Simulation of imputation of censored values by linear regression

Marc Roddis

6/25/2020

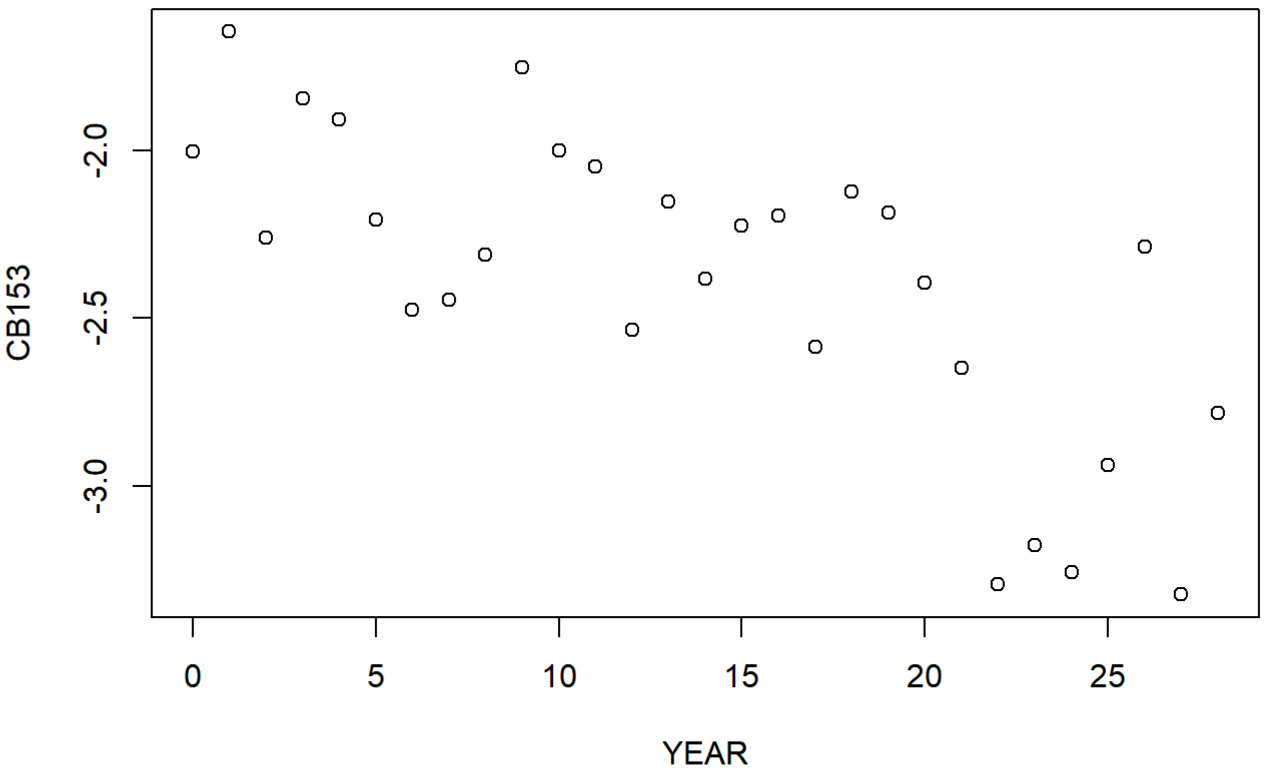
We will first create a test dataset `test_data1` from `pcb.csv` by omitting all missing values of `CB28` and `CB153` , removing all observations except those from herring species, removing all observations prior to 1989, re-indexing 1989 as “year zero”, removing all variables except `YEAR` , `CB28` and `CB153` .

