Confouding adjustment

Michael Cork

1/20/2022

Data generating mechanism

We evaluate the performance of our methods for fitting ERC curves in a variety of data settings. In each setting, we generate six confounders $(C_1, C_2, ..., C_6)$, which include a combination of continuous and categorical variables.

$$C_1, ..., C_4 \sim N(0, I_4), C_5 \sim V\ \{-2, 2\}, C_6 \sim U\ (-3, 3)$$

Where $N(0, I_4)$ denotes a multivariate normal distribution, $V\ \{-2, 2\}$ denotes a discrete uniform distribution, and $U(3, 3)$ denotes a continuous uniform distribution. We generate the exposure based on six different specifications of the relationship between confounding and exposure based on work by Xiao.

The exposure E is generated using six different specifications that rely on the function $\gamma(\mathbf{C}) = -0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)$ C. Specifically,

1. $E = 9 \times \gamma(\mathbf{C}) + 17 + N(0, 5)$;
2. $E = 15 \times \gamma(\mathbf{C}) + 22 + T(2)$;
3. $E = 9 \times \gamma(\mathbf{C}) + 3/2C_3^2 + 15 + N(0, 5)$
4. $E = 49 \times \frac{\exp(\gamma(\mathbf{C}))}{1+\exp(\gamma(\mathbf{C}))} - 6 + N(0, 5)$;
5. $E = 42 \times \frac{1}{1+\exp(\gamma(\mathbf{C})\})} - 18 + N(0, 5)$;
6. $E = 7 \times \log(\gamma(\mathbf{C})) + 13 + N(0, 4)$;

Scenario 1 has a linear relationship between E and C and GPS values are not heavy tailed. In scenario 2, the relationship between E and C stays linear, but GPS values are heavy tailed and include extreme values.The relationship between E and C in scenarios 3-6 are all non-linear, but do not have extreme values like scenario 2.

One additional thing to note is that for scenario 6, in actuality it is using $W = 7 \times \log(|\gamma(\mathbf{C})|) + 13 + N(0, 4)$. I then ran through three different iterations of the sample size (N = 200, 1000, 5000) and then fit the following outcome model:

$$Y|E, C \sim N(\mu(E, C), 10^2)$$
$$\mu(E, C) = 20 + 0.1 * E - (2, 2, 3, -1, 2, 2) * C$$

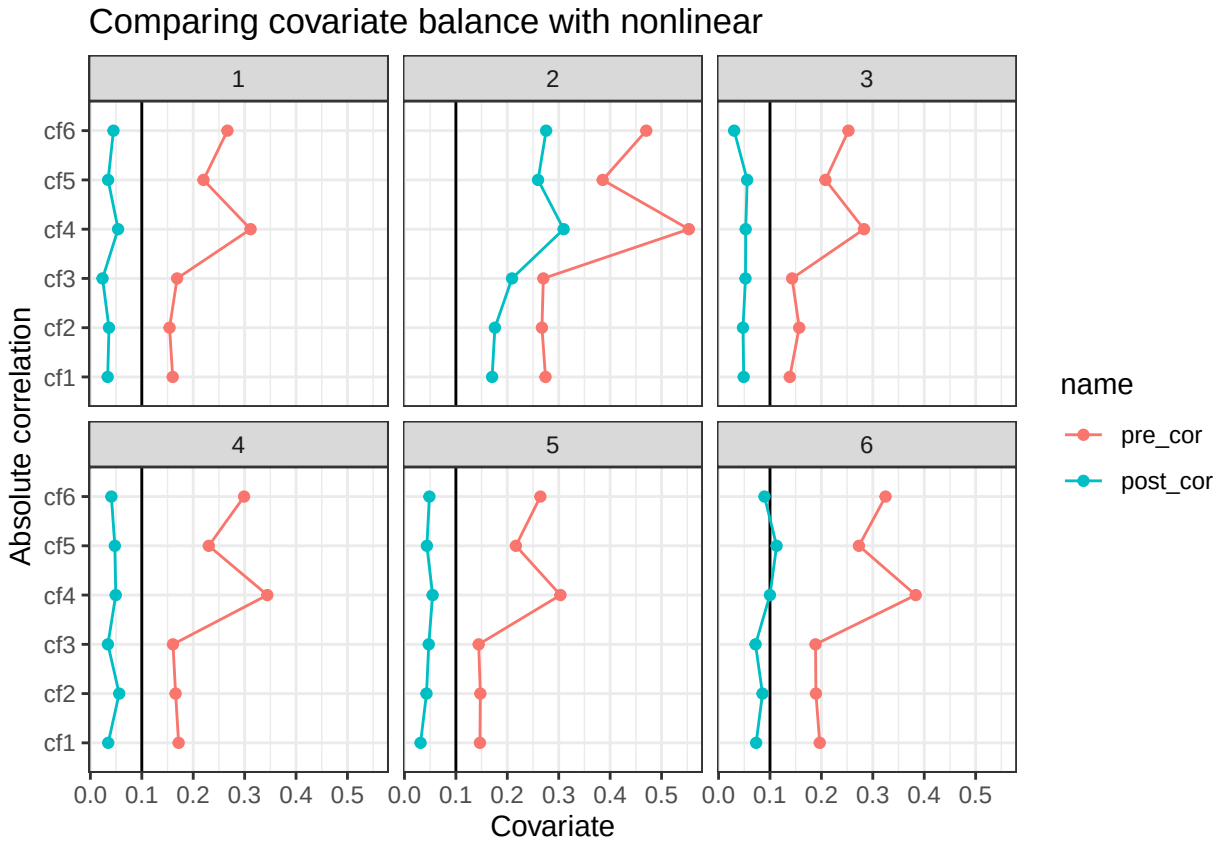
Before fitting both a linear and GAM model mimic a linear and nonlinear exposure-response curve, I wanted to investigate how a linear function of confounders fits the exposure data. First I do not include any random components, so we would expect that a linear model for scenario 1 & 2 would perfectly fit the data and indeed that is what we see:

