Robust testing using Tukey's one-df interaction model

Wenxing Wang, Arnab Maity

Keywords: Hypothesis testing, Robust regression, Tukey's 1-df interaction, Permutation test

1 Introduction

Over recent decades, advances in biomedical research have been significantly propelled by the genome-wide association studies (GWAS). A primary focus within this framework has been on single nucleotide polymorphisms (SNPs), which act as pivotal genetic variants enabling researchers to decode complex trait architectures and disease susceptibilities. For instance, the work of Michailidou et al.[1] investigated associations between 67 SNPs and breast cancer susceptibility, illustrating the prognostic potential of genetic markers and their implications in risk stratification and personalized medicine. Leveraging statistical models, researchers can now quantify SNP effects on disease, enabling the decomposition of phenotypic variance into genetic components and advancing our understanding of traits' heritability. Such models are foundational in linking SNP-level associations to functional characteristics of target genes, providing a deeper understanding of the biological pathways and regulatory mechanisms that drive complex diseases.

A persistent challenge in single nucleotide polymorphism (SNP) research is the complicated nature of gene contributions to disease phenotypes, including both direct gene-gene interactions and indirect protein-protein, gene-environment interactions[2]. Chatterjee et al.[3] introduced a parsimonious model that incorporates gene-gene and gene-environment interactions through Tukey's 1-degree-of-freedom (1-df) interaction model[4]. This logistic

regression framework includes covariates representing two groups of SNPs, as well as an interaction term for these two groups of SNPs. The model's response variable is a binary clinical outcome, indicating the presence or absence of a specific diagnosis, such as cancer. The model facilitates the assessment of associations within a complex genomic context. Additionally, Chatterjee et al.[4] developed a corresponding statistical test to evaluate gene contributions in the presence of multifaceted genetic interactions, thereby advancing statistical tools for disentangling these intricate relationships. However, this model's logistic regression framework limits its applicability to cases with binary clinical outcome. There is a growing need for models that accommodate continuous clinical measurements, like plasma glucose levels, cholesterol level or blood pressure level, which are vital for biomedical research and drug development.

In this study, we introduce TukRobust, a novel statistical testing framework designed to incorporate gene-gene and gene-environment interactions into association studies involving continuous clinical outcomes. Our approach leverages Tukey's 1-degree-of-freedom interaction model to account for complex interaction effects. Unlike binary outcomes, which are typically modeled as 0/1 variables, continuous clinical outcomes frequently exhibit outliers that can significantly distort hypothesis testing results under a normal assumption with squared loss. To address this issue, TukRobust includes an option for Laplace-distributed residuals with absolute and Huber loss, offering improved resilience to outliers and skewness commonly present in clinical data. We evaluate the TukRobust through simulations of case-control data that emulate association study designs with continuous outcomes containing outliers, assessing the model's type I error rates, statistical power, and robustness. Furthermore, we apply our method to genomic and clinical data from a cohort of diabetes patients, demonstrating its advantages in detecting gene-gene and gene-environment interactions in continuous outcomes, particularly with respect to outlier robustness.

2 Method

2.1 Hypothesis testing

The Tukey's 1-df interaction model takes the formula:

$$Y_i = \beta_0 + S_{i1}^T\beta_1 + S_{i2}^T\beta_2 + \gamma(S_{i1}^T\beta_1)(S_{i2}^T\beta_2) + e_i$$

with the null hypothesis

$$H_0 : \beta_1 = 0$$

where Y is the continuous outcome and S_1 and S_2 are two groups of SNPs or environmental factors, with β_1 and β_2 being their corresponding effect sizes.

