Analysis of Drug Consumption Data

Mark de Rooij

Load data

Load data and necessary libraries

LRRR analysis

Before analyzing the data, we throw away everything except **X** and **Y.** We also load the functions for analysis.

```
rm(list=ls()[! ls() %in% c("Y","X")])
source("~/surfdrive/LogitMDA/lrrr-diagnostics/lrrr.R")
source("~/surfdrive/LogitMDA/lrrr-diagnostics/diagnosis.R")
source("~/surfdrive/LogitMDA/lmap-package/new/R/procx.R")
source("~/surfdrive/LogitMDA/lrrr-diagnostics/plot.lrrr.R")
P = ncol(X)
R = ncol(Y)
N = nrow(X)
```

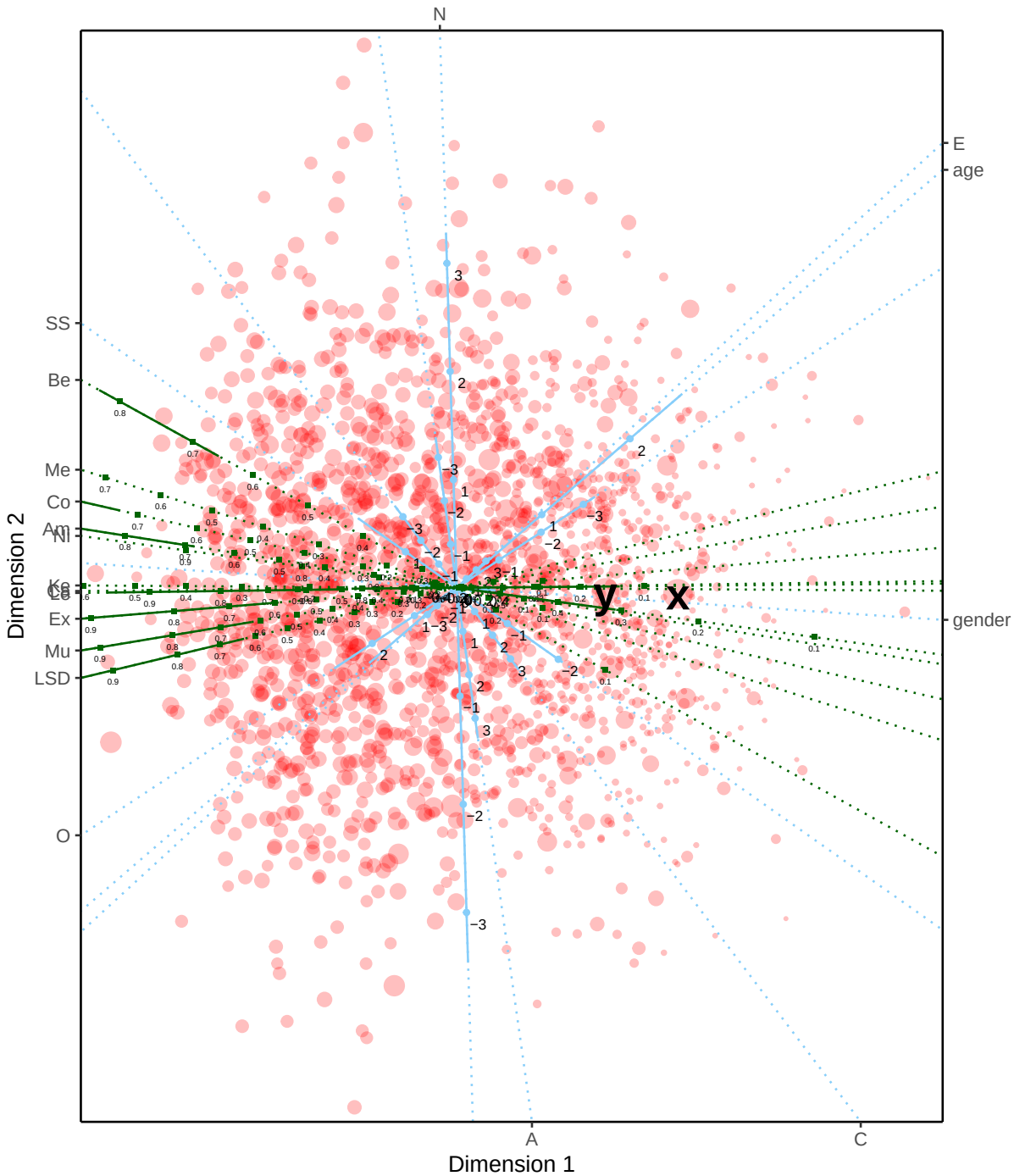
We do the same analysis on this data, as the final one reported in De Rooij and Groenen (2023). That is, a two-dimensional model where the predictor **impulsivity** is left out of the analysis.

```
lrrr.out = lpca(Y = Y, X = X, S = 2)
```

The biplot, where the markersize depends on the deviance residual.

```
plta = plot.lpcah(lrrr.out, res = TRUE, ocol= "red")

Udf = as.data.frame(lrrr.out$U)
colnames(Udf) = c("dim1", "dim2")
plta + geom_text(data = Udf[c(1090, 1300), ], aes(x = dim1, y = dim2, label = c("x", "y"))
```



Diagnosis

In this Section, we will run our diagnostics for the fitted model.

Fit

```r
fit = fit.lrrr(lrrr.out)
```

The deviance of the fitted model is 1.806646×10^4. The overall explained variance equals 0.21. Per response variable, the explained variance is

```r
round(fit$R2.variables, digits = 2)
```

```
  Am   Be   Ca   Co   Ex   Ke   Le  LSD   Me   Mu   Ni
0.19 0.12 0.33 0.14 0.22 0.14 0.28 0.29 0.14 0.26 0.14
```

The quality of representation for the response variables is

```r
round(fit$QoR, digits = 2)
```

```
  Am   Be   Ca   Co   Ex   Ke   Le  LSD   Me   Mu   Ni
1.00 0.96 0.97 0.90 0.96 0.95 1.00 0.98 0.94 0.99 0.98
```

Residuals

```r
res = residuals.lrrr(lrrr.out)
```

```r
round(rbind(apply(res$raw, 2, min),
apply(res$raw, 2, mean),
apply(res$raw, 2, max)), digits = 2)
```

```
        Am    Be    Ca    Co    Ex    Ke    Le   LSD    Me    Mu    Ni
[1,] -0.77 -0.75 -0.96 -0.74 -0.89 -0.58 -0.94 -0.92 -0.67 -0.90 -0.93
[2,]  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
[3,]  0.98  0.96  0.97  0.97  0.97  0.98  0.98  0.99  0.98  0.99  0.94
```

```r
round(rbind(apply(res$pearson, 2, min),
apply(res$pearson, 2, mean),
apply(res$pearson, 2, max)), digits = 2)
```

```
        Am    Be    Ca    Co    Ex    Ke    Le   LSD    Me    Mu    Ni
[1,] -1.83 -1.75 -5.00 -1.69 -2.88 -1.17 -3.83 -3.37 -1.42 -3.08 -3.54
[2,]  0.00  0.00  0.02 -0.01 -0.02 -0.01 -0.01 -0.02  0.00 -0.02  0.01
[3,]  6.54  5.09  5.97  5.70  5.75  6.45  7.06 10.78  6.56 11.23  4.04
```

We see some very large residuals for LSD and Mush. The participant with the highest value for the Pearson residual for LSD is 1090, the value of the residual in 10.780589. The participant with the highest value for the Pearson residual for Mush is 609, the value of the residual in 11.2303841.

```r
round(rbind(apply(res$deviance, 2, min),
apply(res$deviance, 2, mean),
apply(res$deviance, 2, max)), digits = 2)
```

```
        Am    Be    Ca    Co    Ex    Ke    Le   LSD    Me    Mu    Ni
[1,] -1.72 -1.68 -2.55 -1.64 -2.11 -1.31 -2.35 -2.24 -1.49 -2.17 -2.28
[2,] -0.13 -0.12  0.02 -0.14 -0.11 -0.17 -0.09 -0.12 -0.16 -0.12  0.04
[3,]  2.75  2.57  2.68  2.65  2.66  2.74  2.80  3.09  2.75  3.11  2.39
```

We see some very large deviance residuals for LSD and Mush. The participant with the highest value for the Pearson residual for LSD is 1090, the value of the residual in 3.0867659. The participant with the highest value for the deviance residual for Mush is 609, the value of the residual in 3.1129226.

We can also check which participants have high residual values. Therefore, we square the residuals and sum them over the 11 response variables. For the Pearson statistic the 10 observations with the larges residuals are

```r
pres_i = rowSums(res$pearson^2)
tail(sort(pres_i, index.return = T)$ix, n = 10)
```

```
 [1] 1180 1186  679 1017 1329  200 1272  609 1300 1090
```

for which the values are

```r
round(tail(sort(pres_i, index.return = T)$x, n = 10), digits = 2)
```

```
 [1]  81.78  86.94  92.14 100.57 130.78 132.63 141.45 159.19 217.78 381.76
```

We see that specifically observations 1300, 1090 have high residuals.

For the deviance residuals we can compute similar contributions. The 10 observations with the larges residuals are

```
dres_i = rowSums(res$deviance^2)
tail(sort(dres_i, index.return = T)$ix, n = 10)
```

```
 [1] 1870  679 1216 1518 1375  200 1329 1180 1300 1090
```

for which the values are

```
round(tail(sort(dres_i, index.return = T)$x, n = 10), digits = 2)
```

```
 [1] 32.51 33.34 34.19 36.80 37.70 38.53 42.68 42.93 56.76 66.65
```

We see that specifically observations 1300, 1090 have high residuals, the same two as for the Pearson residuals. Verifying their observed and fitted values

```
Y[c(1090, 1300), ]
```

```
     Am Be Ca Co Ex Ke Le LSD Me Mu Ni
[1,]  1  1  1  1  1  0  1   1  1  1  1
[2,]  1  1  0  1  1  1  1   1  1  1  0
```

```
round(lrrr.out$probabilities[c(1090, 1300), ], digits = 2)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
[1,] 0.03 0.09 0.07 0.04 0.03 0.01 0.02 0.01 0.03  0.01  0.23
[2,] 0.05 0.12 0.15 0.07 0.06 0.02 0.05 0.02 0.05  0.03  0.33
```

we see that all estimated probabilities are small, but yet some observations are 1.

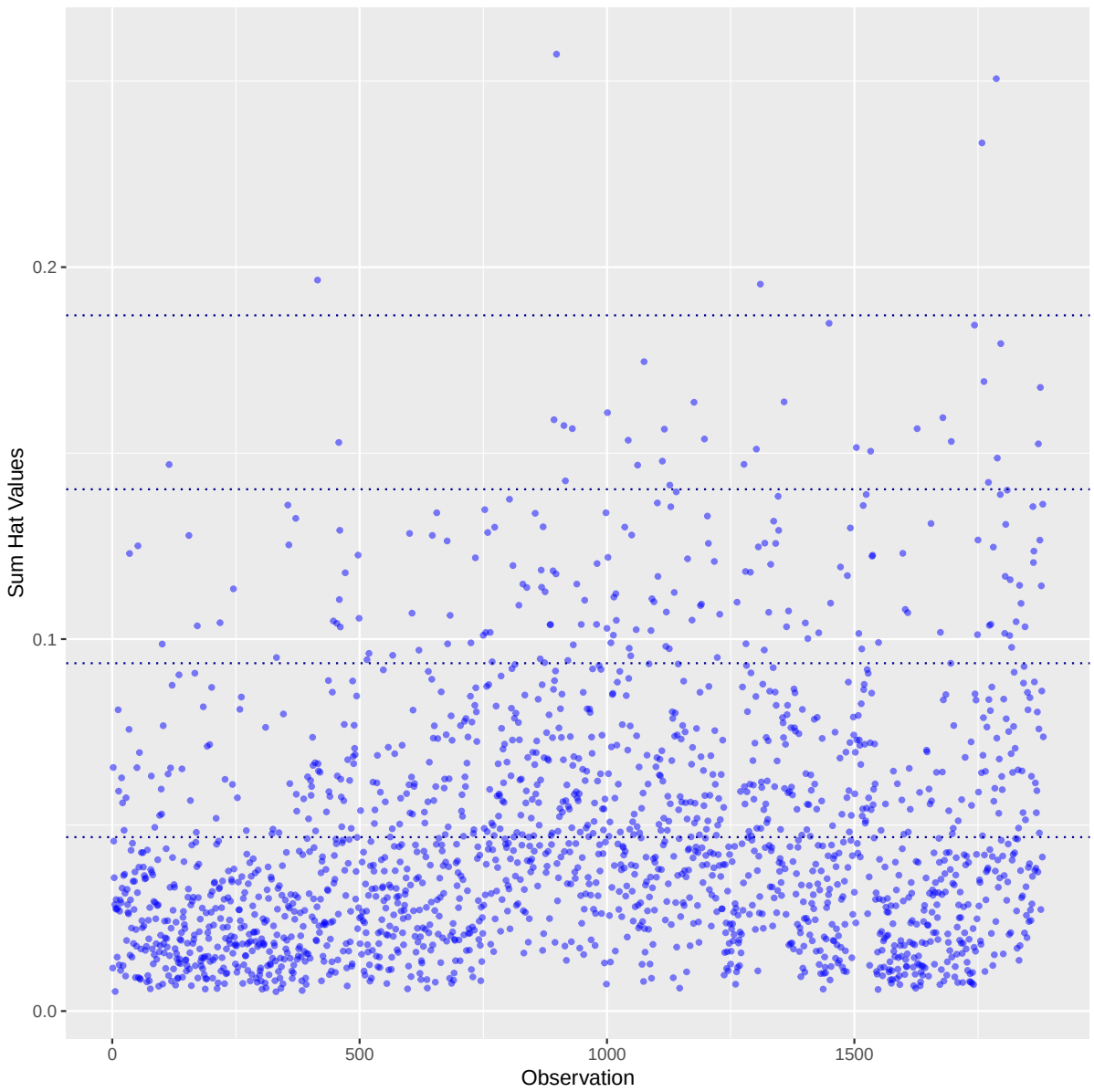
Hat

The average of the hat values per response variable equals P/N. As a hat statistic for a single observation we can take the sum over the hat values for the different responses

$$\sum_{r=1}^{R} h_i^{(r)},$$

the average over the observation of this statistic equals $(P \cdot R)/N$. A scatterplot can be made of case number against this hat value, enriched by a vertical line indicating the average. For logistic regression, usual thresholds are twice or three times this average, these two thresholds are also depicted.

```
hat = hat.lrrr(lrrr.out)
hat$sum = rowSums(hat[, -1])
ggplot(hat, aes(x = idx, y = sum)) +
  geom_point(size = 1, col = "blue", alpha = 0.5) +
  geom_hline(yintercept = c(P*R/N, 2*P*R/N, 3*P*R/N, 4*P*R/N), col = "darkblue", lty = "do
  labs(x = "Observation", y = "Sum Hat Values")
```



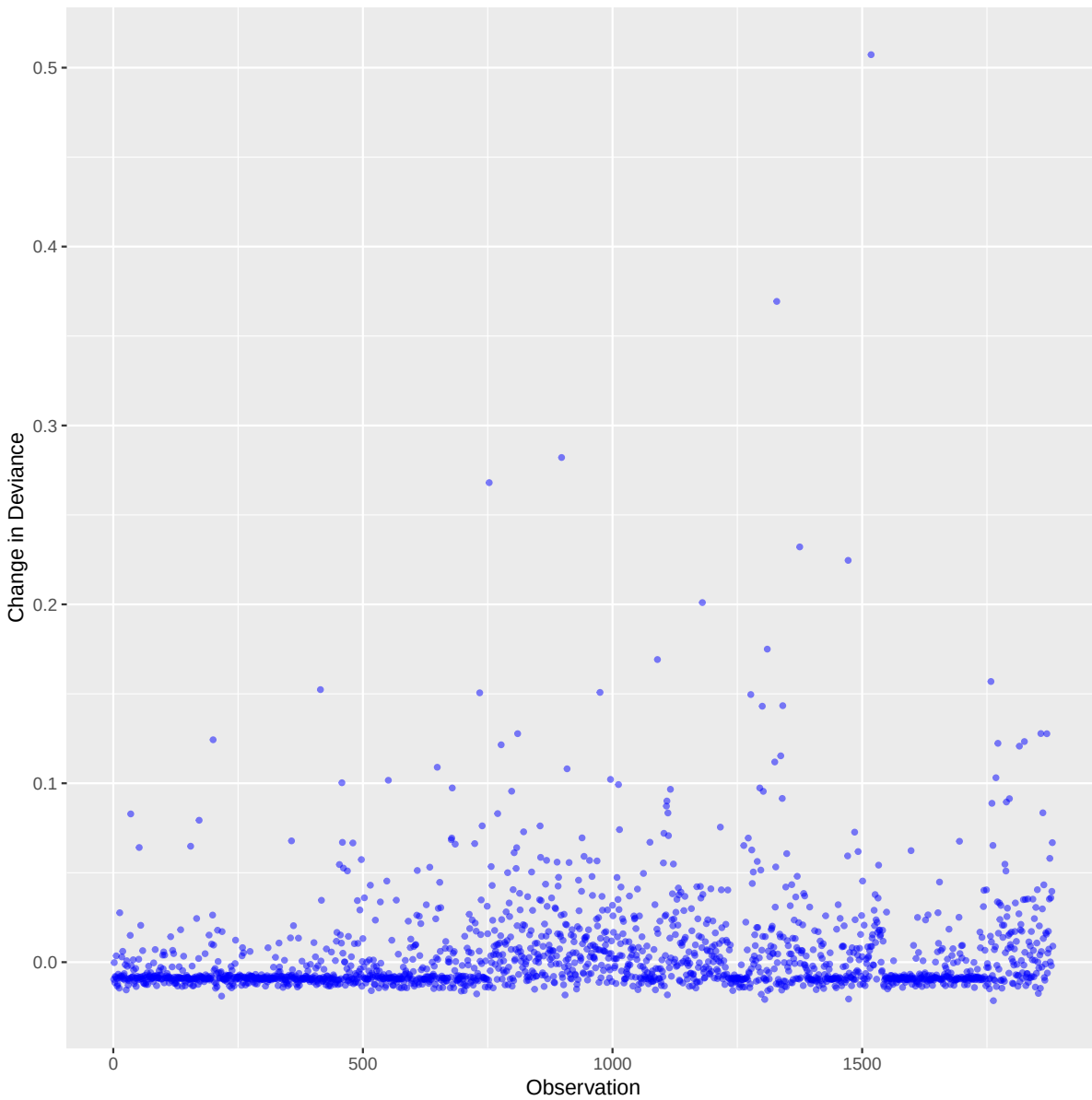
There are several observations exceeding the largest threshold. Remember that influence is leverage times discrepancy, so that, if these points show also discrepancy, they might be influential. The observation with the largest hat value is number 898.

Cases that exceed the *four times average*-threshold are 415, 898, 1310, 1758, 1787.

Influential observations

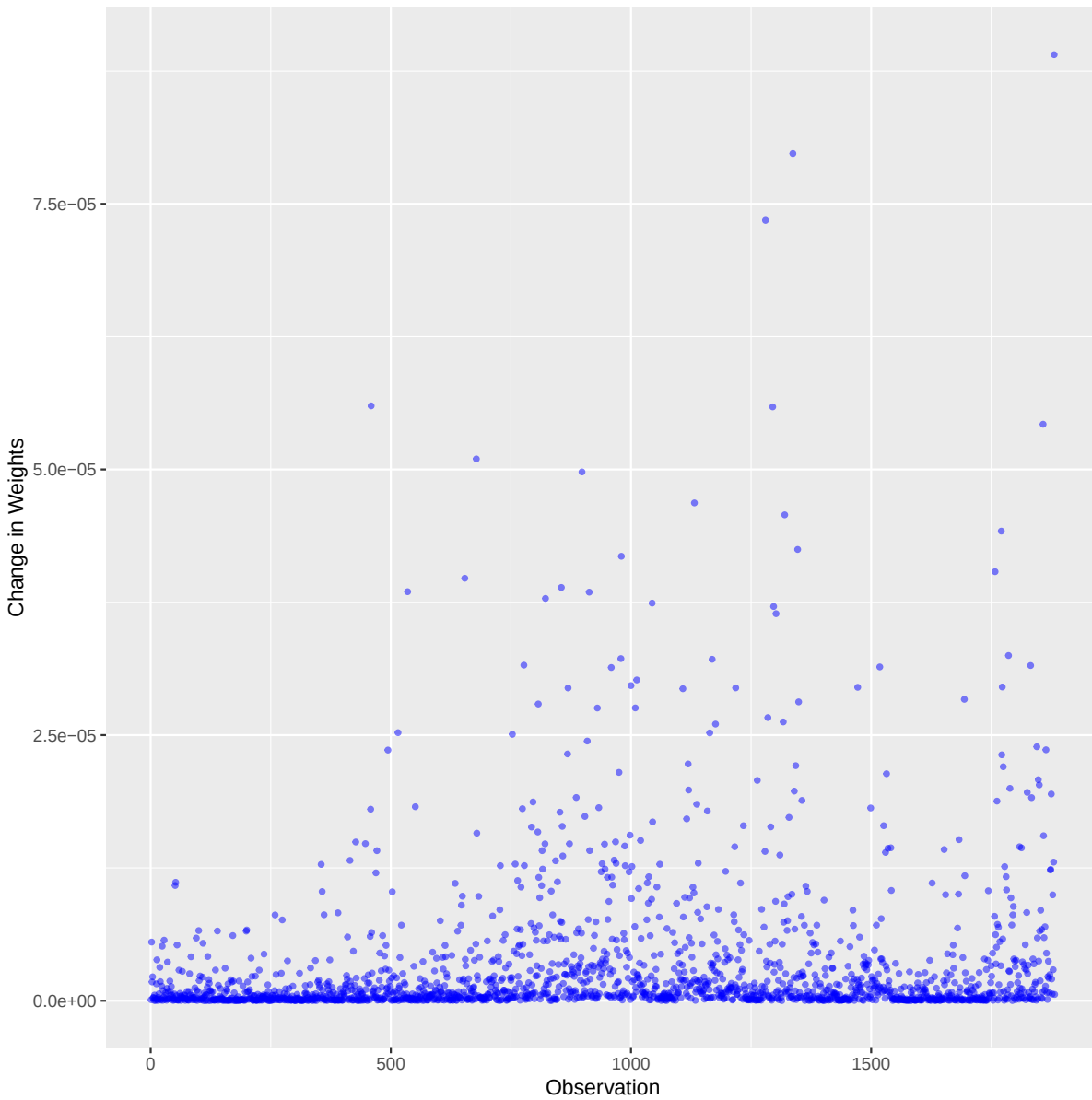
```r
deletion = influence.lrrr(lrrr.out)
```

```r
ggplot(deletion, aes(x = idx, y = deviance)) +
  geom_point(size = 1, col = "blue", alpha = 0.5) +
  labs(x = "Observation",
       y = "Change in Deviance")
```

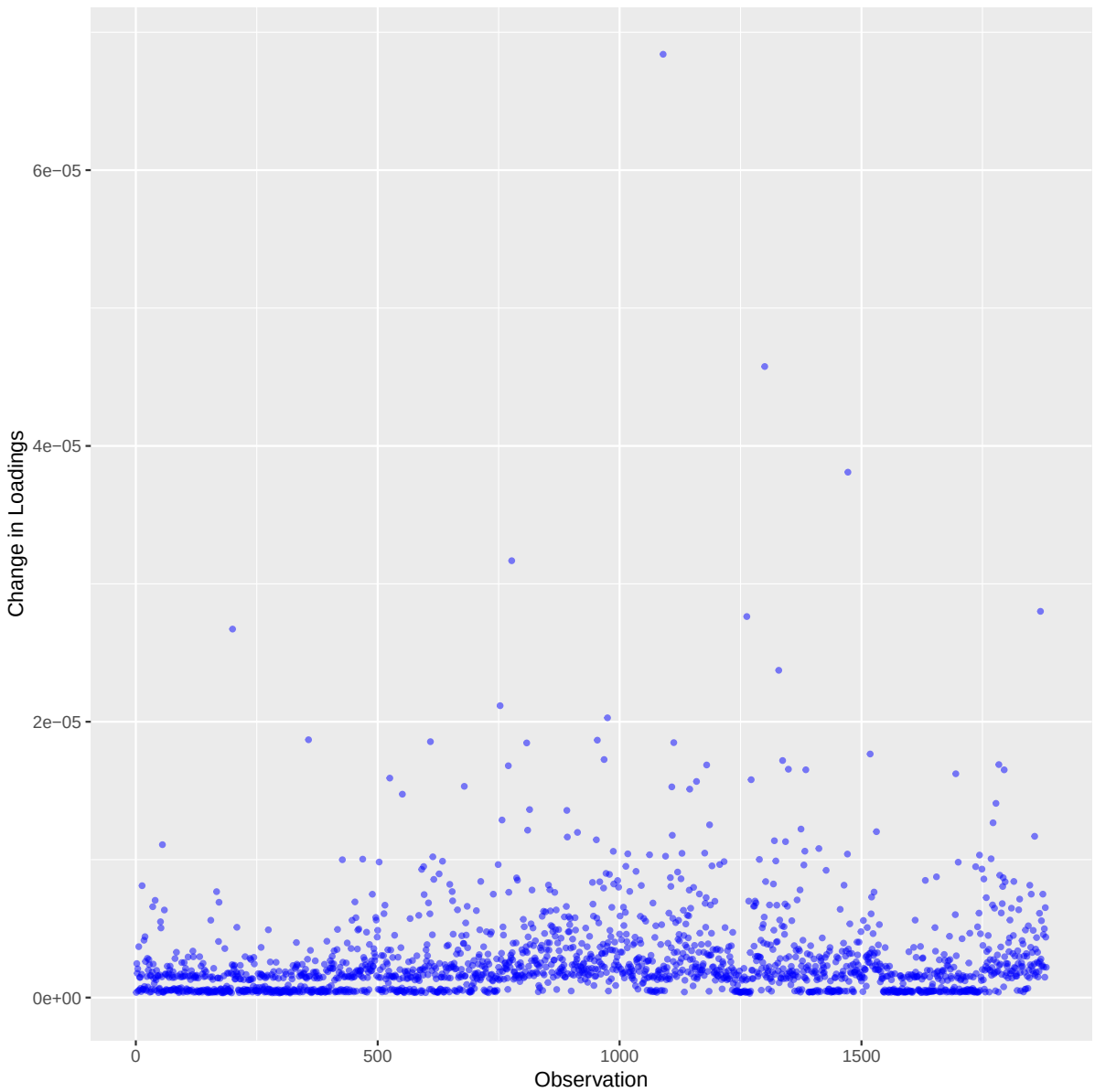


There is one observation with a substantial change in deviance, this is observation 1518. Observations with a change in deviance larger than 0.3 are 1329, 1518.

```
ggplot(deletion, aes(x = idx, y = B)) +
  geom_point(size = 1, col = "blue", alpha = 0.5) +
  labs(x = "Observation",
       y = "Change in Weights")
```




```r
ggplot(deletion, aes(x = idx, y = V)) +
  geom_point(size = 1, col = "blue", alpha = 0.5) +
  labs(x = "Observation",
       y = "Change in Loadings")
```



From these plots, no influential observations can be discovered. All values of the change parameters are small. The observation with largest change in regression weights is 1881, observations with a change in weights larger than 7.5×10^{-5} are 1337, 1881.

Similarly, observations with a change in loadings larger than 4×10^{-5} are 1090, 1300, with the largest loading 1090.

Inspecting cases 1300 and 1090

Observations 1300 and 1090 might be influential, they had high residual values and δ_V. Let us check, how the configuration changes as we delete either of them or both. With all data the estimated regression weights are

```r
round(lrrr.out$B, digits = 3)
```

```
       [,1]   [,2]
[1,]  0.533  0.458
[2,]  0.530 -0.035
[3,] -0.022  0.743
[4,]  0.067  0.061
[5,] -0.289 -0.191
[6,]  0.042 -0.298
[7,]  0.123 -0.162
[8,] -0.337  0.237
```

and the estimated loadings are

```r
round(lrrr.out$V, digits = 3)
```

```
        [,1]   [,2]
 [1,] -1.275  0.201
 [2,] -0.822  0.454
 [3,] -1.844 -0.016
 [4,] -1.038  0.238
 [5,] -1.406 -0.116
 [6,] -1.153  0.007
 [7,] -1.660 -0.023
 [8,] -1.734 -0.416
 [9,] -1.051  0.330
[10,] -1.582 -0.264
[11,] -0.997  0.138
```

Now, we fit the model excluding case 1300 and look at the estimated regression weights and loadings

```r
lrrr.out.1300 = lpca(Y = Y[-1300, ], X = X[-1300, ], S = 2)
round(lrrr.out$B - lrrr.out.1300$B, digits = 3)
```

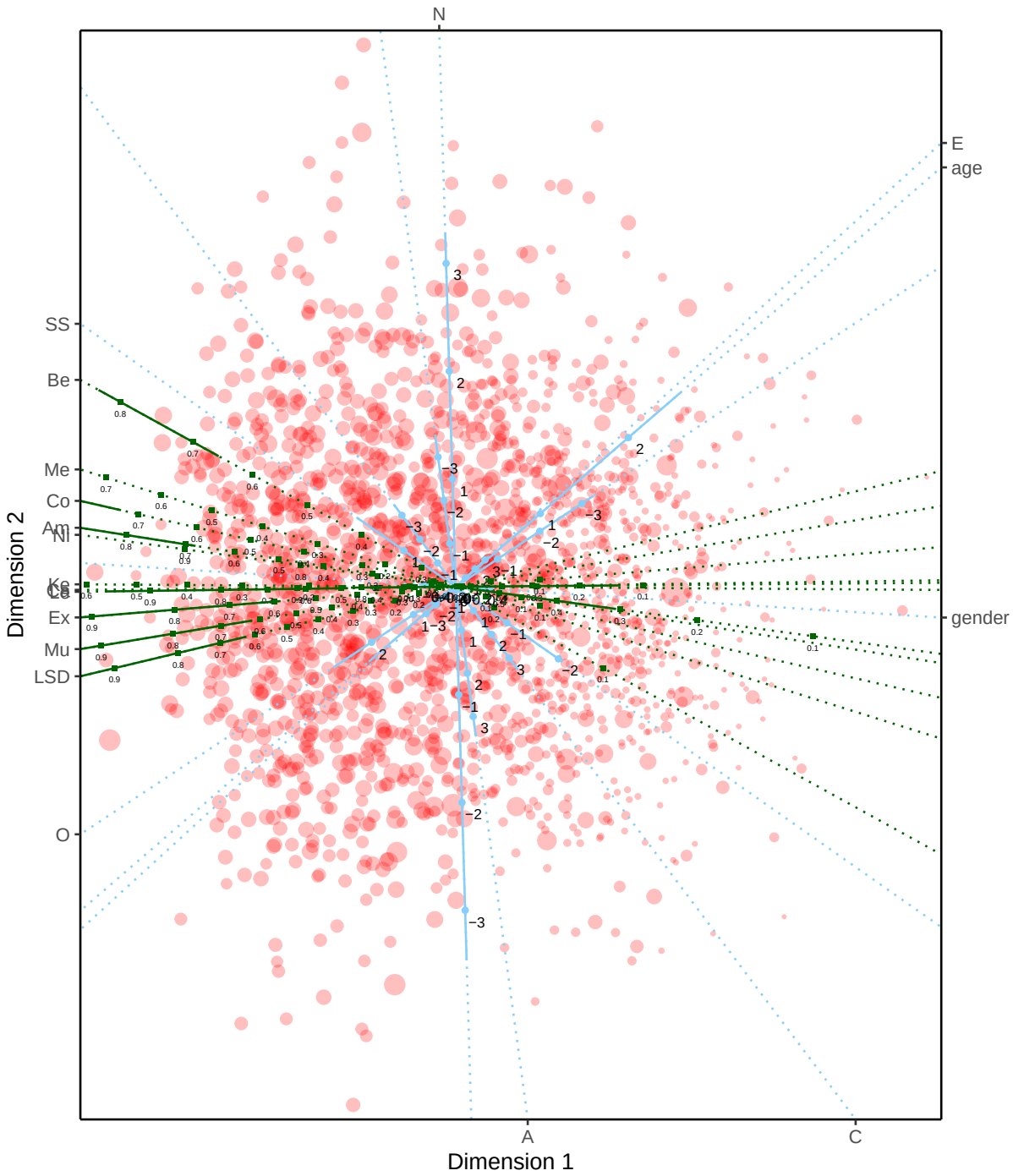
```
       [,1]   [,2]
[1,]  0.001 -0.001
[2,]  0.007 -0.002
[3,] -0.001  0.001
[4,]  0.000  0.000
[5,] -0.001  0.000
[6,]  0.002 -0.001
[7,]  0.000  0.001
[8,]  0.004 -0.002
```

```r
round(lrrr.out$V - lrrr.out.1300$V, digits = 3)
```

```
        [,1]   [,2]
 [1,]  0.009 -0.001
 [2,]  0.004 -0.001
 [3,] -0.001 -0.001
 [4,]  0.008 -0.001
 [5,]  0.009  0.000
 [6,]  0.016  0.000
 [7,]  0.010  0.000
 [8,]  0.018  0.002
 [9,]  0.009 -0.001
[10,]  0.013  0.001
[11,] -0.001  0.000
```

The triplot for this analysis is as follows

```r
plot.lpcah(lrrr.out.1300, res = TRUE, ocol= "red")
```



Now, we fit the model excluding case 1090 and look at the estimated regression weights and loadings

```r
lrrr.out.1090 = lpca(Y = Y[-1090, ], X = X[-1090, ], S = 2)
round(lrrr.out$B - lrrr.out.1090$B, digits = 3)
```

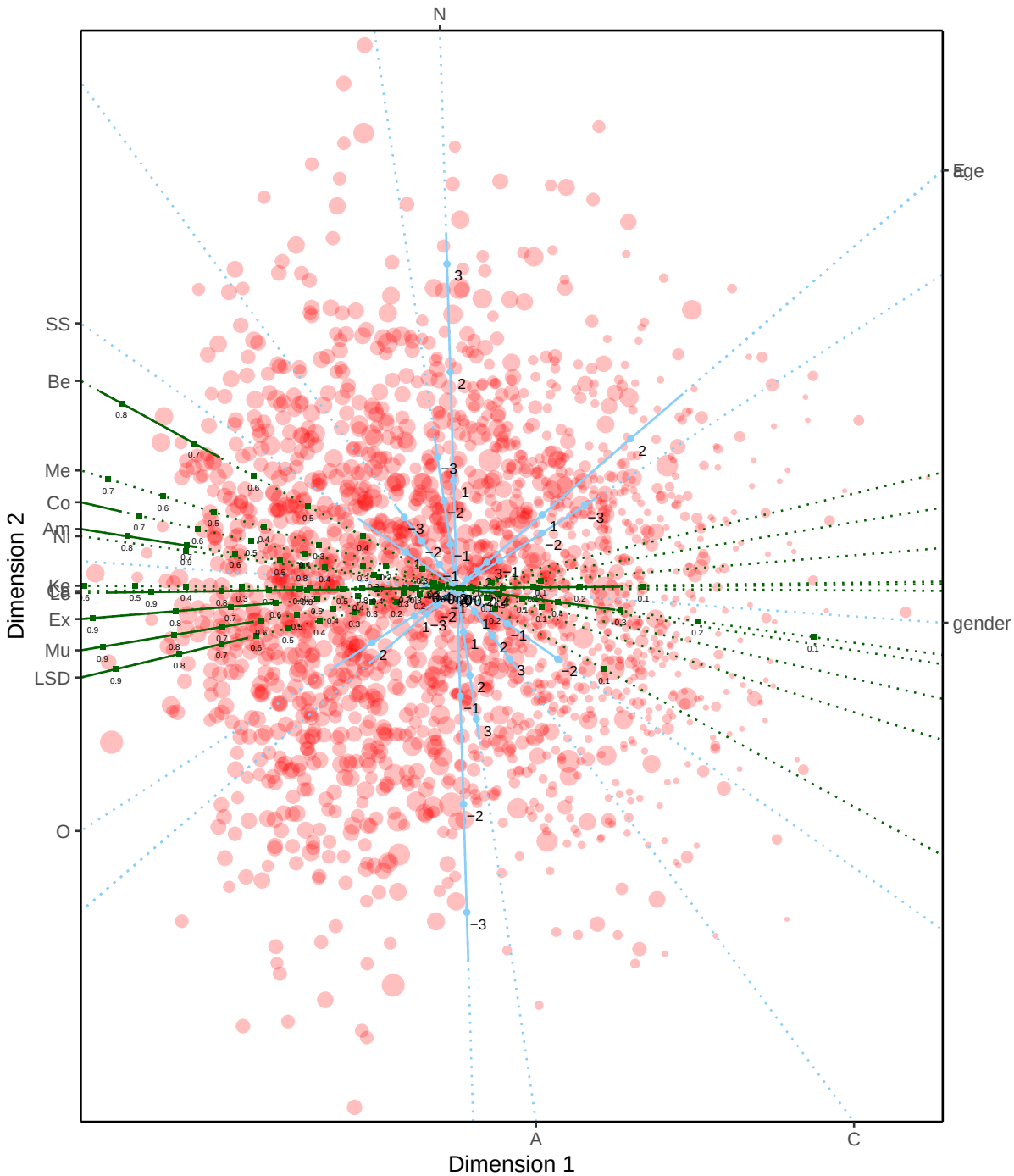
```
       [,1]   [,2]
[1,] -0.001 -0.001
[2,] -0.001  0.003
[3,]  0.000  0.001
[4,] -0.002  0.002
[5,]  0.002 -0.002
[6,] -0.002  0.001
[7,]  0.003 -0.001
[8,] -0.002  0.001
```

```r
round(lrrr.out$V - lrrr.out.1090$V, digits = 3)
```

```
       [,1]   [,2]
 [1,] 0.011 -0.001
 [2,] 0.006 -0.002
 [3,] 0.008 -0.001
 [4,] 0.009 -0.001
 [5,] 0.010  0.001
 [6,] 0.000  0.000
 [7,] 0.013 -0.001
 [8,] 0.023  0.002
 [9,] 0.014 -0.003
[10,] 0.017  0.001
[11,] 0.003  0.000
```

The triplot for this analysis is as follows

```r
plot.lpcah(lrrr.out.1090, res = TRUE, ocol= "red")
```



Deleting both from the data, we obtain

```r
lrrr.out.both = lpca(Y = Y[-c(1090, 1300), ], X = X[-c(1090, 1300), ], S = 2)
round(lrrr.out$B - lrrr.out.both$B, digits = 3)
```

