Three criteria method for logistic regression

by Riley D Richard etc.

Xijin

9/19/2019

Contents

1 Abstract

In medical research, prediction model is usually used to predict individual's risk of disease or certain *AEs* for either *efficacy* or *safety*. Calculation of sample size is needed to ensure the performance of the prediction model on new individuals. Richard D Riley etc. showed three criteria, which would guide us to calculate the required sample size as pre-study power analysis.

Traditional way to calculate sample size focuses on number of events per variable (i.e. 10EPV 'rule of thumb') in logistic regreesion. However, this *three-criteria method* accounts for both the number of candidate predictors and potential for overfitting when developing a prediction model. To put this method into practice, this document shows the transparent process step by step and finally calculate the sample size required for a prediction model with good predicitve ability.

The document is divided into two parts. The first part focus on background information of this method and the second part shows required inputs and the sensitivity of those inputs on outcomes.

2 Background information about the prediction model

$$logit(p) = intercept + X\beta \tag{1}$$

Equation 1 indicates that estimation of the outcomes depends on two parts: intercept of the model and the candidate predictors. In the case of larger outcome proportion, *intercept* of the model would dominate and the candidate predictors do not matter any more. This is the common way to calculate sample size, ignoring the effects from possible predictors.

However, validity of predictions are what we are interested with when developing a prediction model. A key threat, when predicting outcomes on an independent dataset, is overfitting. The reason is the unusual random features of the original data are reflected in the prediction but not be replicated in a set of independent observations.

A widely-accepted solution to overfitting is shrinkage, which help us to shrinkage predictor effects on outcomes. This three-criteria method is based on a *global shrinkage factor* (or *uniform shrinkage factor*), a multiplicative measure of overfitting.

3 Description of three-criteria method

Based on the background information above, there are two motivations of the those three criteria:

1. Minimization of overfitting of the statistical model by shrinkage factor on both relative and absolute scale

Sample size calculation could be calculated based on the performance of prediction model, with controlling of overfitting. Therefore, pre-defined value of shrinkage factor would direct us to work out the required sample size. The larger shrinkage factor (or closer to 1), the better prediction model we would get, thus larger sample size would be required.

There are two approaches to calculate the shrinkage factor, *bootstrap* after estimation and *heuristic formula* before estimation. As our goal is to judge whether the sample size large enough for a prediction model before development of a prediction model, we would mainly focus on the heuristic formula for shrinkage.

The closed-form 'heurisic' shrinkage factor of Van Houweilingen and Le Cessie, defined by

$$S_{VH} = 1 - \frac{p}{LR}$$

where p is **number of predictors** and LR is the **likelihood ratio statistic**, which shows the difference of a null model (model with only intercept) and full model (model with all predictors). As LR is not easy to get and it depends on **sample size**, we re-express in another way in terms of **sample size n** and $R^2_{CS\_app}$.

$$LR = -nln(1 - R^2_{CS\_app})$$

.

$R^2_{CS\_app}$ denots the apparentestimate of a prediction model's Cox-Snell R^2 performance as obtained from the model development data set.

This leads to the close-form solution for shrinkage factor based on **sample size n**, **number of predictors p**, $R^2_{CS\_app}$.

$$S_{VH} = 1 + \frac{p}{nln(1 - R^2_{CS\_app})}$$

Therefore, we could connect the anticipated shrinkage factor to the quantification of performance of the prediction model on development dataset.

- Criterion (i) ensuring a global shrinkage factor of 0.9 (or a larger value if better performance of prediction model is required). With the closed-form solution for **global shrnkage factor**, it is eacy to identify **n** and **p** if we could specify a realistic value for $R^2_{CS_{app}}$. As $R^2_{CS_{app}}$ is an estimated based on the development data set, we would like to adjust the optimism due to *overfitting* for an unbiased estimate in new data. This adjusted (approximately unbiased) estimate of the model's expected R^2_{CS} in new individuals. This adjustment is suggested by Mittlboeck and Heinzl,

$$R^2_{CS_{adj}} = S_{VH} R^2_{CS_{app}}$$

Therefore, we minimized overfiiting on relative scale by criterion (i), and the sample size to satisfy this criterion is

$$S_{VH} = 1 + \frac{p}{nln(1 - \frac{R^2_{CS\_adj}}{S_{VH}})}$$

- Criterion (ii) ensuring a small **absolute difference d** in the apparent and adjusted Nagelkerke R^2 (or a smaller value if better performance of prediction model is required).

