Prediction Accuracy Over Sample Sizes and Added-on Non-basic Predictors

1 Aim

This report aims to use calculations or simulations to demonstrate the relationship between sample size and newly added predictors. We believe that the more variables are added, the smaller the efficient sample size, that is, the prediction accuracy tends to be more stable and does not increase significantly when more predictors are added and using a smaller sample size.

2 Result

Based on the PMSE improvement document and SampleSizeAnalysisEPPICWu. The data is generated with the covariate matrix $\mathbf{\Sigma}$ from Baker TA (2008). The ”Basic” model contains $p = 3$ demographic predictors. The ”non-basic” predictors are added sequentially into the model being evaluated by the rPMSEp. The calculated rPMSEp is shown as 1. The efficient sample size with $\alpha = 0.1$ (i.e. reaches 90% of the largest pPMSEr at n = ∞)is calculated as a reference in 2.

Sample Size	Basic Predictors	Comorbidities	Pain Locations	Medications	Physical Functioning	Depressive Symptoms	Life Satisfaction	LOC -Chance	LOC -Powerful
60	0.4033	0.2884	0.2512	0.2559	0.2041	0.1919	0.1806	0.0549	0.0327
90	0.4419	0.3300	0.2903	0.2898	0.2349	0.2174	0.2004	0.0704	0.0408
120	0.4586	0.3481	0.3072	0.3045	0.2482	0.2285	0.2090	0.0771	0.0443
150	0.4680	0.3582	0.3167	0.3127	0.2557	0.2347	0.2138	0.0809	0.0463
180	0.4739	0.3646	0.3227	0.3180	0.2605	0.2387	0.2169	0.0833	0.0476
210	0.4781	0.3691	0.3269	0.3216	0.2638	0.2414	0.2190	0.0850	0.0484
240	0.4811	0.3724	0.3300	0.3243	0.2662	0.2434	0.2206	0.0862	0.0491
270	0.4834	0.3749	0.3323	0.3264	0.2681	0.2449	0.2218	0.0872	0.0496
300	0.4853	0.3769	0.3342	0.3280	0.2695	0.2462	0.2227	0.0879	0.0500
330	0.4868	0.3785	0.3357	0.3293	0.2707	0.2472	0.2235	0.0885	0.0503
360	0.4880	0.3798	0.3370	0.3304	0.2717	0.2480	0.2241	0.0890	0.0505

Table 1: Sequentially added predictors over the “basic” 3 predictors

- The value in 1 shows the proportion of variation explained by different combinations of the predictor variables The complete model includes an additional predictor, $LOC-internal$, which is not listed in the table. This is because the rPMSEp was compared to the full model, which resulted in a rPMSEp value of 1.

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

Table 2: Efficient sample size n^*

- Once the sample size reached the efficient sample size of n^*, the rPMSEp was expected to remain stable as the sample size continued to increase. This result confirms the formula of efficient sample size n^*.

- When insignificant predictors are incorporated into the model, the efficient sample size n^* will not decrease but rather increase. Including these predictors does not yield additional information but instead introduces random error, which implies that achieving the same prediction accuracy would require a larger sample size.

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).*