Now we include back the noise terms to see how adjusting for confounders linearly does in terms of R-square value:

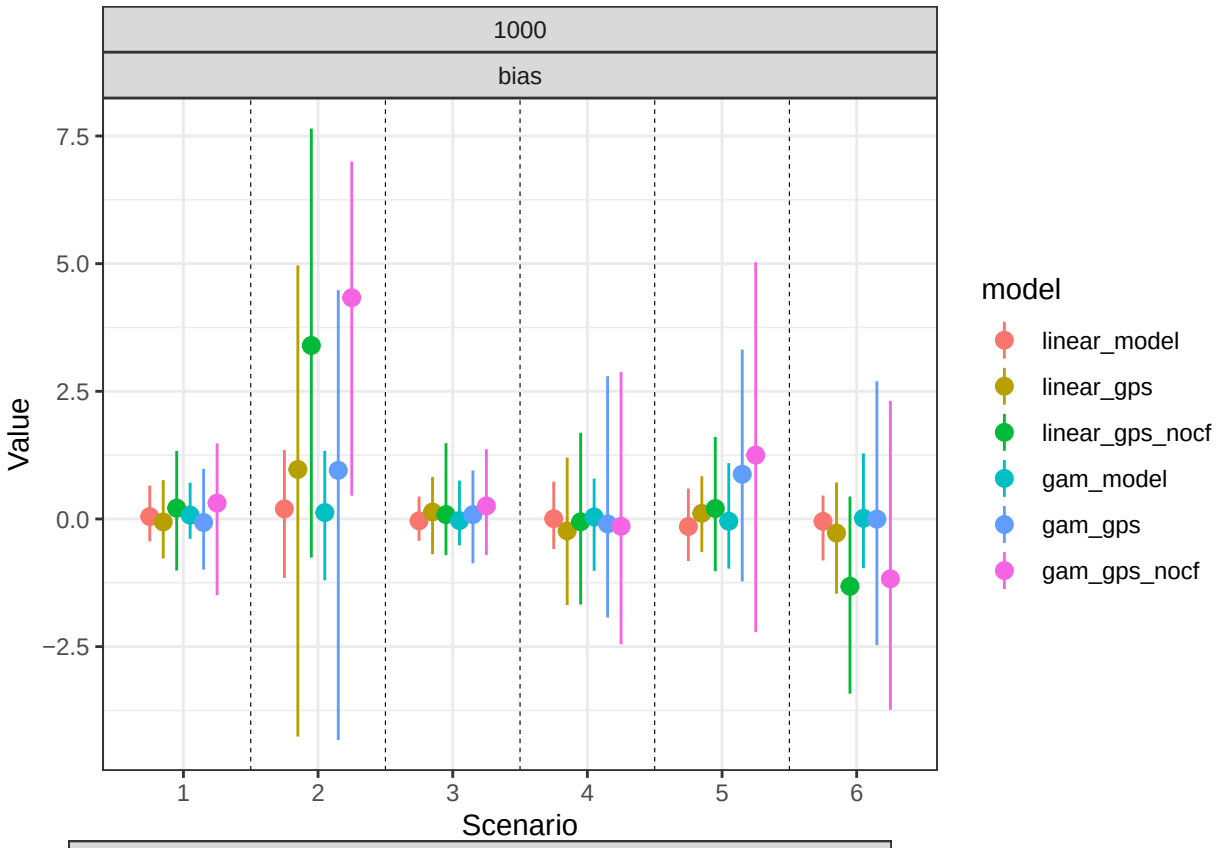
Bias with (previously without) linear adjustment for confounders

