**A note on the sampling variance of the predicted values under the null model**

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

Assuming $y$ is the phenotype, $y \sim N(0,1)$; $x_i$ is the standardized genotypes for the $i^{th}$ locus, and $e \sim N(0,1)$, and heritability is zero. There are $N$ individuals, and each individual has $M$ genotypes.

The single-marker regression coefficient is estimated $\hat{b}_i = \frac{cov(y,x_i)}{var(x_i)} = cov(y,x_i)$.

Assuming the correlation between $x_i$ and $x_j$ is $\rho_{ij}$, and $x_j$ can be decomposed as $x_j = \rho_{ij}x_i + \sqrt{1-\rho_{ij}^2}z$, in which $z$ is a variable from standard normal distribution.

$$\hat{b}_j = \frac{cov(y,x_j)}{var(x_j)} = \frac{cov(y,\rho_{ij}x_i + \sqrt{1-\rho_{ij}^2}z)}{var(x_j)} = cov(y,\rho_{ij}x_i) + cov\left(y,\sqrt{1-\rho_{ij}^2}z\right) = \rho_{ij}\hat{b}_i + \sqrt{1-\rho_{ij}^2}\beta_j$$

Then

$$cov(x_i\hat{b}_i, x_j\hat{b}_j) = cov\left\{\hat{b}_i x_i, \left[\rho_{ij}\hat{b}_i + \sqrt{1-\rho_{ij}^2}\beta_j\right]x_j\right\} = cov(\hat{b}_i x_i, \rho_{ij}\hat{b}_i x_j) + cov\left(\hat{b}_i x_i, \sqrt{1-\rho_{ij}^2}\beta_j x_j\right)$$
$$= \rho_{ij}^2\hat{b}_i^2 + \rho_{ij}\sqrt{1-\rho_{ij}^2}\hat{b}_i\beta_j$$

It should be noticed that $E[\rho_{ij}\sqrt{1-\rho_{ij}^2}\hat{b}_i\beta_j] = 0$

As $\rho_{ij} \sim N(0,1)$, and given the sample size $N$, the sampling variance of $\rho_{ij} = \frac{1}{\sqrt{N}}$, and similar to $\hat{b}_i^2$.

Eventually, $E[cov(x_i\hat{b}_i, x_j\hat{b}_j)] = \frac{1}{N^2}$.

The variance of the predicted value $\tilde{y} = \Sigma_{i=1}^{M}\hat{b}_i x_i$ is

$$var(\tilde{y}) = \Sigma_{i=1}^{M} var(\hat{b}_i x_i) + \Sigma_{i=1}^{M}\Sigma_{j\neq i}^{M} cov(x_i\hat{b}_i, x_j\hat{b}_j) = \Sigma_{i=1}^{M}\hat{b}_i^2 + \Sigma_{i=1}^{M}\Sigma_{j\neq i}^{M}\frac{1}{N^2} = \frac{M}{N} + \frac{M(M-1)}{N^2}$$
$$\approx \frac{M}{N} + \left(\frac{M}{N}\right)^2$$

Summary of the prediction accuracy under the null distribution

| | Dataset | | |
|---|---|---|---|
| | Training | Test | Mixed |
| $cov(\hat{y},y)$ | $\frac{M}{N}$ | 0 | $w\frac{M}{N}$ |
| $var(\hat{y})$ | $\frac{M}{N}(1+\frac{M}{N})$ | $\frac{M}{N}$ | $\frac{M}{N} + w\left(\frac{M}{N}\right)^2$ |
| $R^2$ | $\frac{M}{M+N}$ | 0 | $\frac{w^2\frac{M}{N}}{1+w\frac{M}{N}}$ |

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

$\lambda = N_{tst}\frac{R^2}{1-R^2}$ is the NCP for $\chi_1^2$.

For the test set, $\lambda_T = 0$, and its 95% confidence interval is $\sqrt{\lambda_T} \pm 1.96$.

