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 four specifications of the relationship between confounding and exposure based on work by Xiao.

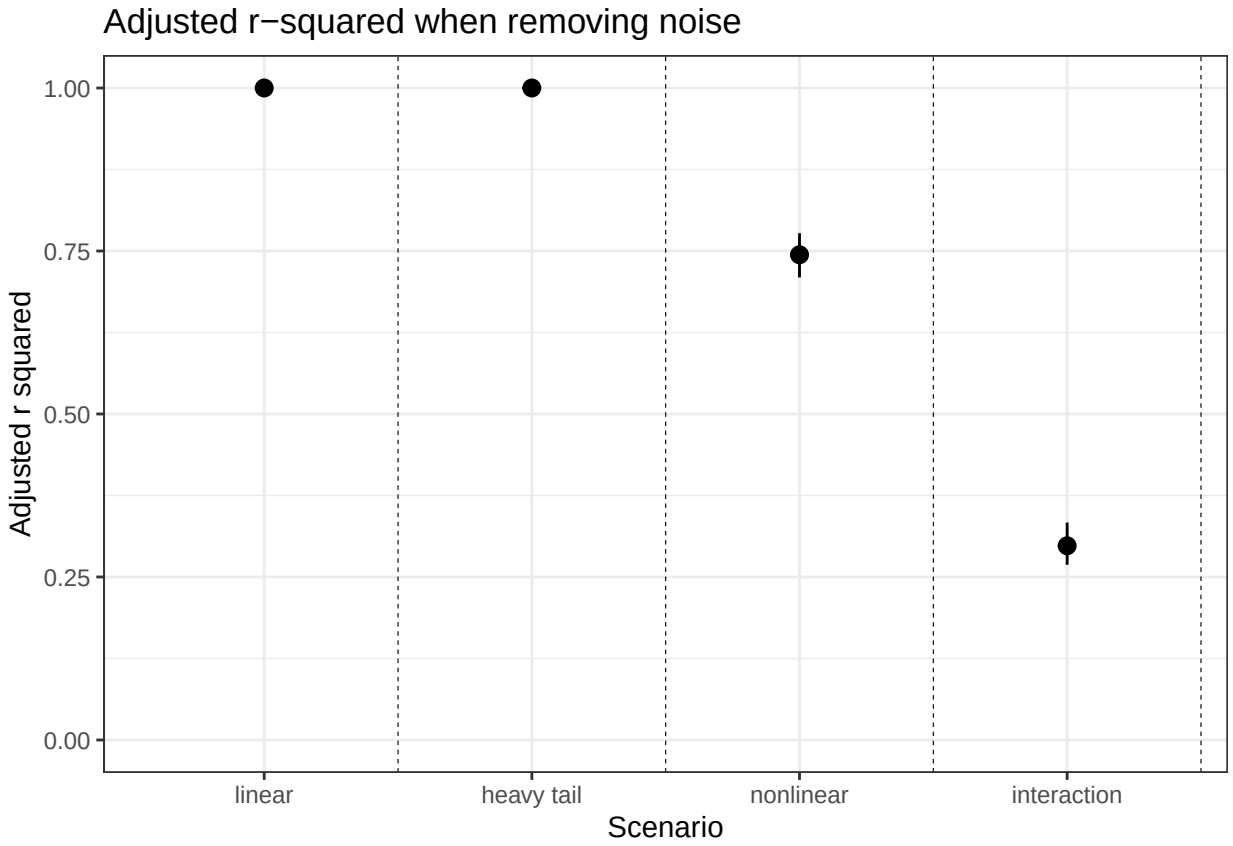
The exposure E is generated using four different specifications that rely on the function $\gamma(\mathbf{C}) = -0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)$ C. Specifically,

1. $E = 9 \times \gamma(\mathbf{C}) + N(0, 5)$; (simple linear)
2. $E = 15 \times \gamma(\mathbf{C}) + T(2)$; (heavy-tailed linear)
3. $E = 7 \times \gamma(\mathbf{C}) + C_3^2 + N(0, 5)$ (nonlinear)
4. $E = 7 \times \gamma(\mathbf{C}) + C_3^2 + C_1 C_6 + N(0, 5)$; (interaction)

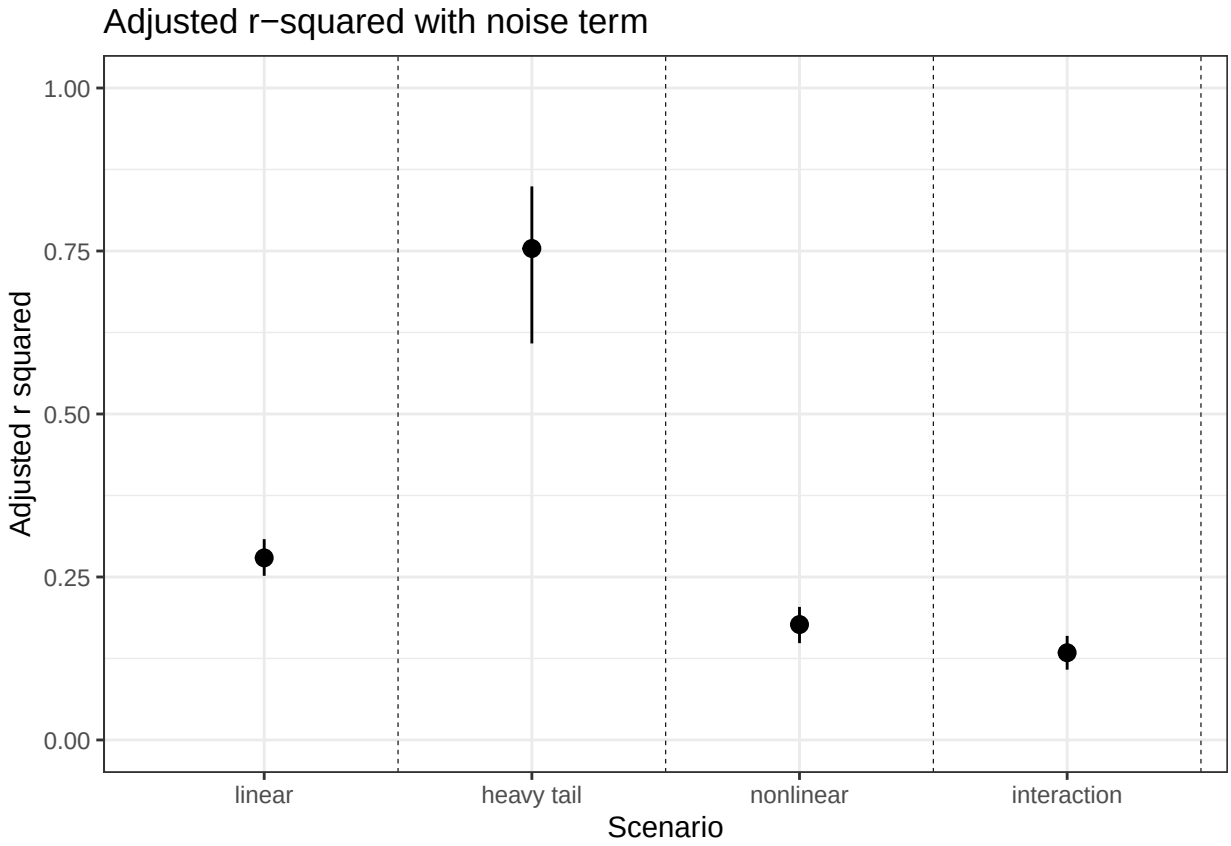
The exposure is then rescaled such that all values lie between 0 and 20, to mimic what is typically seen in air pollution data. Scenario 1 has a linear relationship between E and C and GPS values are not heavy tailed. In scenario 2, the relationship between E and C stays linear, but GPS values are heavy tailed and include extreme values. Scenario 3 include a non-linear relationship between E and C and scenario 4 adds an interaction term to the relationship between the confounder and the exposure.

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

Before fitting the model we can compare how well the covariates predict the exposure. We can first look at this without any noise, and we would expect that in scenario 1 and 2 it would perfectly capture the exposure given it is a linear function of the covariates, whereas the association would get significantly worse in scenario 3 and 4.



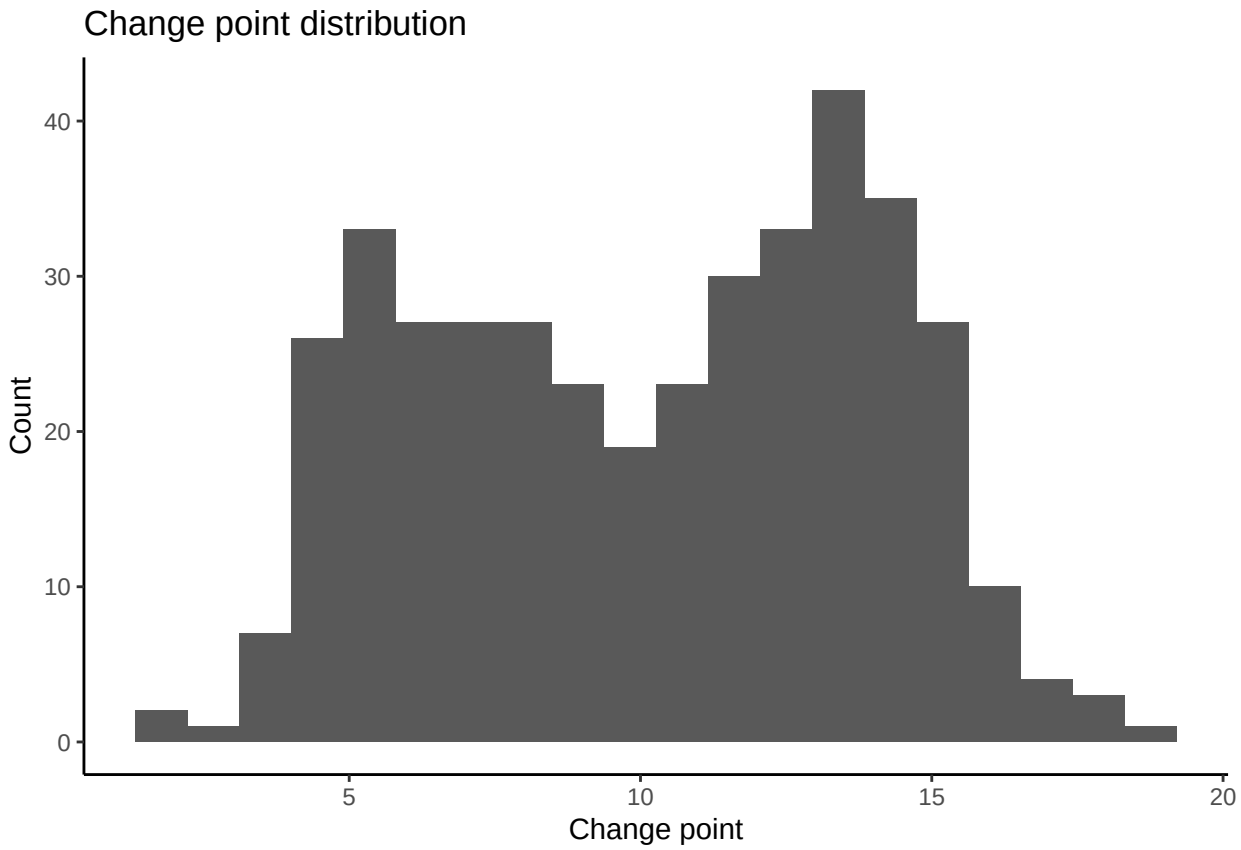
Now we do the same thing but include the noise terms into the model, so that scenario 1 and 2 are no longer perfectly described by the linear relationship between the covariates and exposure:



