Prediction Mean Square Error Calculation based on PMSE Improvement

December 8, 2022

1 Model

Consider a sample of n independent individuals, for the i^{th} individual, the vector of all k predictos is $\boldsymbol{z_i'} = (z_{i1}, ..., z_{ik}), i = 1, 2, ..., n$. Assume we have total k predictors, of which p predictors are "basic predictors" proven to have the information to explain the variation of the response from prior studies, corresponding to $\boldsymbol{z_{1i}'} = (z_{i1}, ..., z_{ik})$. The rest $k - p$ are "new predictors" expected to account for more variation of the response in newly proposed study, corresponding to $\boldsymbol{z_{2i}'} = (z_{i,(p+1)}, ..., z_{ik})$. For i^{th} individual, the vector of predictors is partitioned into two parts $\boldsymbol{z_i'} = (\boldsymbol{z_{1i}'}, \boldsymbol{z_{2i}'})$.

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

Similarly, the covariance vector could be partitioned into two parts $\boldsymbol{\sigma} = Cov(y_i, \boldsymbol{z_i}) = (\boldsymbol{\sigma_1'}, \boldsymbol{\sigma_2'})'$, and the variance matrix of predictors could be partitioned into

$$Var(\boldsymbol{z_i}) = \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}$$

.

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 "full-model effects" are

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

Consider partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, we have

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

Consider p "basic predictors" with partitioned effects, 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 "reduced-model effects" are

$$\boldsymbol{\beta}_1^\sharp = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_1. \tag{2}$$

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

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

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

$$\boldsymbol{\sigma} = (\Sigma_{11}\beta_1^*, ..., \Sigma_{kk}\beta_k^*)' = \text{diag}(\boldsymbol{\Sigma})\boldsymbol{\beta}^*. \tag{3}$$

Following this equation, the coefficients/effects in joint models (1) and (2) can be obtained.

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\text{diag}(\boldsymbol{\Sigma})\boldsymbol{\beta}^*; \tag{4}$$
$$\boldsymbol{\beta}_1^\sharp = \boldsymbol{\Sigma}_{11}^{-1}\text{diag}(\boldsymbol{\Sigma}_{11})\boldsymbol{\beta}_1^*. \tag{5}$$

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

2 Calculation Method

2.1 PMSE

According to **?**, the predictive mean squared error (PMSE) based on the least squares estimator (LSE) of the full regression with k predictors is

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

The PMSE based on the LSE of the reduced regression with p predictors is

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

The "improvement" of prediction by adding the new $k-p$ predictors $\boldsymbol{z_{2i}}$ can be measured by the "percentage of PMSE reduction":

$$pPMSEr = \left(\frac{PMSE_1 - PMSE}{PMSE_1}\right) \times 100\% = \left(1 - \frac{\sigma_k^2}{\sigma_p^2} \cdot \frac{n-p-2}{n-k-2}\right) \times 100\%. \quad (6)$$

The "error variance ratio" is defined as

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

And, the inflation factor

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

where $p_2 = k - p$ is the number of "new" predictors added to the reduced model to form the full model.

2.1.1 Efficient sample size

since pPMSEr is increasing in n, we define "efficient sample size" n^* as the smallest sample size such that

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

where we call $1-\alpha$ the "efficiency" (e.g., 90% of the largest pPMSEr at $n = \infty$). By

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

we get

$$n^* = p + 2 + p_2 \cdot \frac{\lambda^*}{\lambda^* - 1}, \text{ where } \lambda^* = 1 + \alpha(\frac{1}{EVR} - 1). \quad (7)$$

2.2 Cohen's f^2

For measuring effect size, Cohen's f^2 is a representative measure 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 response's variance accounted for by all k predictors. $R_1^2 = \frac{\sigma_{00} - \sigma_p^2}{\sigma_{00}} = \frac{\boldsymbol{\sigma_1}' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\sigma_1}}{\sigma_{00}}$ be the proportion of response's variance accounted for by p known predictors.

Cohen's f^2 for the effects of all predictors is $f^2 = \frac{R^2}{1-R^2}$. Cohen's f^2 for the effects of new predictors conditional on the known predictors is

$$f_2^2 = \frac{R^2 - R_1^2}{1 - R^2} = \frac{\sigma_p^2 - \sigma_k^2}{\sigma_k^2} = \frac{1 - \sigma_k^2/\sigma_p^2}{\sigma_k^2/\sigma_p^2}. \tag{8}$$

which gives

$$\frac{\sigma_k^2}{\sigma_p^2} = \frac{1}{f_2^2 + 1}. \tag{9}$$

3 Parameter Setting

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, Buchanan NT, Corson N. Factors influencing chronic pain intensity in older black women: examining depression, locus of control, and physical health." Since the paper gives only the correlation matrix but we cannot obtain the standard deviation to get the covariance matrix. We used the empirical covariance matrix from the simulation for calculation.