For the mixed set,

Given $h^2 = 0$, the accuracy of prediction when testing containing $w \times N_{Tst}$ samples from training.

| $N_{Tr}$ | $N_{Tst}$ | $M$ | $w$ | $R^2$ (theoretical) | $R^2$ (simulation) |
|---|---|---|---|---|---|

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

**When $h^2 \neq 0$**

$$cov\left(\sum_{m=1}^{M} \hat{b}_m x_m, y\right) = h^2 + \frac{M}{N}$$

and

$$var\left(\sum_{m=1}^{M} \hat{b}_m x_m\right) = \sum \hat{b}_m^2 \, var(x_m) + \sum_{m_1=1}^{M} \sum_{m_2 \neq m_1}^{M} \hat{b}_{m_1} \hat{b}_{m_2} cov(x_{m_1}, x_{m_2})$$

$$= [h^2 + \frac{M}{N}] + \sum_{m_1=1}^{M} \sum_{m_2 \neq m_1}^{M} [b_{m_1} b_{m_2} + b_{m_1} e_2 + b_{m_2} e_1 + e_1 e_2] cov(x_{m_1}, x_{m_2})$$

For $[b_{m_1} b_{m_2} + b_{m_1} e_2 + b_{m_2} e_1 + e_1 e_2] cov(x_{m_1}, x_{m_2})$, assume all the markers are independent, $cov(x_{m_1}, x_{m_2}) = r_{m_1 m_2}$, and $E(r_{m_1 m_2}) = \frac{1}{\sqrt{N}}$.

$[b_{m_1} b_{m_2} + b_{m_1} e_2 + b_{m_2} e_1 + e_1 e_2] cov(x_{m_1}, x_{m_2}) = [b_{m_1} b_{m_2} + b_{m_1} e_2 + b_{m_2} e_1 + e_1 e_2] r_{m_1, m_2}$

$E(b_{m_1} b_{m_2} r_{m_1, m_2}) = 0$, $E(b_{m_1} e_2 r_{m_1} r_{m_2}) = b_{m_1}^2 r_{m_1 m_2}^2 = \frac{h^2}{M} \frac{1}{N}$, $E(b_{m_2} e_1 r_{m_1} r_{m_2}) = b_{m_2}^2 r_{m_1 m_2}^2 = \frac{h^2}{M} \frac{1}{N}$, and $E(e_1 e_2 r_{m_1} r_{m_2}) = \frac{1}{N^2}$

So, $\sum_{m_1=1}^{M} \sum_{m_2 \neq m_1}^{M} [b_{m_1} b_{m_2} + b_{m_1} e_2 + b_{m_2} e_1 + e_1 e_2] cov(x_{m_1}, x_{m_2}) = M(M-1)(\frac{2h^2}{MN} + \frac{1}{N^2})$.

In total, $var(\sum_{m=1}^{M} \hat{b}_m x_m) = \left[h^2 + \frac{M}{N}\right] + M(M-1)\left(\frac{2h^2}{MN} + \frac{1}{N^2}\right) = h^2\left(1 + 2\frac{M}{N}\right) + \frac{M}{N}(1 + \frac{M}{N})$.

**Summary of the prediction accuracy**

| | Dataset | | |
|---|---|---|---|
| | Training | Test | Mixed |
| $cov(\hat{y}, y)$ | $h^2 + \frac{M}{N}$ | $h^2$ | $h^2 + w\frac{M}{N}$ |
| $var(\hat{y})$ | $h^2\left(1 + 2\frac{M}{N}\right) + \frac{M}{N}(1 + \frac{M}{N})$ | $h^2 + \frac{M}{N}$ | $wh^2\frac{M}{N} + (h^2 + \frac{M}{N})(1 + w\frac{M}{N})$ |
| $R^2$ | $\frac{\left(h^2 + \frac{M}{N}\right)^2}{h^2\left(1 + 2\frac{M}{N}\right) + \frac{M}{N}(1 + \frac{M}{N})}$ | $\frac{(h^2)^2}{h^2 + \frac{M}{N}}$ | $\frac{\left(h^2 + w\frac{M}{N}\right)^2}{wh^2\frac{M}{N} + (h^2 + \frac{M}{N})(1 + w\frac{M}{N})}$ |

