Yearly predictions results FV3

Marc Roddis

7/12/2020

Results

Estimation of the regression coefficient `beta28year`

Our first goal will be to screen our 11 methods for the estimation of `beta28year` to determine which methods we will use in our main analysis in a later section. We will assess these estimates from their MSE, squared-bias and variance in each case.

We first choose parameter values for this screening study: cprop = 0.3, beta153year = -0.02, sd28_153= 0.3. These values for cprop and beta153year are equal to our estimates from our real dataset `pcb.csv` , whereas this value for sd28_153 is equal to the mean of two estimates: one which is unconditional, and a second which is conditional on the variable `year` .

Evaluation of methods for smaller sample sizes

We will first obtain results from datasets with different sample sizes in order to decide an appropriate sample size for all our subsequent work. Our real dataset has approximately 100 observations per year for CB28 and CB153 from herring in years 2003-2017. However these observations are from various locations and have differences for various other variables such as age, fat-percentage etc., which means that any statistical analysis which controls for such variables would have a smaller sample size. We will test sample sizes that differ by a factor of 2: we do this by generating datasets by simulation using 10000 iterations, with sample sizes 50, 25, 12 and 6 respectively. The squared-bias of the estimates of `beta28year` from all 11 methods and all 4 sample sizes is shown below; note that all values shown in the table are 100000 times bigger than the actual values (to make them easier to read and compare). The column names `bias_ss50` , `bias_ss25` , … denote sample sizes 50, 25, … respectively.

The following table below is the same as the previous one, except that it shows the variance of the estimates.

Allowing for random error from using only 10000 iterations, we can conclude that the squared-bias is independent of sample size, whereas the variance is inversely proportional sample size. Moreover since the bias_variance decomposition $ MSE = Bias^2 + Variance$, always holds, we need not look at the MSE values for the purpose of choosing sample size.

We find in additional experiments (details not shown) that the standard error of the estimates is inversely proportional to the square root of the number of simulation iterations, so we have three factors to balance:

1. We want our results to be potentially applicable for real data.
2. We want sample size to be sufficiently large to avoid MSE being dominated by variance alone.
3. We want the number of iterations to be sufficiently large that our estimates have sufficiently low standard error.

We therefore decide to use sample size = 12 for all of our subsequent experiments.

```
##                mse_beta bias_beta variance_beta
## omit            6.73529   6.26334       0.47242
## subst2          1.02097   0.16359       0.85824
## subst1          2.63096   2.24608       0.38527
## censReg1        0.67592   0.00015       0.67645
## censReg2        0.73239   0.00016       0.73296
## censReg0impute  0.73307   0.00020       0.73360
## best            0.71821   0.00119       0.71774
## subst4          6.98697   5.32509       1.66354
## censReg1naive   1.13477   0.69481       0.44040
## subst2lmimpute  6.19523   5.59525       0.60058
## omitlmimpute    7.39523   6.50439       0.89173
```

```
##                mse_beta bias_beta variance_beta
## omit            7.30307   6.31339       0.99067
## subst2          1.82651   0.13937       1.68883
## subst1          3.05353   2.30050       0.75378
## censReg1        1.32626   0.00127       1.32632
## censReg2        1.43939   0.00156       1.43926
## censReg0impute  1.43625   0.00133       1.43636
## best            1.37819   0.00010       1.37947
## subst4          8.41458   5.12293       3.29495
## censReg1naive   1.57556   0.71425       0.86217
## subst2lmimpute  6.65171   5.83757       0.81496
## omitlmimpute    7.78183   6.66494       1.11800
```

```
##                mse_beta bias_beta variance_beta
## omit            8.41972   6.35020       2.07160
## subst2          3.82340   0.12135       3.70576
## subst1          3.97220   2.34445       1.62938
## censReg1        2.87939   0.00449       2.87778
## censReg2        3.14083   0.00496       3.13900
## censReg0impute  3.15323   0.00456       3.15182
## best            2.90273   0.00305       2.90258
## subst4         12.20168   4.96336       7.24557
## censReg1naive   2.55612   0.79263       1.76526
## subst2lmimpute  7.52554   6.19608       1.33079
## omitlmimpute    8.84331   7.15221       1.69280
```

```
##                mse_beta bias_beta variance_beta
## omit           10.93907   5.58942       5.35500
## subst2          8.11560   0.41070       7.71261
## subst1          5.49398   1.79078       3.70691
## censReg1        6.28797   0.04032       6.25390
## censReg2        6.81454   0.04134       6.77997
## censReg0impute  6.84659   0.03893       6.81448
## best            5.77008   0.00002       5.77584
## subst4         21.29759   6.86397      14.44806
## censReg1naive   4.60166   0.41858       4.18726
## subst2lmimpute  8.86108   6.20702       2.65672
## omitlmimpute    9.17897   6.22616       2.95576
```

Selection of censoring methods for further study

We will now use simulations with just 1000 iterations for all 10 methods (and also for our reference method `best`) to estimate `beta28year` for four sets of parameter values:

beta28year = -0.02 is held fixed.

a “low” and a “high” value for each of `cprop` and `sd28_153` are used. Concretely: (0.1, 0.1), (0.7, 0.1), (0.1, 0.5) and (0.7, 0.5) were used for (cprop, sd28_153) respectively.

The following four tables show the MSE, squared-bias, and variance of estimates of `beta28year` from all 11 methods, for the four sets of parameter values, respectively.

We see that there is a much bigger difference between different methods in the amount of bias than in the amount of variance. We will therefore focus primarily on the results for bias; we will use terms such as high and low to compare the bias from different methods. We see that the amount of bias for:

`best` serves as a reference value; a gold standard that we compare the other methods with.

`omit` is high for (0.1, 0.1) and (0.1, 0.5), and is very high for (0.7, 0.1) and (0.7, 0.5). It makes sense that there is higher bias with higher proportion of censored values since a higher proportion of the data has been omitted. Moreover, these generally high values are commensurate with our prior expectations (ref: Helsel’s book) that `omit` is a poor method, so we will not study this method further.

Very high for: `subst1` for (0.7, 0.1) and (0.7, 0.5); `subst2` for (0.7, 0.5); `subst4` for (0.1, 0.1) and (0.7, 0.1). However, all three substitution methods also have low bias for at least one set of parameter values. This is intriguing and merits further investigation.

Very low for: `censReg1`, `censReg2` and `censReg0impute` for all four parameter value sets.

Highest of all four censReg methods for all four parameter sets for `censReg1naive`. This illustrates the necessity of conditioning on both the cb153 value and the condition cb28 < cb28_cprop by using a truncated normal distribution, and verifies the results we presented in our previous chapter on mathematical theory. `censReg1`, `censReg2` and `censReg0impute` all do this, whereas in contrast, `censReg1naive` conditions solely on the cb153 value, and thus uses a (non-truncated) normal distribution; this results in significant bias because cb28 values can be erroneously imputed to be higher than `cb28_cprop`. Consequently we will not discuss censReg1naive any further: it has served its purpose in showing the importance of conditioning oth the cb153 value and the condition cb28 < cb28_cprop.

The hybrid methods `subst2lmimpute` and `omitlmimpute` first use substitution and omission as in `subst2` and `omit` respectively, followed by imputation. Therefore `subst2lmimpute` should be compared with `subst2`, and `omitlmimpute` with `omit`. `subst2lmimpute` has higher bias (and also MSE) than `subst2` for all parameter value sets, and `omitlmimpute` has higher bias than `omit` for all sets except (0.1, 0.5). We have already rejected the `omit` method so we must also reject `omitlmimpute` since it performs no better than `omit`. Similarly we reject `subst2lmimpute`, since this method performed worse than `subst2` in all four cases.

In summary, we have rejected 4 of our 10 methods. We will limit our attention to six methods for all our subsequent work: the three substitution methods `subst1`, `subst2`, `subst4`, and the three censReg methods `censReg1`, `censReg2` and `censReg0impute`. We will use `best` as our reference method throughout.

```
##                mse_beta bias_beta variance_beta
## omit            1.94573   1.50568       0.44049
## subst2          3.28387   2.47542       0.80926
## subst1          0.62231   0.19495       0.42779
## censReg1        0.48293   0.00001       0.48340
## censReg2        0.48707   0.00000       0.48756
## censReg0impute  0.48612   0.00000       0.48661
## best            0.49985   0.00002       0.50033
## subst4         14.44620  12.87532       1.57245
## censReg1naive   0.74552   0.31287       0.43308
## subst2lmimpute  4.33385   3.80822       0.52616
## omitlmimpute    2.22148   1.66736       0.55467
```

```
##                mse_beta bias_beta variance_beta
## omit           14.13397  13.45675       0.67790
## subst2          1.49817   0.62596       0.87308
## subst1         12.62066  12.52761       0.09314
## censReg1        0.74680   0.00039       0.74716
## censReg2        0.96849   0.00129       0.96816
## censReg0impute  1.02894   0.00175       1.02822
## best            0.49985   0.00002       0.50033
## subst4         28.93460  26.23278       2.70452
## censReg1naive   1.03778   0.25186       0.78671
## subst2lmimpute 10.28893  10.15964       0.12942
## omitlmimpute   18.82095  18.70101       0.12007
```

```
##                mse_beta bias_beta variance_beta
## omit            8.16691   2.32836       5.84439
## subst2          6.97480   0.01114       6.97063
## subst1          6.08940   0.44654       5.64852
## censReg1        6.64952   0.03803       6.61811
## censReg2        6.76917   0.03933       6.73658
## censReg0impute  6.76973   0.03983       6.73664
## best            7.76117   0.01682       7.75210
## subst4          9.06100   0.20899       8.86088
## censReg1naive   6.55698   1.37954       5.18262
## subst2lmimpute  7.42060   2.45574       4.96983
## omitlmimpute    6.95762   1.84795       5.11478
```

```
##                mse_beta bias_beta variance_beta
## omit           21.18031  12.48679       8.70222
## subst2          8.37304   5.15594       3.22032
## subst1         13.19163  11.98684       1.20600
## censReg1        7.66589   0.00451       7.66905
## censReg2       11.34171   0.00386      11.34921
## censReg0impute 11.34744   0.00429      11.35451
## best            7.76117   0.01682       7.75210
## subst4          7.81905   1.16453       6.66118
## censReg1naive   9.35157   0.90675       8.45328
## subst2lmimpute 12.84661  11.45671       1.39129
## omitlmimpute   19.55833  17.22978       2.33089
```

Evaluation of methods for larger absolute values of `beta28year`

We will now focus our six chosen methods `subst1`, `subst2`, `subst4`, `censReg1`, `censReg2`, `censReg0impute`. We will use these methods to estimate `beta28year` from four simulations that use the same parameter values ss = 12, cprop = 0.3, sd28_153 = 0.3, n_iter = 10000 as before. We will use the four `cb153year` parameter values -0.02, -0.04, -0.08, -0.16 in these four simulations, respectively. The results from the simulations are shown in the four tables below.

We see that for these parameters value sets, censReg method give estimates that have much lower bias, in general. The `subst1` method is designed as a reference that gives biased estimates, since it substitutes cb28 values that are observed to be below LOD with the LOD value itself, so the substituted values will always be larger than the real values. Moreover, we have chosen to maintain a constant LOD level for all years of the same dataset. We are also simulating cb28 and cb153 data using a linear (degree 1 polynomial) function with a negative slope and a fixed constant (intercept) term. This means that the cb28 values decrease faster with years for larger values of abs(beta153year). This all means that it is an inevitable consequence of our design that the bias from `subst1` increases as abs(beta153year) increases, which is precisely what we see in these results.

In contrast, the bias from `subst4` first increases from abs(beta153year) = 0.02 to 0.08 and then decreases for abs(beta153year) = 0.16. This suggests that since `subst4` substitutes censored values with $\frac{LOD}{2}$, which are lower than the true values on average for low values of abs(beta153year) but not lower for the highest value abs(beta153year) = 0.16. This is also supported by the fact that the bias from `subst2` is much lower than that from `subst1` or `subst4`, which suggests that the real values of the censored data mostly lie between LOD and $\frac{LOD}{2}$.

The three censReg methods all give very similar results to one another; the values of MSE, squared-bias and variance are very similar from these methods for both abs(beta153year) = 0.08 and 0.16. However, for the lowest value abs(beta153year) = 0.02, the variance from `censReg1` is approximately 10% lower than from `censReg2`, which is a statistically significant difference. This fits with our prior knowledge that a more complex model generally has higher variance than the corresponding less complex one. Moreover, we also expected that the estimates from `censReg2` would improve relative to those from `censReg1` as the value of abs(beta153year) increases, since the difference between these two methods is that `censReg2` uses `year` as additional predictor variable. However, the fact that `censReg0impute` has a higher variance than `censReg1` seems puzzling in this respect, so perhaps our interpretation of model complexity is wrong is this context.

Our prior expectation was that `censReg0impute` would perform relatively worse compared to the other `censReg` methods for larger values of abs(beta153year). This is because we conjecture that the imputations from the predictor variables carry more information about cb28 as the absolute value of `beta28year` increases, and `censReg0impute` does not use imputation at all. However, these results fail to support our conjecture here too.

If we now compare the best performing models from each category, i.e. `subst2` and `censReg1`, we see that `censReg1` has much lower bias for all parameter values. However, MSE for `subst2` is lower for one of the values, abs(beta153year) = 0.08. In conclusion, we can say that `censReg1` gives better estimates than `subst2` for most, but not necessarily all, values of abs(beta153year).

```
##                mse_beta bias_beta variance_beta
## subst1           3.9272    2.2607        1.6667
## subst2           3.8331    0.1603        3.6732
## subst4          12.4075    5.3096        7.0986
## censReg1         2.8892    0.0003        2.8891
## censReg2         3.1412    0.0006        3.1409
## censReg0impute   3.1444    0.0006        3.1441
## best             2.8939    0.0001        2.8941
```

```
##                mse_beta bias_beta variance_beta
## subst1          10.3914    8.6084        1.7833
## subst2           4.0038    0.4912        3.5130
## subst4          25.2352   18.7980        6.4379
## censReg1         3.1584    0.0001        3.1586
## censReg2         3.2579    0.0002        3.2580
## censReg0impute   3.2621    0.0002        3.2622
## best             2.8074    0.0001        2.8076
```

```
##                mse_beta bias_beta variance_beta
## subst1          32.8102   30.5608        2.2497
## subst2           3.5467    0.3798        3.1672
## subst4          50.4127   45.7078        4.7055
## censReg1         3.6276    0.0002        3.6278
## censReg2         3.6445    0.0002        3.6446
## censReg0impute   3.6540    0.0003        3.6540
## best             2.8808    0.0006        2.8805
```

```
##                mse_beta bias_beta variance_beta
## subst1          99.3248   96.2660        3.0591
## subst2           6.0901    2.8591        3.2313
## subst4          44.8933   41.3417        3.5519
## censReg1         4.4961    0.0020        4.4945
## censReg2         4.4988    0.0020        4.4972
## censReg0impute   4.5284    0.0024        4.5265
## best             2.8542    0.0017        2.8528
```

Evaluation of methods for other values of `sd28_153`

We will now hold `beta28year` and `cprop` fixed at their original values (-0.02 and 0.3) and investigate the effect of larger `sd28_153` values, specifically: 0.1, 0.3, 0.5, and 0.7.

We see again that for these parameters value sets, censReg method give estimates that have very much lower bias, in general. Since the three censReg methods all give very similar results to one another and very different results from the three substitution methods, we will again begin by interpreting the results for these two method categories separately.

The bias from `subst4` decreases greatly as the value of `sd28_153` increases, whereas the bias from `subst1` is relatively independent of the value of `sd28_153` . The bias from `subst2` again follows a trend intermediate between that of `subst1` and `subst4` , since it decreases from sd28_153 = 0.1 to 0.5 and then decreases for sd28_153 = 0.7. Our interpretation is that since the censored values lie closer on average to LOD for smaller values of sd28_153, and further away for larger values. The low bias from `subst4` for sd28_153 = 0.7 indicates that the real values for the censored data lie close to $\frac{LOD}{2}$ on average for this parameter value.

The large gap between the uncensored cb28 data and the $\frac{LOD}{2}$ value means that `subst4` gives higher variance than all other methods for all values of sd28_153. Similarly, the smallest possible gap between LOD and the uncensored cb28 data explains the fact that `subst1` always gives the lowest variance. We conjecture that the same logic would also hold for other possible substitution values; the larger the gap between this value and LOD, the larger the resulting variance.

Again, we see that the variance from `censReg1` is approximately 10 % lower than that from `censReg2` for all four values of sd28_153. Surprisingly the results from `censReg2` and `censReg0impute` are almost identical. Is this a bug?

In conclusion, substitution methods give much higher bias than cenreg methods. Moreover, all three cenreg methods gave lower MSE than all three substitution methods for both sd28_153 = 0.1 and sd28_153 = 0.3. However, the variance from cenreg methods increases faster than from substitution methods as sd28_153 increases; in fact for higher values of sd28_153, `subst1` and `subst2` gave the lowest and second lowest MSE values, respectively. This relative failure of cenreg methods for relatively high values of sd28_153 makes sense, here is our explanation: A higher `sd28_153` value means that the correlation between cb28 and cb153 is weaker, which results in less accurate imputation by ´censReg1 `and` censReg2`, since the accuracy of imputation by these methods relies on the strength of correlation between cb28 and cb153.

```
##                mse_beta bias_beta variance_beta
## subst1           2.4314    2.1164        0.3150
## subst2           9.3323    8.0908        1.2416
## subst4          54.2390   51.0318        3.2075
## censReg1         0.5166    0.0000        0.5167
## censReg2         0.5478    0.0001        0.5478
## censReg0impute   0.5575    0.0001        0.5574
## best             0.4887    0.0000        0.4887
```

```
##                mse_beta bias_beta variance_beta
## subst1           3.9272    2.2607        1.6667
## subst2           3.8331    0.1603        3.6732
## subst4          12.4075    5.3096        7.0986
## censReg1         2.8892    0.0003        2.8891
## censReg2         3.1412    0.0006        3.1409
## censReg0impute   3.1444    0.0006        3.1441
## best             2.8939    0.0001        2.8941
```

```
##                mse_beta bias_beta variance_beta
## subst1           6.5713    2.2449        4.3269
## subst2           7.3837    0.1013        7.2831
## subst4          12.4701    0.7425       11.7287
## censReg1         7.6126    0.0001        7.6133
## censReg2         8.3107    0.0001        8.3114
## censReg0impute   8.3088    0.0001        8.3095
## best             7.4450    0.0002        7.4455
```

```
##                mse_beta bias_beta variance_beta
## subst1          10.7825    2.2270        8.5564
## subst2          12.9888    0.4014       12.5887
## subst4          18.2077    0.0507       18.1587
## censReg1        15.2039    0.0002       15.2052
## censReg2        16.5719    0.0001       16.5735
## censReg0impute  16.5751    0.0001       16.5767
## best            14.8431    0.0035       14.8411
```

Further comparisons between `subst2` and `censReg1`

From our previous results, `subst2` is generally the best performing substitution method and `censReg1` is the best censReg method. In the previous section, these methods gave similar MSE values for sd28_153 = 0.5, so we will fix this parameter at this value and investigate these estimation methods for four values of cprop: 0.1, 0.3, 0.5, 0.7. These cprop values correspond to censoring 10 %, 30 %, 50 %, and 70% of the data respectively, so they correspond to decreasing values of LOD, which is our variable of primary interest.

We see that `censReg1` gives estimates with very low bias for all values of `cprop` , whereas the bias from `subst2` increases greatly as `cprop` increases. We interpret this as meaning that the real cb28 values are unchanged when LOD is lowered, which means that a higher proportion are likely to lie closer to LOD for larger values of `cprop` which means that substituted values are increasingly biased towards being too small as `cprop` increases. Since `censReg1` fits a model to all the data (censored and uncensored) it maintains low bias as the LOD decreases, whilst the variance remains approximately constant. However, as a greater proportion of values are substituted for the same constant value by the `subst2` method, the variance decreases because a higher proportion of the data values are identical.

In conclusion, `censReg1` gives similar bias and variance for different values of cprop whereas `subst2` does not. From `subst2` the bias increases and the variance decreases as `cprop` increases.

```
##          mse_beta bias_beta variance_beta
## subst2     8.0879    0.0041        8.0846
## censReg1   7.6265    0.0011        7.6261
## best       7.7061    0.0002        7.7067
```

```
##          mse_beta bias_beta variance_beta
## subst2     6.7352    1.3153        5.4205
## censReg1   7.4995    0.0001        7.5002
```

```
##          mse_beta bias_beta variance_beta
## subst2     6.8178    1.3197        5.4987
## censReg1   7.6068    0.0002        7.6074
```

```
##          mse_beta bias_beta variance_beta
## subst2     8.6654    5.3095        3.3562
## censReg1   8.0983    0.0001        8.0991
```

The MSE, squared-bias and variance of predictions of `cb28` annual means from various censoring methods

All the graphs in this section will shows MSE, squared-bias, or variance on the y-axis and year on the x-axis for the simulated 15-year period. We begin by looking at variance of predictions from our best three substitution methods, best three cenreg methods. We will again use `best` as our gold standard.