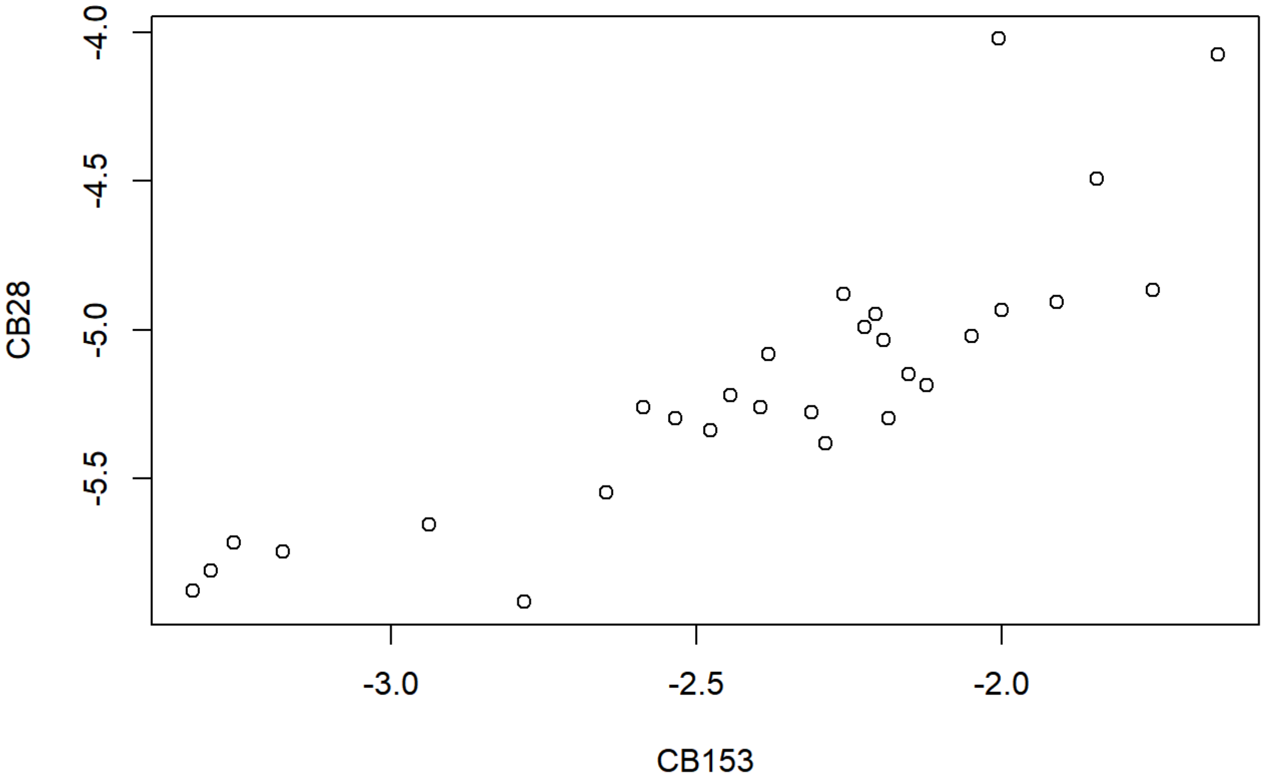
We now create `testdata_cen_omit` by omitting all censored observations and replacing concentrations with log-concentrations.

```
##  num [1:29] -4.02 -4.07 -4.88 -4.49 -4.91 ...
```

```
##  num [1:29] -2.01 -1.65 -2.26 -1.84 -1.91 ...
```

```
##  num [1:29] 0 1 2 3 4 5 6 7 8 9 ...
```


CB153
-2.0
-2.5
-3.0
0
5
10
15
20
25
YEAR


CB28
-4.0
-4.5
-5.0
-5.5
-3.0
-2.5
-2.0
CB153

```
##
## Call:
## lm(formula = CB28 ~ YEAR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52725 -0.18164  0.02054  0.17732  0.51372
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.536673   0.091845 -49.395  < 2e-16 ***
## YEAR        -0.045869   0.005631  -8.145  9.5e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2537 on 27 degrees of freedom
## Multiple R-squared:  0.7107, Adjusted R-squared:    0.7
## F-statistic: 66.34 on 1 and 27 DF,  p-value: 9.5e-09
```

```
##
## Call:
## lm(formula = CB153 ~ YEAR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58821 -0.21957  0.01893  0.23303  0.57364
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.845763   0.115547 -15.974 2.78e-15 ***
## YEAR        -0.039098   0.007085  -5.519 7.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3192 on 27 degrees of freedom
## Multiple R-squared:  0.5301, Adjusted R-squared:  0.5127
## F-statistic: 30.46 on 1 and 27 DF,  p-value: 7.584e-06
```

