**A note on the overfitting for the predicted values**

**By Guo-Bo Chen, 2020/06/24**

Assuming is the phenotype, ; is the standardized genotypes for the locus, and , and heritability is zero. There are individuals, and each individual has genotypes.

The single-marker regression coefficient is estimated .

Assuming the correlation between and is , and can be decomposed as , in which is a variable from standard normal distribution.

Then

It should be noticed that

As , and given the sample size , the sampling variance of , and similar to .

Eventually, .

The variance of the predicted value is

Summary of the prediction accuracy under the null distribution

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dataset | | |
|  | Training | Test | Mixed |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Notes: is the number of markers, and is the sample size of the training set, and is the proportion of the samples in the test set but eventually from the training set.

is the NCP for .

For the test set, , and its 95% confidence interval is .

For the mixed set,

Given , the accuracy of prediction when testing containing samples from training.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | (theoretical) | (simulation) |
| 1000 | 500 | 100 | 0.1 | 0.001 | 0.001 0.0044 |
|  |  |  | 0.25 | 0.006 | 0.006 0.020 |
|  |  |  | 0.5 | 0.024 | 0.024 0.014 |
|  |  |  |  |  |  |
| 1000 | 500 | 1000 | 0.1 | 0.009 | 0.0089 0.0096 |
|  |  |  | 0.25 | 0.05 | 0.051 0.019 |
|  |  |  | 0.5 | 0.167 | 0.168 0.03 |
|  |  |  |  |  |  |
| 2000 | 500 | 100 | 0.1 | 0.0005 | 0.00034 0.0034 |
|  |  |  | 0.25 | 0.0031 | 0.0032 0.0062 |
|  |  |  | 0.5 | 0.012 | 0.012 0.0088 |
|  |  |  |  |  |  |
| 1000 | 500 | 5000 | 0.1 | 0.033 | 0.035 0.018 |
|  |  |  | 0.25 | 0.139 | 0.134 0.023 |
|  |  |  | 0.5 | 0.357 | 0.352 0.034 |

**When**

and

For , assume all the markers are independent, , and .

, , , and

So, .

In total, .

**Summary of the prediction accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dataset | | |
|  | Training | Test | Mixed |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Notes: is the number of markers, and is the sample size of the training set, and is the proportion of the samples in the test set but eventually from the training set.

, under the null hypothesis of no mixed samples.

The p-value

The statistical power given type I error rate of is .

Given , the accuracy of prediction when testing containing samples from training.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | (theoretical) | (simulation) |
| 1000 | 500 | 100 | 0.1 | 0.427 | 0.432 0.047 |
|  |  |  | 0.25 | 0.439 | 0.448 0.046 |
|  |  |  | 0.5 | 0.462 | 0.456 0.048 |
|  |  |  |  |  |  |
| 1000 | 500 | 1000 | 0.1 | 0.212 | 0.210 0.035 |
|  |  |  | 0.25 | 0.281 | 0.278 0.035 |
|  |  |  | 0.5 | 0.40 | 0.401 0.034 |
|  |  |  |  |  |  |
| 2000 | 500 | 100 | 0.1 | 0.459 | 0.455 0.044 |
|  |  |  | 0.25 | 0.466 | 0.461 0.048 |
|  |  |  | 0.5 | 0.478 | 0.478 0.052 |
|  |  |  |  |  |  |
| 1000 | 500 | 5000 | 0.1 | 0.118 | 0.118 0.026 |
|  |  |  | 0.25 | 0.236 | 0.238 0.030 |
|  |  |  | 0.5 | 0.439 | 0.439 0.033 |

Case-control design

For case-control studies, the accuracy can be measure by , in which is the difference between the mean of the risk scores between the cases and controls, and are sampling variance for risk scores for the cases and controls, respectively.

For a case-control study, which has cases and controls, the odds ratio of a locus can be estimated as

in which and are the frequency of the reference allele in the cases and controls, , and . , and .

In prediction, is used. When is close to 1, . So we have . The sampling variance for

can be calculated as below if the training and the testing are the same data.

, but for a null locus, , ., which is determined by the number of loci and the numbers of the cases and the controls.

As the real genetic effect of each locus is zero, the estimated effect is due to sampling variance..

in which .

For null model, if the testing set is independent from the training set,

in which is the frequency in the testing set.

The z score test for the different between two risk scores are , and , in which and are the numbers of cases and controls in the testing set.

For significant test, the p-value is , in which .

Inflation of AUC under the null for case-control study

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data |  |  |  |  |
| Training |  |  |  |  |
| Testing | 0 |  | 0.5 |  |

MAF=0.5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | AUC (theory) | AUC (simulation)  MAF=0.5 | AUC (simulation)  MAF=0.05~0.5 |
| 100 | 100 | 100 | 1 | 0.841 |  |  |
| 100 | 200 | 100 | 0.75 | 0.807 |  |  |
| 200 | 100 | 100 | 0.75 | 0.807 |  |  |
|  |  |  |  |  |  |  |
| 1000 | 1000 | 100 | 0.1 | 0.624 |  |  |
| 1000 | 1000 | 1000 | 1 | 0.841 |  |  |
|  |  |  |  |  |  |  |
| 1000 | 2000 | 100 | 0.075 | 0.608 |  |  |
| 1000 | 2000 | 1000 | 0.75 | 0.807 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

From GWAS meta-analysis to GWAS mega-analysis

in which

when

when

in which .

When