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

$$\lambda = N_{tst} \frac{R^2}{1 - R^2}$$

$\lambda_0 = N_{tst} \frac{R_{tst}^2}{1 - R_{tst}^2} = \frac{N_{tst} \frac{(h^2)^2}{h^2 + \frac{M}{N}}}{1 - \frac{(h^2)^2}{h^2 + \frac{M}{N}}} = N_{tst} \frac{(h^2)^2}{h^2 + \frac{M}{N} - (h^2)^2}$, under the null hypothesis of no mixed samples.

$$\lambda_M = N_{tst} \frac{R_M^2}{1 - R_M^2}$$

$$\lambda_M = \frac{N_{tst} \frac{\left(h^2 + w\frac{M}{N}\right)^2}{wh^2\frac{M}{N} + (h^2 + \frac{M}{N})(1 + w\frac{M}{N})}}{1 - \frac{\left(h^2 + w\frac{M}{N}\right)^2}{wh^2\frac{M}{N} + (h^2 + \frac{M}{N})(1 + w\frac{M}{N})}} = \frac{N_{tst}}{\frac{wh^2\frac{M}{N} + (h^2 + \frac{M}{N})(1 + w\frac{M}{N})}{\left(h^2 + w\frac{M}{N}\right)^2} - 1} \approx wN_{tst}$$

The p-value $\chi^2_{\lambda_0}(\lambda_M)$

The statistical power given type I error rate of $\alpha$ is $1 - \phi\left(\phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \sqrt{\lambda_0} - \sqrt{\lambda_1}\right) + \phi(\phi^{-1}\left(\frac{\alpha}{2}\right) + \sqrt{\lambda_0} - \sqrt{\lambda_1})$.

Given $h^2 = 0.5$, the accuracy of prediction when testing containing $w \times N_{Tst}$ samples from training.

| $N_{Tr}$ | $N_{Tst}$ | $M$ | $w$ | $R^2$ (theoretical) | $R^2$ (simulation) |
|---|---|---|---|---|---|
| 1000 | 500 | 100 | 0.1 | 0.427 | $0.432 \pm 0.047$ |
| | | | 0.25 | 0.439 | $0.448 \pm 0.046$ |
| | | | 0.5 | 0.462 | $0.456 \pm 0.048$ |
| | | | | | |
| 1000 | 500 | 1000 | 0.1 | 0.212 | $0.210 \pm 0.035$ |
| | | | 0.25 | 0.281 | $0.278 \pm 0.035$ |
| | | | 0.5 | 0.40 | $0.401 \pm 0.034$ |
| | | | | | |
| 2000 | 500 | 100 | 0.1 | 0.459 | $0.455 \pm 0.044$ |
| | | | 0.25 | 0.466 | $0.461 \pm 0.048$ |
| | | | 0.5 | 0.478 | $0.478 \pm 0.052$ |
| | | | | | |
| 1000 | 500 | 5000 | 0.1 | 0.118 | $0.118 \pm 0.026$ |
| | | | 0.25 | 0.236 | $0.238 \pm 0.030$ |
| | | | 0.5 | 0.439 | $0.439 \pm 0.033$ |

Case-control design

For case-control studies, the accuracy can be measure by $AUC = \phi(\frac{D_s}{\sqrt{\sigma_{cs}^2+\sigma_{cl}^2}})$, in which $D_s$ is the difference between the mean of the risk scores between the cases and controls, $\sigma_{cs}^2$ and $\sigma_{cl}^2$ are sampling variance for risk scores for the cases and controls, respectively.