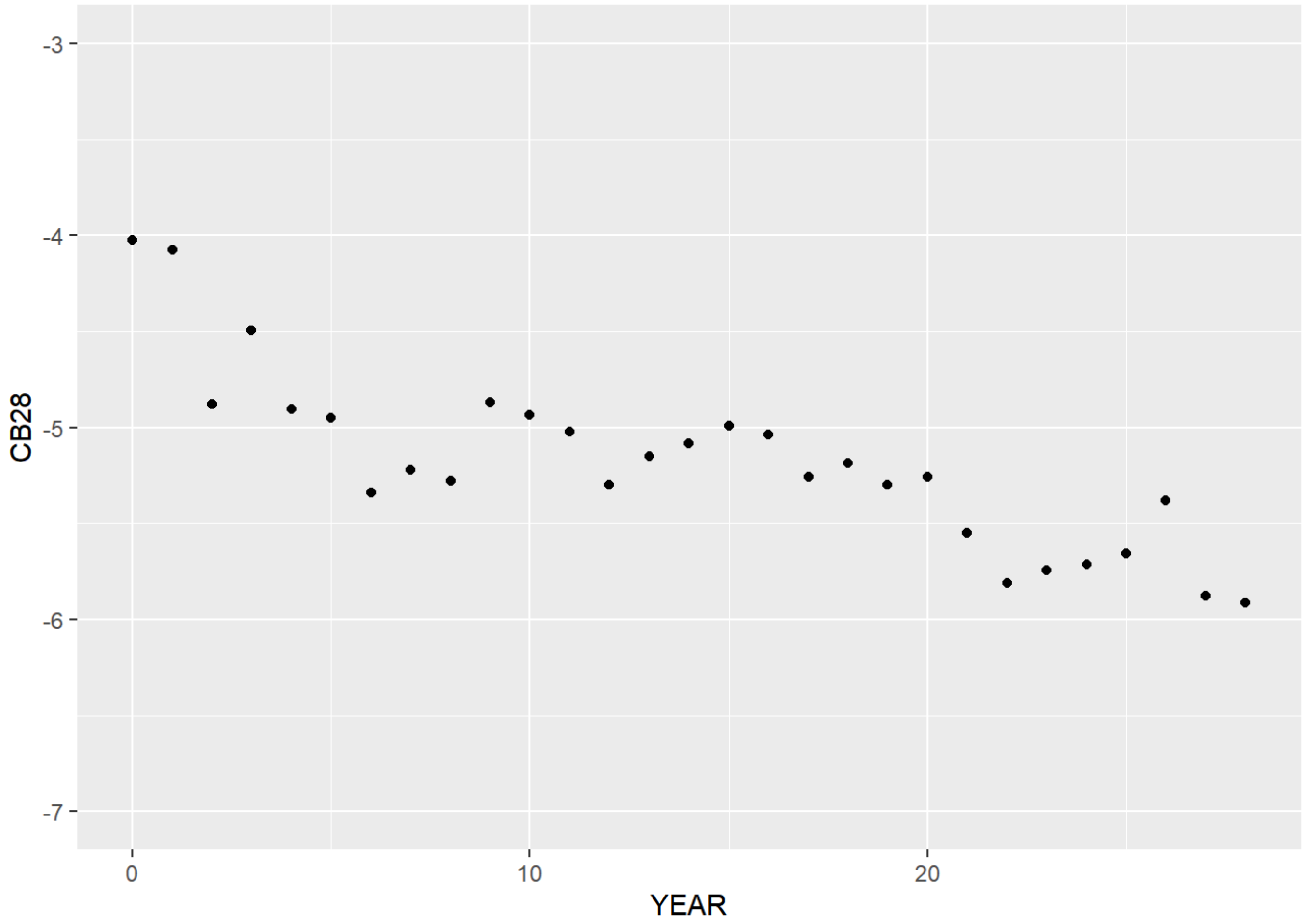
```
##
## Call:
## lm(formula = CB28 ~ CB153)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4008 -0.1487 -0.0077  0.1020  0.8222
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.1192     0.2505 -12.452 1.06e-12 ***
## CB153         0.8606     0.1029   8.365 5.63e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2489 on 27 degrees of freedom
## Multiple R-squared:  0.7216, Adjusted R-squared:  0.7113
## F-statistic: 69.98 on 1 and 27 DF,  p-value: 5.633e-09
```

