Chapter 1

Summary

1. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes

   When we are using linear regression to predict contimuous outcomes,Four criteria for obtaining a suitable sample size are presented in this article.

   - ensure a shrinkage factor

   $$S_c = 1 + \frac{p-2}{nln(1-(\frac{R^2_{adj}(n-p-1)+p}{(n-1)}))} \leq 0.9$$

   - ensure a small absolute difference in R^2_{adj} and R^2_{app}

   $$\frac{p(1-R^2_{adj})}{(n-1)} \leq \delta$$

   - precise estimate of the residual standard deviation.
     MMOE for estimating σ_{model})

   $$MMOE = \sqrt{max(\frac{\chi^2_{1-\frac{\alpha}{2},n-p-1)}}{n-p-1}, \frac{n-p-1}{\chi_{\frac{\alpha}{2},n-p-1}})}$$

   - precise estimate of the mean predicted outcome value $1.0 \leq MMOE \leq 1.1$

   $$MMOE = t_{1-\frac{0.05}{2},n-p-1}\sqrt{\frac{\sigma^2_{null}(1-R^2_{adj})}{n}}$$

   This paper starts with avoiding overfitting while preserving accurate prediction performance. The global shrinkage factor is applied to all estimated predictor effects to adjust for overfitting. As the calculation of sample size shoule be done before the model fitting, we use R^2_{adj} as an unbiased estimate of R^2_{app}. The difference between R^2_{adj} and R^2_{app} represents the optimism in the developed model's apparent proportion of variance explained.

2. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes

   Basically the same idea as in continuous outcome. For binary outcome, the author considered the events per predictor parameter, EPP.

- Small optimism in predictor effect estimates as defined by a global shrinkage factor of ≤ 0.9
- Small absolute difference of ≤ 0.05 in the model's apparent and adjusted Nagelkerke's R^2

$$R^2_{Nagelkerke_app} - R^2_{Nagelkerke_adj} = \frac{R^2_{CS\_app}}{max(R^2_{CS\_app})} - \frac{R^2_{CS\_adj}}{max(R^2_{CS\_adj})}$$

The $R^2_{CS\_adj}$ can be derived by LR statistic, $pseudoR^2$ statistics, C statistic or directly $R^2_{CS\_adj}$ from existing models.

- Precise estimation of the overall risk or rate in the population. the margin of error in outcome proportion estimates $\hat{\phi}$ for a null model is (in 95%)

$$1.96\sqrt{\frac{\hat{\phi}(1-\hat{\phi})}{n}}$$

we recommend a more stringent margin of error ≤ 0.05.

3. A note on estimating the Cox-Snell R^2 from a reported C-statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome

- Simulate a large dataset (eg, one million participants)
- Assign an outcome of $Y_i = 0$ (no event) or $Y_i = 1$ (event) based on sampling from a $Bernoulli(\phi)$ distribution, where ϕ is the outcome proportion in the article reporting the existing prediction model
- Simulate LP_i values for every participant assuming $LP_i \sim N(0,1)$ in the non-events group and $LP_i \sim N(\mu,1)$ in the events group, where $\mu = \sqrt{2}\phi^{-1}(C)$.
- Fit a logistic regression to the simulated data; that is,

$$Y_i \sim Bernoulli(p_i), logit(p_i) = \alpha + \beta LP_i$$

This fitted model will have the same C statistic. The estimated values of coefficients ensure a perfect calibration-in-the-large (=0) and calibration slope (=1), respectively, in new data from the same assumed target population. We can obtain the $R^2_{CS}$for this fitted logistic regression model post estimation

4. Copas Using regression models for prediction: shrinkage and regression to the mean

When we are using the standard multiple regression model with response variable Y distributed as

$$Y \sim N(\alpha + \beta^{\mathrm{T}} X, \sigma^2)$$

The reason of shrinkage occurrence is explained by 'There is a sense in which the estimated model fits the data too well - shrinkage occurs because any unusual random features of the original data will be reflected in the predictions but not be replicated in a set of independent observations'. This may be because of the uniqueness of each set of data. The values of response variable for

the new patients will be closer to the overall mean than we would expect from an uncritical application of least squares or maximum likelihood.