For a case-control study, which has $N_{cs}$ cases and $N_{cl}$ controls, the odds ratio of a locus can be estimated as

$$OR = \frac{p_{cs}}{p_{cl}}\frac{q_{cl}}{q_{cs}}$$

in which $p_{cs}$ and $p_{cl}$ are the frequency of the reference allele in the cases and controls, $q_{cs} = 1 - p_{cs}$, and $q_{cl} = 1 - p_{cl}$. $p_{cs} \sim N(p_{cs}, \frac{p_{cs}q_{cs}}{2N_{cs}})$, and $p_{cl} \sim N(p_{cl}, \frac{p_{cl}q_{cl}}{2N_{cl}})$.

In prediction, $\beta = \log_e(OR)$ is used. When $OR$ is close to 1, $\log_e(OR) \approx OR - 1$. So we have $\beta \approx OR - 1 = \frac{p_{cs}-p_{cl}}{p_{cl}q_{cs}}$. The sampling variance for $\sigma_{\beta_i}^2 = (\frac{1}{2N_{cs}p_{cs}} + \frac{1}{2N_{cs}q_{cs}} + \frac{1}{2N_{cl}p_{cl}} + \frac{1}{2N_{cl}q_{cl}})$

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

$$D_s = \sum_{i=1}^{M} \beta_i(2p_{cs.i} - 2p_{cl.i}) = \sum_{i=1}^{M} \frac{2(p_{cs.i} - p_{cl.i})^2}{p_{cl.i}q_{cs.i}}$$

$(p_{cs} - p_{cl}) \sim N(p_{cs} - p_{cl}, \frac{p_{cs}q_{cs}}{2N_{cs}} + \frac{p_{cl}q_{cl}}{2N_{cl}})$, but for a null locus, $p_{cs} \approx p_{cl}$, $(p_{cs} - p_{cl}) \sim N(0, p_{cs}q_{cs}\frac{N_{cl}+N_{cs}}{2N_{cs}N_{cl}})$. $D_s \approx M\frac{N_{cl}+N_{cs}}{N_{cs}N_{cl}}$, 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.$var(\hat{\beta}_i x_i) = var(e_i x_i)$.

$$\sigma_{cs}^2 = \sum_{i=1}^{M} var(x_i e_i) = \sum_{i=1}^{M} 2p_{cs.i}q_{cs.i}var(e) = \sum_{i=1}^{M} 2p_{cs.i}q_{cs.i}\left(\frac{1}{2N_{cs}p_{cs.i}} + \frac{1}{2N_{cs}q_{cs.i}} + \frac{1}{2N_{cl}p_{cl.i}} + \frac{1}{2N_{cl}q_{cl.i}}\right)$$
$$= M\frac{N_{cl}+N_{cs}}{N_{cs}N_{cl}}$$

$$AUC = \phi\left(\frac{D_s}{\sqrt{\sigma_{cs}^2+\sigma_{cl}^2}}\right) = \phi(T)$$

in which $T = \sqrt{\frac{M}{2}\frac{N_{cl}+N_{cs}}{N_{cs}N_{cl}}}$.

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

$$D_s = \sum_{i=1}^{M} \beta_i(2\tilde{p}_{cs.i} - 2\tilde{p}_{cl.i}) = 0$$

in which $\tilde{p}$ is the frequency in the testing set.

The z score test for the different between two risk scores are $\frac{D_s}{\sqrt{\frac{\sigma_{cs}^2}{\tilde{N}_{cs}}+\frac{\sigma_{cl}^2}{\tilde{N}_{cl}}}} \approx T\sqrt{\frac{2\tilde{N}_{cs}\tilde{N}_{cl}}{\tilde{N}_{cl}+\tilde{N}_{cs}}}$, and $T \sim N(0, \sqrt{\frac{\tilde{N}_{cl}+\tilde{N}_{cs}}{2\tilde{N}_{cs}\tilde{N}_{cl}}})$, in which $\tilde{N}_{cs}$ and $\tilde{N}_{cl}$ are the numbers of cases and controls in the testing set.