2.2 Steps for deriving the test statistics

The steps for deriving the test statistics are

1. Obtain the maximum-likelihood estimates for β_0, β_2 and σ under null hypothesis H_0. The model becomes a standard linear regression with the estimates $\hat{\psi} = (\hat{\beta}_0, \hat{\sigma}, \hat{\beta}_2)$

2. For a fixed value θ, evaluate the score functions for the parameters β_1 at $\beta_1 = 0$, $\psi = \hat{\psi}$ and $\gamma = \theta$, that is
$$S_{\beta_1}(\theta) = \frac{dl}{d\beta_1} \mid_{\beta_1=0,\psi=\hat{\psi},\gamma=\theta}$$

3. Estimate the inverse of the variance-covariance matrix for $S_{\beta_1}(\theta)$ with

$$I^{\beta_1\beta_1}(\theta) = \left[I_{\beta_1\beta_1}(\theta) - I_{\beta_1\psi}(\theta)I_{\psi\psi}^{-1}I_{\psi\beta_1}(\theta)\right]^{-1}$$

where the expressions for the component information matrices $I_{\beta_1\beta_1}(\theta) = \partial l/\partial\beta_1\partial\beta_1^T$, $I_{\beta_1\psi}(\theta) = \partial l/\partial\beta_1\partial\psi^T$ and $I_{\psi\psi} = \partial l/\partial\psi\partial\psi^T$, evaluated at $\beta_1 = 0$ and $\psi = \hat{\psi}$

4. For given value of θ, obtain the score statistics

$$T_1(\theta) = S_{\beta_1}(\theta)^T I^{\beta_1\beta_1}(\theta) S_{\beta_1}(\theta).$$

Compute the final test statistics as $T_1^* = \max_{L\leq\theta\leq U} T(\theta)$, where L and U denote some pre-specified values for lower and upper limits of θ, respectively.

2.2.1 Compute $T_1(\theta)$ with normal residual

To compute the test statistics with the given θ, score function and its variance-covariance matrixs are needed. Under normal assumptions on the residual when e_i are i.i.d $N(0, \sigma^2)$ errors, the log likelihood function is

$$l = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}[Y_i - \beta_0 - S_{i1}^T\beta_1 - S_{i2}^T\beta_2 - \gamma(S_{i1}^T\beta_1)(S_{i2}^T\beta_2)]^2/(2\sigma^2)$$

The score function for β_1 is

$$S_{\beta_1}(\theta) = \frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} S_{i1}\{1 + \gamma S_{i2}^T\beta_2\}e_i/\sigma^2$$

when

$$e_i = Y_i - \beta_0 - S_{i1}^T\beta_1 - S_{i2}^T\beta_2 - \theta(S_{i1}^T\beta_1)(S_{i2}^T\beta_2)$$

for the given

$$\gamma = \theta$$

The variance-covariance matrix of the estimated score function is

$$I = \begin{bmatrix} I_{\beta_1\beta_1} & I_{\beta_1\beta_0} & I_{\beta_1\beta_2} & I_{\beta_1\sigma^2} \\ I_{\beta_1\beta_0} & I_{\beta_0\beta_0} & I_{\beta_0\beta_2} & I_{\beta_0\sigma^2} \\ I_{\beta_2\beta_1} & I_{\beta_2\beta_0} & I_{\beta_2\beta_2} & I_{\beta_2\sigma^2} \\ I_{\sigma^2\beta_1} & I_{\sigma^2\beta_0} & I_{\sigma^2\beta_2} & I_{\sigma^2\sigma^2} \end{bmatrix}$$

Each part of the matrix are

$$I_{\beta_0\beta_1} = E[\sum_{i=1}^{n} S_{i1}\{1+\theta S_{i2}^T\beta_2\}e_i^2/\sigma^4] = \sum_{i=1}^{n} S_{i1}\{1+\theta S_{i2}^T\beta_2\}/\sigma^2$$

$$I_{\beta_0\beta_2} = E[\sum_{i=1}^{n} S_{i2}\{1+\theta S_{i1}^T\beta_1\}e_i^2/\sigma^4] = \sum_{i=1}^{n} S_{i2}\{1+\theta S_{i1}^T\beta_1\}/\sigma^2$$

$$I_{\beta_1\beta_1} = E[\sum_{i=1}^{n} S_{i1}S_{i1}^T\{1+\theta S_{i2}^T\beta_2\}^2e_i^2/\sigma^4] = \sum_{i=1}^{n} S_{i1}S_{i1}^T\{1+\theta S_{i2}^T\beta_2\}^2/\sigma^2$$

