Effect Estimation Project

Brandon Avila

September 6, 2016

1

Sorry we didn't get to talk this weekend. I pretty much had what I wanted to do already done by yesterday, but being the holiday, I had a few errands to do after finally finishing moving.

I elected to type up a LaTeX document because it was just easier to insert my figures and special characters and whatnot. I figured a PDF would be easy enough for you to read even if you didn't have time to video chat.

2

I'll quickly outline the methods here and then give the results of what I did. All methods were performed on sets of 500000 simulated individuals, a disease with $h^2 = 0.3$, baseline prevalence 0.1, and 200 common known-effect variants with 2 rare unknown-effect variants. Rare variants were always protective with frequency 0.005 and effect size -1.23. Basically, it's the same as when you were looking at it. Each study only estimates the effect of one of the variants, namely, β_{201}.

2.1 Method A

This one took the cases and controls and assumed we had all of the individuals. It simply compared the prevalence of disease in those with one protective allele to that in those with zero protective alleles. With only the individuals I generated (significantly fewer cases than controls), I took this to be a decent estimate of our best guess on the effect size. The assumption here was that if the method gave an inaccurate estimate, it probably was because we didn't get enough cases or enough individuals with the protective allele or something like that. While less optimal than performing a true case-control study, this gave reasonable results, and was held as the "silver standard" for the rest of my tests.

2.2 Method B

Here, we select out the healthy individuals (as we assume we only have controls). Then we do a linear fit of $\widehat{PRS}$ to genotype. As we discussed, the slope is the negative of $\hat{\beta}_{201}$.

2.3 Method C

Was an attempt to do something similar to Method B with cases rather than controls, but I abandoned this idea and just stuck with all control-based tests. No results from this method.

2.4 Method D

Essentially, this is a pseudo-case-control study on the tails of the $\widehat{PRS}$ distribution of the healthy individuals. Take the top and bottom tails (tail size can be altered, but commonly 10%), and compare the "prevalence" of disease in those with zero, one, and two allele genotypes for our rare variant. If no individuals in the tails have genotype 2, we discard the result, as two points cannot give us a p-value, so the estimate is not meaningful. Then, translate the prevalences into risk scores, and do a linear fit on the scores vs genotypes. If the prevalence in those with genotype $G_{201} = 2$ is 0 or 1, the score for that point is $\pm\infty$, so the linear fit will make no sense, and the estimate is not meaningful.

For this method, only the estimates that did not fail in either of those ways were included in the results.

2.5 Method E

Perform Method B, but rather than fitting the entire $\widehat{PRS}$ distribution to the genotypes, select out only the middle 50% of the distribution for each genotype. The thought here was that we would still maintain a distribution of $\widehat{PRS}$ for each genotype, but would not allow outliers to contribute as much to the estimate.