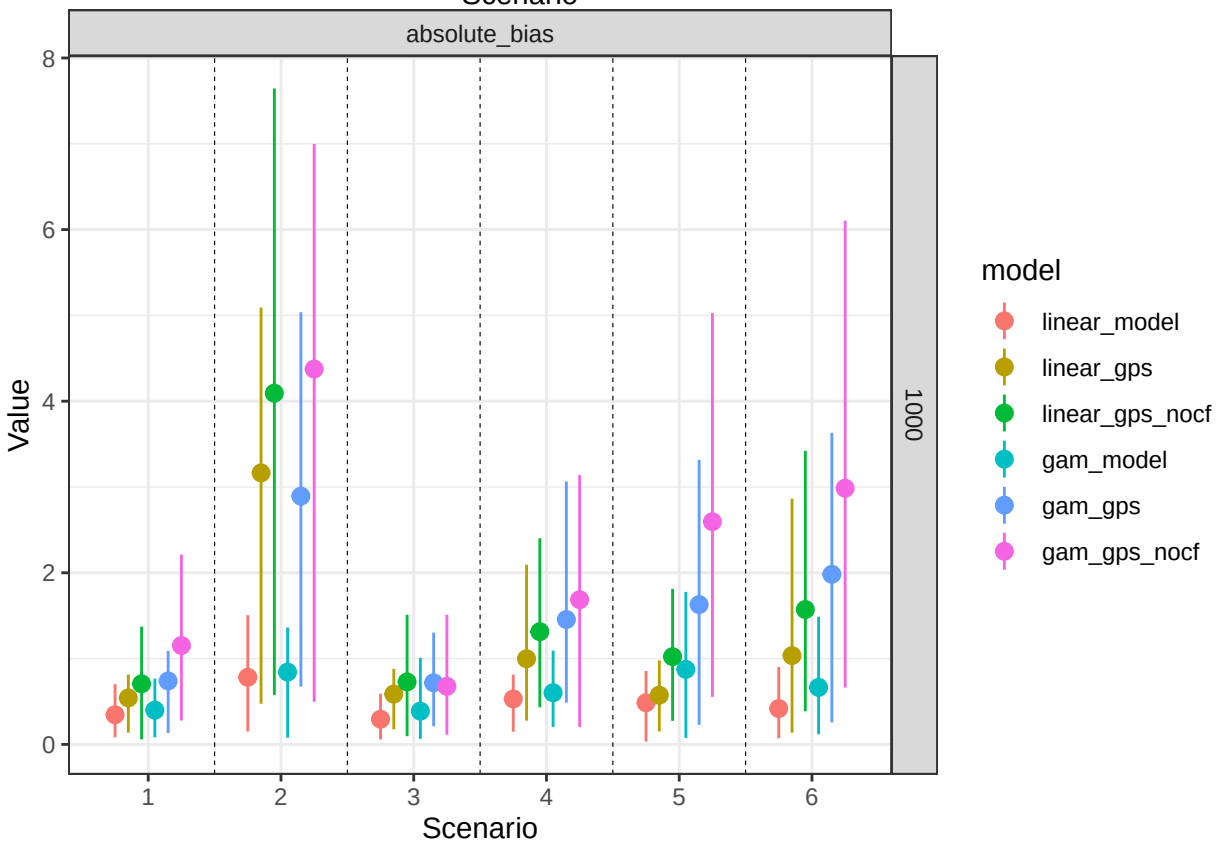
I first ran through each model to see the bias without adjustment for confounders to see the baseline of how biased the results might be without proper adjustment.