$$\begin{aligned} I_{\beta_1\beta_2} &= E[\sum_{i=1}^{n} S_{i1}S_{i2}^T\{1+\theta S_{i1}^T\beta_1\}\{1+\theta S_{i2}^T\beta_2\}e_i^2/\sigma^4] \\ &= \sum_{i=1}^{n} S_{i1}S_{i2}^T\{1+\theta S_{i1}^T\beta_1\}\{1+\theta S_{i2}^T\beta_2\}/\sigma^2 \end{aligned}$$

$$I_{\beta_2\beta_2} = E[\sum_{i=1}^{n} S_{i2}S_{i2}^T\{1+\gamma S_{i1}^T\beta_1\}^2e_i^2/\sigma^4] = \sum_{i=1}^{n} S_{i2}S_{i2}^T\{1+\theta S_{i1}^T\beta_1\}^2/\sigma^2$$

$$I_{\beta_0\sigma^2} = 0,\ I_{\beta_1\sigma^2} = 0,\ I_{\beta_2\sigma^2} = 0$$

$$I_{\sigma^2\sigma^2} = E[\{-\frac{n}{2\sigma^2} + \sum_{i=1}^{n} e_i^2/(2\sigma^4)\}^2] = \mathrm{var}[\sum_{i=1}^{n} e_i^2/(2\sigma^4)] = \frac{n}{2\sigma^4}$$

and also

$$I_{\beta_1\psi} = \begin{bmatrix} I_{\beta_1\beta_0} & I_{\beta_1\beta_2} & I_{\beta_1\sigma^2} \end{bmatrix}$$

$$I_{\psi\psi} = \begin{bmatrix} I_{\beta_0\beta_0} & I_{\beta_0\beta_2} & I_{\beta_0\sigma^2} \\ I_{\beta_2\beta_0} & I_{\beta_2\beta_2} & I_{\beta_2\sigma^2} \\ I_{\sigma^2\beta_0} & I_{\sigma^2\beta_2} & I_{\sigma^2\sigma^2} \end{bmatrix}$$

2.2.2 Compute $T_1(\theta)$ with Laplace residual

TukRobust also have the Laplace residual option with Laplace assumption on e_i, expecting more robust performance when dealing with outcomes containing outliers. Under this assumption e_i follows Laplace distributions with scale $b = \sigma$. Denote

$$\mu_i = \beta_0 + S_{i1}^T\beta_1 + S_{i2}^T\beta_2 + \gamma(S_{i1}^T\beta_1)(S_{i2}^T\beta_2)$$

the log likelihood function is

$$l = -\log\sigma - \sum_{i=1}^{n} |Y_i - \mu_i| / 2\sigma^2$$

and the score function is

$$S_{\beta_1}(\theta) = -\sum_{i=1}^{n} \text{sign}\,(Y_i - \mu_i)\, S_{1i}\,(1 + \theta S_{2i}\beta_{2i})\,/2\sigma^2$$

To resolve the issue that the derivative of the score function will be 0 due to the sign function, we consider the asymptotic test for large samples. When sample size is sufficiently large, the conclusion holds that

$$S_{\beta_1}(\theta)\mid_{\hat{\psi}} \rightarrow S_{\beta_1}(\theta)\mid_{\psi}$$

when $\hat{\psi}$ is the maximum likelihood estimation of ψ under null hypothesis. The variance-covariance matrix of the score function is

$$\text{Var}\left[S\left(\beta_{1p}\right)\right] = \sum_{i=1}^{n} S_{1pi}^2\left(1 + \theta S_{2i}\beta_{2i}\right)^2 / 4\sigma^4$$

for the p^{th} value in vector β_1 and also

$$\text{Cov}[S(\beta_{1p}), S(\beta_{1q})] = \sum_{i=1}^{N} S_{1pi}S_{1qi}\left(1 + \theta S_{2i}\beta_{2i}\right)^2 / 4\sigma^4$$

for the p^{th} value and q^{th} in vector β_1 and also and $p \neq q$.