2.6 Method F

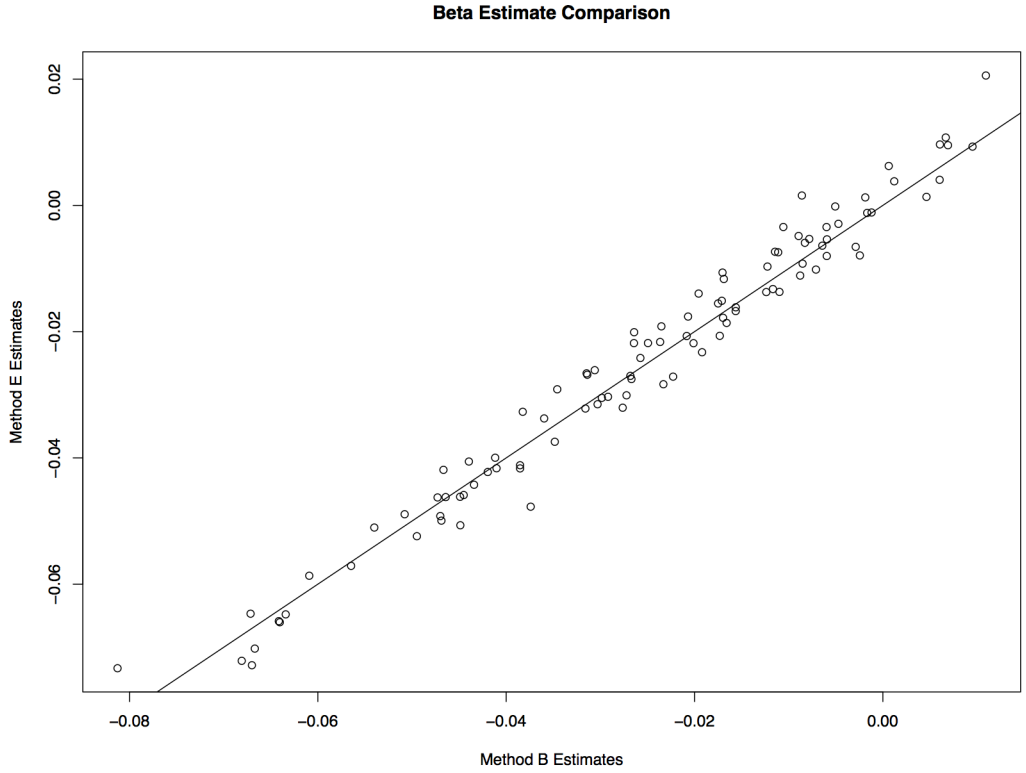
Perform Method D, but instead of using three equally weighted points, weight each genotype point by the number of individuals (in the tails) with that genotype. Now, for example, a prevalence estimate from one person with $G_{201} = 2$ should count for less than a prevalence estimate from 15000 people with $G_{201} = 0$.

3

I did a lot of playing with things and tweaking parameters, so I'll try to just give you the rundown on how the methods compare, and what I think the best choices are.

3.1 B-E Comparison

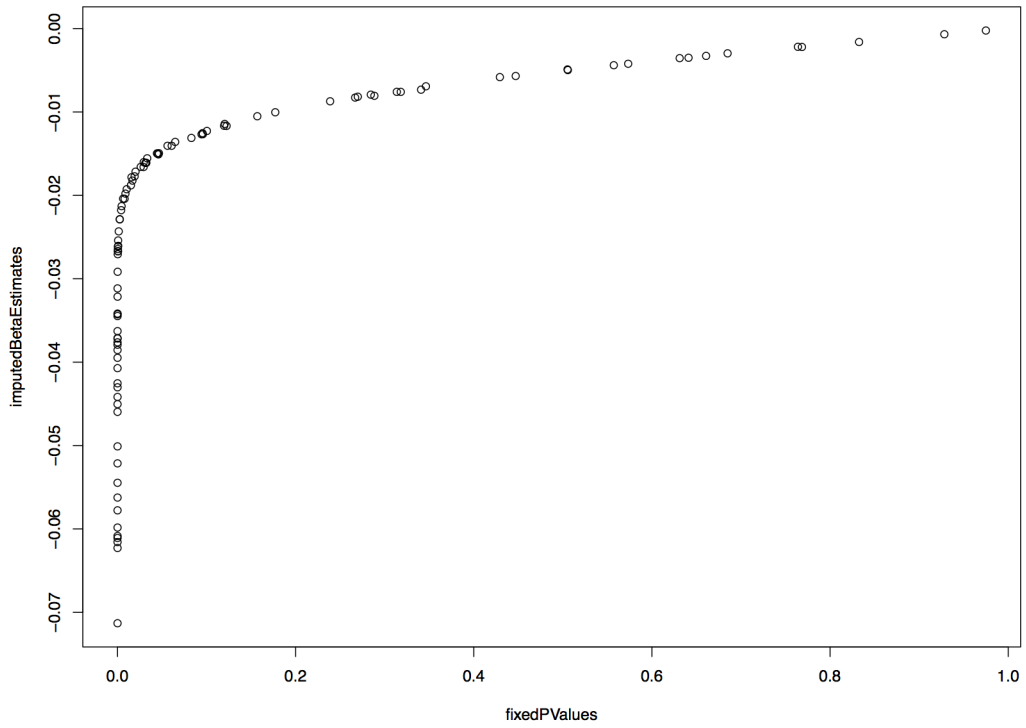
Obviously we need to do a comparison of these two, since E is a variation on B.