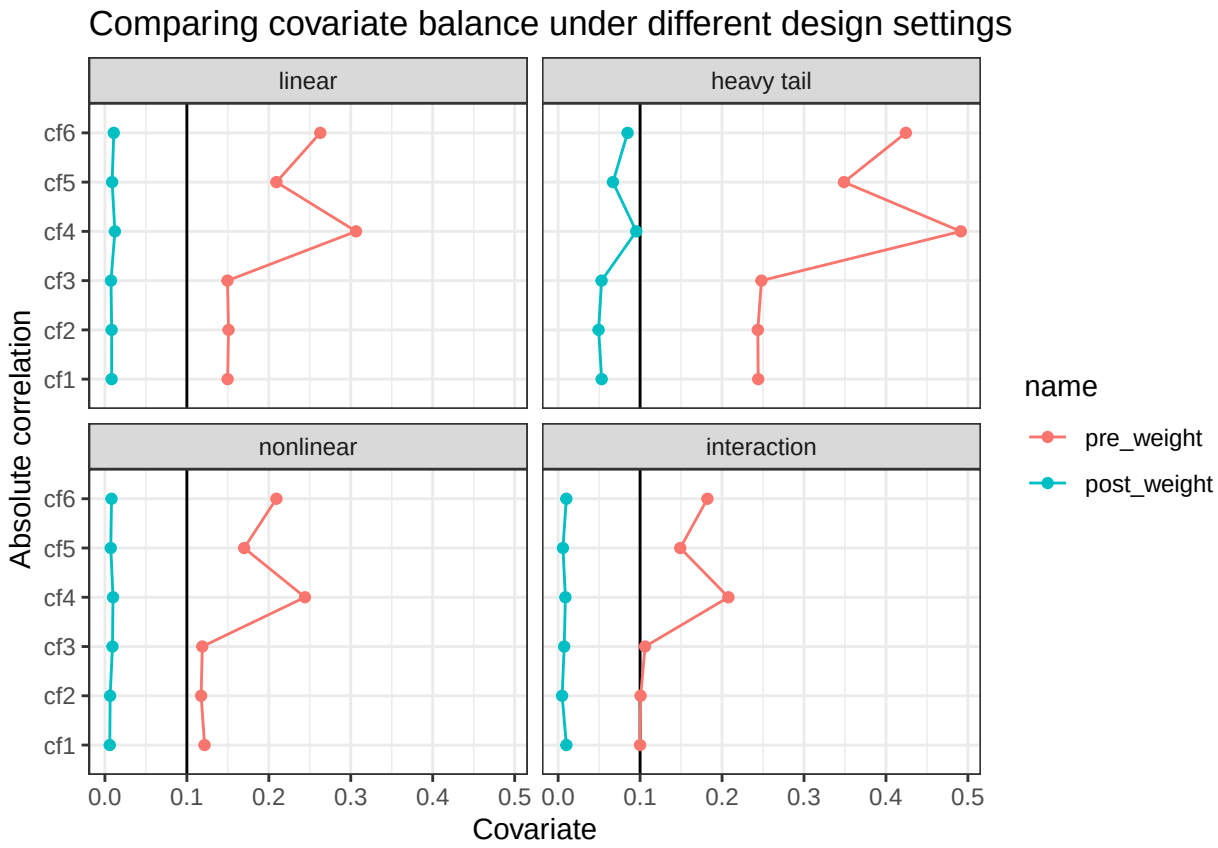
Results

We can now look at this simulation and see how several different methods perform. We compare

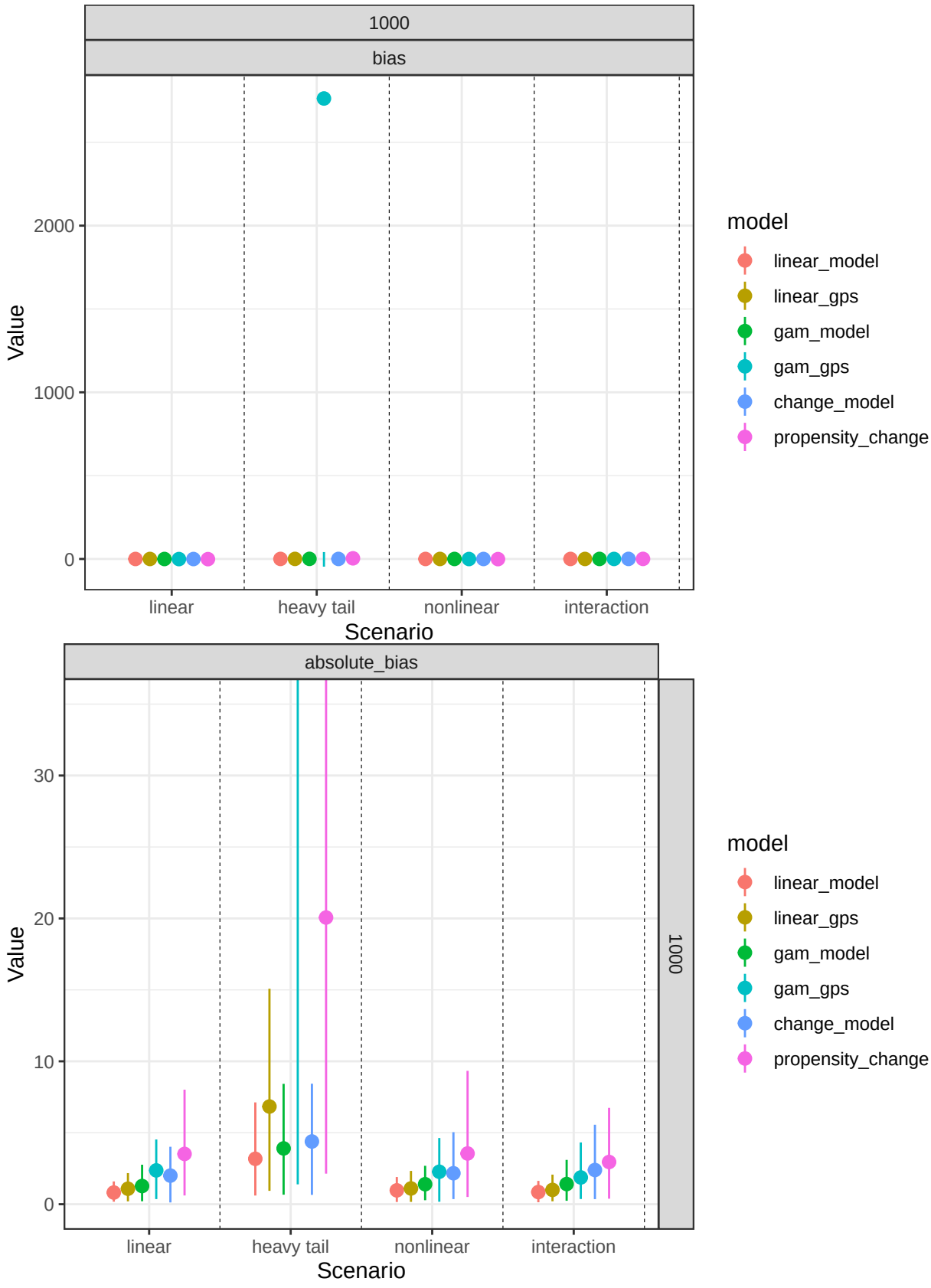
Add nonlinear case for confounders to see propensity score do better than not propensity model. Make segmented model. Make to see how many cover 0 with correct threshold.

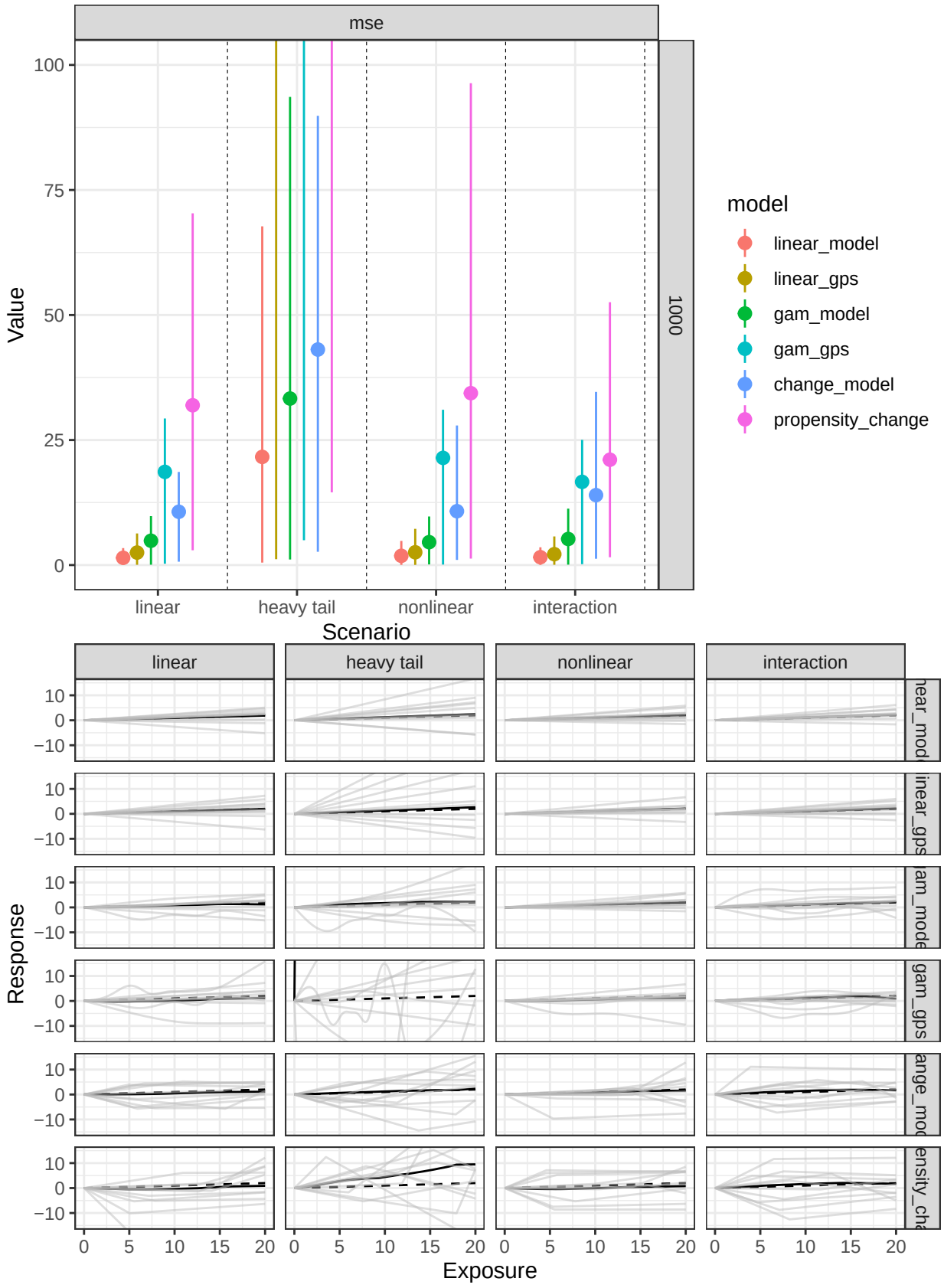



```
## `summarise()` has grouped output by 'gps_mod'. You can override using the
## `.groups` argument.
```