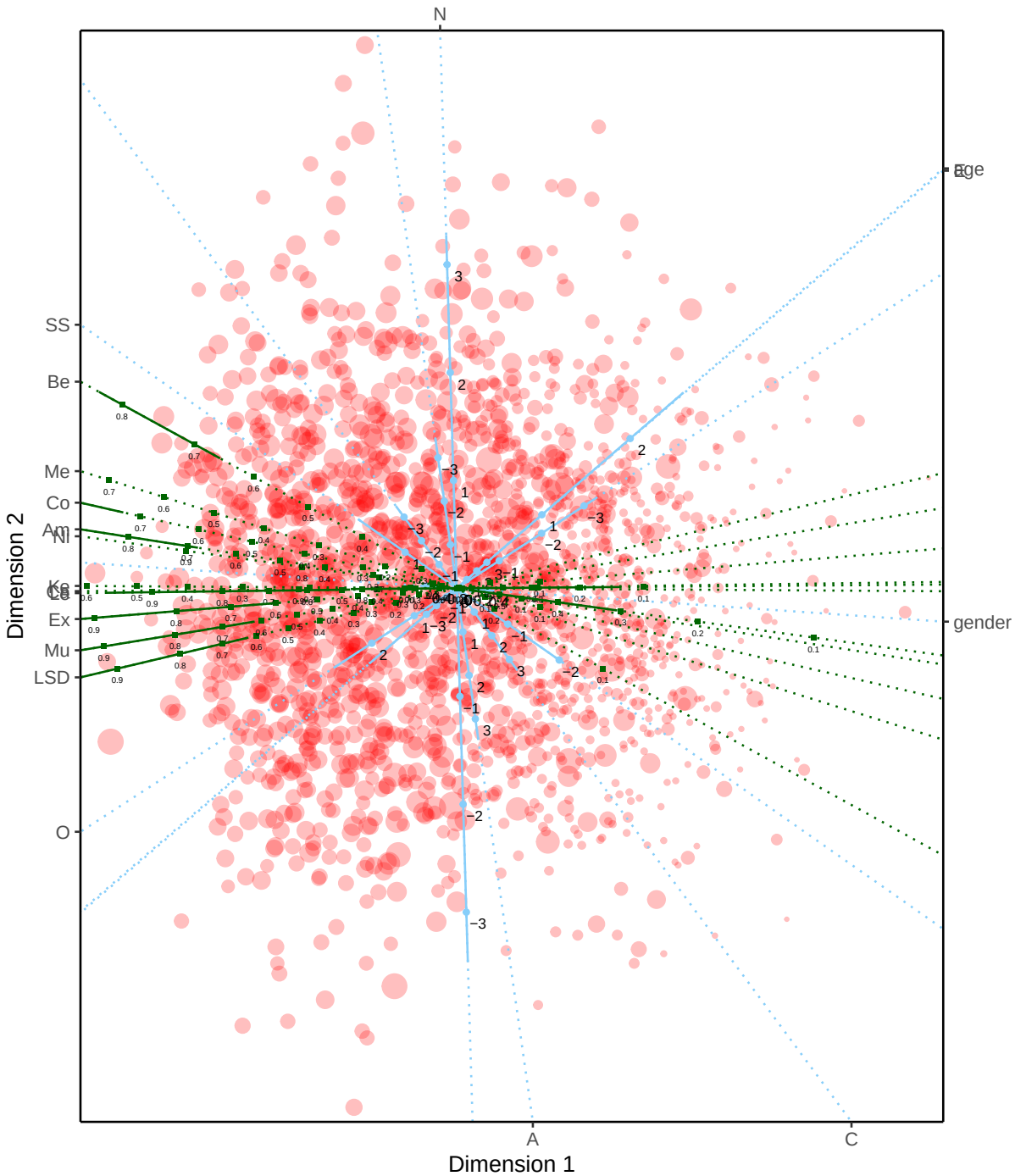
```
       [,1]   [,2]
[1,]  0.000 -0.002
[2,]  0.006  0.001
[3,]  0.000  0.002
[4,] -0.002  0.002
[5,]  0.001 -0.002
[6,] -0.001  0.000
[7,]  0.003  0.000
[8,]  0.002  0.000
```

```r
round(lrrr.out$V - lrrr.out.both$V, digits = 3)
```

```
       [,1]   [,2]
 [1,] 0.020 -0.002
 [2,] 0.011 -0.003
 [3,] 0.008 -0.002
 [4,] 0.017 -0.002
 [5,] 0.019  0.001
 [6,] 0.015  0.000
 [7,] 0.023 -0.001
 [8,] 0.040  0.004
 [9,] 0.022 -0.004
[10,] 0.029  0.002
[11,] 0.002 -0.001
```

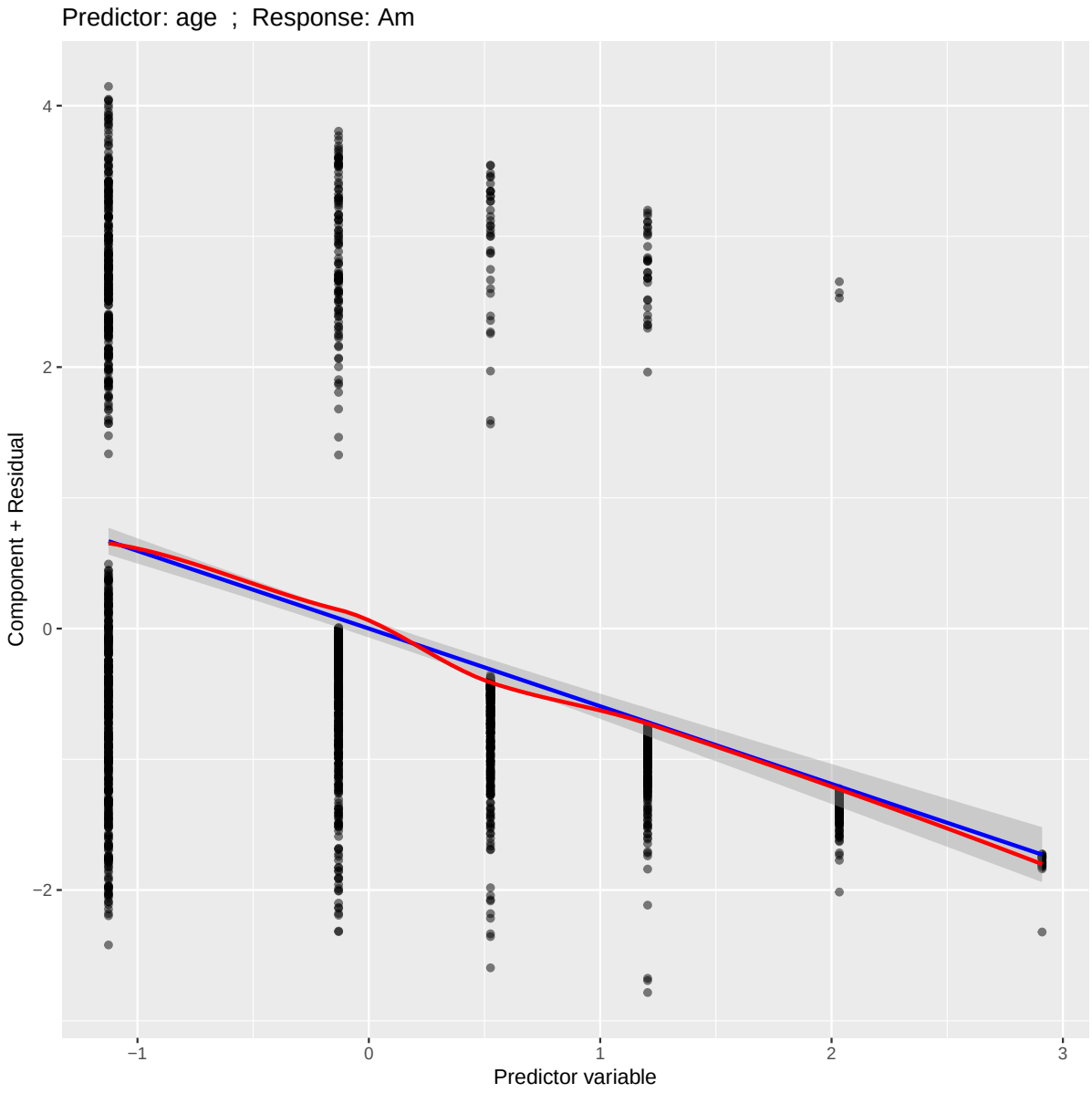
with triplot

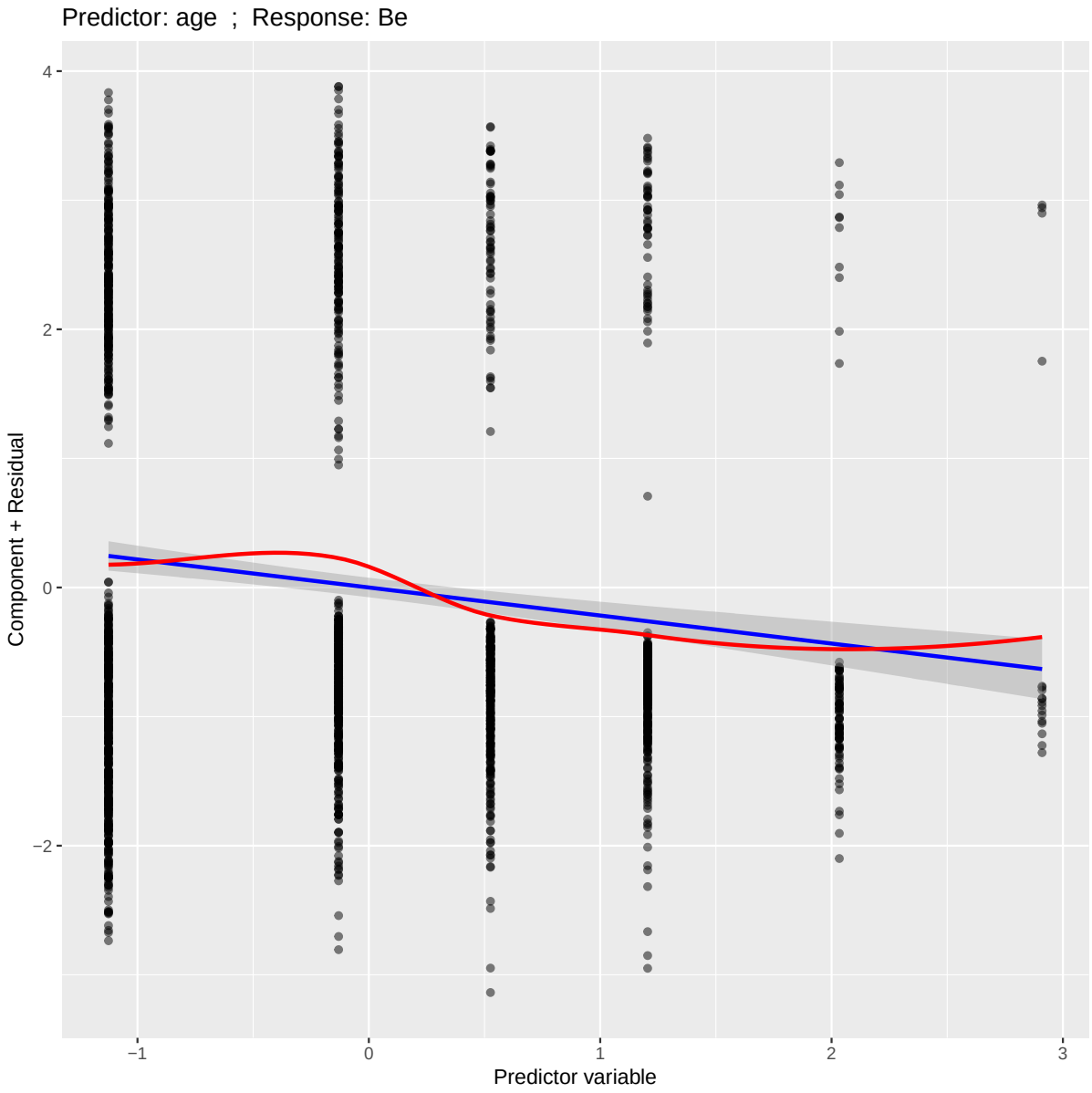
```r
plot.lpcah(lrrr.out.both, res = TRUE, ocol= "red")
```

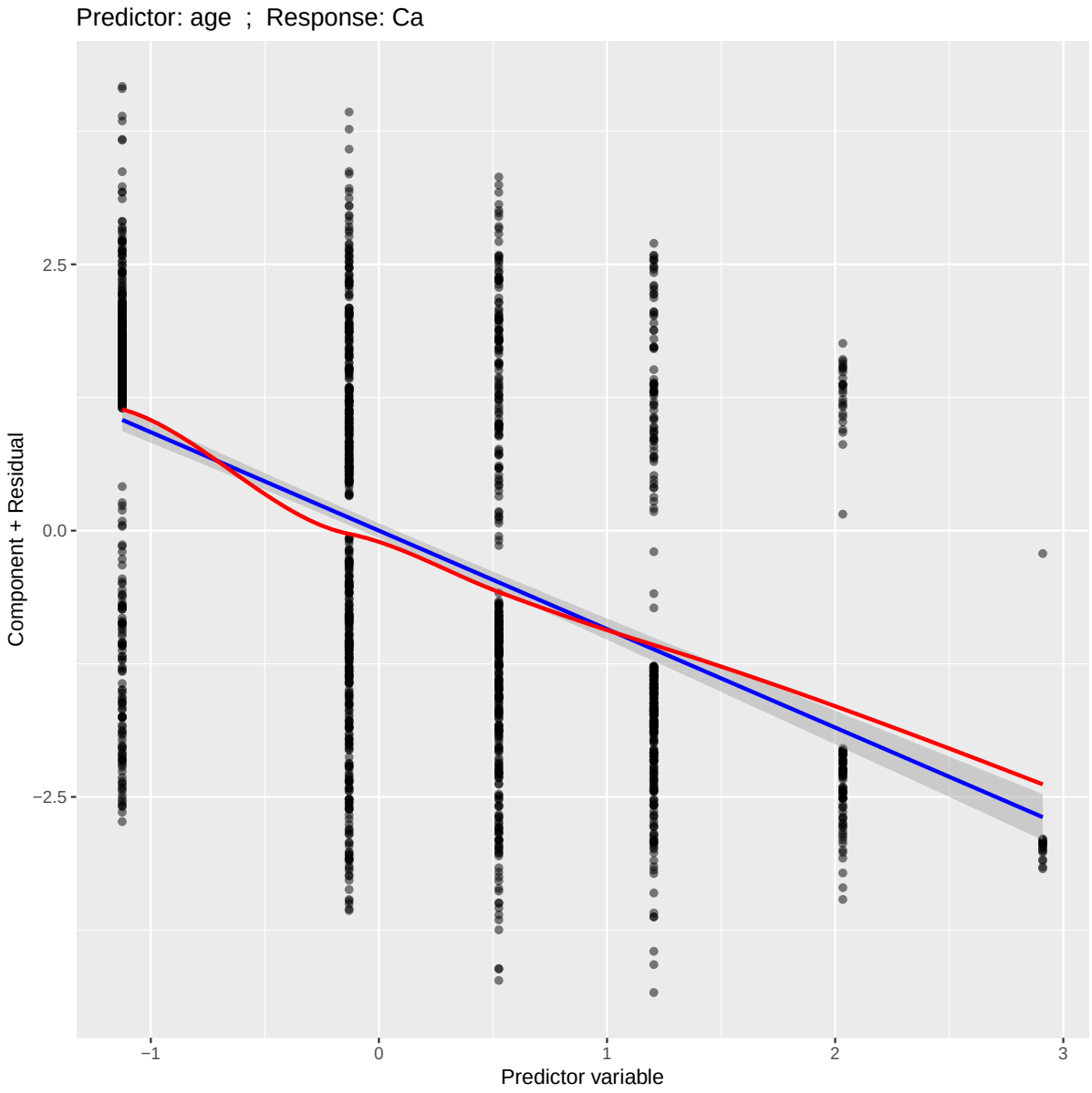


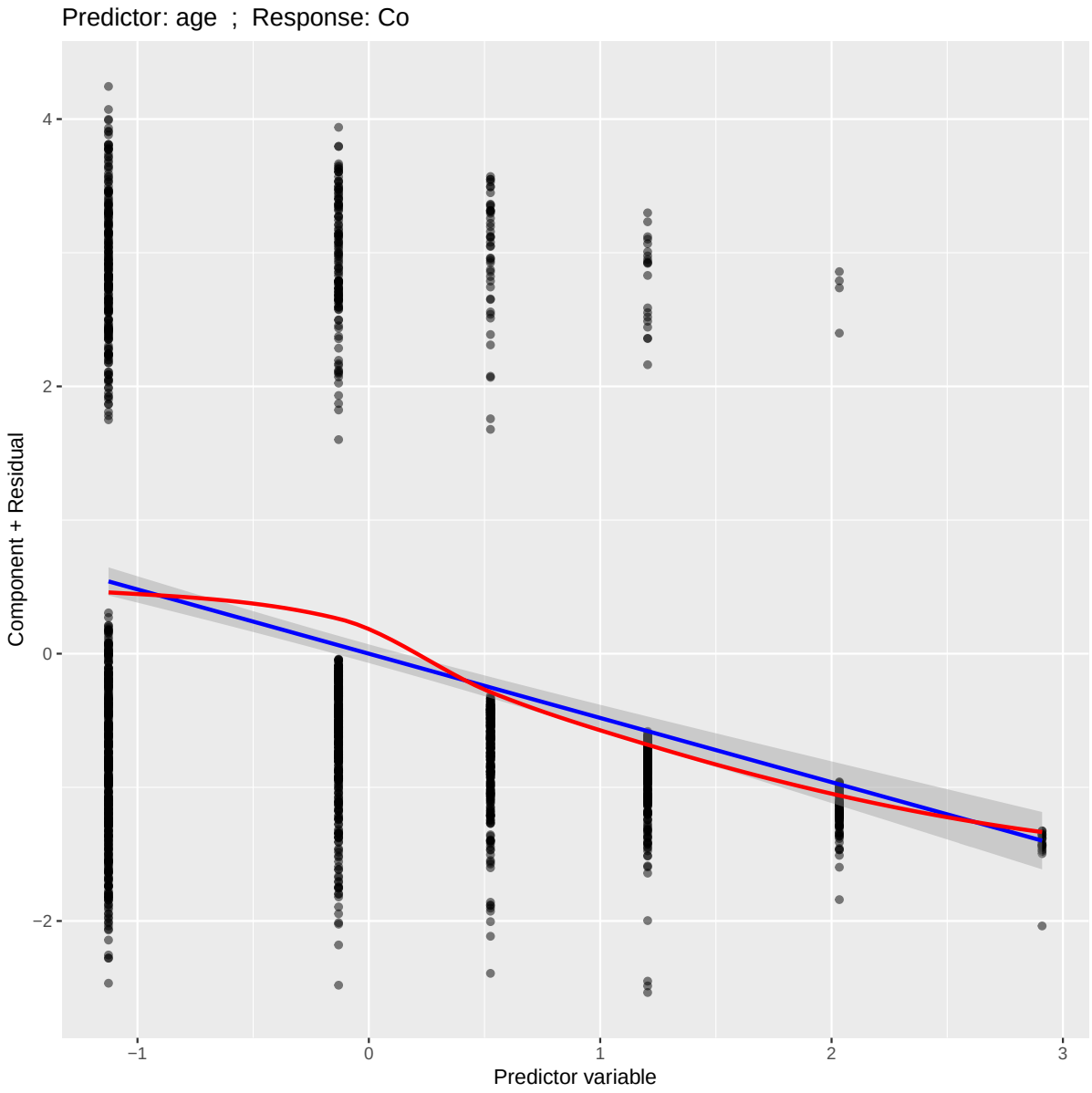
Non-linearity

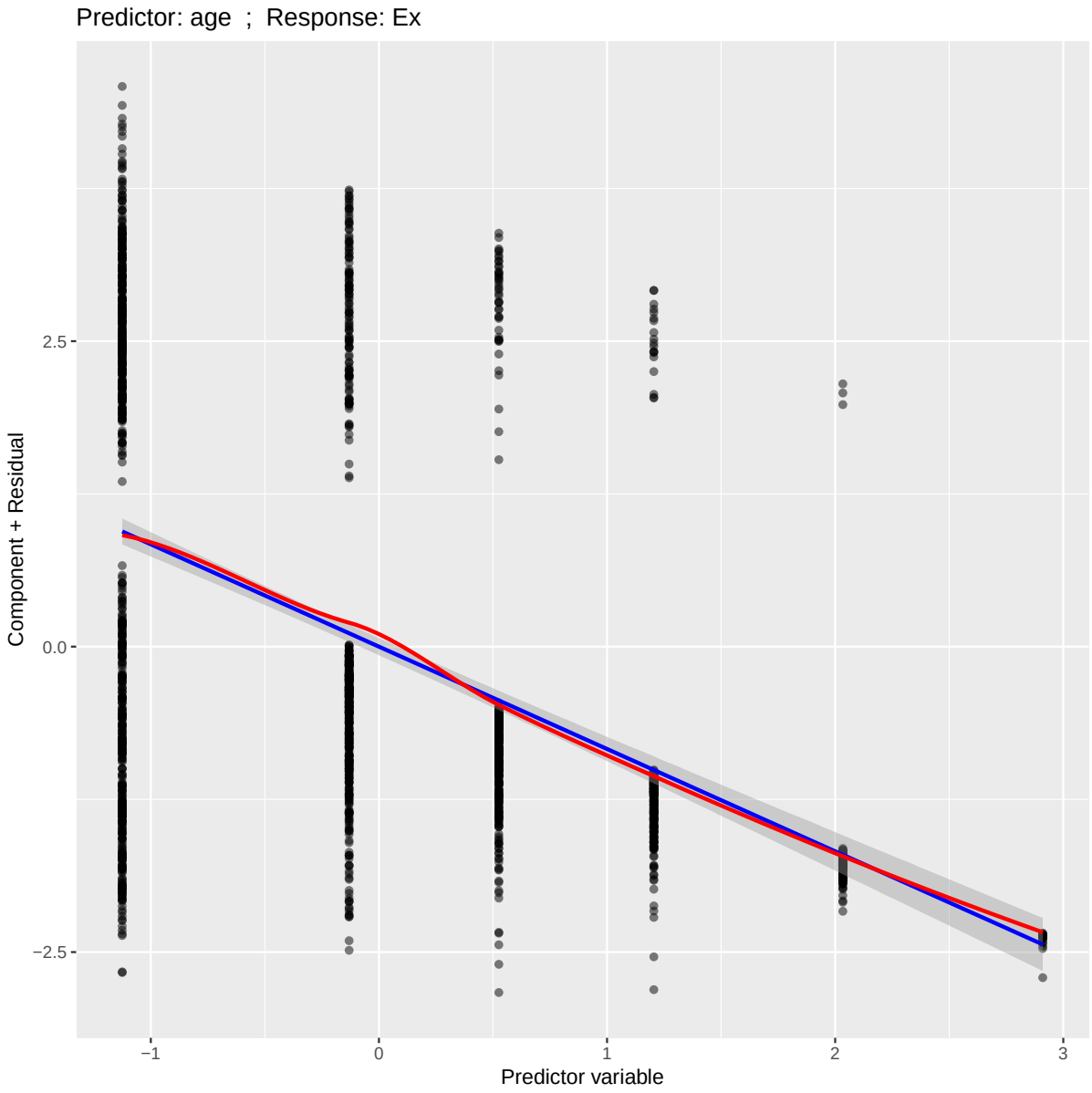
First predictor

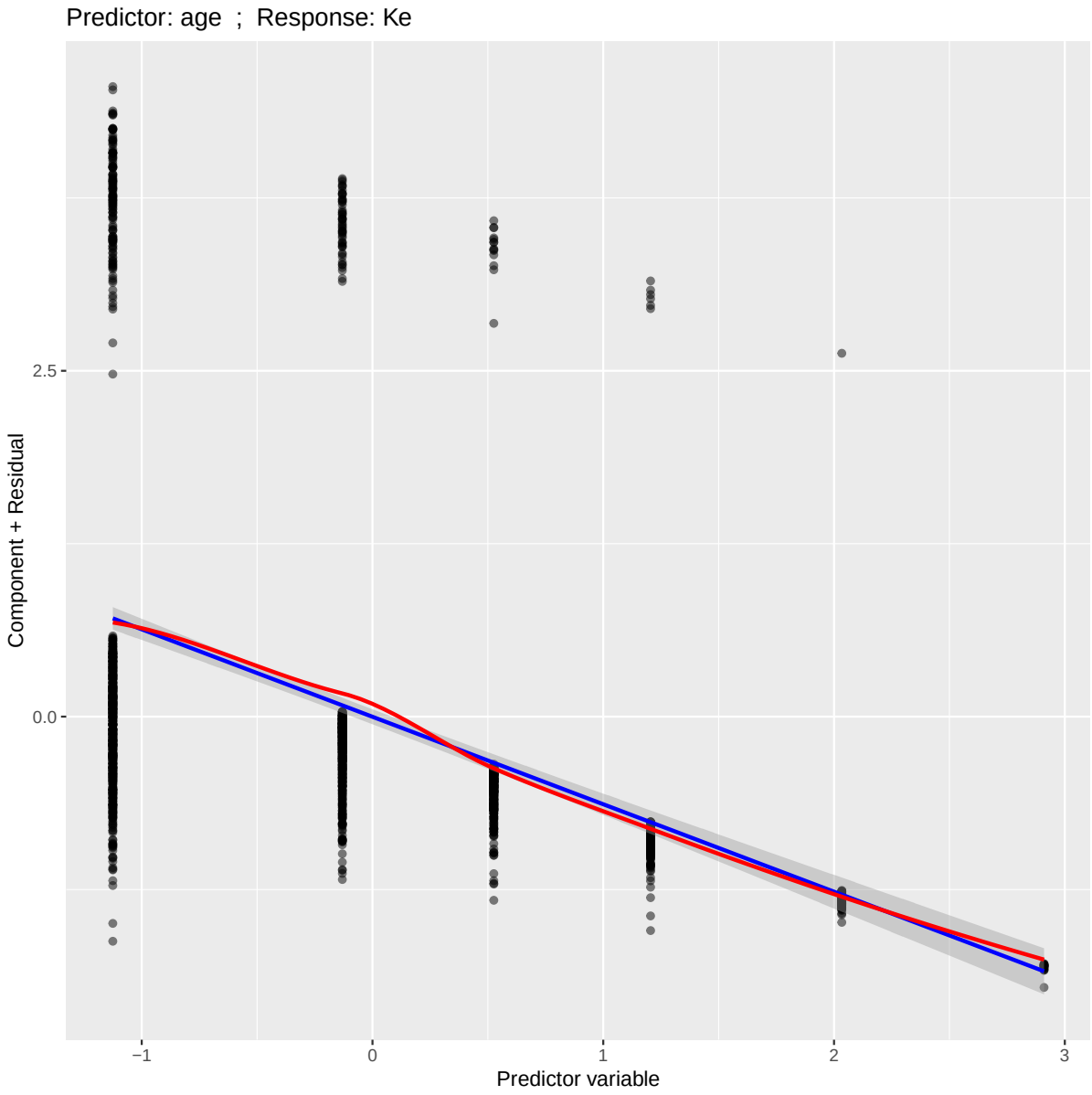

Predictor: age ; Response: Am
Component + Residual
4
2
0
−2
−1
0
1
2
3
Predictor variable

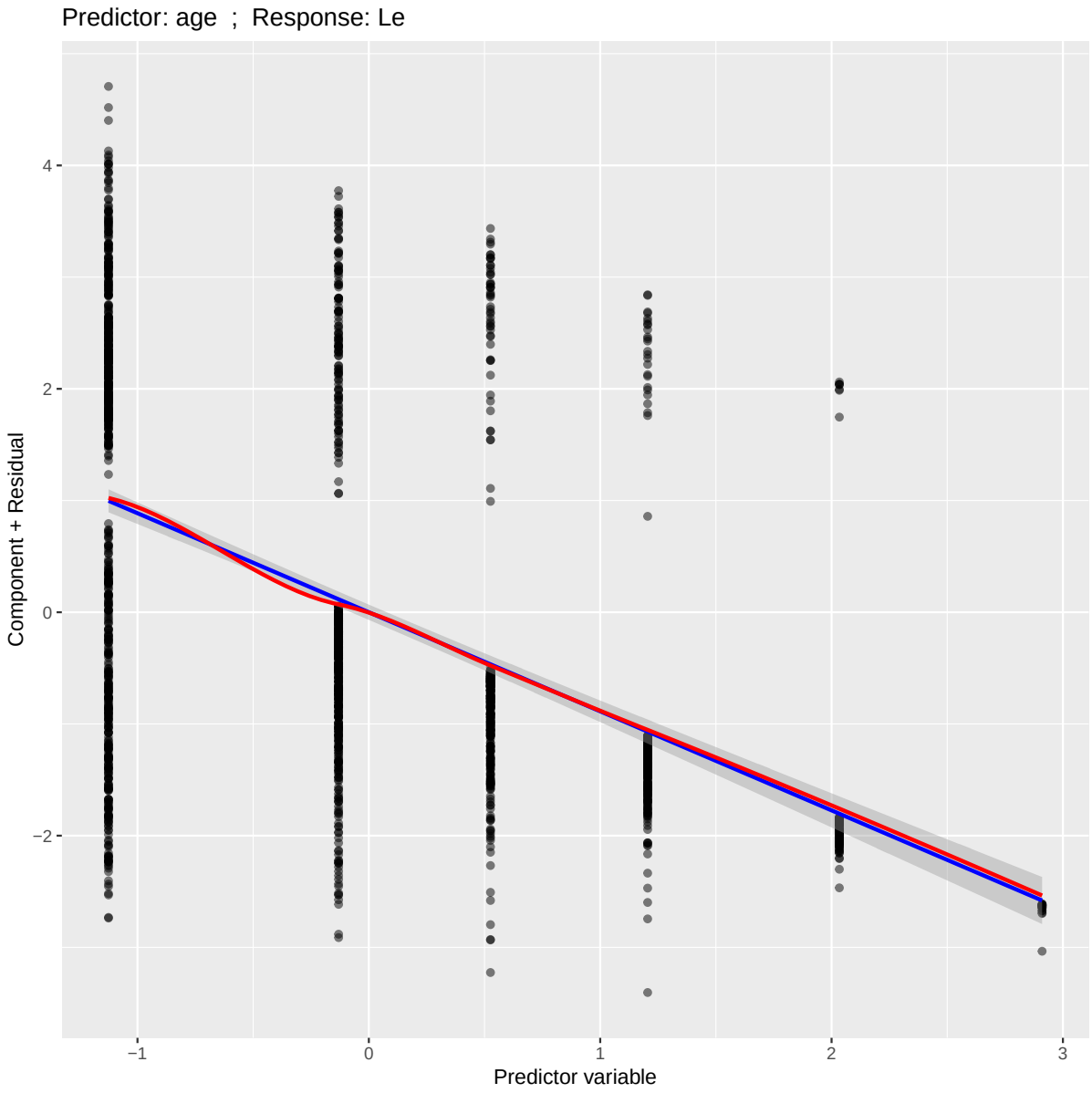

Predictor: age ; Response: Be
Component + Residual
4
2
0
−2
−1
0
1
2
3
Predictor variable

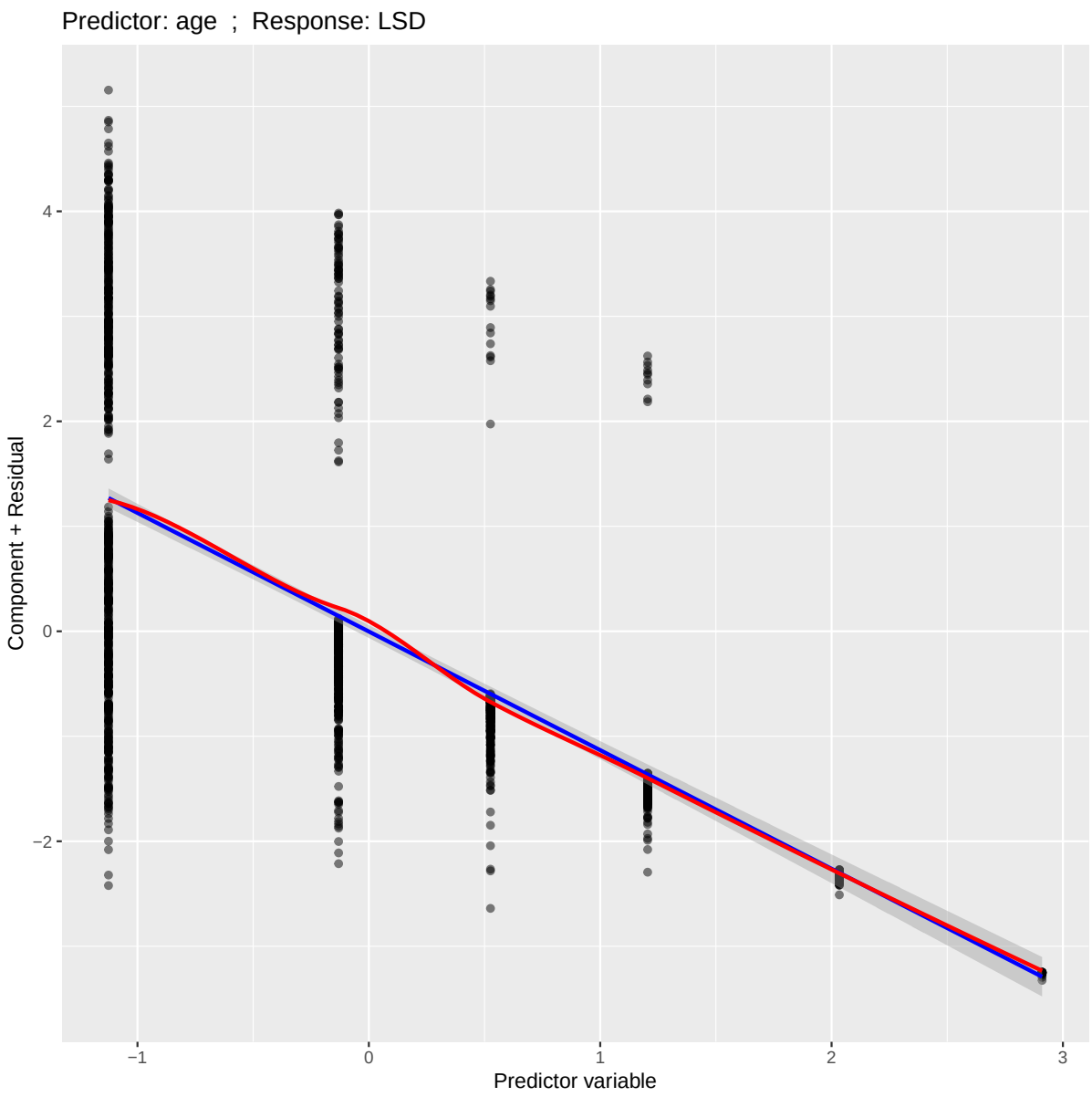

Predictor: age ; Response: Ca
Component + Residual
2.5
0.0
−2.5
−1
0
1
2
3
Predictor variable

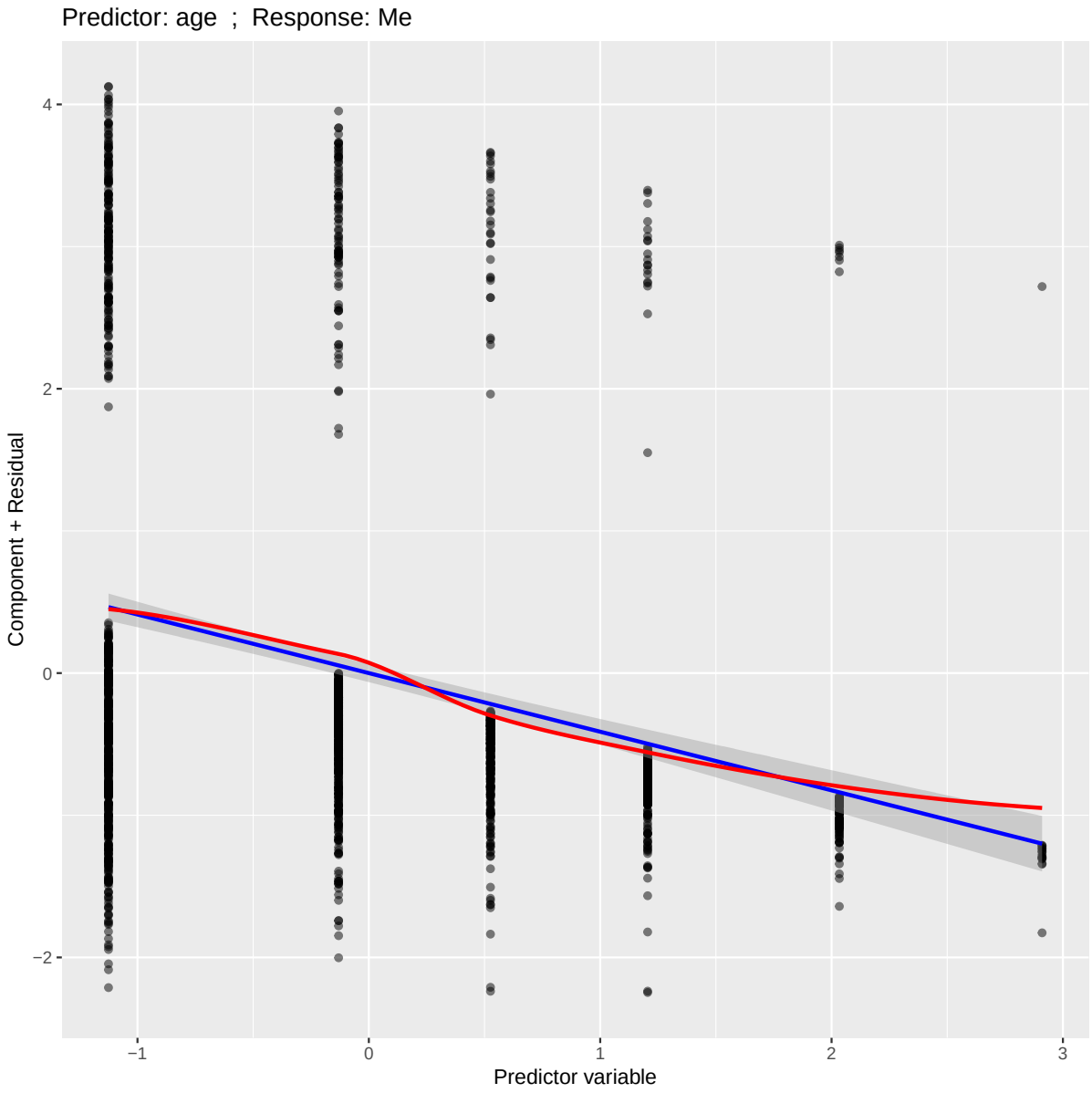

Predictor: age ; Response: Co
Component + Residual
4
2
0
−2
−1
0
1
2
3
Predictor variable

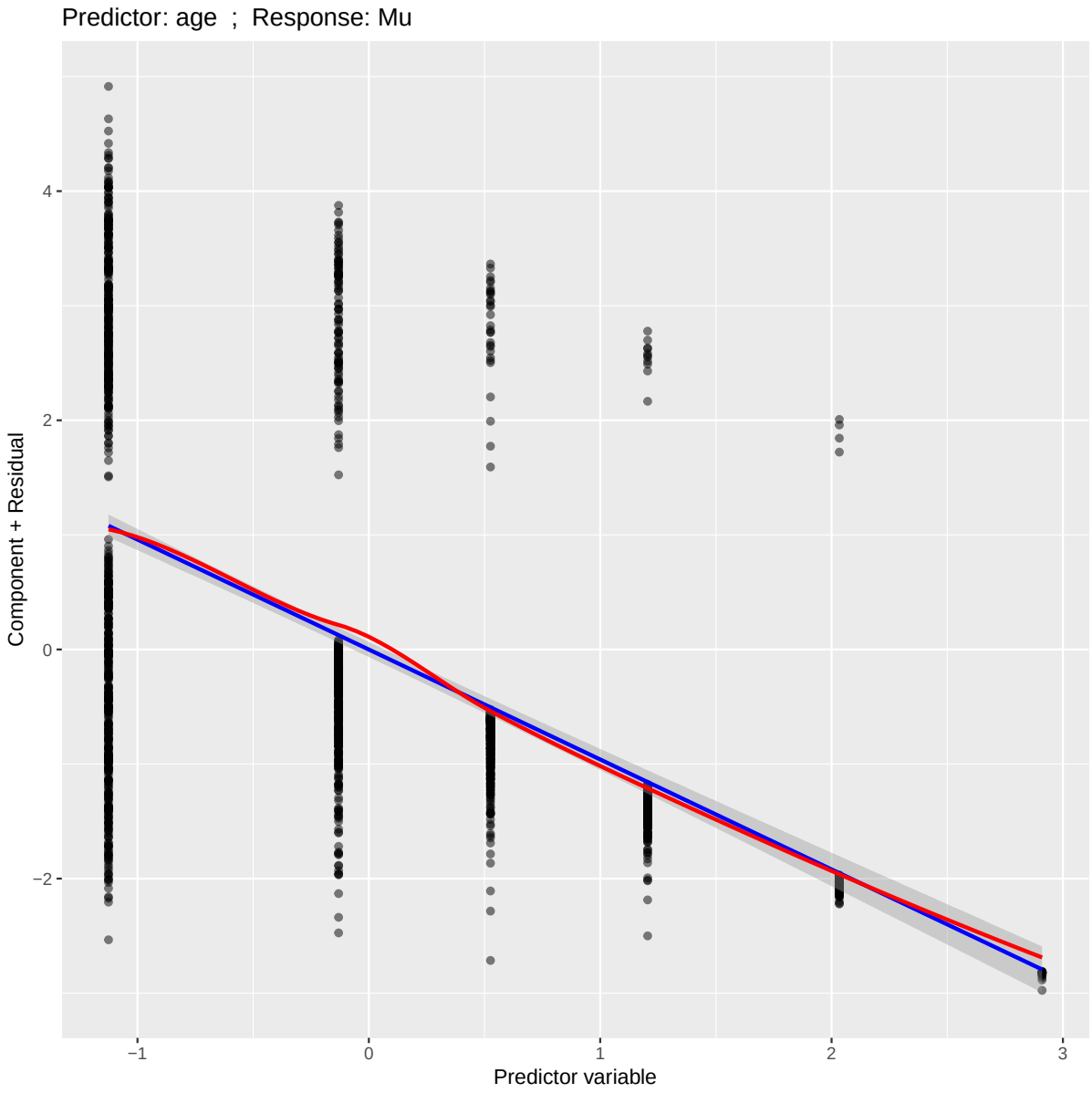

Predictor: age ; Response: Ex
Component + Residual
2.5
0.0
−2.5
−1
0
1
2
3
Predictor variable