The above figure has estimates from B and E on 100 tests. My takeaway from this figure was that it basically doesn't matter whether we use B or E. For me, it was actually more computationally intensive to select out the middle 50% of the distributions rather than just run them all through the fit, so I rejected E and used B from here on out.

3.2 Method B p-Values

It was important to check the p-Values on the estimates from Method B to see when we could definitively say that we had detected an effect, and make sure that when we did, we had at least the right sign. It turns out that the relation between p-Value and estimate was super super strong.

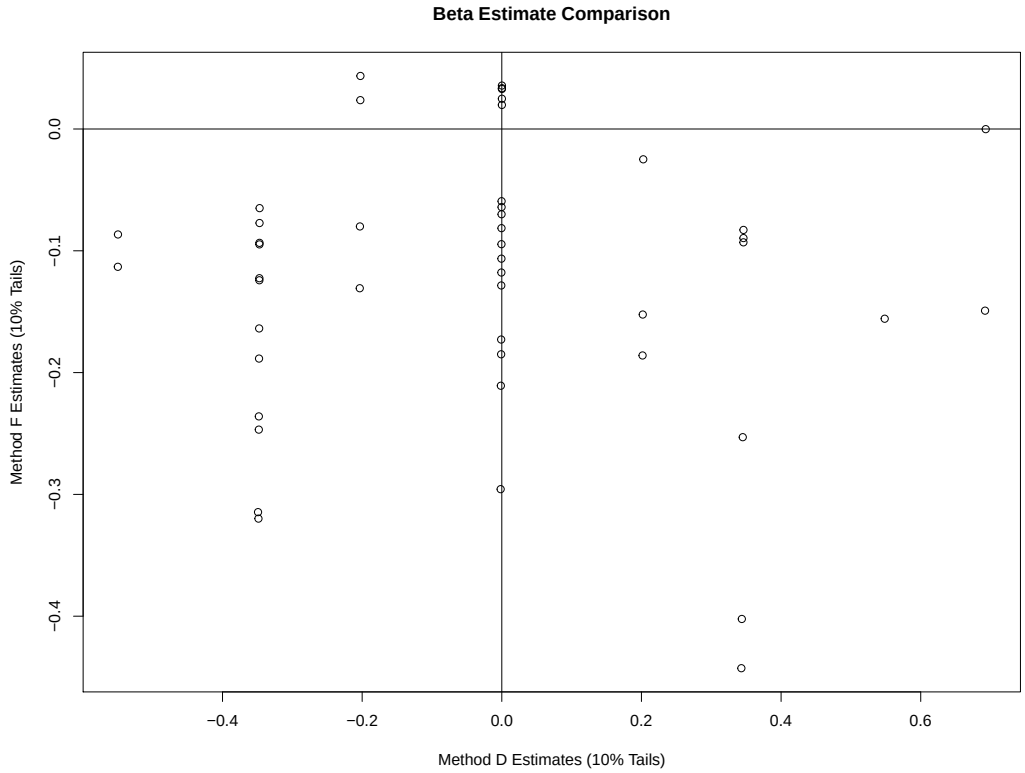


Sorry for the bad axis labels. This particular graph actually included some estimates that were positive (there were one or two between p-values 0.6 and 1.0), but I flipped them over the x-axis, and they fit right along with this curve. I also excluded one or two outliers. I have no idea what's going on here. I became briefly obsessed with trying to fit this curve, thinking that if I found the y-intercept, I would have a really good estimate of β_{201}. It turns out that it's not a very easy curve to fit, so I sort of gave up on that eventually.

Anyway, all the points with $p < 0.5$ had the right sign, so this method is at least pretty good at detecting that.