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

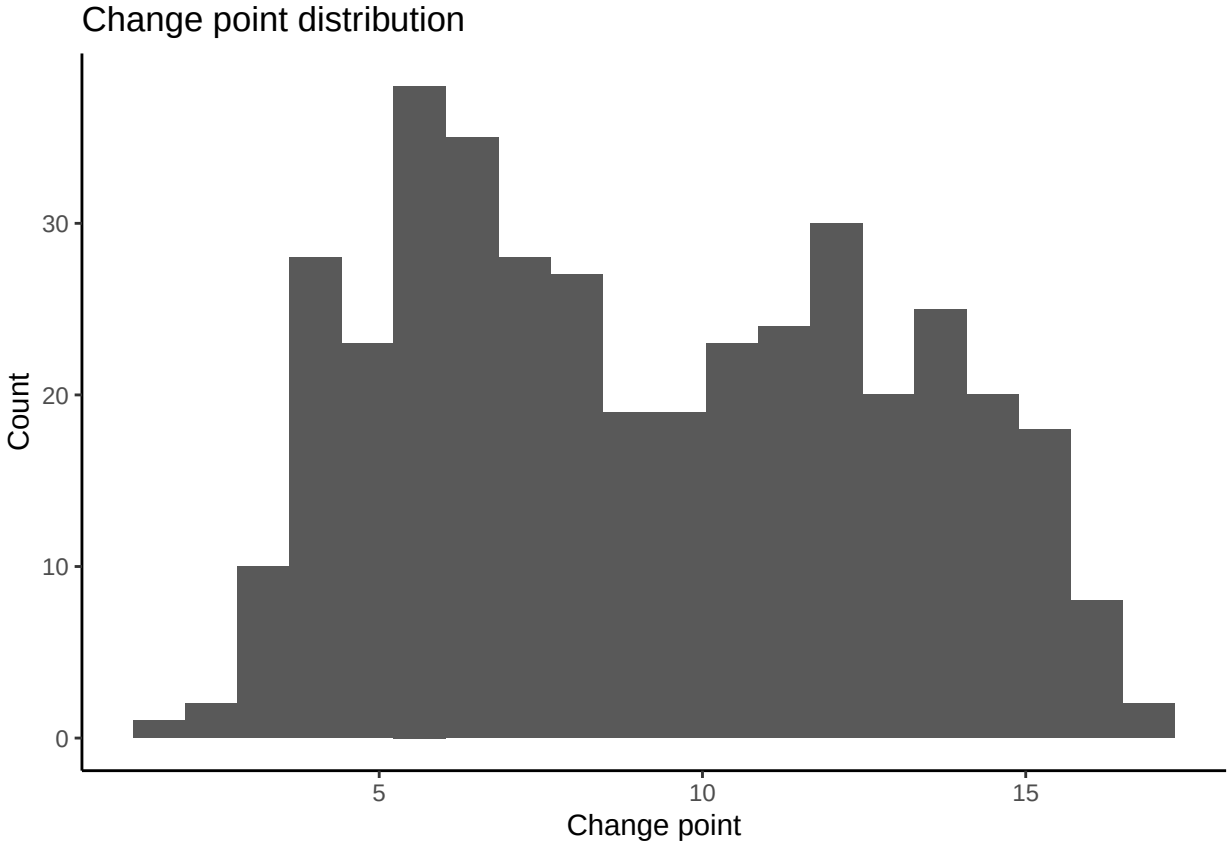

1000
bias
2000
1000
0
Value
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_ps
change_model
propensity_change
absolute_bias
30
20
10
0
Value
1000
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_ps
change_model
propensity_change


mse
1000
Value
100
75
50
25
0
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_ps
change_model
propensity_change
Response
Exposure
0
5
10
15
20

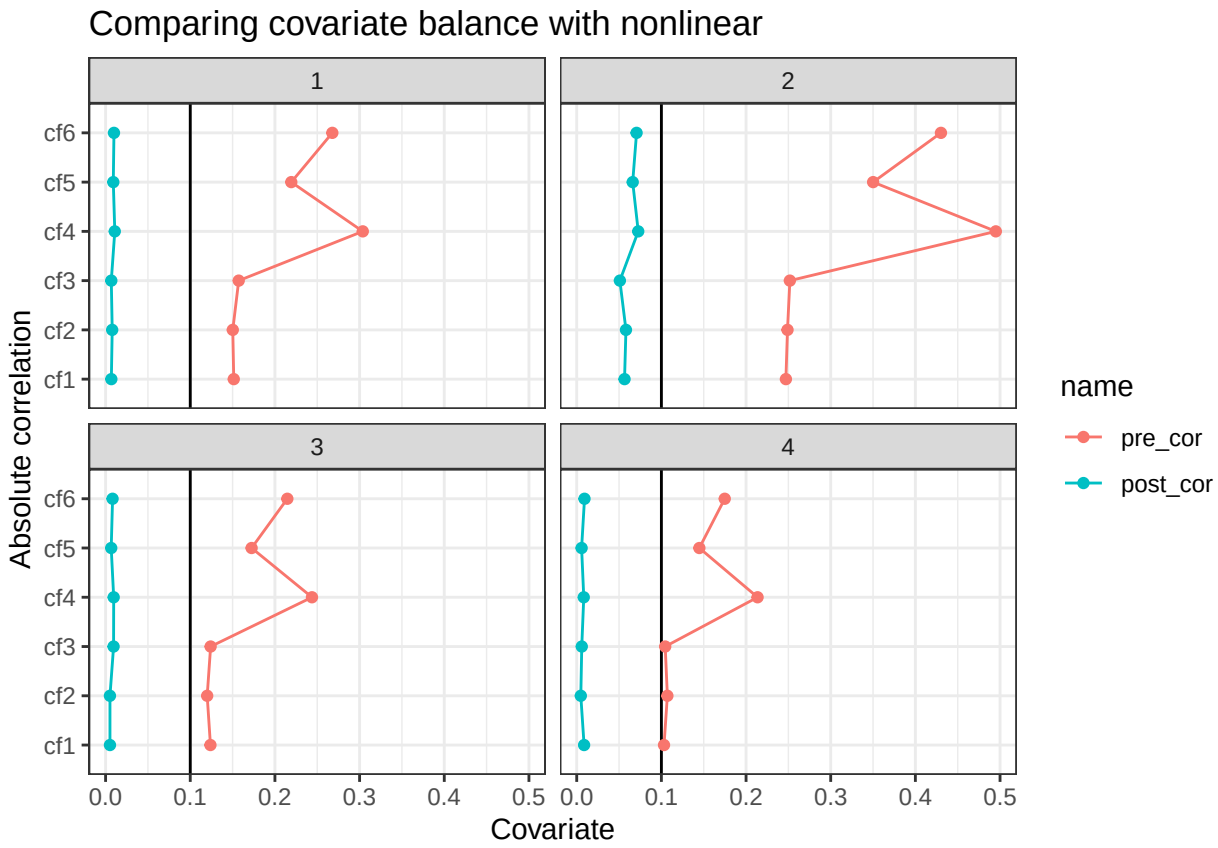
Nonlinear scenario (supralinear)

We repeat our analysis with a nonlinear relationship between the exposure and outcome to compare how these methods perform. The outcome model we fit is described below:

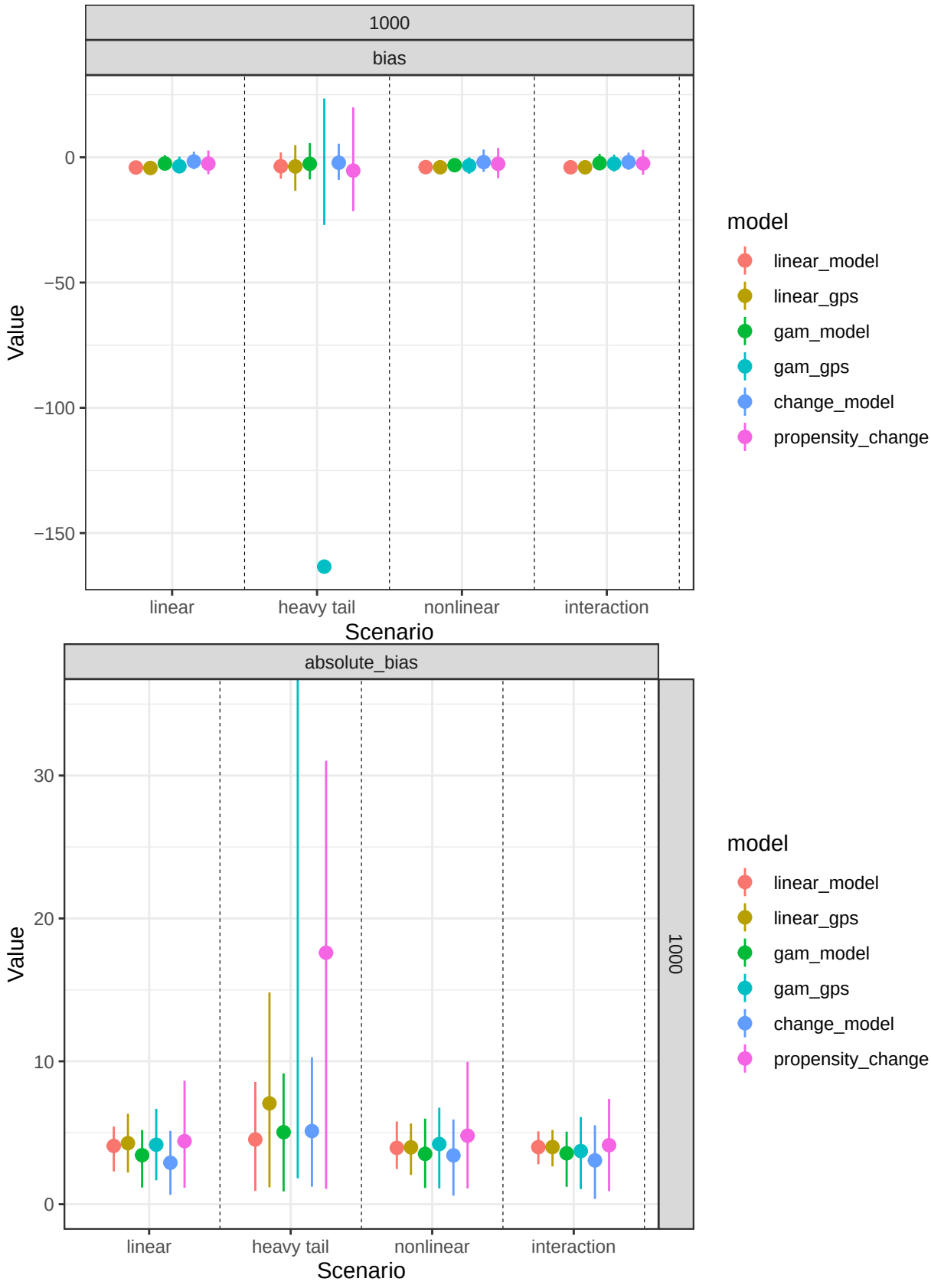
$$Y|E, C \sim N(\mu(E, C), 10^2)$$
$$\mu(E, C) = 20 + 8 * log_{10}(E + 1) - (2, 2, 3, -1, 2, 2) * C$$