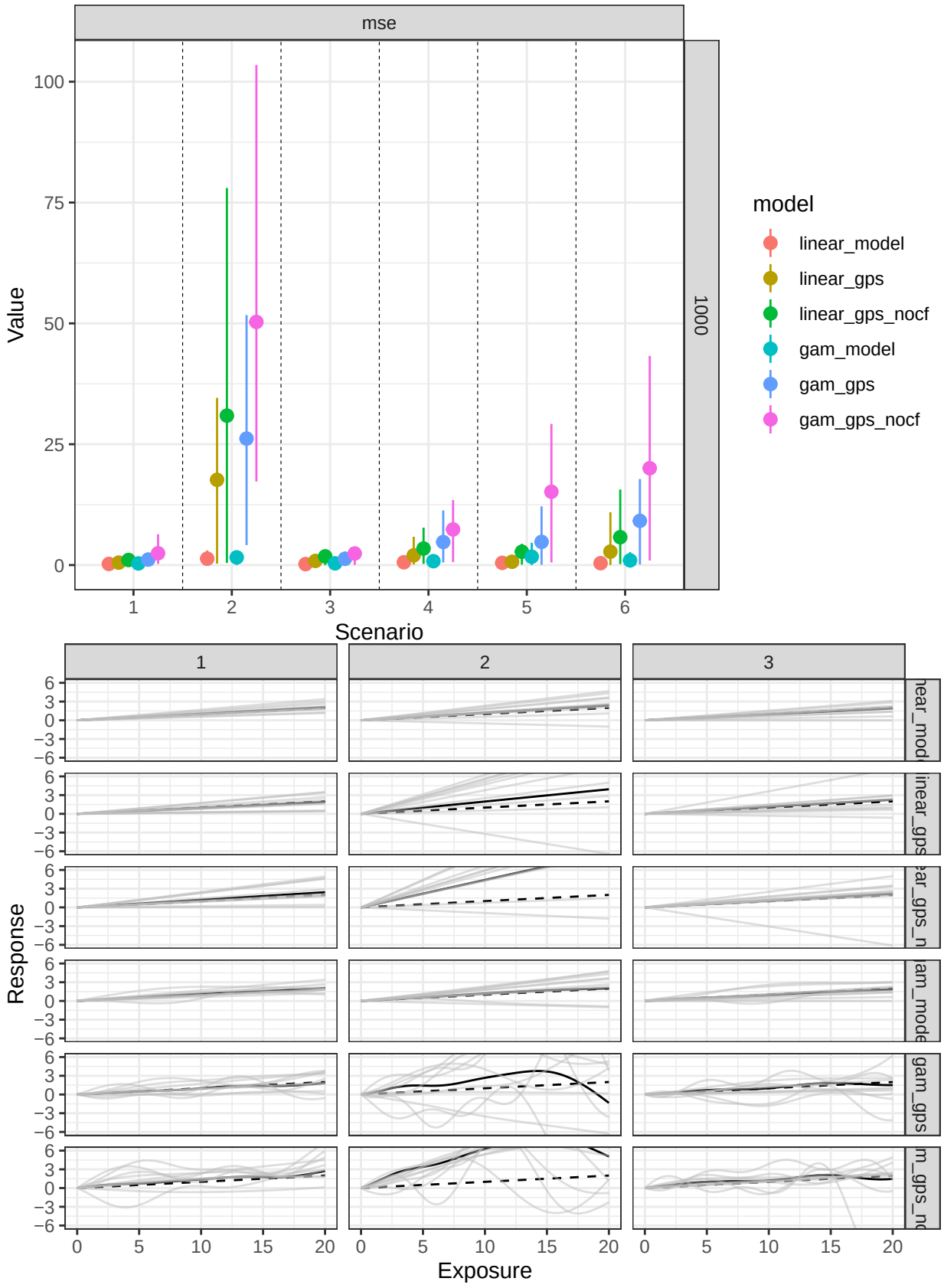
```
## `summarise()` has grouped output by 'gps_mod'. You can override using the `.groups` argument.
```

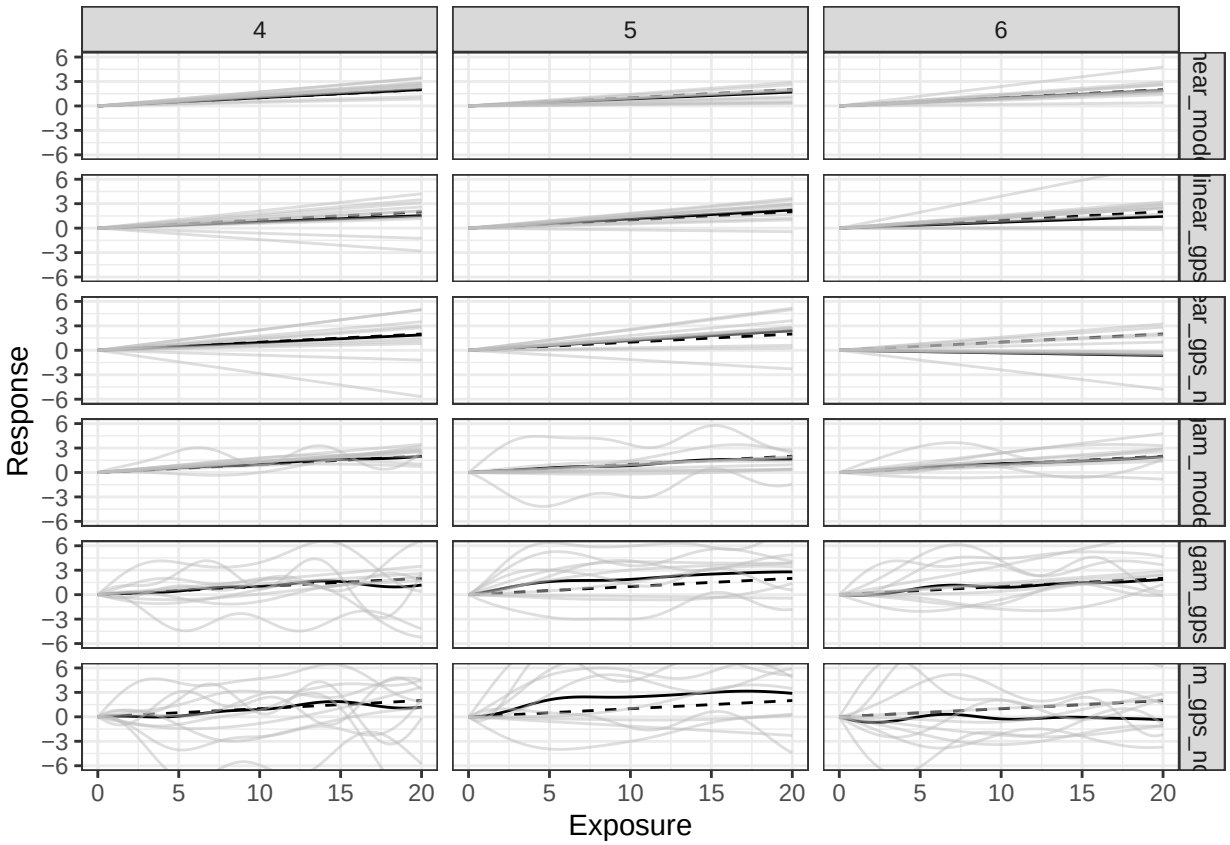



```
## `summarise()` has grouped output by 'model', 'gps_mod', 'sample_size'. You can override using the `.g
```