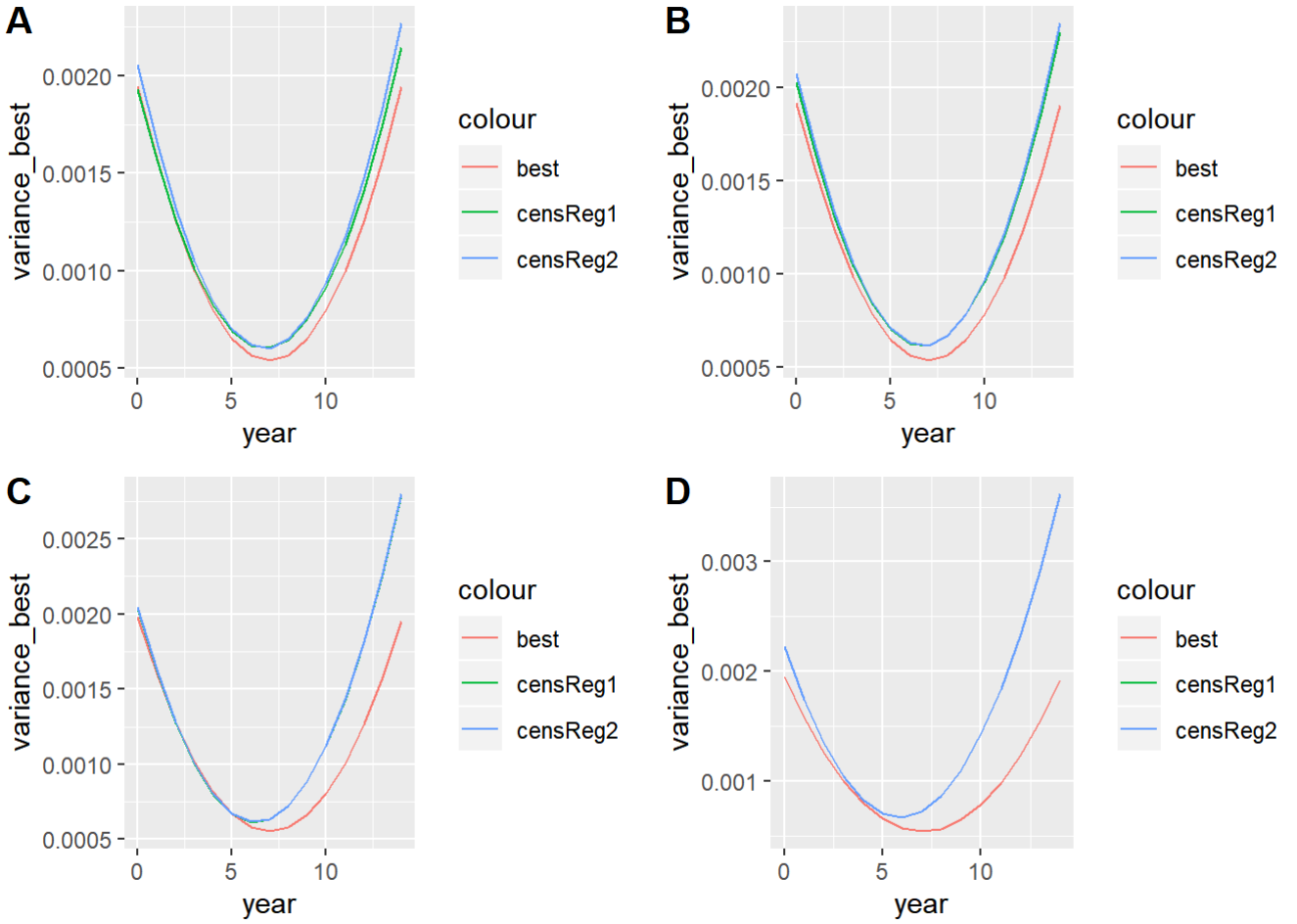
Variance of predictions of cb28 annual means from different methods

Predictions for different values of beta153year

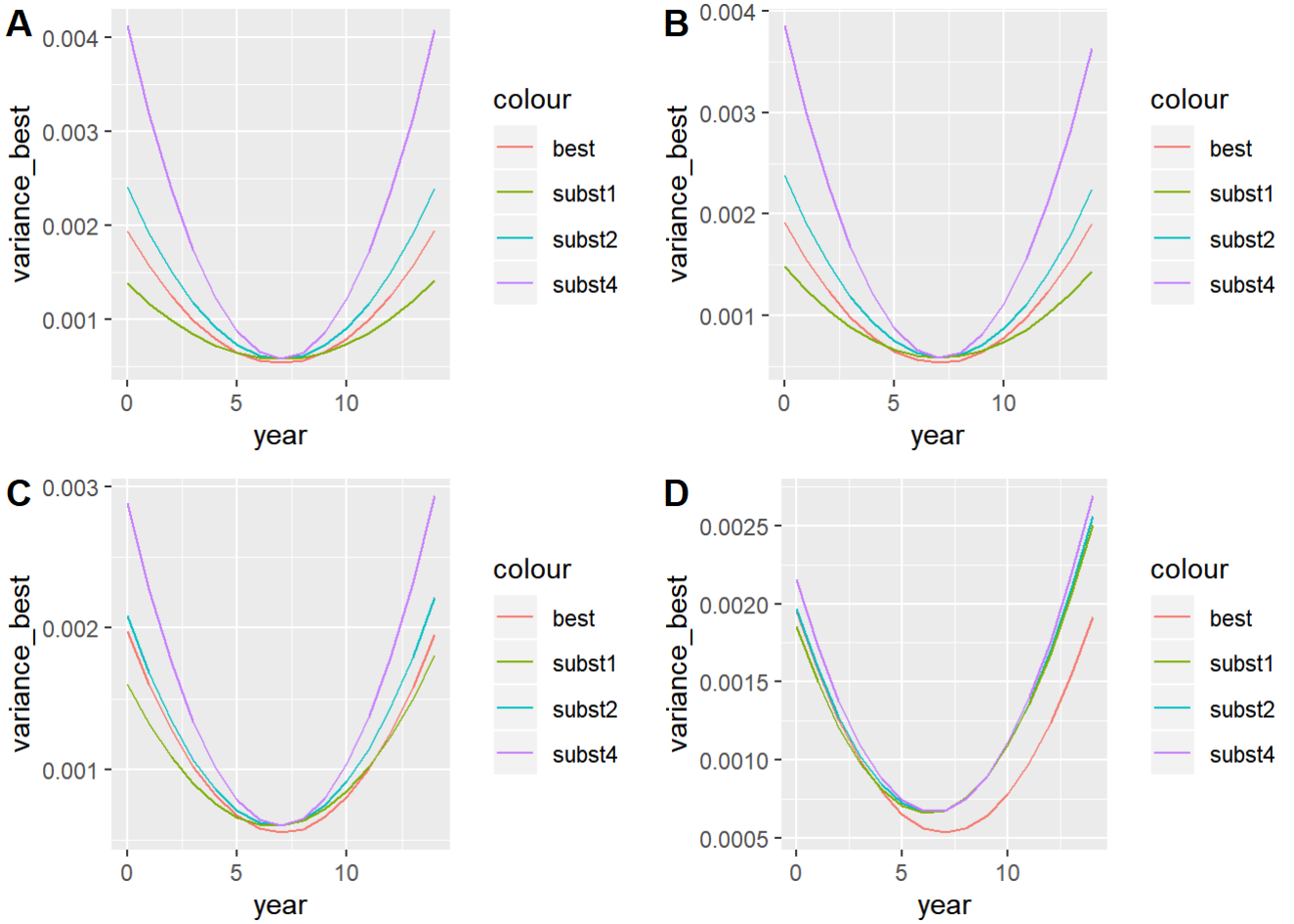
We will begin by using the same parameter values we used in our earlier section “Evaluation of methods for larger absolute values of `beta28year` ”. These parameters are fixed: cprop = 0.3, sd28_153 = 0.3, whilst cb153year is given four values: -0.02, -0.04, -0.08 and -0.16 respectively.

We begin by showing graphs of the variance of predictions of cb28 annual means from our chosen censoring methods. A common feature of all these graphs is that they typically have an approximately parabolic “U” shape, with higher variance at each end of the time period than in the middle of the period. This is in accordance with our prior expectations because this is generally the case.

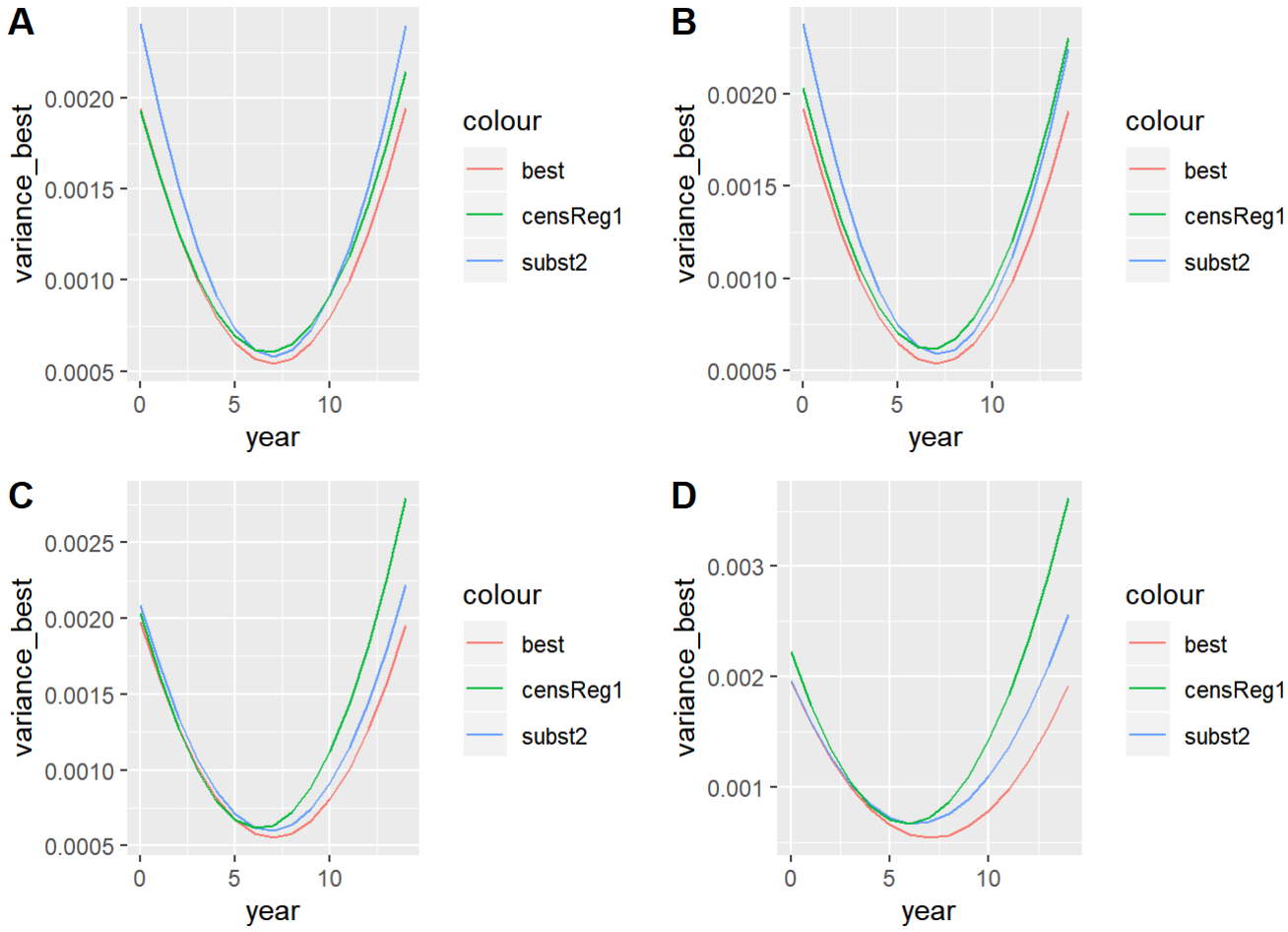
Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for beta153year equal to -0.02, -0.04, -0.08, -0.16, respectively.



Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for beta153year equal to -0.02, -0.04, -0.08, -0.16, respectively.

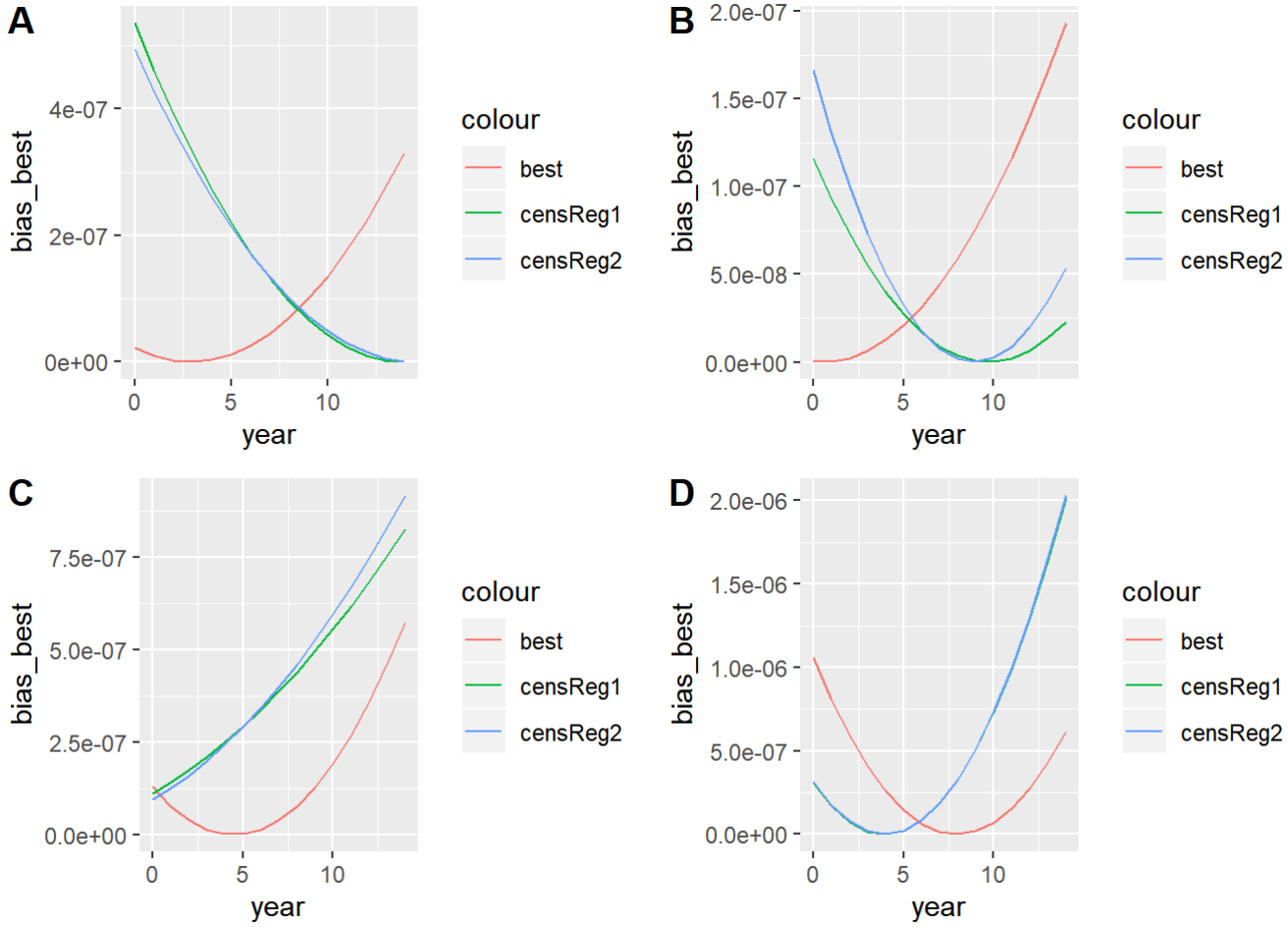


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

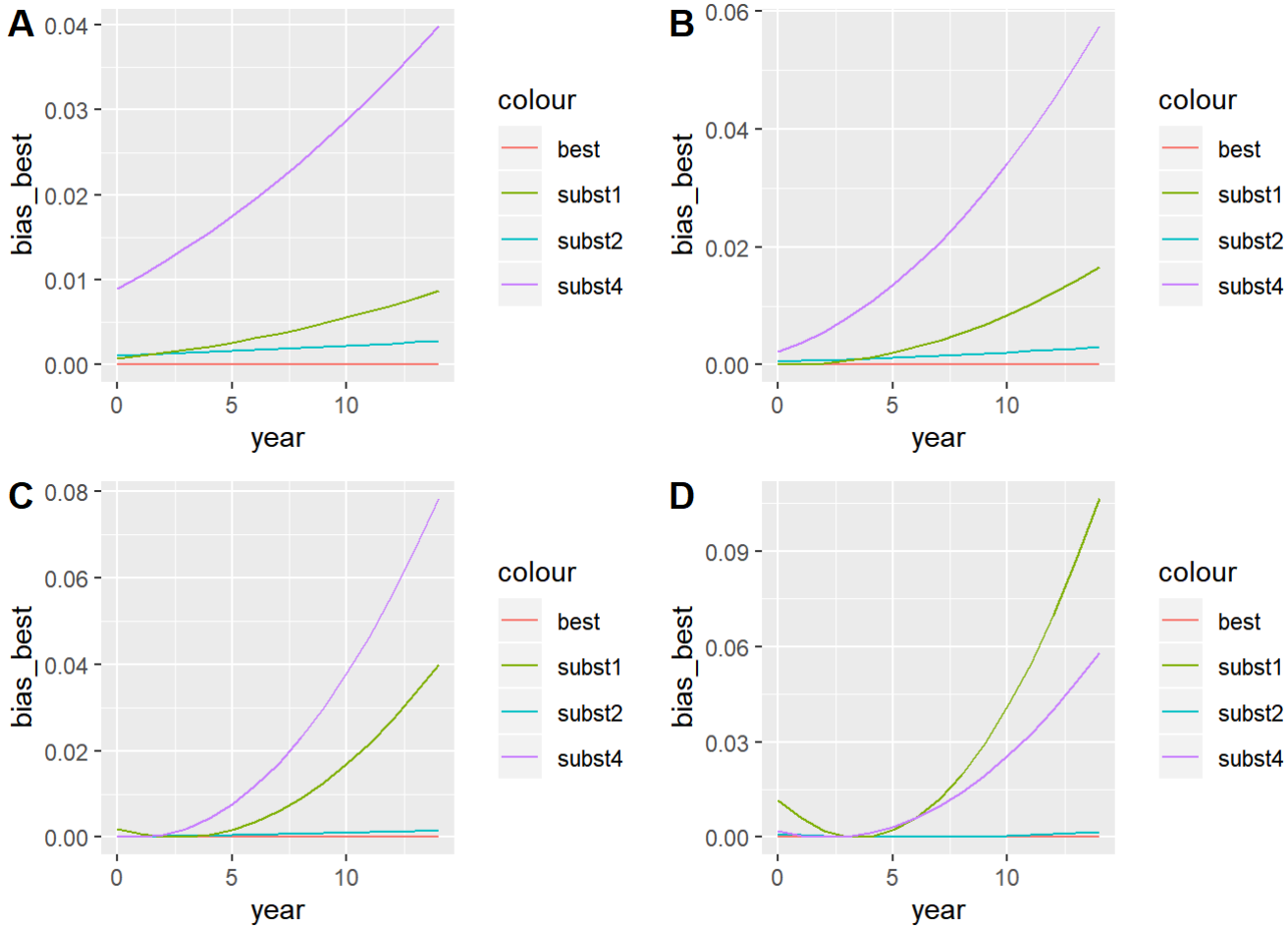


We will now show graphs of the bias of predictions of cb28 annual means from our chosen censoring methods.