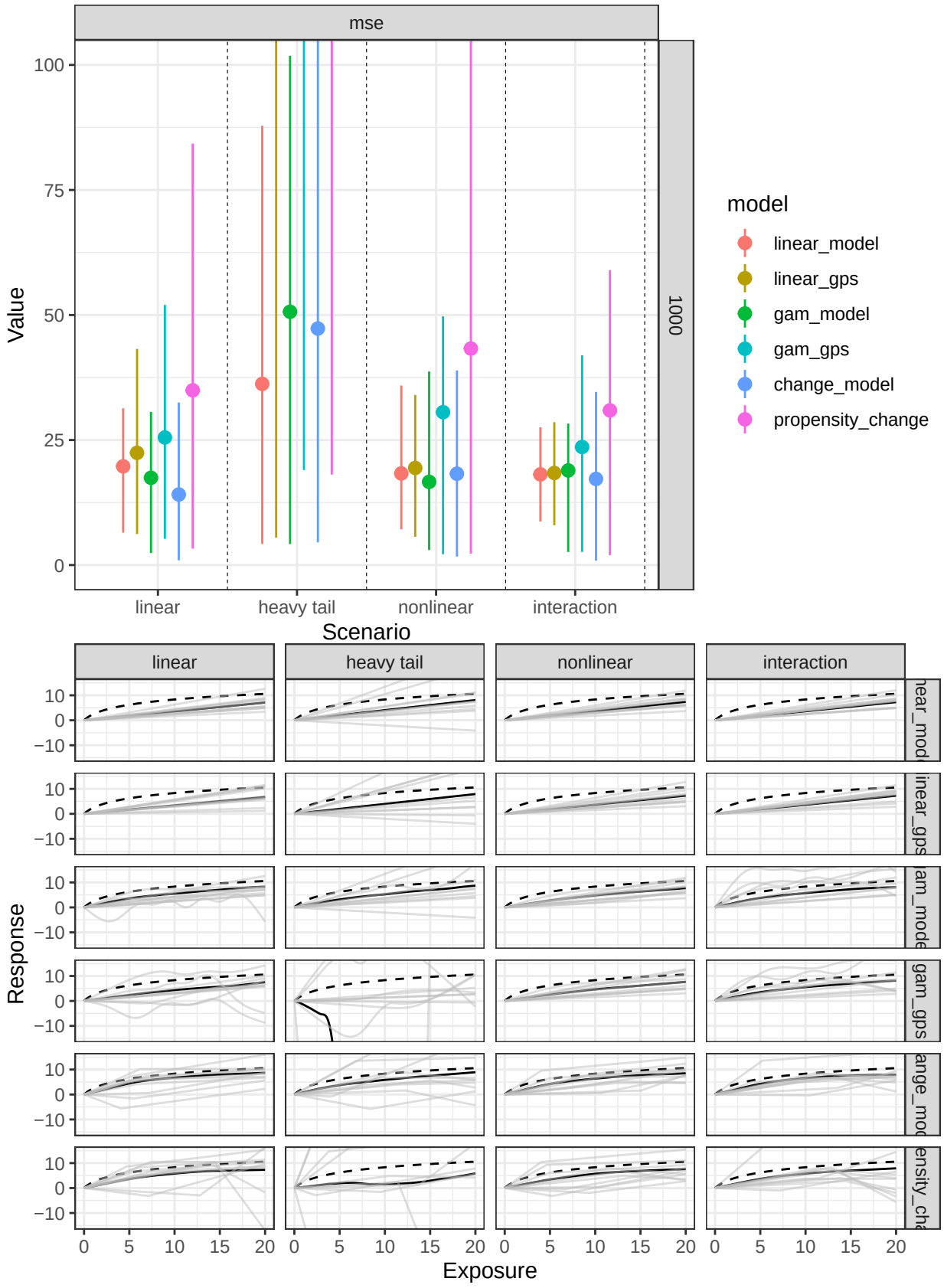



```
## `summarise()` has grouped output by 'gps_mod'. You can override using the
## `.groups` argument.
```




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


1000
bias
0
−50
−100
−150
Value
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change
absolute_bias
30
20
10
0
Value
1000
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change

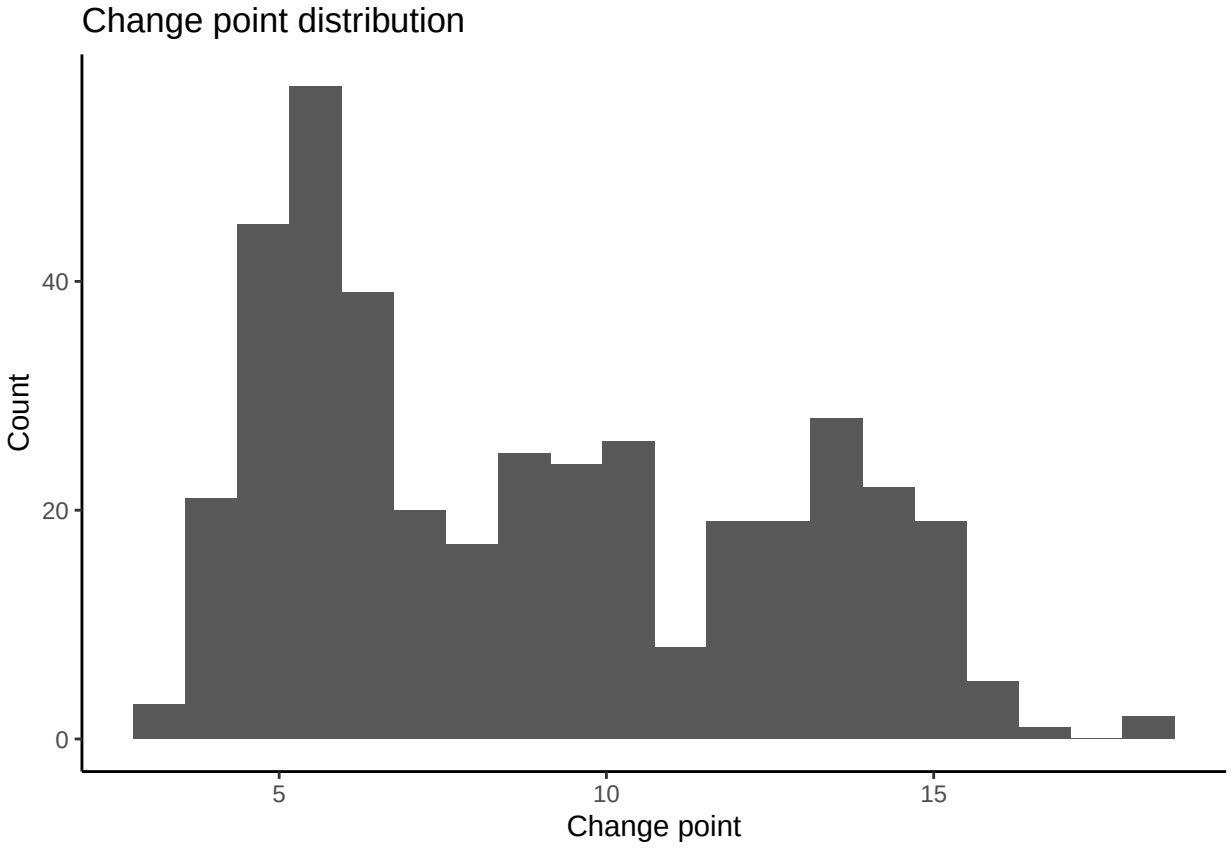

mse
Value
100
75
50
25
0
1000
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change
Response
Exposure
0
5
10
15
20