For significant test, the p-value is $\chi_1^2(\lambda)$, in which $\lambda = \left(\frac{D_s}{\sqrt{\sigma_{cs}^2+\sigma_{cl}^2}}\right)^2 = M\frac{N_{cl}+N_{cs}}{2N_{cs}N_{cl}}$.

Inflation of AUC under the null for case-control study

| Data | $D_s$ | $var(s)$ | $AUC$ | |
|---|---|---|---|---|
| Training | $M\frac{N_{cl}+N_{cs}}{N_{cs}N_{cl}}$ | $M\frac{N_{cl}+N_{cs}}{N_{cs}N_{cl}}$ | $\phi(\sqrt{M\frac{N_{cl}+N_{cs}}{2N_{cs}N_{cl}}})$ | |
| Testing | 0 | $M\frac{N_{cl}+N_{cs}}{N_{cs}N_{cl}}$ | 0.5 | |

MAF=0.5

| $N_{cs}$ | $N_{cl}$ | $M$ | $\lambda$ | AUC (theory) | AUC (simulation) MAF=0.5 | AUC (simulation) MAF=0.05∼0.5 |
|---|---|---|---|---|---|---|
| 100 | 100 | 100 | 1 | 0.841 | $0.843 \pm 0.018$ | $0.838 \pm 0.019$ |
| 100 | 200 | 100 | 0.75 | 0.807 | $0.805 \pm 0.020$ | $0.808 \pm 0.020$ |
| 200 | 100 | 100 | 0.75 | 0.807 | $0.808 \pm 0.019$ | $0.809 \pm 0.023$ |
| | | | | | | |
| 1000 | 1000 | 100 | 0.1 | 0.624 | $0.625 \pm 0.0092$ | $0.623 \pm 0.0099$ |
| 1000 | 1000 | 1000 | 1 | 0.841 | $0.842 \pm 0.0064$ | $0.842 \pm 0.0078$ |
| | | | | | | |
| 1000 | 2000 | 100 | 0.075 | 0.608 | $0.608 \pm 0.0070$ | $0.608 \pm 0.0078$ |
| 1000 | 2000 | 1000 | 0.75 | 0.807 | $0.807 \pm 0.0064$ | $0.807 \pm 0.0058$ |
| | | | | | | |
| | | | | | | |

From GWAS meta-analysis to GWAS mega-analysis

$$\beta_{meta} = \frac{\sum_{i=1}^{C} \beta_i W_i}{\sum_{i=1}^{C} W_i}$$

in which $W_i = \frac{1}{\sigma^2_{\beta_i}} = n_i \frac{var(x_i)}{var(y_i)}$

$$\beta_{meta} = \frac{\sum_{i=1}^{C} \frac{cov(x_i, y_i)}{var(x_i)} n_i \frac{var(x_i)}{var(y_i)}}{\sum_{i=1}^{C} n_i \frac{var(x_i)}{var(y_i)_i}} = \frac{\sum_{i=1}^{C} n_i \frac{cov(x_i, y_i)}{var(y_i)}}{\sum_{i=1}^{C} n_i \frac{var(x_i)}{var(y_i)}}$$

when $var(y_i) = var(y_j)$

$$\beta_{meta} = \frac{\sum_{i=1}^{C} n_i cov(x_i, y_i)}{\sum_{i=1}^{C} n_i var(x_i)}$$

when $var(x_i) = var(x_j)$

$$\beta_{meta} = \frac{\sum_{i=1}^{C} n_i cov(x_i, y_i)}{n var(x_i)} = \frac{\sum_{i=1}^{C} \omega_i cov(x_i, y_i)}{var(x)}$$

in which $\omega_i = \frac{n_i}{n}$.

When $cov(x_i, y_i) = cov(x_j, y_j)$

$$\beta_{meta} = \frac{cov(x, y)}{var(x)} = \beta_{mega}$$