1000
bias
7.5
5.0
2.5
0.0
−2.5
Value
1
2
3
4
5
6
Scenario
model
linear_model
linear_gps
linear_gps_nocf
gam_model
gam_gps
gam_gps_nocf


absolute_bias
8
6
4
2
0
Value
1
2
3
4
5
6
Scenario
1000
model
linear_model
linear_gps
linear_gps_nocf
gam_model
gam_gps
gam_gps_nocf


mse
1000
Value
100
75
50
25
0
1
2
3
4
5
6
Scenario
model
linear_model
linear_gps
linear_gps_nocf
gam_model
gam_gps
gam_gps_nocf
1
2
3
Response
Exposure
0
5
10
15
20



Now do this with nonlinear model (and see how they each do)

With adjustement for confounders

Now I linearly adjust for confounders in the additive and linear model and we see improvements, especially in scenario 1 and 2. I also include a weighted propensity score analysis that uses a linear model with all interactions included to generate a propensity score. I trim extreme values for the weights (95th percentile) beofre the analysis stage. (Xgboost didn't seem to perform well)

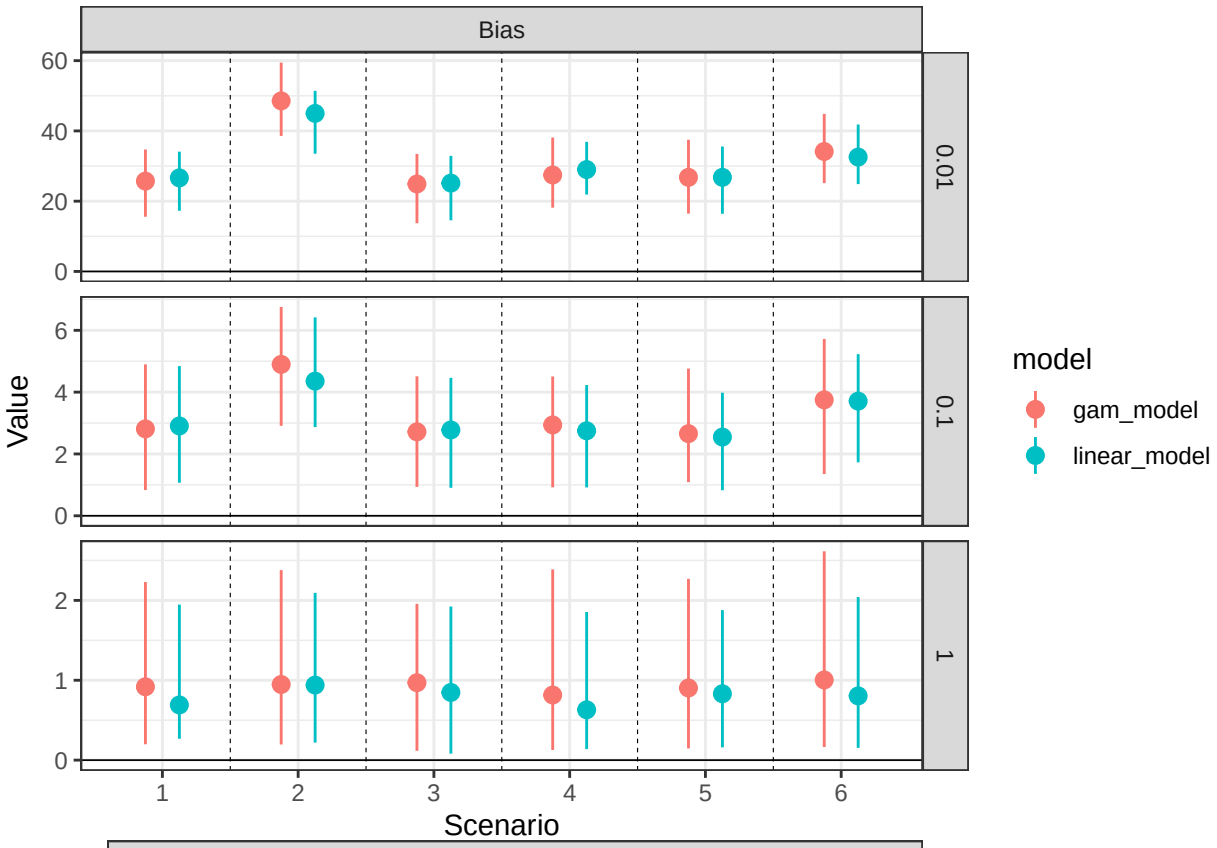
No noise term

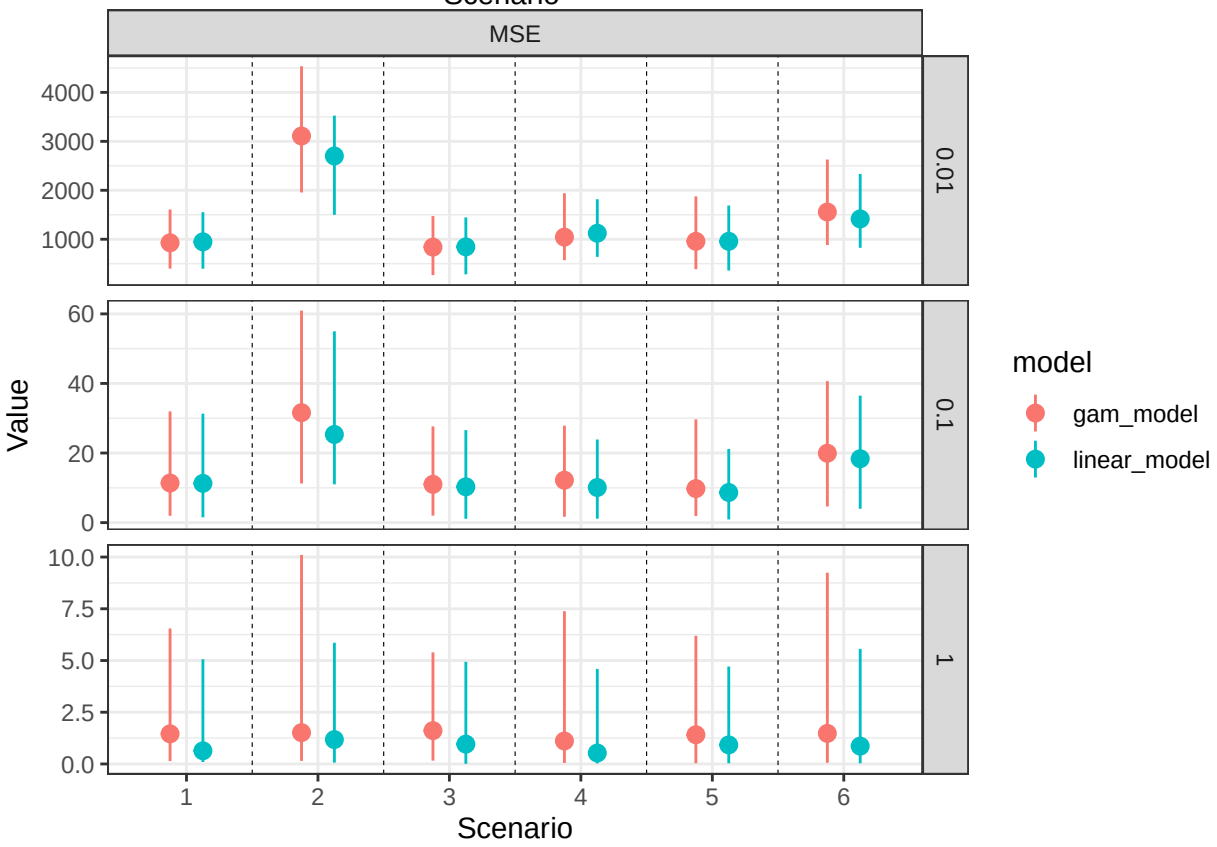
Now I fit without any noise terms included in the model

Increase effect of confounding

Now we adjust for the ratio between the standard deviation of the exposure and the standard deviation of the confounding, keeping the exposure relationship constant at 0.1. Here we don't adjust for confounders at all. Smaller values for this ratio indicate that the standard deviation of the confounding is much larger than the effect of the exposure.

