1 The Case for a "From Scratch" Experimental Design

Hi team, after looking at this piazza post: `https://piazza.com/class/im6wk9z189a2ha?cid=17` (Historical Data Significance) I have a new interpretation of the task we're facing here. According to the response from the professor, it sounds like he *is changing* the model from the historical data. So the historical data is useful only for identifying what terms may be significant in the future and maybe not for determining coefficient values.

So, armed with this new information, I think we should revisit the idea of generating our own experimental design matrix from scratch. As such, I have exported 3 csv files (as *scratch_X.csv*) in the "experiments" folder of hw4. Each one of these is a single run of the optFederov optimizing either the D, A or I criterion (see Table 1).

Table 1: Criterion for Designs

	Name	File	D	A	I
1	Scratch-A	scratch_A.csv	0.1230	11.6048	NA
2	Scratch-D	scratch_D.csv	0.1311	12.6859	NA
3	Scratch-I	scratch_I.csv	0.1290	13.3875	40.4978
4	Non-Augmented (Histdat)	NA	0.1442	10.7394	32.1914
5	Augmented (Histdat+15)	augmented_no_inter.csv	0.1443	10.7378	32.1764
6	Non-Augmented (Interact) (Histdat)	NA	0.0208	119.8062	490.7088
7	Augmented (Interact) (Histdat+15)	augmented_with_inter.csv	0.0208	119.6222	489.6645
8	Scratch-Interact-D	scratch_interact_1se.csv	0.0217	248.7593	NA

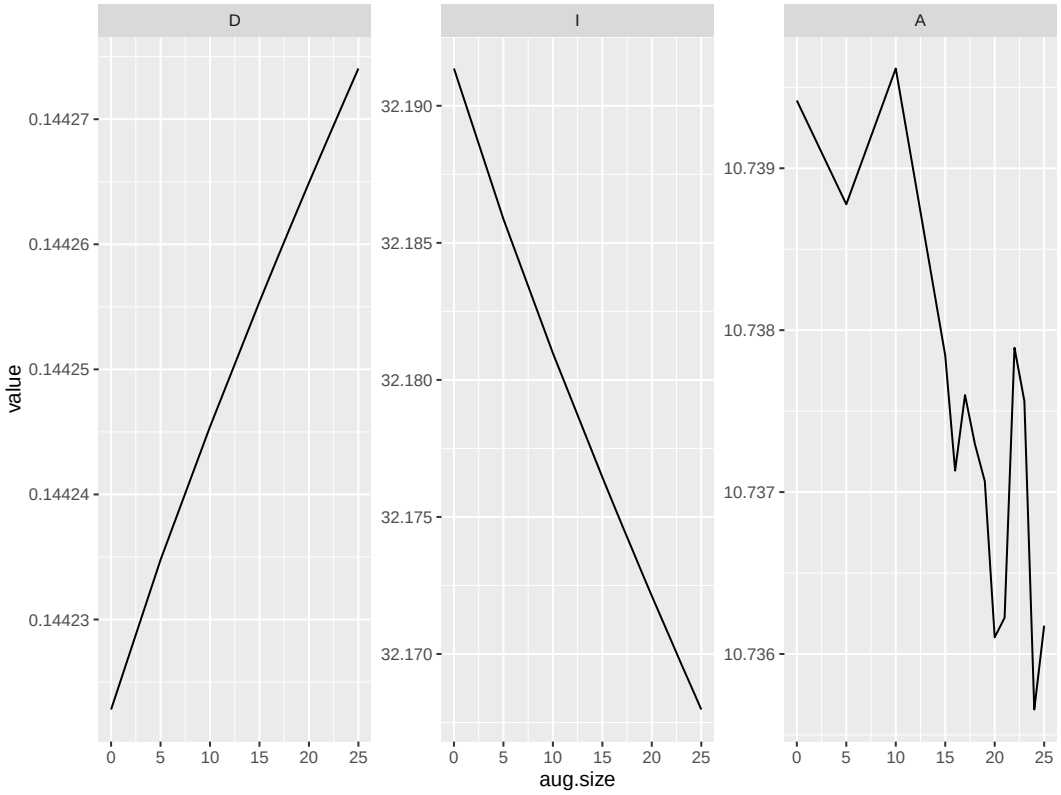
2 Augmented Experimental Design

Additionally, if we decide that we don't agree with this new interpretation, I have also generated two sets of augmented experiments. Each of these contains 15 additional experiments. However, one only looks at linear combinations (*augment_no_inter.csv*) and the other looks at interactions between every element (*augment_with_inter.csv*). While Figures 1 and 2 on the next page and on page 3 show that the criterion improve as experiments are added, the magnitude is quite small. Indeed, comparing the magnitudes in Table 1 between the "from scratch" (first three rows) to the "augmented experiments" (last two) indicates that the additional experiments doesn't move the needle too much.

3 Other Ideas

Given that we want to use the historical data in some capacity, I wanted to explore the idea of using our priors about which variables interact in designing an experiment. To that end, I took the significant (non-zero) coefficients from the glmnet run and spat out the terms that interacted.