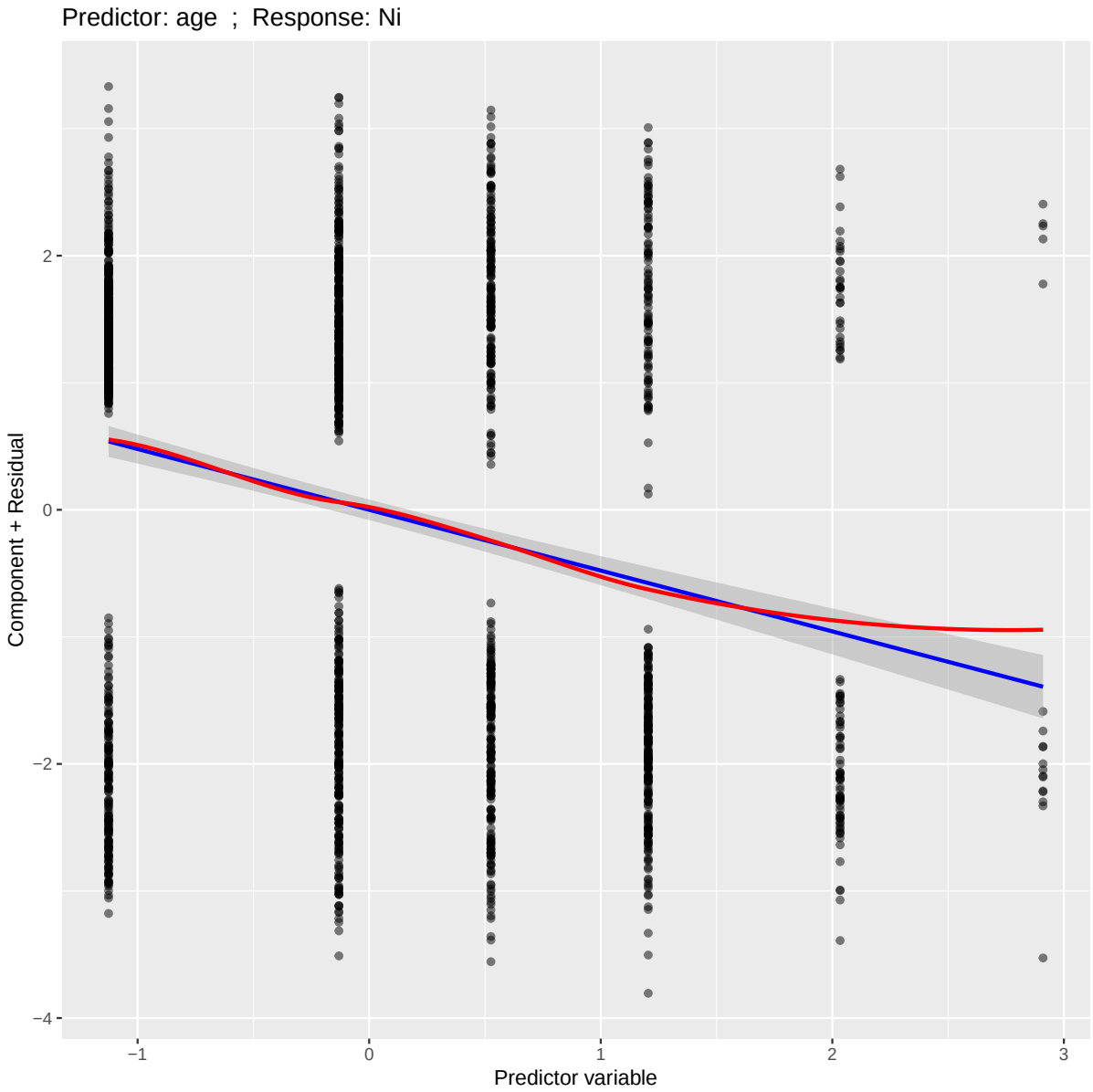

Predictor: age ; Response: Ke
Component + Residual
2.5
0.0
−1
0
1
2
3
Predictor variable


Predictor: age ; Response: Le
Component + Residual
4
2
0
−2
−1
0
1
2
3
Predictor variable

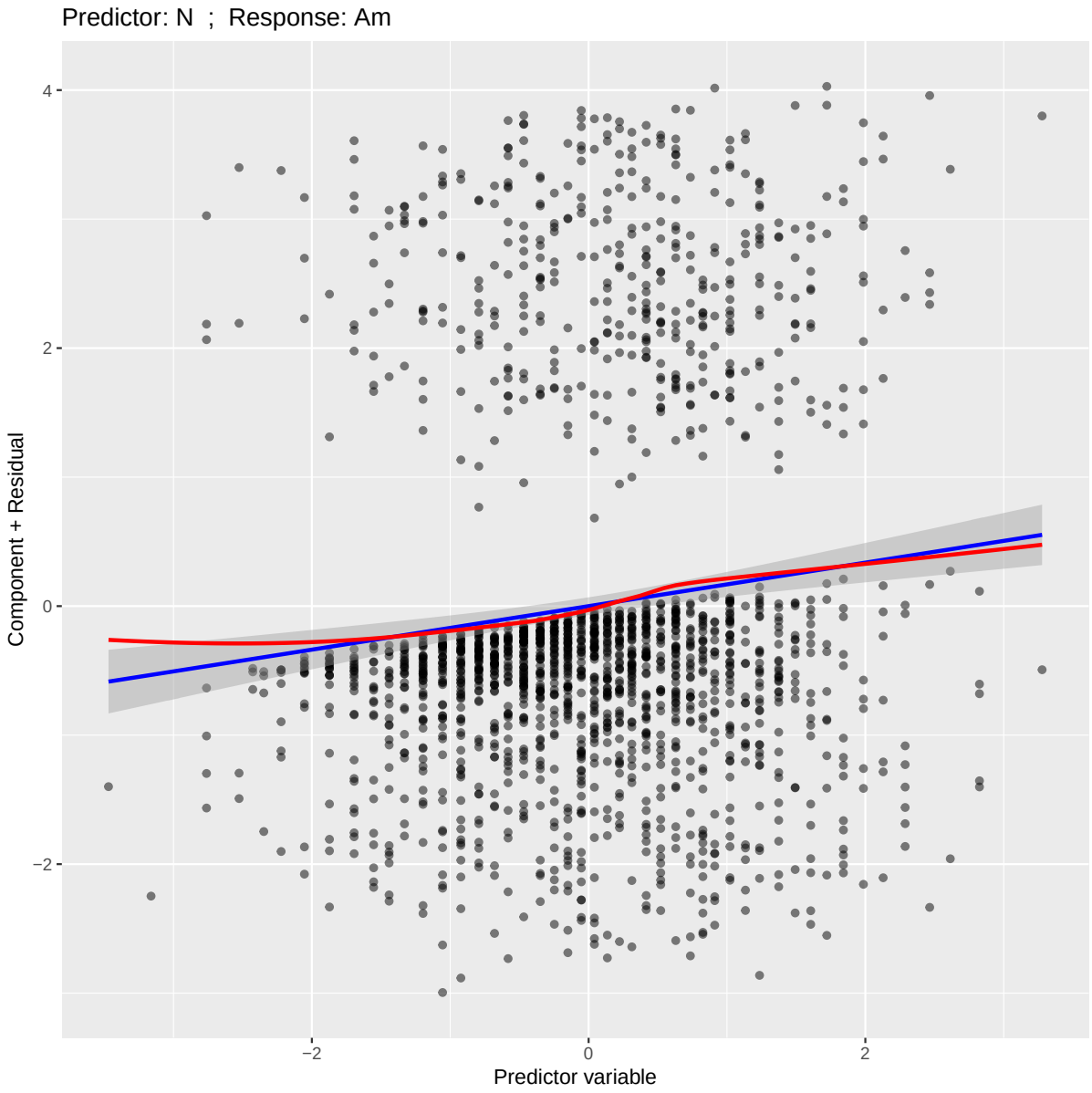

Predictor: age ; Response: LSD
Component + Residual
4
2
0
−2
−1
0
1
2
3
Predictor variable

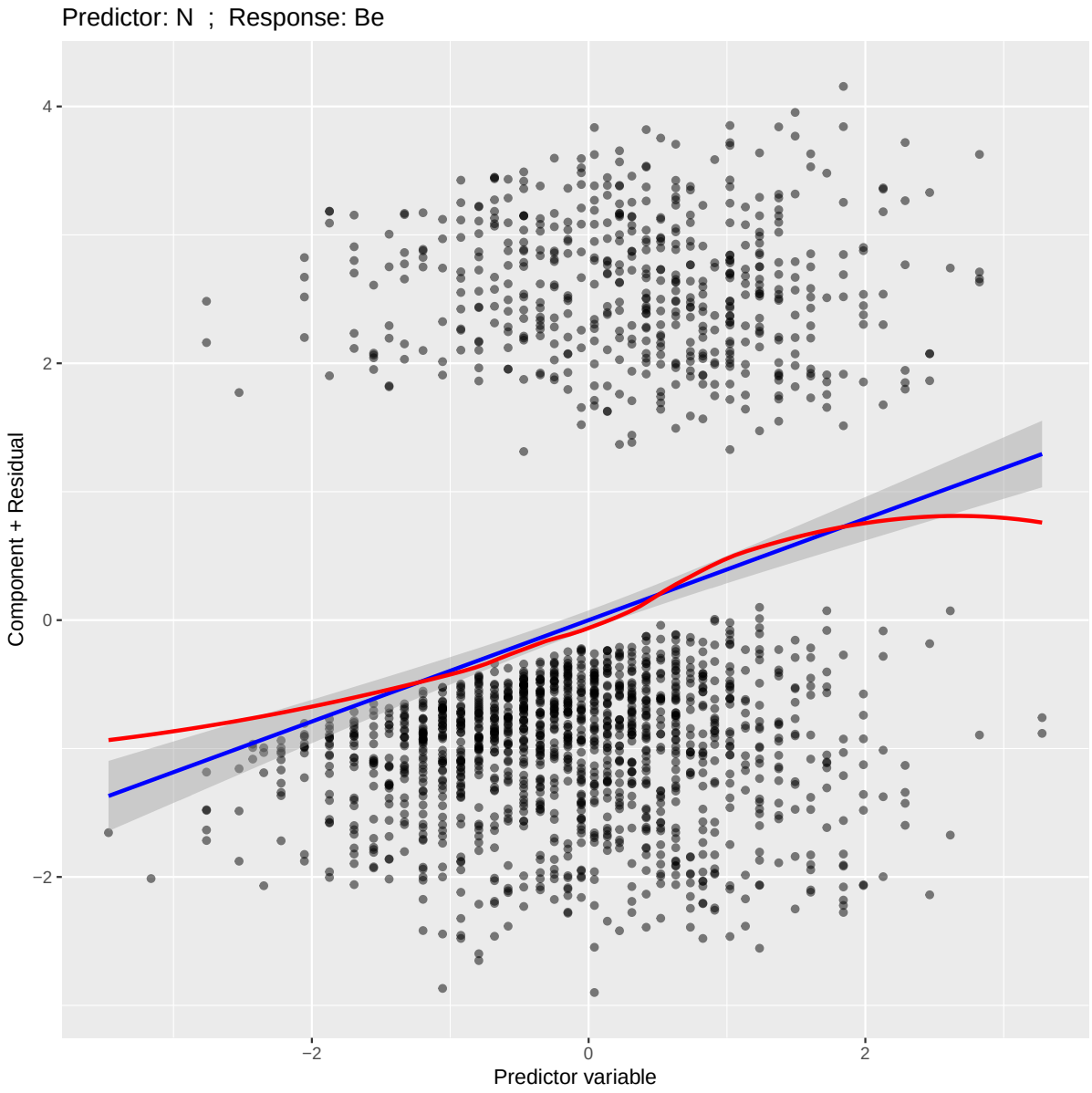

Predictor: age ; Response: Me
Component + Residual
4
2
0
−2
−1
0
1
2
3
Predictor variable

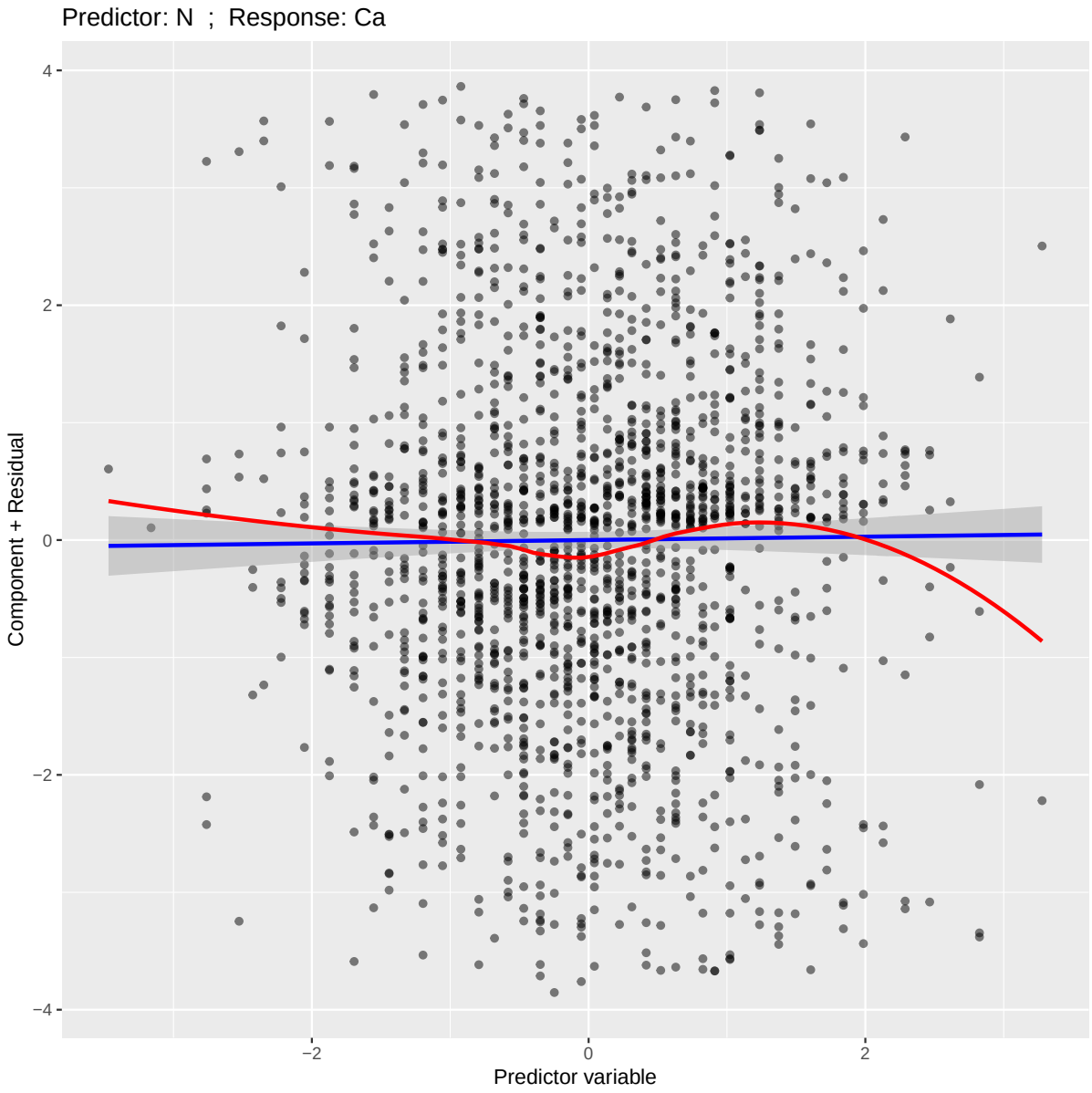

Predictor: age ; Response: Mu
Component + Residual
4
2
0
−2
−1
0
1
2
3
Predictor variable

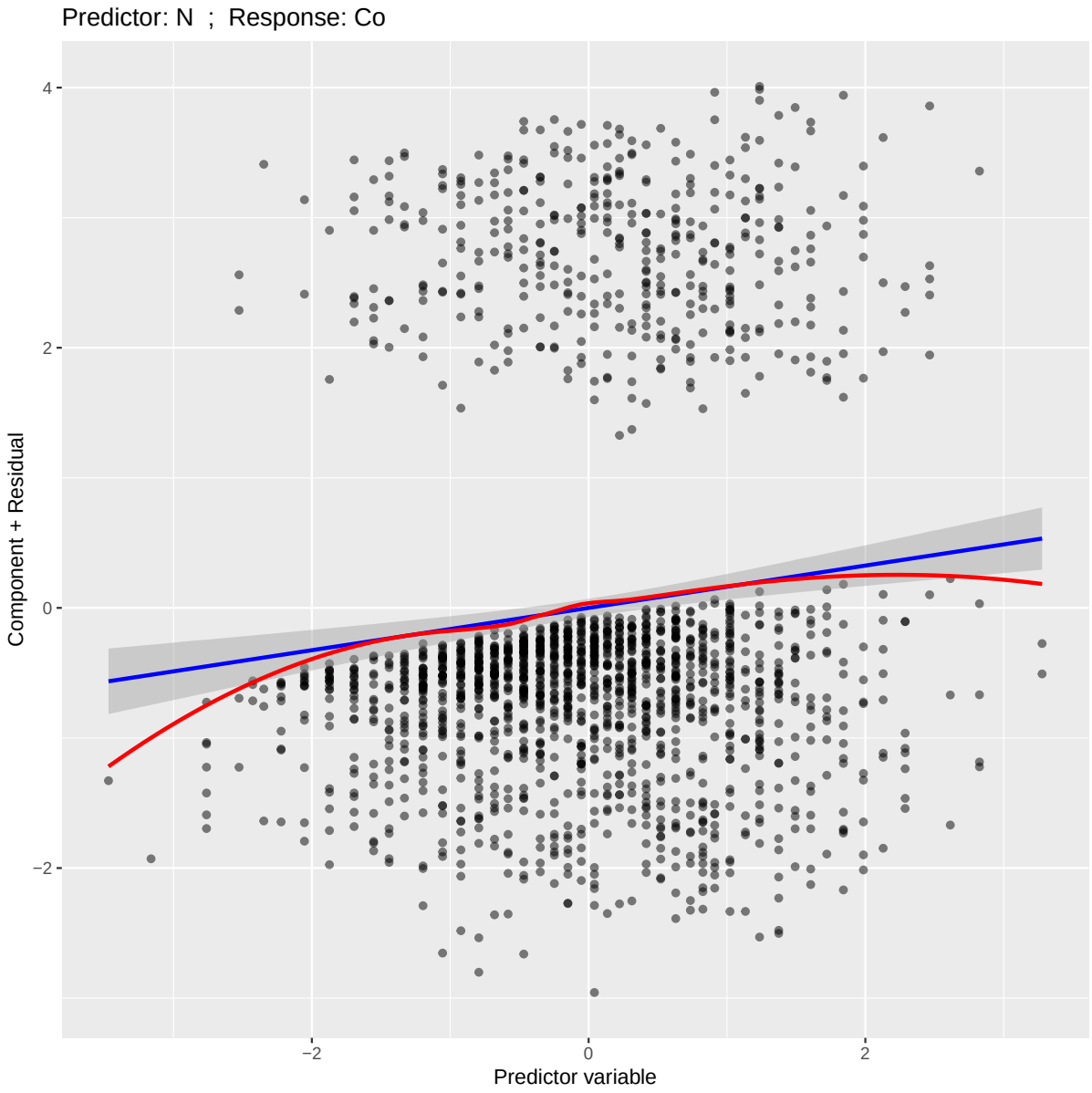


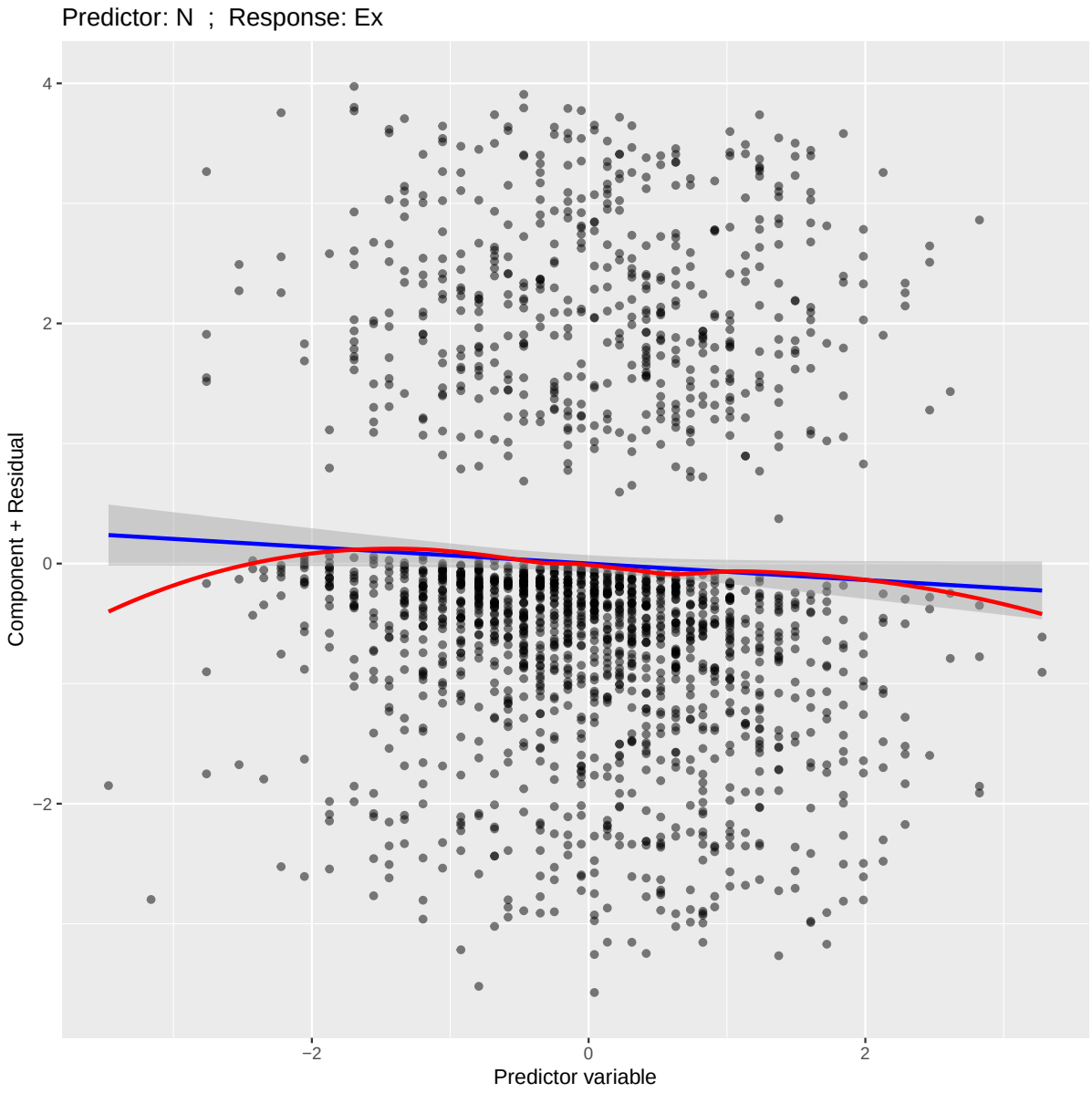
Third predictor

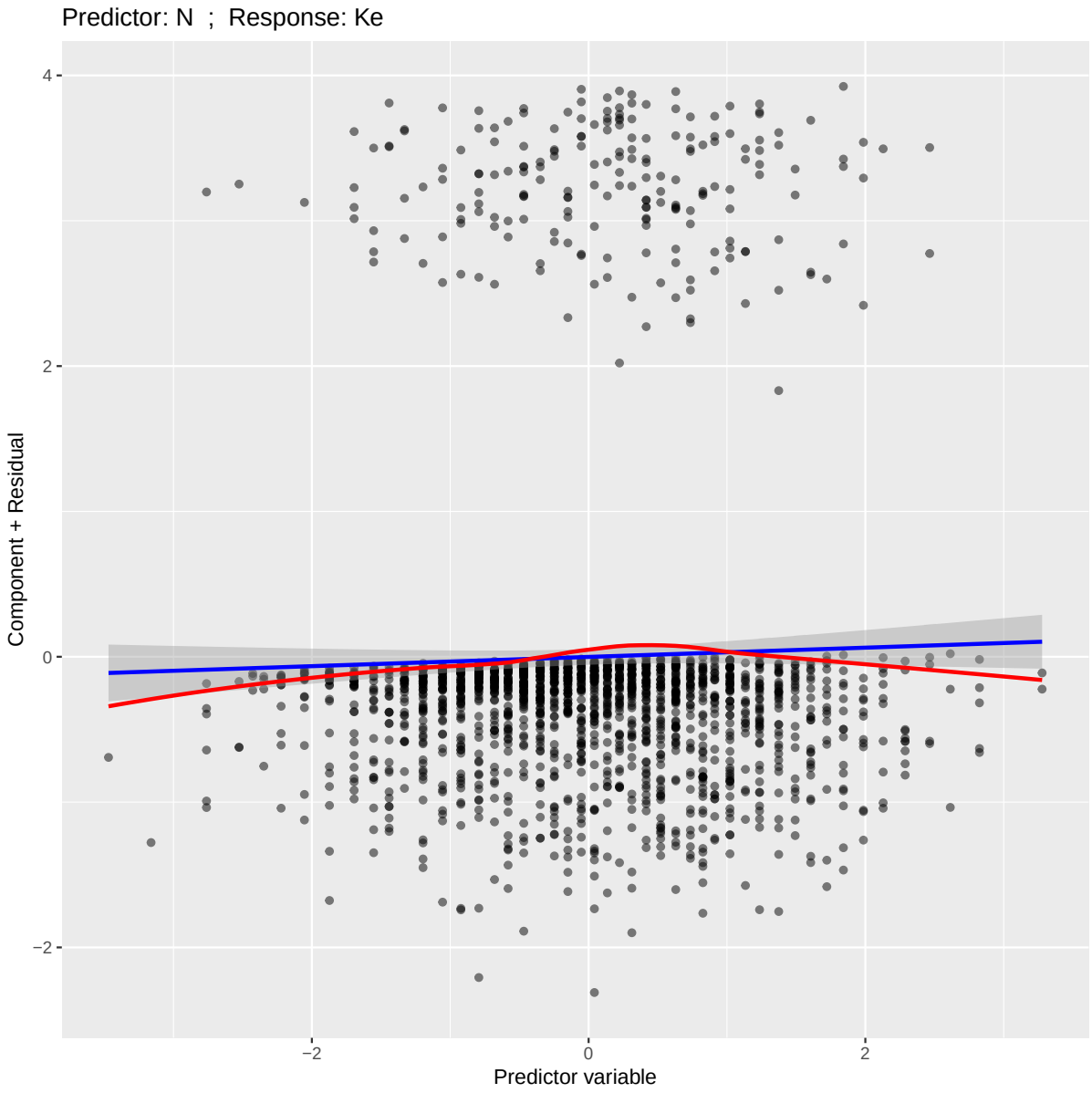

Predictor: N ; Response: Am
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

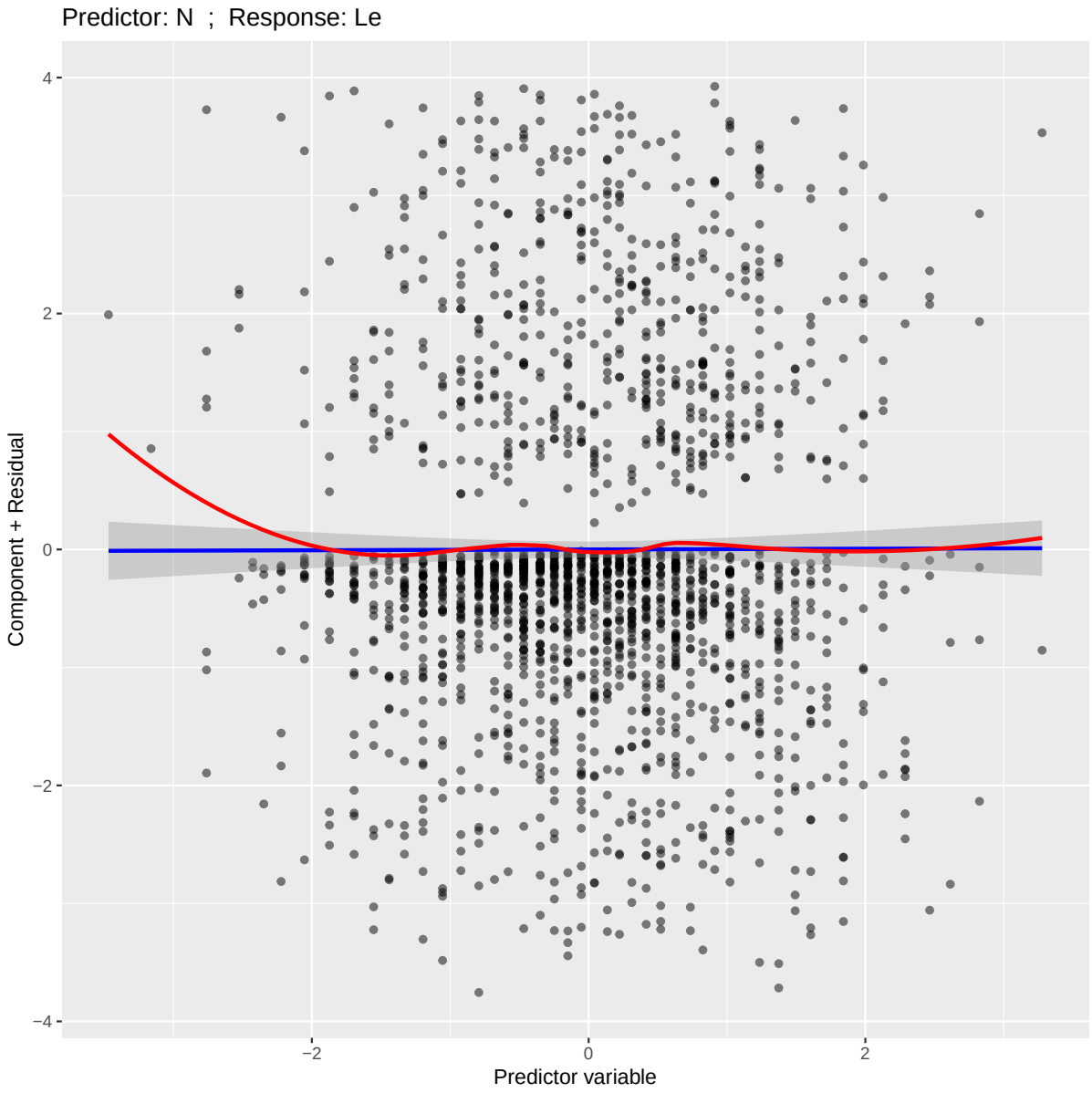

Predictor: N ; Response: Be
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

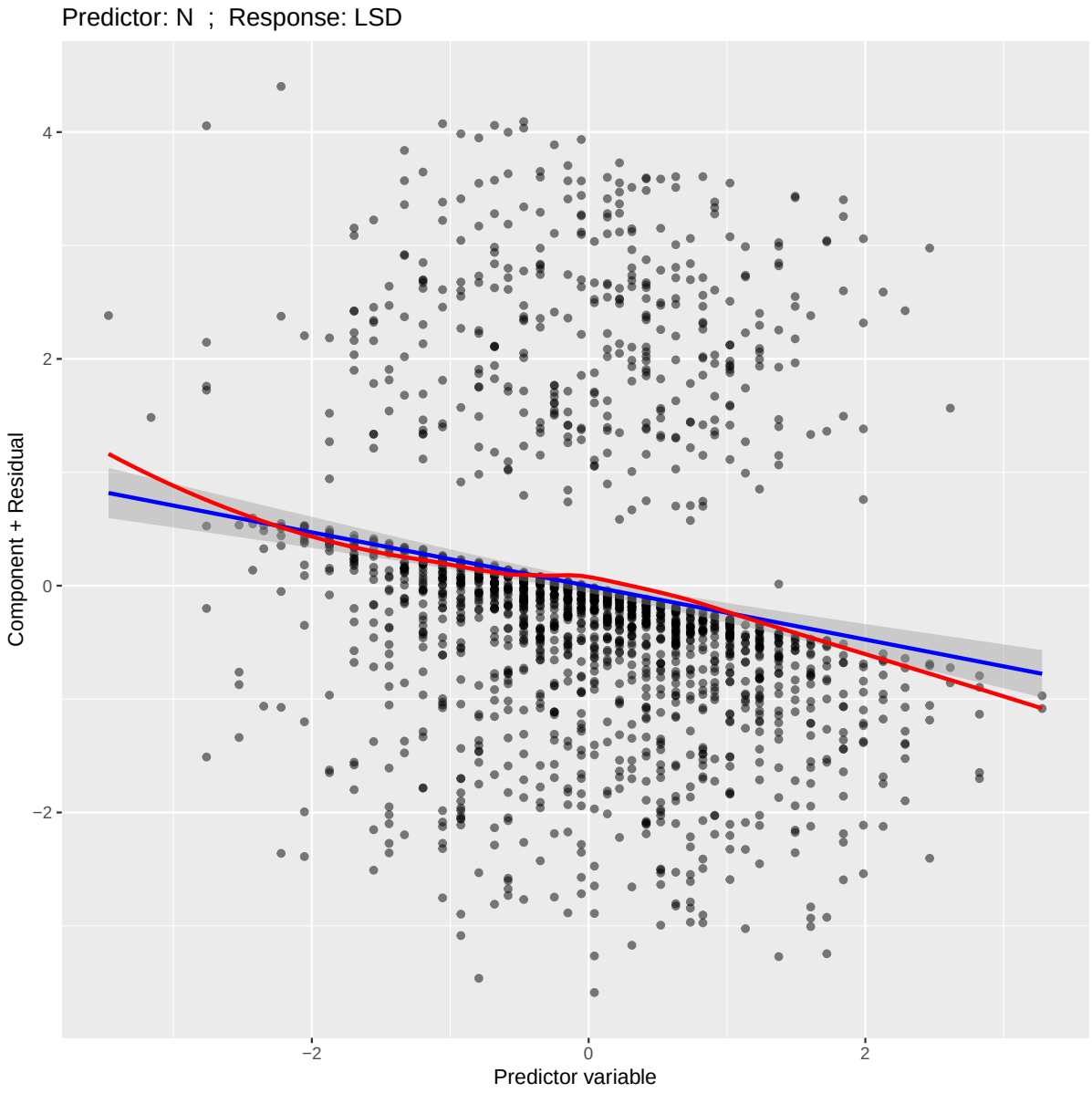

Predictor: N ; Response: Ca
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

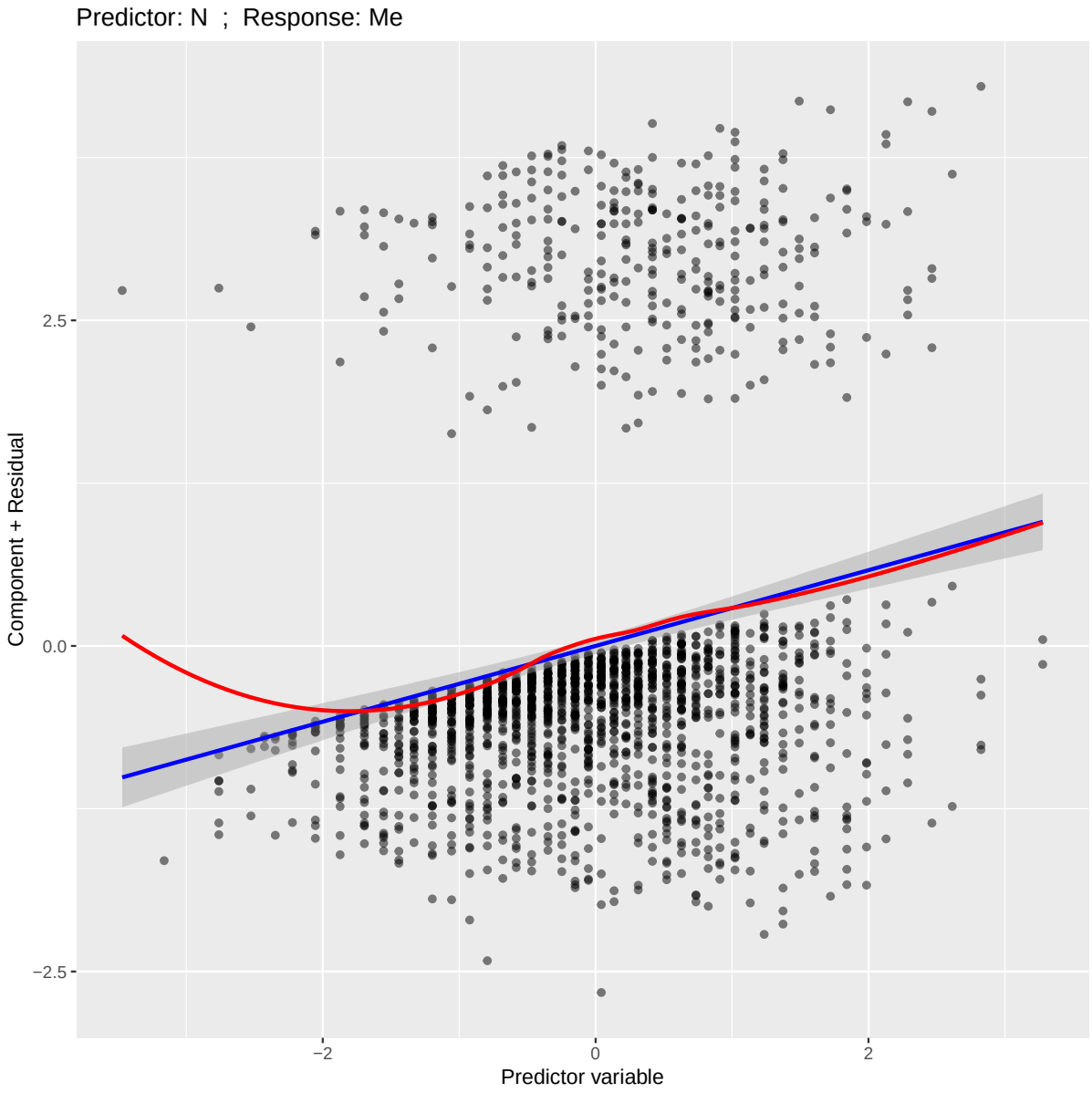

Predictor: N ; Response: Co
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