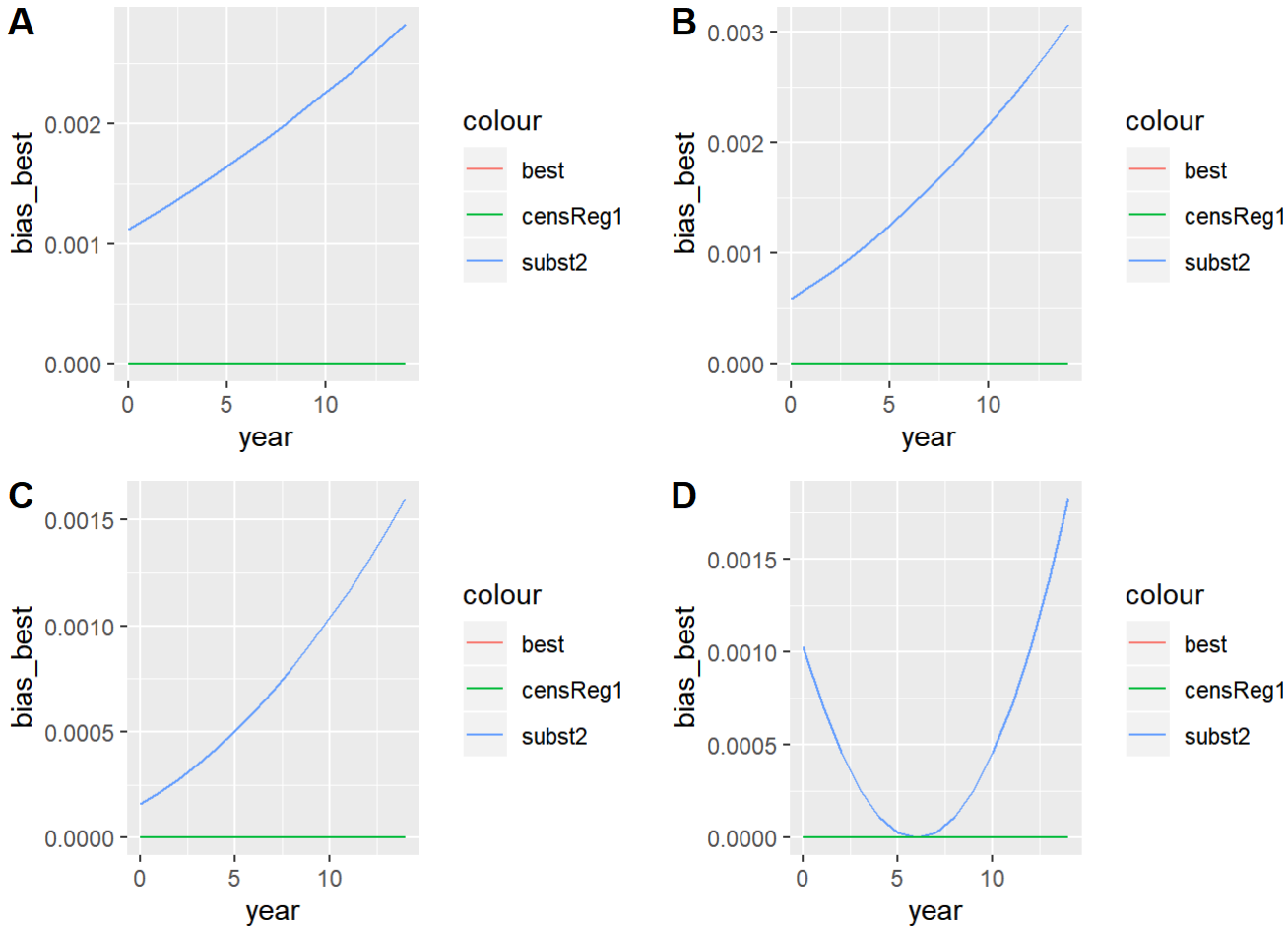
Our first set of four graphs show the bias of `censReg1` and `censReg2` methods relative to `best` method for beta153year equal to -0.02, -0.04, -0.08, -0.16, respectively.



Our second set of four graphs show the bias of `subst1`, `subst2` and `subst4` methods relative to `best` method for beta153year equal to -0.02, -0.04, -0.08, -0.16, respectively.

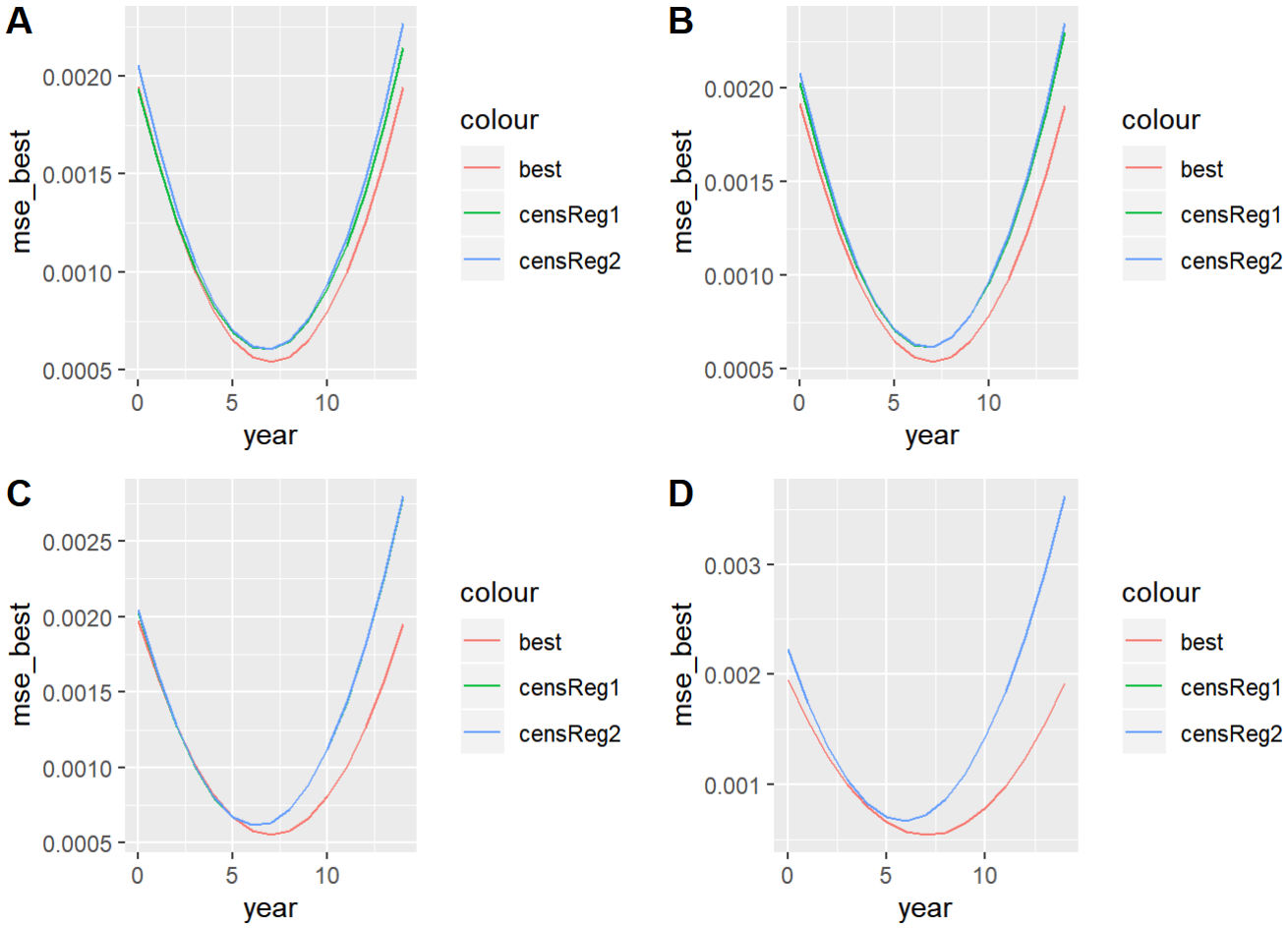


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

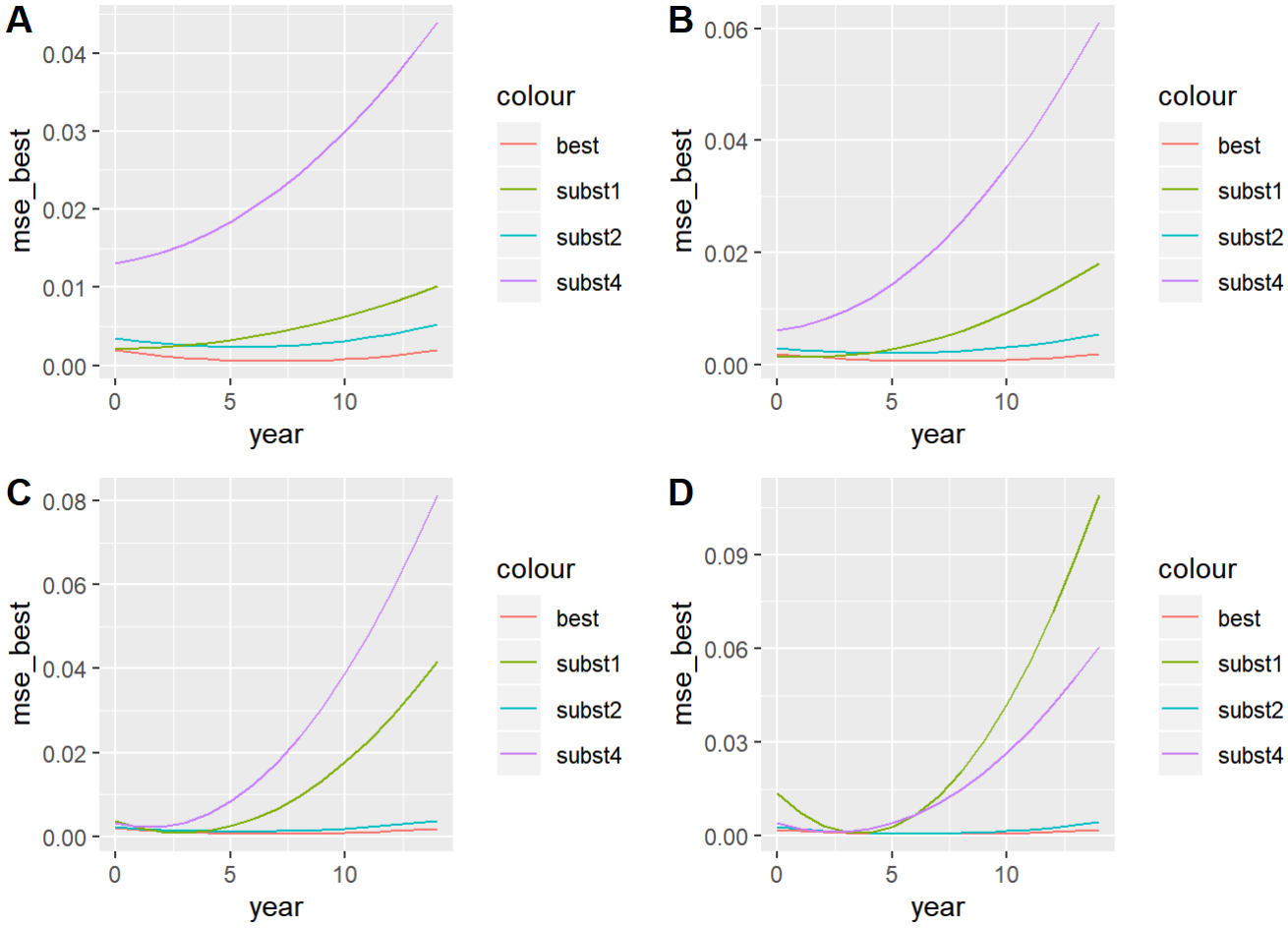


We will now show graphs of the MSE of predictions of cb28 annual means from our chosen censoring methods.

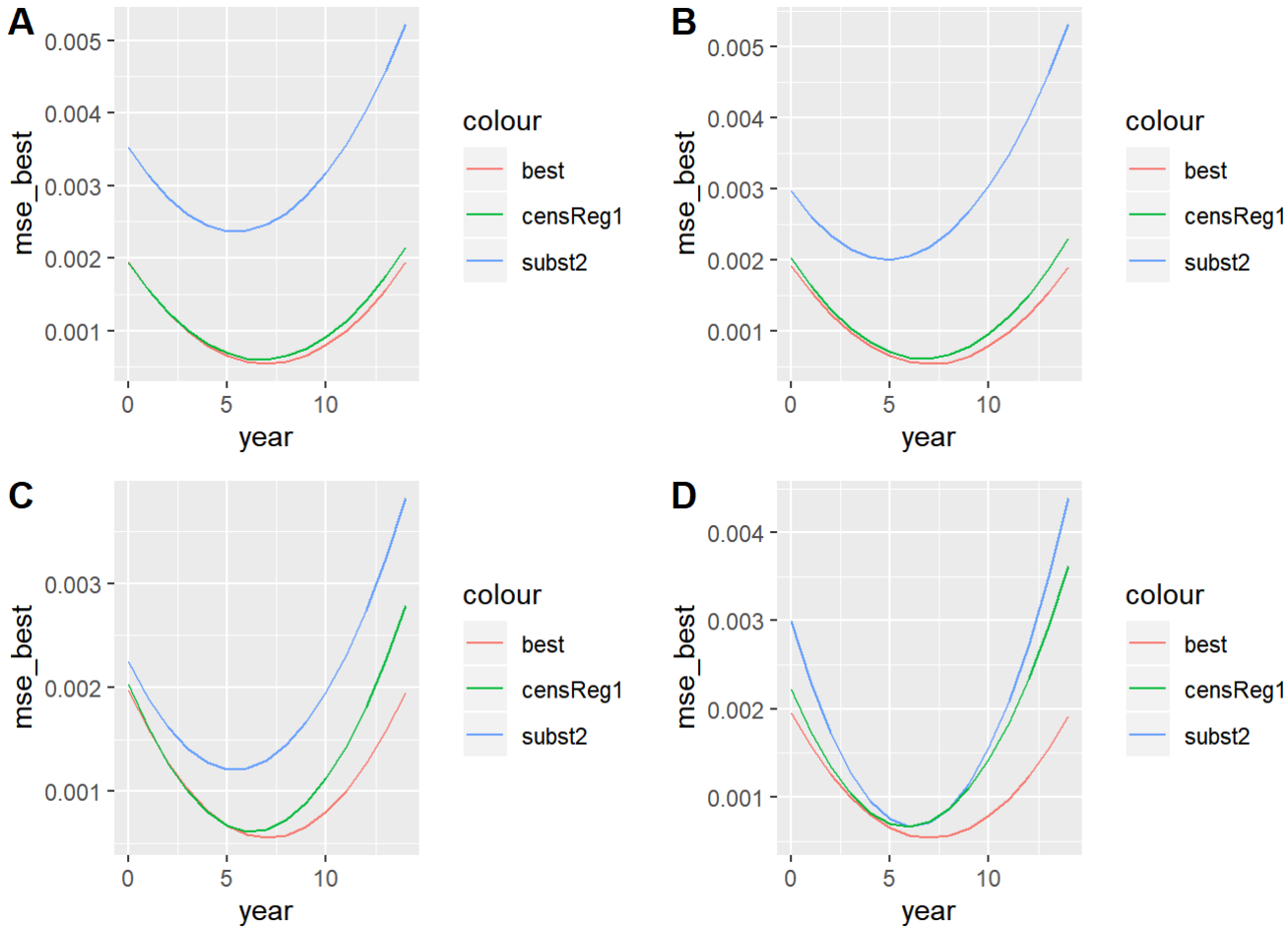
Our first set of four graphs show the MSE of `censReg1` and `censReg2` methods relative to `best` method for beta153year equal to -0.02, -0.04, -0.08, -0.16, respectively.



Our second set of four graphs show the MSE of `subst1`, `subst2` and `subst4` methods relative to `best` method for beta153year equal to -0.02, -0.04, -0.08, -0.16, respectively.



Our third set of four graphs simply displays the MSE from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

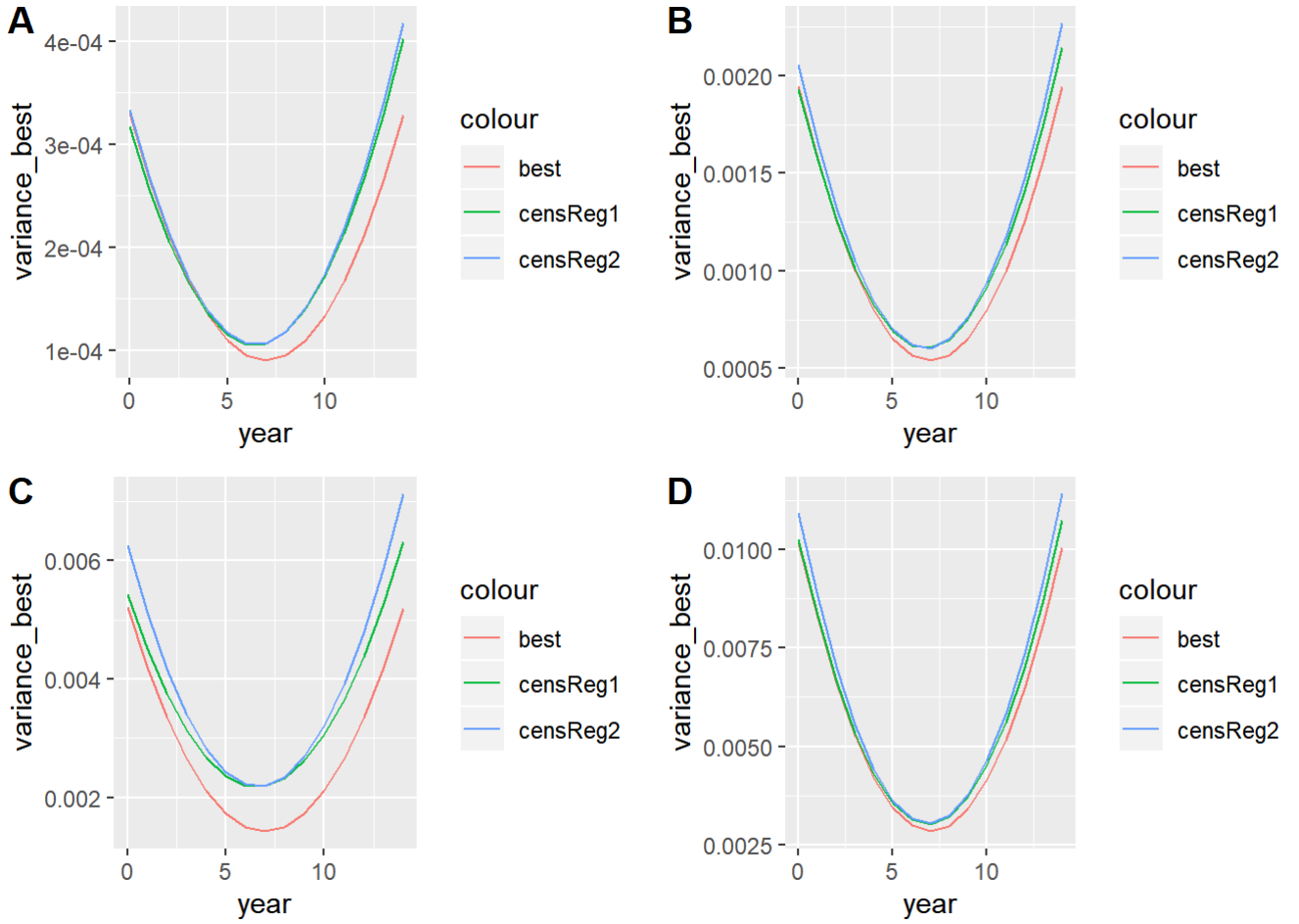


Predictions for different values of sd28vs153

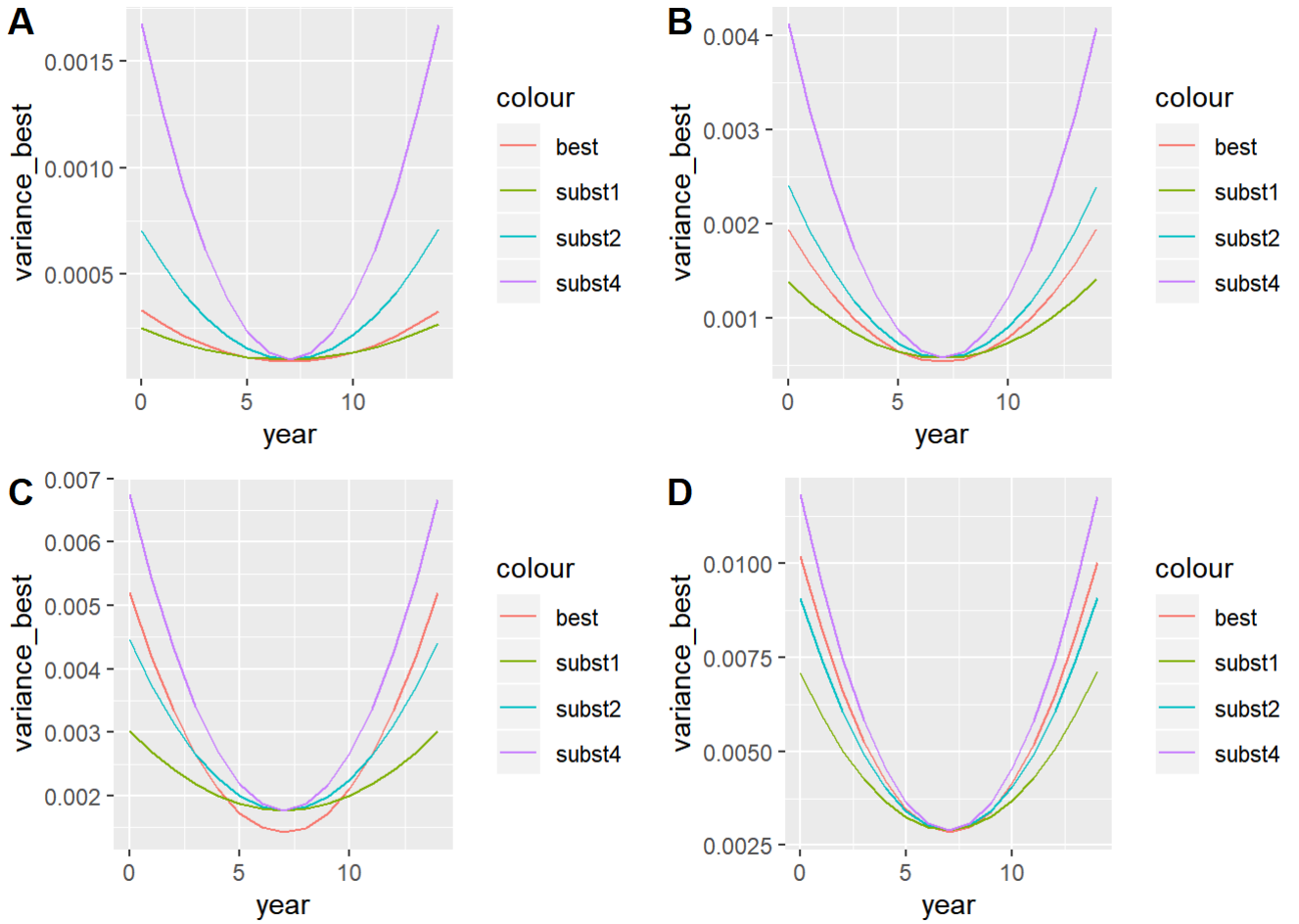
For all our predictions in this section, these parameters are fixed: cprop = 0.3, cb153year = -0.02, whilst sd28_153 is given four values: 0.1, 0.3, 0.5 and 0.7 respectively.