2.3 Permutation test

When hard to derive the null distribution of the test statistic in small finite sample case and large sample case, we consider the permutation test to finish the hypothesis testing. We adopted the method is proposed by Freedman and Lane[5] with steps:

1. Estimate the value of β_2 and β_0 under the null hypothesis that $\beta_1 = 0$, denoted as $\hat{\beta}_2$ and $\hat{\beta}_0$
2. Calculate the residuals $R_{\text{null}} = Y - \hat{\beta}_0 - S_2\hat{\beta}_2$
3. Permute the residues randomly and produce R^*_{null}
4. New values of Y^* are calculated by adding R^*_{null} back to the fitted means, with $Y^* = \hat{\beta}_0 + S_2\hat{\beta}_2 + R^*_{\text{null}}$
5. Compute the test statistic with Y^* against S_1 and S_2

The process above will be repeated sufficiently many times to produce the distribution of the test statistic T_1^* under null hypothesis. The p-value is calculated as

$$p = \sum_{b=1}^{B} I(T_b > T_{obs})/B$$

Here, B denotes the number of permutation repetitions, T_b represents the test statistic calculated for each permutation with permuted outcomes Y^*, and T_{obs} is the test statistic calculated using the observed outcome Y.

3 Simulations

We compared the Type I error and powers of the proposed hypothesis testing and several different statistical methods. In the simulated datasets, the response is generated by

$$Y_i = \beta_0 + S_{i1}^T\beta_1 + S_{i2}^T\beta_2 + \gamma(S_{i1}^T\beta_1)(S_{i2}^T\beta_2) + \epsilon_i$$

where S_1 and S_2 are two groups of covariates generated from a standard normal distribution and β_1 and β_2 are two vectors with the length of 5. All values within the same vector of β_1 or β_2 are set to be identical, where $\beta_2 = 0.2 * 1_5$ and 1_k represents a vector of length k with each element being 1. We set $\beta_1 = \delta * 1_5$ for some constant δ. We set $\delta = 0$ to simulate the null hypothesis to evaluate type I error, and increase δ from 0.1 to 0.8 to evaluate the power of TukRobust. Additinally we set $\beta_0 = 3$ and $\gamma = 1$. To determine the minimum sample size required for the convergence of the robust test statistic, given its reliance on large-sample properties, we created four groups of sample sizes of 50, 100, 200, and 500, respectively.

To evaluate the robustness of TukRobust in the presence of outliers within outcomes, we generated the error term, ϵ, using four distinct distributions known for their propensity to produce outliers: the Laplace distribution, logistic distribution, Cauchy distribution, and 3 mixed distributions comprising two normal distributions with different variances. These scenarios have included some extreme cases such as Mix 3. Detailed specifications of ϵ generation are provided in Table 1. When conducting the permutation test, we performed 200 repeats of permutation to obtain the distribution of T_1^* under null hypothesis.

To estimate the Type I error rate, for each sample size, we conducted 1000 simulations with $\delta = 0$, recording the p-value for each iteration. The Type I error rate was then calculated as the percentage of simulations in which the p-value was less than or equal to a specified significance level, such as 0.05. To evaluate statistical power, we conducted 300 simulations for each sample size and for each value of δ ranging from 0.1 to 0.8, recording the p-value for each simulation. The power was calculated as the percentage of simulations in which the

p-value was less than or equal to the given significance level of 0.05.

Distribution	Settings
Laplace	Location: 0, Scale: $3/\sqrt{2}$
Logistic	Location: 0, Scale: $3\sqrt{3}/\pi$
Cauchy	Location: 0, Scale: 3
Mix 1	90% $N(0, 3^2)$ and 10% $N(0, 10^2)$
Mix 2	80% $N(0, 3^2)$ and 10% $N(0, 20^2)$
Mix 3	80% $N(0, 3^2)$ and 10% $N(0, 70^2)$

Table 1: Configurations for simulation studies

Table 2 illustrates the observed Type I error rate of all simulation settings with the significance level of $\alpha = 0.05$. It reveals that in scenarios with extreme outliers in the residuals, specifically under the Cauchy and Mix 3 distributions, hypothesis testing under the assumption of normal residuals exhibited a substantial inflation of Type I error rates. In contrast, when using the robust alternative with a Laplace residual assumption, Type I error rates were well-controlled across these settings. This suggests that the robust method with Laplace residuals is effective in controlling the impact of extreme outliers on Type I error, enhancing the reliability of hypothesis testing in the presence of heavy-tailed distributions on outcome residuals.