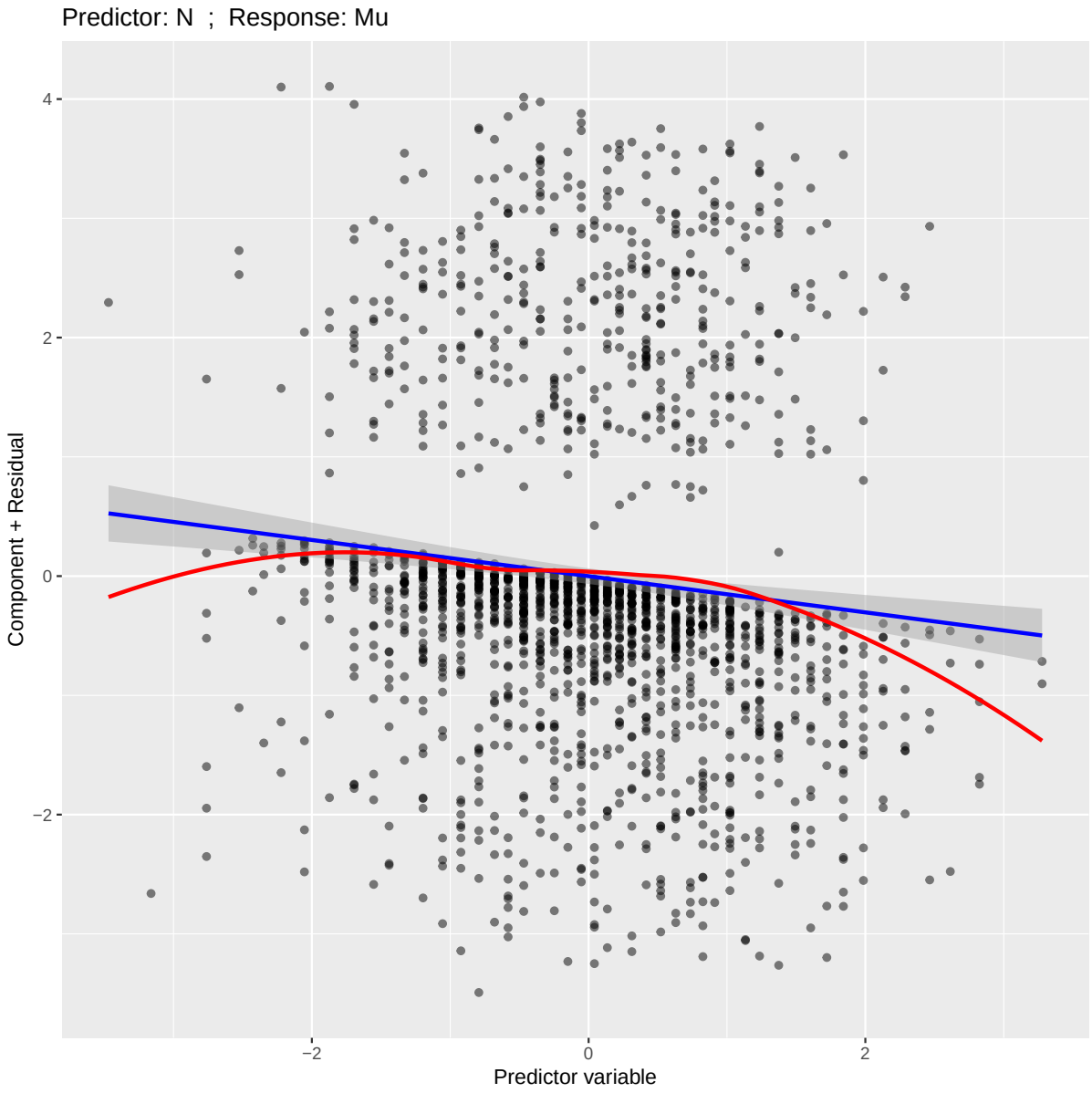

Predictor: N ; Response: Ex
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

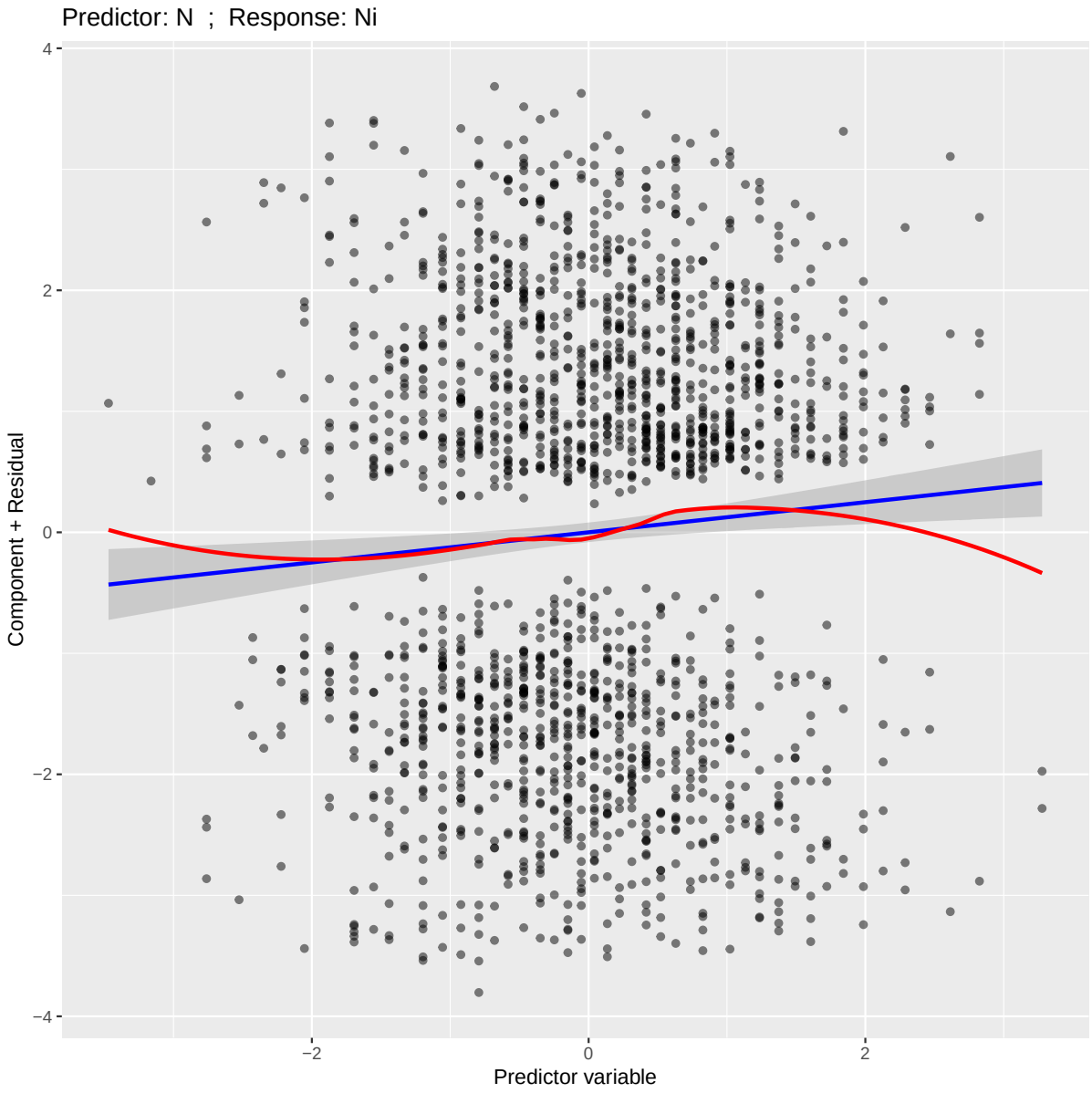

Predictor: N ; Response: Ke
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2


Predictor: N ; Response: Le
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

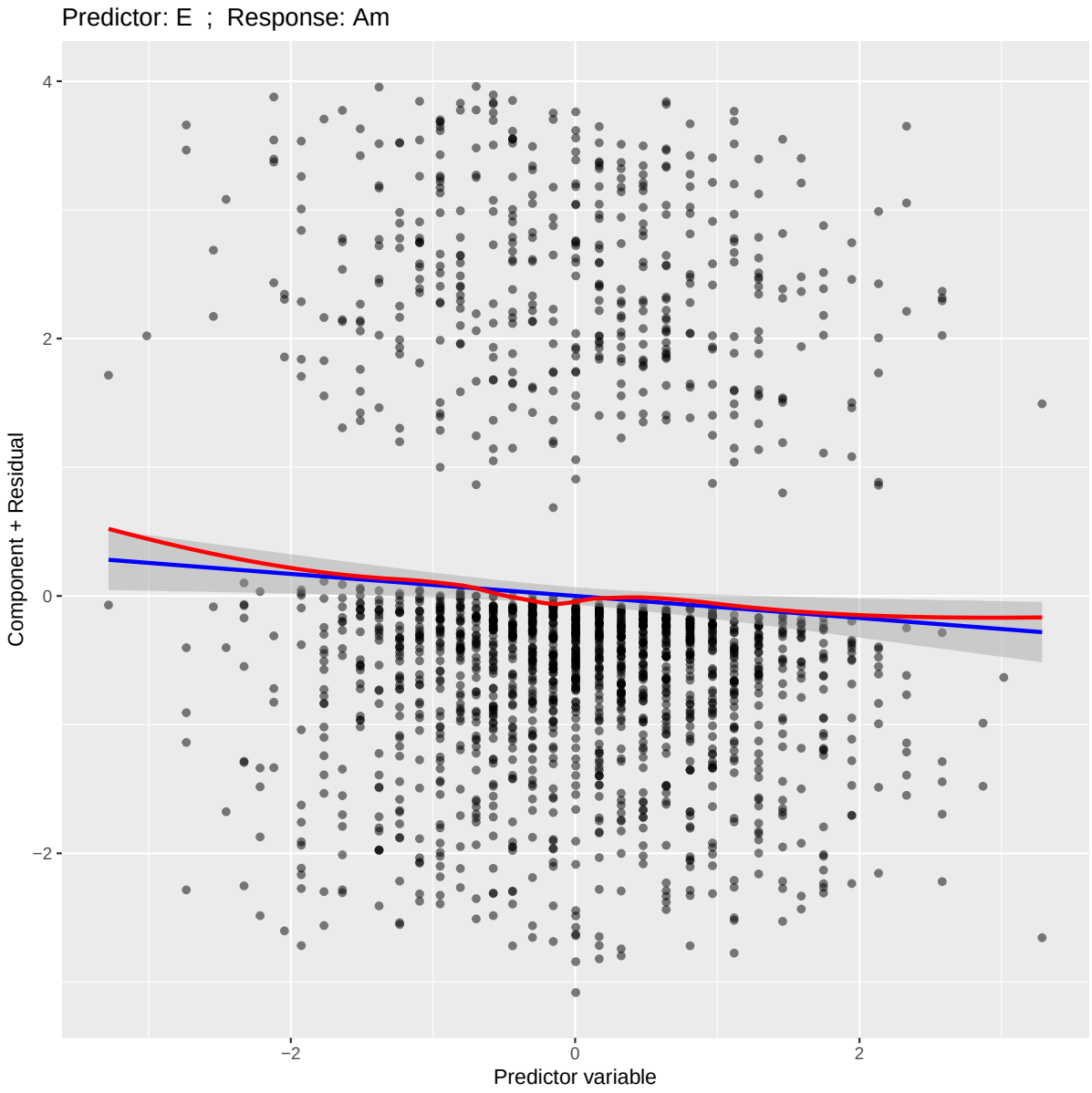

Predictor: N ; Response: LSD
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

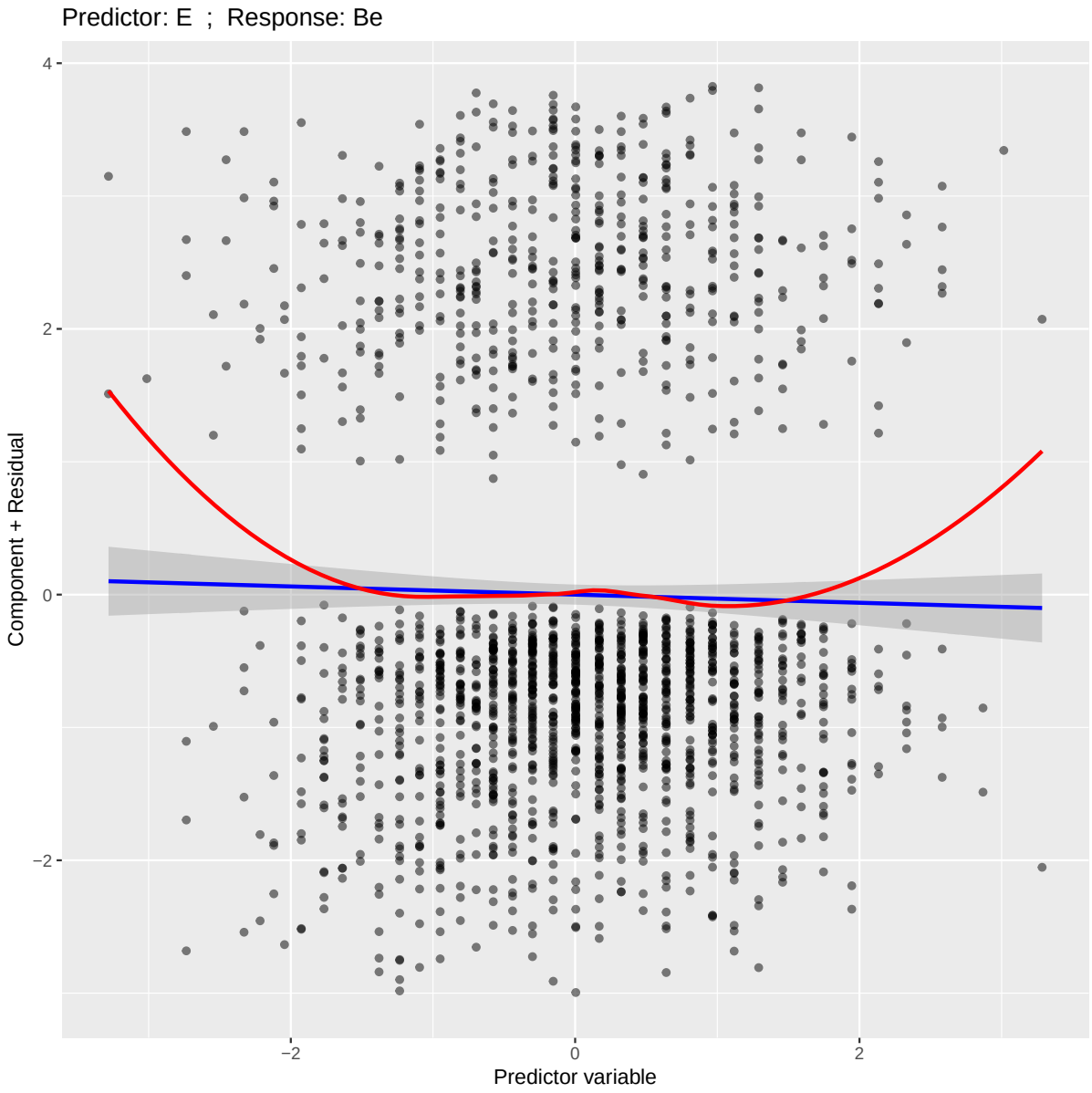

Predictor: N ; Response: Me
Component + Residual
2.5
0.0
−2.5
−2
0
2
Predictor variable

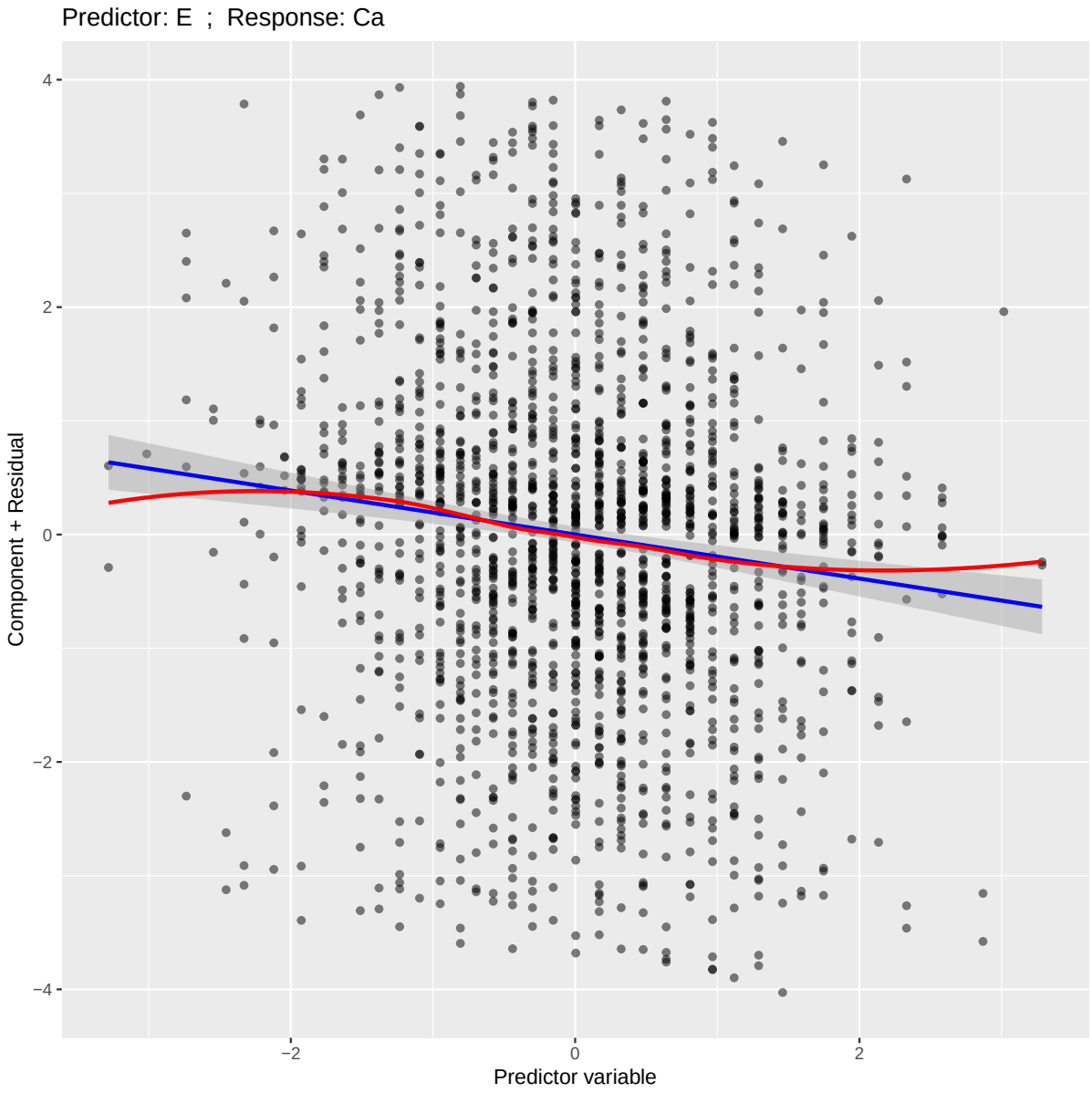

Predictor: N ; Response: Mu
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

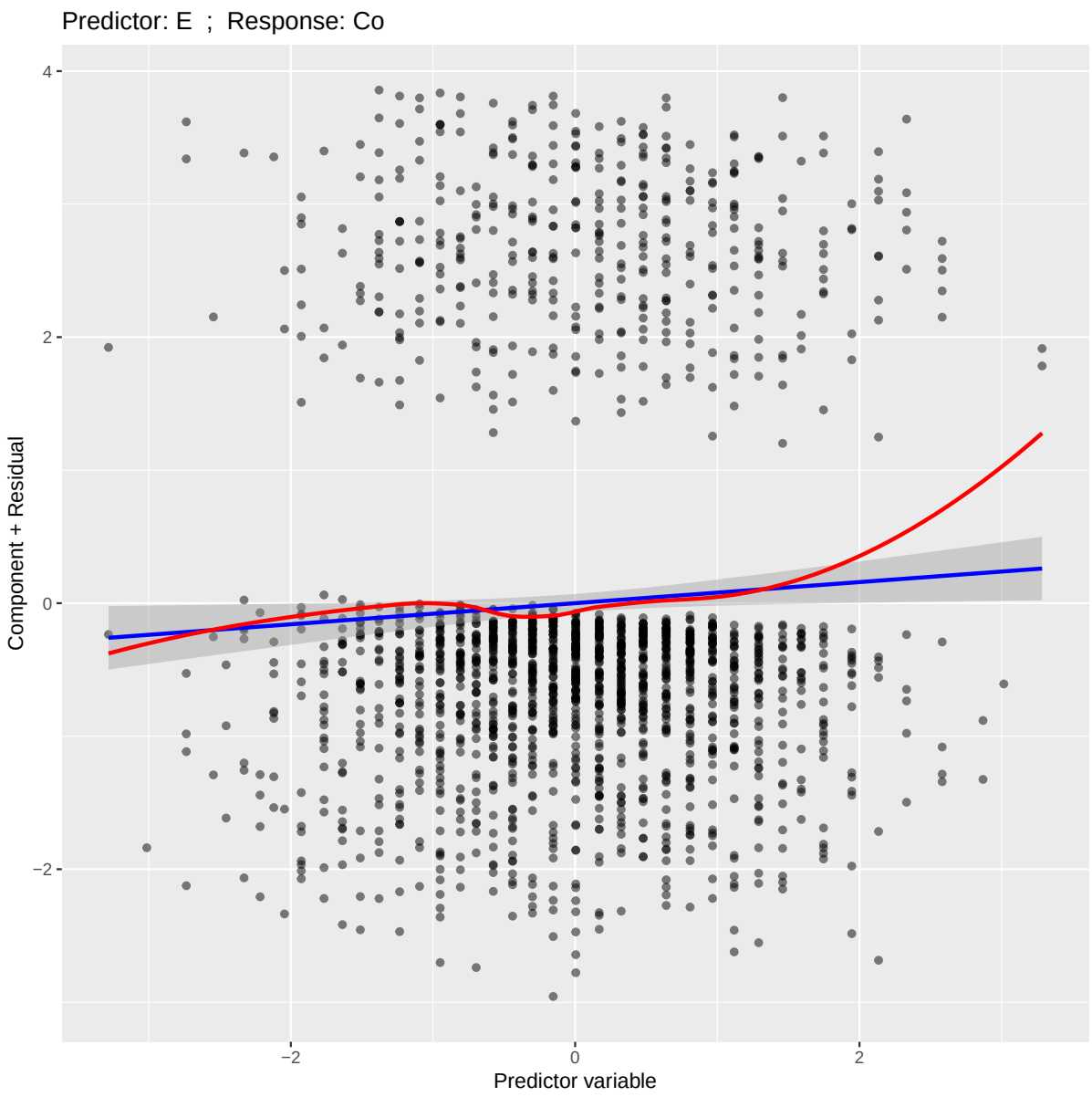


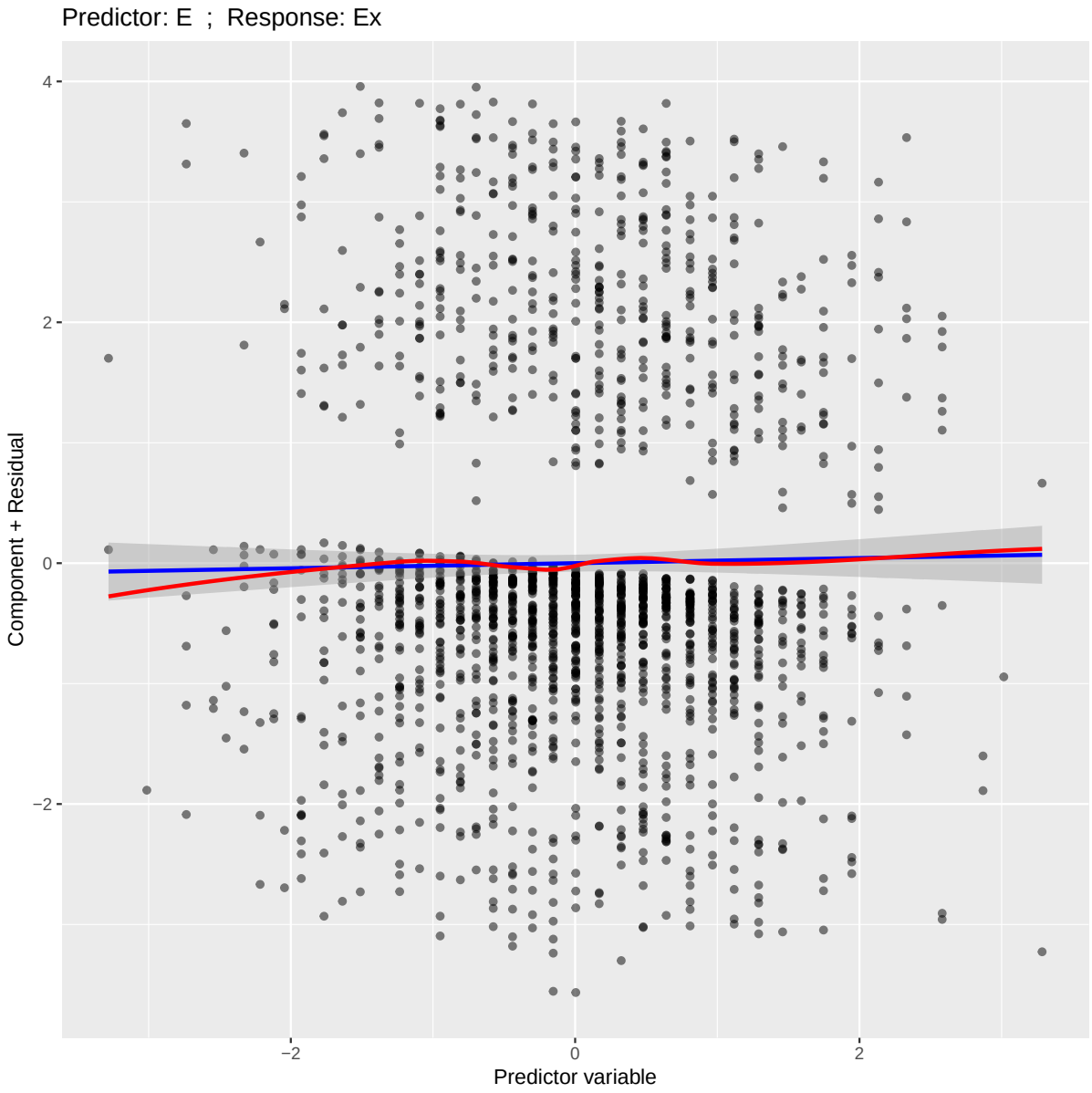
Fourth predictor

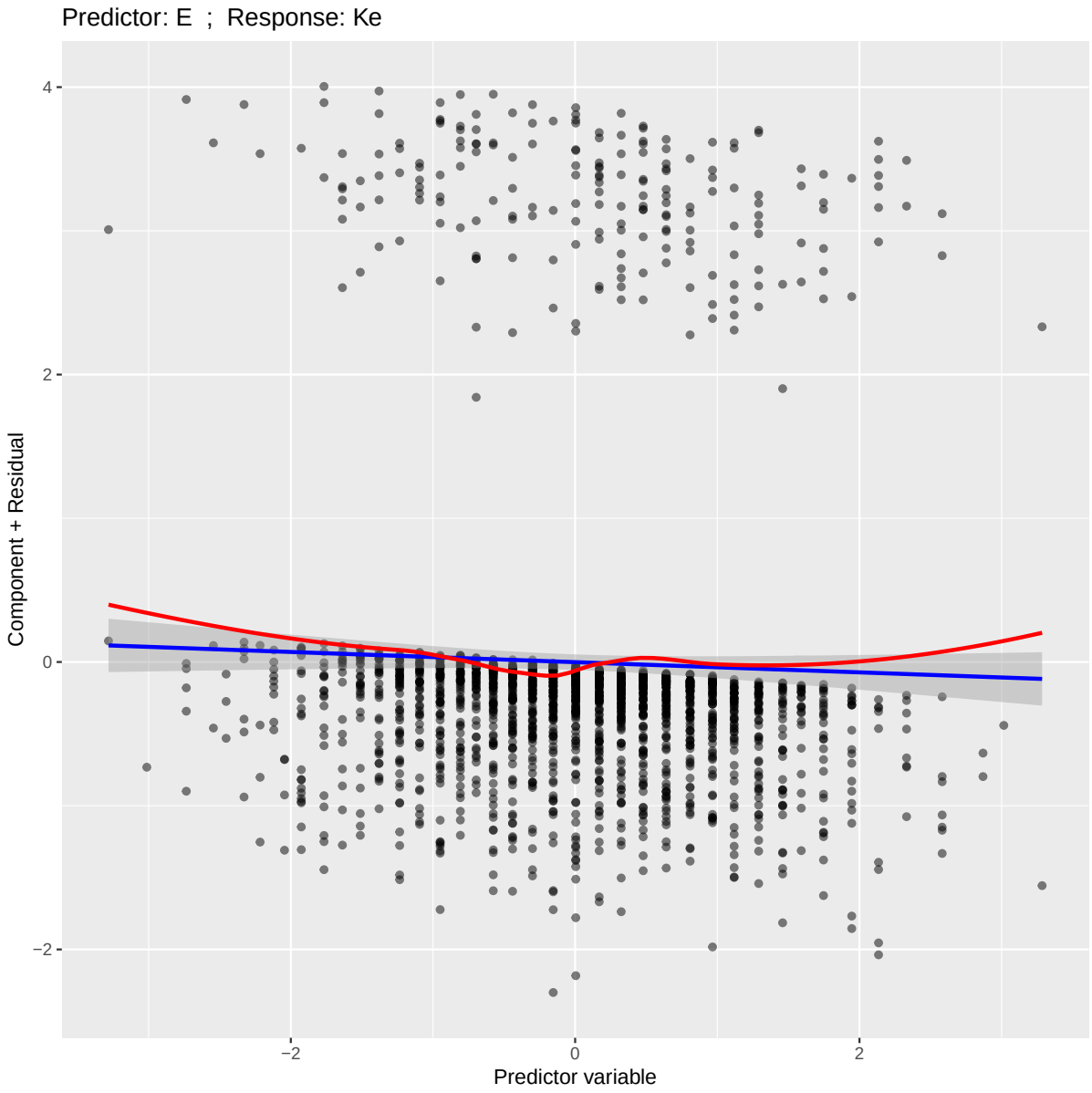

Predictor: E ; Response: Am
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

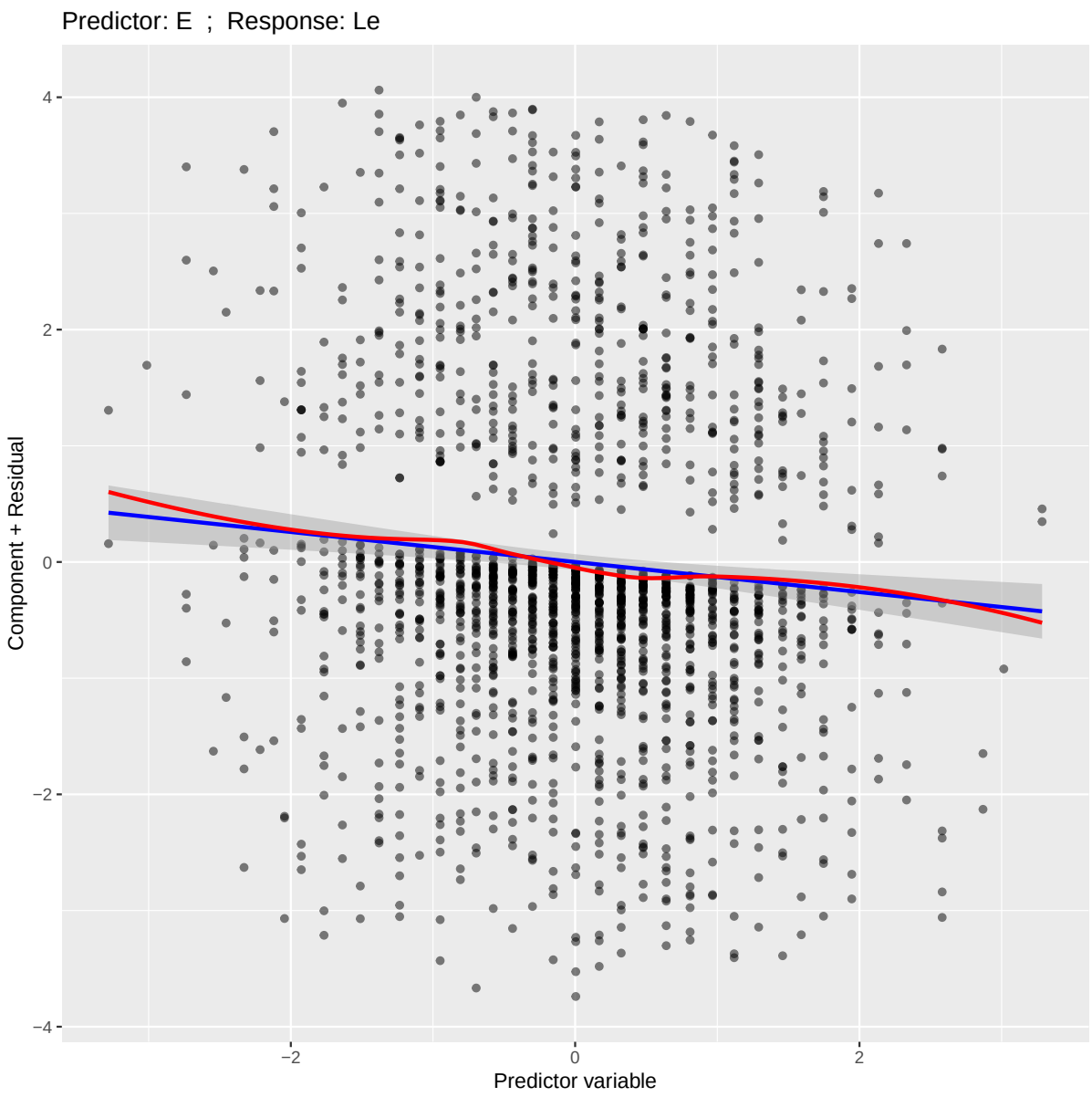

Predictor: E ; Response: Be
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

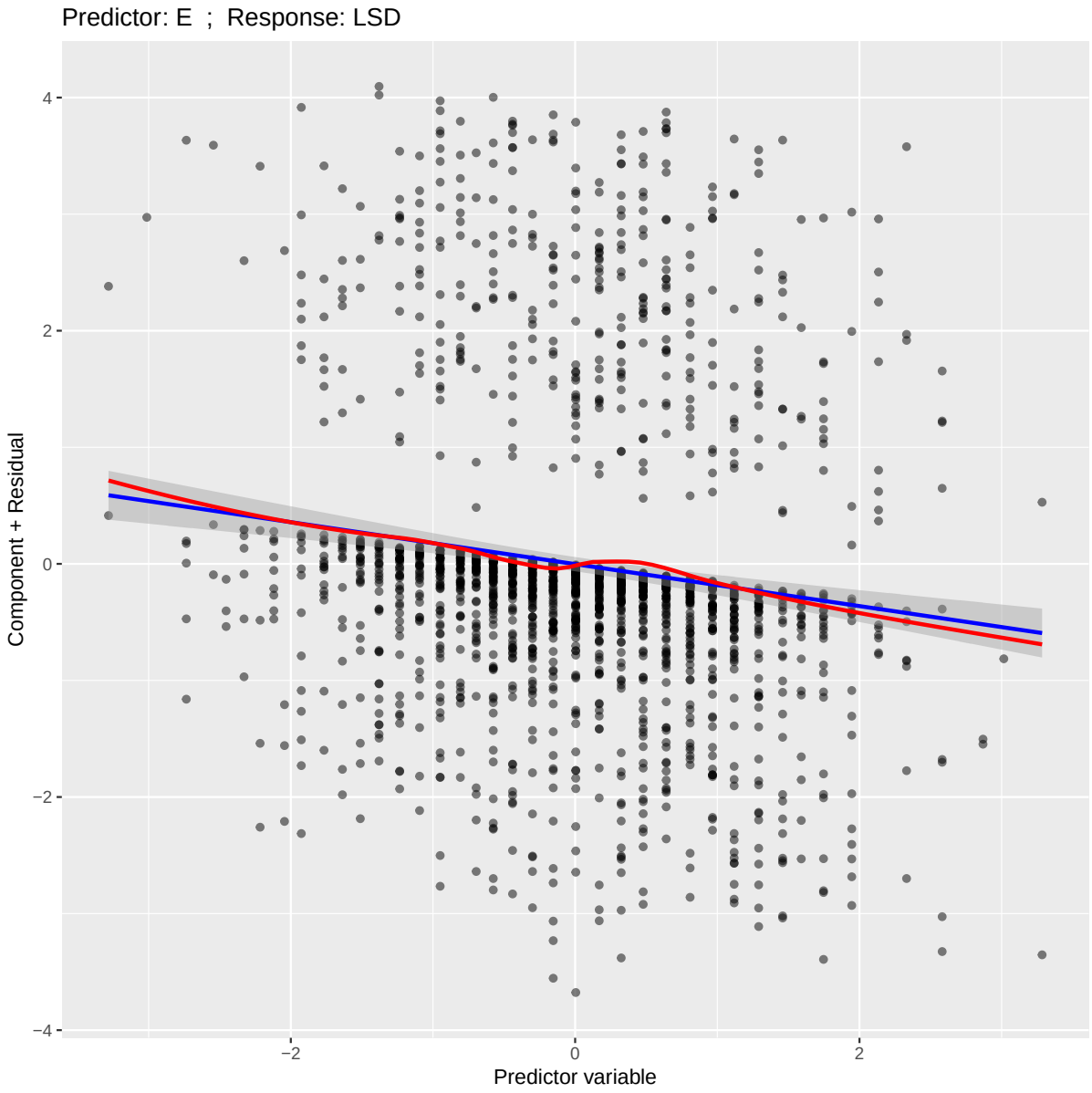

Predictor: E ; Response: Ca
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

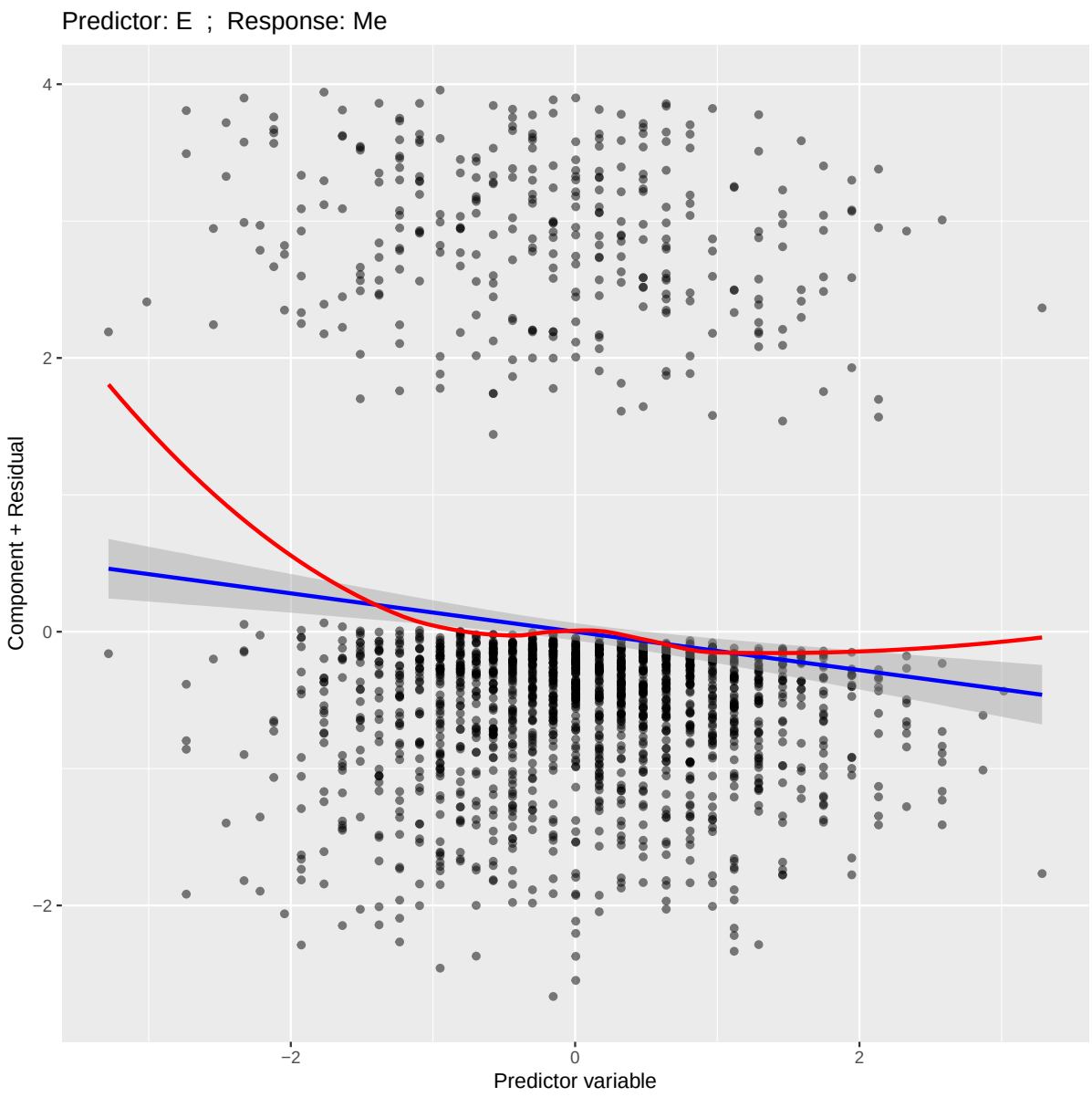

Predictor: E ; Response: Co
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