We begin by showing graphs of the variance of predictions of cb28 annual means from our chosen censoring methods. A common feature of all these graphs is that they typically have an approximately parabolic “U” shape, with higher variance at each end of the time period than in the middle of the period. This is in accordance with our prior expectations because this is generally the case.

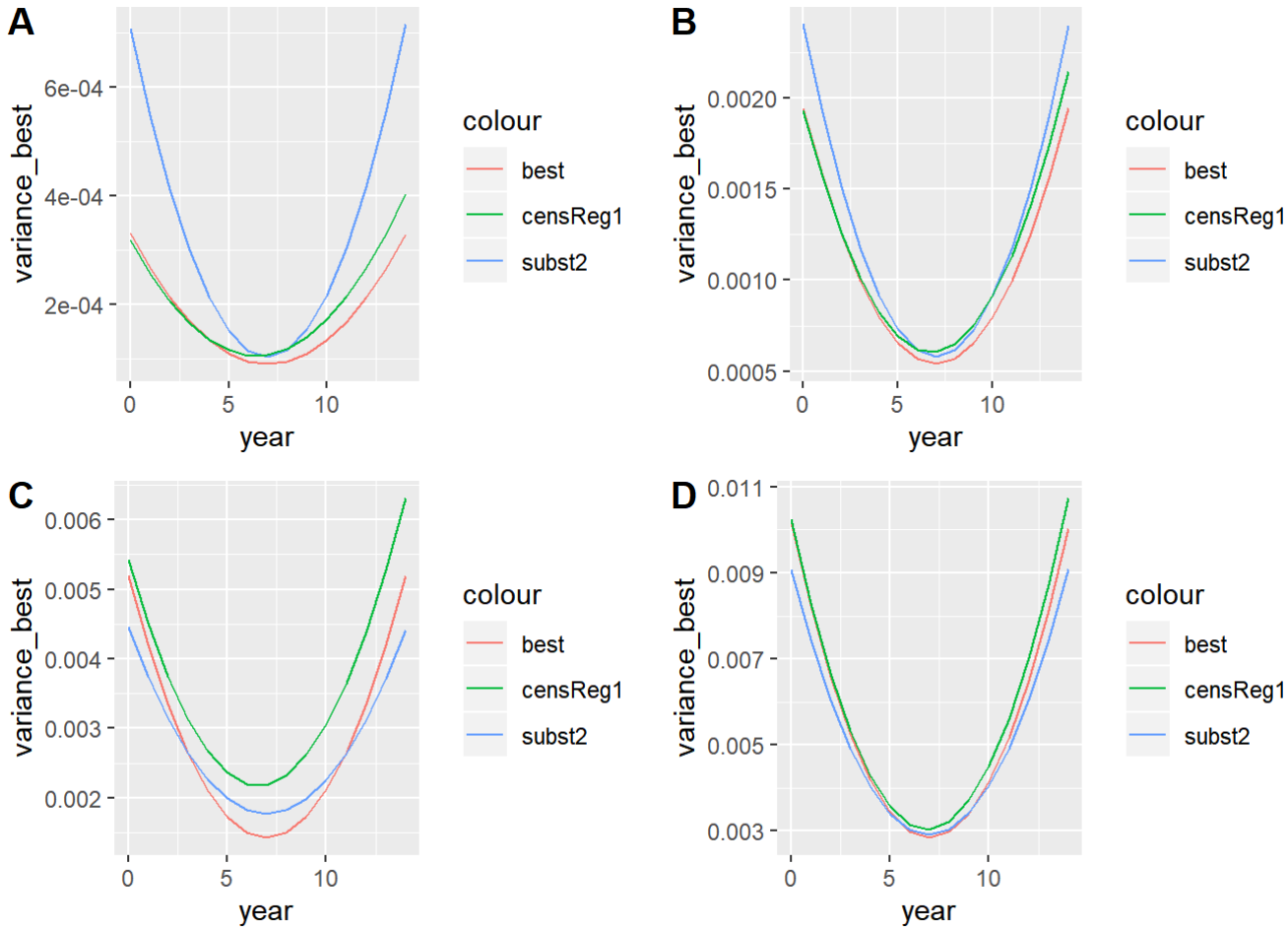
Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for sd28_153 equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for sd28_153 equal to 0.1, 0.3, 0.5, 0.7, respectively.

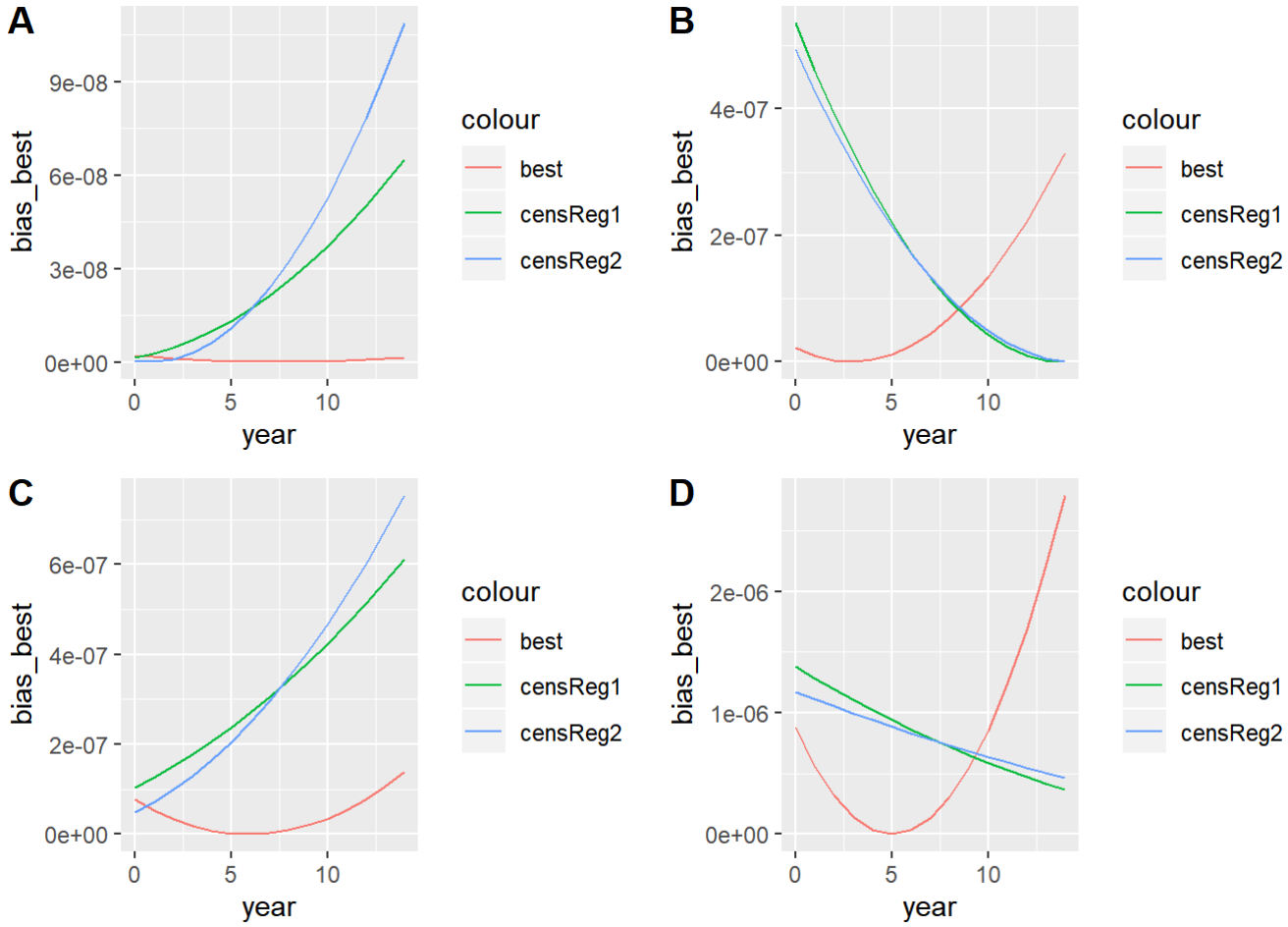


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

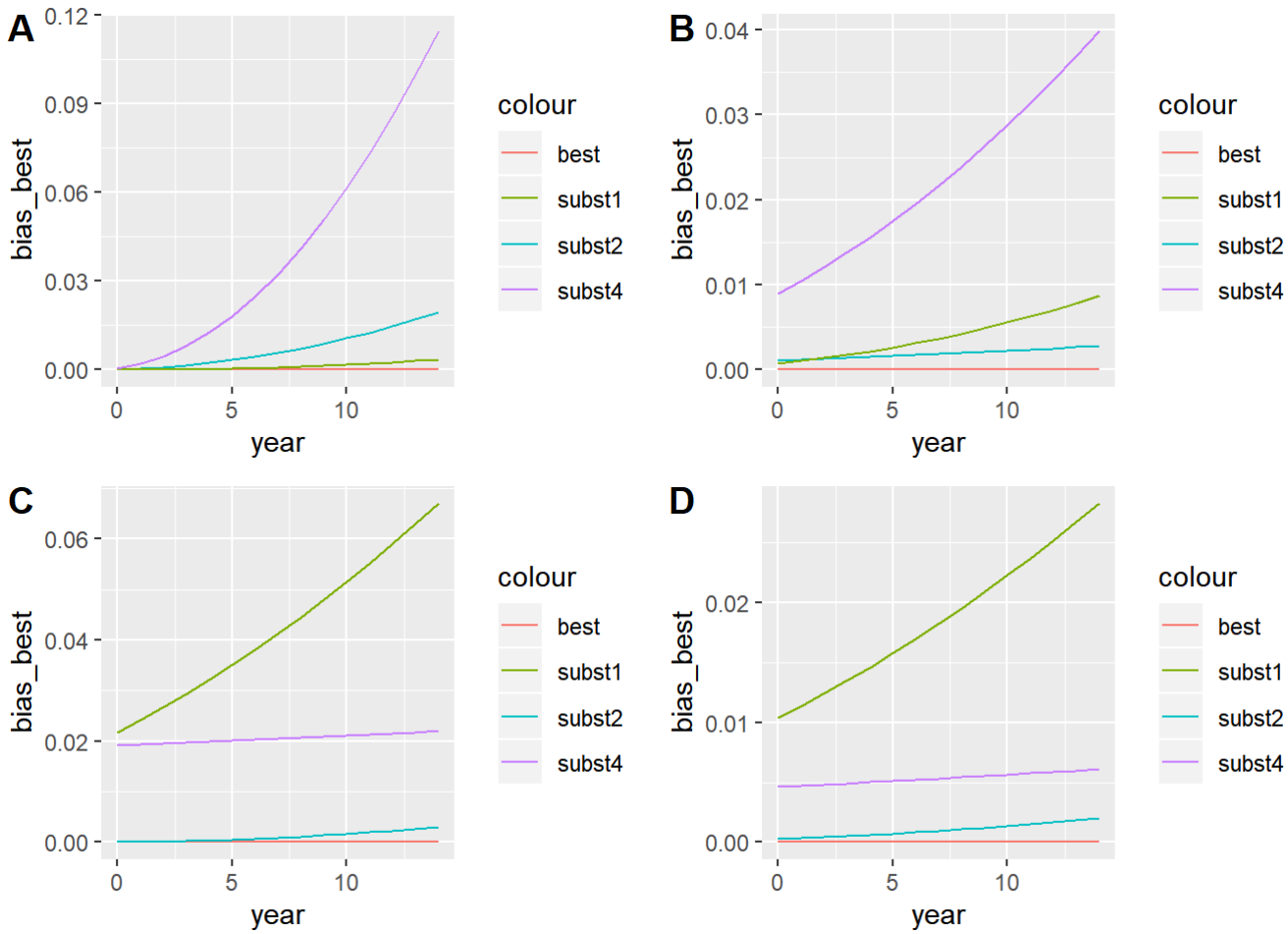


We will now show graphs of the bias of predictions of cb28 annual means from our chosen censoring methods.

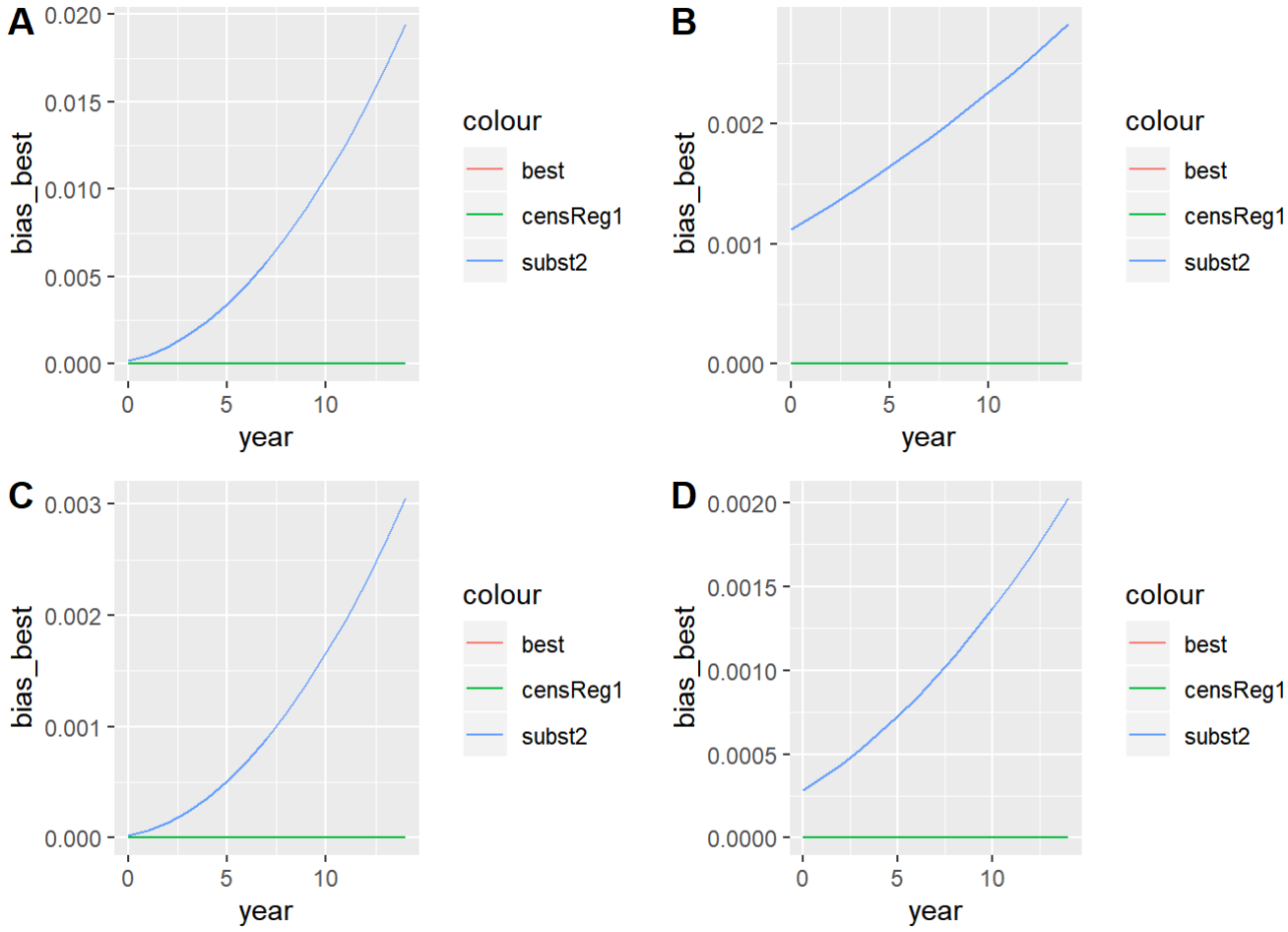
Our first set of four graphs show the bias of `censReg1` and `censReg2` methods relative to `best` method for sd28_153 equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the bias of `subst1`, `subst2` and `subst4` methods relative to `best` method for sd28_153 equal to 0.1, 0.3, 0.5, 0.7, respectively.