If the same set of data is used both to select the covariate and to fit the model, then the values of the coefficients will be biased and the shrinkage of the predictor will be even more marked. This is because that if a regression coefficient is by chance overestimated (in absolute value) then it will be more likely to be selected than if it happened to be underestimated, hence the covariates which end up being selected for the model are likely to have larger coefficients. This means that the LS predictor will give too wide a range of values, high values of prediction$\hat{y}$ will be too high and low values of $\hat{y}$ will be too small.

5. Sample size and the accuracy of predictors made from multiple regression equations- RICHARD SAWYER

   The paper gives approximation function of measure prediction accuracy(mean absolute error) in terms of sample size and the number of predictors. The accuracy in estimating multiple regression coefficients depends on sample size and estimation error in estimating the coefficients propagates error in prediction. With multivariable normal assumption, we have approximations

$$MAE \doteq \sigma\prime\sqrt{\frac{2}{\pi}}$$

   where $\sigma\prime = \sqrt{MSE}$, and an inflation factor $K = \sqrt{\frac{(n+1)(n-2)}{[n(n-p-2)]}} > 1$ from moment approximation with $M = 1$. Controlling inflation factor K between 5% and 10% corresponds to the subject-to-variable ratio of 10 to 1, which is a well-known rulf of thumb.

6. Sample size for prediction of quantitative and binary outcomes based on cohort study

   - Cohen's f^2: one of the most common method for calculating the effect size of each of the variables or construct. Cohen categorized effect size as small($\geq 0,02$), medium(≥ 0.15) or large(≥ 0.35).

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

   - measure how well the factors explains / contributes to the model: f^2 for quantiative outcomes, AUC and Cox and Snell R^2 for binary ourcome.

7. Sample Sizes When Using Multiple Linear Regression for Prediction

   Multiple correlation coefficient is a measure of how well a given variable can be predicted using a linear function of a set of other variables. Higher values indicate higher predictability of the dependent variable from the independent variables, with a value of 1 indicating that the predictions are exactly correct and a value of 0 indicating that no linear combination of the independent variables is a better predictor than is the fixed mean of the dependent variable. The basic idea of finding the minimum sample size in this paper is to find the sample regression(with replications) that is most similar to the population regression. The criteria in this paper is the proportion of the correlation coefficients produced by sample regression replications which expected to ba large

enough than τ. The higer τ means better prediction. It turns out that as the value of ρ^2 decreases and as the number of predictor variables p increases, the recommended sample size n increases at an increasing rate.

8. COMPUTER EXPERIMENTS: PREDICTION ACCURACY, SAMPLE SIZE AND MODEL COMPLEXITY REVISITED

   In this paper the author provide the interrelationship of sample size n, complexity of the model represented by vector of hyperparameters θ and prediction accuracy. The integrated MSPE (IMSPE) is used as measure of prediction accuracy here. By minimizing the IMSPE we can expect to improve the predictive ability of the Kriging predictor. The Root Average Unexplained Variability (RAUV) of predictor $\hat{y}$ is proposed as a measure of expected prediction error on designing a computer experiment.

$$RAUV(\hat{y}; \mathscr{D}, \theta) = \sqrt{\frac{IMSPE(\hat{y}; \mathscr{D}, \theta)}{\sigma^2}}$$

   requiring $RAUV \leq 0.05$ means that we want the square root of the average squared length of our prediction intervals to shrink by 95% once data is observed, explaining at least $100(1-\varepsilon^2)\%$ of the variability of y(x) by the model.

   - Let λ_k be the set of eigenvalues of $R(x, \prime x; \theta)$. Let n_c be the critical sample size required to achieve $RAUV \leq \varepsilon$ for some $\varepsilon > 0$. Then

$$n_c \geq min\{n : \sqrt{\sum_{k \geq n+1} \lambda_k} \leq \varepsilon\} = min\{n : \sum_{k=1}^{n} \lambda_k \geq 1 - \varepsilon^2\}.$$

     we can derive analytically the required sample size for a given average level of prediction accuracy. The result of simulation, for a fixed sample size, the average unexplained variability grows fairly rapidly as the number of active factors increases. The more complex the response surface and the more active inputs influencing the response, the larger the sample size required. When the model is simple, the Gaussian process does an admirable job at computer model emulation.