Expectation of R^2 with stratified phenotypes and genotypes

September 3, 2015

Let Y_i, $i = 1, \ldots, n$ be the stratification-adjusted phenotype centered around 0 and X_i, $i = 1, \ldots, n$ be the stratification-adjusted genotype centered around 0.

We have

$$\begin{aligned}
E[X_i] &= 0 \forall i \\
E[Y_i] &= 0 \forall i \\
Var[X_i] &= \sigma^2_{X,i} \\
Var[X_i] &= \sigma^2_{Y,i}
\end{aligned}$$

The Pearson correlation coefficient is

$$\begin{aligned}
\hat{\rho} &= \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}} \\
\hat{\rho} &= \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n (X_i)^2}\sqrt{\sum_{i=1}^n (Y_i)^2}} \\
R^2 &= \frac{\left(\sum_{i=1}^n X_i Y_i\right)^2}{\sum_{i=1}^n (X_i)^2 \sum_{i=1}^n (Y_i)^2} \\
&= \frac{\left(\sum_{i=1}^n X_i Y_i\right)^2}{\sum_{i=1}^n Var(X_i) \sum_{i=1}^n Var(Y_i)} \\
E\left(R^2\right) &= E\left[\frac{\left(\sum_{i=1}^n X_i Y_i\right)^2}{\sum_{i=1}^n \sigma^2_{X,i} \sum_{i=1}^n \sigma^2_{Y,i}}\right] \\
&= \frac{E\left[\left(\sum_{i=1}^n X_i Y_i\right)^2\right]}{\sum_{i=1}^n \sigma^2_{X,i} \sum_{i=1}^n \sigma^2_{Y,i}} \\
&= \frac{Var\left[\sum_{i=1}^n X_i Y_i\right] + E\left[\sum_{i=1}^n X_i Y_i\right]^2}{\sum_{i=1}^n \sigma^2_{X,i} \sum_{i=1}^n \sigma^2_{Y,i}} \\
&= \frac{Var\left[\sum_{i=1}^n X_i Y_i\right]}{\sum_{i=1}^n \sigma^2_{X,i} \sum_{i=1}^n \sigma^2_{Y,i}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{n} Var\left(X_i Y_i\right)}{\sum_{i=1}^{n} \sigma_{X,i}^2 \sum_{i=1}^{n} \sigma_{Y,i}^2} \\
&= \frac{\sum_{i=1}^{n} \left[E\left(X_i^2\right) E\left(Y_i^2\right) - E\left(X_i\right)^2 E\left(Y_i\right)^2\right]}{\sum_{i=1}^{n} \sigma_{X,i}^2 \sum_{i=1}^{n} \sigma_{Y,i}^2} \\
&= \frac{\sum_{i=1}^{n} \left[E\left(X_i^2\right) E\left(Y_i^2\right)\right]}{\sum_{i=1}^{n} \sigma_{X,i}^2 \sum_{i=1}^{n} \sigma_{Y,i}^2} \\
&= \frac{\sum_{i=1}^{n} \left[Var\left(X_i\right) Var\left(Y_i\right)\right]}{\sum_{i=1}^{n} \sigma_{X,i}^2 \sum_{i=1}^{n} \sigma_{Y,i}^2} \\
&= \frac{\sum_{i=1}^{n} \left[\sigma_{X,i}^2 \sigma_{Y,i}^2\right]}{\sum_{i=1}^{n} \sigma_{X,i}^2 \sum_{i=1}^{n} \sigma_{Y,i}^2}
\end{aligned}
$$

For a binary phenotype and a diploid genotype we can express this as

$$
E\left(R^2\right) = \frac{\sum_{i=1}^{n} \left[2\hat{p}_{X,i}\left(1-\hat{p}_{X,i}\right)\hat{p}_{Y,i}\left(1-\hat{p}_{Y,i}\right)\right]}{\sum_{i=1}^{n} 2\hat{p}_{X,i}\left(1-\hat{p}_{X,i}\right)\sum_{i=1}^{n} \hat{p}_{Y,i}\left(1-\hat{p}_{Y,i}\right)}
$$