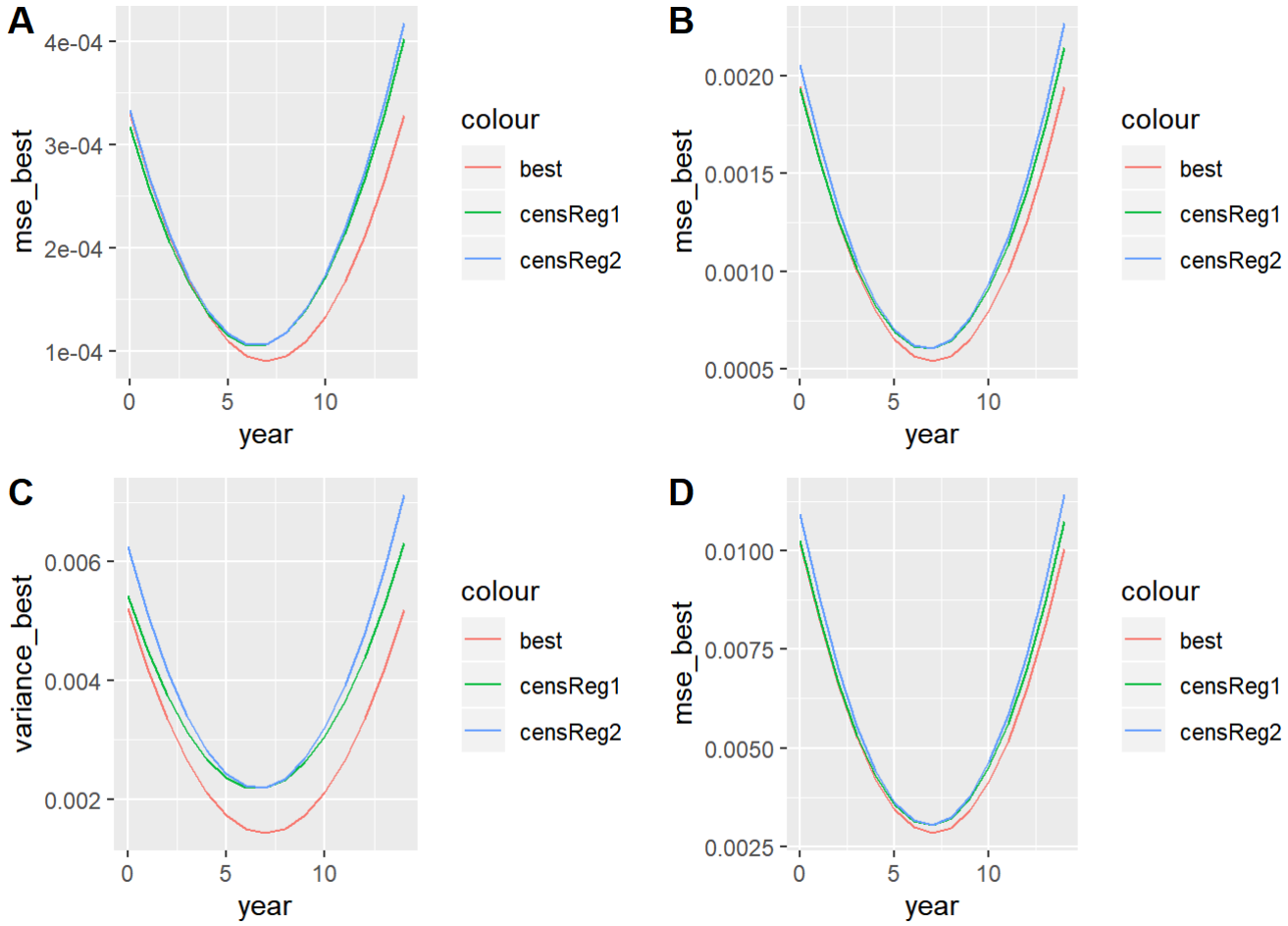


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

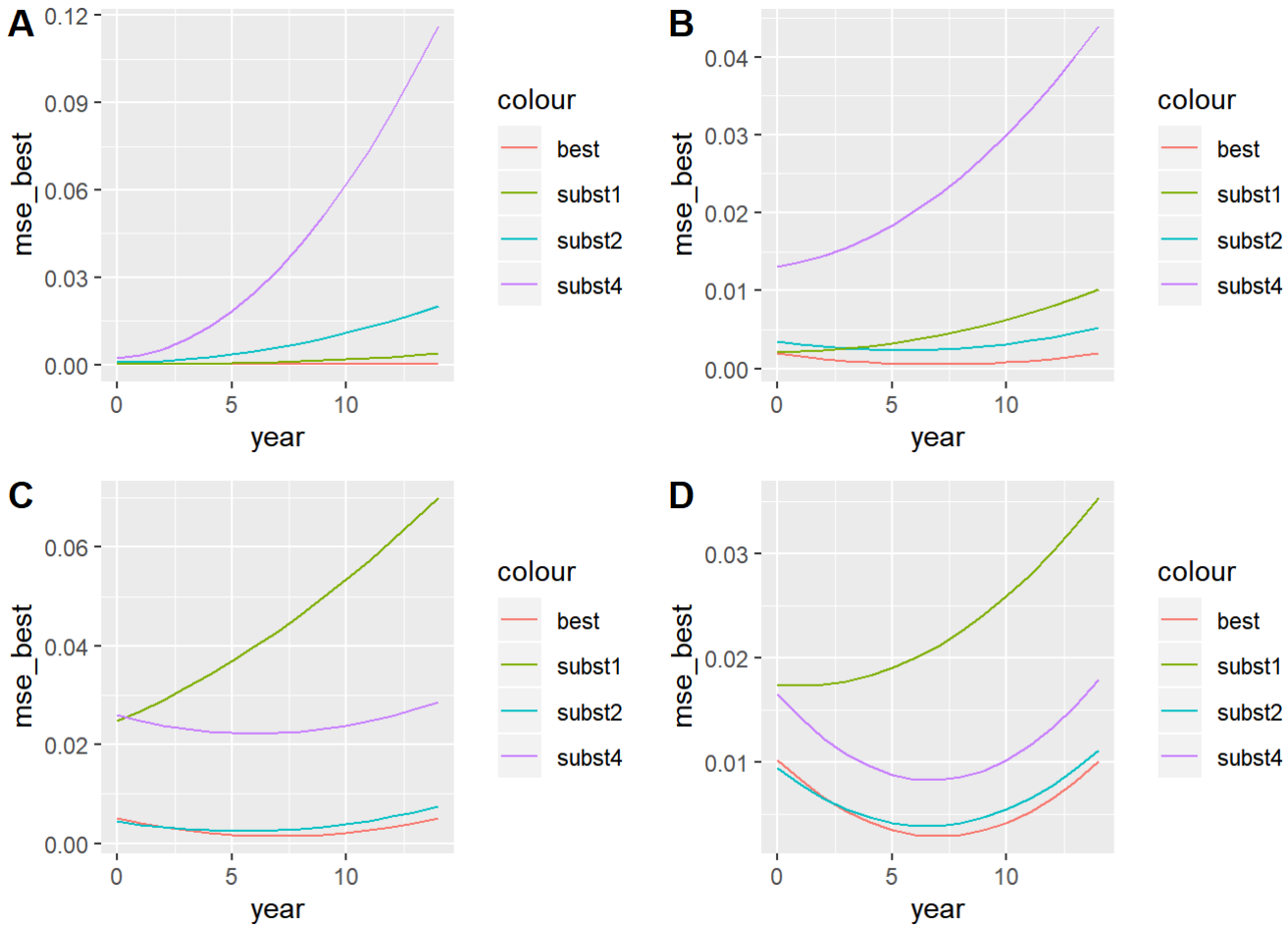


We will now show graphs of the MSE of predictions of cb28 annual means from our chosen censoring methods.

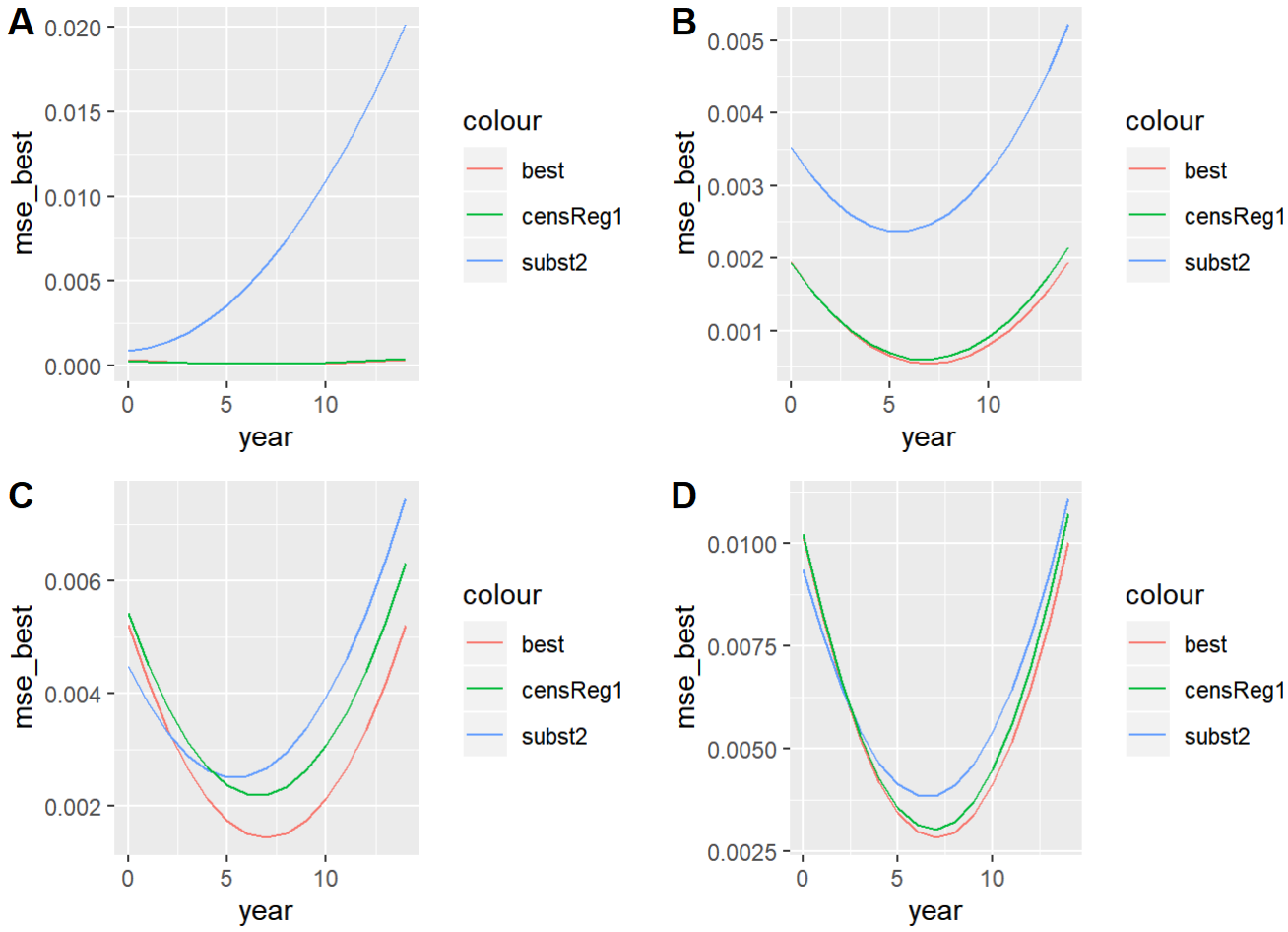
Our first set of four graphs show the MSE of `censReg1` and `censReg2` methods relative to `best` method for sd28_153 equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the MSE of `subst1`, `subst2` and `subst4` methods relative to `best` method for sd28_153 equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our third set of four graphs simply displays the MSE from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

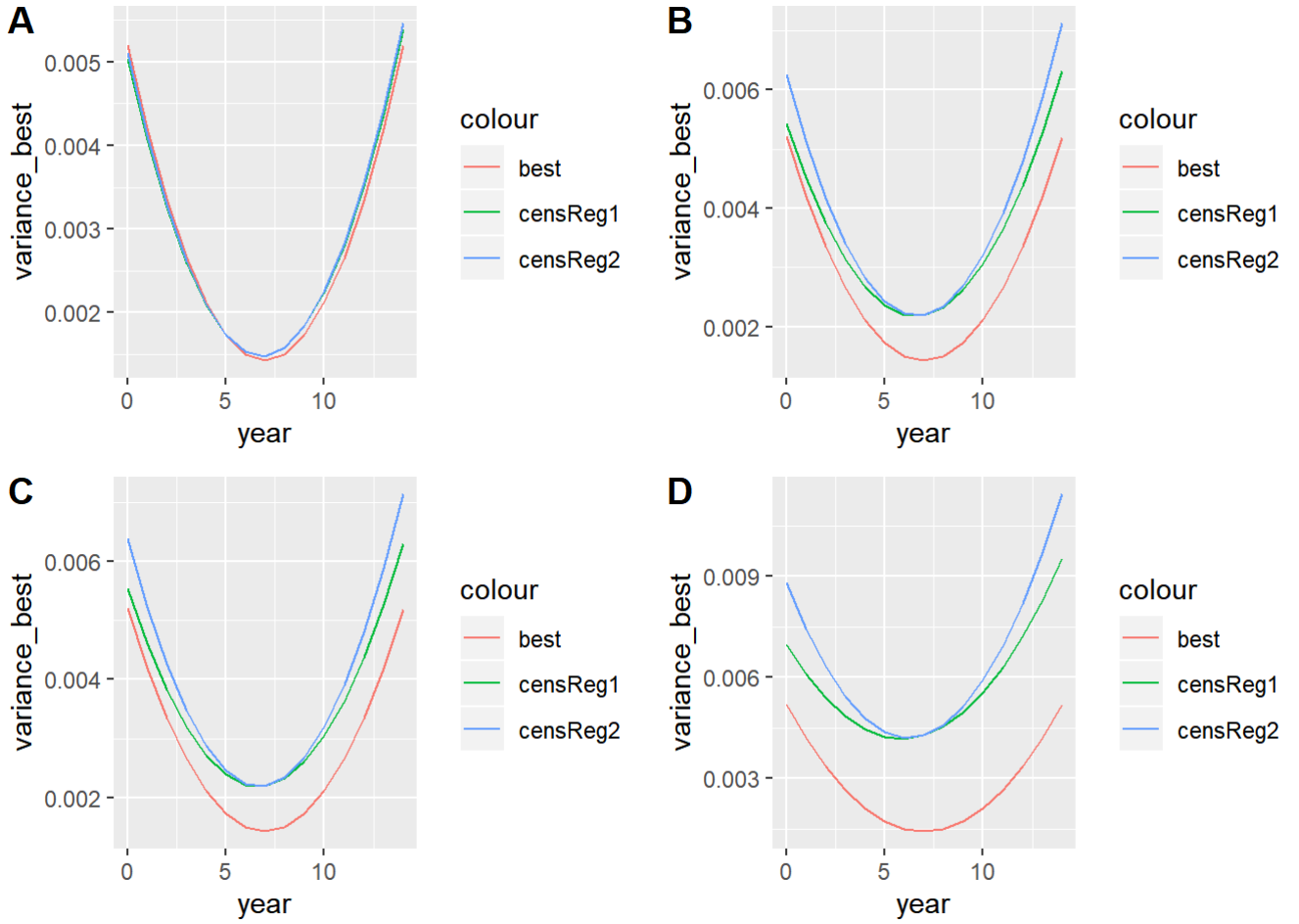


Predictions for different values of cprop

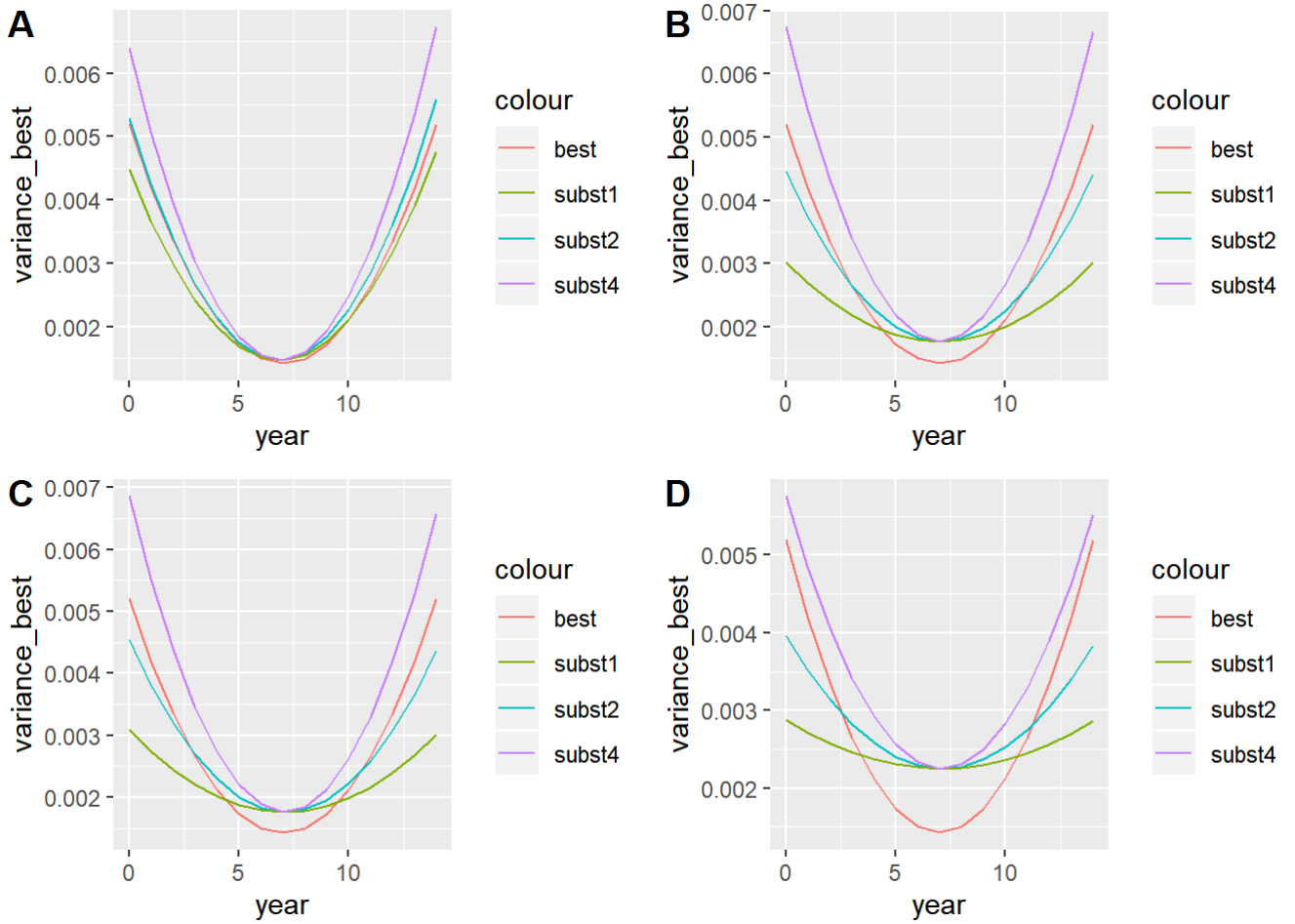
For all our predictions in this section, these parameters are fixed: sd28_153 = 0.5, cb153year = -0.02, whilst cprop is given four values: 0.1, 0.3, 0.5 and 0.7 respectively.

We begin by showing graphs of the variance of predictions of cb28 annual means from our chosen censoring methods.

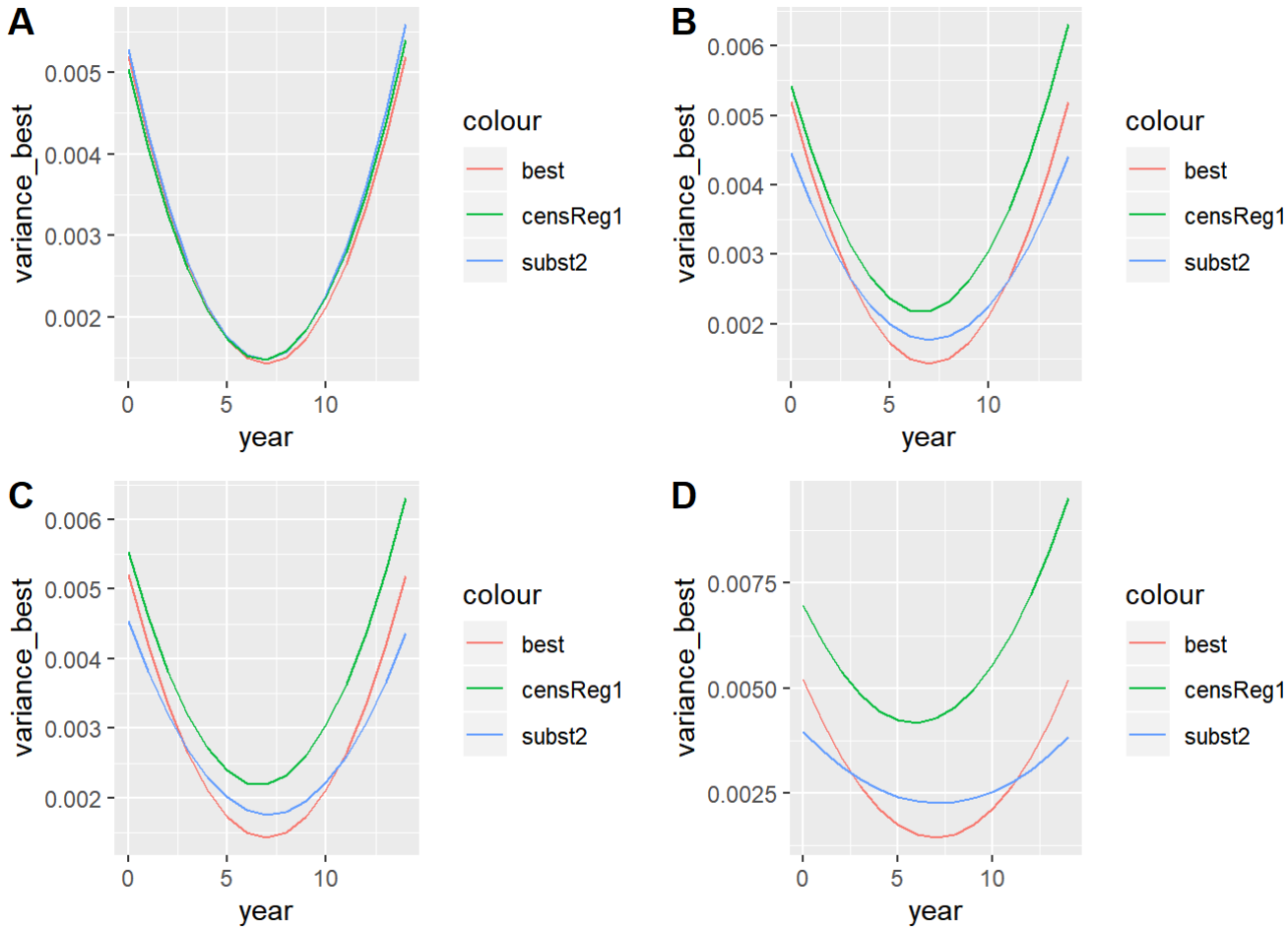
Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for cprop equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for cprop equal to 0.1, 0.3, 0.5, 0.7, respectively.