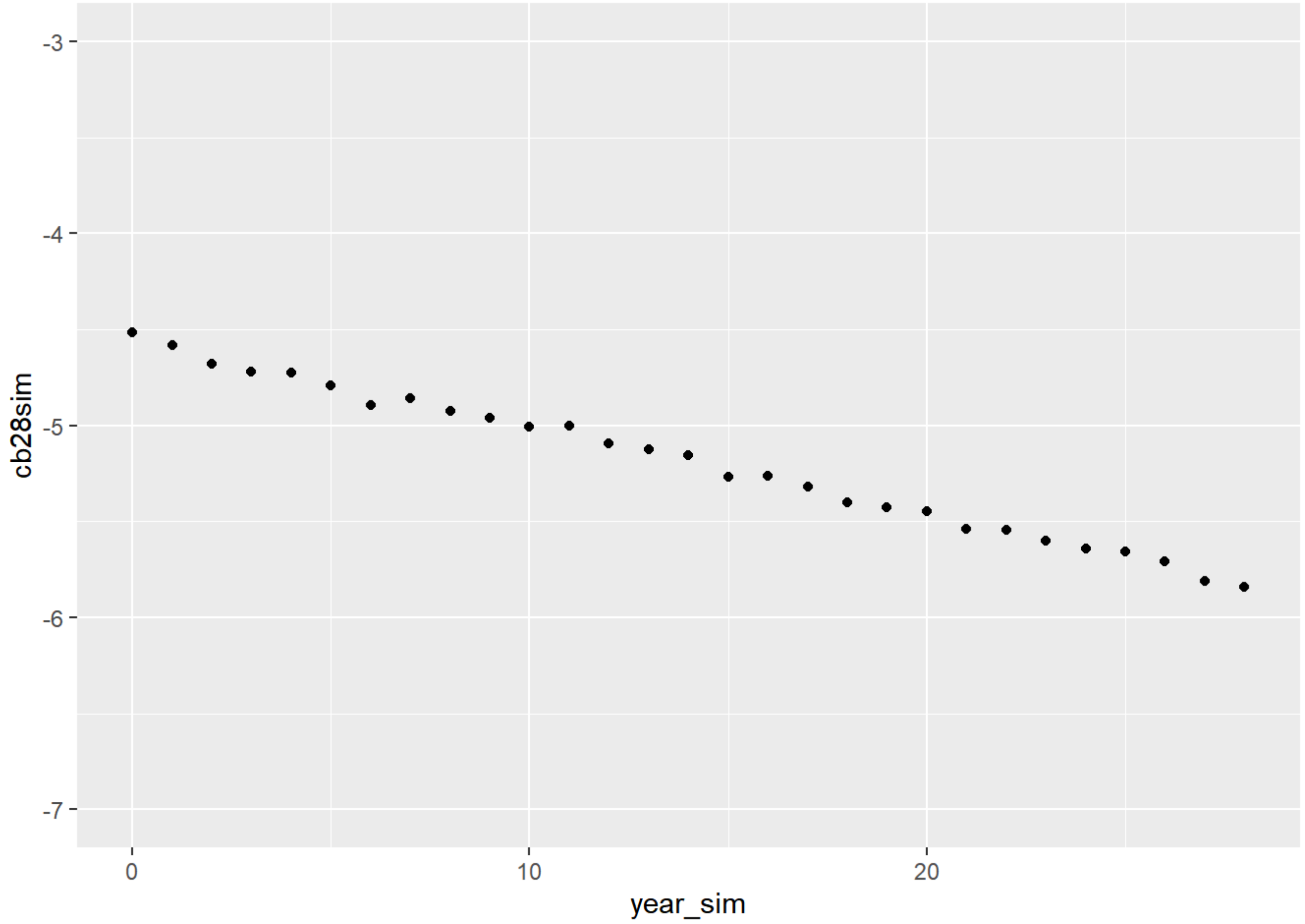
```
##
## Call:
## lm(formula = CB28 ~ CB153 + YEAR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32576 -0.09966 -0.03952  0.09403  0.59485
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.598867   0.232690 -15.466 1.26e-14 ***
## CB153        0.508086   0.119884   4.238 0.000251 ***
## YEAR        -0.026003   0.006438  -4.039 0.000422 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1988 on 26 degrees of freedom
## Multiple R-squared:  0.8289, Adjusted R-squared:  0.8158
## F-statistic: 62.99 on 2 and 26 DF,  p-value: 1.075e-10
```

```
## [1] 0.463263
```

```
## [1] 0.4572533
```