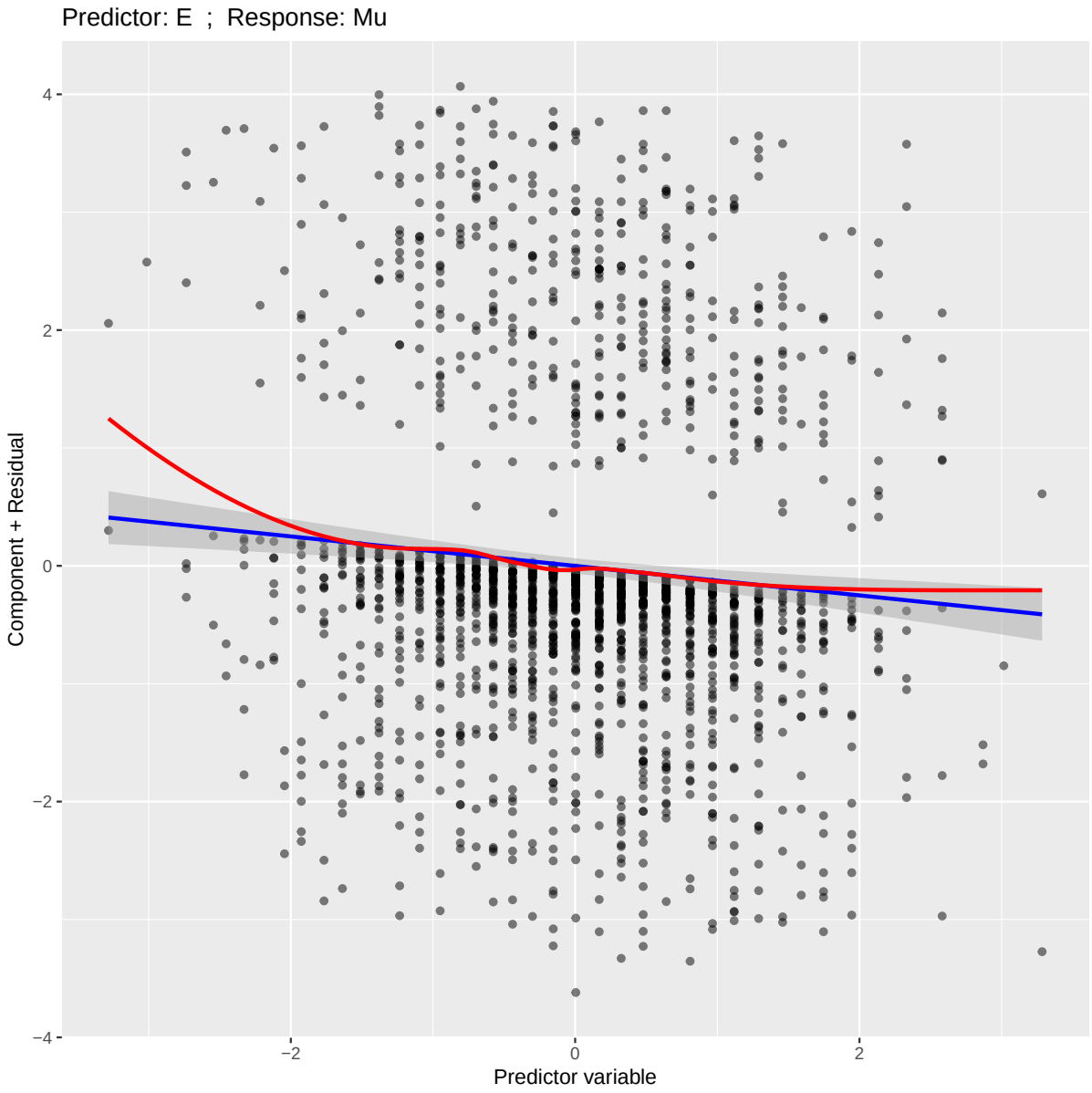

Predictor: E ; Response: Ex
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

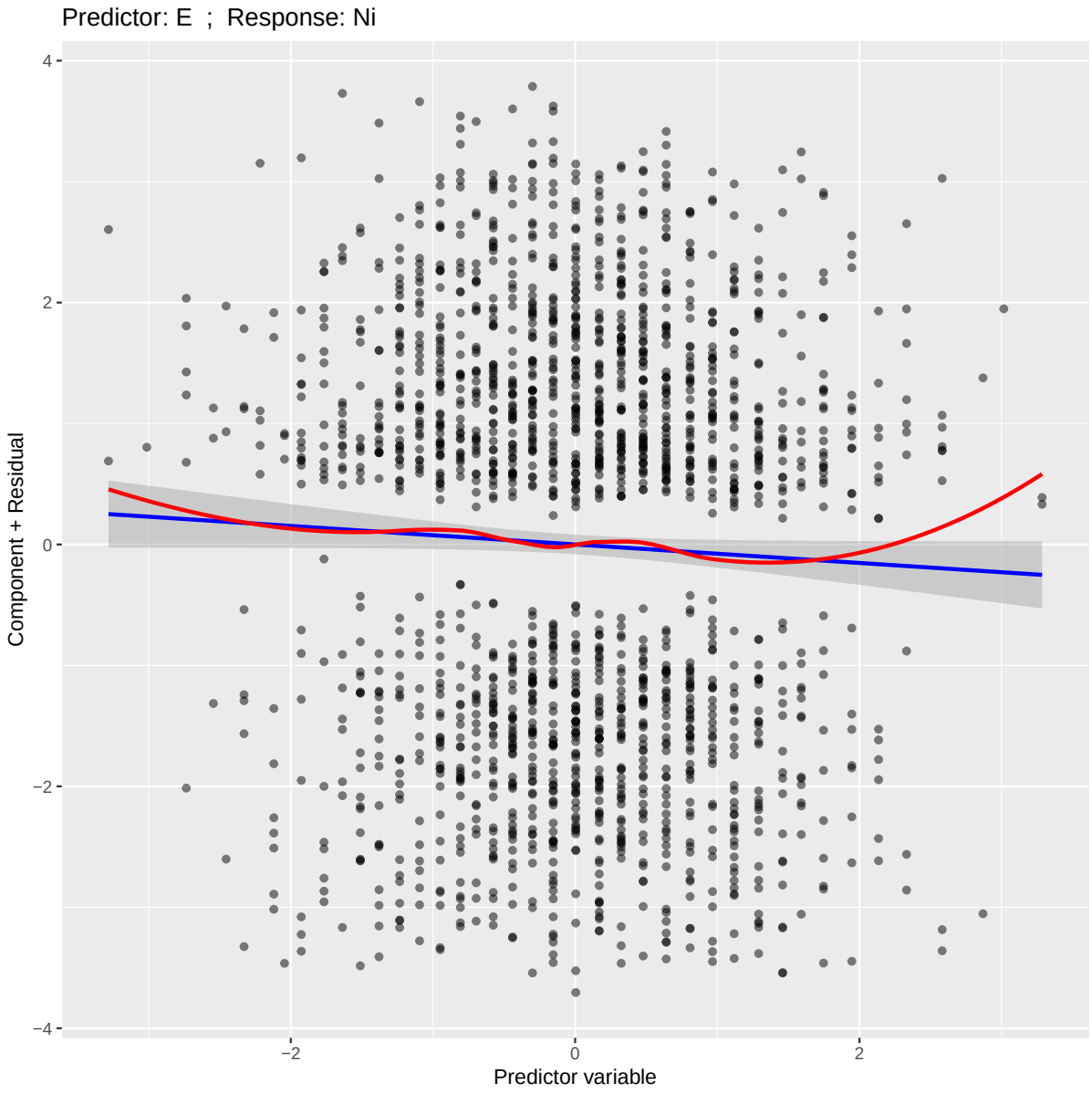

Predictor: E ; Response: Ke
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2


Predictor: E ; Response: Le
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

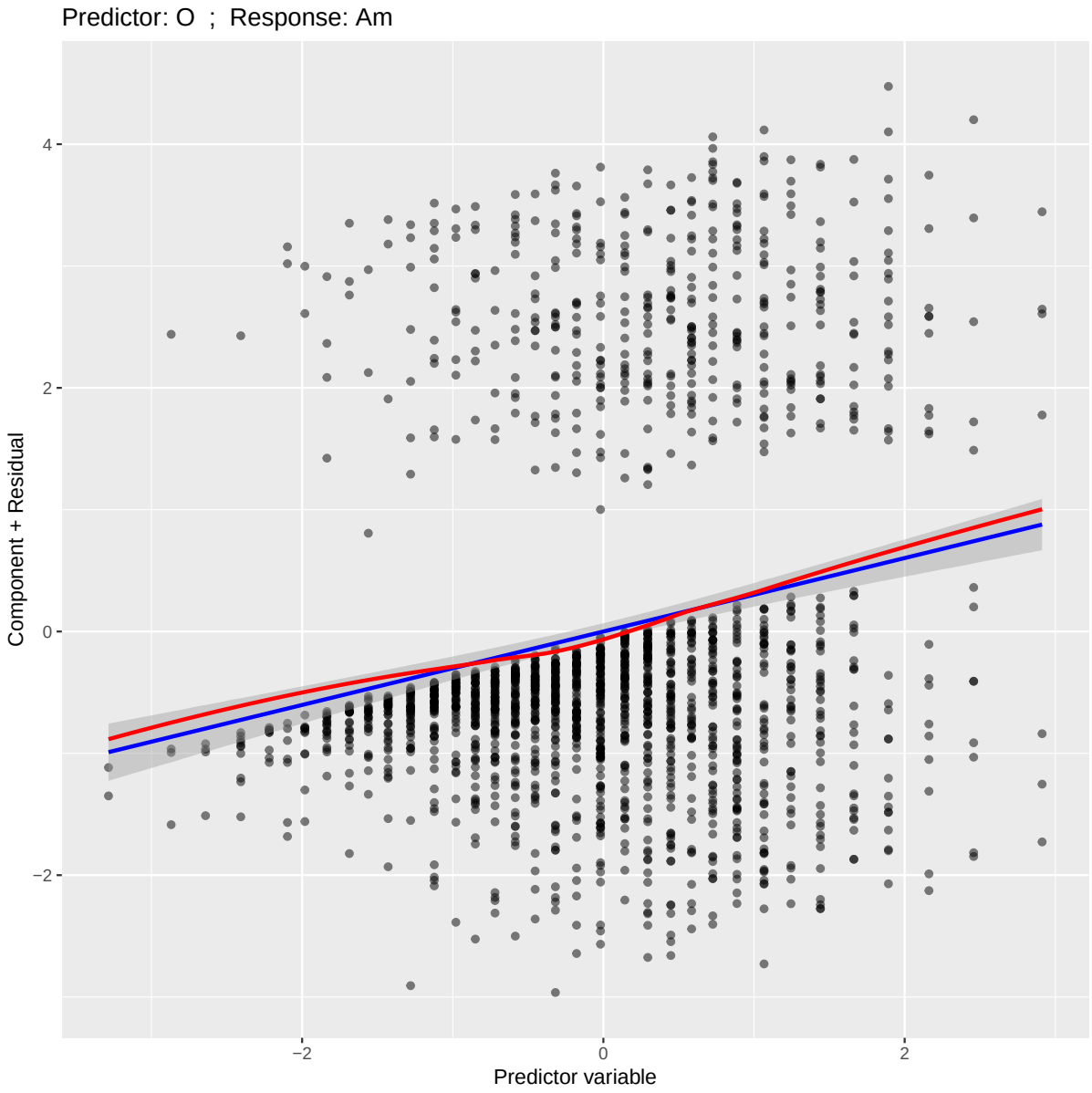

Predictor: E ; Response: LSD
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

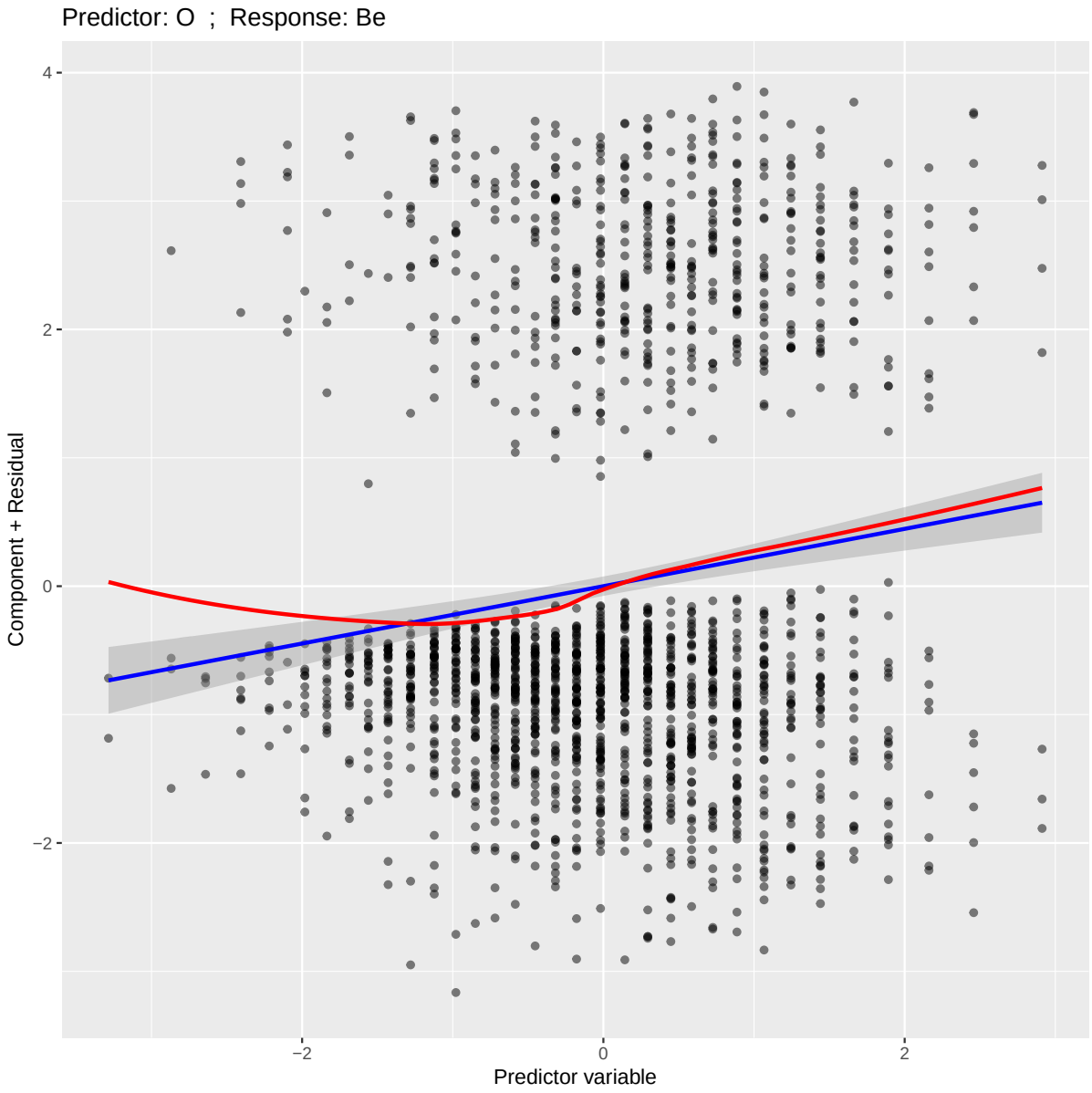

Predictor: E ; Response: Me
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

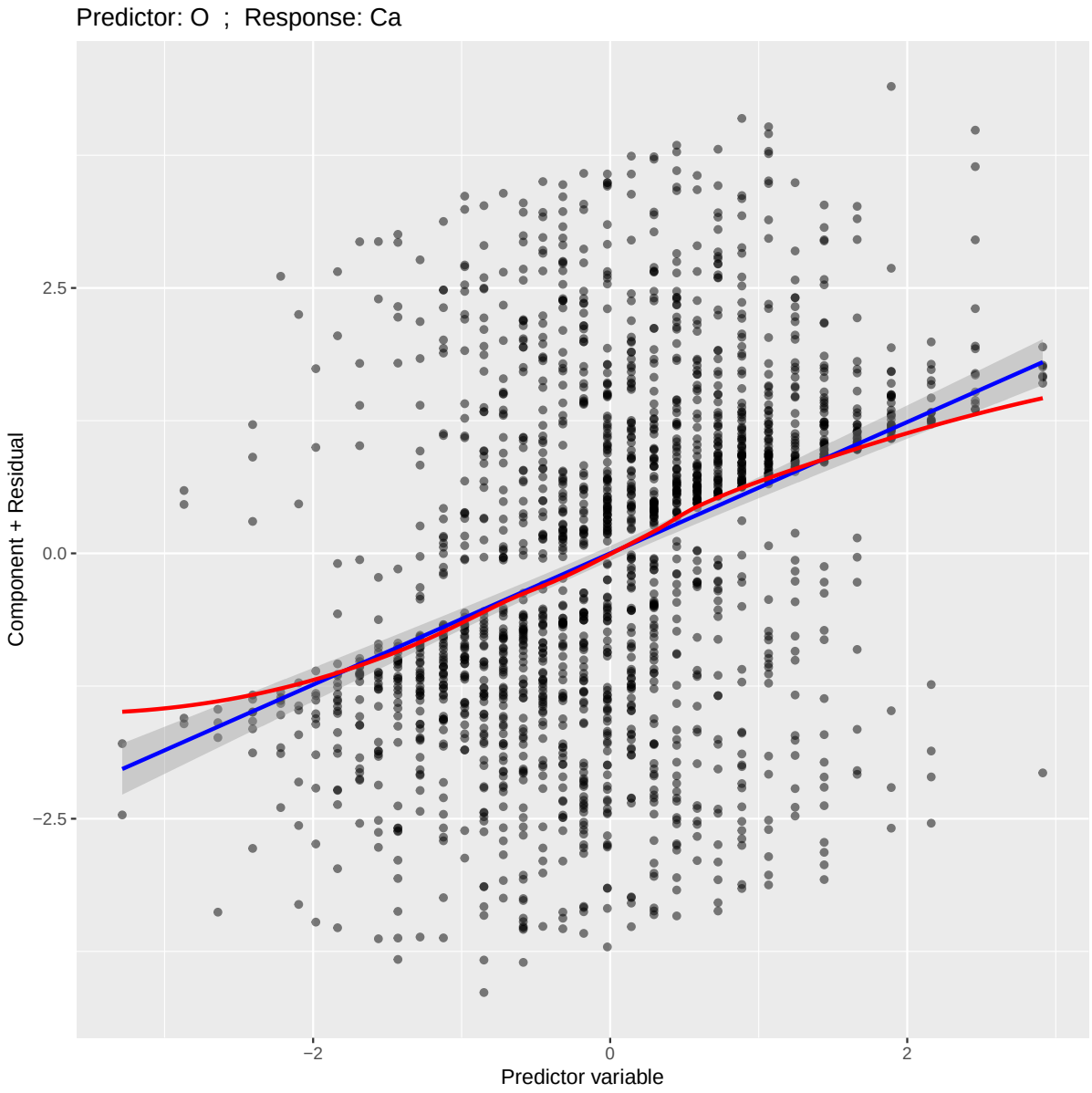

Predictor: E ; Response: Mu
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

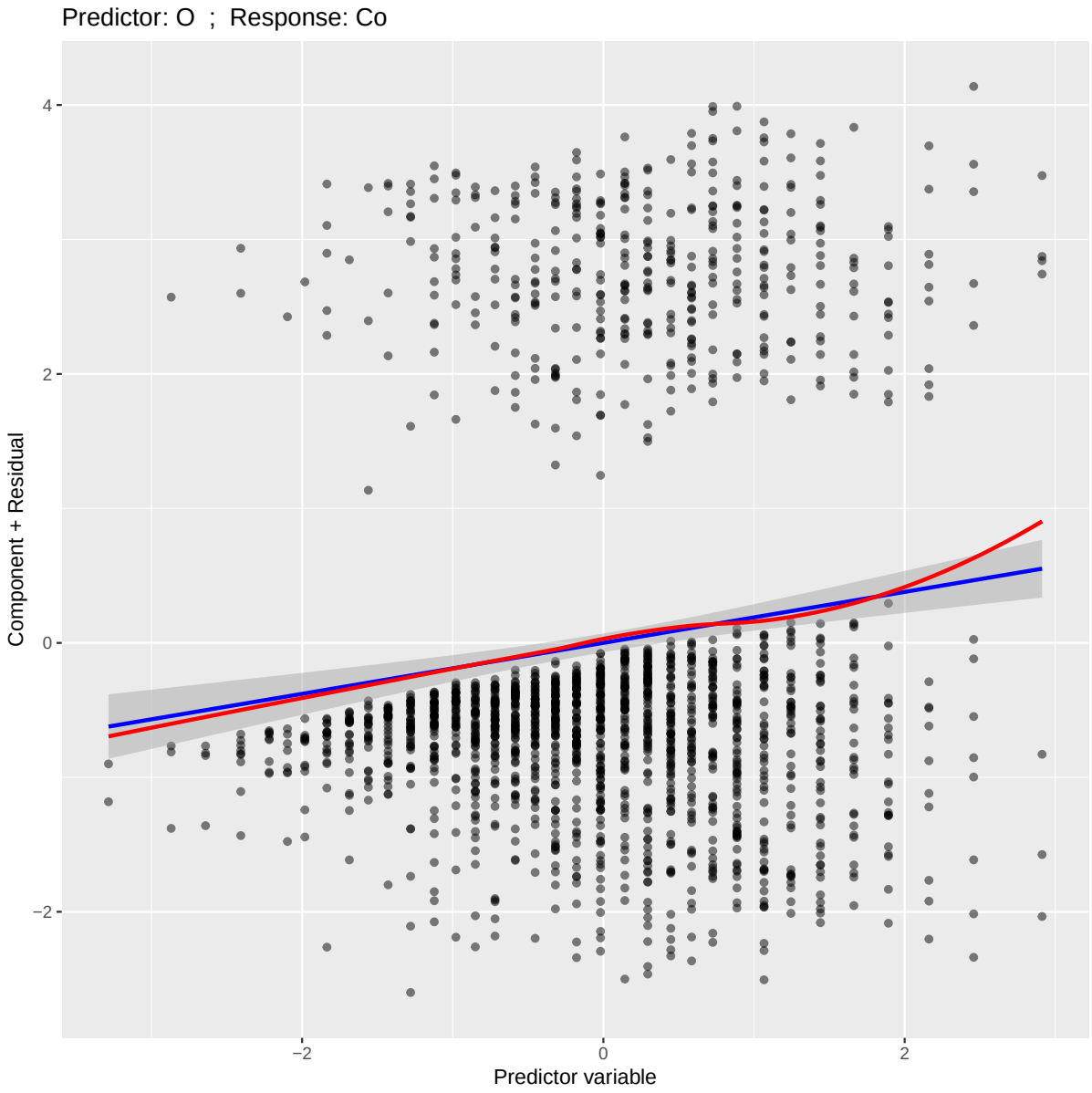


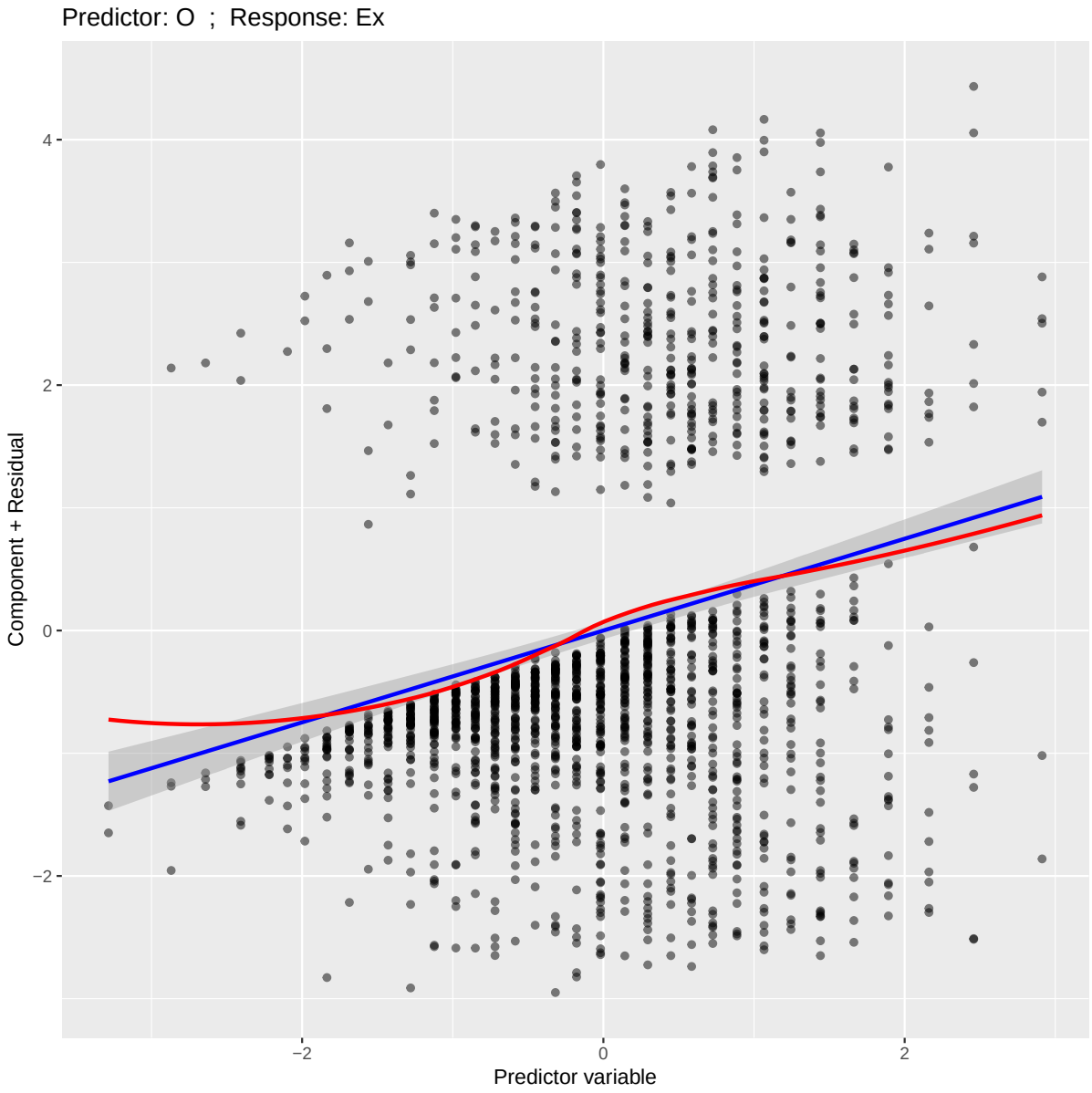
Fifth predictor

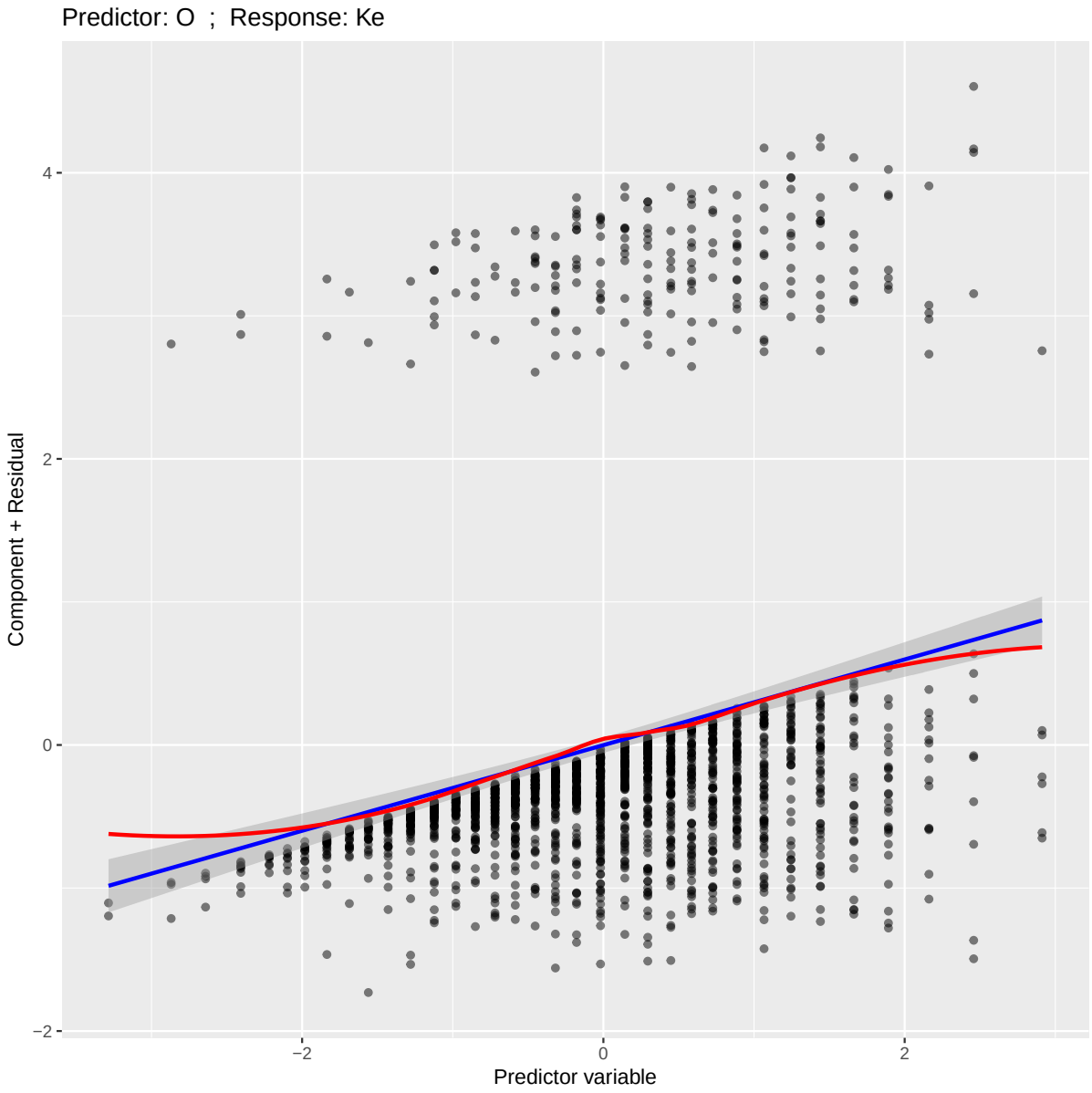

Predictor: O ; Response: Am
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

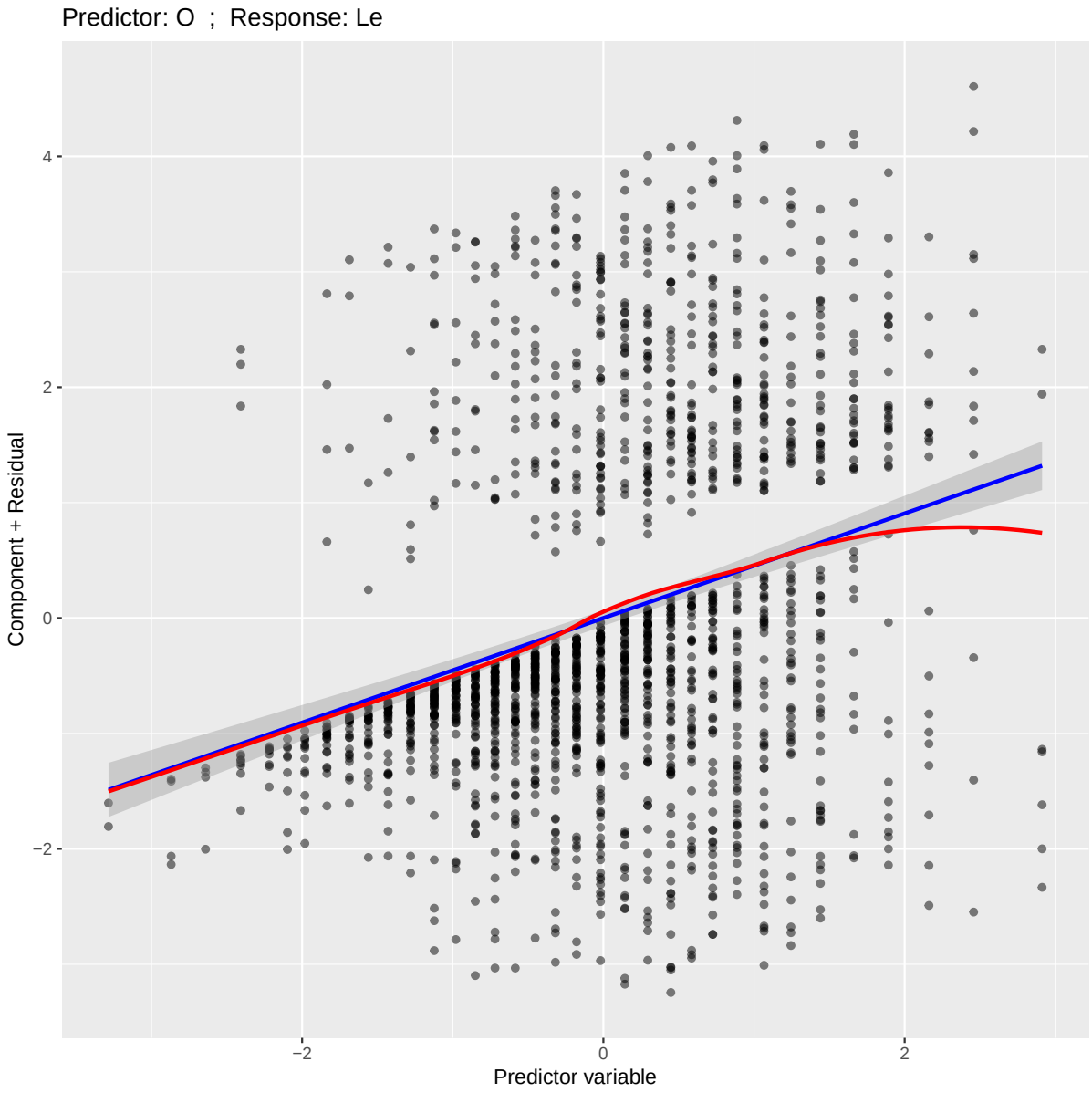

Predictor: O ; Response: Be
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

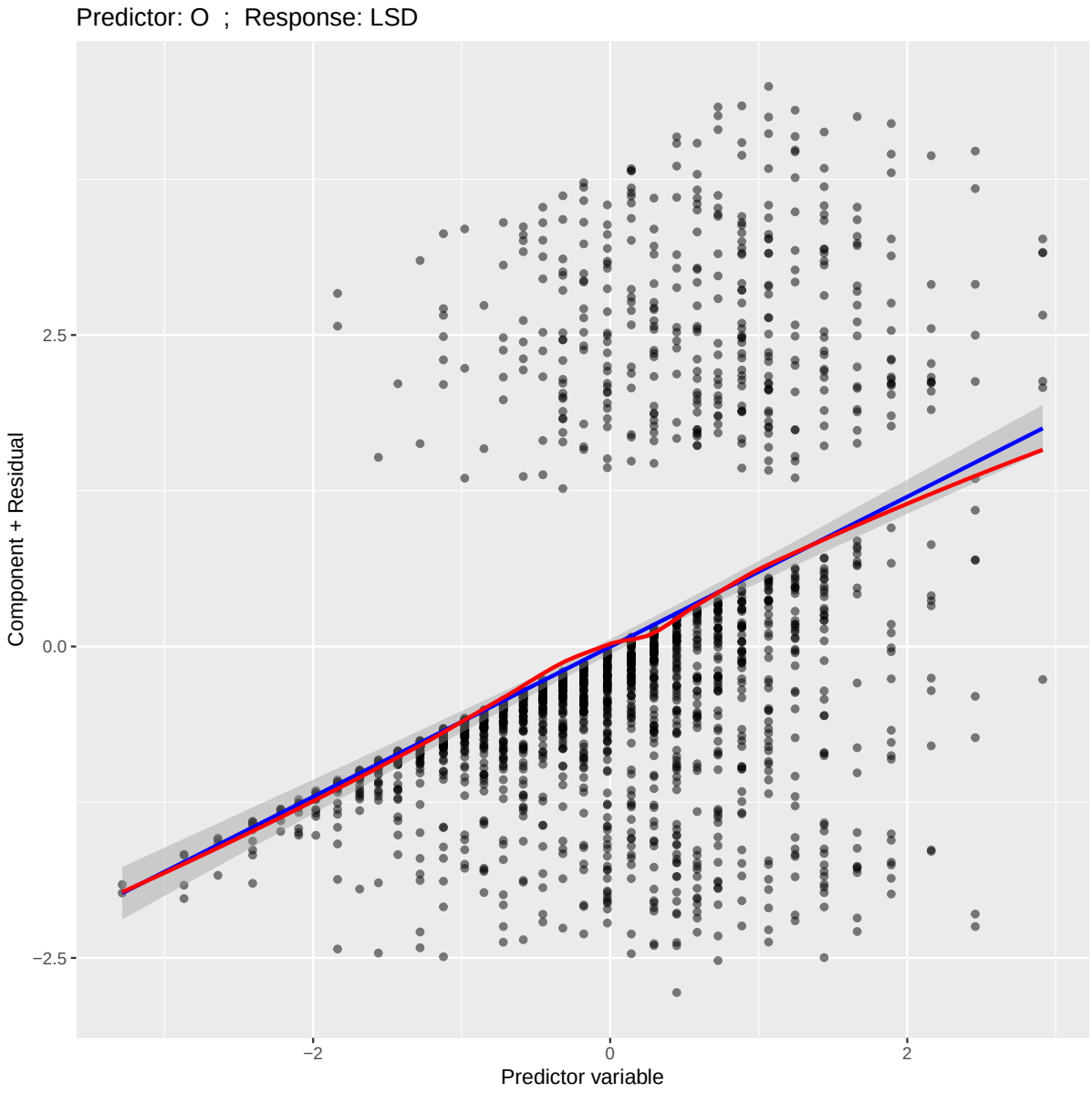

Predictor: O ; Response: Ca
Component + Residual
2.5
0.0
−2.5
−2
0
2
Predictor variable