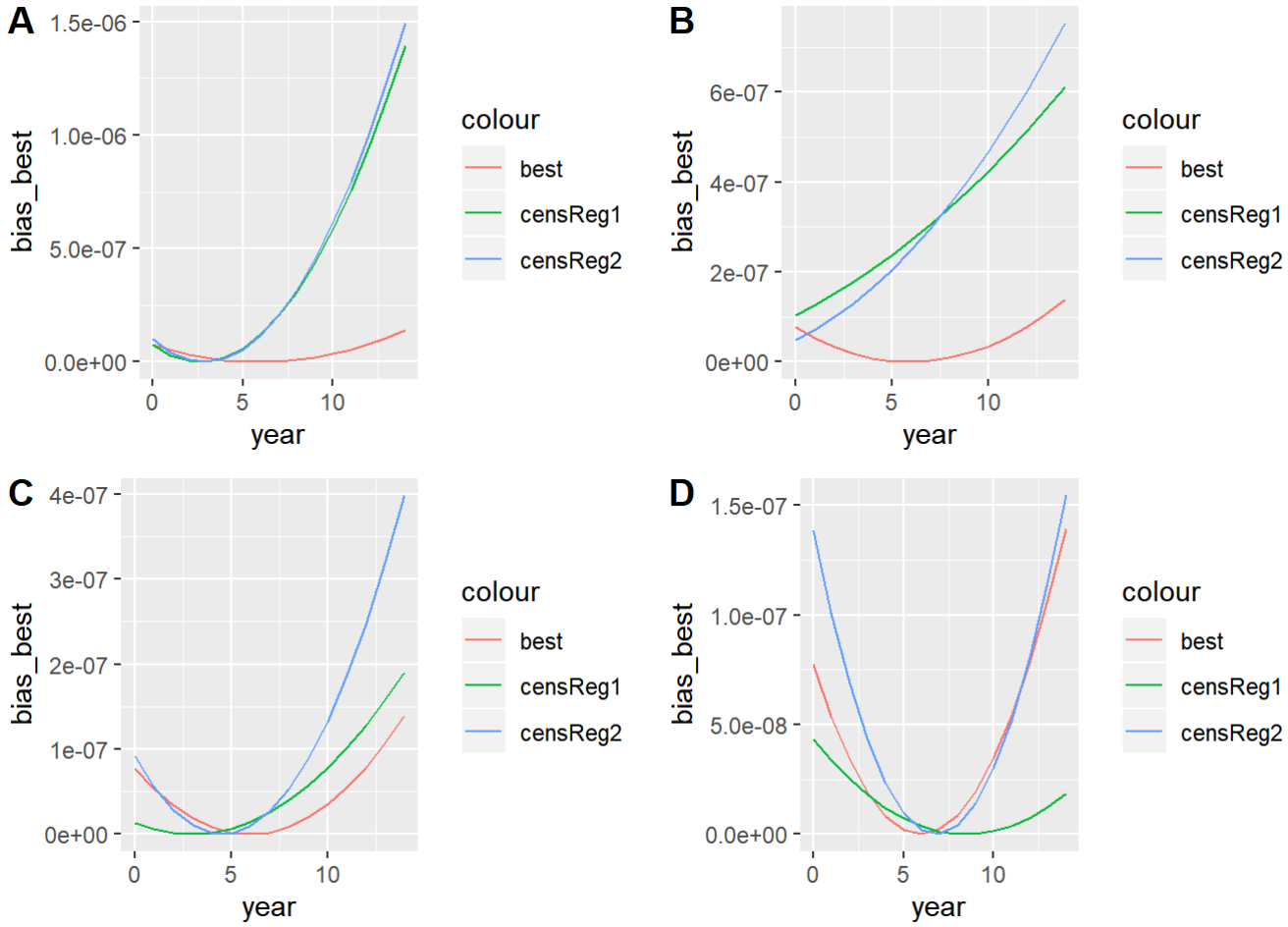


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

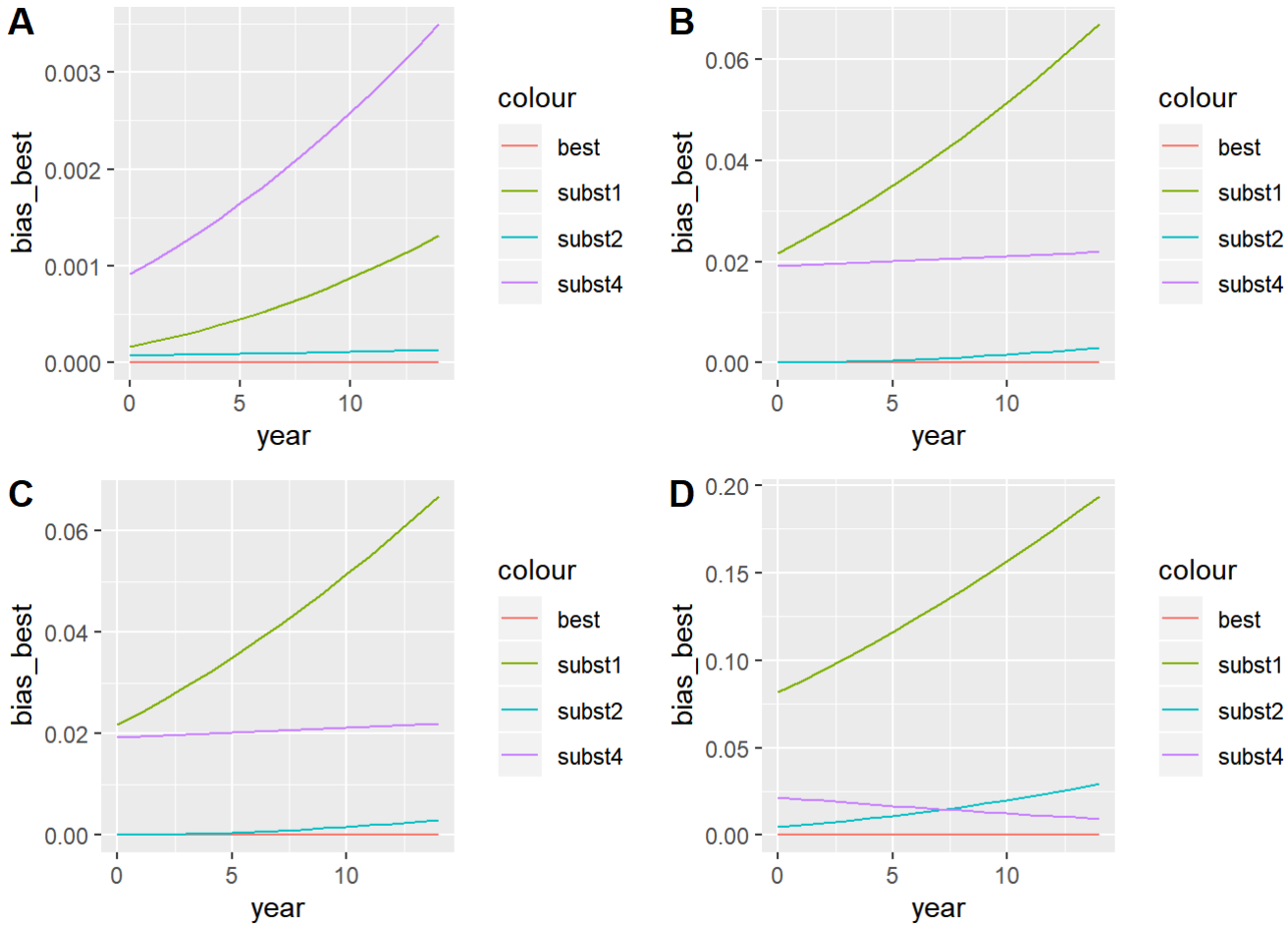


We will now show graphs of the bias of predictions of cb28 annual means from our chosen censoring methods.

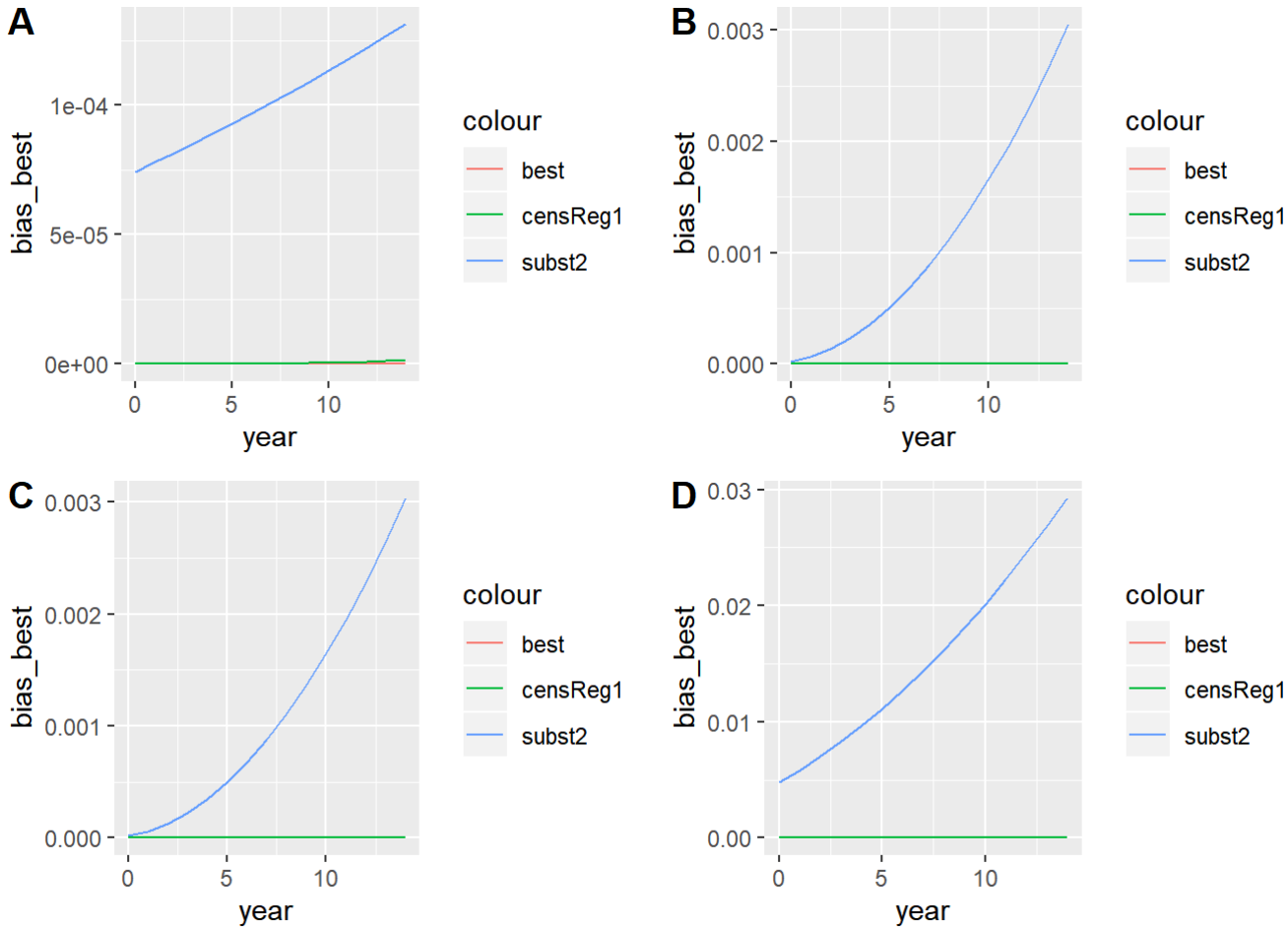
Our first set of four graphs show the bias of `censReg1` and `censReg2` methods relative to `best` method for cprop equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the bias of `subst1`, `subst2` and `subst4` methods relative to `best` method for cprop equal to 0.1, 0.3, 0.5, 0.7, respectively.

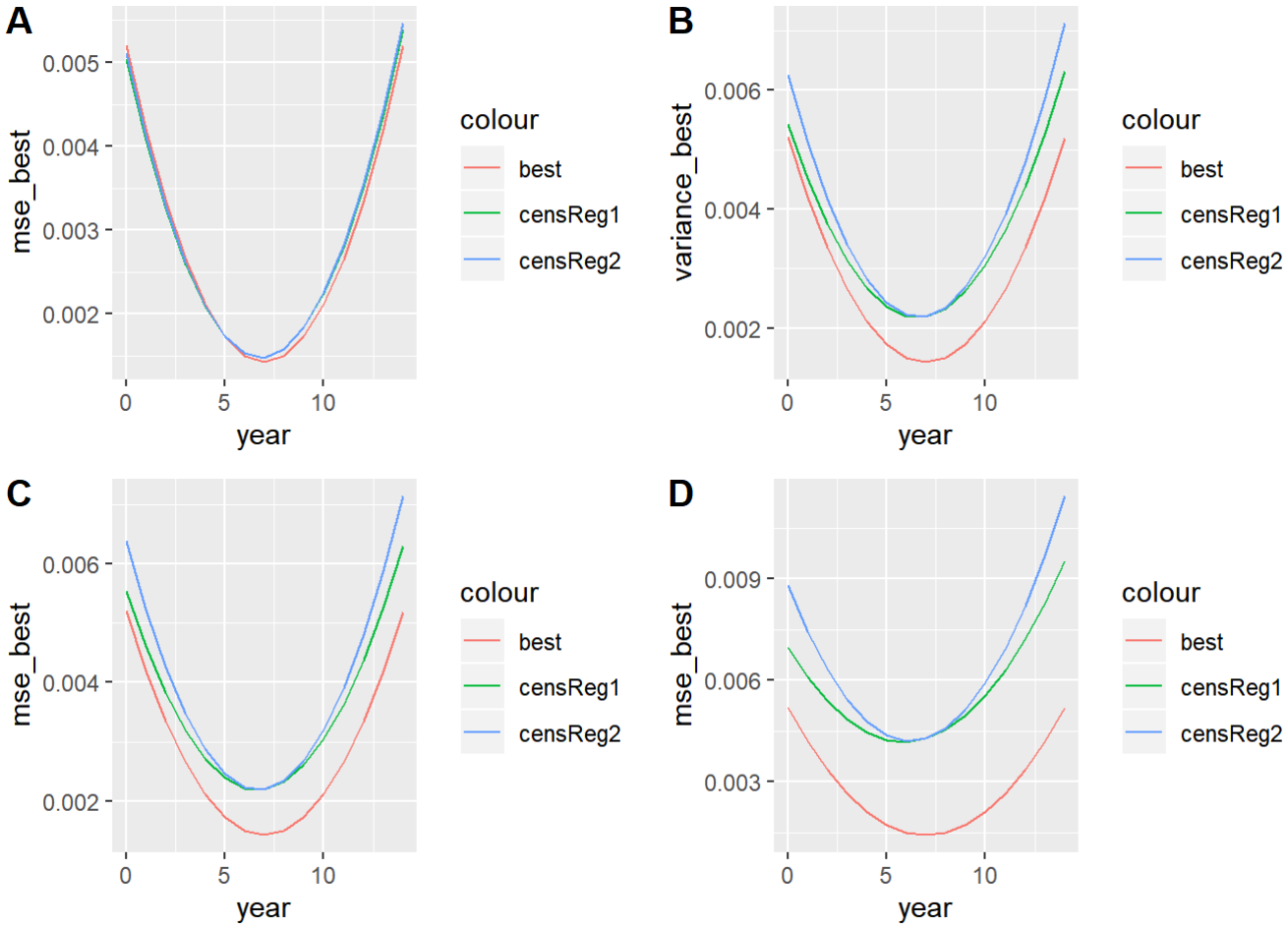


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

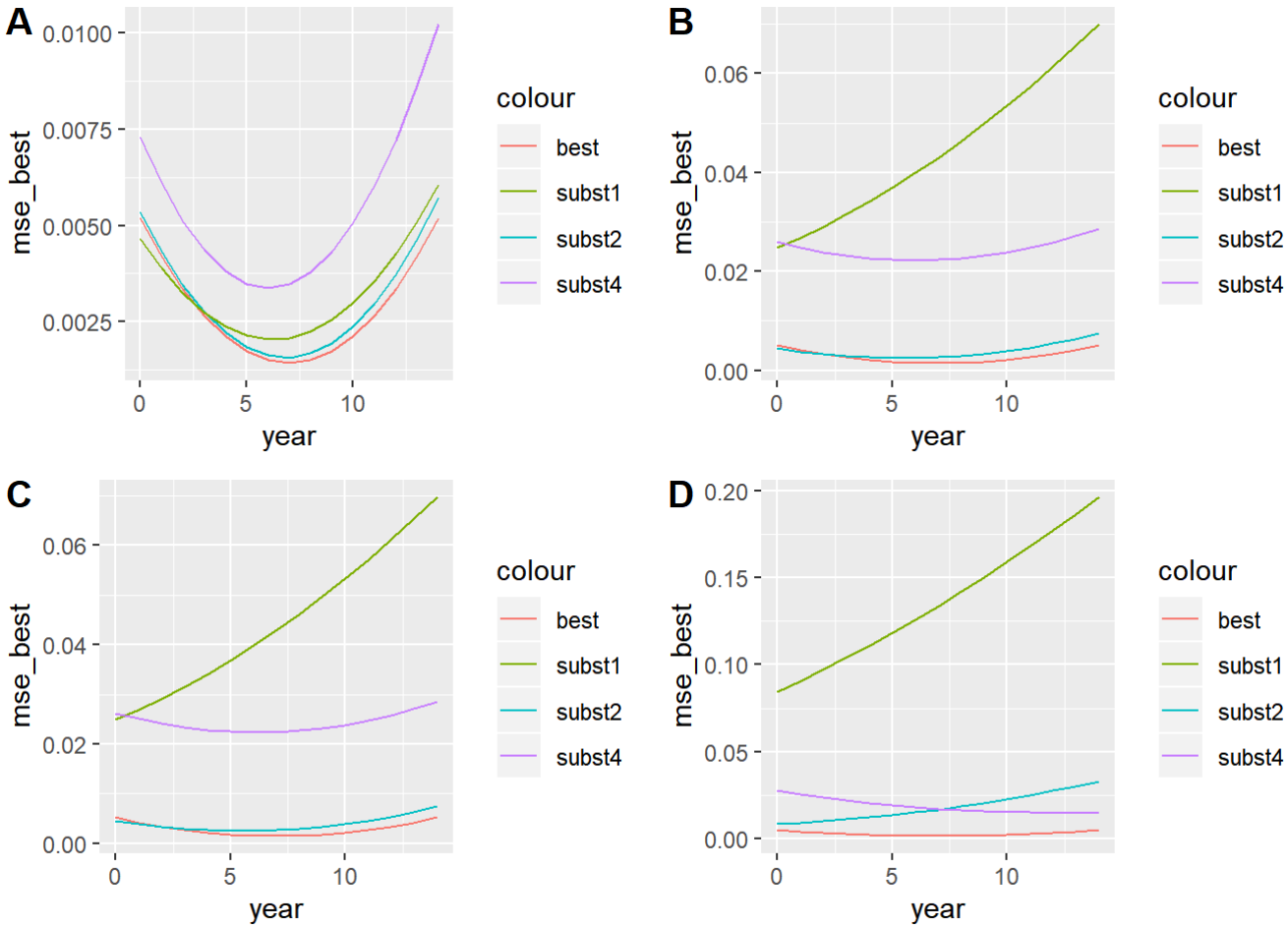


We will now show graphs of the MSE of predictions of cb28 annual means from our chosen censoring methods.

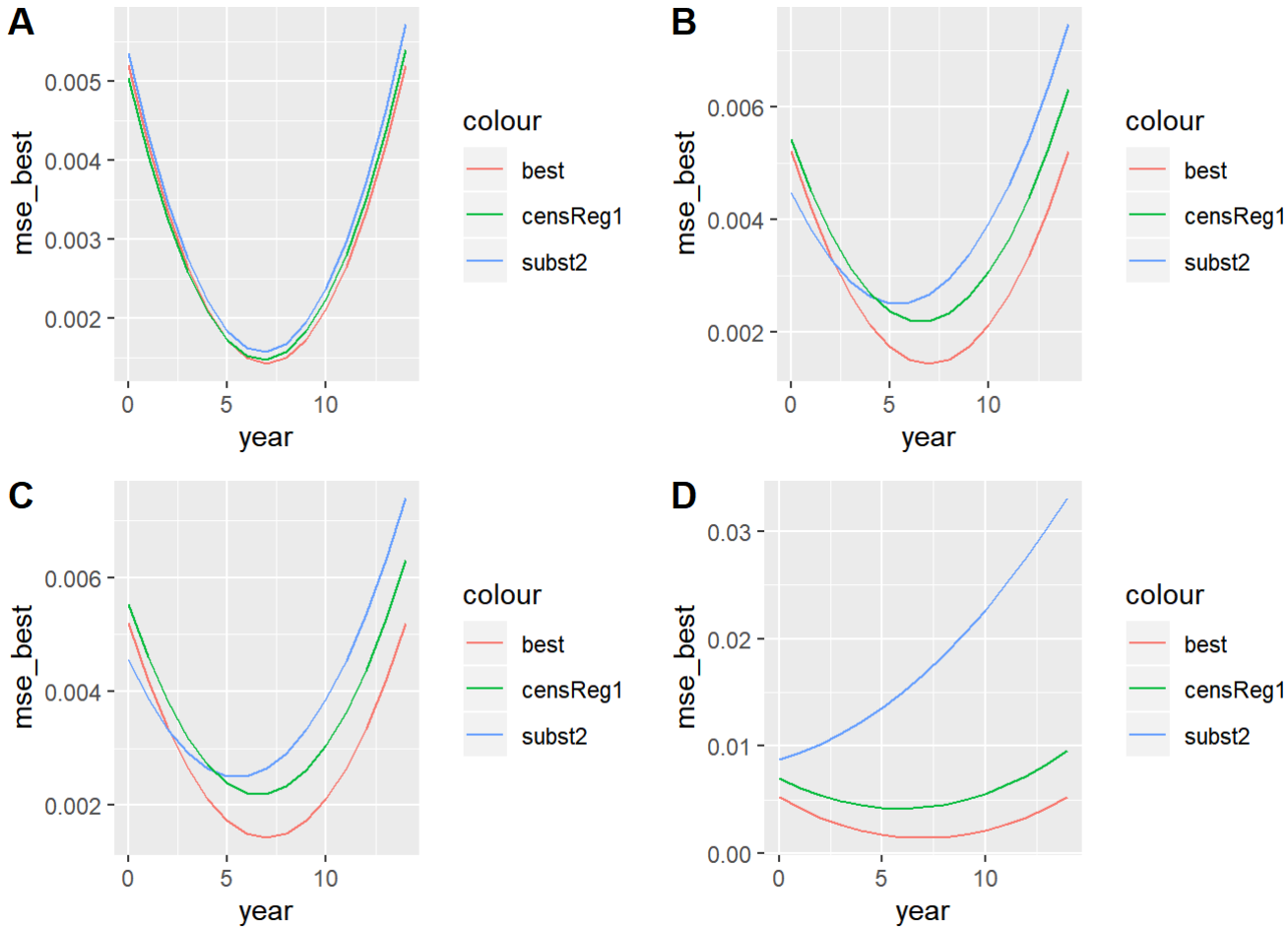
Our first set of four graphs show the MSE of `censReg1` and `censReg2` methods relative to `best` method for cprop equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the MSE of `subst1`, `subst2` and `subst4` methods relative to `best` method for cprop equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our third set of four graphs simply displays the MSE from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

