Trait value conditional on polygenic score

Alex Bloemendal, Patrick Turley, Eric Lander

April 5, 2019

What is the distribution of an individual's trait value conditional on their value of a polygenic score for the trait? We consider the canonical model of a quantitative trait and then a binary trait.

1 Quantitative trait

Consider a complex quantitative trait following a standard normal distribution $y \sim \mathcal{N}(0,1)$. Suppose we have a standardized polygenic score $z \sim \mathcal{N}(0,1)$ that predicts the trait, and assume further that y, z are jointly bivariate normal with correlation $R = \mathrm{Corr}(y,z) = \mathbb{E}\, yz$, i.e. z explains R^2 of the trait variance. Then the conditional distribution of y given z is

$$y \mid z \sim \mathcal{N}(Rz, 1 - R^2). \tag{1}$$

Unsurprisingly the conditional mean is the regression of y onto z and the conditional variance is attenuated by the portion explained.

Under fairly general conditions the distributional assumptions are plausible and the behavior of R^2 is understood. Imagine y as a sum of many small independent additive effects both genetic and environmental. Suppose we measure a large number M of genomic markers x_j (e.g. SNPs), and let β_j be the true average effect of allelic substitution of x_j on y conditional on $\{x_k\}_{k \neq j}$. Let $g_M = \sum_j \beta_j x_j$ be the combined effect of these markers and $h^2_M = \mathrm{Var}(g_M)$ the proportion of trait variance they explain; they may not completely tag all causal variants, so in general $h^2_M \leq h^2$, the narrow sense heritability.

Take z to be the standardized version of a linear predictor $\hat{g}_M = \sum_j \hat{\beta}_j x_j$, where the coefficients are fit by some method on a discovery sample. The errors $\hat{\beta}_j - \beta_j$ will be independent of a new target sample. Independence and the central limit theorem motivate the joint normality of z and y. Furthermore, the prediction error $\varepsilon = \hat{g}_M - g_M$ will be uncorrelated with y, so

$$R^2 \;=\; (\mathbb{E}\, yz)^2 \;=\; \frac{(\mathbb{E}\, y\hat{g}_M)^2}{\mathrm{Var}(\hat{g}_M)} \;=\; \frac{(\mathbb{E}\, g_M^2)}{\mathrm{Var}(g_M) + \mathrm{Var}(\varepsilon)} \;=\; \frac{h_M^2}{1 + \frac{\sigma_\varepsilon^2}{h_M^2}}, \tag{2}$$

where $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon)$. In particular we have the bound $R^2 \leq h_M^2$, saturated in the large discovery sample limit of perfect effect size estimation.

Daetwyler et al. (2008), see also Lee and Wray (2013), Lee et al. (2017), effectively approximate the error variance as $\sigma_\varepsilon^2 \approx M_e/N_d$, where N_d is the discovery sample size and M_e is the effective number of independent markers or chromosome segments. The latter can be estimated from a genotype-derived GRM or LD scores [refs]. With a fitting method based on an infinitesimal model for the genetic architecture (i.e. the distribution of β_j), e.g. BLUP or LDPredInf [refs], this result does not depend on the actual genetic architecture, but in some circumstances a good prior can be leveraged to do better (Wray et al. 2013). A more robust approach is therefore to simply assume the behavior $\sigma_\varepsilon^2 \sim C/N_d$ and calibrate C to one or more observations of R^2, recognizing that C may depend on both the trait and the method [refs?].

2 Binary trait

Consider now a complex binary trait $s \sim \mathrm{Bernoulli}(K)$, where K is the population prevalence. Under the liability threshold model we have

$$s = \mathbf{1}_{y>T}$$

where $y \sim \mathcal{N}(0,1)$ is a latent liability and T is a threshold. The threshold is related to the prevalence by

$$K \;=\; \mathbb{E}\, s \;=\; \mathbb{P}(y > T) \;=\; 1 - \Phi(T),$$

where Φ the cumulative distribution function of the standard normal.