Nagelkerke's $R^2_{Nagelkerke}$ is a widely-used measurement for the performance of non-linear model (ie. logistic model and time-to-event model). It is a correction form of Cox-Snell R^2_{CS} to ensure the range from 0 to 1.

Therefore, ensuring small difference between apparent and adjusted Nagelkerke's R^2, which is adjusted for the optimism, would be able to minimize overfitting on absolute scale.

$$R^2_{Nagelkerkes\_app} - R^2_{Nagelkerkes\_adj} \leq \delta$$

2. Precise estimation of outcome proportion under the situation of univariate comparison tests.

- Criterion (iii) ensure precise estimation of the overall risk in the population (margin of error would be 0.05 or smaller).

In normal case of univariate comparison tests, where different covariates are not taken into consideration, we would usually set a benchmark for **margin of error** δ and get the corresponding sample size.

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

Different criteria ensure the performance of prediction model on new objects from different aspects, thus sample size calculation requires different inputs and have different outputs with respect to these three criteria.

4 Steps by three criteria method

Three-criteria method requires many inputs, which show not only the information of the development dataset but also some anticipated values or desired values (i.e. shrinkage factor S_{VH}, pseudo R^2). The choice of those inputs have great influence on the output of required sample size. In order to have a better guide of our inputs, the sensitivity of those inputs on calculated sample size is illustrated below.

4.1 *Inputs based on the characteristics of dataset*

1. **p: number of predictors.**
2. **E: number of events.**
3. **n: number of patients**

4.2 *Inputs based on our expectations*

4. **S: desired shrinkage factor.**

In an attempt to minimize overfitting, seldom do we choose a shrinkage factor smaller than 0.9. The larger expected *shrinkage factor S* is, the lager calculated sample size would be and the less predictors would be allowed in the prediction model. Influence from shrinkage factor on calculated sample size is more obvious when the **Cox-Snell** R^2 (or the **signal-to-noise ratio**) is small.

From the closed form 'heuristic' shrinkage factor of Van Houselingen and Le Cessie, shrinkage factor is defined by