	$N = 50$		$N = 100$		$N = 200$		$N = 500$	
	Normal	Laplace	Normal	Laplace	Normal	Laplace	Normal	Laplace
Laplace	0.063	0.051	0.088	0.046	0.04	0.053	0.162	0.049
Logistic	0.068	0.046	0.064	0.054	0.051	0.046	0.047	0.048
Cauchy	0.186	0.046	0.159	0.051	0.133	0.06	0.141	0.035
Mix 1	0.064	0.062	0.068	0.058	0.052	0.054	0.063	0.054
Mix 2	0.083	0.056	0.078	0.063	0.065	0.056	0.045	0.063
Mix 3	0.158	0.05	0.09	0.059	0.097	0.066	0.065	0.049

Table 2: Type I error rate of all simulation settings at significance level of $\alpha = 0.05$

Figure 1 illustrates the statistical power of the hypothesis tests conducted under the assumption of normal residuals at significance level of $\alpha = 0.05$. The power curve generally exhibits a desirable form across most conditions; however, a notable deviation occurs in scenarios involving Mix 3 residuals, the most extreme case, where the power is diminished. This outcome indicates that the presence of Mix 3 residuals adversely impacted the power, highlighting limitations in the test's robustness under this residual structure. In contrast, Figure 2 presents the power of the hypothesis tests conducted under the assumption of Laplace residuals. The power curve maintains a consistent and desirable shape across all cases, including those with extreme outliers from the Cauchy and Mix 3 distributions. This suggests that the Laplace residual assumption enhances the robustness of the hypothesis test's power, even in the presence of heavy-tailed distributions and substantial outliers.

Figures 1 and 2 demonstrate that the statistical power of the hypothesis tests increases with larger sample sizes. This trend aligns with the expected asymptotic properties of the test, where larger samples enhance power, thus validating the theoretical assumptions underlying the test's effectiveness as sample size grows.

