $R_1^2 = \frac{\sigma_{00} - \sigma_p^2}{\sigma_{00}} = \frac{\boldsymbol{\sigma}_1' \boldsymbol{\Sigma} 11^{-1} \boldsymbol{\sigma}_1}{\sigma 00}$ be the proportion of the response's variance accounted for by p basic predictors.
- 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

$$f_2^2 = \frac{R^2 - R_1^2}{1 - R^2} = \frac{\sigma_p^2 - \sigma_k^2}{\sigma_k^2} = \frac{1 - \sigma_k^2/\sigma_p^2}{\sigma_k^2/\sigma_p^2}. \tag{7}$$

Example

The coefficients utilized in this study were based on the findings reported by Baker et al. (2008). However, as the paper only provided a correlation matrix, a standard deviation of $SD = 1$ was used to derive the covariance matrix. Given the correlation matrix provided in the original paper, and assuming the covariates were standardized, the covariance matrix can be obtained using the formula:

$$\mathbf{Cov} = \text{diag}(\boldsymbol{SD}) \times \mathbf{Cor} \times \text{diag}(\boldsymbol{SD})$$

The response variable, Pain intensity, and the $k = 12$ predictors were assumed to follow a multivariate normal distribution.

Example

The full regression model included $k = 12$ predictors, while the reduced regression model included the first $p = 3$ predictors. 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. The relationship between pPMSEr, sample size, and Cohen's f^2 was examined.
The variances of the error terms in the full and reduced regression models were calculated as follows: $\sigma_k^2 = 0.4687399$ and $\sigma_p^2 = 0.9393167$. The full model effects were obtained using:

$$\begin{aligned}\boldsymbol{\beta} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} \\ &= (-0.13, 0.07, 0.10, 0.50, 0.23, -0.15, 0.18, -0.05, -0.47, 0.43, 0.14, 0.21).\end{aligned} \tag{8}$$

Example

pPMSEr

The "improvement" of prediction by adding the new $k - p = 9$ health and psychological predictors can be measured by the "percentage of PMSE reduction":

$$\begin{aligned} pPMSEr &= \left(\frac{PMSE_1 - PMSE}{PMSE_1} \right) \times 100\% \\ &= \left(1 - \frac{\sigma_k^2}{\sigma_p^2} \cdot \frac{n-p-2}{n-k-2} \right) \times 100\% = 47.41\%. \end{aligned}$$

in this example, the introduction of psychological predictors into the model increased the model's prediction accuracy by 47%.

Example

Efficient Sample Size

It shows that the prediction accuracy is stable when the sample size equals or exceeds a threshold. 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 *pPMSEr* should be greater than 0.1. On the flip side, with 181 sample size, the "efficiency" $1 - \alpha = 0.953$.

Example

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 et al. written below:

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

- By the definition of the squared multiple correlations R^2 and EVR, Cohen's f^2 can be calculated $f_2^2 = 0.3328571$.

Example

Simulation

- Simulation investigates impact of sample size and inclusion of "non-basic" predictors on prediction accuracy.
- Hypothesis: Adding more variables will decrease efficient sample size, resulting in stable prediction accuracy that does not significantly improve with larger sample size.
- Data generated using covariate matrix Σ from Equation Table 2 by Bakers et al..
- "Non-basic" predictors were sequentially added to the model.
- Prediction Mean Square Error (PMSE) and correlation between true value and prediction calculated using mean of 1000 iterations.