$$S_{VH} = 1 - \frac{p}{LR}. \tag{2}$$

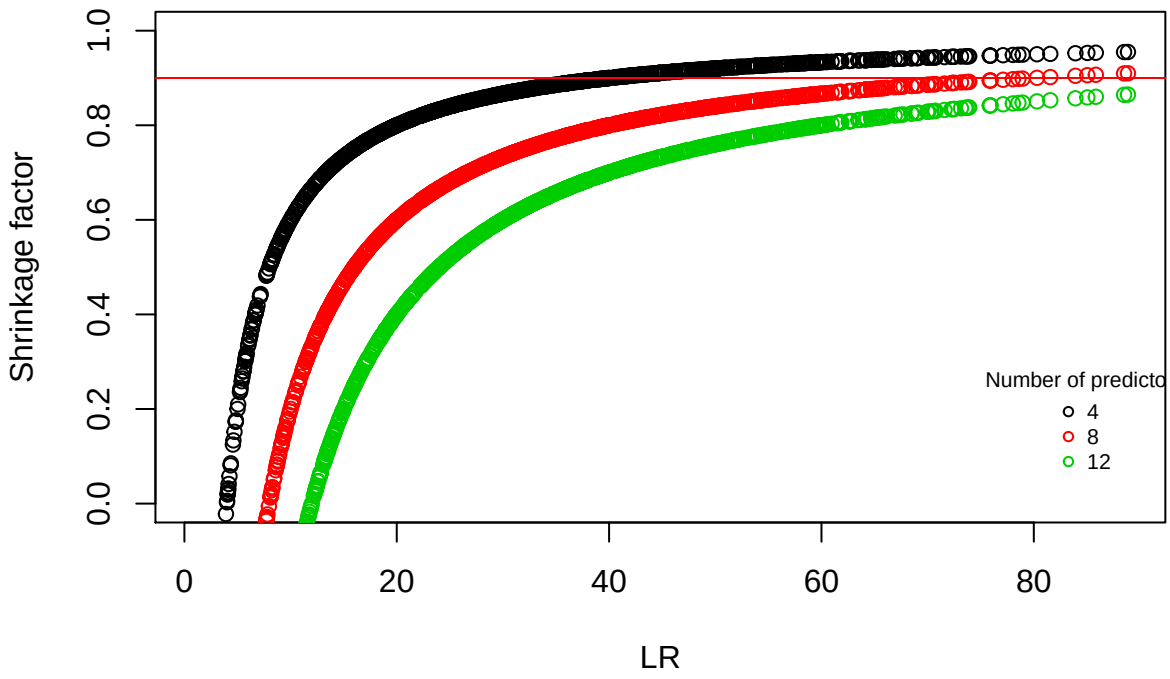


Figure 1: Shrinkage factor and likelihood ratio

5. $R^2_{CS\_adj}$**: estimated Cox-Snell** R^2 **of our model on new individuals.**

Cox-Snell R^2 is a generalization of the general R^2 in linear regression. It is a measure of *signal-to-noise ratio* in non-linear regression model.

$$R^2 = 1 - (\frac{L_{M\_Intercept}}{L_{M\_Full}})^{\frac{2}{N}}. \tag{3}$$

We could also express R^2_{CS} by *LR statistic*,

$$R^2_{CS} = 1 - exp(\frac{LR}{n}), \tag{4}$$

Equation 4 reflects the improvement of estimate of the likelihood of each Y value between the model with only intercept and all the candidate covariates. It is easy to conclude some properties of R^2_{CS} just by its definition. In an attempt to make it clearer, this document would show it by simulation results or some figures.

- i) Properties of R^2_{CS}

1. Largest value of R^2_{CS} is below 1.

It is not difficult to come to the smaller-than-one upper bound by equation 4. Besides, it is clear that the value of maximum value depended only on the proportion of outcome, the largest maximum value is about 0.75, occurs when the outcome proportion is 0.5.

Therefore, values of R^2_{CS} are not comparable when the dataset are not the same (specifically, when events fraction is not the same). Low values of R^2_{CS} do not necessarily indicate poor performance of prediction model.

