Supplementary Note

Multivariate statistical test

Consider a regression model:

$$Y = f(\boldsymbol{X}\beta + \boldsymbol{C}\gamma + (\boldsymbol{X} * \boldsymbol{C})\delta) + \varepsilon$$

where Y is a $N\times1$ vector of observations of either a continuous or a binary phenotype, $\boldsymbol{X}$ is a $N\times K$ genotype matrix, $\boldsymbol{C}$ is a $N\times P$ matrix of covariates and $\varepsilon\sim\mathcal{N}(0,\sigma)$ is a $N\times1$ vector of residuals. The $K\times P$ interactions between genotypes and covariates are denoted by the $\boldsymbol{X} * \boldsymbol{C}$ matrix. Genotype effects, covariates effects and interaction effects are respectively denoted by the $K\times1$vector β, the $P\times1$ vector γ and the $(K\times P)\times1$ vector δ. Depending on the nature of $\boldsymbol{Y}$, the function f is either the identity if Y is continuous, or the expit function if Y is binary.

Without loss of generality, we will focus here on testing the interaction effects δ. Several statistical tests can be performed to test the null hypothesis (H_0) $\delta = \delta_0$ against the alternative (H_1) $\delta \neq \delta_0$.

<u>*Wald test:*</u> Let $\boldsymbol{\Sigma}$ denote the variance covariance matrix of the interaction effects. The $Wald$ (or Omnibus) statistics is defined:

$$T_{Wald} = \delta^T\boldsymbol{\Sigma}^{-1}\delta$$

Under the null, T_{Wald} follows a χ^2 distribution with $K\times P$ degrees of freedom.

<u>*Likelihood Ratio Test:*</u> Let $L(\hat{\delta})$ denote the likelihood of the model when $\delta = \hat{\delta}$, the estimated coefficients (either using Maximum Likelihood Estimators or Ordinary Least

Squares). The statistics used in the Likelihood Ratio Test (*LRT*) is defined as:

$$T_{LRT} = 2\left(\log(L(\delta_0)) - \log\left(L(\hat{\delta})\right)\right)$$

The statistics T_{LRT} follows a χ^2 distribution with $K \times P$ degrees of freedom under the null.

Rao's Score Test (Lagrange Multiplier): The Rao's *Score* Test statistics is defined as:

$$T_{Score} = U^T(\delta_0)I^{-1}(\delta_0)U(\delta_0)$$

where $U(0) = \left.\frac{\partial \log(L)}{\partial \delta}\right|_{\delta=\delta_0}$ is the value of the derivative of the log-likelihood when $\delta = \delta_0$, and $I(0) = -\mathbb{E}\left[\left.\frac{\partial^2 \log(L)}{\partial\delta\partial\delta'}\right|_{\delta=\delta_0}\right]$ is the Fisher Information. Under the null, T_{Score} follows a χ^2 distribution with $K \times P$ degrees of freedom.

More details on those tests can be found in the literatures (Buse 1982; Engle 1984). Asymptotically, the three tests are equivalent even though some discrepancies can be observed in finite samples. In the case where the log-likelihood is quadratic, the three statistics of linear regression are equal but it has been proven that $T_{Wald} \geq T_{LRT} \geq T_{Score}$ (Buse 1982). This trend look different from our result in Figure 1 ($T_{LRT} \geq T_{Wald} \geq T_{Score}$) but it is mainly due to the inflated residual variance estimate of *LRT* (Details are described below).

LRT inflation observed in the linear model

An important inflation can be observed with the *LRT* (QQ plot for null model) when testing the simultaneous nullity of the interaction effects, whereas the distributions of the p-values obtained with the *Wald* test and the *Score* test are in adequacy with the expected uniform distribution. The explanation is that the *LRT* requires the estimation of the residual variance of the model. In the *lrtest*, this residual variance is estimated by $\hat{\sigma} = \sqrt{\sum \varepsilon_i^2 / n}$, which corresponds to the maximum likelihood estimator of the residual variance. However, this estimation is biased and an unbiased estimation of this residual variance is given by $\hat{\sigma} = \sqrt{\sum \varepsilon_i^2 / (n - r)}$, where r denotes the rank of the covariate matrix (here, $r = K \times P$. This different scaling of the two estimators leads to the inflation observed when performing the *LRT*. Still, substituting the Maximum Likelihood Estimator of the residual variance by its Ordinary Least Squares estimation corrects for this inflation and yields the same results as the two other tests (Figure S1).

Inflation of LRT and deflation of the Wald test in the logistic model

In the simulation setting of Figure 1 (N = 20,000; 30% disease prevalence, Nsnp = 100; Nexp = 10), the logistic model with interaction terms includes 1,111 (including intercept) parameters to be estimated. Thus, the number of Events Per Variable (EPV), defined as the ratio of the number of cases over the number of parameters, equals $6000/1111 = 5.4$. A commonly admitted rule of thumb in logistic model is to have at least 10 EPV. This rule has been defined based on previous simulation studies demonstrating that small EPV lead to bias in the estimation of both the parameters value and variance (Peduzzi et al. 1996), although other studies (van Smeden et al. 2016; Vittinghoff and McCulloch 2007) argued that this rule may be too conservative and highlighted that other factors (such as sample size or number of variables) might play a role in estimation accuracy. Overall, these previous works and other (Hauck and Donner 1977) –which focused on situation where a single parameter is tested–reported qualitatively similar behavior as we observed in our simulation where multiple parameters are tested jointly, showing that: i) the *Wald* test was too conservative (deflated) in the case of logistic regression with low EPV, and ii) that *LRT* test will tend to display inflated statistics.

The rationale for this behavior is likely explained by difference in the estimates required to perform each of the three tests. The *LRT* requires the likelihoods from both the saturated model (with the interaction terms) and the constrained model (with the interaction terms set to 0). The *Wald* test requires the coefficient and their standard error estimates in the saturated model. Finally, the *Score* test requires the coefficient and their standard error estimates in the unsaturated model. As discussed above, because of low EPV, estimates from the saturated model are biased. It follows that tests using information from the unsaturated model, *i.e.* the *LRT* and the *Wald* test, are more likely to perform poorly. On the other hand, the unsaturated model uses less parameters and thus the number of EPV mechanically increases (in our simulation scenario, 111 parameters and EPV $6000/111 = 55$). Consequently, parameters in the unsaturated model are less impacted by the aforementioned bias, explaining the better behavior of the *Score* test.

Reference

Buse, A. 1982. 'The Likelihood Ratio, Wald, and Lagrange Multiplier Tests - an Expository Note', *American Statistician*, 36: 153-57.

Engle, Robert F. 1984. 'Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics.' in, *Handbook of Econometrics* (Elsevier).

Hauck, Walter W., and Allan Donner. 1977. 'Wald's Test as Applied to Hypotheses in Logit Analysis', *Journal of the American Statistical Association*, 72: 851-53.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. 1996. 'A simulation study of the number of events per variable in logistic regression analysis', *J Clin Epidemiol*, 49: 1373-9.

van Smeden, M., J. A. de Groot, K. G. Moons, G. S. Collins, D. G. Altman, M. J. Eijkemans, and J. B. Reitsma. 2016. 'No rationale for 1 variable per 10 events criterion for binary logistic regression analysis', *BMC Med Res Methodol*, 16: 163.

Vittinghoff, E., and C. E. McCulloch. 2007. 'Relaxing the rule of ten events per variable in logistic and Cox regression', *Am J Epidemiol*, 165: 710-8.