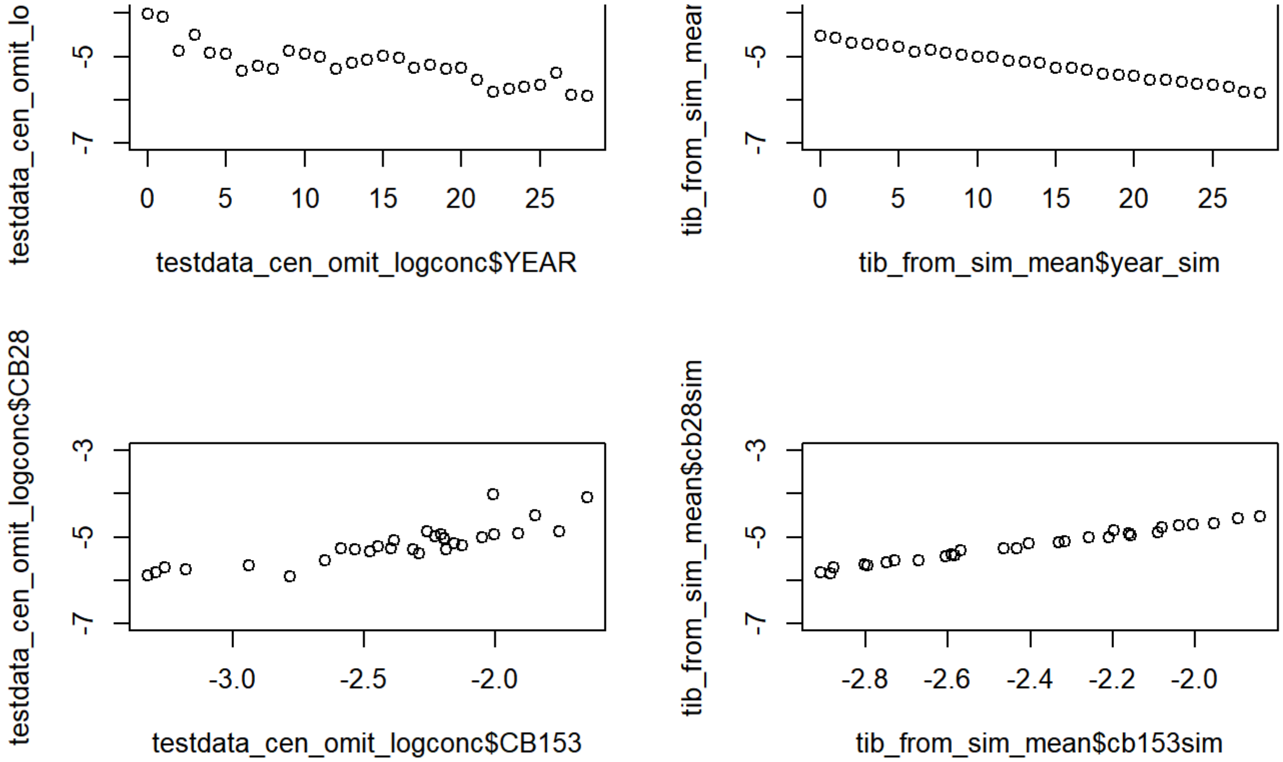
```
##       10%       30%       50%       90%
## -5.757512 -5.322811 -5.221356 -4.791596
```