3.3 D-F Comparison

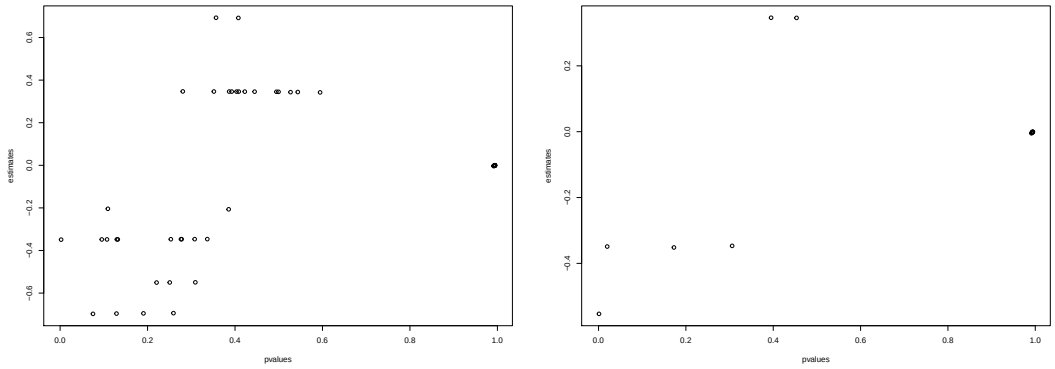
This is the next natural comparison to make, since F is a variation on D.



The comparison here shows that in general, Method F gives more accurate sign detection than Method D. The p-values are similar, being uniformly slightly lower for Method F, but I don't have a graph of that. I decided, based on this graph and the similar p-values, that Method F was better than Method D for detecting the sign of the effect, and rejected D in favor of F.

3.4 Tail Size

Deciding on the tail size is a trade-off between statistical power and the ability to actually come up with a result that isn't rejected based on the criteria listed under the description of Method D.

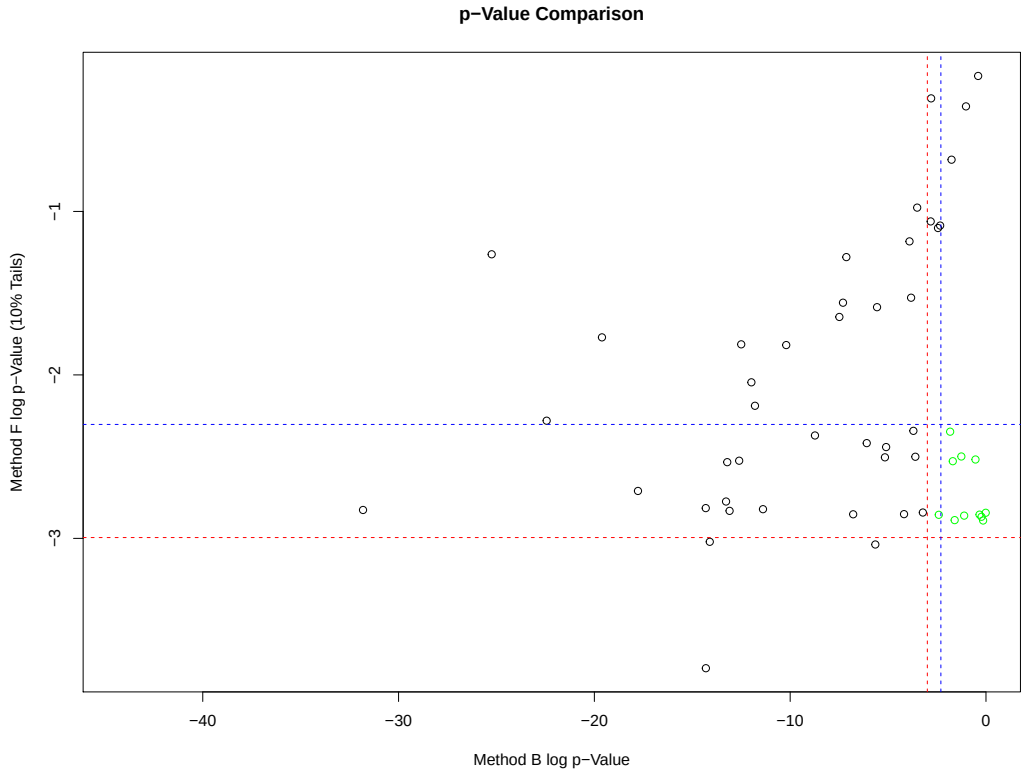


In both of the above graphs, 100 tests were done, and only those results that met the aforementioned criteria were kept. The one on the left used 10% tails, and the one on the right used 5% tails. Because the spread of p-values were pretty similar, and for sufficiently low p-values, the estimates were the same, I decided that using 10% tails was better than using 5%, since we could get meaningful estimates way more often. I didn't look into using even bigger tails, but I wonder how big we would have to make the tails before this became ineffective.