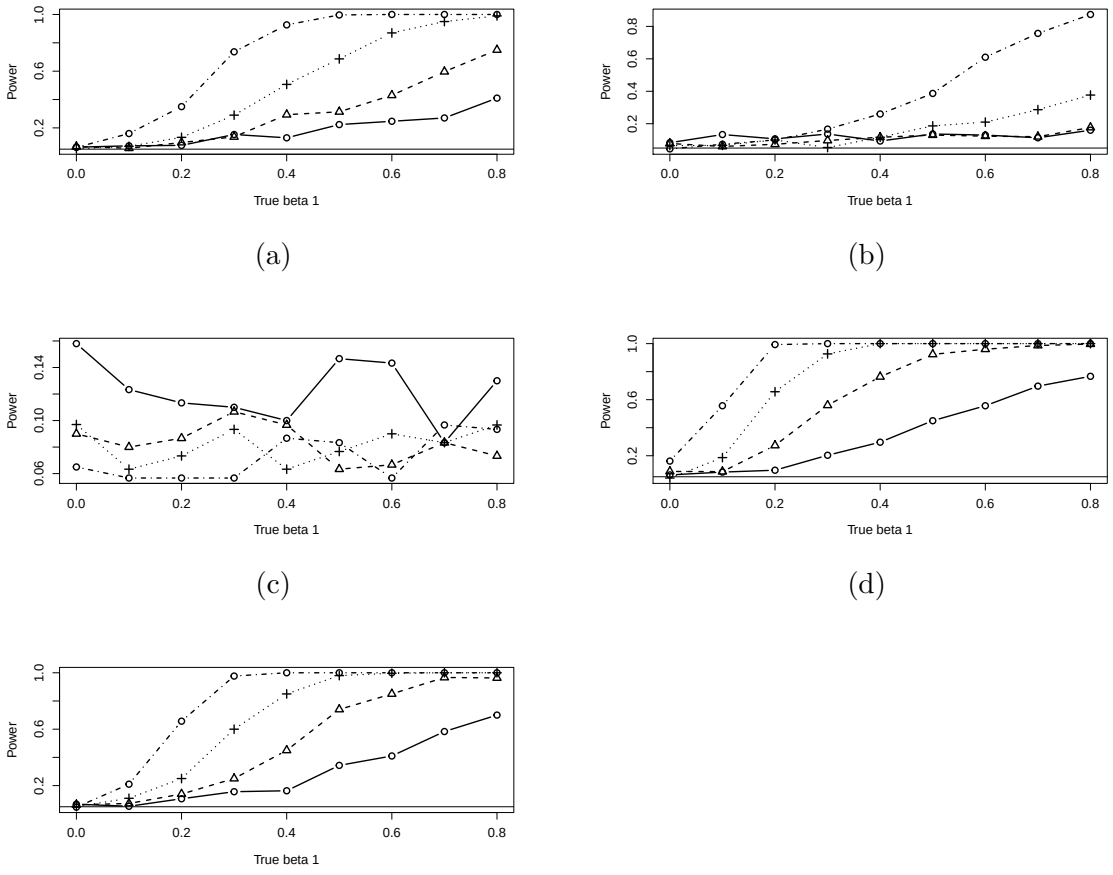

Figure 1: Power of the hypothesis testing under the assumption of normal residuals. Simulation settings on residuals: (a) Mix 1, (b) Mix 2, (c) Mix 3, (d) Laplace distribution, (e) Logistic distribution. The four curves in each figure represent four sample sizes: Solid line with hollow circle: $N = 50$, Dashed line with triangle: $N = 100$, Dashed line with cross: $N = 200$, Dot-dashed line with hollow circles: $N = 500$.