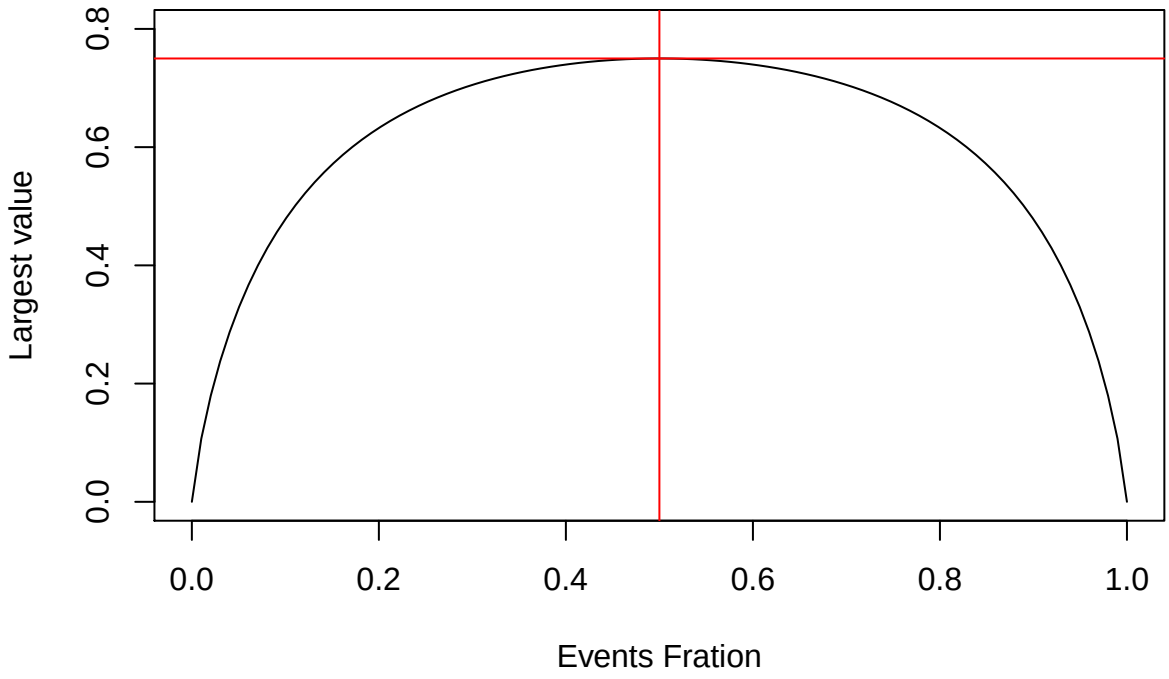


Figure 2: Maximum value of Cox-Snell R-squared

2. Anticipated value depends on *sample size n*.

In order to explore relationships between values of $Cox-SnellR^2$ and *sample size n*, *LR statistics* as well as *events fraction*. We did 4032 simulations based on 6 design factors.

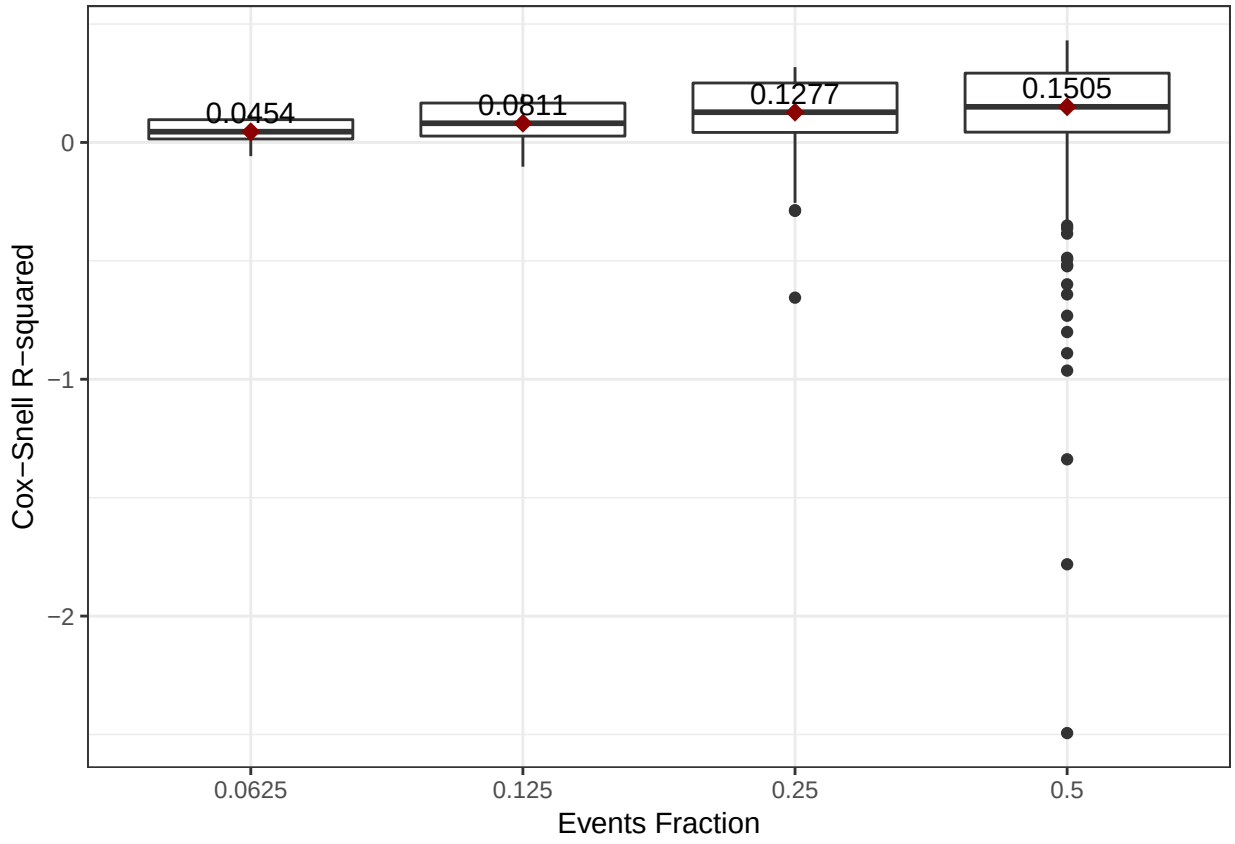


Figure 3: Cox-Snell R-squared values and events fraction (based on simulations)