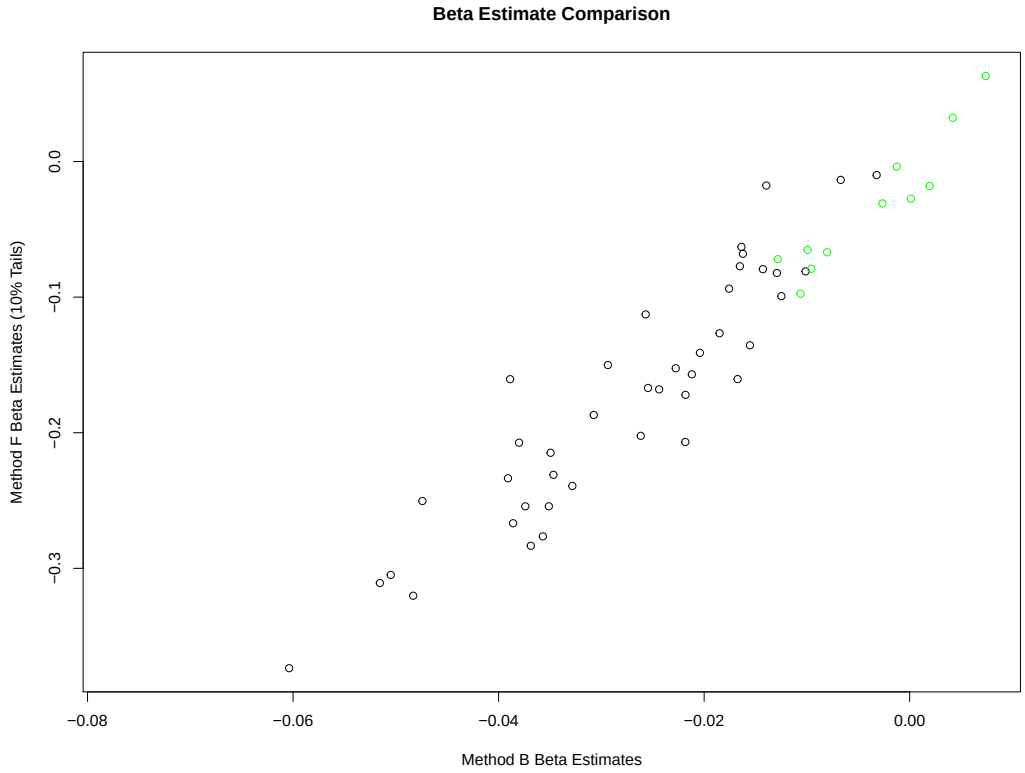
3.5 B-F Comparison

I decided that B was easier to use than E, and F gave better results than D, so the last thing to do was to compare B and F.