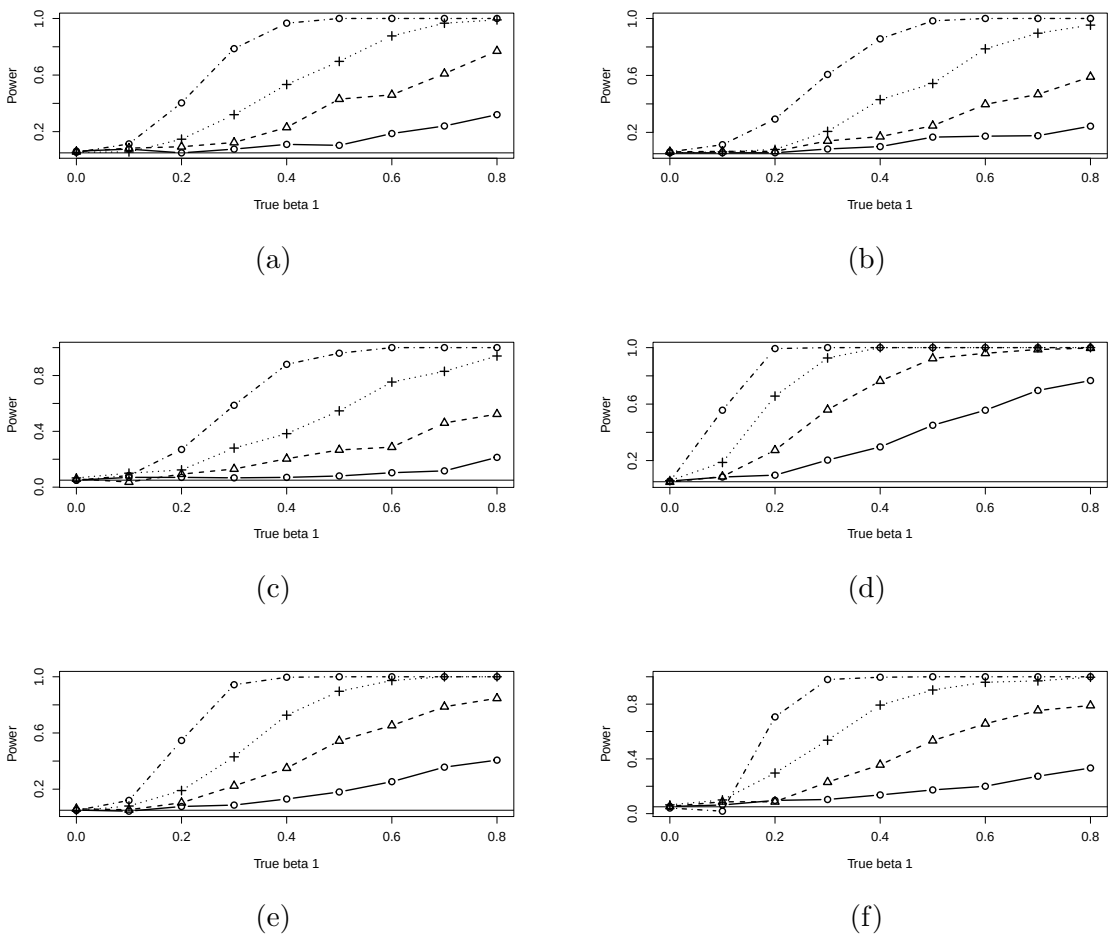

Figure 2: Power of the hypothesis testing under the assumption of Laplace residuals. Simulation settings on residuals: (a) Mix 1, (b) Mix 2, (c) Mix 3, (d) Laplace distribution, (e) Logistic distribution, (f) Cauchy distribution. The four curves in each figure represent four sample sizes: Solid line with hollow circle: $N = 50$, Dashed line with triangle: $N = 100$, Dashed line with cross: $N = 200$, Dot-dashed line with hollow circles: $N = 500$.

4 Real Data Analysis

5 Discussion

In this study, we developed a robust hypothesis testing method tailored for a Tukey 1-degree-of-freedom interaction linear regression model. This approach is well-suited for assessing the contributions of specific target genes to continuous clinical measurements, accounting for

gene-gene and gene-environment interactions. The derivation of our test statistic involved constructing a permutation test by permuting residuals under the reduced model. To mitigate the potential adverse effects of extreme outliers in the dataset, we introduced an option of Laplace assumption on residuals.

We conducted extensive simulation studies to evaluate the performance of our test statistic and to compare the efficacy of using the normal residual assumption versus Laplace residual assumption in handling outliers. Results indicated that under the permutation test framework, the test statistic demonstrated reliable performance, maintaining acceptable Type I error rates and desirable powers. While both the normal residual assumption and Laplace residual assumption were effective with moderate outliers, severe outliers in the response variable significantly impacted the test using normal residual assumption, resulting in dimished power and inflated Type I error rates. In contrast, the Laplace residual assumption approach retained robust power, demonstrating its advantage in scenarios with extreme outliers.

References

[1] Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., ... & Devilee, P. (2013). *Large-scale genotyping identifies 41 new loci associated with breast cancer risk*, Nature genetics, 45(4), 353-361.

[2] Jia, Y., McAdams, S. A., Bryan, G. T., Hershey, H. P., & Valent, B. (2000). *Direct interaction of resistance gene and avirulence gene products confers rice blast resistance*, The EMBO journal, 19(15), 4004–4014.

[3] Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., & Wacholder, S. (2006). *Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions*, The American Journal of Human Genetics, 79(6), 1002-1016.

[4] Tukey, J. W. (1949). *One degree of freedom for non-additivity*, Biometrics, 5(3), 232-242.

[5] Freedman, D., & Lane, D. (1983). *A nonstochastic interpretation of reported significance levels*, Journal of Business & Economic Statistics, 1(4), 292-298.

[6] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*, Proceedings of the National Academy of Sciences, 106(23), 9362-9367.

[7] Jablonski, K. A., McAteer, J. B., de Bakker, P. I., Franks, P. W., Pollin, T. I., Hanson, R. L., ... & Diabetes Prevention Program Research Group. (2010). *Common variants in 40 genes assessed for diabetes incidence and response to metformin and lifestyle intervention in the diabetes prevention program*, Diabetes, 59(10), 2672-2681.

[8] Mukherjee, B., Ahn, J., Gruber, S. B., & Chatterjee, N. (2012). *Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons*, American journal of epidemiology, 175(3), 177-190.

[9] Titeca, K., Lemmens, I., Tavernier, J., & Eyckerman, S. (2019). *Discovering cellular protein-protein interactions: Technological strategies and opportunities*, Mass spectrometry reviews, 38(1), 79-111.

[10] Wang, Y., Li, D., & Wei, P. (2015). *Powerful Tukey's One Degree-of-Freedom Test for Detecting Gene-Gene and Gene-Environment Interactions*, Cancer informatics, 14, CIN-S17305.