In the calculation, the number of predictors in full regression is $k = 12$, that in reduced regression is $p = 7$, corresponding the predictors, Age, Education, Income, Comorbidities, Pain locations, Medications, Physical functioning, Depressive symptoms, Life satisfaction, LOC-chance, LOC-powerful, LOC-internal, in which Age, Education, Income, Comorbidities, Pain locations, Medications, Physical functioning as "basic" predictors and the rest $p_2 = k - p = 5$ as "new" predictors. The empirical variance of response, the empirical covariance matrix, corresponding $\hat{\sigma_k^2}$ and $\hat{\sigma_p^2}$ are estimated by generated multinormal distribution. The regression model is fitted by $lm()$ in R with LSE by default.

Though the correlation matrix is not available. The R^2 for reduced regression and full regression are given,

$$R_1^2 = 0.31, R_2^2 = 0.44$$

By the definition of the squared multiple correlations R^2 and (9), Cohen's f^2 can be calculated $f_2^2 = 0.2321$.

The relationship between pPMSEr, sample size, and Cohen's f^2 has been investigated. The sample size ranges from 50 to 500. The variance of the error term in full regression is assumed to be 1 and predictor data are standardized with a mean is 0. Each result of the calculation is the mean of 1000 iterations.

4 Calculation Result

Table 1 summarizes the result of the calculation. In the table, the $\hat{\sigma_k^2}$ and $\hat{\sigma_p^2}$ are estimated value of σ_k^2 and σ_p^2 for each sample size. $PMSEres$ and $PMSE_1res$ are the residuals

between the empirical value and estimated value by $PMSE = \hat{\sigma_k^2}\frac{(n+1)(n-2)}{n(n-k-2)}, PMSE_1 = \hat{\sigma_p^2}\frac{(n+1)(n-2)}{n(n-p-2)}$. $pPMSEr$ represents the percentage of PMSE reduction to measure the "improvement" of prediction by adding the new predictors $\boldsymbol{z_{2i}}$. The effect size measure Cohen's f^2 for new predictors conditional on the known predictors is given. The efficient sample size n^* with efficiency $1 - \alpha = 90\%$ is given for reference.

sample size	$\hat{\sigma_k^2}$	$\hat{\sigma_p^2}$	$PMSEres$	$PMSE_1res$	pPMSEr	n^*	cohen's f^2
50	0.7584647	0.9717771	0.30140513	0.169383114	0.1067495	275.0206	0.2975529
75	0.8383694	1.0239119	0.2078505	0.1486635	0.1108186	328.7663	0.2285396
100	0.8777050	1.0490777	0.15651218	0.095599615	0.1125238	347.8343	0.1999096
150	0.9200495	1.0793241	0.06765876	0.037431969	0.1146989	365.3131	0.1755574
200	0.9397086	1.0928815	0.08291560	0.047955323	0.1159664	368.4875	0.1647422
250	0.9517625	1.0998664	0.06522052	0.049993444	0.1154573	376.4712	0.1569149
300	0.9598517	1.1060478	0.01849940	0.009434911	0.1162766	377.3857	0.1533460
350	0.9660424	1.1101825	0.03174622	0.011635561	0.1162782	379.1989	0.1500834
400	0.9694059	1.1112636	0.03645961	0.016546566	0.1158297	383.1952	0.1470855
450	0.9737744	1.1149450	0.03596313	0.023135953	0.1161016	384.1469	0.1456229
500	0.9769832	1.1170268	0.02686947	0.026124696	0.1159573	385.4763	0.1439371

Table 1: Result

By table 1, we observe

- The estimated value $\hat{\sigma_k^2}$ is approaching assumption $\sigma_k^2 = 1$ when larger sample size is given. Also, when the sample size increases, the error variance ratio (EVR) is getting closer to 1, which gives a larger efficient sample size n^*.
- The PMSE estimation for reduced regression is closer to its empirical value than that for full regression. With a larger sample size, the estimation performs better.
- When the actual sample size exceeds the efficient sample size n^*, the increase in $pPMSEr$ is not significant. The pPMSEr has already reached efficiency may be the reason for this phenomenon. For this part, if we want to validate the reliability of efficient sample size, we should use a larger EVR to generate the predictors.
- Based on the true value of Cohen's f^2, the sample size used in the paper might be around 75