Clearly, R^2_{CS} depends on *events fraction* by the simulation results, corresponding to what we have discussed above. This proporty tells us that the value of R^2_{CS} is decided by not only *predictive ability* of the developed prediction model, but also the *events fraction* of the available dataset. One model with same R^2_{CS} compared with another one, could perfrom better just because of a larger events fraction.

When coming back to the relation between the value of $R^2_{CS_{adj}}$ and *required sample size*:

Larger anticipated adjusted R^2_{CS} (denoting higher *SNR*) would require smaller sample size, since the model performance is good enough.

Shrinkage factor also has an influence on the calculation of sample size. For the same anticipated $R^2_{CS_{adj}}$, larger value of *shrinkage factor* denotes more optimism in our model and require larger *sample size*.

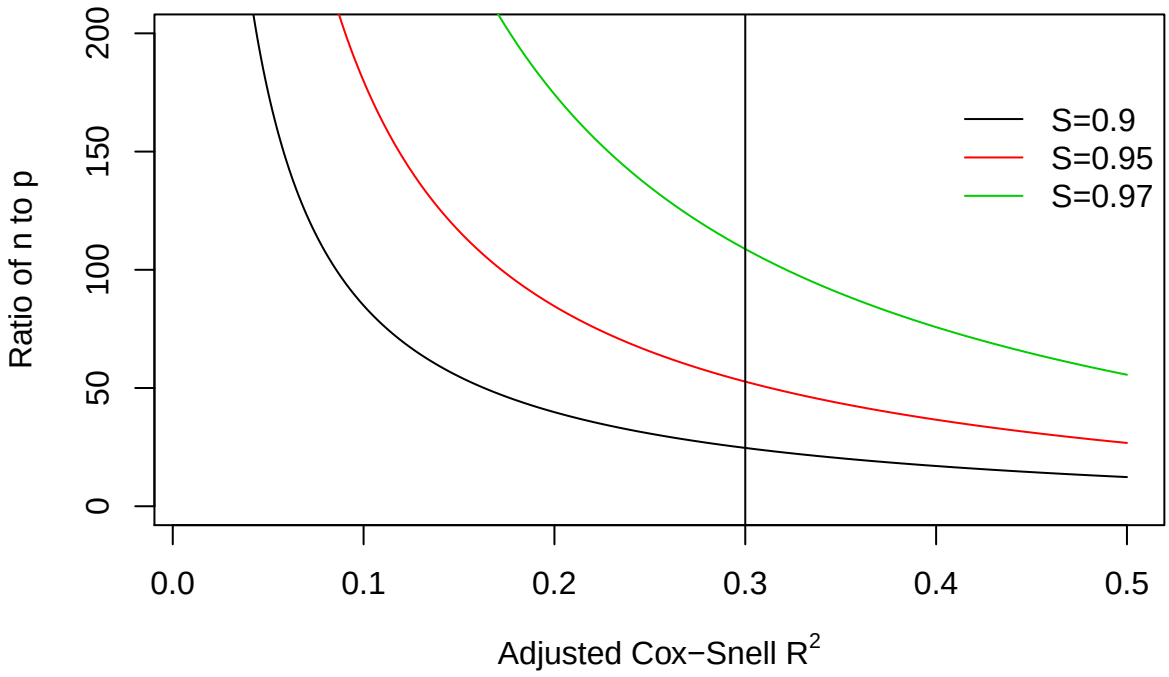
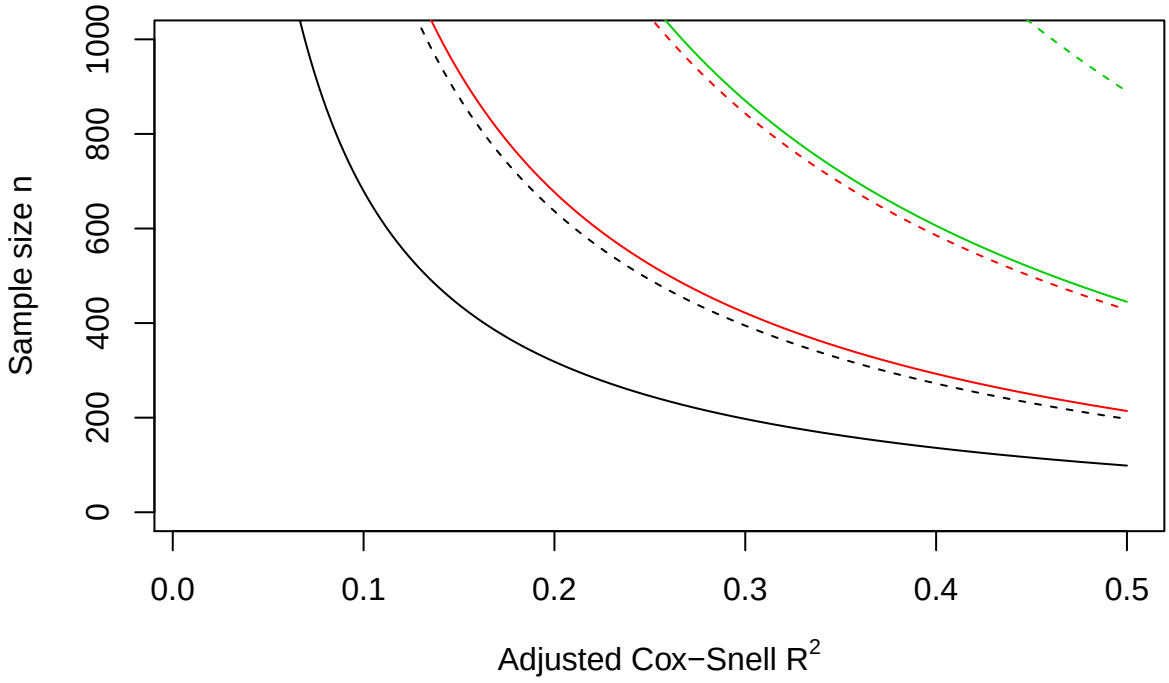