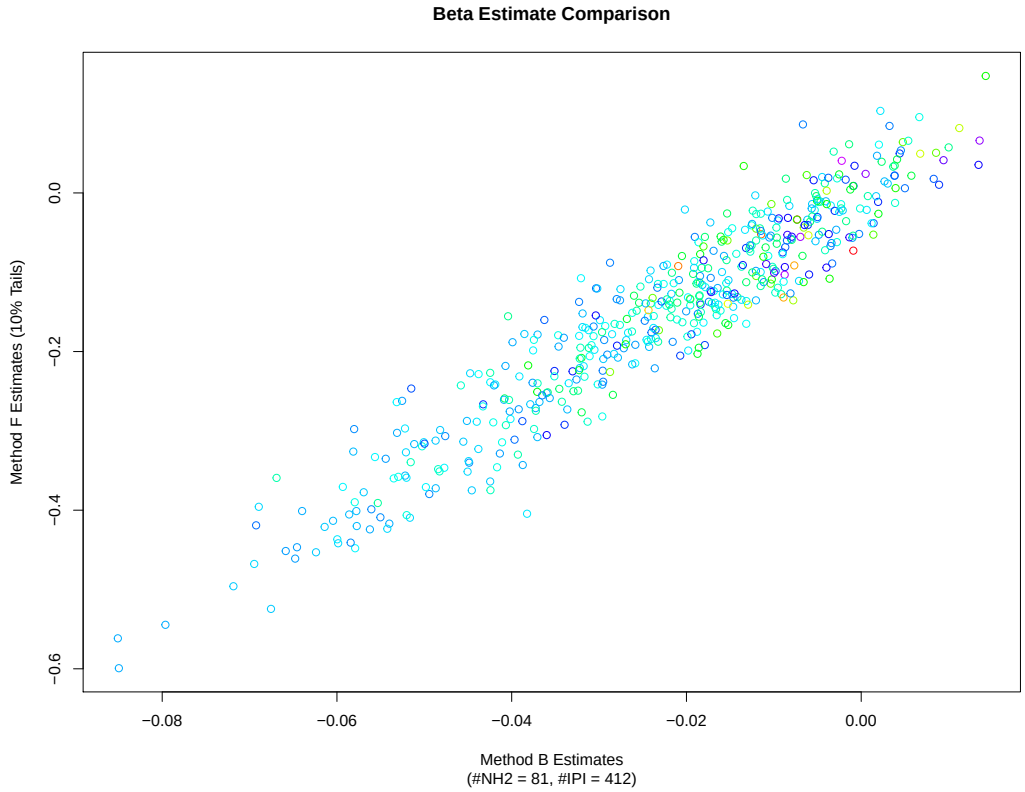
First, I tried 100 tests to compare B to F and to look at their power for detecting effects.



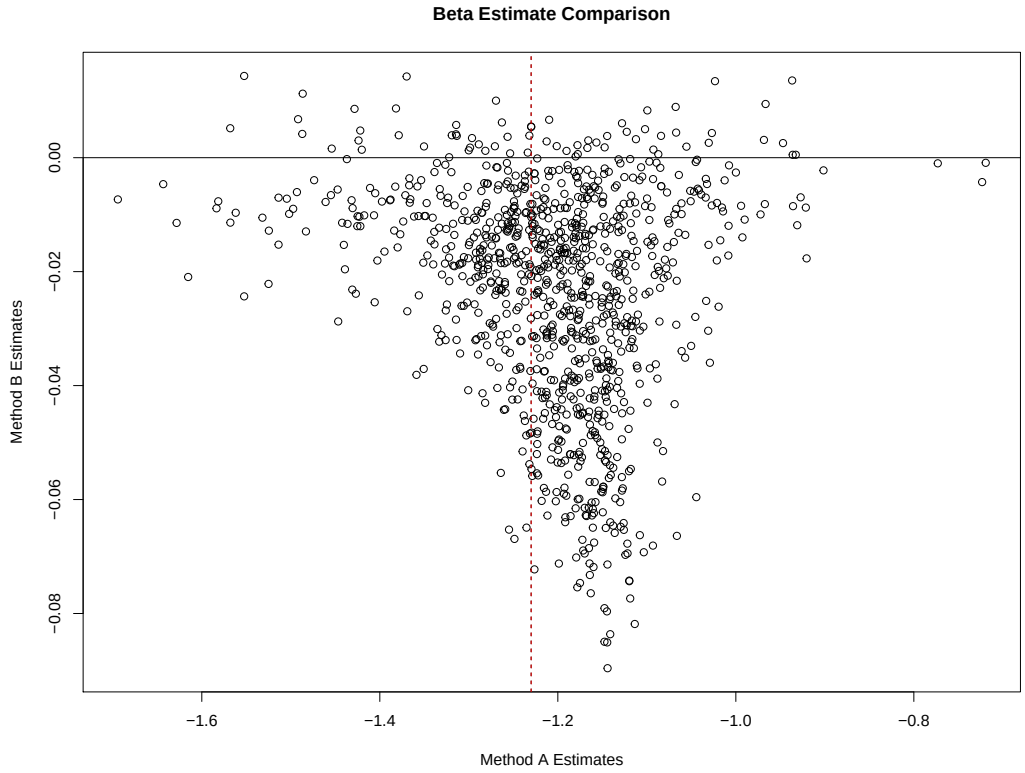
In the above graph, the p-values for Methods B and F are compared for the tests for which Method F produced a meaningful result. The blue lines represent a p-value of $p = 0.10$, and the red lines represent $p = 0.05$. The green highlighted points are those for which (given *some* p-value) Method F may detect while Method B does not (that is, $p_F < p_B$.