Figure 1: "Augmented Experiment Criterion (No Interactions)"



3a lambda.1se

For the "1se" selection rule, the following coefficients were non-zero:

```
[1] "V15:V23" "V16:V23" "V14:V72" "V26:V72" "V41:V55" "V41:V72" "V55:V72" "V55:V84" "V55:V94"
[10] "V63:V72" "V72:V84" "V72:V85" "V72:V92" "V72:V94"
```

After post-processing, this yields following formula equation for optFederov:

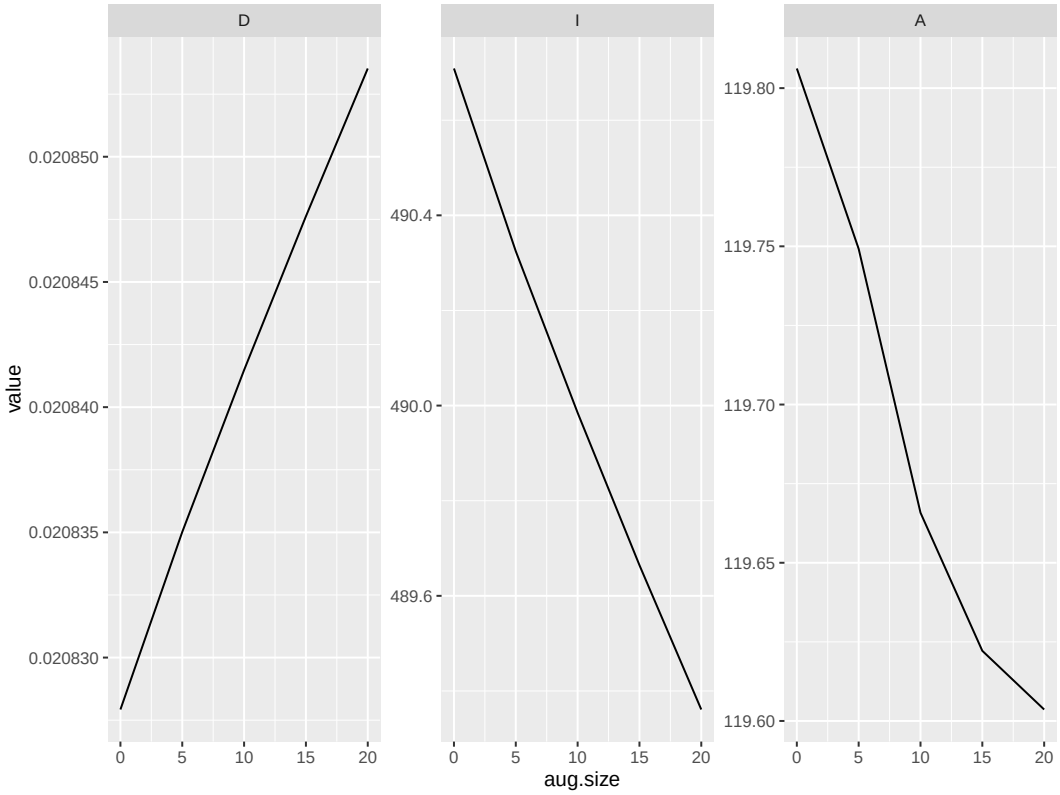
```
~. + V1:V2 + V1:V7 + V2:V7 + V4:V5 + V4:V7 + V5:V7 + V5:V8 +
    V5:V9 + V6:V7 + V7:V8 + V7:V9
```

Using this, I then calculated the minimum number of experiments to run = 131 + 5 (for safety margin). I'll run an optFederov overnight to see if I can get something out, but it's going to take a non-trivial amount of time. However, is far too many experiments for us to stomach – costing over $26k.

3b lambda.min

For the "min" selection rule:

Figure 2: "Augmented Experiment Criterion (With Interactions)"




```
[1] "V12:V22" "V14:V22" "V11:V23" "V13:V23" "V14:V23" "V15:V23" "V16:V23" "V15:V25"
[9] "V15:V26" "V14:V55" "V15:V55" "V14:V63" "V14:V72" "V15:V72" "V15:V84" "V22:V72"
[17] "V22:V85" "V33:V41" "V33:V55" "V33:V72" "V33:V85" "V41:V55" "V41:V72" "V55:V63"
[25] "V55:V72" "V55:V84" "V55:V85" "V55:V94" "V61:V72" "V63:V72" "V63:V84" "V61:V94"
[33] "V72:V81" "V72:V84" "V72:V85" "V72:V92" "V72:V94" "V85:V92"
```

Converted to a formula, yields:

```
~. + V1:V2 + V1:V5 + V1:V6 + V1:V7 + V1:V8 + V2:V7 + V2:V8 +
  V3:V4 + V3:V5 + V3:V7 + V3:V8 + V4:V5 + V4:V7 + V5:V6 + V5:V7 +
  V5:V8 + V5:V9 + V6:V7 + V6:V8 + V6:V9 + V7:V8 + V7:V9 + V8:V9
```

And, this requires 287 + 5 (for margin) experiments to cover the space. This one is too infeasible so I'm not going to run it.

3c lambda.1se with a twist

I thought about taking the most significant two interaction terms and using those to generate experiments. However, this cause optFederov to fail with "Singular design" errors, which is egregious

and frustrating at 3 in the morning. The number of required experiments is still super high, on the order of 60-80, which is still too expensive to run. Right now I'm running the top two Positive interaction terms (the top magnitude ones were giving me singular design errors).

4 Predicted Quantile Sampling

A totally random idea (that I realize is terrible) that actually uses the historical data is to take our predicted click probabilities and sample them at quantiles. That's in *sampled_quantile.csv*. Basically this equates to sample the ones we think are good, and some of the bad ones as well. It totally discounts any of the structure we'd like to use in choosing good experiments and I cannot in good conscience suggest it.

5 Conclusion/Recommendation

Given the instructors comments on the aforementioned piazza post, I feel that the augmented experimental design will not yield the additional information we were hoping to get. Additionally, identifying the interaction terms yields an experimental design that is too costly to implement. So unfortunately, that doesn't leave us with much use for the historical data other than knowing the levels of each column. Maybe Section 3c on page 3 will yield something useful.