Supporting information

Approximating the relationship between regression coefficients and relative risks for continuous and discretised characteristics

The following approximation is used to relate a genotype's regression coefficient, β, to the log relative risk associating the genotype to a threshold characteristic defined by *t*:

$$\ln \mathrm{RR} \approx \frac{\beta}{(1-p)s}\left(0.8+0.68\frac{t}{s}+0.064\frac{t^2}{s^2}\right)-\frac{\beta^2}{(1-p)^2 s^2}\left(0.34+0.064\frac{t}{s}-0.0128\frac{t^2}{s^2}\right).$$

This is derived as follows. The relative risk is defined in this case as

$$\mathrm{RR}=\frac{\Pr(Y>t\,|\,g=1)}{\Pr(Y>t\,|\,g=0)}\ , \qquad \text{(S1)}$$

where *Y* is the continuous phenotype and *g* is the genotype status (1 if the individual possesses the 'risk' genotype and 0 otherwise). For simplicity we assume that *g*=1 implies aa and *g*=0 implies ab/bb (or vice versa), or *g*=1 implies ab and *g*=0 implies aa/bb (over or underdominance).

We assume that *Y* is normally distributed with variance 1 and mean 0. If *g*=1, *Y* is normal with mean μ_1, and if *g*=0, *Y* has mean $u_0=-u_1\Pr(g=1)/\left(1-\Pr(g=1)\right)$. Each distribution has variance $s^2=1-\mathrm{var}(g)(\mu_1-\mu_0)^2$. It can be shown that the regression coefficient β, also known as the average effect, is related to μ_1 via $\mu_1=\beta(1-p)$, where *p* is the allele frequency, when *g*=1 is dominant. Similar relationships can be obtained under recessive and overdominant inheritance models. We proceed using the dominant model for now (*g*=1 implies aa or ab and *g*=0 implies bb), which is also a good approximation of the additive model when the dominant allele is not too common. We next produce a second order Taylor series approximation for the relative risk. The following is made easier if we analyse $\tilde{Y}=(Y-\mu_0)/s$ in place of *Y* by making the substitutions $\tilde{t}=(t-\mu_0)/s$, $\tilde{\mu}_1=(\mu_1-\mu_0)/s$, and $\tilde{\mu}_0=(\mu_0-\mu_0)/s=0$. This forces the denominator in Equation S1 to be independent from $\tilde{\mu}_1$, and it would otherwise depend on μ_1 through *s*.

www.pnas.org/cgi/doi/10.1073/pnas.2005634117

The first derivative of the log relative risk with respect to $\tilde{\mu}_1$ is then

$$f_t'(\tilde{\mu}_1) = \frac{\phi(\tilde{t}-\tilde{\mu}_1)}{1-\Phi(\tilde{t}-\tilde{\mu}_1)} \ ,$$

where $\phi(x)$ is the standard normal density function, $\Phi(x)$ the standard cumulative normal distribution and $f_t(\tilde{\mu}_1) = \ln \mathrm{RR}$ (Equation S1).

The following approximation to the cumulative normal distribution by Hart[1],

$$\Phi(x) = 1 - \frac{\phi(x)}{x + 0.8e^{-0.4x}} \ ,$$

then allows one to write

$$f_t'(\tilde{\mu}_1) \approx \tilde{t} - \tilde{\mu}_1 + 0.8e^{-0.4(\tilde{t}-\tilde{\mu}_1)} .$$

Differentiating a second time gives

$$f_t''(\tilde{\mu}_1) \approx 0.32e^{-0.4(\tilde{t}-\tilde{\mu}_1)} - 1 .$$

Combining the derivatives into a Taylor series for $\beta \approx 0$ gives

$$\ln \mathrm{RR} \approx (0.8\tilde{\mu}_1 + 0.16\tilde{\mu}_1^2)e^{-0.4\tilde{t}} + \tilde{t}\tilde{\mu}_1 - 0.5\tilde{\mu}_1^2 .$$

Further simplification is achieved via second order Taylor approximation of $e^{-0.4\tilde{t}}$ with $\tilde{t}$ in the neighborhood of zero:

$$e^{-0.4\tilde{t}} \approx 1 - 0.4\tilde{t} + 0.08\tilde{t}^2 ,$$

giving

$$\ln \mathrm{RR} \approx \tilde{\mu}_1(0.8 + 0.68\tilde{t} + 0.064\tilde{t}^2) - \tilde{\mu}_1^2(0.34 + 0.064\tilde{t} - 0.0128\tilde{t}^2) .$$

Converting β to $\tilde{\mu}_1 = (\beta(1-p) - \mu_0)/s$, which equals $\beta(1-p)\left(1 - \frac{\Pr(g=1)}{\Pr(g=1)-1}\right)/s$, under the dominance model gives

$$\ln \mathrm{RR} \approx \frac{\beta}{(1-p)s}(0.8+0.68\tilde{t}+0.064\tilde{t}^2)-\frac{\beta^2}{(1-p)^2 s^2}(0.34+0.064\tilde{t}-0.0128\tilde{t}^2),$$

as $\Pr(g=1)=1-(1-p)^2$. In practice, using *t/s* in place of $\tilde{t}$ maintains reasonable accuracy, as shown in Figure 2 and Equation 1.

The similarity of this approximation to results obtained using distribution functions, when $\beta>0$, is demonstrated in Figure 2. The approximation is invalid when *t* becomes moderately large and negative, but in this case the relative risk can be inverted.

References

1 Hart, R. G. A formula for the approximation of definite integrals of the normal distribution function. *Math. Comp.* **11**, 265 (1957).