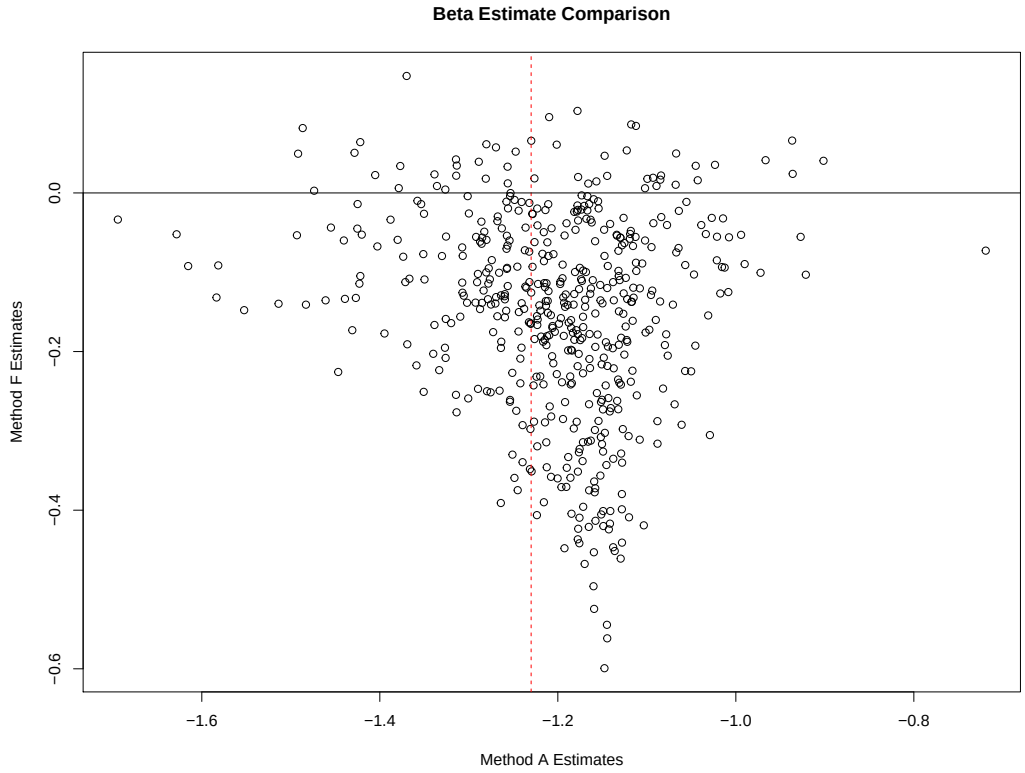
I plotted the actual estimates to see where these green points fell. While not detected at $p = 0.05$, many of these green points were detected by Method F at $p = 0.10$. It seems that when using Method F, in order to have a significant rate of effect detection, we need to adjust the p-value such that we end up with many false sign errors. Nevertheless, it's interesting to note that estimates from B and F do correlate linearly, and an extreme estimate from Method F corresponds to an extreme estimate from Method B, which in turn has a low p-value. It is possible that Method F is more useful than its reported p-values lead us to believe.



Above is a plot of the estimates from Methods B and F over 1000 simulations. The colors correspond to the estimates provided by Method A with reds and purples being the most extreme estimates. It's a bit puzzling at first why more extreme estimates from A on *both* sides correspond to closer-to-zero estimates for Methods B and F, but this is likely because these extreme estimates come from less favorably distributed data.



Here, I plotted just the B estimates against the A estimates. The shape is a bit confusing, since the red line is the true value of β_{201}. I'm not sure why we get the lowest estimates (and hence, best p-values) when A decides to estimate a little above the true effect size.



This is F vs A. Fewer points because a lot of them got rejected, but still the same shape, which is to be expected, as B and F are pretty linear.

4

Ultimately, I chose not to consider the method where we simulate cases and controls by coin-flipping from $\widehat{PRS}$ because every individual has such a similar estimated score that it wouldn't give a detectable effect unless we had an obscene number of individuals.