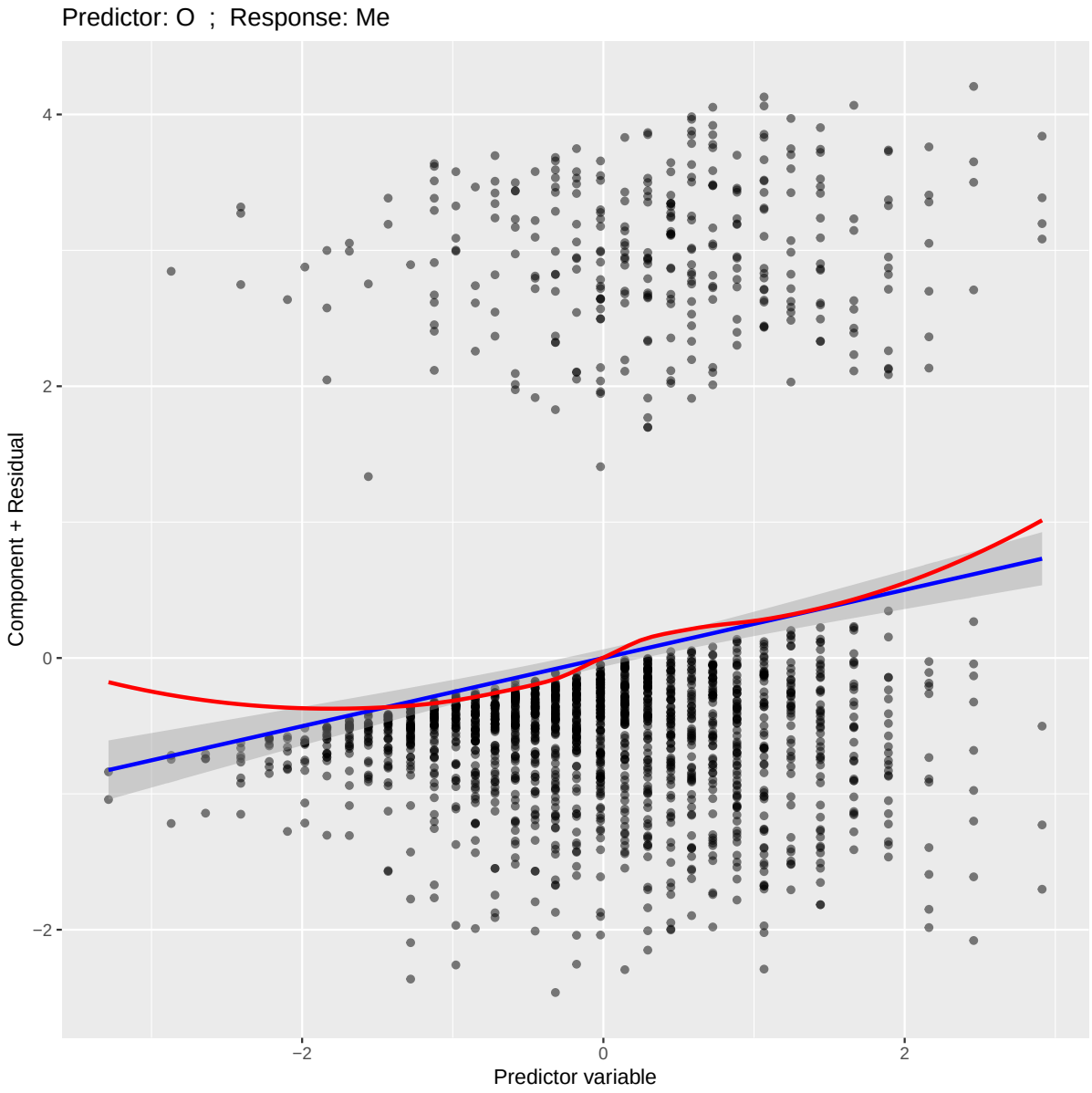

Predictor: O ; Response: Co
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

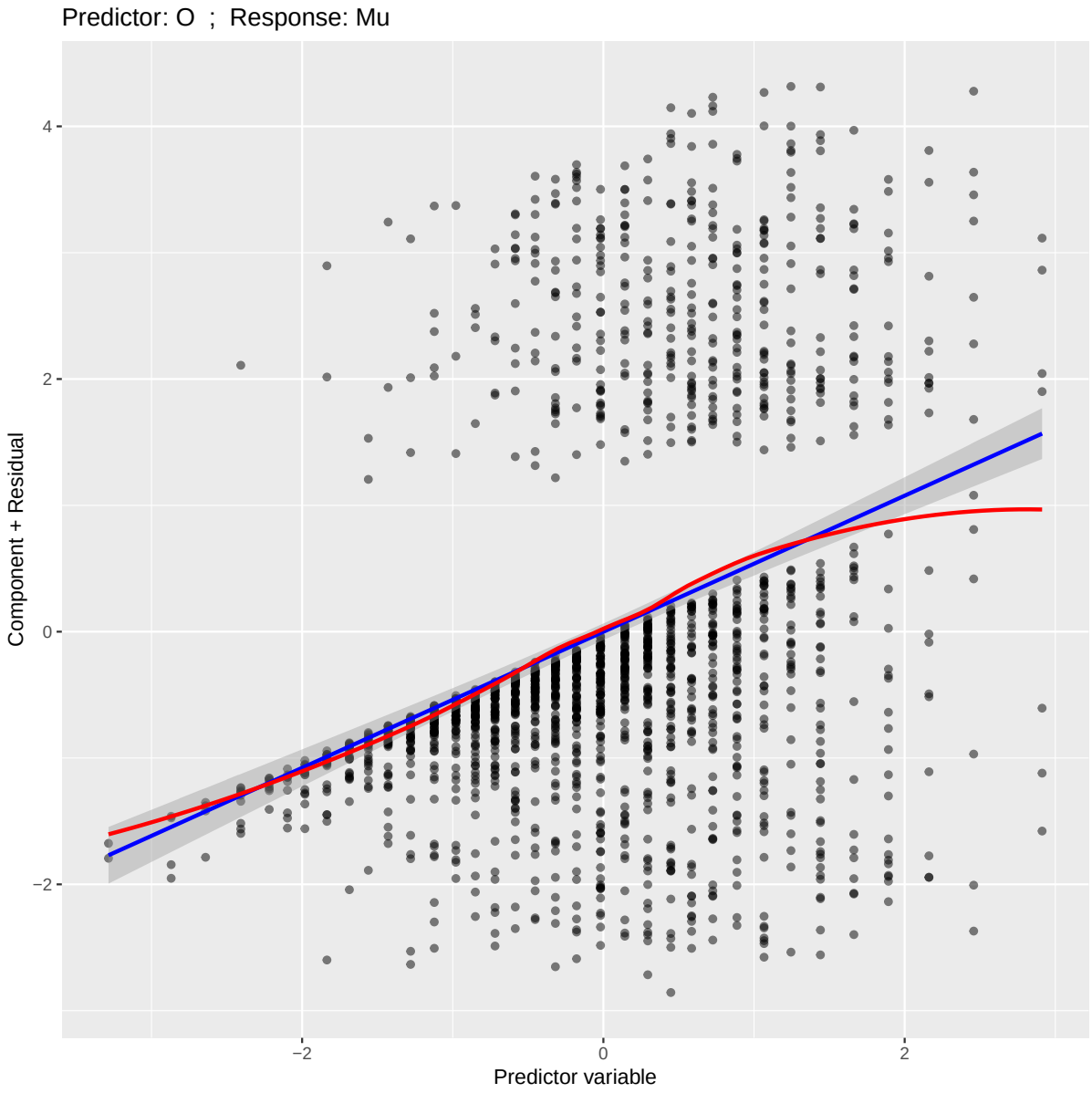

Predictor: O ; Response: Ex
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

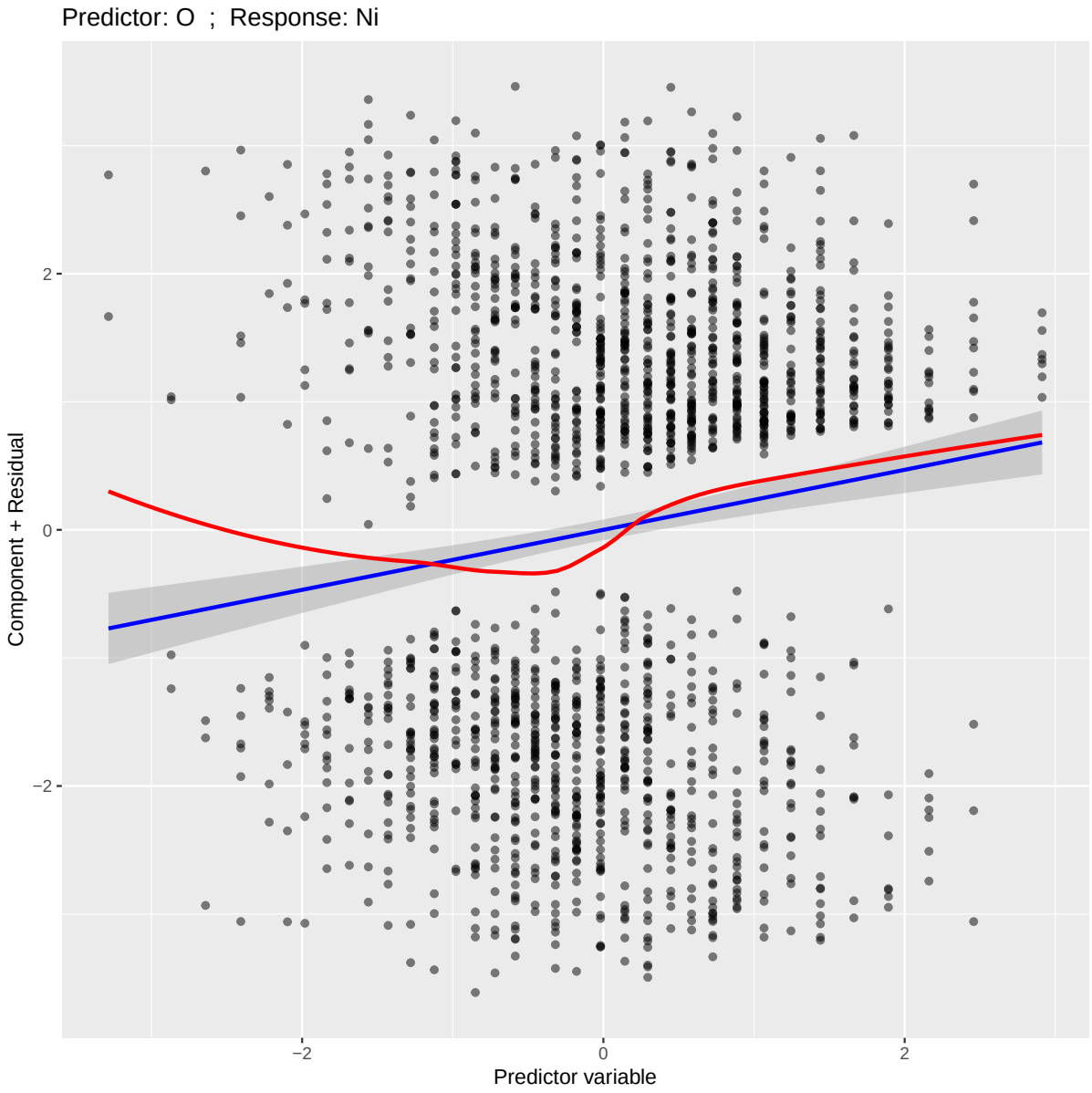

Predictor: O ; Response: Ke
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable


Predictor: O ; Response: Le
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

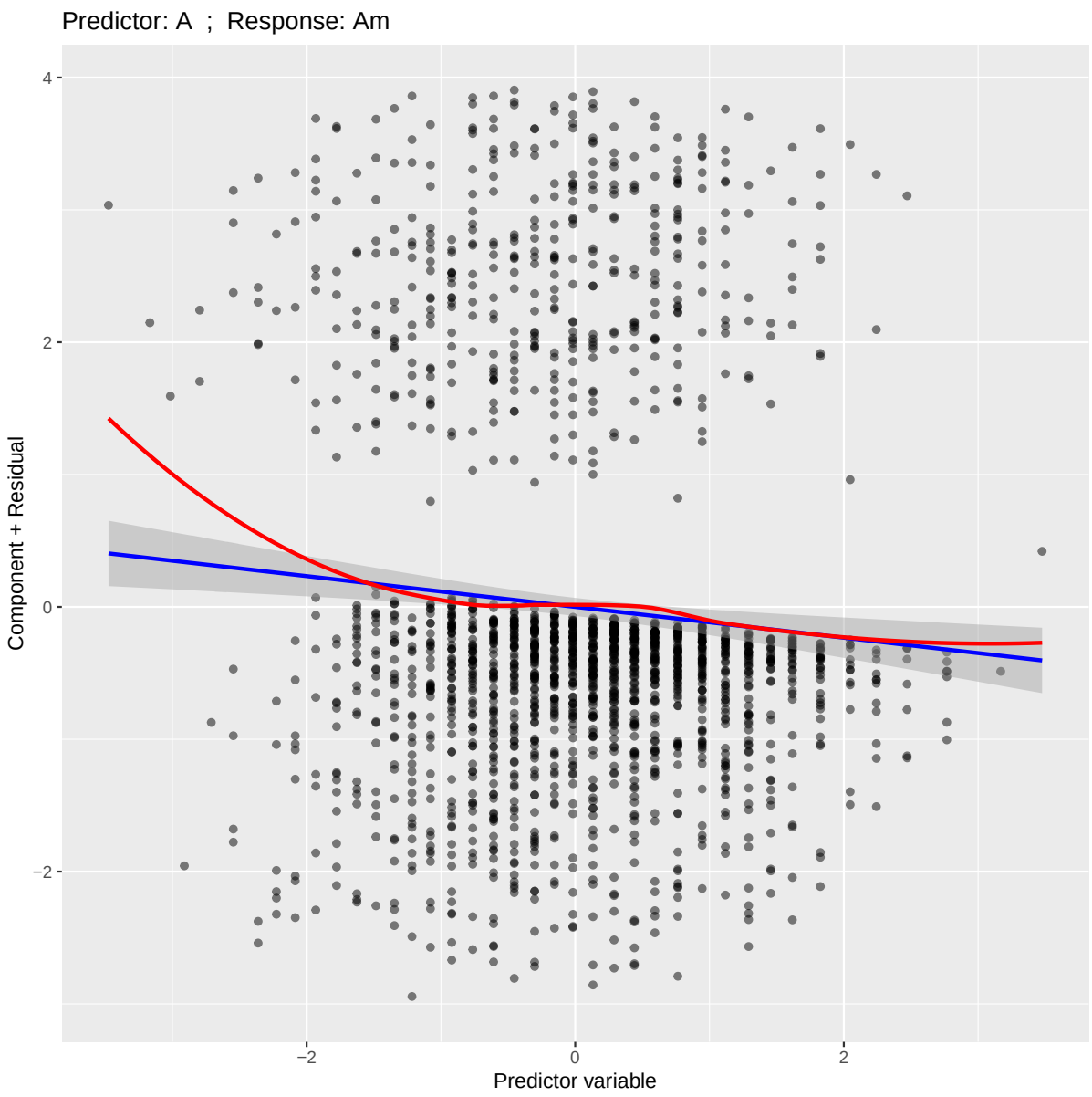

Predictor: O ; Response: LSD
Component + Residual
2.5
0.0
−2.5
−2
0
2
Predictor variable

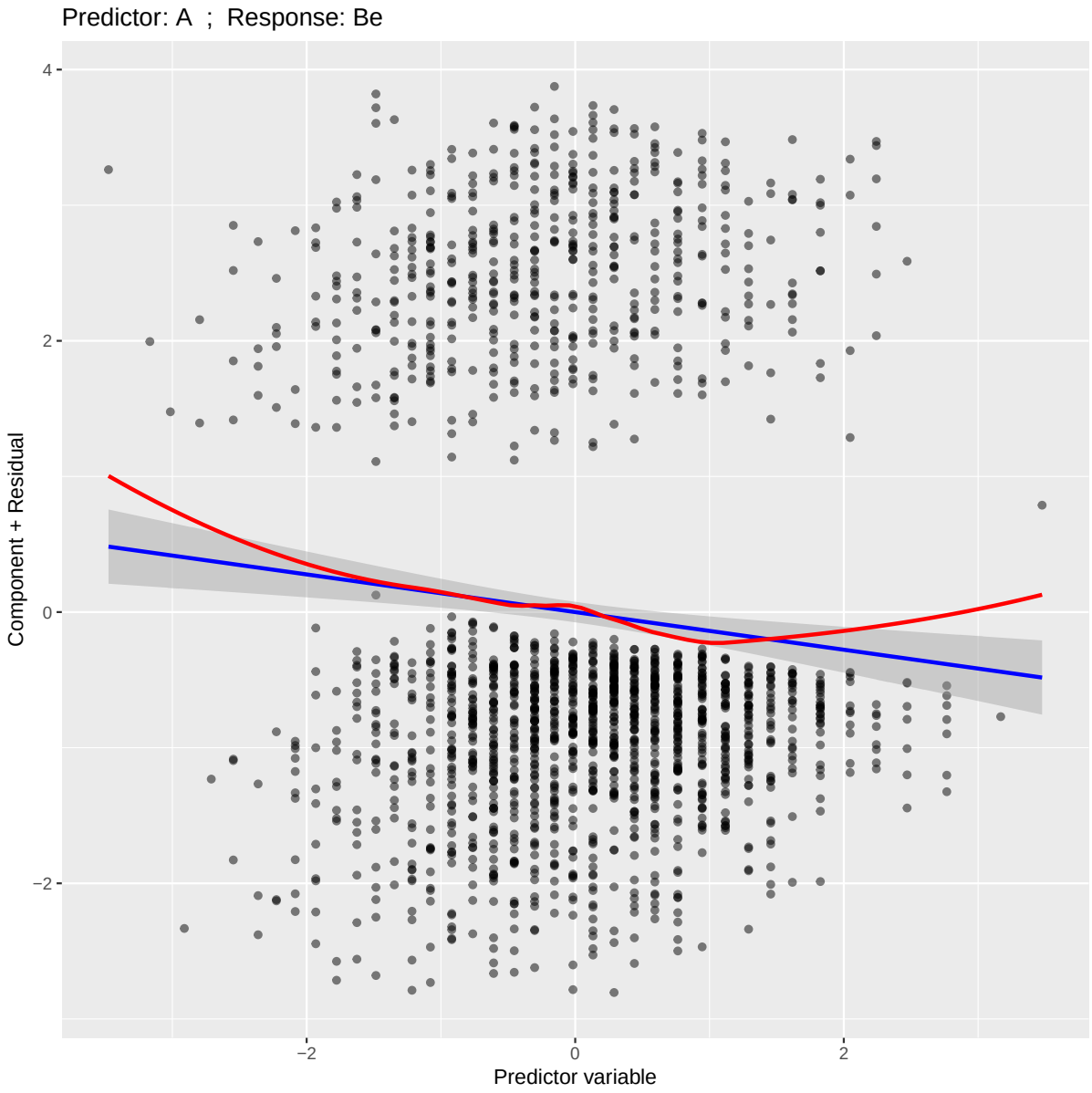

Predictor: O ; Response: Me
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

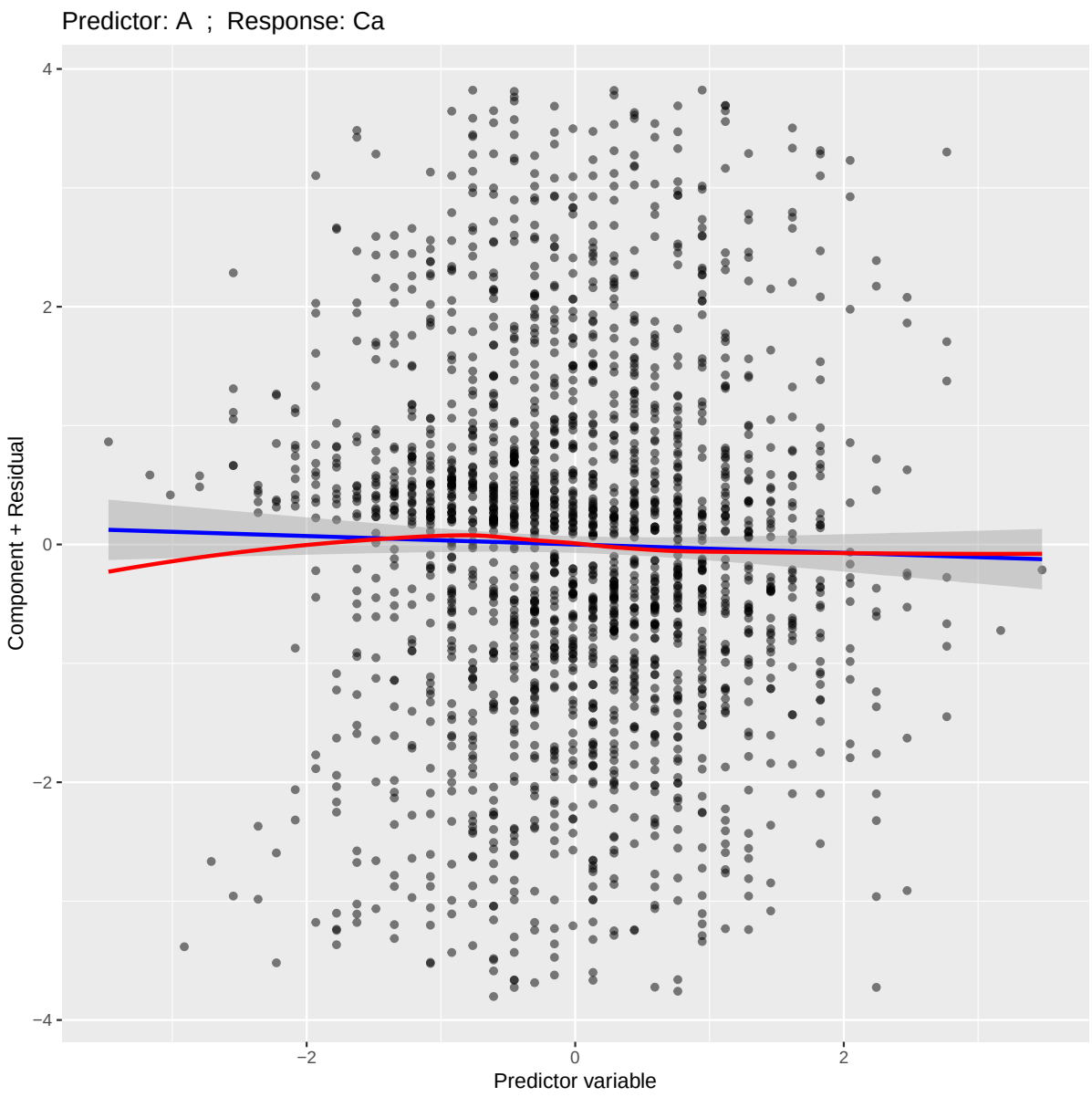

Predictor: O ; Response: Mu
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

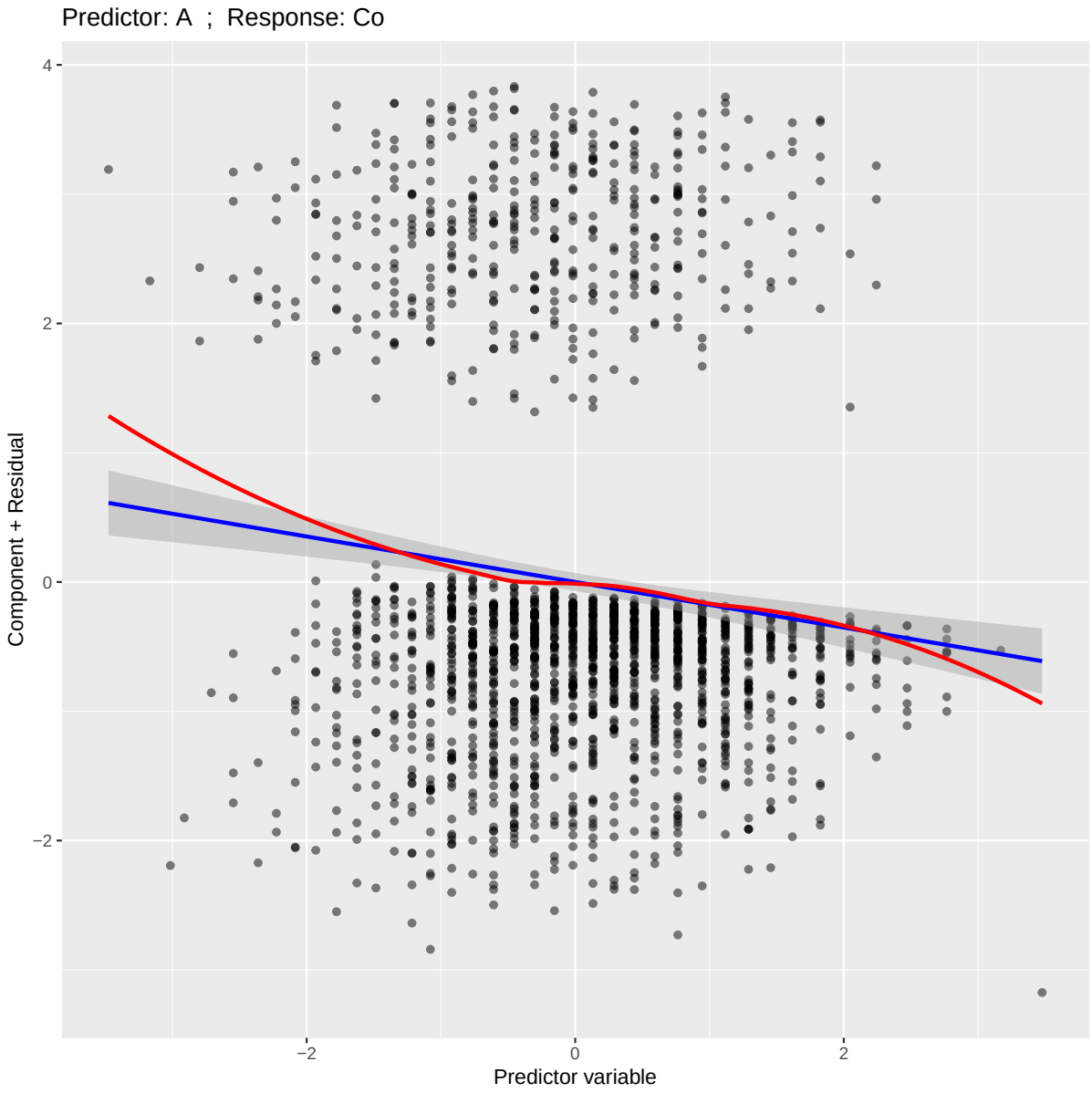


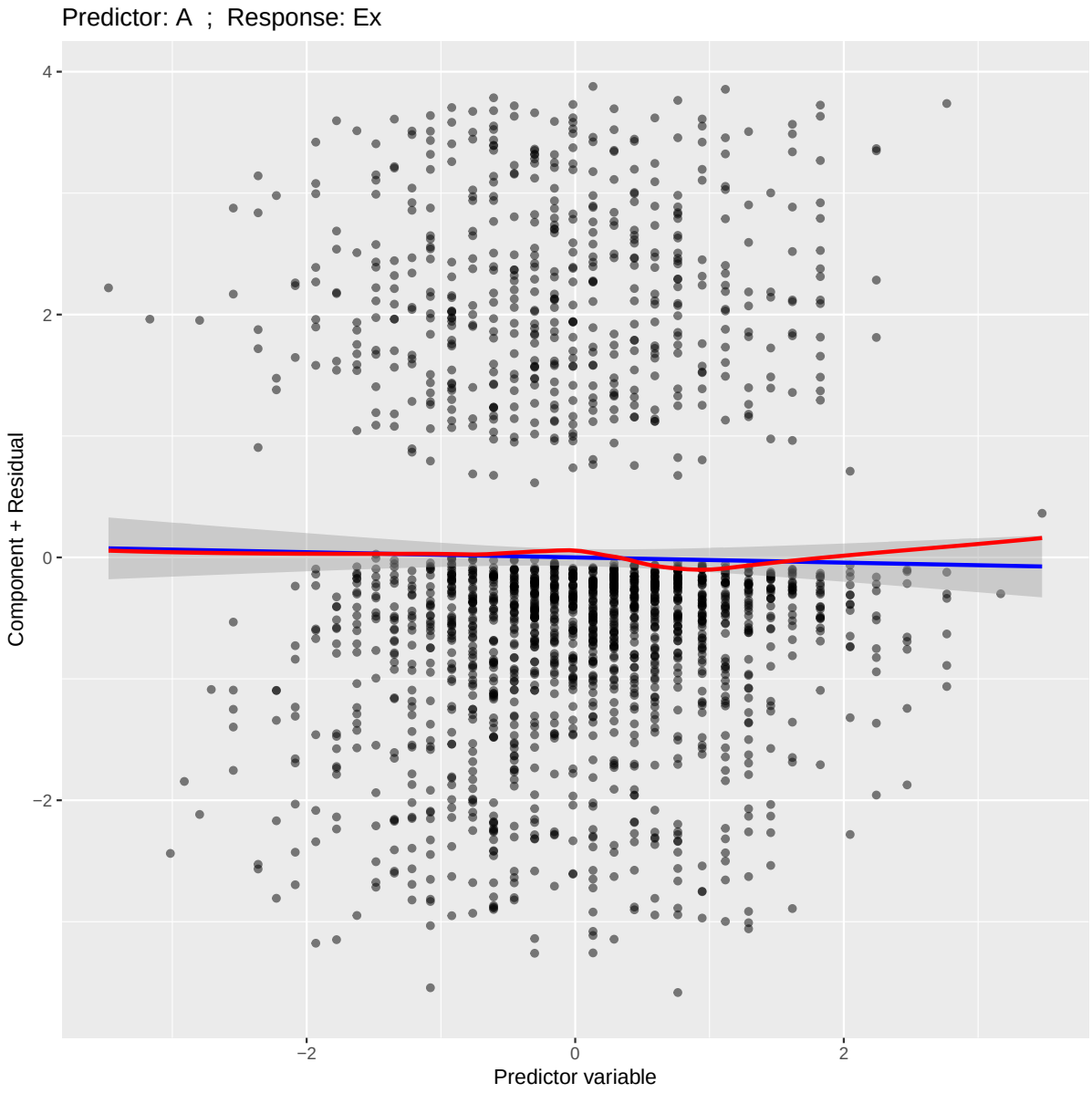
Sixth predictor

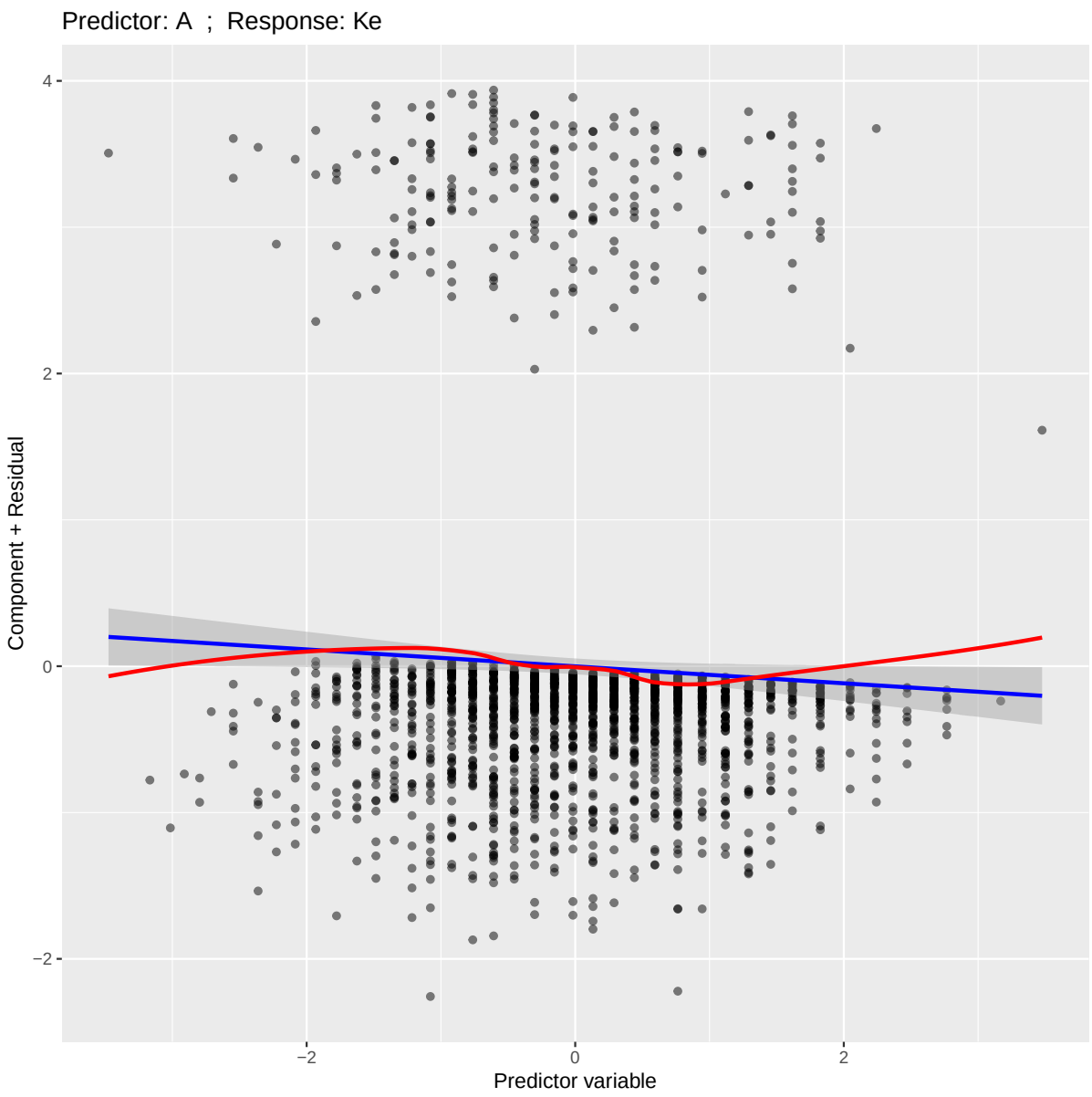

Predictor: A ; Response: Am
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

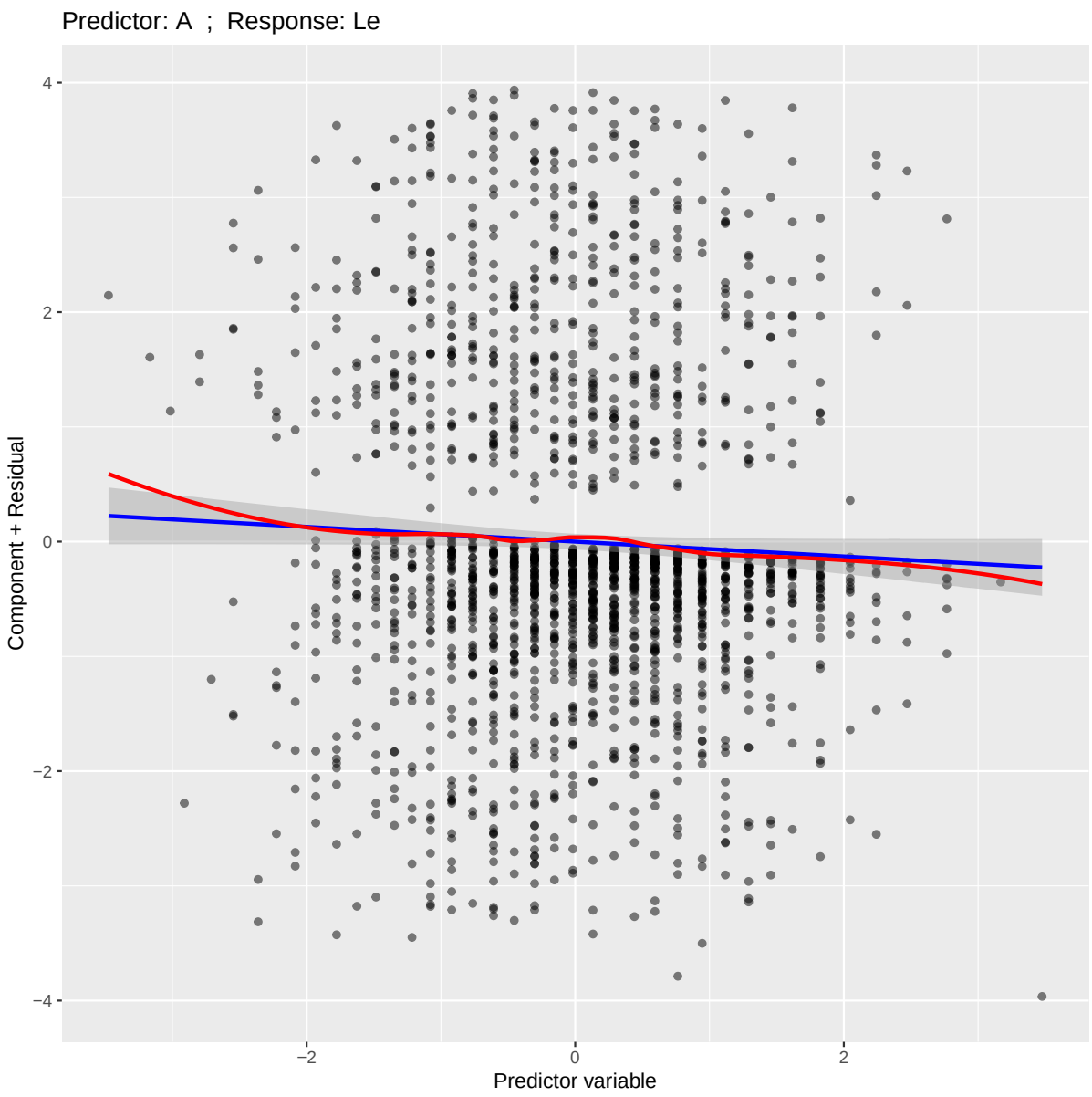

Predictor: A ; Response: Be
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

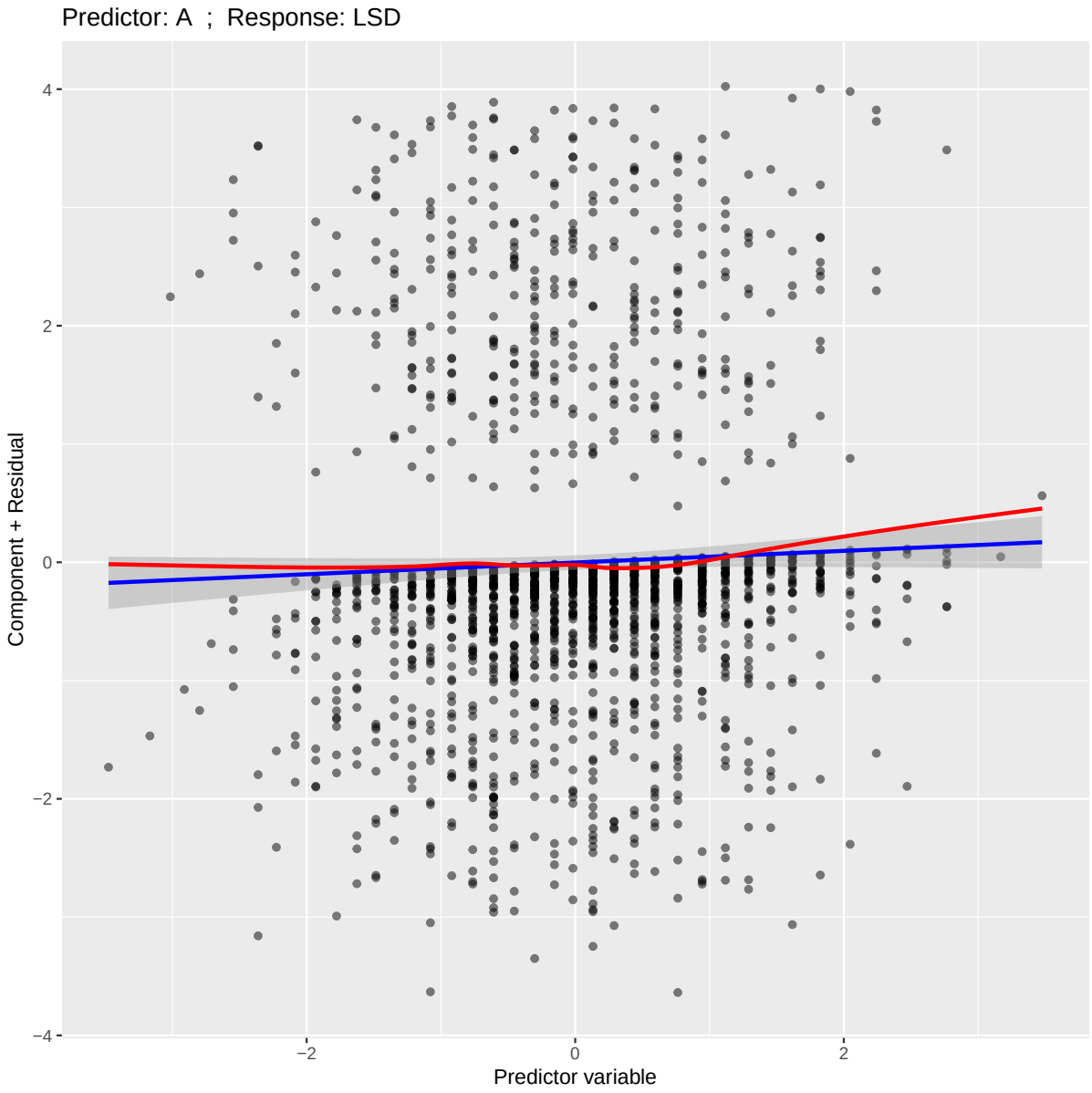

Predictor: A ; Response: Ca
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