Threshold model

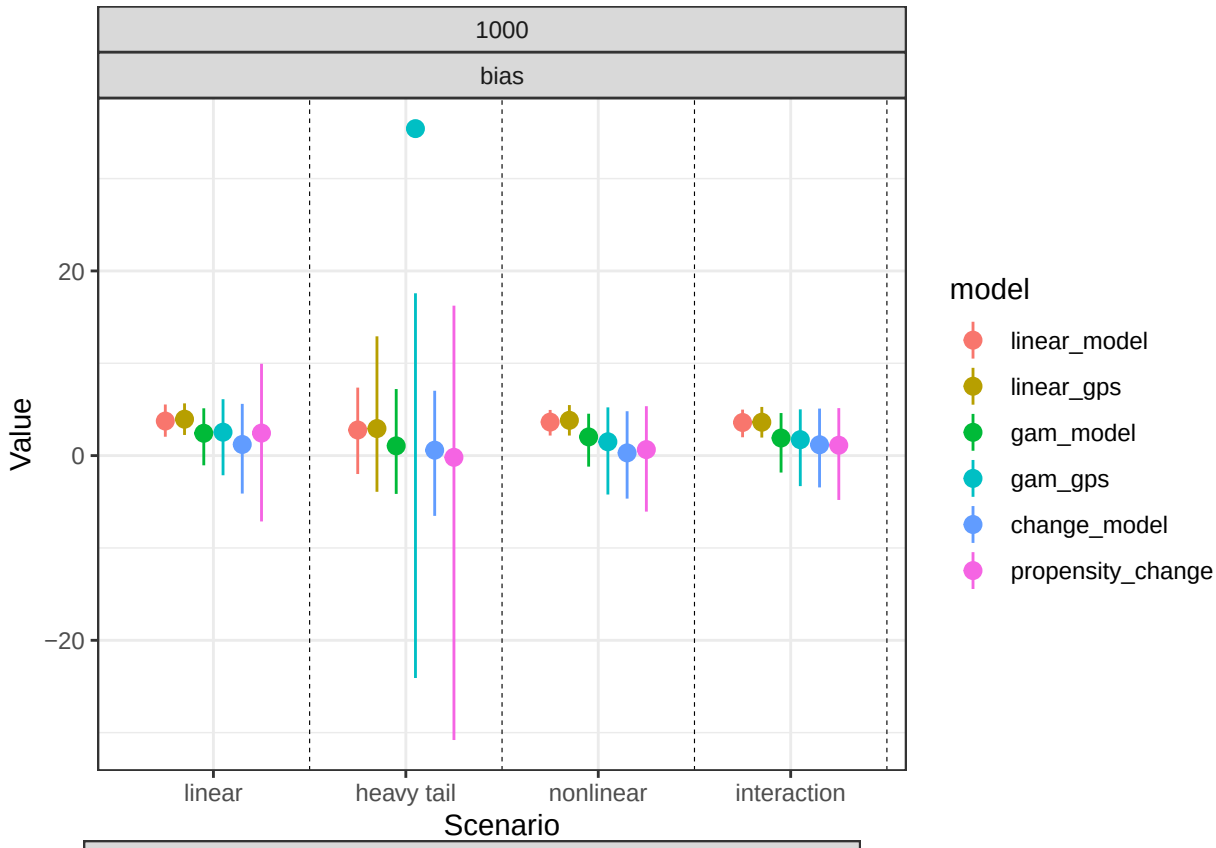
Finally we fit a threshold model and repeat our analysis. The following outcome model is fit:

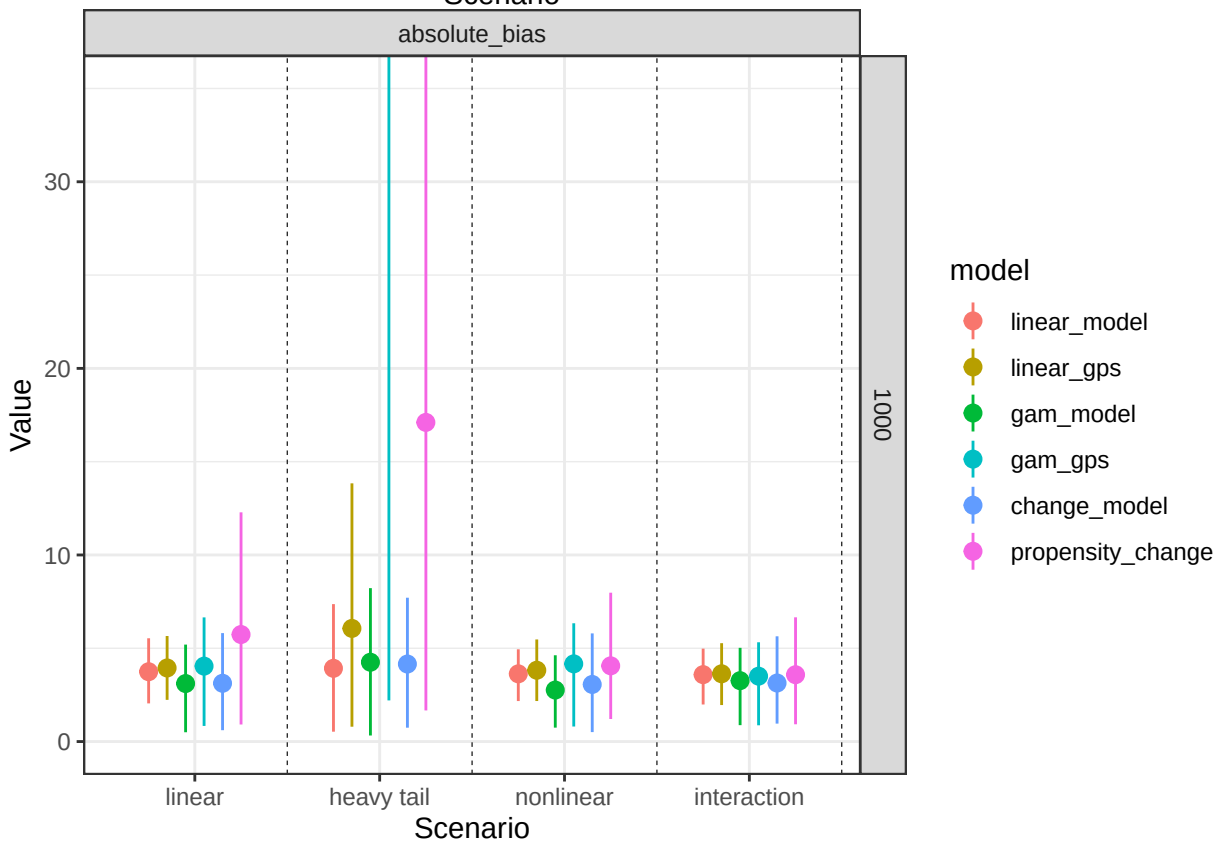
$$Y|E, C \sim N(\mu(E, C), 10^2)$$
$$\mu(E, C) = 20 + E[E > 5] - (2, 2, 3, -1, 2, 2) * C$$

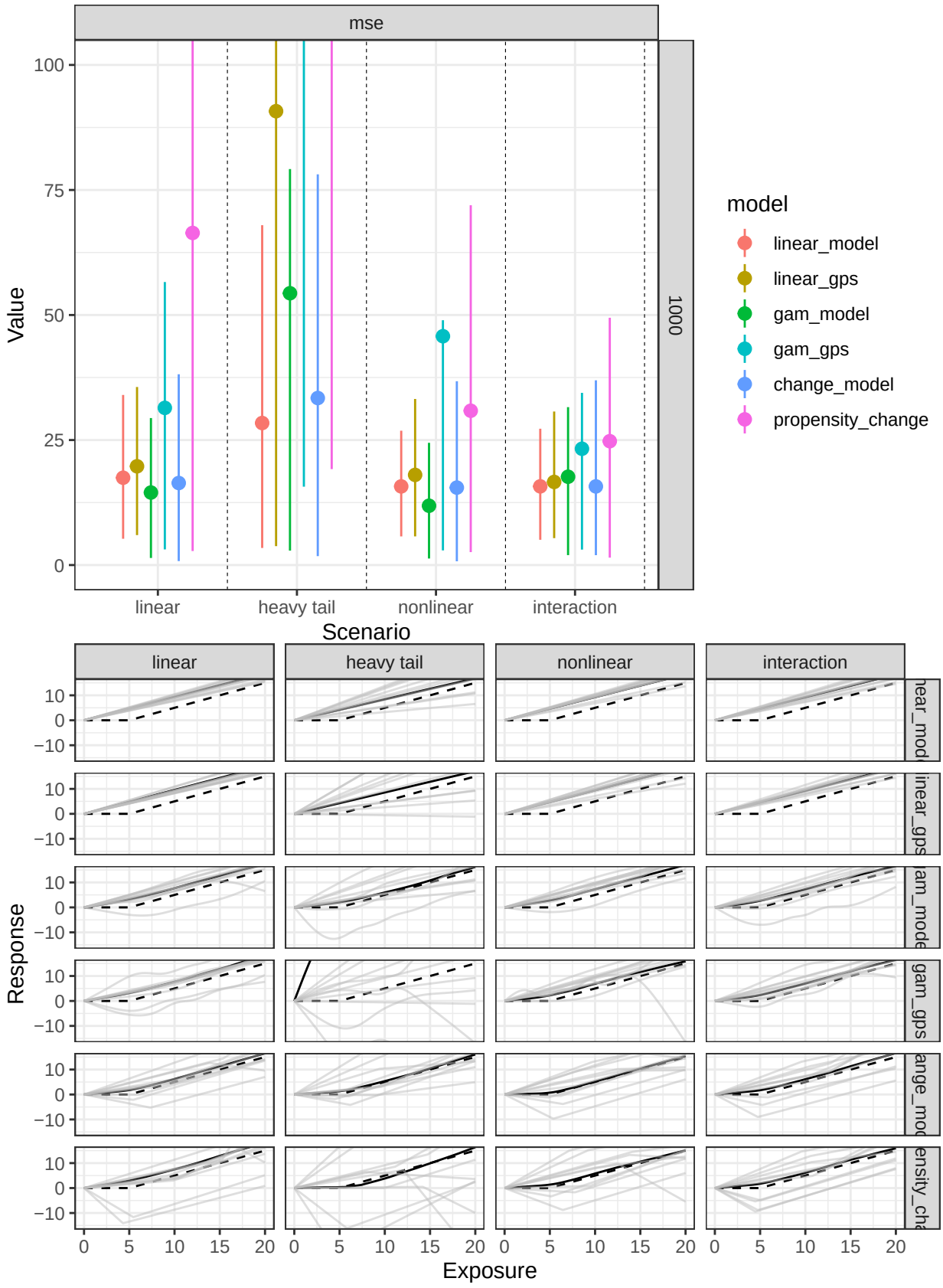
I am fitting a hinge threshold here, which is the correctly identified case here. Information on how this is fit can be found here.




```
## `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 `.groups` argument.
```