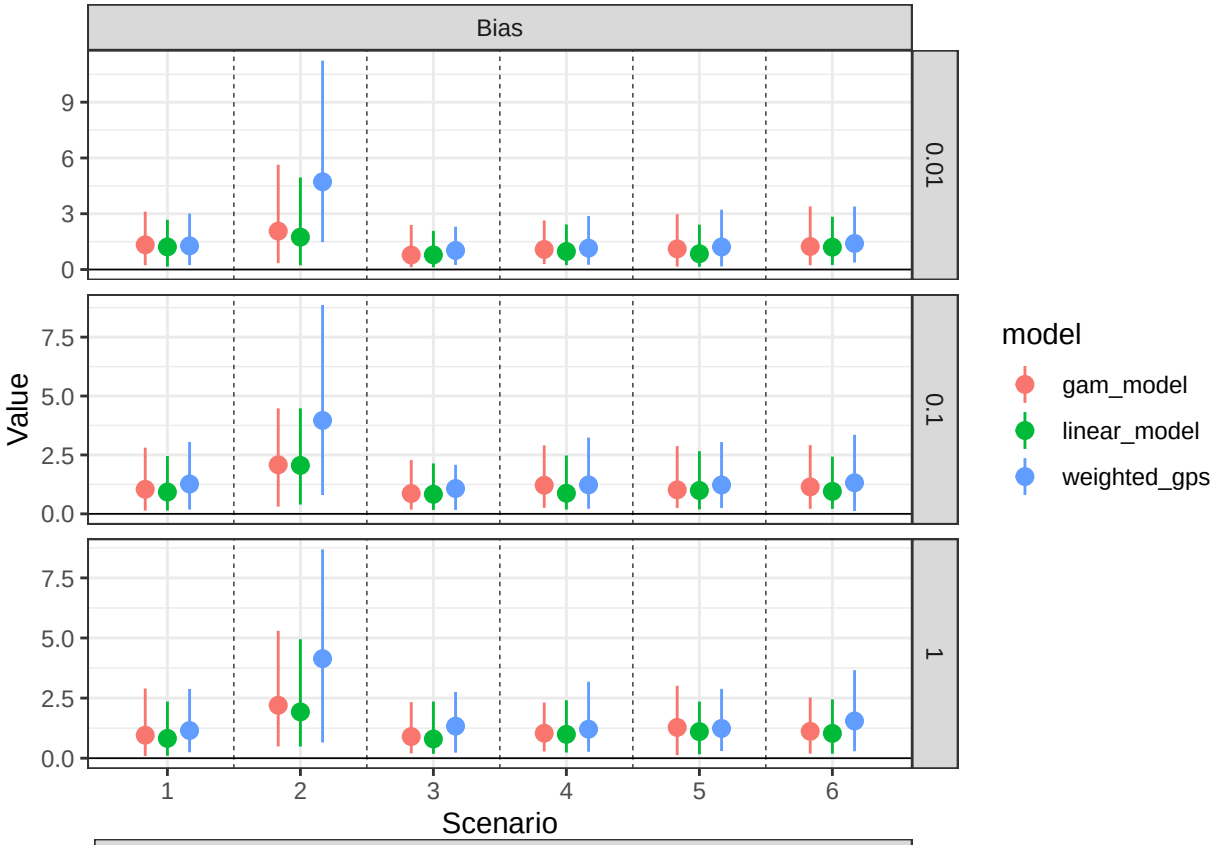
```
## `summarise()` has grouped output by 'model', 'gps_mod', 'sample_size', 'exp_conf_ratio'. You can ove
```