Suppose as before that $z \sim \mathcal{N}(0,1)$ is a standardized polygenic score for the trait and that y, z are jointly bivariate normal with $R = \mathrm{Corr}(y, z)$. Using (1), we compute the conditional prevalence K_z of the trait given the score z as

$$K_z \;=\; \mathbb{E}\, s|z \;=\; \mathbb{P}(y > T \mid z) \;=\; 1 - \Phi\left(\frac{T - Rz}{\sqrt{1-R^2}}\right).$$

Using the symmetry $\Phi(-T) = 1 - \Phi(T)$, we obtain

$$\Phi^{-1}(K_z) \;=\; \frac{Rz + \Phi^{-1}(K)}{\sqrt{1-R^2}}. \tag{3}$$

The normal quantile function Φ^{-1} is also called the probit transformation; observe that on the probit scale, the conditional prevalence is in a simple linear relationship with the score. The slope

$$m(R^2) \;=\; \frac{R}{\sqrt{1-R^2}} \tag{4}$$

is a monotonic function of the liability variance explained. In particular, we have the bounds

$$m(R^2) \;\leq\; m(h^2_M) \;\leq\; m(h^2).$$

Again, the first bound saturates in the large discovery sample limit of perfect effect size estimation; the second saturates in the limit of complete tagging of causal variation.

Here heritability is on the liability scale, the most natural and interpretable scale for a binary trait, invariant with respect to population prevalence and case ascertainment. We argue as others have that R^2, the proportion of liability variance explained by z, is similarly the most natural and interpretable figure of merit for a binary trait polygenic score (Wray et al. 2010, Lee et al. 2012). Furthermore, the linear relationship (3) suggests a natural way to plot empirical risk gradients and probe model assumptions.

Equation (2) holds as before; Lee and Wray (2013) (see also Daetwyler et al. (2008), Lee et al. (2017)) treat the liability threshold setting as well, obtaining now $\sigma^2_\varepsilon \approx cM_e/N_d$ with c the conversion factor from heritability on the observed scale h^2_o:

$$h^2_M = ch^2_o\,, \qquad c = \frac{K^2(1-K)^2}{\varphi(T)^2 P(1-P)}$$

where P is the fraction of cases ascertained in the discovery sample and φ is the

standard normal density (Lee et al. 2011). The error variance therefore scales as

$$\sigma^2_\varepsilon \sim \frac{C}{N_d P(1-P)} = C\left(\frac{1}{N_{\text{case}}} + \frac{1}{N_{\text{control}}}\right), \tag{5}$$

where C may depend in a complicated way on the set of variants, genetic architecture, fitting method, and population prevalence; as before, however, C can be calibrated to one or more observations [elaborate on how to use observed R^2]. In this way one can extrapolate to discovery samples with any number of cases and controls.

References

Daetwyler, H. D., Villanueva, B. and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach, *PloS one* **3**(10): e3395.

Lee, S. H., Goddard, M. E., Wray, N. R. and Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis, *Genetic epidemiology* **36**(3): 214–224.

Lee, S. H., Weerasinghe, W. S. P., Wray, N. R., Goddard, M. E. and Van Der Werf, J. H. (2017). Using information of relatives in genomic prediction to apply effective stratified medicine, *Scientific reports* **7**: 42091.

Lee, S. H. and Wray, N. R. (2013). Novel genetic analysis for case-control genome-wide association studies: quantification of power and genomic prediction accuracy, *PloS one* **8**(8): e71494.

Lee, S. H., Wray, N. R., Goddard, M. E. and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies, *The American Journal of Human Genetics* **88**(3): 294–305.

Wray, N. R., Yang, J., Goddard, M. E. and Visscher, P. M. (2010). The genetic interpretation of area under the roc curve in genomic profiling, *PLoS genetics* **6**(2): e1000864.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E. and Visscher, P. M. (2013). Pitfalls of predicting complex traits from snps, *Nature Reviews Genetics* **14**(7): 507.