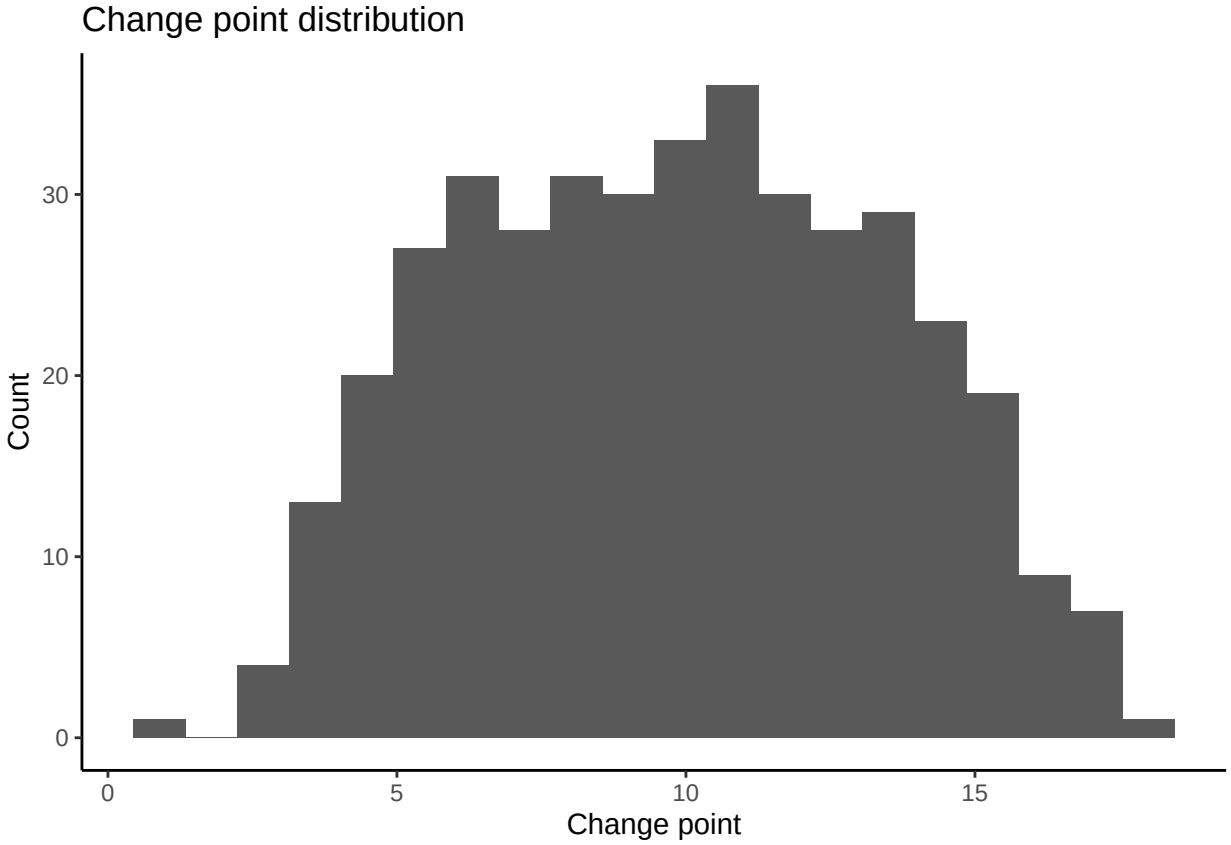

1000
bias
20
0
−20
Value
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change


absolute_bias
1000
30
20
10
0
Value
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change

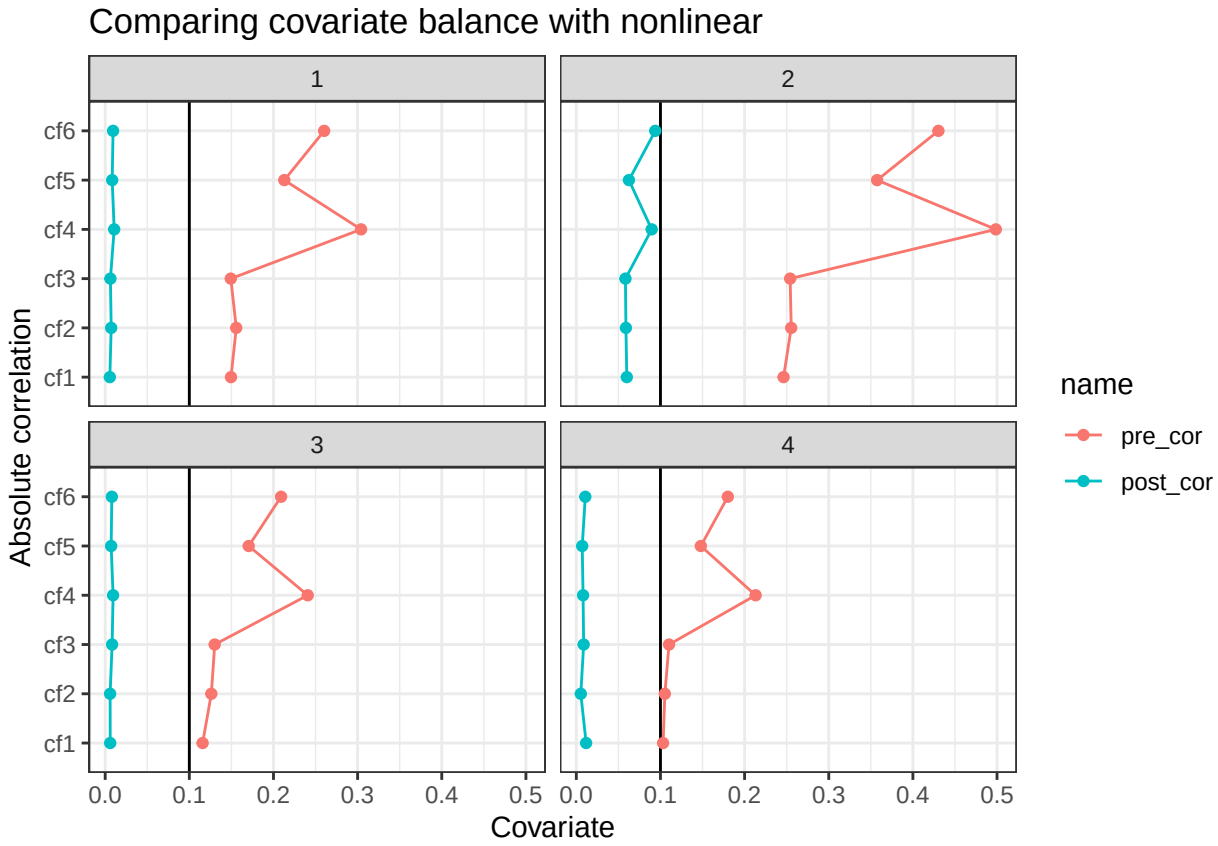

mse
Value
100
75
50
25
0
1000
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change
Response
Exposure

Nonlinearity in outcome model

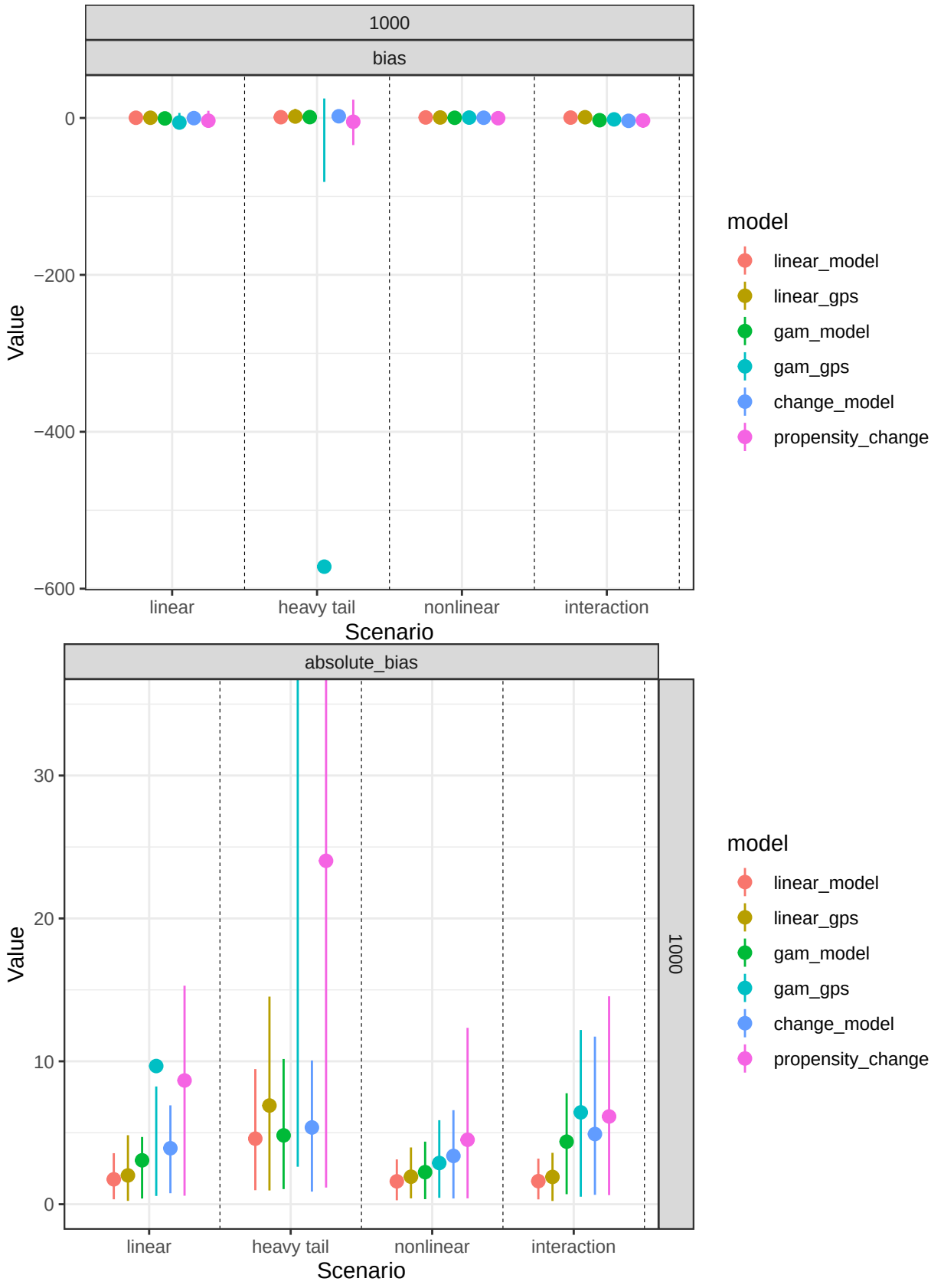
$$Y|E, C \sim N(\mu(E, C), 10^2)$$
$$\mu(E, C) = 20 + 0.1 * E - (2, 2, 3, -1, 2, 2) * C + 2 * C_1^2$$