Figure 4: Sensitivity of adjusted Cox-Snell R-squared to n to p ratio



- ii) Prespecify R^2_{CS}

From discussion above, we could see that these criteria require some anticipated values (i.e. $R^2_{CS_{adj}}$) to ensure model performance on independent samples. The choice of those anticipated seems to be important for our final calculation.

a). With prior information

- Studies with same or similar population and proportion of outcomes, we use *LR statistic*, other *pseudo-*R^2 or *C statistic* to calculate R^2_{CS}. Since pseudo-R^2 use *LR statistic* as an estimation method, we could use LR related statistics, or other pseudo-R^2 to calculate Cox-Snell R^2 for sample size calculation.

b). Without prior information

- Borrow form other kinds of studies (like predictor finding studies) to get *C statistics* or *pseudo* R^2. In such case, we could assume that our model is similar to the model in other predictor finding studies, thus getting reliable statistics from those models.
- "Rule of thumb": Medical diagnosis and prediction of health-related outcomes are, generally speaking, low signal-to-noise ratio situations. Therefore, the anticipated $R^2_{CS\_adj}$ is always smaller than 0.3.

1) assume that $R^2_{Nagelkerke} = 0.15$ as in general medical studies are always low signal-to-noise ratio situations.

2) Specially, $R^2_{Nagelkerke} = 0.5$ is appropriate when predictors include directly measurements, or direct process measures is involved.

6. **d: absolute difference δ between apparent and adjusted Nagelkerke R^2 .**

$NagelkerkerR^2$ is another *Pseudo* R^2, a correction of R^2_{CS},

$$R^2_{Nagelkerke} = \frac{R^2_{CS}}{max(R^2_{CS})} = \frac{(1 - exp(-\frac{LR}{n}))}{(1 - exp(- - 2\frac{LL\_0}{n}))} \tag{5}$$

We could also explore relationships between $NagelkerkerR^2$ and and *sample size n*, *LR statistics* as well as *events fraction* based on the results from 4032 simulations mentioned above.