Example

Result

Sample Size	Basic Model	Comorbidities	Pain Locations	Medications	Physical Functioning	Depressive Symptoms	Life Satisfaction	LOC chance	LOC powerful	LOC internal
30	1.6148	0.9757	0.9690	1.0203	1.0174	1.0791	1.1203	1.0558	1.0558	1.1025
60	1.5183	0.9319	0.9095	0.9421	0.9089	0.9463	0.9571	0.8822	0.8760	0.8866
90	1.4622	0.9004	0.8709	0.8946	0.8500	0.8699	0.8724	0.7894	0.7793	0.7788
120	1.4317	0.8765	0.8443	0.8616	0.8146	0.8247	0.8225	0.7369	0.7239	0.7187
150	1.4124	0.8611	0.8264	0.8388	0.7883	0.7935	0.7891	0.7011	0.6873	0.6773
180	1.3997	0.8518	0.8149	0.8240	0.7731	0.7742	0.7676	0.6781	0.6629	0.6503
210	1.3865	0.8438	0.8051	0.8122	0.7606	0.7595	0.7514	0.6614	0.6457	0.6311
240	1.3786	0.8364	0.7965	0.8021	0.7496	0.7458	0.7368	0.6466	0.6299	0.6139
270	1.3727	0.8304	0.7902	0.7943	0.7418	0.7363	0.7260	0.6359	0.6181	0.6010
300	1.3678	0.8268	0.7853	0.7886	0.7354	0.7284	0.7173	0.6266	0.6084	0.5904
330	1.3628	0.8240	0.7815	0.7837	0.7301	0.7217	0.7100	0.6189	0.6003	0.5814
360	1.3575	0.8212	0.7779	0.7792	0.7252	0.7156	0.7034	0.6123	0.5934	0.5738
390	1.3532	0.8181	0.7744	0.7750	0.7208	0.7105	0.6978	0.6063	0.5872	0.5670
420	1.3495	0.8161	0.7719	0.7720	0.7176	0.7067	0.6935	0.6018	0.5825	0.5617
450	1.3463	0.8144	0.7696	0.7692	0.7146	0.7032	0.6895	0.5977	0.5782	0.5570

Table: Prediction Mean Square Error by sequentially added predictors over the "basic" 3 predictors

Example

Result

Sample Size	Basic Model	Comorbidities	Pain Locations	Medications	Physical Functioning	Depressive Symptoms	Life Satisfaction	LOC chance	LOC powerful	LOC internal
30	0.0870	0.3263	0.3723	0.3457	0.3724	0.3711	0.3761	0.4334	0.4624	0.4466
60	0.0830	0.3564	0.4012	0.3893	0.4297	0.4319	0.4447	0.5082	0.5253	0.5307
90	0.0826	0.3800	0.4273	0.4181	0.4647	0.4709	0.4838	0.5531	0.5697	0.5784
120	0.0907	0.3968	0.4446	0.4376	0.4856	0.4945	0.5080	0.5785	0.5945	0.6045
150	0.0927	0.4087	0.4566	0.4516	0.5017	0.5123	0.5252	0.5962	0.6119	0.6239
180	0.0968	0.4182	0.4663	0.4631	0.5127	0.5241	0.5373	0.6084	0.6241	0.6373
210	0.1020	0.4243	0.4727	0.4703	0.5204	0.5318	0.5454	0.6166	0.6321	0.6462
240	0.1045	0.4295	0.4776	0.4763	0.5270	0.5391	0.5526	0.6238	0.6397	0.6541
270	0.1075	0.4336	0.4811	0.4808	0.5312	0.5438	0.5579	0.6289	0.6453	0.6601
300	0.1111	0.4365	0.4847	0.4849	0.5354	0.5484	0.5627	0.6339	0.6502	0.6655
330	0.1138	0.4394	0.4880	0.4889	0.5394	0.5529	0.5672	0.6385	0.6547	0.6704
360	0.1162	0.4413	0.4903	0.4919	0.5425	0.5565	0.5709	0.6420	0.6583	0.6741
390	0.1182	0.4432	0.4923	0.4942	0.5449	0.5590	0.5735	0.6449	0.6611	0.6773
420	0.1198	0.4446	0.4937	0.4959	0.5468	0.5610	0.5757	0.6471	0.6634	0.6798
450	0.1214	0.4459	0.4954	0.4979	0.5487	0.5630	0.5779	0.6494	0.6656	0.6821

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

Example

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 for Correlation and pPMSEr, as demonstrated in Table 3.

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: Efficient sample size n

Discussion

- The study examined the relationship between *pPMSEr*, sample size n, and Cohen's f^2, finding a significant relationship where *pPMSEr* increased as sample size and effect size increased.
- Non-basic predictors significantly contributed to explaining the variability in pain intensity. Physical functioning, Depressive symptoms, and LOC-internal were identified as the most significant predictors of pain intensity.
- The study's limitations included the consideration of only one quantitative outcome variable, a limited set of predictors, and the use of a correlation matrix rather than individual-level data. Future research should consider general linear models and explore the interaction between predictors.

Reference

Baker TA, C. N., Buchanan NT (2008). Factors influencing chronic pain inten- sity 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 develop- ing 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 develop- ing 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 dichoto- mous 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.

Acknowledge

Thanks for you time!