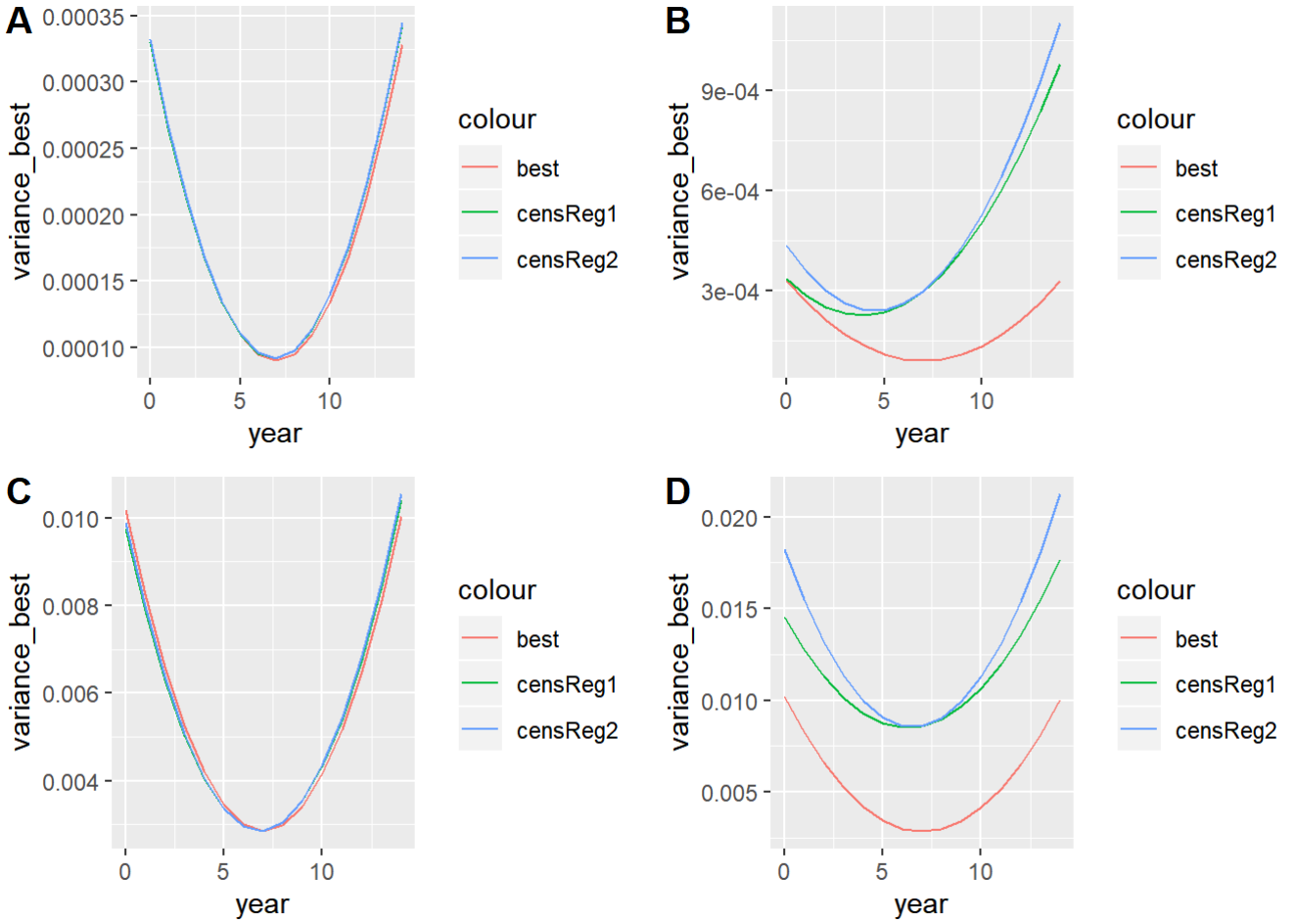


Predictions for low-low, high-low, low-high, high-high values of sd28vs153-cprop

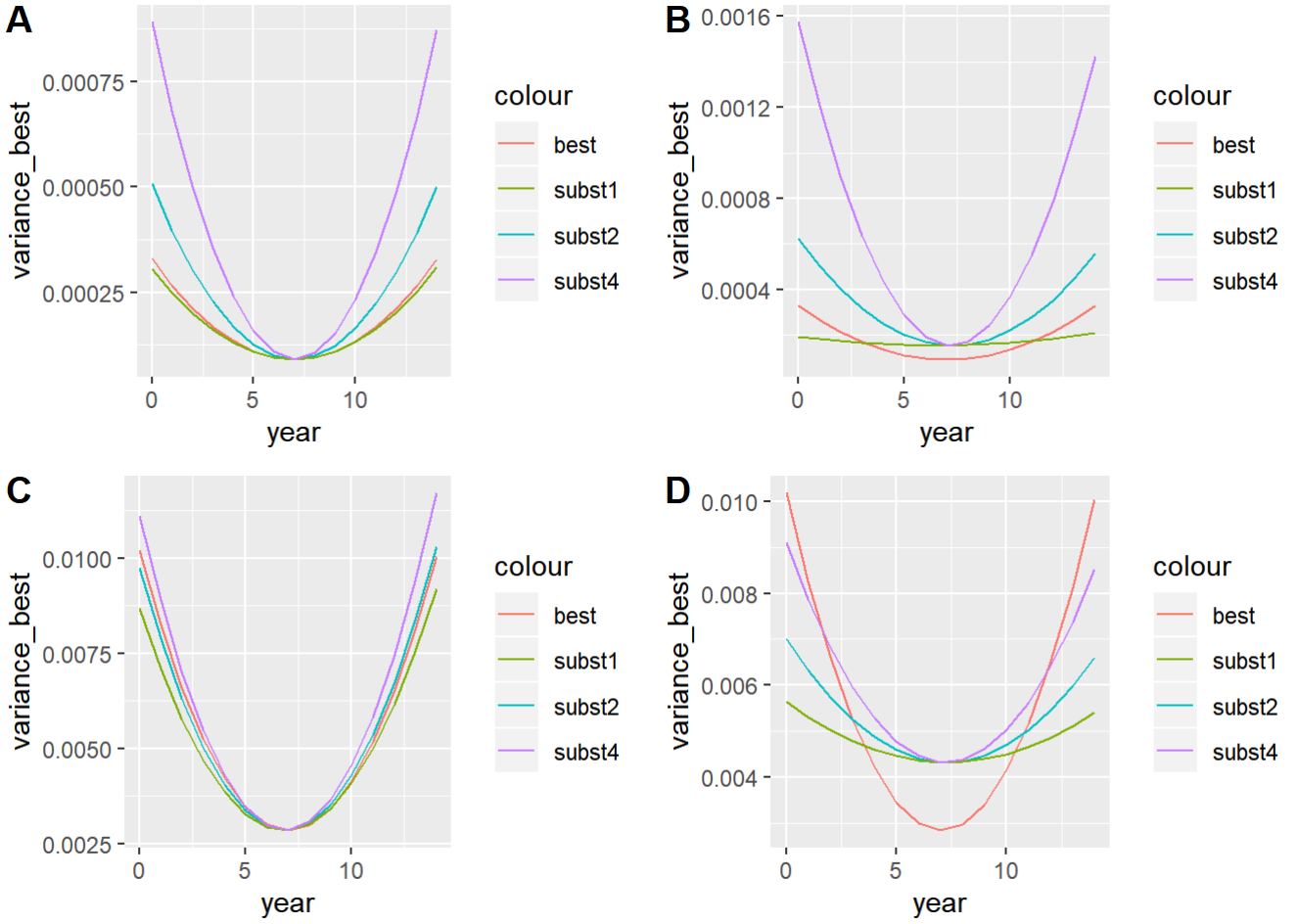
We will now use the same sets of parameter values that we used in our earlier section “Selection of censoring methods for further study”. Concretely: beta28year = -0.02 is held fixed, whilst a “low” and a “high” value for each of `cprop` and `sd28_153` are used. Concretely: (0.1, 0.1), (0.7, 0.1), (0.1, 0.5) and (0.7, 0.5) were used for (cprop, sd28_153) respectively.

We begin by showing graphs of the variance of predictions of cb28 annual means from our chosen censoring methods.

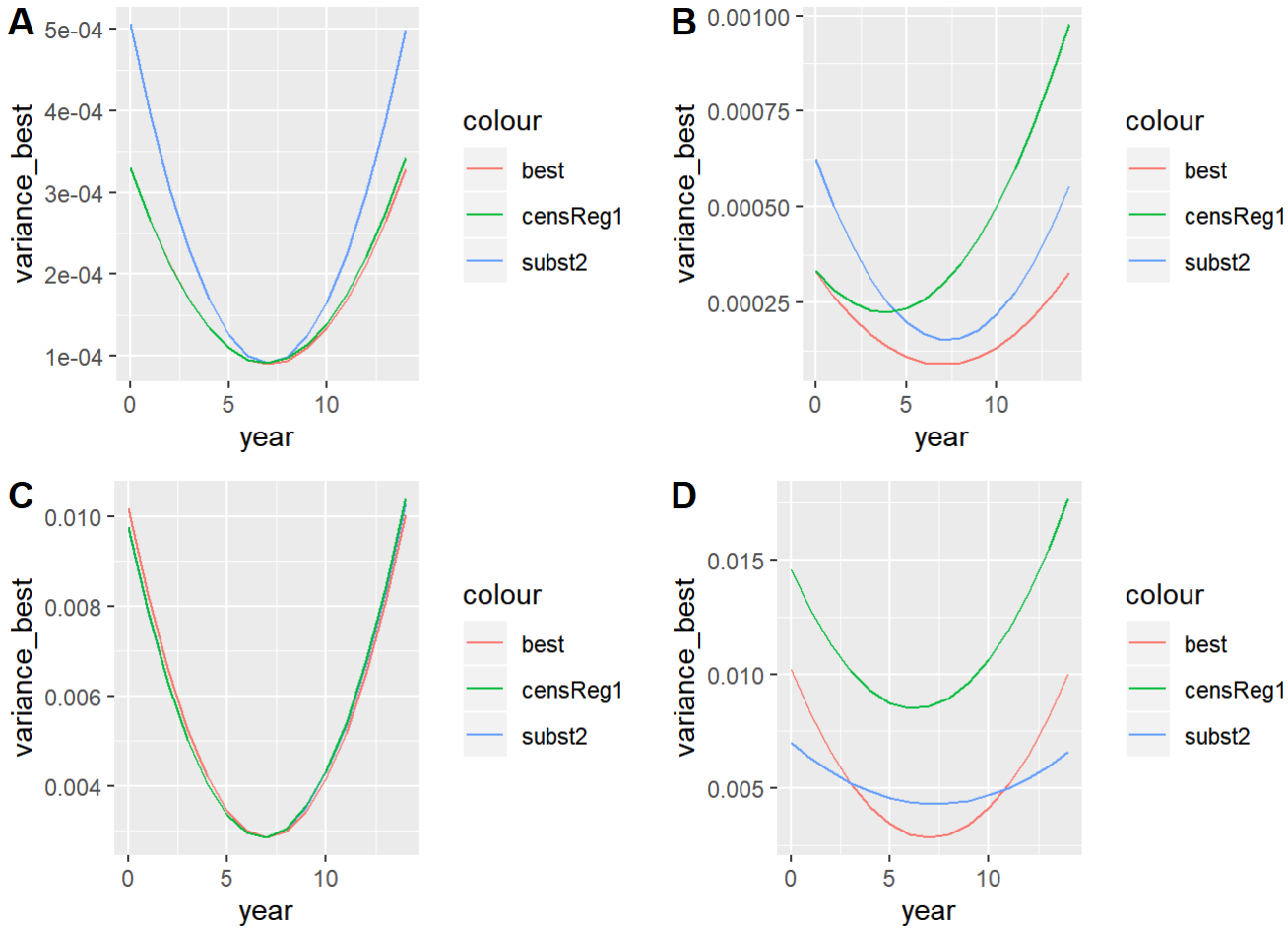
Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for (sd28_153, cprop) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.



Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for (sd28_153, cprop) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.

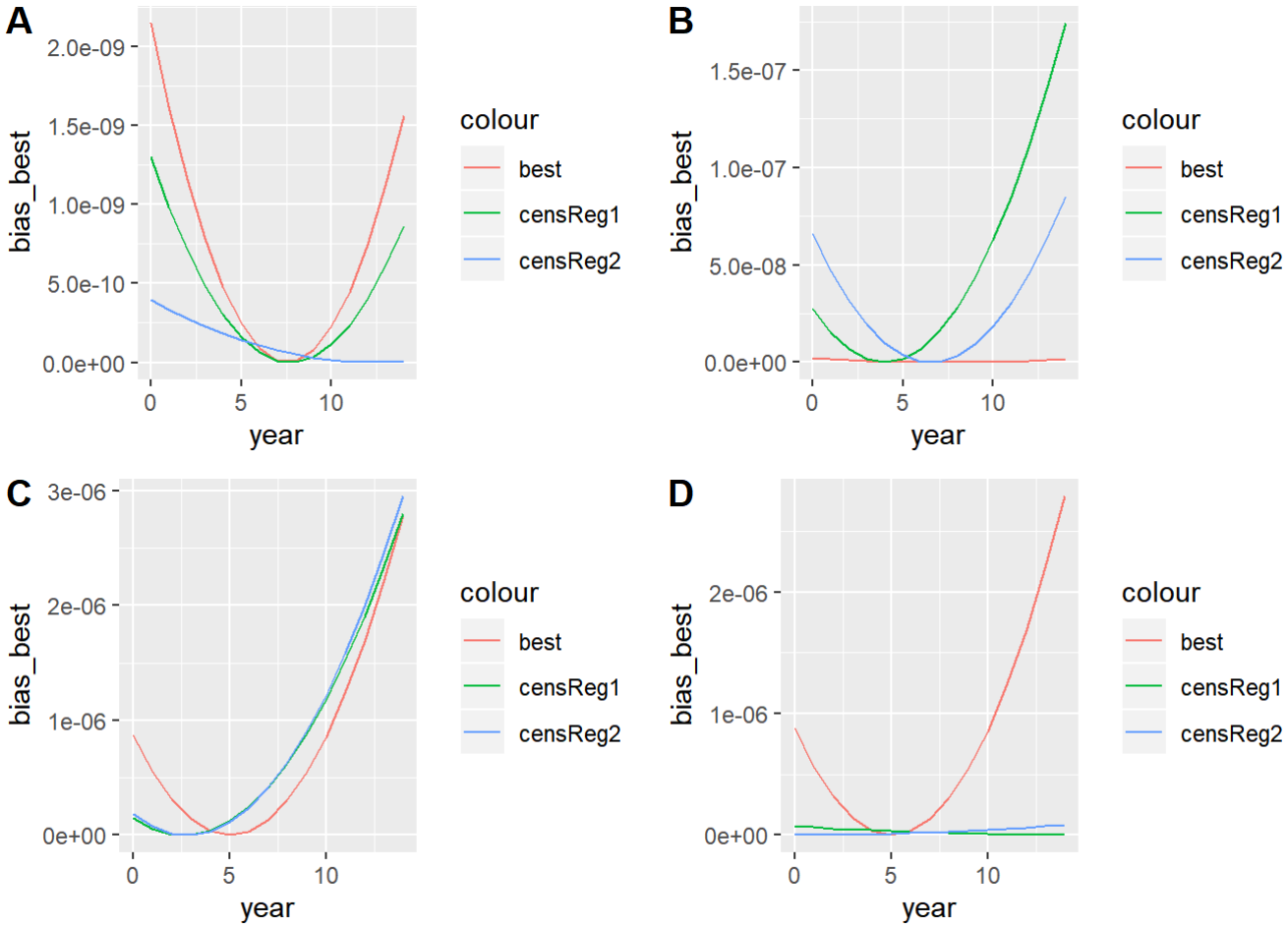


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

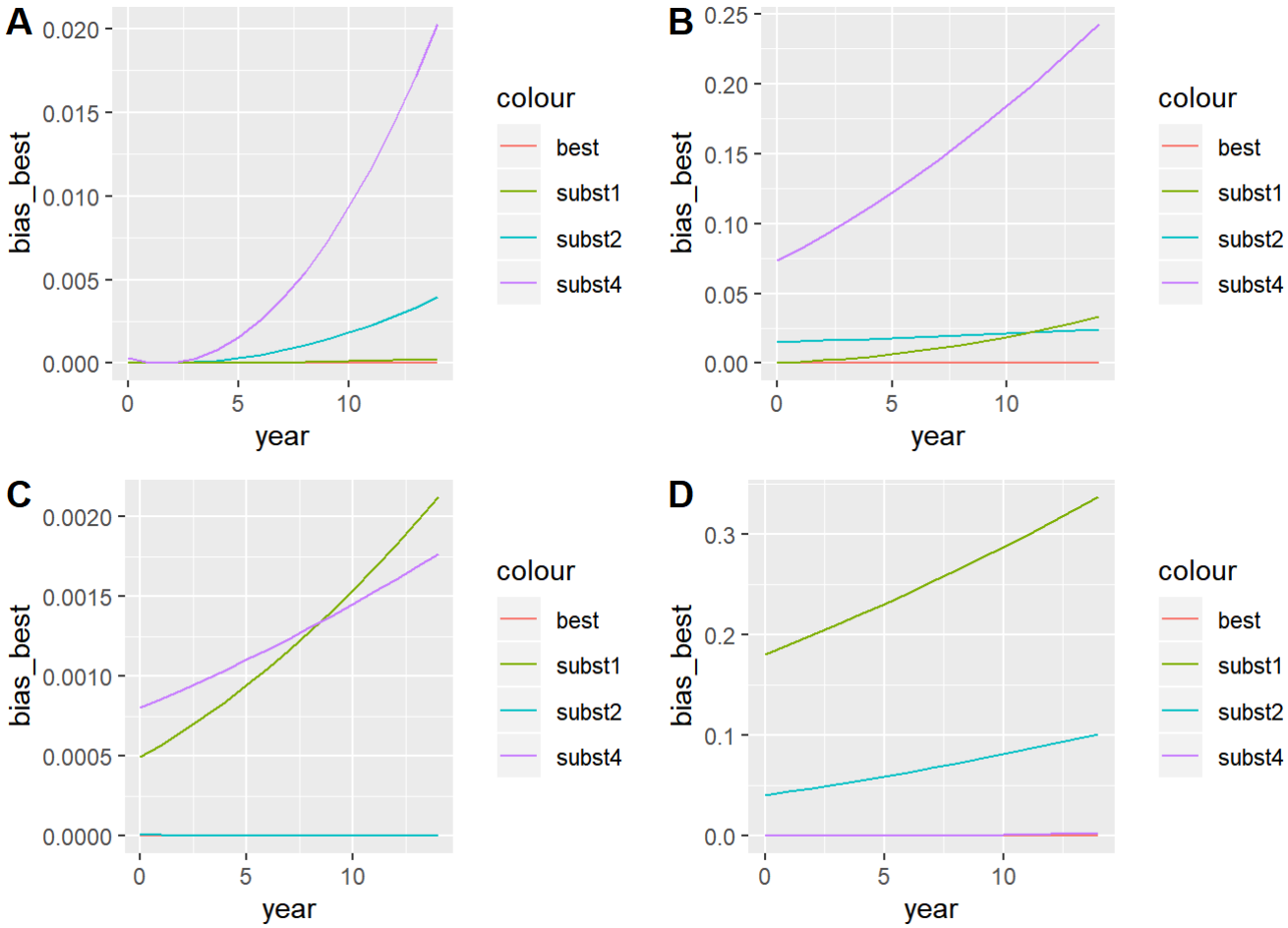


We will now show graphs of the bias of predictions of cb28 annual means from our chosen censoring methods.