It is not strange that properties of $R^2_{Nagelkerke}$ are more or less similar with R^2_{CS}, and $R^2_{Nagelkerke}$ are always larger than R^2_{CS} (when R^2_{CS} is positive). However, from equation 5, it is clear that, unlike R^2_{CS}, $R^2_{Nagelkerke}$ is related to *events fraction* (intercept of model). To make it clearer, we could focus on sensitivity of *events fraction* of these 4032 simulations.

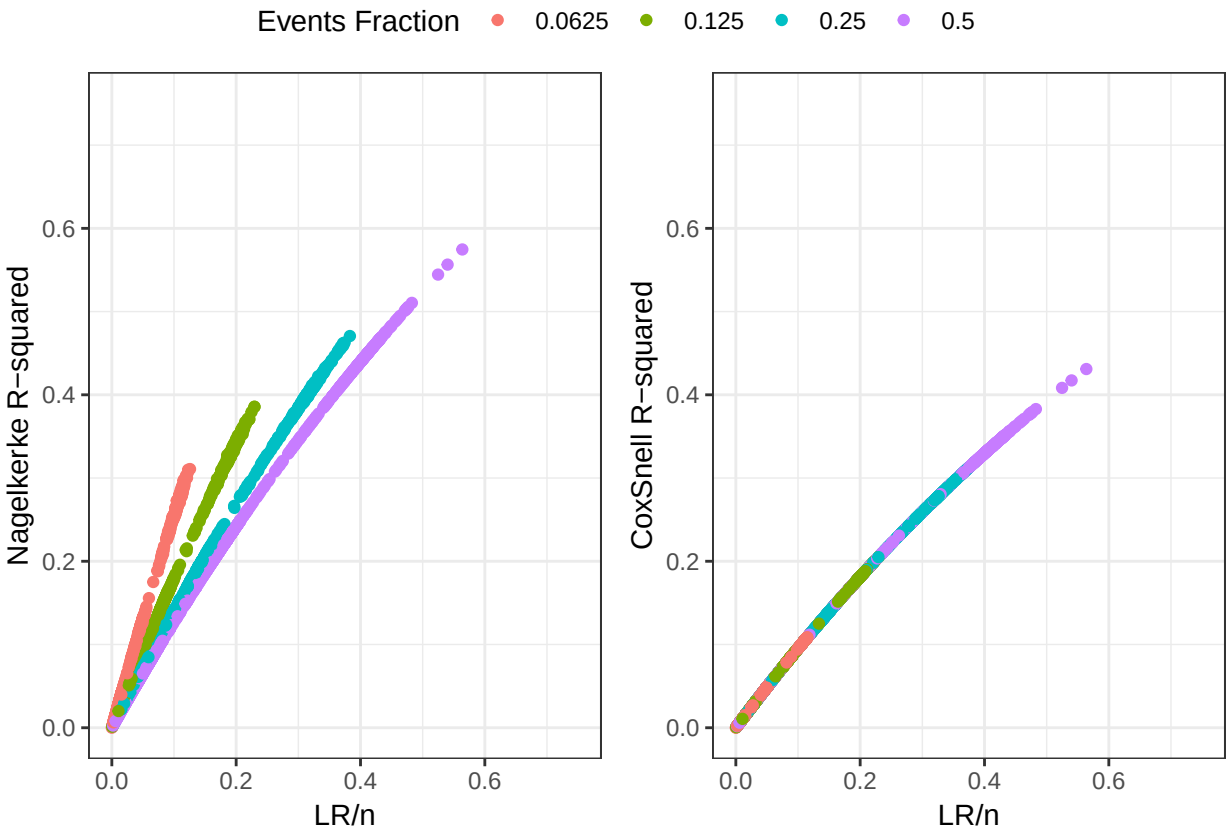


Figure 5: Pseudo R-squared w.r.t. events fraction

Therefore, when compared with R^2_{CS}:

a) The value of Nagelkerke $R^2_{Nagelkerke}$ ranges from 0 to 1;

b) It is not surprising that the value of $R^2_{Nagelkerke}$ is larger than that of R^2_{CS};

c) values $R^2_{Nagelkerke}$ is related to *events fraction*.