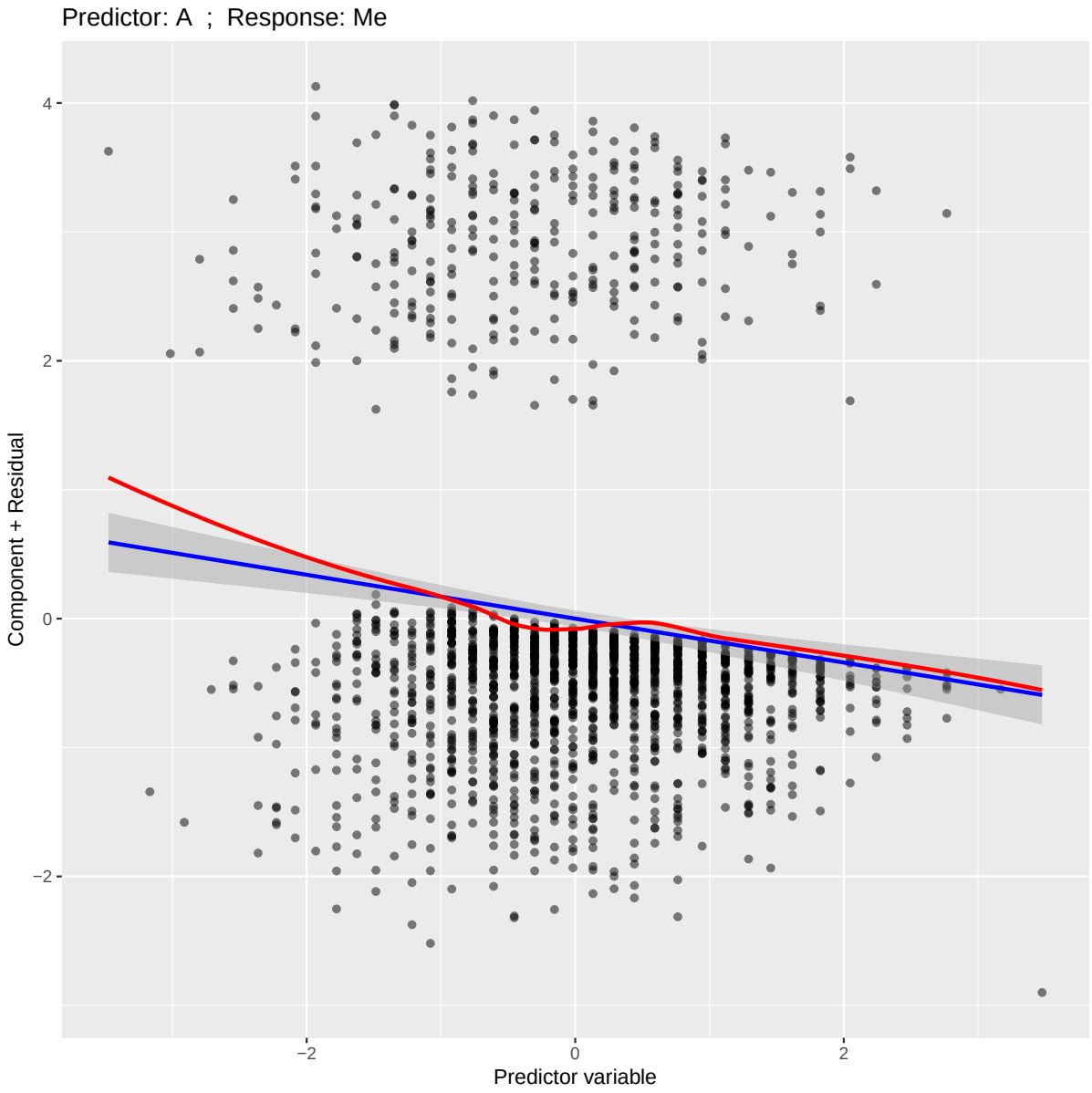

Predictor: A ; Response: Co
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

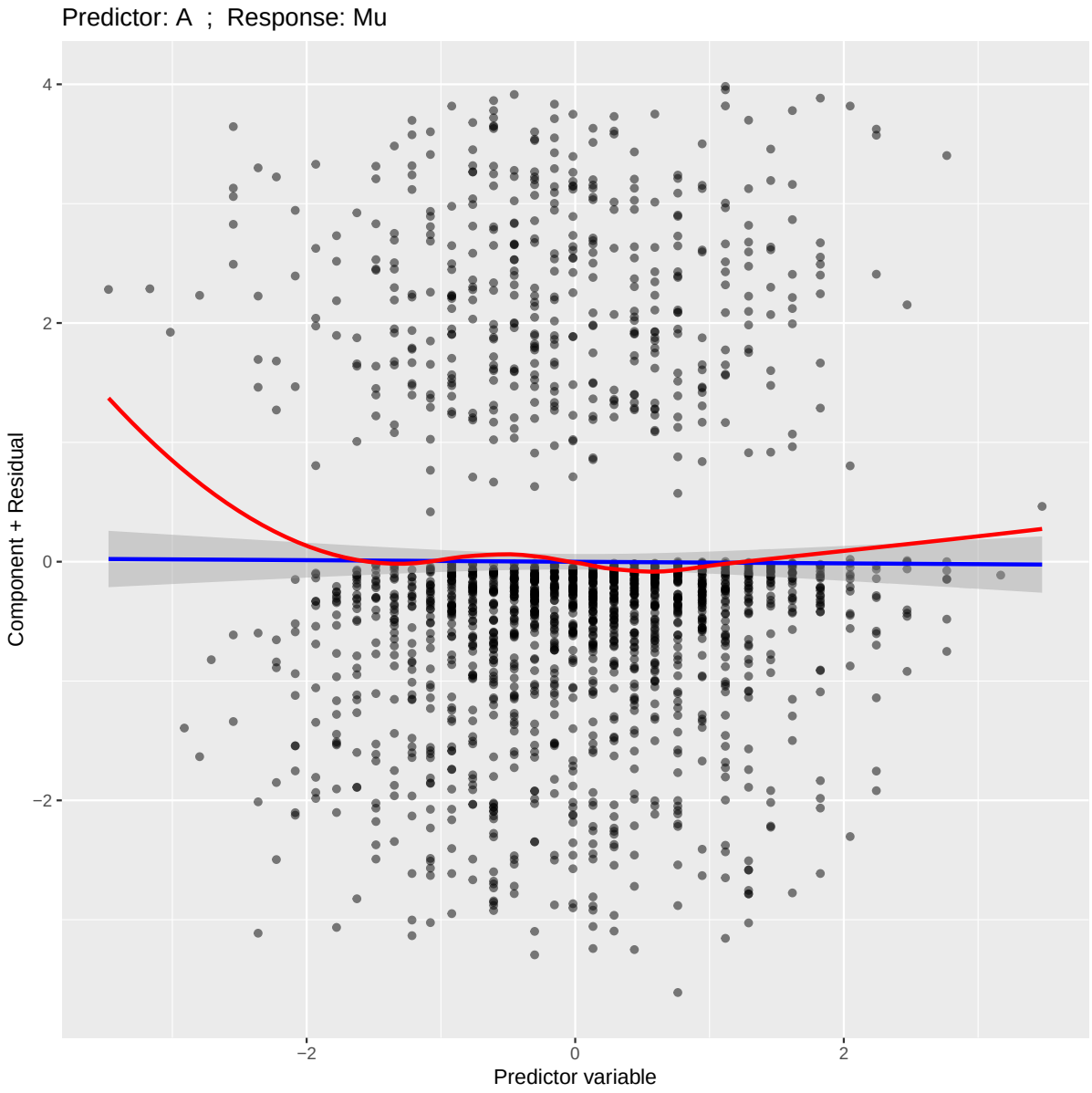

Predictor: A ; Response: Ex
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

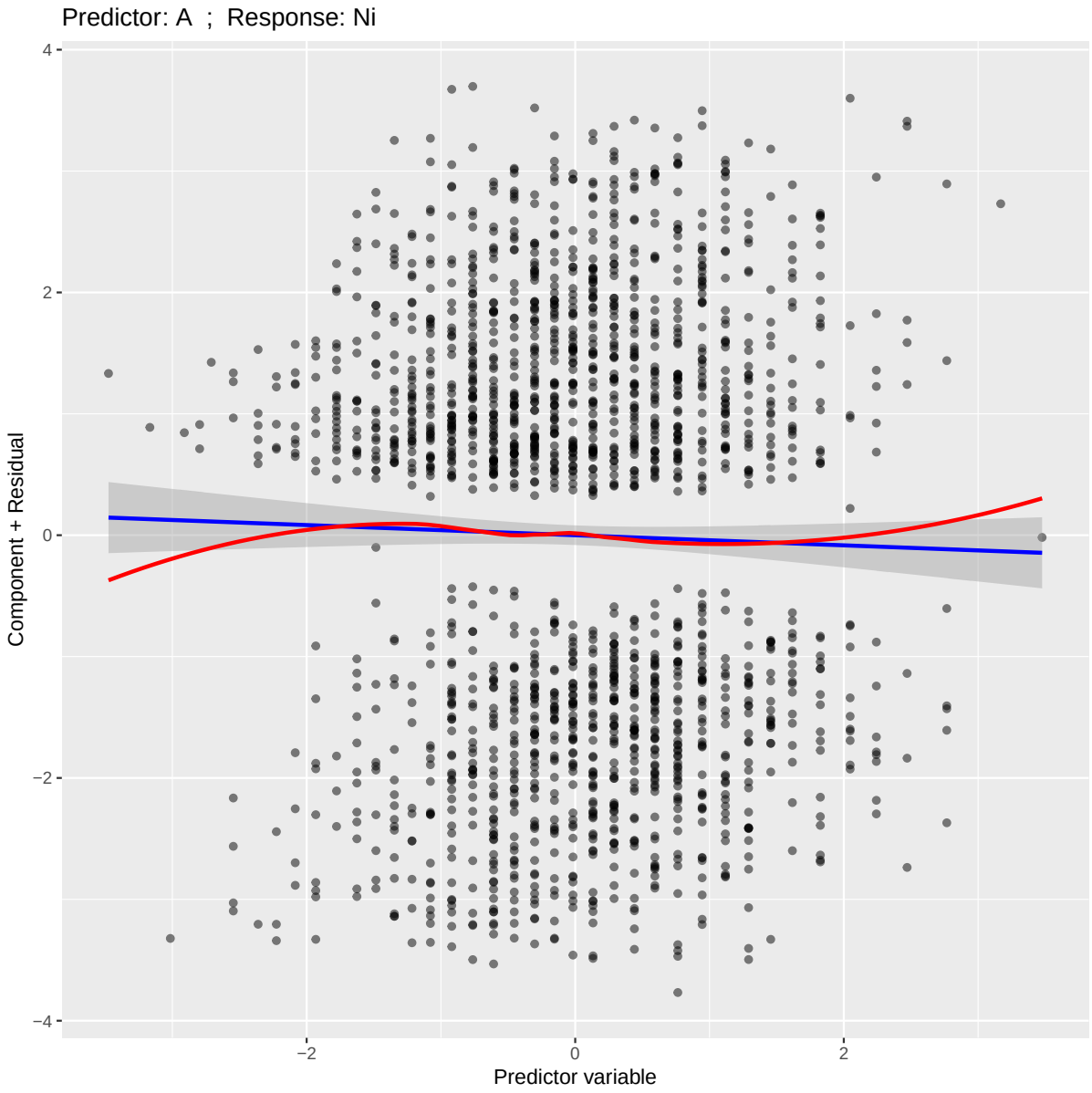

Predictor: A ; Response: Ke
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2


Predictor: A ; Response: Le
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

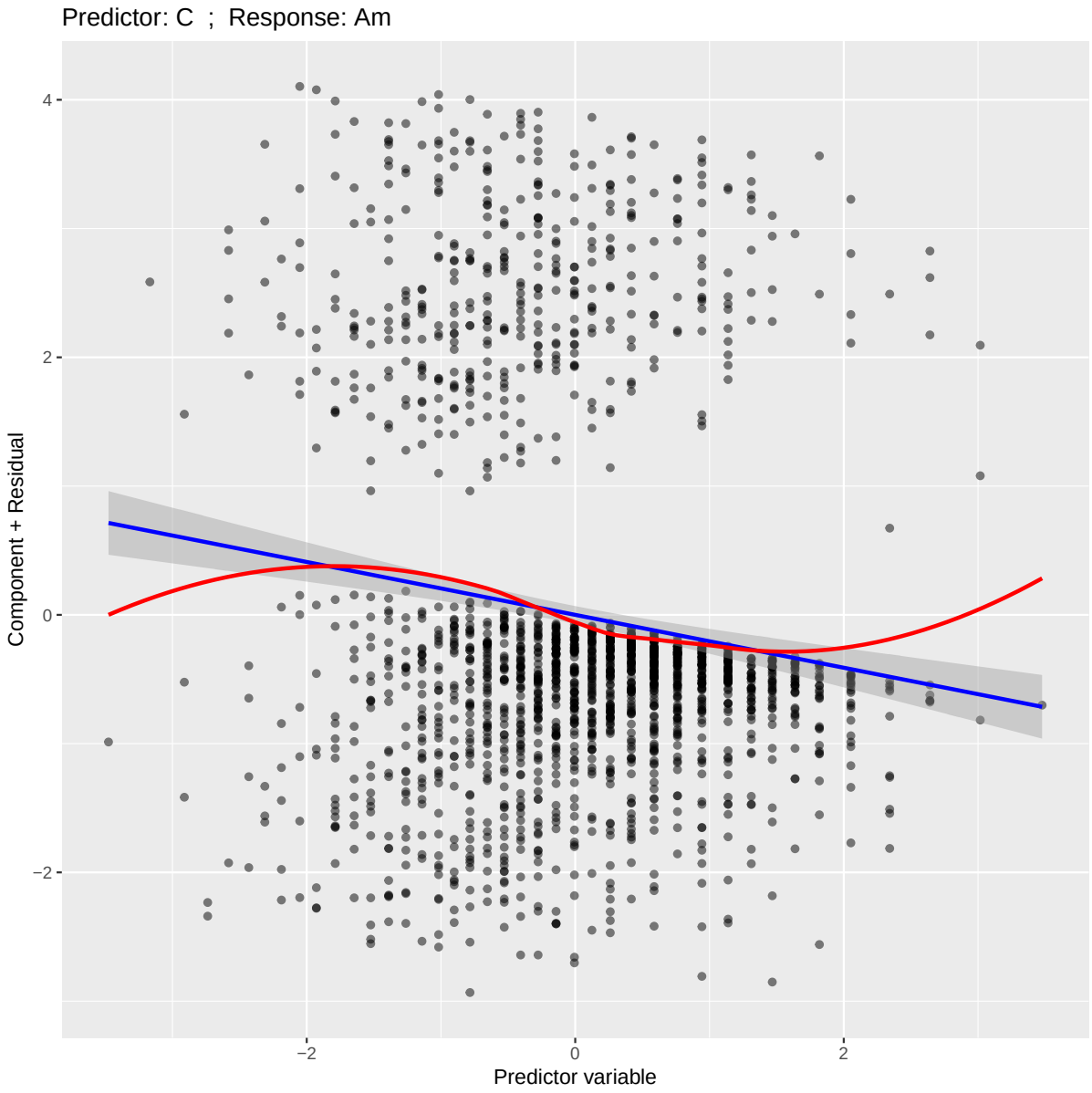

Predictor: A ; Response: LSD
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

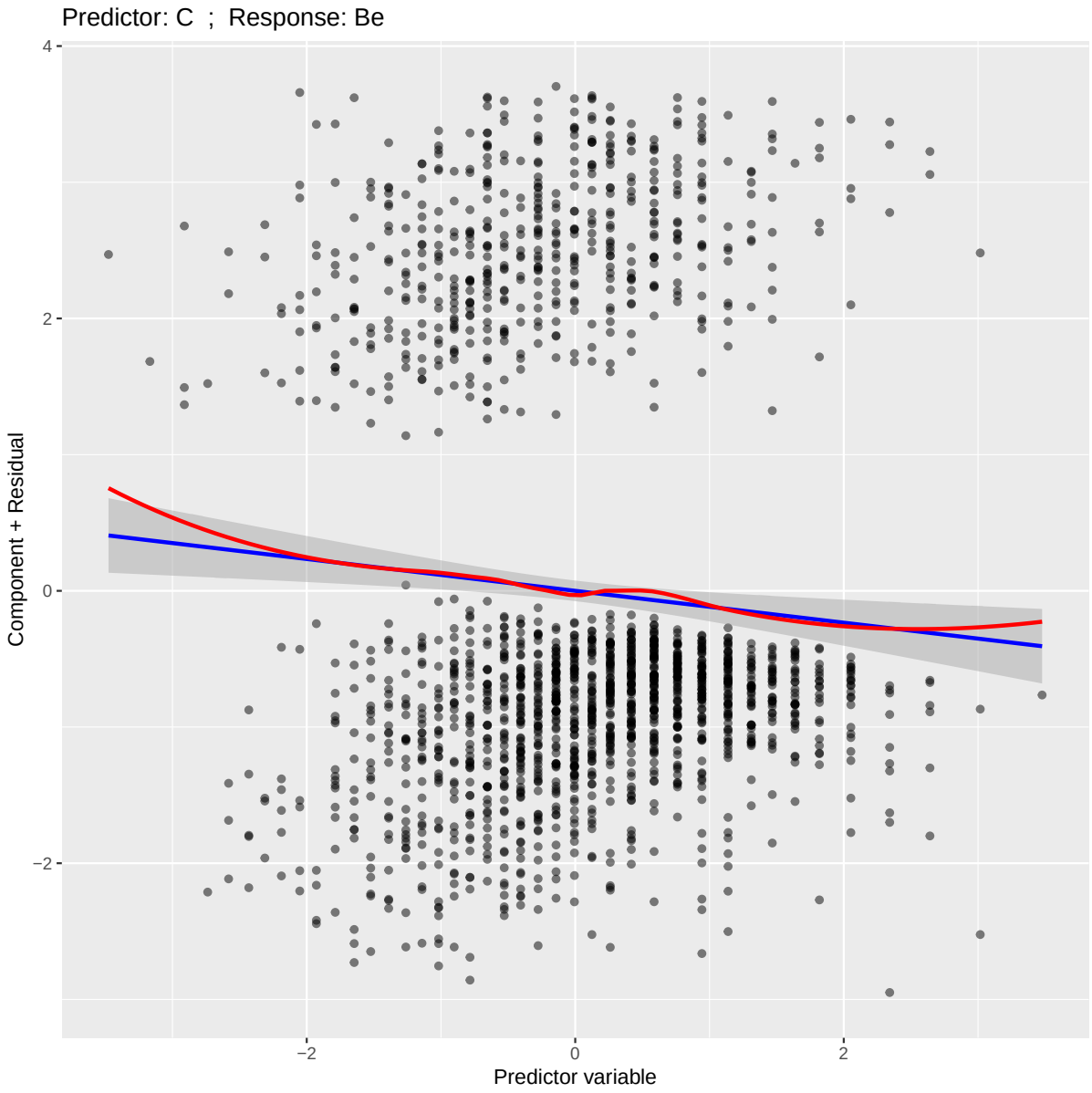

Predictor: A ; Response: Me
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

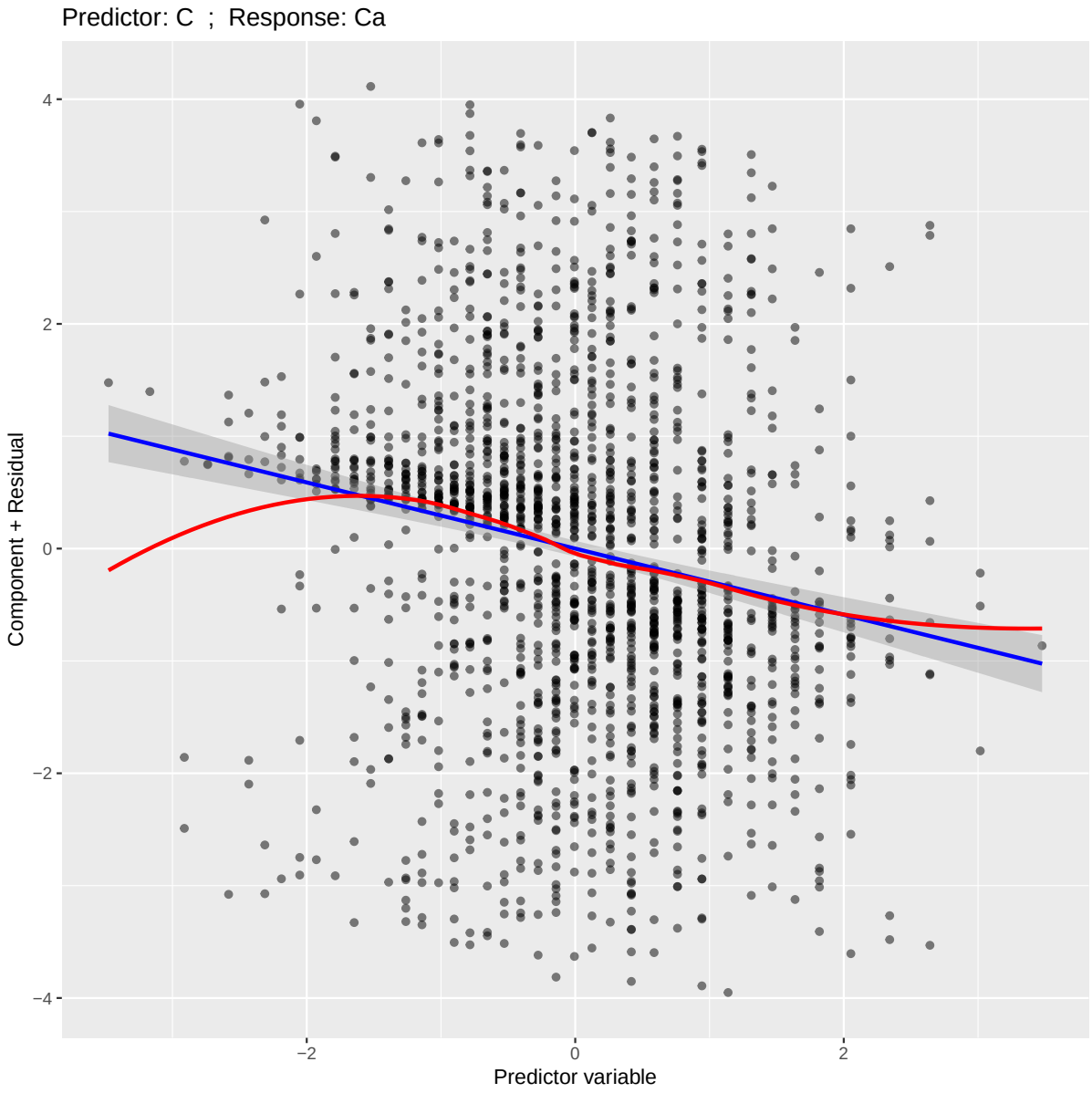

Predictor: A ; Response: Mu
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

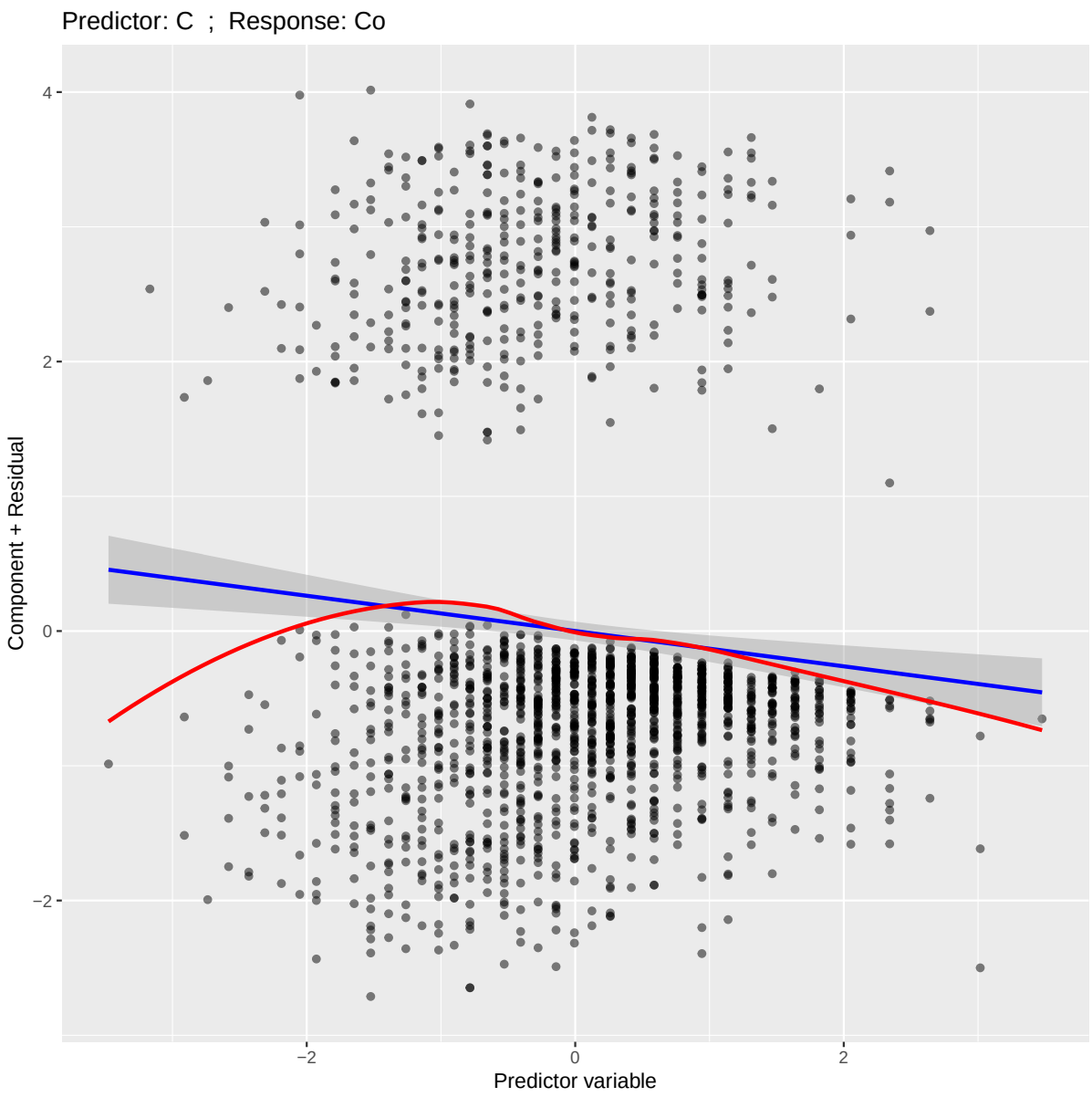


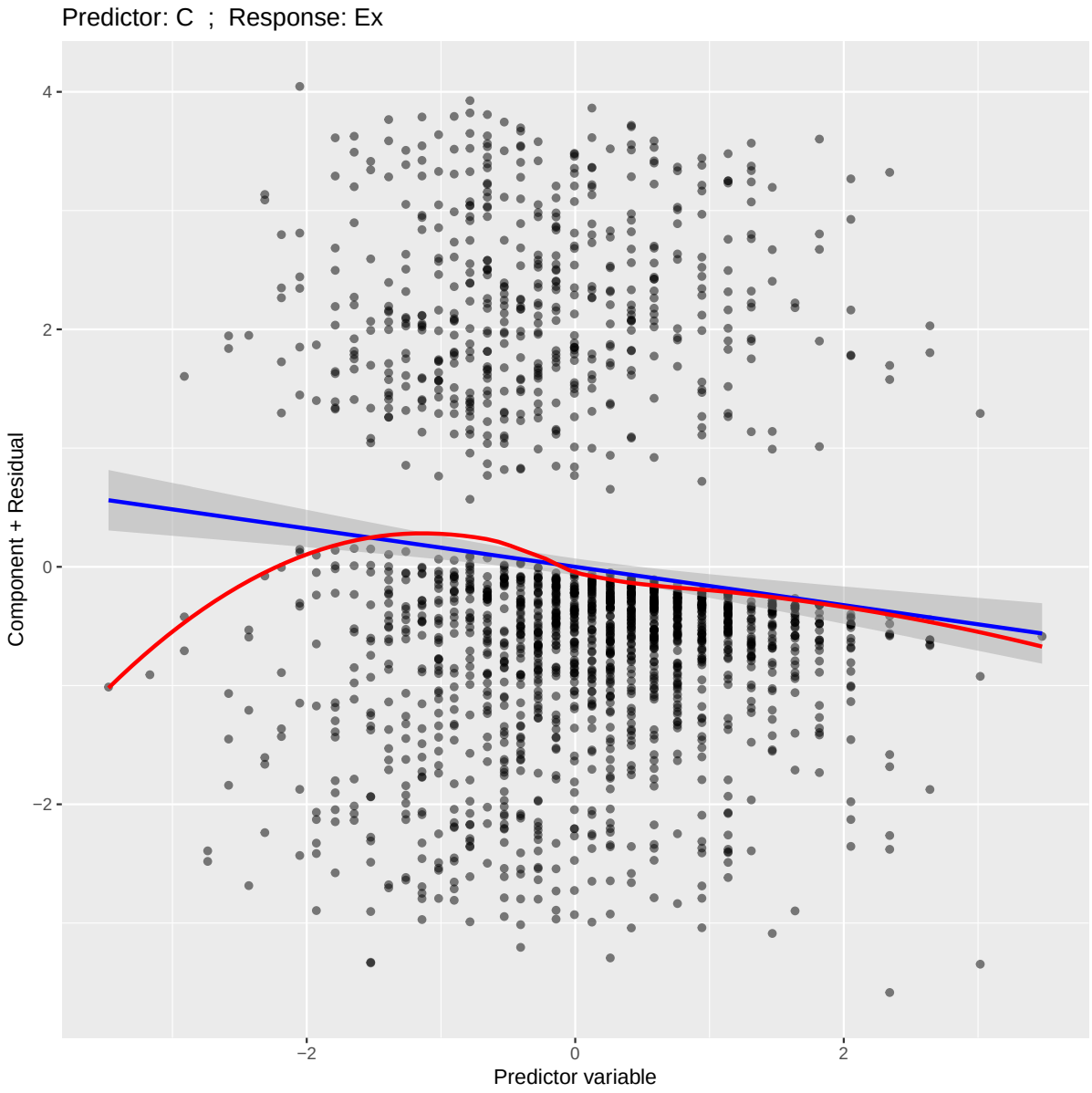
Seventh predictor

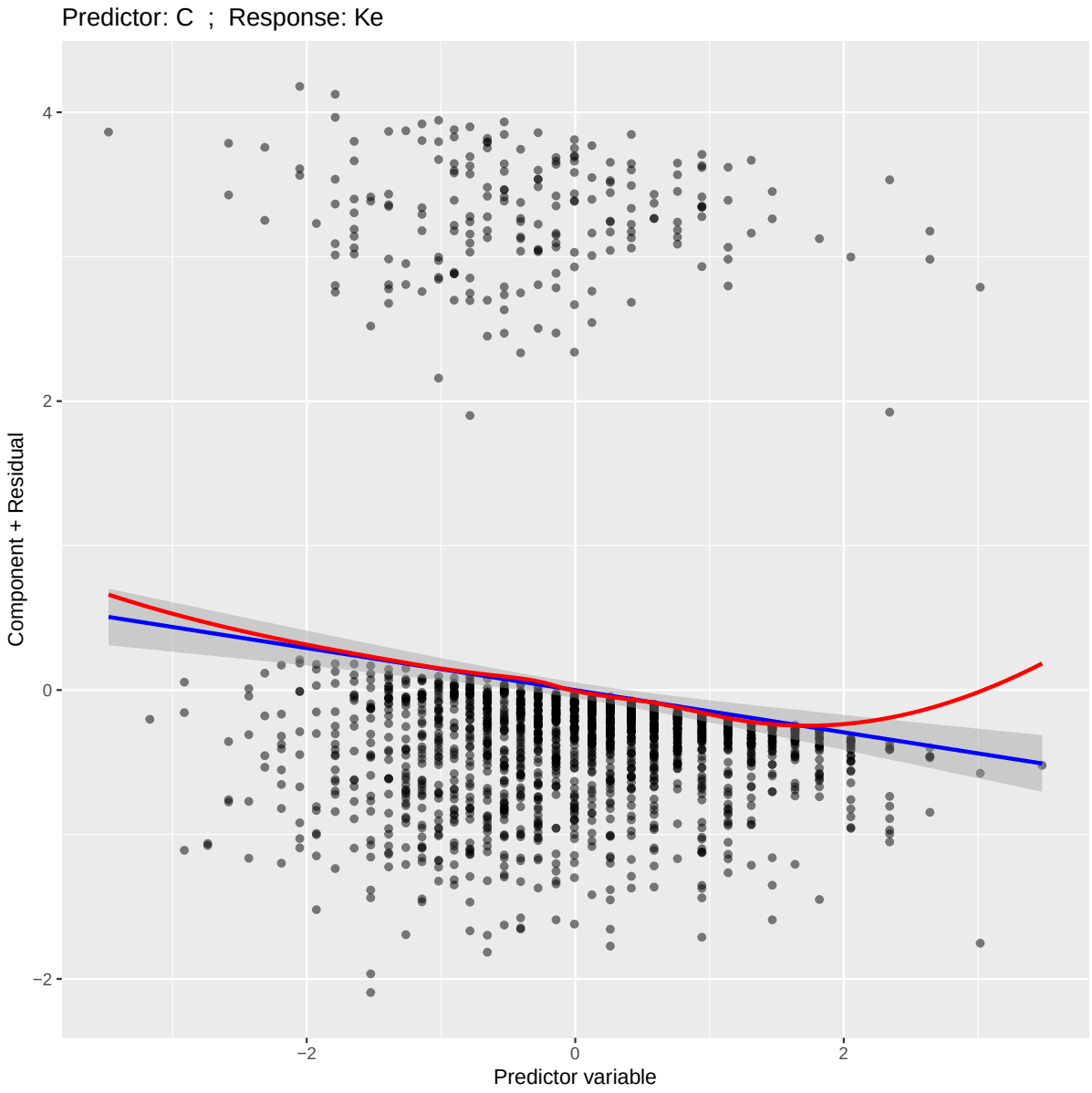

Predictor: C ; Response: Am
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

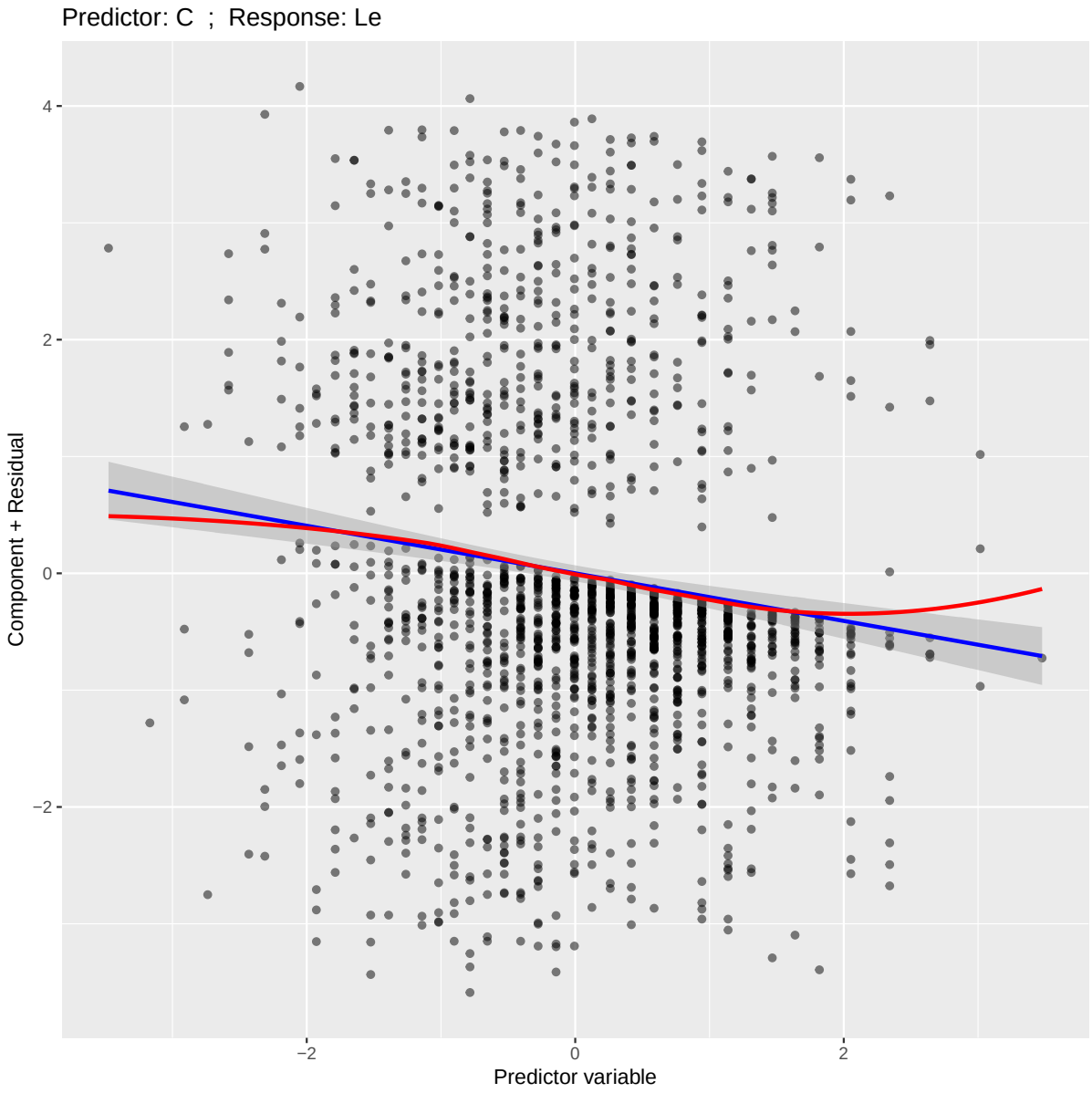

Predictor: C ; Response: Be
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

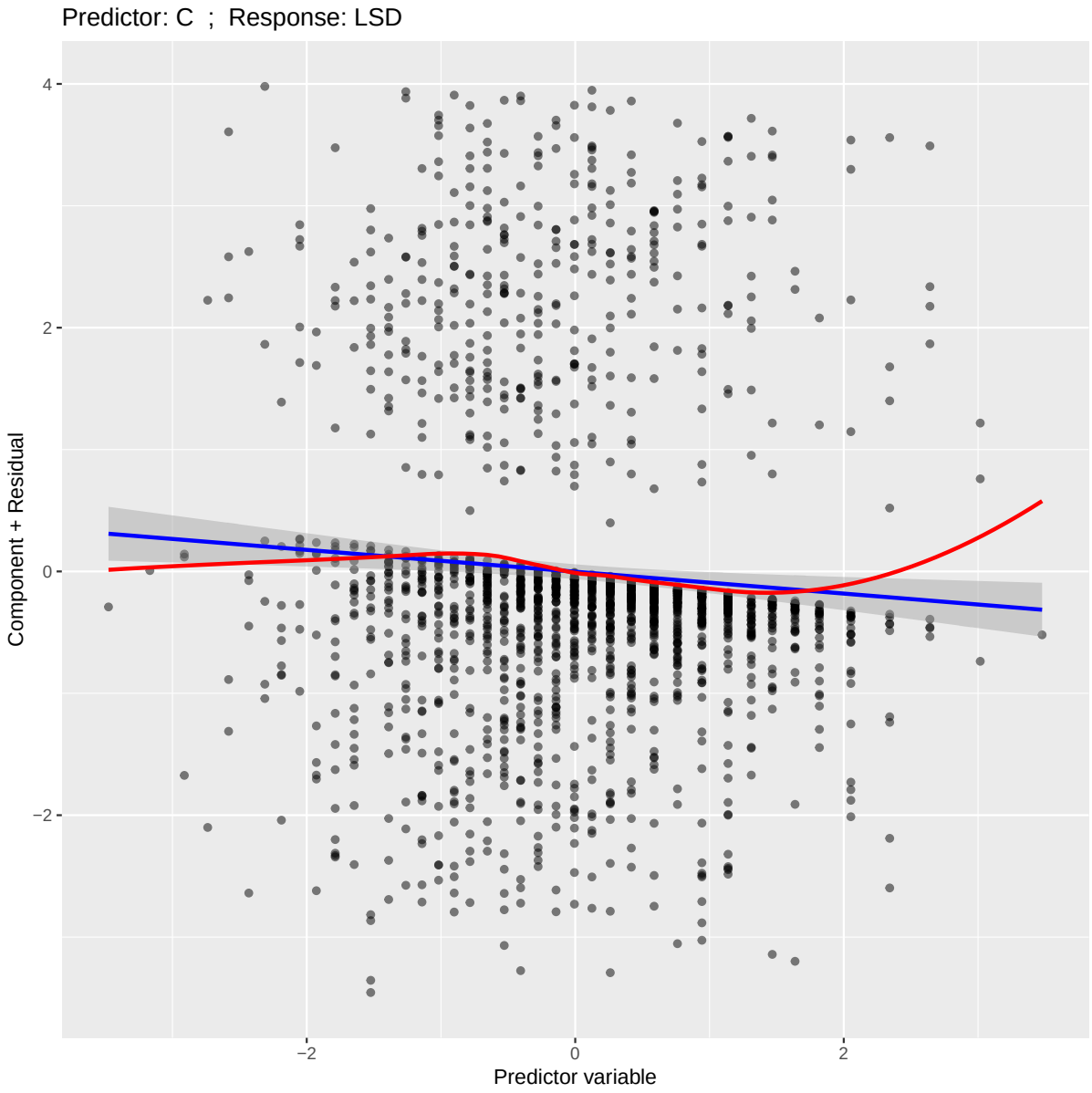

Predictor: C ; Response: Ca
Component + Residual
Predictor variable
4
2
0
−2
−4
−2
0
2