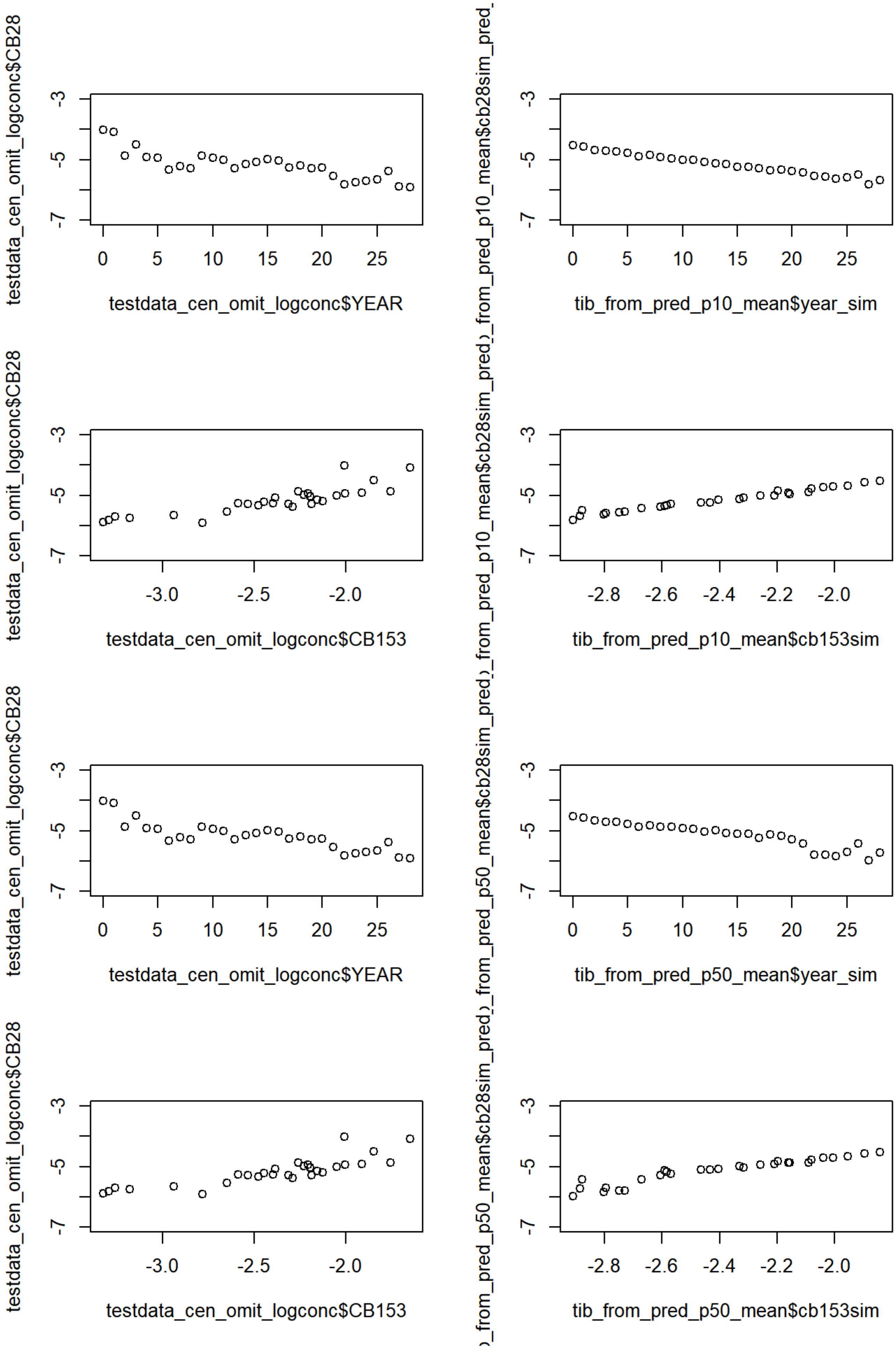

CB28
-3
-4
-5
-6
-7
0
10
20
YEAR

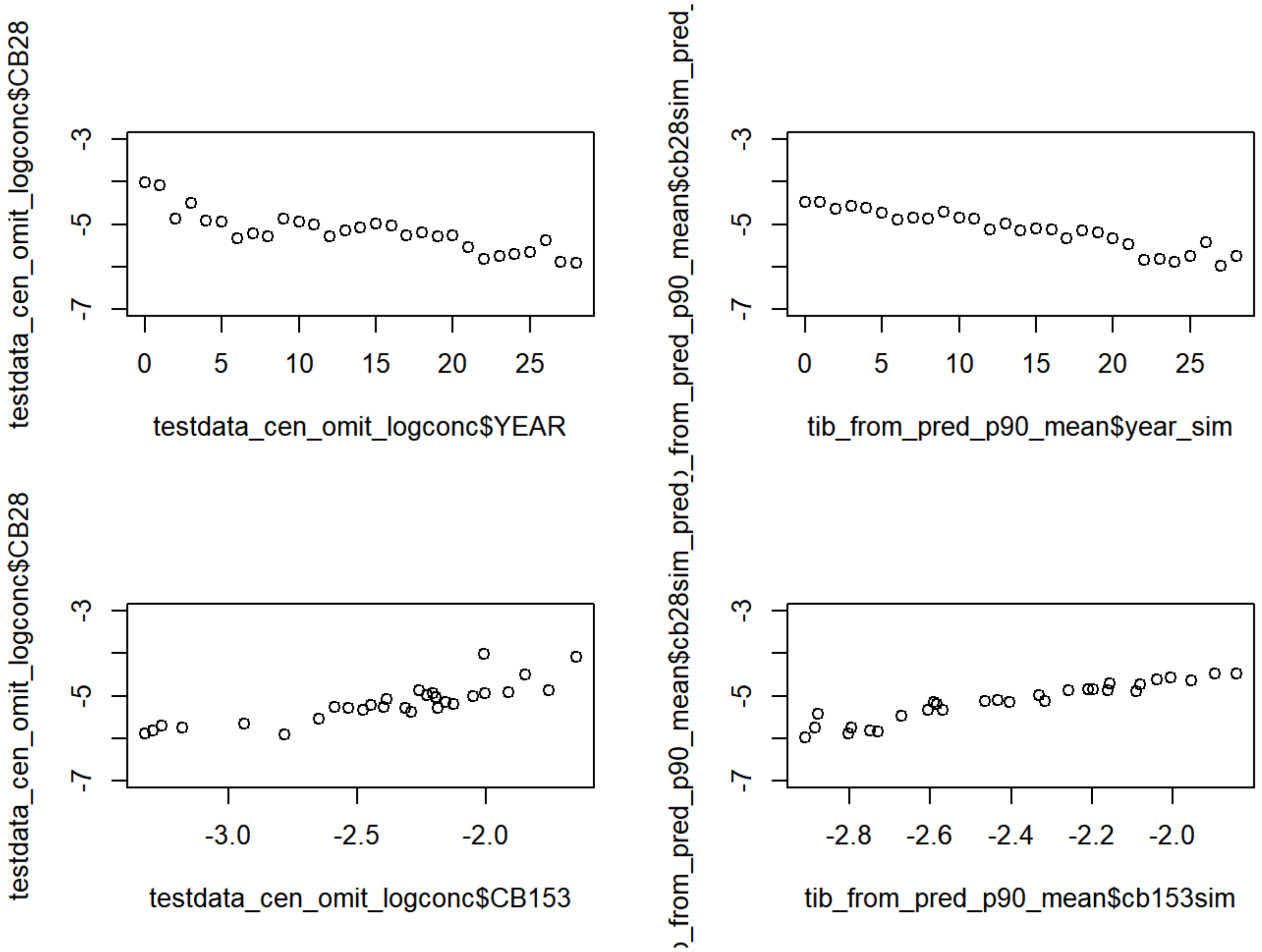

cb28sim
-3
-4
-5
-6
-7
0
10
20
year_sim

gconc$CB28
-3

n$cb28sim
-3


testdata_cen_omit_lo
-7
-5
0
5
10
15
20
25
testdata_cen_omit_logconc$YEAR
tib_from_sim_mear
-7
-5
0
5
10
15
20
25
tib_from_sim_mean$year_sim
testdata_cen_omit_logconc$CB28
-7
-5
-3
-3.0
-2.5
-2.0
testdata_cen_omit_logconc$CB153
tib_from_sim_mean$cb28sim
-7
-5
-3
-2.8
-2.6
-2.4
-2.2
-2.0
tib_from_sim_mean$cb153sim