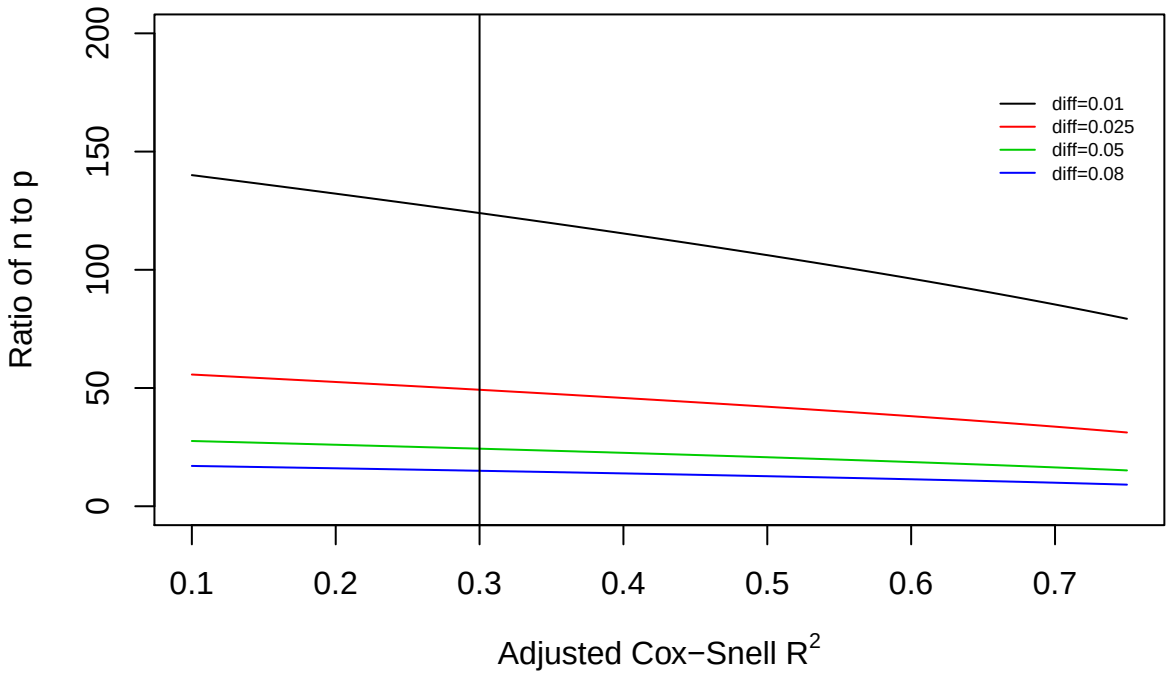


Figure 6: Ratio of n to p with respect to expected Cox-Snell R-squared S=0.9

Relatively, the "benchmark" would have slight influence the calculation of sample size (or n to p ratio). Smaller absolute difference requires larger required sample size, as overfitting would be more of a problem.

7. **error_mar: margin of error, indicating the anticipated precision of our estimation.**

We could chose δ to ensure the margin of error ≤ 0.1, or more stringer margin of error ≤ 0.05.

After values of those inputs being fixed, we would calculate three minimized sample size with respect to each criterion. The largest one would be the one we want.

5 *Data reduction*

If the calculated sample size is too large to achieve, we use variable selection method to reduce the number of predictors p. Ideally, we would like to choose the technique blind to the estimated predictor effects so as to avoid the selection bias caused by "data-drive process" (ie. principle component analysis PCA).

After variables selection, we use the reduced number of predictors to calculate the sample size by all of the criteria again (Note that data reduction techniques only work when criterion (iii) works).

Notes: simulation settings metioned in this document

The simulation settings (4032) of the following plot are:

a) *Events per variable* (7)

b) *Events fraction* (4)

c) *Number of candidate predictors (P)* (3)

d) *Model discrimination (AUC)* (4)

e) *Distribution of predictor variables* (3)

f) *Predictor effects* (4)