Comparison between GxEScanR and GLM for 2df/3df models

Updated on 2020-04-07

Goal

For post-hoc analyses, need to validate hits by fitting various models (GxE, joint models) with additional covariates. In the process, noticed that 3DF results have very small differences between GLM and GxEScanR

In here, I'm summarizing 2DF/3DF results for 10 SNPs from the Type II Diabetes GWIS, for both GxEScanR and GLM

Models for likelihood ratio test

2DF joint test

- base model: $Outcome \sim SNP + E + age + sex + pc1 + pc2 + pc3 + study$
- interaction model: $Outcome \sim SNP + E + SNP * E + age + sex + pc1 + pc2 + pc3 + study$

3DF joint test

- base model: $SNP \sim age + sex + pc1 + pc2 + pc3 + study$
- interaction model: $SNP \sim Outcome + E + Outcome * E + age + sex + pc1 + pc2 + pc3 + study$

P-values

SNP	Subjects	pval_gxescan_2df	pval_glm_2df	pval_gxescan_3df	pval_glm_3df
1:71040166:G:T	74390	1.77074e-07	1.77072e-07	1.25551e-09	1.57517e-09
10:114754071:T:C	74390	6.01701e-01	6.01701e-01	2.45775e-49	1.89116e-49
10:114784926:C:T	74390	2.82608e-01	2.82608e-01	4.83087e-08	4.21936e-08
12:4384844:T:G	74390	6.61997e-04	6.61980e-04	4.42451e-10	6.31504e-10
16:53811788:A:G	74390	9.87860e-01	9.87860e-01	4.72788e-08	4.76705e-08
20:6442961:G:A	74390	2.67526e-07	2.67523e-07	3.89925e-08	5.62870e-08
3:185510884:A:C	74390	2.57710e-01	2.57710e-01	1.02726e-11	1.31501e-11
6:20688121:T:A	74390	3.75289e-05	3.75294e-05	2.75143e-11	2.28556e-11
7:28189411:T:C	74390	4.66363e-01	4.66362e-01	5.74195e-10	4.32742e-10
8:118185025:G:A	74390	2.75634e-03	2.75633e-03	7.87174e-11	6.06649e-10

chi-square values

SNP	Subjects	chiSqGE	chiSq2df	chiSq_glm_2df	chiSq3df	chiSq_glm_3df
1:71040166:G:T	74390	13.28280	31.0934000	31.0934149	44.37620	43.91250
10:114754071:T:C	74390	227.82900	1.0159900	1.0159894	228.84499	229.37138
10:114784926:C:T	74390	34.37270	2.5273900	2.5273899	36.90009	37.17790
12:4384844:T:G	74390	31.86660	14.6405000	14.6405494	46.50710	45.78045
16:53811788:A:G	74390	36.91990	0.0244291	0.0244291	36.94433	36.92739
20:6442961:G:A	74390	7.07173	30.2681000	30.2681184	37.33983	36.58628
3:185510884:A:C	74390	51.46790	2.7118400	2.7118440	54.17974	53.67686
6:20688121:T:A	74390	31.79210	20.3808000	20.3807719	52.17290	52.55087
7:28189411:T:C	74390	44.44920	1.5255800	1.5255850	45.97478	46.55241
8:118185025:G:A	74390	38.24250	11.7877000	11.7877140	50.03020	45.86247

You can see results for 2DF are essentially the same, while there are slight differences for 3DF. What do you think? Is this small difference acceptable?

Fixed

Using GLM I'll simply calculate 3DF pvalues the same exact way as gxescan e.g. adding E|G (linear) to 2DF chiSq and calculalate 3df p values. Once i do that the chiSq 3DF values are consistent:

SNP	Subjects	chiSqGE	chiSq2df	chiSq_glm_2df	chiSq3df	chiSq_glm_3df
1:71040166:G:T	74390	13.28280	31.0934000	31.0934149	44.37620	44.37617
10:114754071:T:C	74390	227.82900	1.0159900	1.0159894	228.84499	228.84546
10:114784926:C:T	74390	34.37270	2.5273900	2.5273899	36.90009	36.90013
12:4384844:T:G	74390	31.86660	14.6405000	14.6405494	46.50710	46.50714
16:53811788:A:G	74390	36.91990	0.0244291	0.0244291	36.94433	36.94436
20:6442961:G:A	74390	7.07173	30.2681000	30.2681184	37.33983	37.33984
3:185510884:A:C	74390	51.46790	2.7118400	2.7118440	54.17974	54.17970
6:20688121:T:A	74390	31.79210	20.3808000	20.3807719	52.17290	52.17289
7:28189411:T:C	74390	44.44920	1.5255800	1.5255850	45.97478	45.97475
8:118185025:G:A	74390	38.24250	11.7877000	11.7877140	50.03020	50.03022