testdata_cen_omit_logconc$CB28
-3
-5
-7
0
5
10
15
20
25
testdata_cen_omit_logconc$YEAR
tib_from_pred_p10_mean$cb28sim_pred_
tib_from_pred_p10_mean$year_sim
testdata_cen_omit_logconc$CB28
-3.0
-2.5
-2.0
testdata_cen_omit_logconc$CB153
tib_from_pred_p10_mean$cb28sim_pred
-2.8
-2.6
-2.4
-2.2
-2.0
tib_from_pred_p10_mean$cb153sim
testdata_cen_omit_logconc$CB28
testdata_cen_omit_logconc$YEAR
tib_from_pred_p50_mean$cb28sim_pred
tib_from_pred_p50_mean$year_sim
testdata_cen_omit_logconc$CB28
testdata_cen_omit_logconc$CB153
tib_from_pred_p50_mean$cb28sim_pred
tib_from_pred_p50_mean$cb153sim



Linear models for $(x, y) = (CB28, YEAR)$ and for $(x, y) = (CB28, CB153)$ respectively were each fitted to data with 10%, 50%, 90% censored observations substituted by imputed values, respectively. The adjusted R^2 values decreased as the proportion of censored observations increased, which reflects the fact that the non-censored values were simulated whereas the censored values were predicted from the linear model that was fitted to the observed data.

```
## 
## Call:
## lm(formula = tib_from_pred_p10_mean$cb28sim_pred_p10 ~ tib_from_pred_p10_mean$year_sim)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.117286 -0.017568  0.001234  0.019296  0.169494 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     -4.576558   0.018485 -247.58   <2e-16 ***
## tib_from_pred_p10_mean$year_sim -0.041561   0.001133  -36.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.05107 on 27 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9796 
## F-statistic:  1345 on 1 and 27 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = tib_from_pred_p10_mean$cb28sim_pred_p10 ~ tib_from_pred_p10_mean$cb153sim)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.104266 -0.031301 -0.003007  0.018464  0.189727
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.55067    0.07795  -32.72   <2e-16 ***
## tib_from_pred_p10_mean$cb153sim  1.08633    0.03219   33.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05538 on 27 degrees of freedom
## Multiple R-squared:  0.9768, Adjusted R-squared:  0.976
## F-statistic:  1139 on 1 and 27 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = tib_from_pred_p50_mean$cb28sim_pred_p50 ~ tib_from_pred_p50_mean$year_sim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28113 -0.06429  0.01591  0.07134  0.26902
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -4.497699   0.049769  -90.37  < 2e-16 ***
## tib_from_pred_p50_mean$year_sim -0.045907   0.003052  -15.04  1.2e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1375 on 27 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.8895
## F-statistic: 226.3 on 1 and 27 DF,  p-value: 1.199e-14
```

```
##
## Call:
## lm(formula = tib_from_pred_p50_mean$cb28sim_pred_p50 ~ tib_from_pred_p50_mean$cb153sim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25200 -0.04654 -0.01059  0.06682  0.29371
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.24819    0.18971  -11.85 3.30e-12 ***
## tib_from_pred_p50_mean$cb153sim  1.20484    0.07834   15.38 7.02e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1348 on 27 degrees of freedom
## Multiple R-squared:  0.8975, Adjusted R-squared:  0.8938
## F-statistic: 236.5 on 1 and 27 DF,  p-value: 7.015e-15
```

```
##
## Call:
## lm(formula = tib_from_pred_p90_mean$cb28sim_pred_p90 ~ tib_from_pred_p90_mean$year_sim)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.303987 -0.068756  0.004924  0.101642  0.303897
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -4.434944   0.052262  -84.86  < 2e-16 ***
## tib_from_pred_p90_mean$year_sim -0.050260   0.003204  -15.69 4.35e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1444 on 27 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8974
## F-statistic:   246 on 1 and 27 DF,  p-value: 4.349e-15
```

```
## 
## Call:
## lm(formula = tib_from_pred_p90_mean$cb28sim_pred_p90 ~ tib_from_pred_p90_mean$cb153sim)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.27048 -0.08091  0.01442  0.06854  0.33329 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     -1.96029    0.19160  -10.23 8.68e-11 ***
## tib_from_pred_p90_mean$cb153sim  1.32402    0.07912   16.73 8.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1361 on 27 degrees of freedom
## Multiple R-squared:  0.9121, Adjusted R-squared:  0.9088 
## F-statistic:   280 on 1 and 27 DF,  p-value: 8.853e-16
```