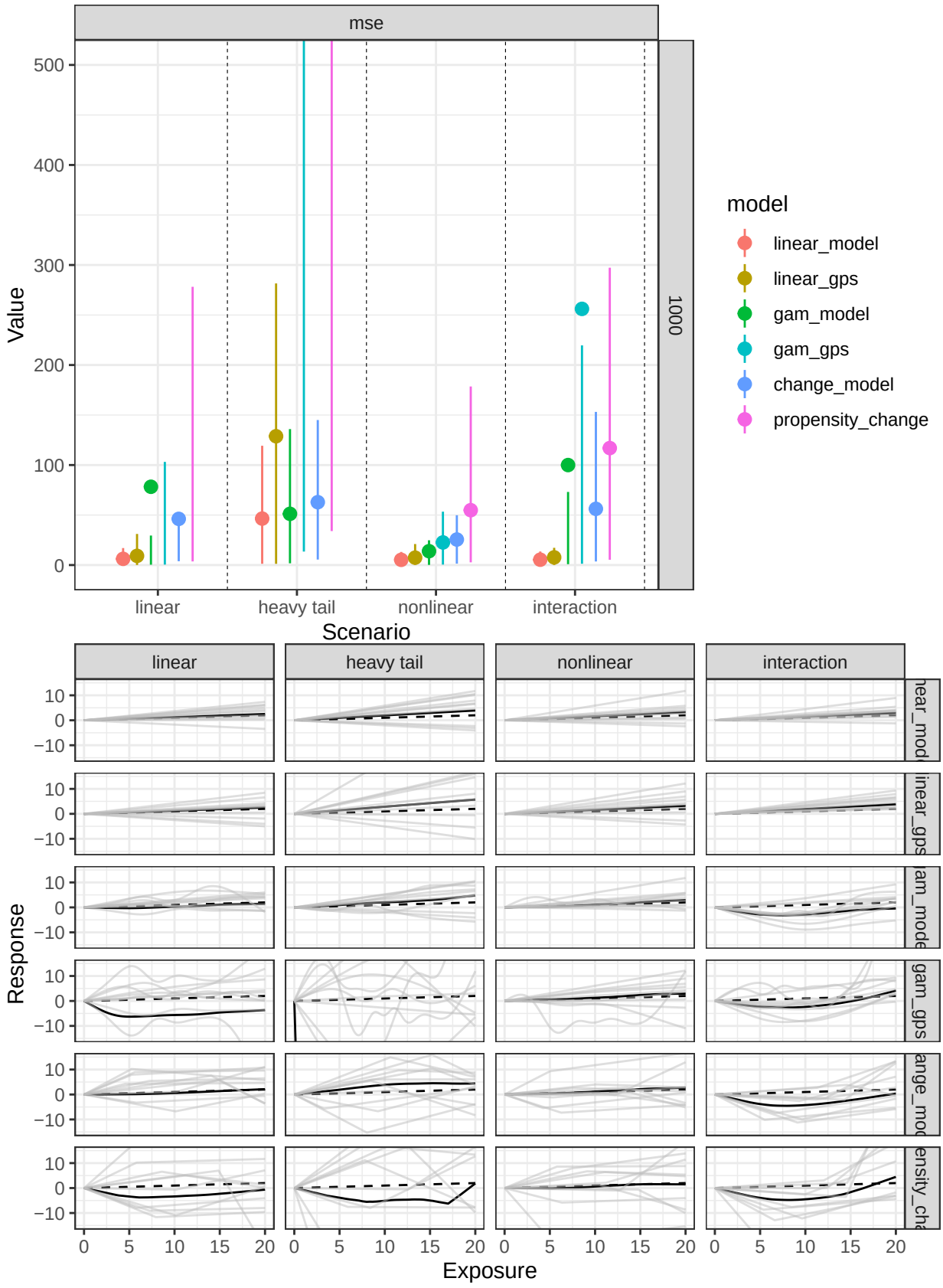



```
## `summarise()` has grouped output by 'gps_mod'. You can override using the
## `.groups` argument.
```

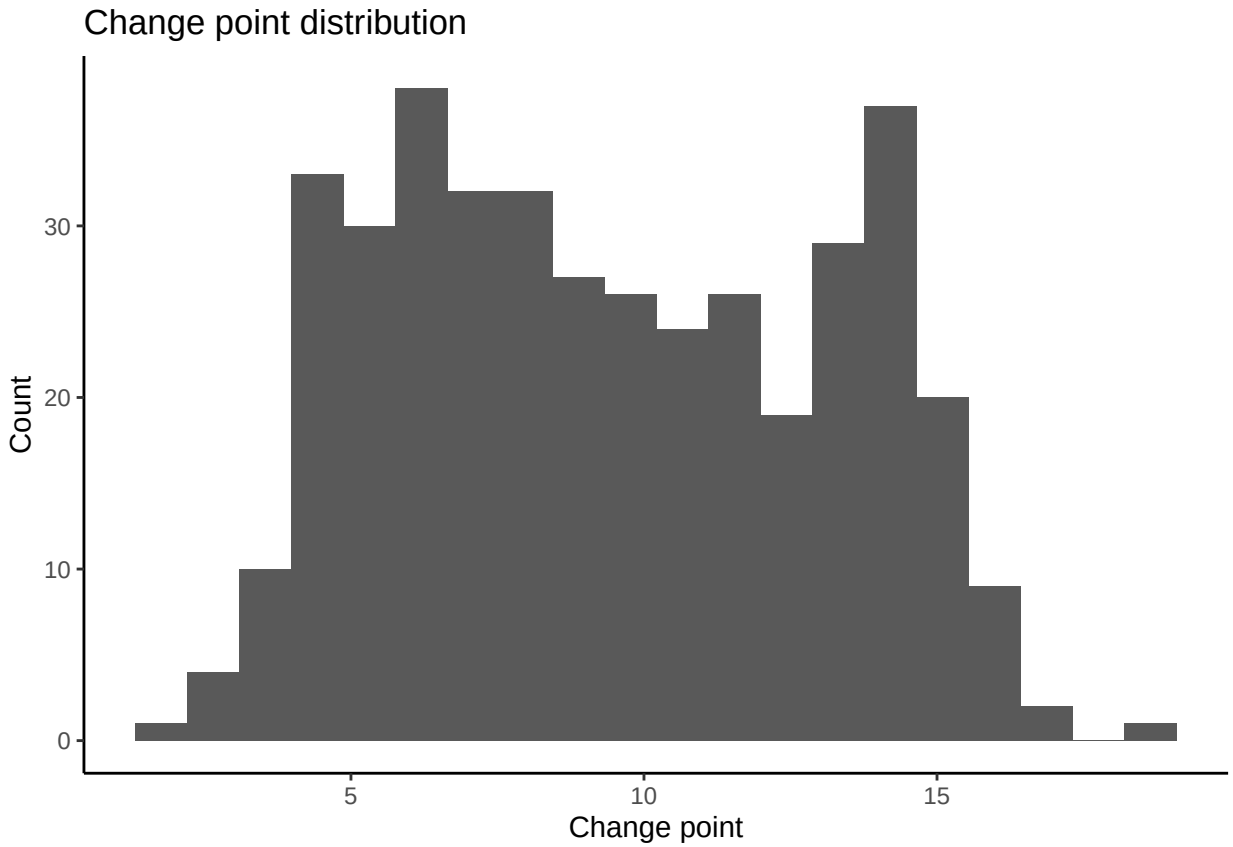



```
## `summarise()` has grouped output by 'model', 'gps_mod', 'sample_size'. You can
## override using the `.groups` argument.
```

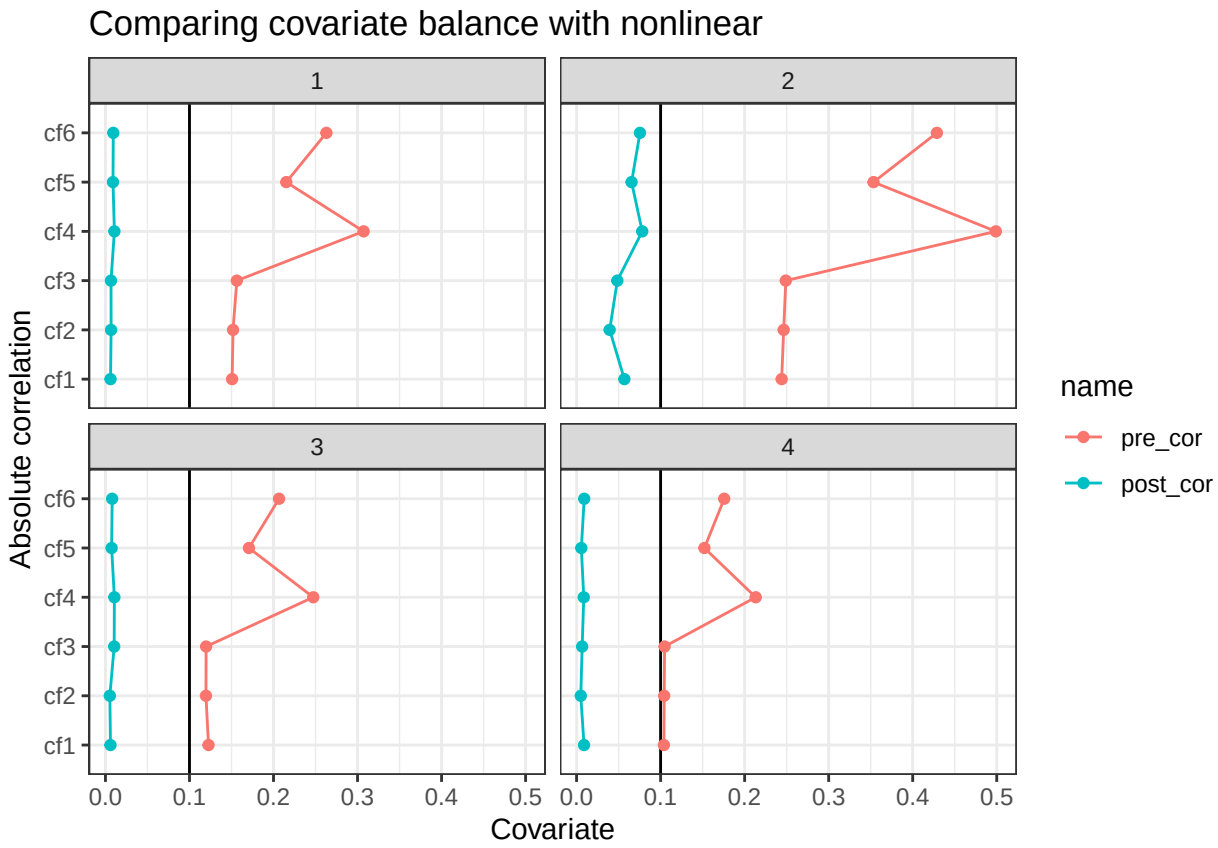

1000
bias
0
−200
−400
−600
Value
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change
absolute_bias
30
20
10
0
Value
1000
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change


mse
1000
Value
500
400
300
200
100
0
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
propensity_change
Response
Exposure
10
0
−10
0
5
10
15
20