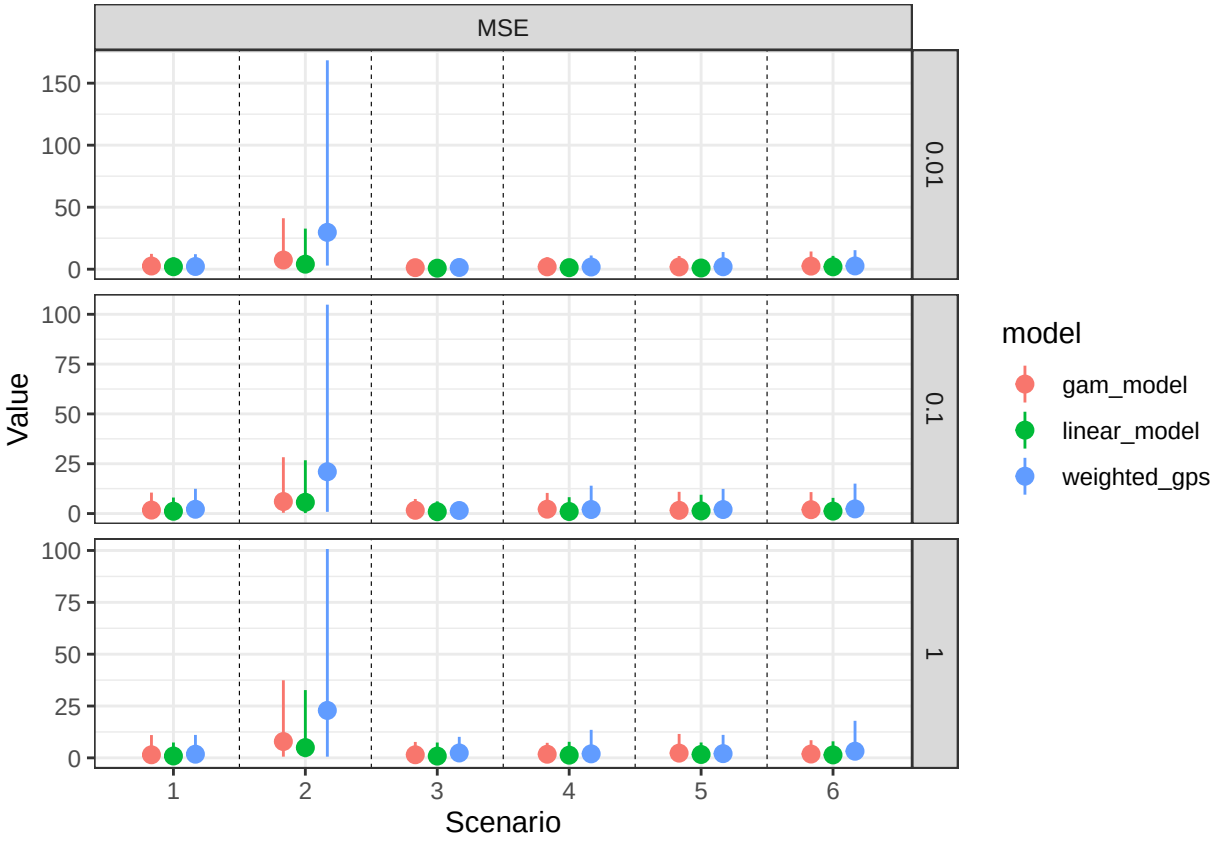




I then repeat with adjustment for linear confounding:

```
## `summarise()` has grouped output by 'model', 'gps_mod', 'sample_size', 'exp_conf_ratio'. You can ove
```





Now use GPS weighted adjustement as well:

- Add causalGPS to type of fit here, and see how it does in terms of bias compared to others (should see difference since it is correctly accounting in all cases)
- Move to Poisson
- Move forward on getting causalGPS package to work
- Add more propensity score models to analysis

different relationships: linear, Three different methods: unadjusted, linear, gam, different GPS methods metrics like RMSE and bias from a descriptive standpoint they don't tell you about the fit and how it looks. Cool to have a plot of a bunch of different fits: columns are three different exposure effects, rows are different methods to fit it, true efect and then overlay 20 estimates from the simulation. Show how well the simulation estimates relate to the true across a simulation of the sample replicates. Visualize what is happening also behond the RMSE and bias values. Later on it would be a useful plot to have. Overlay the mean the replicates to little information, unbiased in the long run the mean will look similar. If you took 10 or 20 of the simulation replicates and put them randomly, so its not quite linear but there are different fits around linear. For linear you might see lines are not exact but close. Useful to see what the models are doing in trying to fit the exposure respone function. On top of that have a table of RMSE bias and coverage for these different settings.

A couple of other metrics in there as well, the plots would be a nice supplement to the table b

Cases where GAM creates a weird biased estimate because it is strange curve, a lot of uncertainty near the extremes because there are not a lot of data near.

Be it the Bayesian causal response function, Implement to the continuous will help you make sure you are doing it right. The Bayesian one you can't transfer over right now without recreating a continuous outcome.

Keep what you have and try playing with Poisson model in new document. Dan said remember that your second year you wont make a ton of progress on research, that's just the way it goes

Even if you are generating from a linear outcome but nonlinear confounder but in the outcome model it is linear, I would guess that if you reverse the siuation where you assume linearlity between exposure and confounder but nonlinearity in the outcome model, the amount of bias from linear model would be higher

Misspecification of the outcome model will have larger consequences that confounding in the exposure model. Something interesting to look at, if you revert the situation and see which outcome is in the outcome vs exposure.

MIspecify which one and how will the consequences be for the exposure model. The quesiton is whether you need to be less in assymptopia for the outcome model vs the expsoure. Very interesting theoretical result to look at. You learn

Throw any fancer estimator that you want and you won't see big differences. However if you revert the situation and see nonlinearity in the outcome and linear in the outcome model. some methods will look better than others and eventually the double robust will work the best.

change to predict at the same points from 0 to 15 for this paper