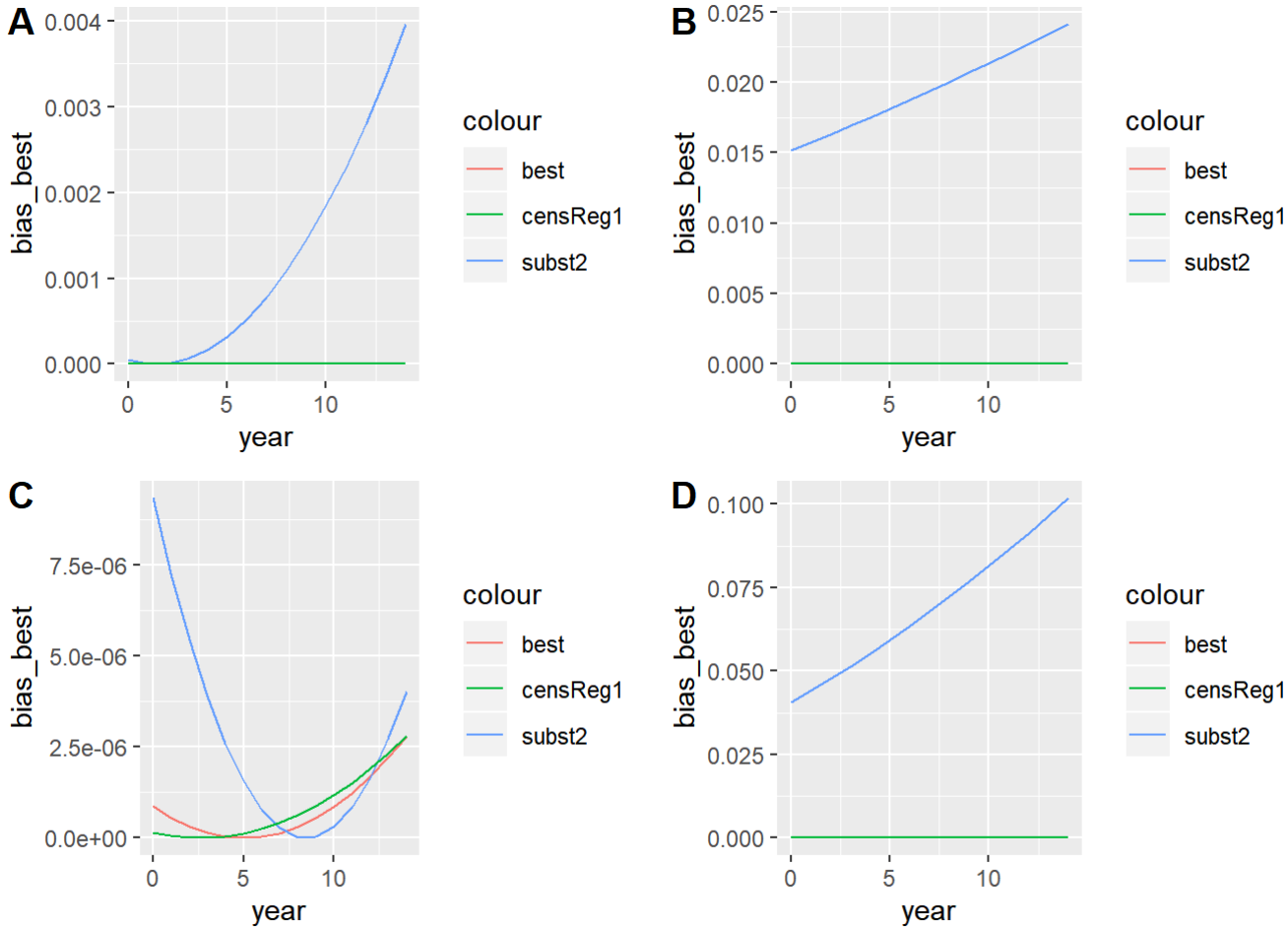
Our first set of four graphs show the bias of `censReg1` and `censReg2` methods relative to `best` method for (sd28_153, cprop) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.



Our second set of four graphs show the bias of `subst1`, `subst2` and `subst4` methods relative to `best` method for (sd28_153, cprop) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.

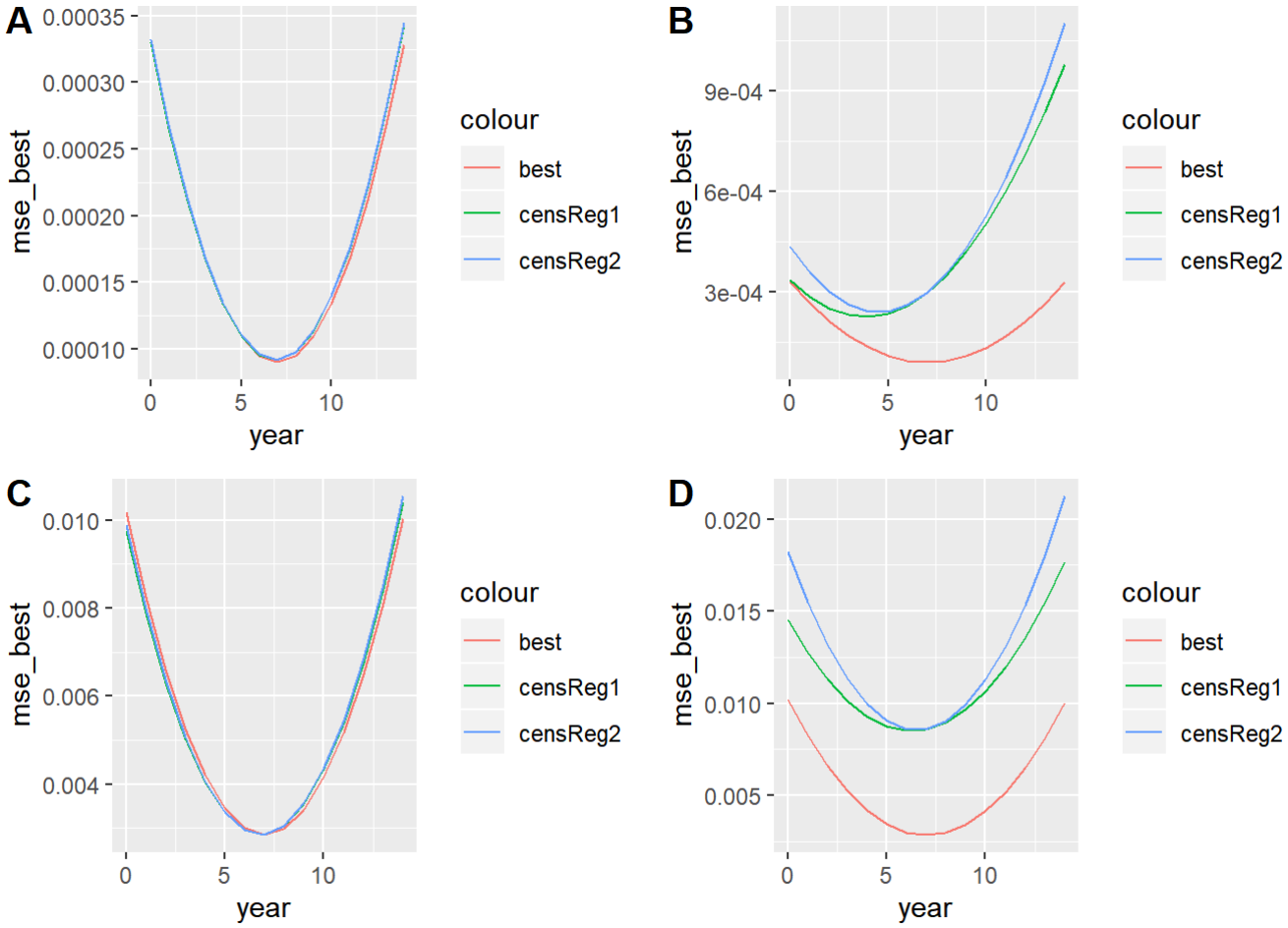


Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

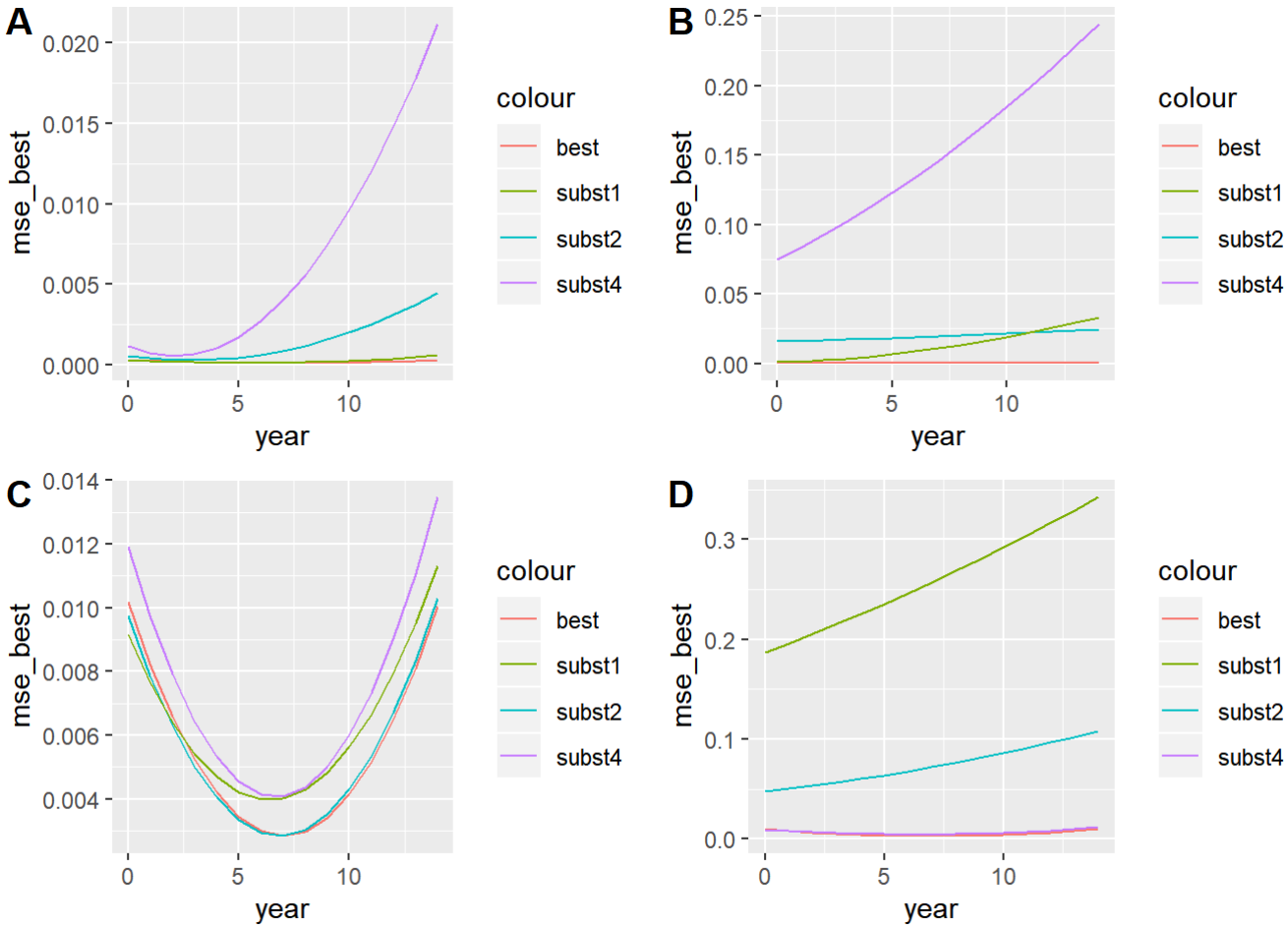


We will now show graphs of the MSE of predictions of cb28 annual means from our chosen censoring methods.

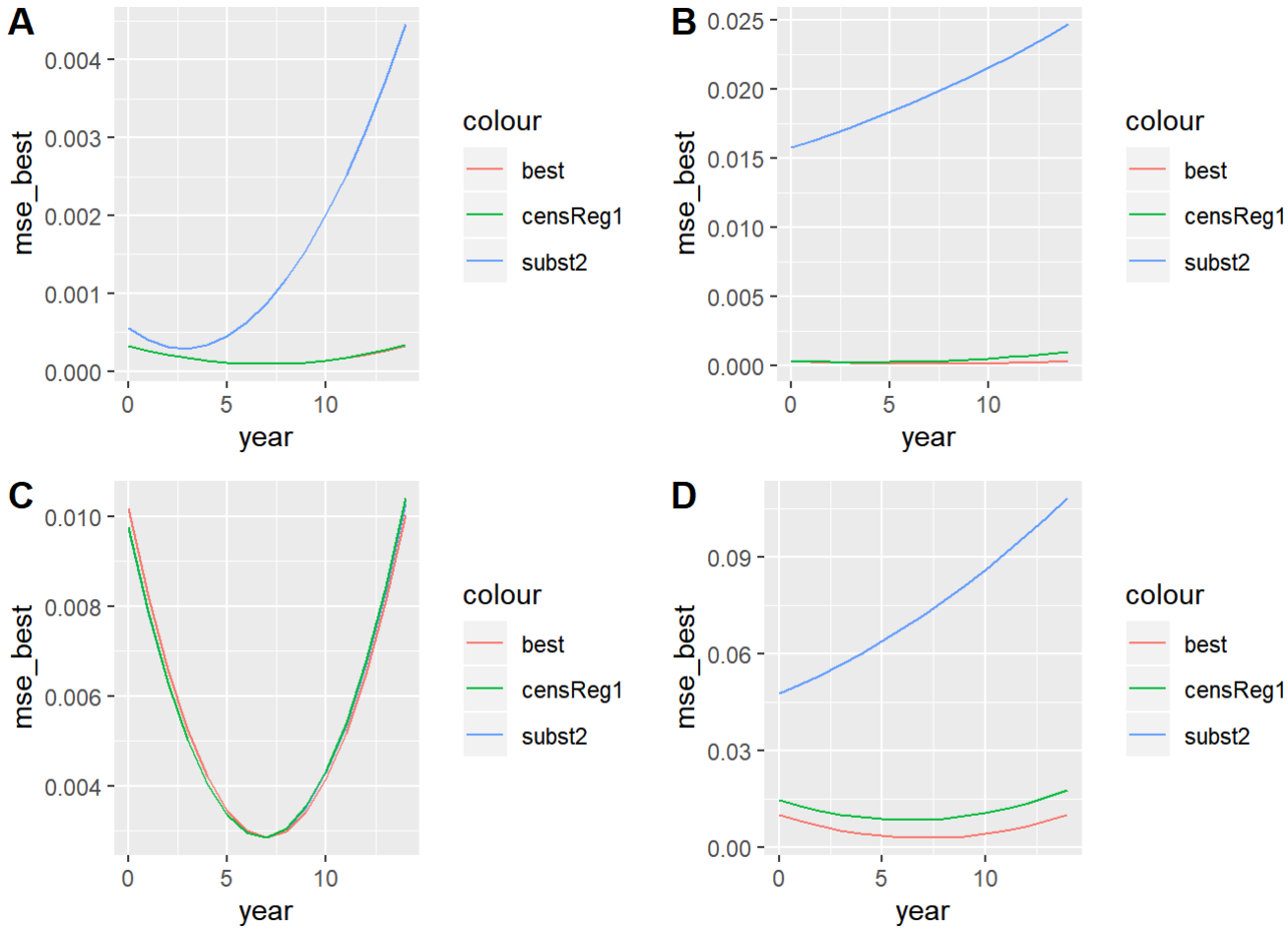
Our first set of four graphs show the MSE of `censReg1` and `censReg2` methods relative to `best` method for (sd28_153, cprop) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.



Our second set of four graphs show the MSE of `subst1`, `subst2` and `subst4` methods relative to `best` method for (sd28_153, cprop) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.



Our third set of four graphs simply displays the MSE from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.



Obsolete content

The remainder of this document stored content that I now view as obsolete.

Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for $(cprop, sd)$ equal to $(0.1, 0.1)$, $(0.7, 0.1)$, $(0.1, 0.5)$, $(0.7, 0.5)$, respectively. We see that the `censReg2` method consistently gives higher variance than `censReg1`. This is in accordance with our prior expectations because models with more predictor variables generally have higher variance than models with fewer predictors.

Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for $(cprop, sd)$ equal to $(0.1, 0.1)$, $(0.7, 0.1)$, $(0.1, 0.5)$, $(0.7, 0.5)$, respectively. In addition, the `omit` method can be used to obtain results for `cprop = 0.1` but not for `cprop = 0.7`, so the variance from those methods is shown on the two graphs for which `cprop = 0.1`.

Our third set of four graphs simply displays the previous two sets together on the same plot, which is displayed below.

Bias of predictions of cb28 annual means from all of our censoring methods

The `censReg1naive` method gives extremely biased predictions

The `censReg1naive` method gives extremely biased predictions for all 12 parameter value-pairs (four cprop levels, and three sd28_153 levels) .
Two graphs are presented below to illustrate this: they show the bias of predictions of cb28 annual means from the lowest levels (cprop = 0.1, sd28_153 = 0.1), and highest levels (cprop = 0.7, sd28_153 = 0.5), for each of these parameters, respectively.

This shows the necessity of conditioning on both the cb153 value and the condition cb28 < cb28_cprop by using a truncated normal distribution as presented in our previous chapter on mathematical theory. In contrast, censReg1naive conditions on the cb153 value only and uses a non-truncated normal distribution which results in significant bias because imputed cb28 values can be higher than cb28_cprop. Consequently we will not discuss censReg1naive any further: we expected this method to give biased estimates and it did.

Bias of predictions of cb28 annual means from all of our censoring methods except for `censReg1naive`

Every graph in this section shows the squared-bias of predictions of cb28 annual means from some of our censoring methods.

Our first set of four graphs show the bias of `censReg1` and `censReg2` methods relative to `best` method for $(cprop, sd)$ equal to $(0.1, 0.1)$, $(0.7, 0.1)$, $(0.1, 0.5)$, $(0.7, 0.5)$, respectively.

Our second set of four graphs show the bias of `subst1` , `subst2` and `subst4` methods relative to `best` method for $(cprop, sd)$ equal to $(0.1, 0.1)$, $(0.7, 0.1)$, $(0.1, 0.5)$, $(0.7, 0.5)$, respectively. In addition, the `omit` method can be used to obtain results for `cprop = 0.1` but not for `cprop = 0.7` , so the variance from those methods is shown on the two graphs for which `cprop = 0.1` .

We see that (ref: my Google Doc)

Our third set of four graphs simply displays the previous two sets together on the same plot, which is displayed below.

The remainder of this report is only VERY PRELIMINARY. BIG CHANGE IS GONNA COME.

Accuracy of yearly predictions for cprop = 0.1 and sd28_153 = 0.1.

Accuracy of yearly predictions for cprop = 0.3 and sd28_153 = 0.1.

Accuracy of yearly predictions for cprop = 0.5 and sd28_153 = 0.1.

Accuracy of yearly predictions for cprop = 0.7 and sd28_153 = 0.1.

Accuracy of yearly predictions for cprop = 0.1 and sd28_153 = 0.3.

Accuracy of yearly predictions for cprop = 0.3 and sd28_153 = 0.3.

Accuracy of yearly predictions for cprop = 0.5 and sd28_153 = 0.3.

Accuracy of yearly predictions for cprop = 0.7 and sd28_153 = 0.3.

Accuracy of yearly predictions for cprop = 0.1 and sd28_153 = 0.5.

Accuracy of yearly predictions for cprop = 0.3 and sd28_153 = 0.5.

Accuracy of yearly predictions for cprop = 0.5 and sd28_153 = 0.5.

Accuracy of yearly predictions for cprop = 0.7 and sd28_153 = 0.5.

Miscellaneous brain-storming-type notes

Cenreg did not work reliably (see v1 of this doc), so all censored regression will be done with censReg (followed by etruncnorm).

Sqrt(2) seems to be the best denominator. I could also try other numbers denominators and compare.

Found mse, squared-bias and variance with respect to estimation of beta for:

best, omit, subst2 (substitution with $\frac{LOD}{\sqrt{(2)}}$), subst1, subst4, censReg1, censReg2,

Simulation study reference: `Tekindal2017_EvaluatingLeft-CensoredDataBySimulationStudy.pdf`).

Could find the boundaries of the parameter space, especially:

`cprop` # censoring proportion

`true_beta28year` #beta for cb28 ~ year

Could try censReg with or without `year`.

Could try different substitutions:

LOD, LOD/sqrt(2), LOD/2, 0.

Math concepts to include

How many iterations to run?

Preliminary screening (fewer iterations), followed by in-depth assessment of the best performing models.

Confidence interval or prediction interval or neither?

Estimated prediction error (EPE).

Assessing model accuracy: need for test dataset and training dataset? Use RMSE instead of MSE so that this metric has same units as y?

Appendices

Appendix 1

The two graphs A, B below show the variation of MSE (red curve) and squared-bias-plus-variance (black curve) from `best_fit` and `omit_fit` respectively, over the simulated 15-year period. The famous result "Bias-variance decomposition" states

$$MSE = Bias^2 + Variance$$

so we expect the black and red curves to coincide (be superposed); happily they are :)

Appendix 2

We now generate the dataset `omit_yearly_mean` from the fixed parameters as follows:

12 values for the log-concentration of CB153 per year, for ten years, were generated and denoted as `cb153sim`.

From every such CB153 value, the corresponding value for CB28 was generated.

`median(cb28)` was used as the level of quantification `LOQ_p50`.

Observations with CB28 < LOQ_p50 were removed from the dataset.

Annual geometric means for CB28 and CB153 concentrations were generated.

The code chunk below generates `omit_yearly_mean` for 1000 iterations, fits a linear model `omit_fit` at each iteration, and computes the corresponding mse, squared-bias and variance.

Appendix 3