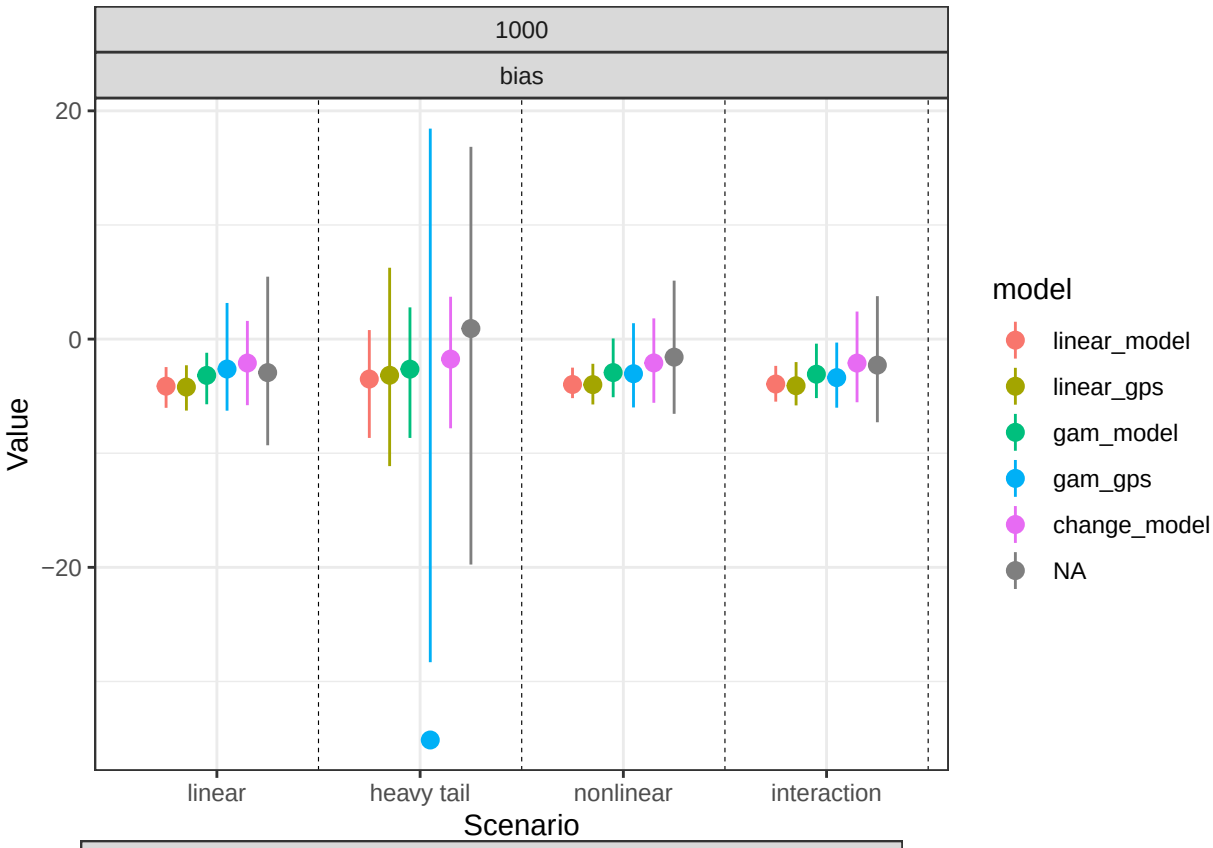
$$Y|E, C \sim N(\mu(E, C), 10^2)$$
$$\mu(E, C) = 20 + 8 * log_{10}(E + 1) - (2, 2, 3, -1, 2, 2) * C + 2 * C_1^2$$

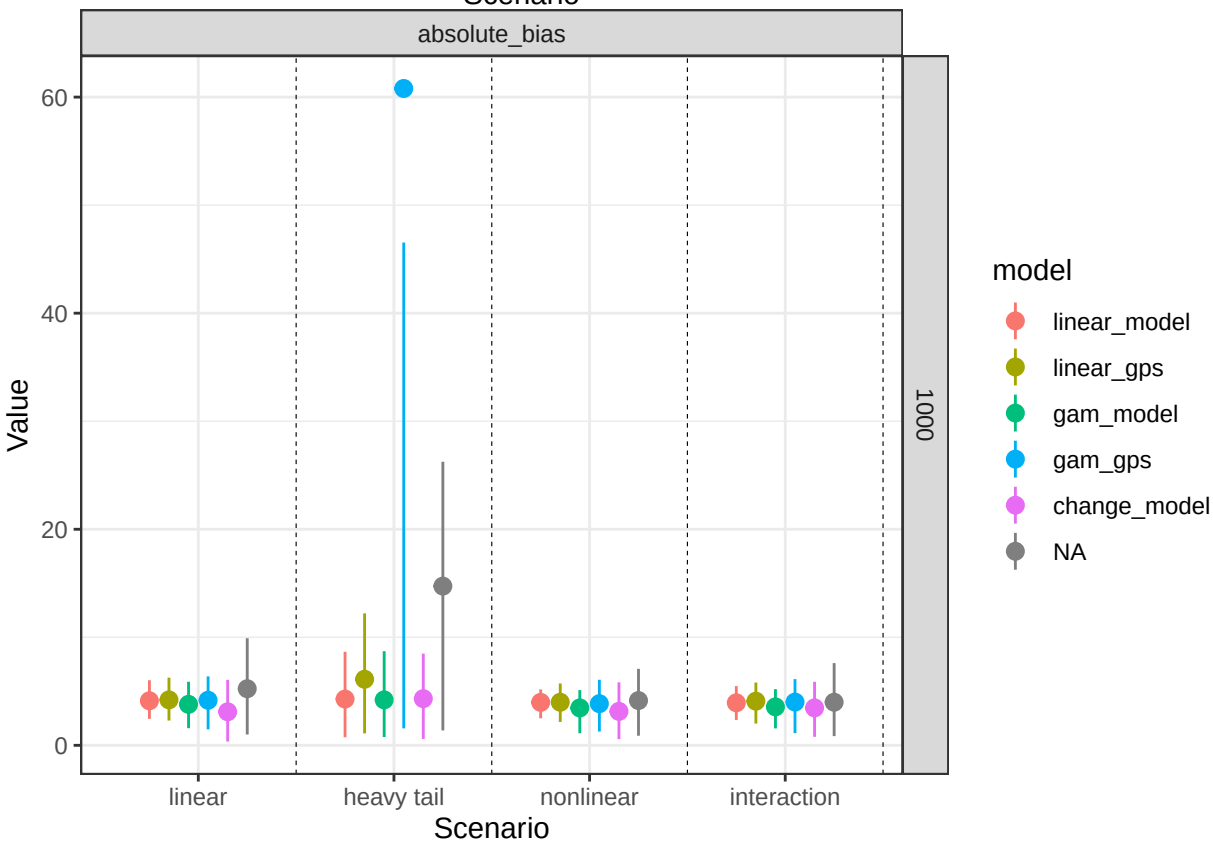


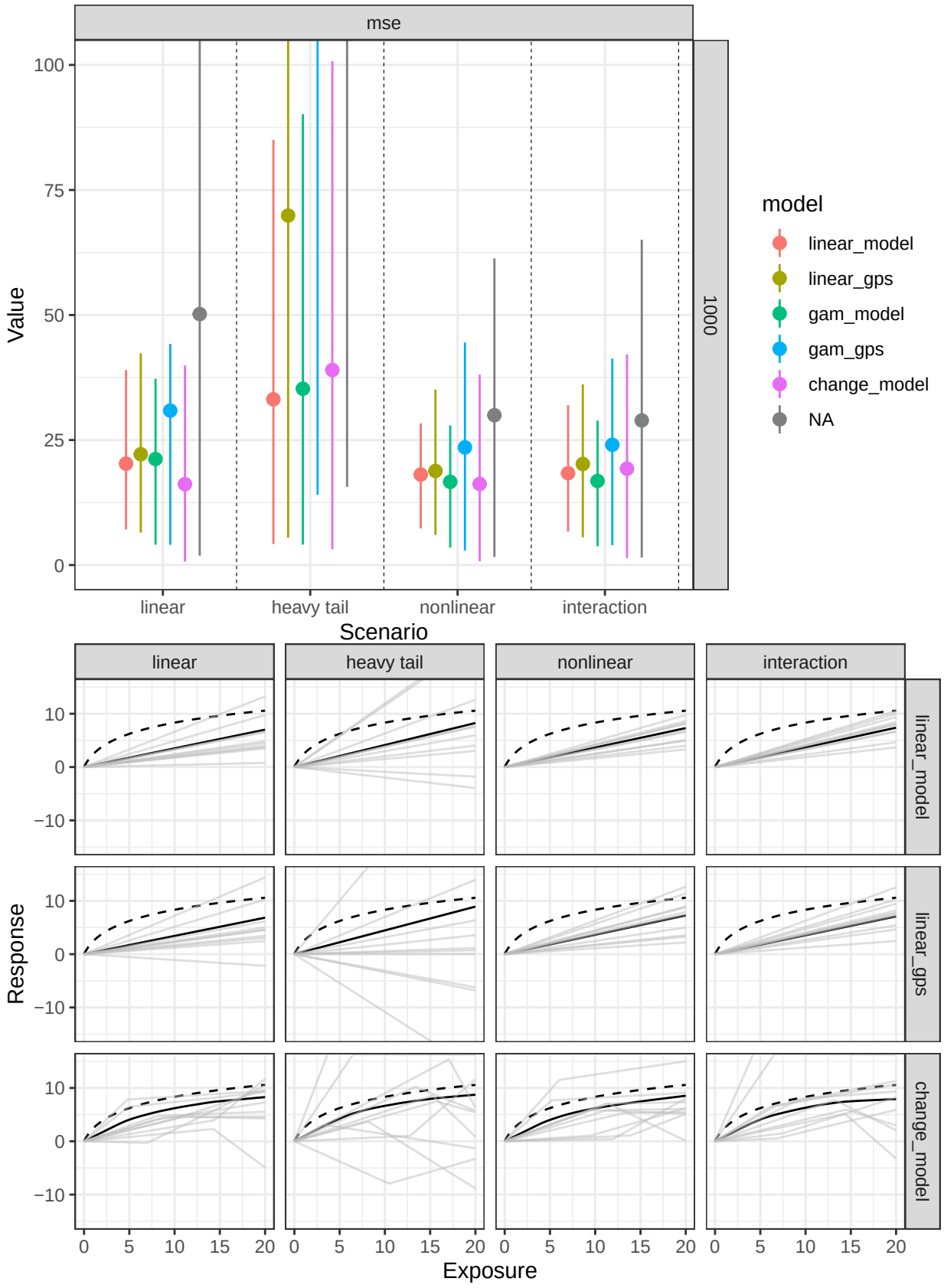

```
## `summarise()` has grouped output by 'gps_mod'. You can override using the
## `.groups` argument.
```




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


1000
bias
20
0
−20
Value
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
NA


absolute_bias
1000
60
40
20
0
Value
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_gps
change_model
NA


mse
1000
Value
100
75
50
25
0
linear
heavy tail
nonlinear
interaction
Scenario
model
linear_model
linear_gps
gam_model
gam_ops
change_model
NA
linear
heavy tail
nonlinear
interaction
linear_model
linear_gps
change_model
Response
10
0
−10
0
5
10
15
20
Exposure

End of document

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.

I then repeat with adjustment for linear confounding:

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.

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