Prediction Mean Square Error Calculation based on PMSE Improvement

1 Aim

The report will summarize the degree of influence of the new predictors on the prediction accuracy of the model, reflected in $rPMSEp$ or $cohen'sf^2$, corresponding to the efficient sample size required to achieve a certain prediction accuracy as a reference. Calculations based on data in Baker TA (2008) are provided.

2 Calculation

The calculation is based on the document "Sample size, the number of predictors and effects influence PMSE". The coefficients used to generate data refer to the result of the paper Baker TA (2008). Since the paper gives only the correlation matrix and the regression coefficients are seen as standardized, the correlation matrix is used as the standardized covariance matrix

$$\boldsymbol{\Sigma}^* = diag(SD) * Corr * diag(SD)$$

where SD is standard deviation vector.

Consider a sample of $n = 181$, total $k = 12$ predictors, of which $p = 3$ predictors are "basic predictors".The response and the predictors follow a multivariate normal distribution. That is,

$$(\boldsymbol{Y}, \boldsymbol{Z}) \sim MVN(\boldsymbol{\mu^*}, \boldsymbol{\Sigma}^*),$$

where unknown mean vector$\boldsymbol{\mu^*} = (\mu_0, \boldsymbol{\mu}')'$, and unknown covariance matrix $\boldsymbol{\Sigma^*} = \begin{pmatrix} \sigma_{00} & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{pmatrix}$.

Based on the distribution of $(Y, \boldsymbol{Z})$, the "full regression" model containing k predictors is

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

where the error term $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_k^2)$ is independent of $\boldsymbol{z_i}$ for each i, and $\sigma_k^2 = \sigma_{00} - \boldsymbol{\sigma'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$. The reduced model can be written as

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

where, the error term $\epsilon_i^\sharp \overset{\text{iid}}{\sim} N(0, \sigma_p^2)$ is independent of $\boldsymbol{z_{1i}}$ for each i, and $\sigma_p^2 = \sigma_{00} - \boldsymbol{\sigma_1'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_1$.

The response *Painintensity* and the $k = 12$ predictors follow a multivariate normal distribution with $\boldsymbol{\mu}^* = \mathbf{0}$ and

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

The predictors are divided into demographic, health, and psychological factors according to Baker TA (2008). The reduced regression model contains demographic factors. The full regression model considered health and psychological factors while controlling the reduced predictors. The variance of the error term in full regression σ_k^2 and that in reduced regression σ_p^2 is calculated by $\sigma_k^2 = \sigma_{00} - \boldsymbol{\sigma'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} = 0.4687399$ and $\sigma_p^2 = \sigma_{00} - \boldsymbol{\sigma_1'}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\sigma_1} = 0.9393167$ in which the variance of response $\sigma_{00} = 1$ as standardized.

2.1 Effects

The "full model effects" by Narula (1974)are

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

The coefficients calculated in (1) are not the same as the coefficients given in Baker TA (2008) Table 2 written below:

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

since $\boldsymbol{\beta}$ is calculated considering all of the predictors in the full model. Whereas, $\boldsymbol{\beta^*}$ is calculated with added-up predictors controlling the prior sets of predictors.

Consider $p = 3$ "basic predictors" i.e. demographic predictors with a partitioned covariance matrix, the "reduced-model effects" are

$$\begin{aligned} \boldsymbol{\beta}_1^\sharp &= \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_1 \\ &= (-0.25, -0.04, -0.02). \end{aligned} \tag{2}$$

The effects for demographic factors in $\boldsymbol{\beta}_1^\sharp$ is closer to the effects for demographic factors in $\boldsymbol{\beta}^*$.

2.2 PMSE

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

$$\begin{aligned} pPMSEr &= \left(\frac{PMSE_1 - PMSE}{PMSE_1}\right) \times 100\% \\ &= \left(1 - \frac{\sigma_k^2}{\sigma_p^2} \cdot \frac{n-p-2}{n-k-2}\right) \times 100\% = 47.41\%. \end{aligned} \tag{3}$$

The prediction mean square error measures the expected squared distance between what your predictor predicts for a specific value and what the true value is whereas $pPMSEr$ is used to demonstrate how much accuracy the new $k-p=9$ predictors bring to the model. In this example, the introduction of psychological predictors into the model increased the model's prediction accuracy by 47%.

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 illustrated by Prof. Wu in the report for SampleSizeAnalysis EPPIC. 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" $1-\alpha=0.1$(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 181 could reach 95.3% of the largest pPMSEr at $n=\infty$.

2.3 Cohen's f^2

Cohen's f^2 for the effects of new predictors conditional on the known predictors is defined as

$$f_2^2 = \frac{R^2 - R_1^2}{1-R^2} = \frac{\sigma_p^2 - \sigma_k^2}{\sigma_k^2} = \frac{1 - \sigma_k^2/\sigma_p^2}{\sigma_k^2/\sigma_p^2}. \tag{4}$$

which gives

$$\frac{\sigma_k^2}{\sigma_p^2} = \frac{1}{f_2^2 + 1}. \tag{5}$$

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:

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

By the definition of the squared multiple correlations R^2 and (5), 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 $f^2 \geq 0.15$.

References

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)*.

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.