S2 Appendix. The partial $F$ test in a fully adjusted full-stage INT procedure

A mathematical detail for the partial $F$ test in a fully adjusted full-stage INT procedure is provided for explaining why the fully adjusted full-stage INT procedure has a robust performance on control of the type I error specially under the situation in which genotypes are uncorrelated with covariates, in comparison with the fully adjusted two-stage INT procedure.

The partial $F$ test in a fully adjusted full-stage INT procedure

Processing Stages 1-4 of the fully adjusted full-stage INT approach, in the fifth stage of the fully adjusted full-stage INT approach, the INT-transformed residuals, $\text{RN}(\hat{e}), i = 1, 2, \ldots n$, which have a standard normal distribution, are immediately regressed on the SNP genotype $g$ and the covariates $X$. Precisely, The relationship between $\text{RN}(\hat{e}), X$ and $g$ is given by

$$\text{RN}(\hat{e}) = X\alpha^\ast + g\beta^\ast + \epsilon^\ast$$

(S4)

where $\alpha^\ast = (\alpha_0^\ast, \alpha_1^\ast, \ldots, \alpha_p^\ast)^T$ is a $p \times 1$ vector of regression coefficients of the covariates, $\beta^\ast$ is the regression coefficient of the SNP genotype $g$, $\epsilon^\ast = (\epsilon_1^\ast, \epsilon_2^\ast, \ldots, \epsilon_n^\ast)^T$ is an $n \times 1$ vector of random errors with each component independently from the normal distribution $N(0, \sigma^2)$. Let

$$\tilde{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n,j})^T, j = 1, 2, \ldots, p-1.$$ The regression form in Equation (S4) is re-expressed by

$$\text{RN}(\hat{e}) = \alpha_0^\ast + \tilde{x}_1\alpha_1^\ast + \tilde{x}_2\alpha_2^\ast + \cdots + \tilde{x}_{p-1}\alpha_{p-1}^\ast + g\beta^\ast + \epsilon^\ast.$$  

(S5)
According to correlation and regression theory, covariate effects $\alpha'' = (\alpha''_0, \alpha''_1, \ldots, \alpha''_{p-1})^T$ and the SNP (genetic) effect $\beta''$ can be written in terms of the relationship between standard deviations and partial correlation coefficients [1-3]. Precisely, the SNP (genetic) effect $\beta''$ in Equation (S5) is given by

$$\beta'' = \rho_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}} \frac{\sigma_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}}{\sigma_{g|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}}$$

where the partial correlation coefficient $\rho_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}$ stands for the correlation between $\text{RN}(\varepsilon)$ and $g$ adjusting for covariate effects $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{p-1}$ as well as $\sigma_{\text{RN}(\varepsilon)|x_1, x_2, \ldots, x_{p-1}}$ and $\sigma_{g|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}$ represent conditional standard deviations of $\text{RN}(\varepsilon)$ and $g$ given $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{p-1}$, respectively [1-3]. Correspondingly, the null hypothesis of $H_0: \beta'' = 0$ is equal to the null hypothesis of $H_0: \rho_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}} = 0$. The partial $F$ test under the null hypothesis of $H_0: \rho_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}} = 0$, i.e., under the null hypothesis of $H_0: \beta'' = 0$ is given by

$$F\left( g \mid \bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{p-1} \right) = \frac{\hat{\rho}_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}^2}{1 - \hat{\rho}_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}^2 / (n - p - 1)}$$  \hspace{1cm} (S6)

where $\hat{\rho}_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}^2$ is the estimate of $\rho_{\text{RN}(\varepsilon)|\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{p-1}}^2$ and the null distribution of the partial $F$ test follows a $F$ distribution with one and $n - p - 1$ degrees of freedom [1].

Alternatively, the partial $F$ test in Equation (S6) is equivalently computed by

$$F\left( g \mid \bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{p-1} \right) = T^2 = \left( \hat{\beta}'' / \hat{\sigma}_p'' \right)^2$$
where \( T = \hat{\beta}^\ast / \hat{\sigma}_{\beta}^\ast \) as well as \( \hat{\beta}^\ast \) and \( \hat{\sigma}_{\beta}^\ast \) are the estimates of \( \beta^\ast \) and its standard deviation \( \sigma_{\beta}^\ast \), respectively [1]. The null distribution of \( T = \hat{\beta}^\ast / \hat{\sigma}_{\beta}^\ast \) follows a \( t \) distribution with \( n - p - 1 \) degrees of freedom [1].

As had been shown above, \( \beta^\ast = 0 \) is equal to \( \rho_{RN(\varepsilon)g|x_1,x_2,x_3,...,x_p} = 0 \). The partial correlation \( \rho_{RN(\varepsilon)g|x_1,x_2,x_3,...,x_p} \) is equal to zero which indicates that the SNP genotype \( g \) is uncorrelated with the INT-transformed residuals \( RN(\hat{\varepsilon}) \). This in turn means that the SNP genotype \( g \) are uncorrelated with the covariates \( X \) and the traits \( y \), because the INT-transformed residuals \( RN(\hat{\varepsilon}) \) comprise the elements \( X \) and \( y \). Therefore, under the null hypothesis of no association between the outcomes and the SNP (genetic) effect, the INT-transformed residuals, \( RN(\hat{\varepsilon}_i), i = 1,2,...,n \), in Equation (S5) can prove more valid inference for the SNP (genetic) effect, when the SNP genotype \( g \) are uncorrelated with the covariates \( X \). Moreover, the INT-transformed residuals, \( RN(\hat{\varepsilon}_i), i = 1,2,...,n \), in the fifth stage of the fully adjusted full-stage INT method have sufficiently made for following a standard normal distribution, when the INT-transformed residuals, \( RN(\hat{\varepsilon}_i), i = 1,2,...,n \), in the second of the fully adjusted two-stage INT method don’t satisfy the assumption of the normal distribution with zero mean and finite variance. Therefore, in comparison with the fully adjusted two-stage INT method, when the SNP genotype is uncorrelated with the covariates, the fully adjusted full-stage INT method, under the null hypothesis of no SNP (genetic) effect, can provide valid statistical
inference for the SNP (genetic) effect, although the distribution of the traits (or the residuals) of
the regression that is not normal is heavily skewed or the traits involve the outliers.

References

1. Kleinbarum DG, Kupper LL, Nizam A, Muller KE. Applied regression analysis and other
2. Demissie S, Cupples LA. Bias due to 2-stage residual-outcome regression analysis in genetic
3. Che R, Motsinger-Reif AA, Brown CC. Loss of power in two-stage residual-outcome