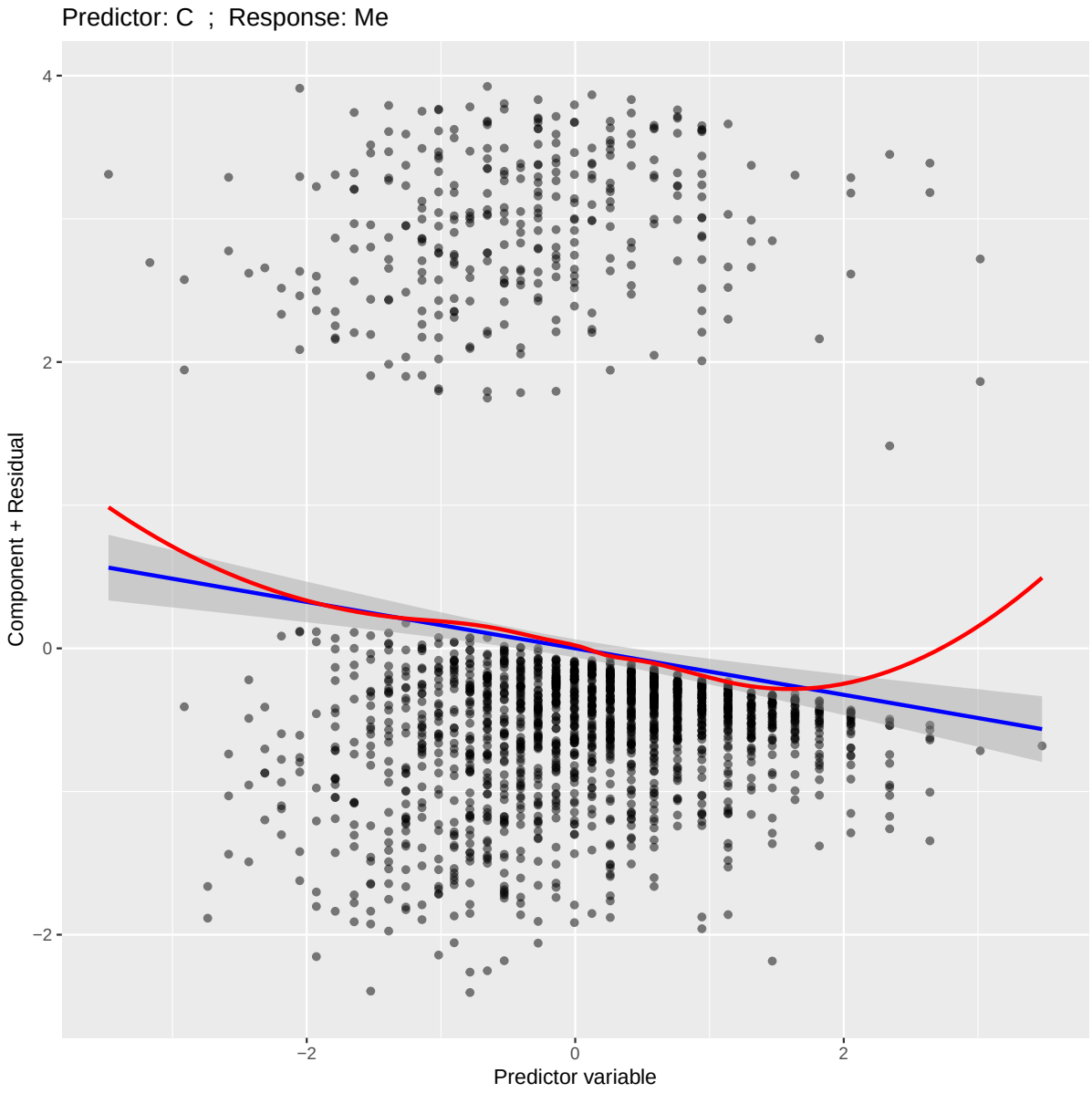

Predictor: C ; Response: Co
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

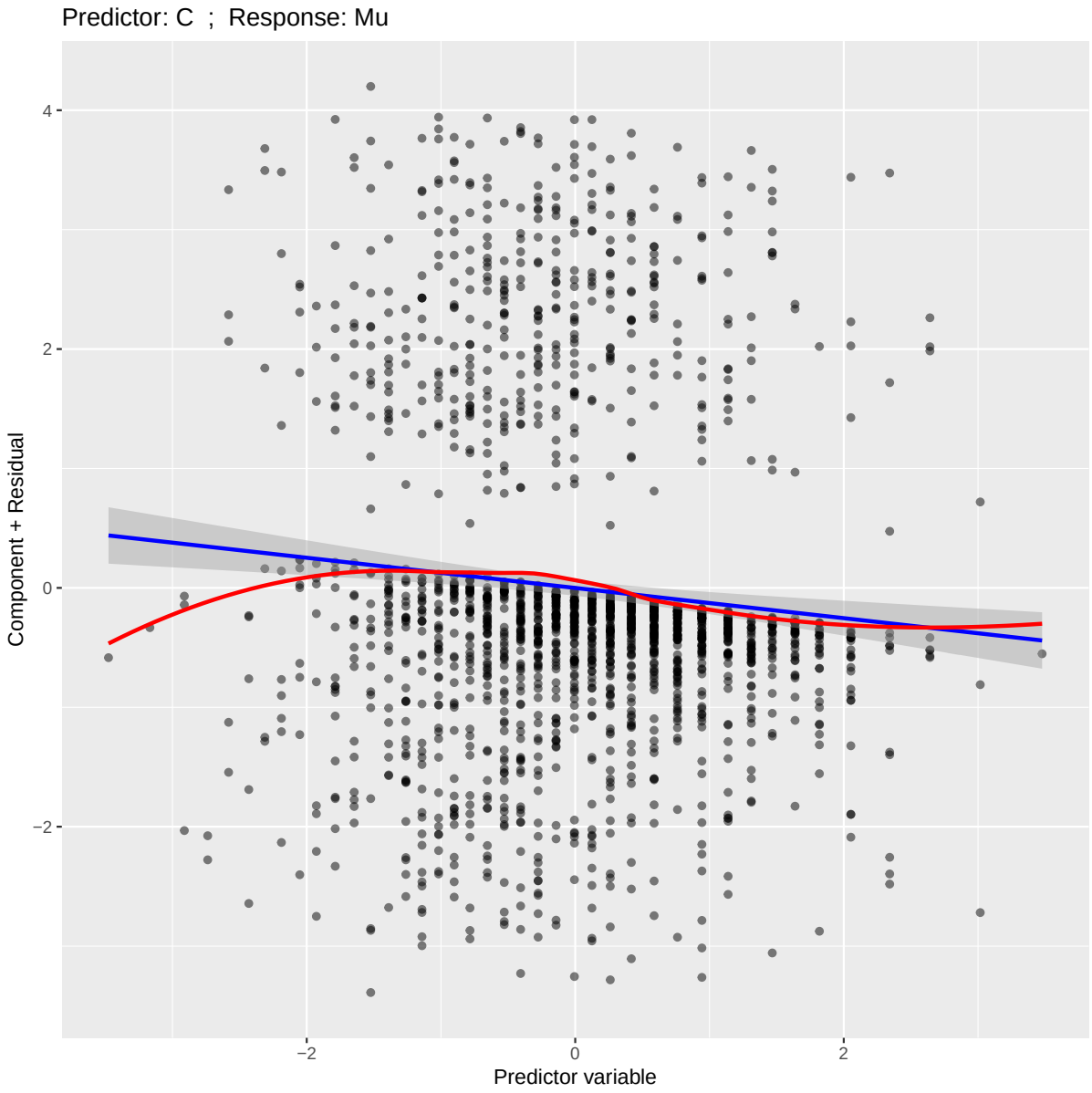

Predictor: C ; Response: Ex
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

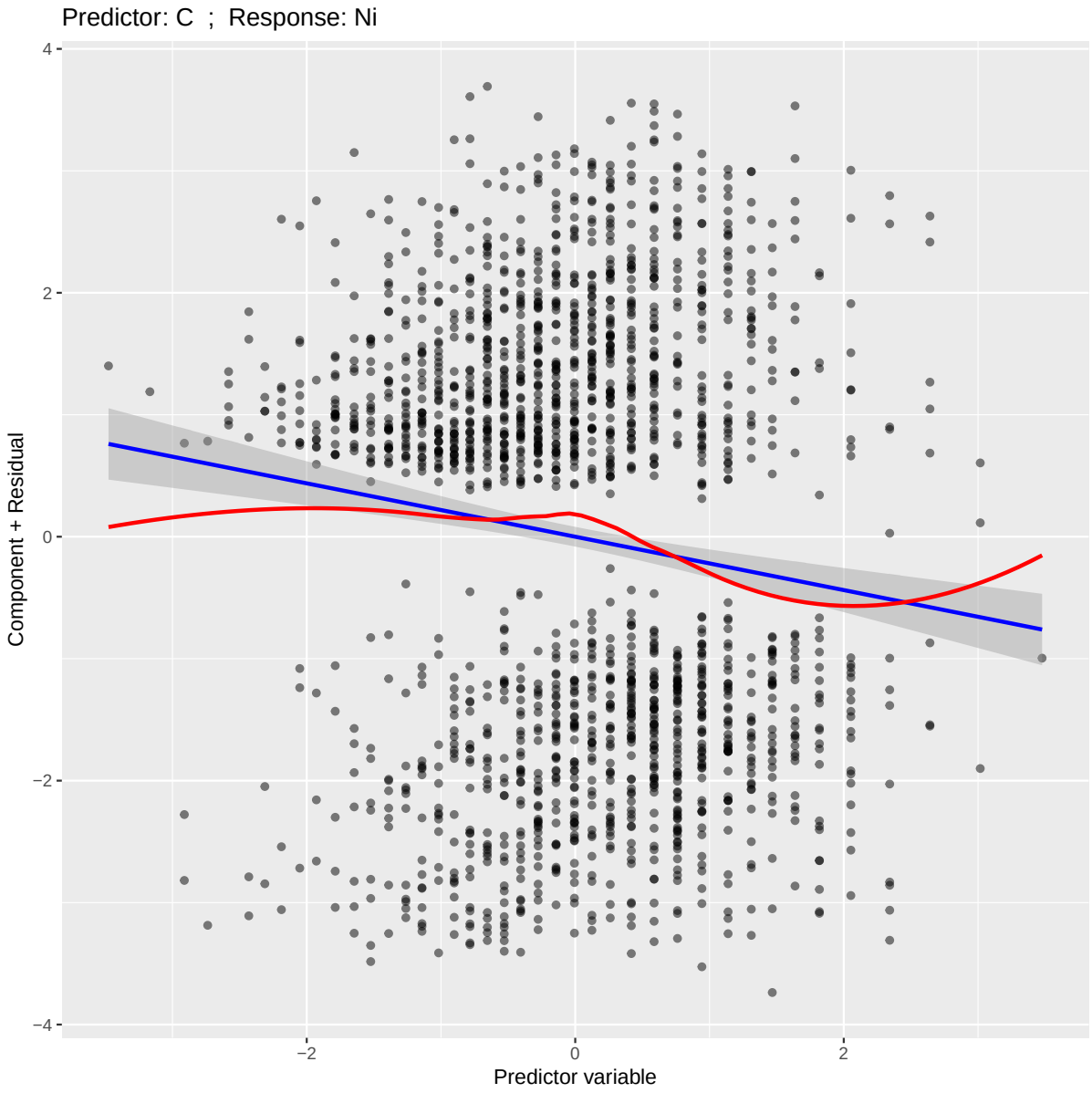

Predictor: C ; Response: Ke
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2


Predictor: C ; Response: Le
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

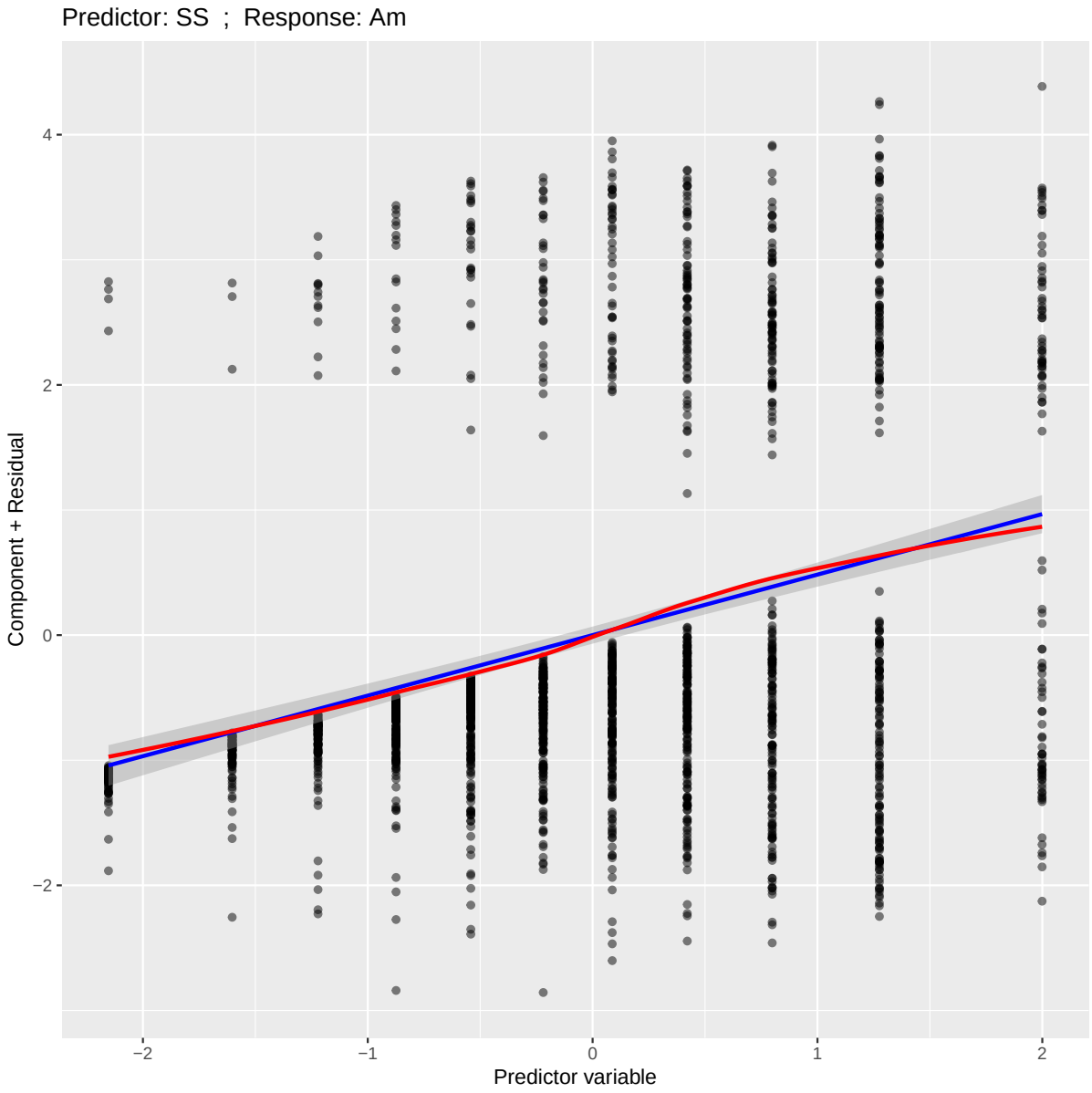

Predictor: C ; Response: LSD
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

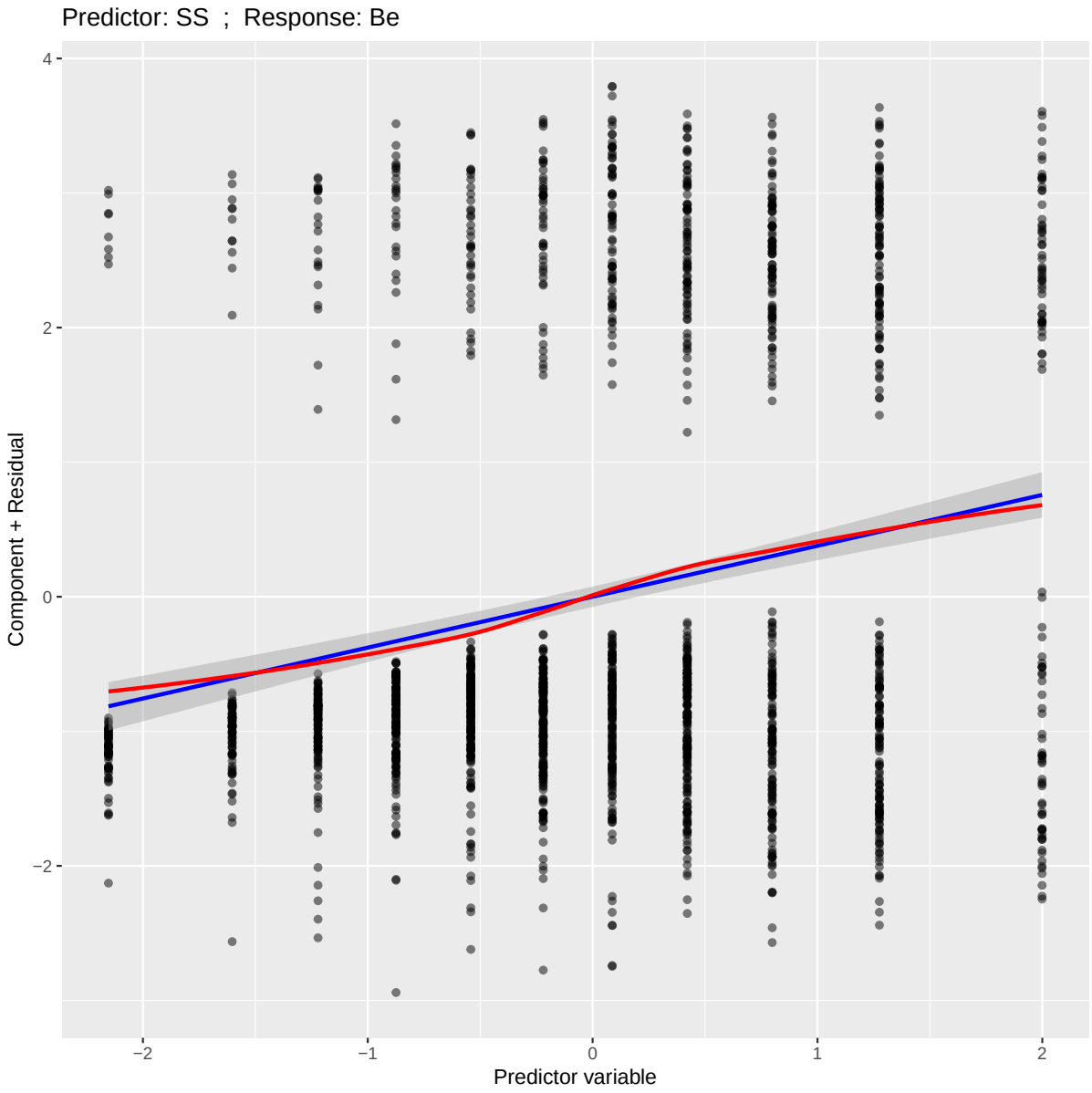

Predictor: C ; Response: Me
Component + Residual
Predictor variable
4
2
0
−2
−2
0
2

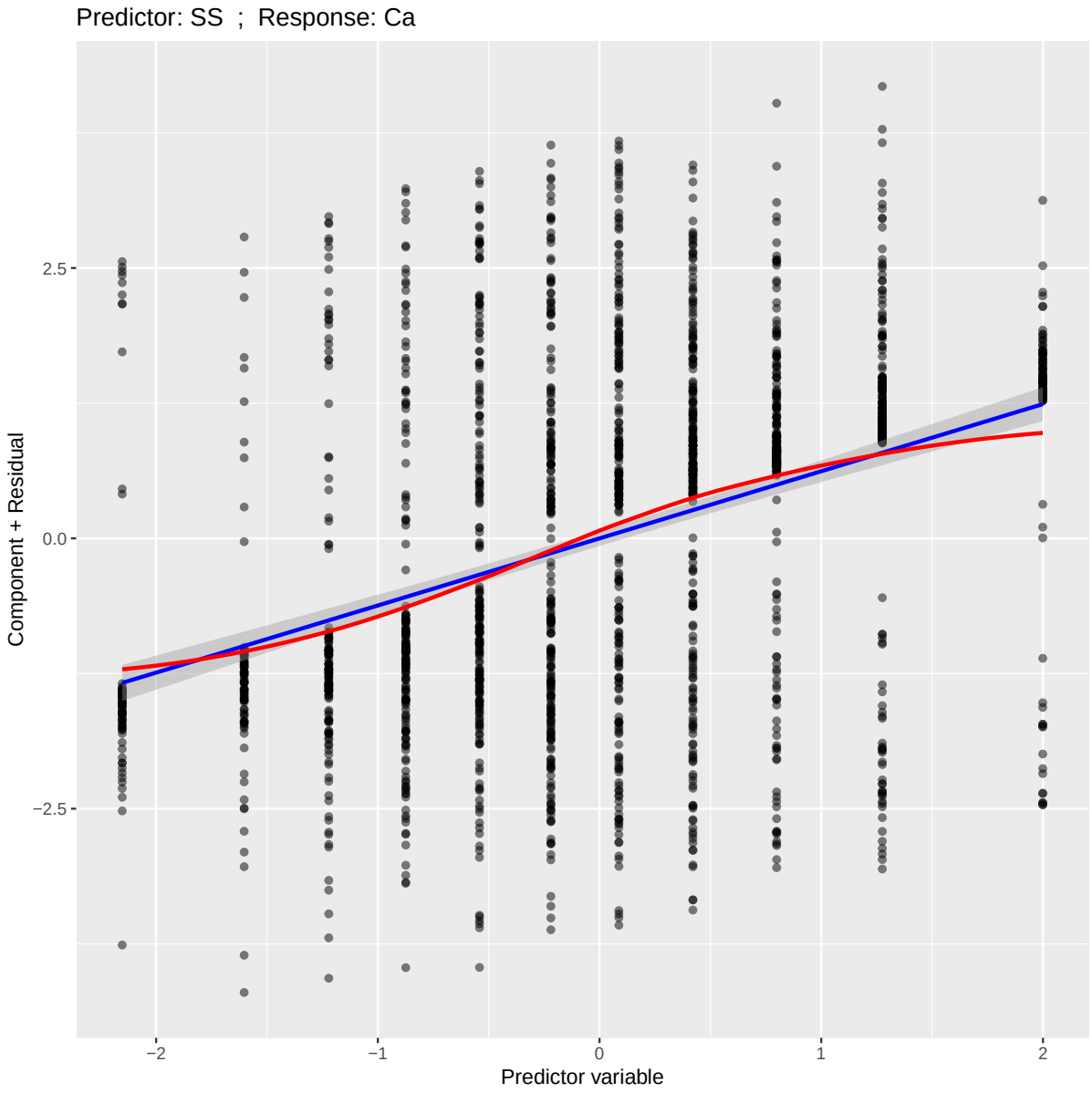

Predictor: C ; Response: Mu
Component + Residual
4
2
0
−2
−2
0
2
Predictor variable

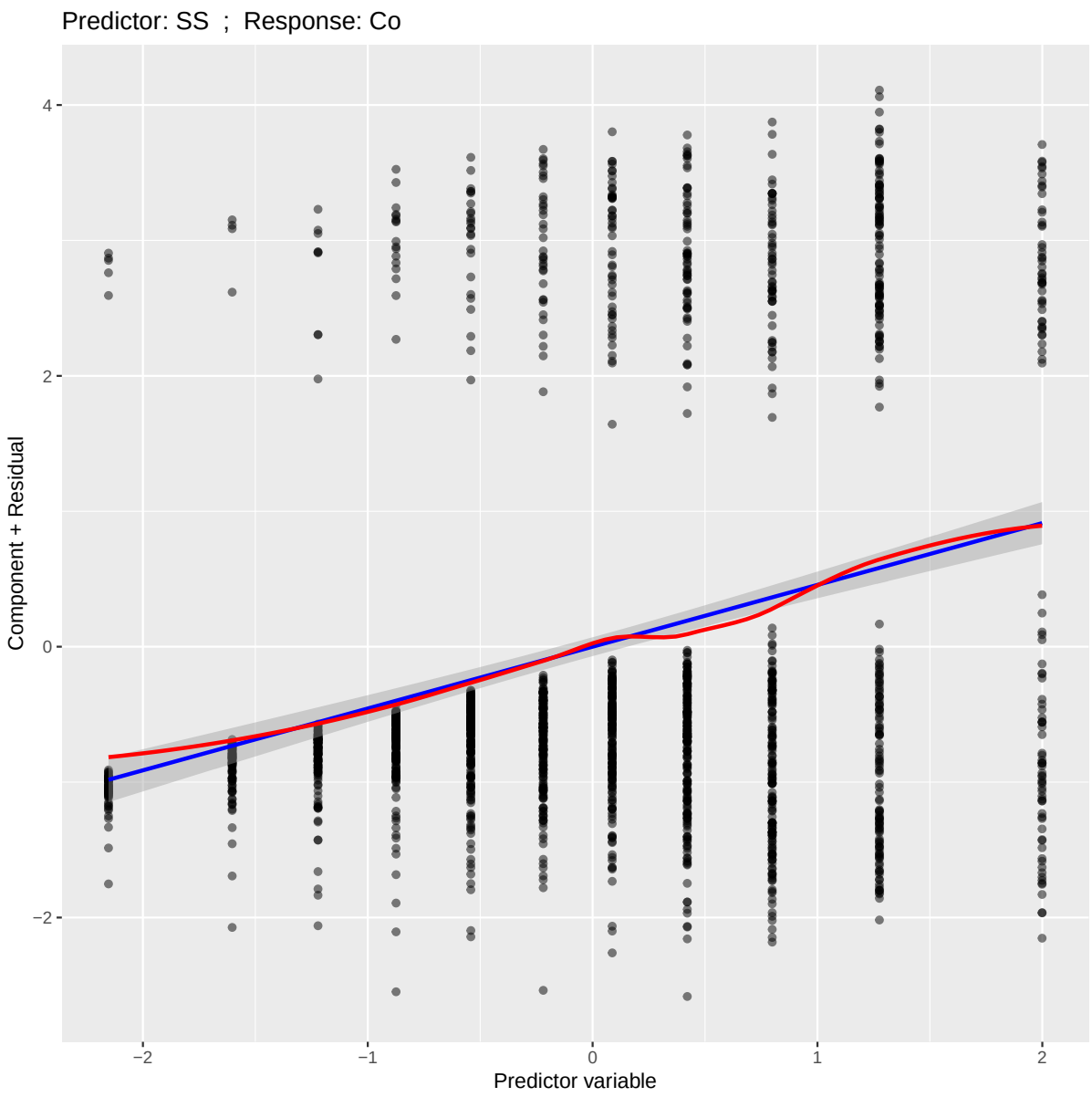


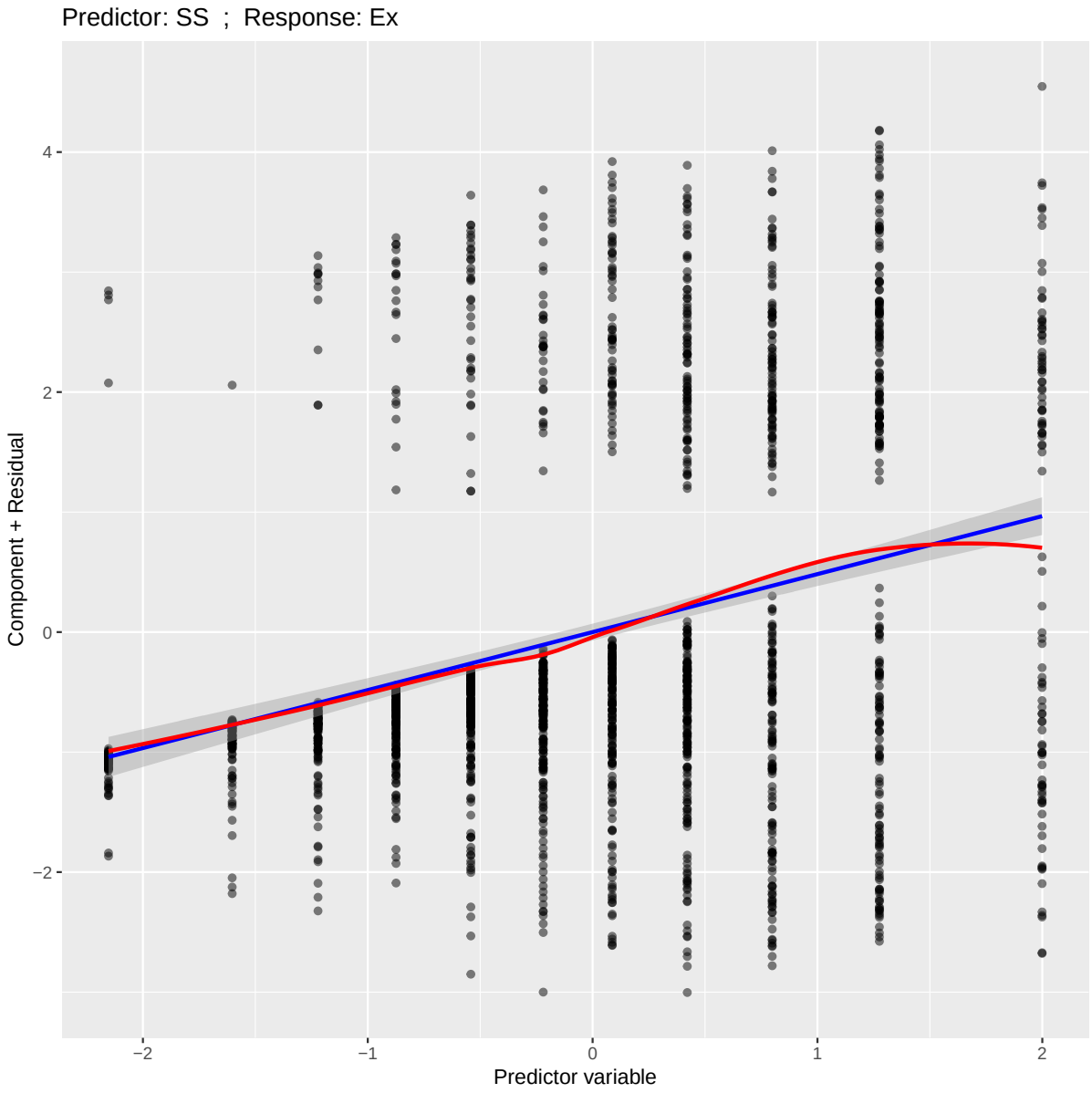
Eighth predictor

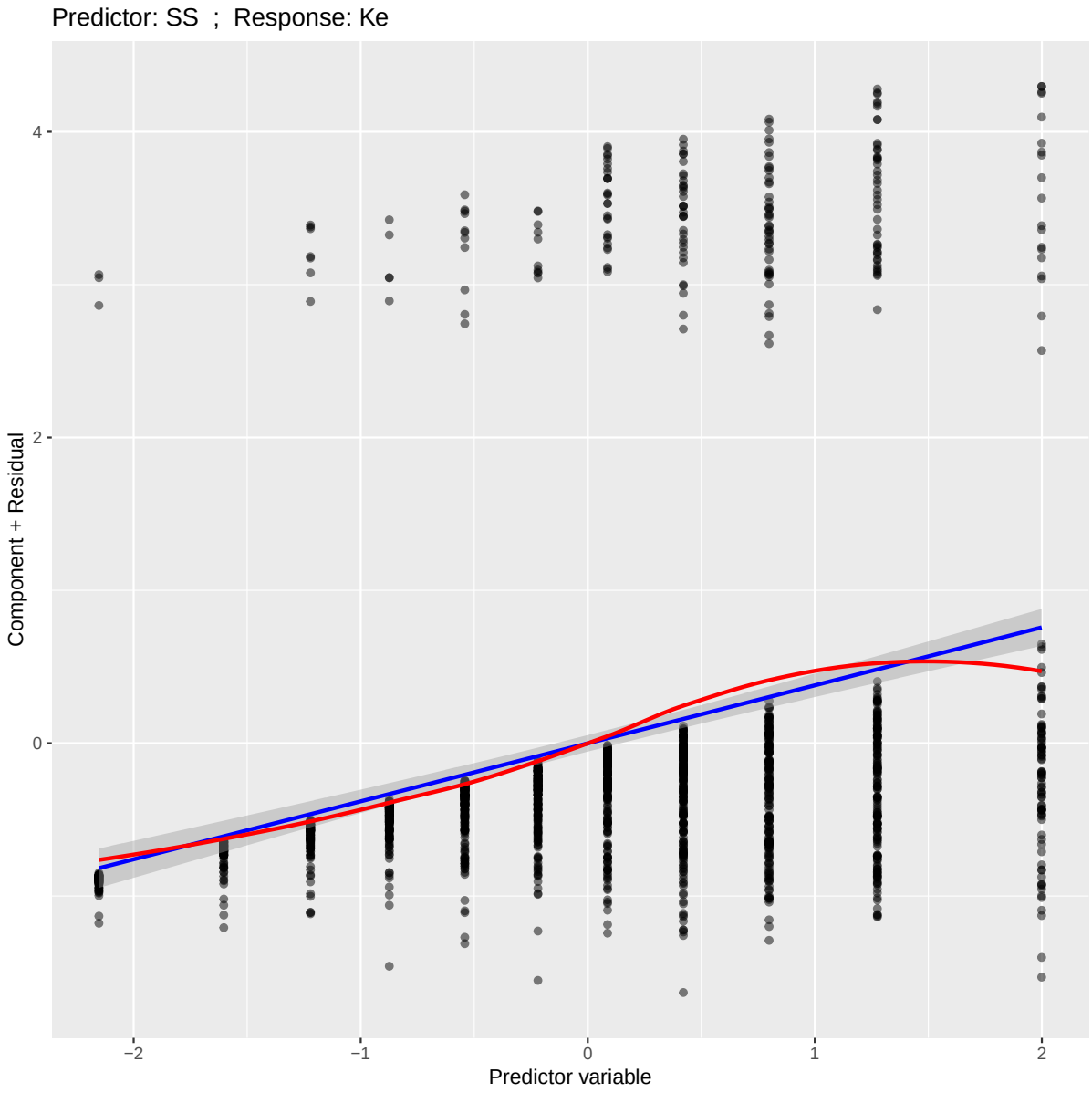

Predictor: SS ; Response: Am
Component + Residual
4
2
0
−2
−2
−1
0
1
2
Predictor variable

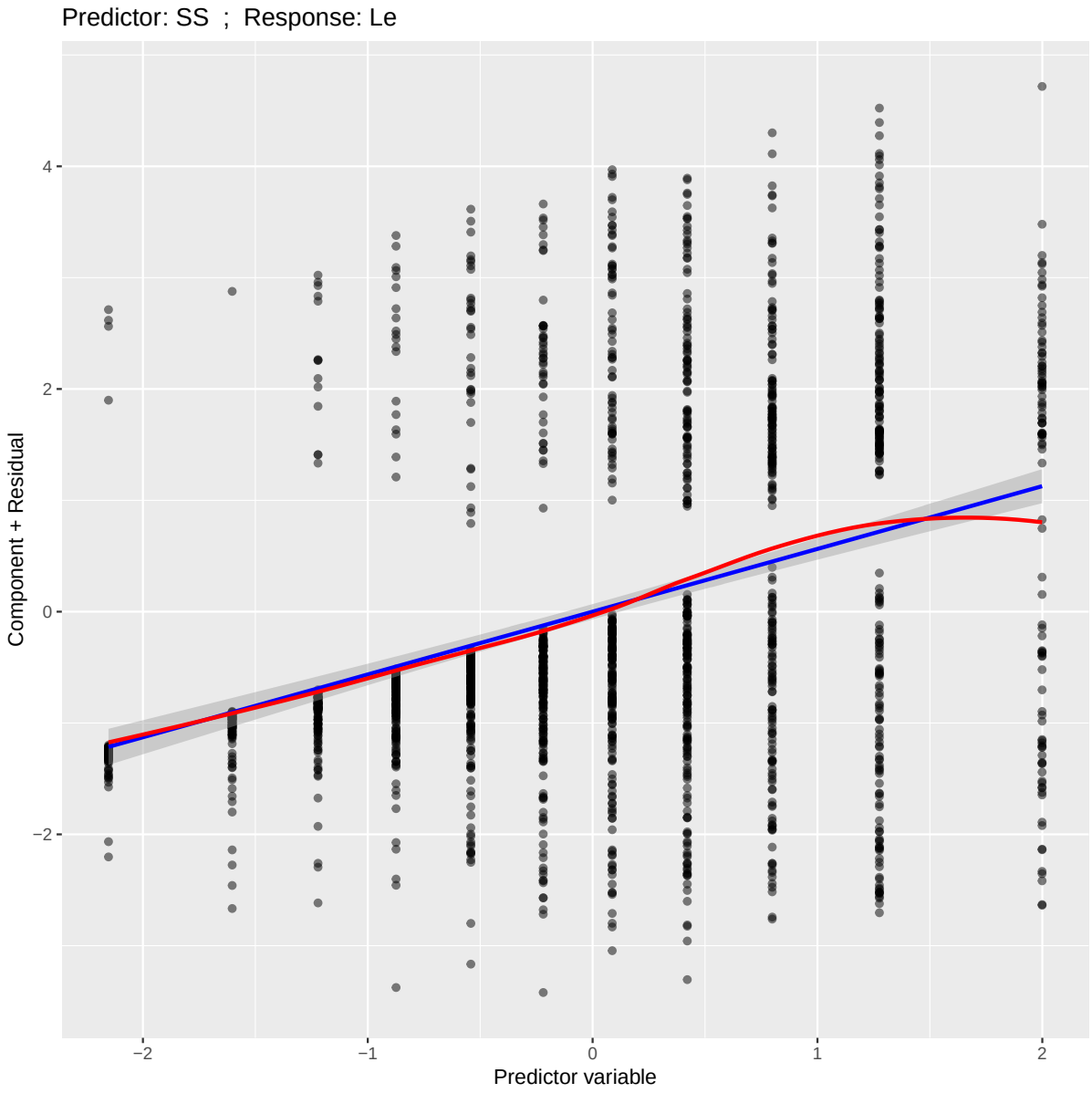

Predictor: SS ; Response: Be
Component + Residual
Predictor variable
4
2
0
−2
−2
−1
0
1
2

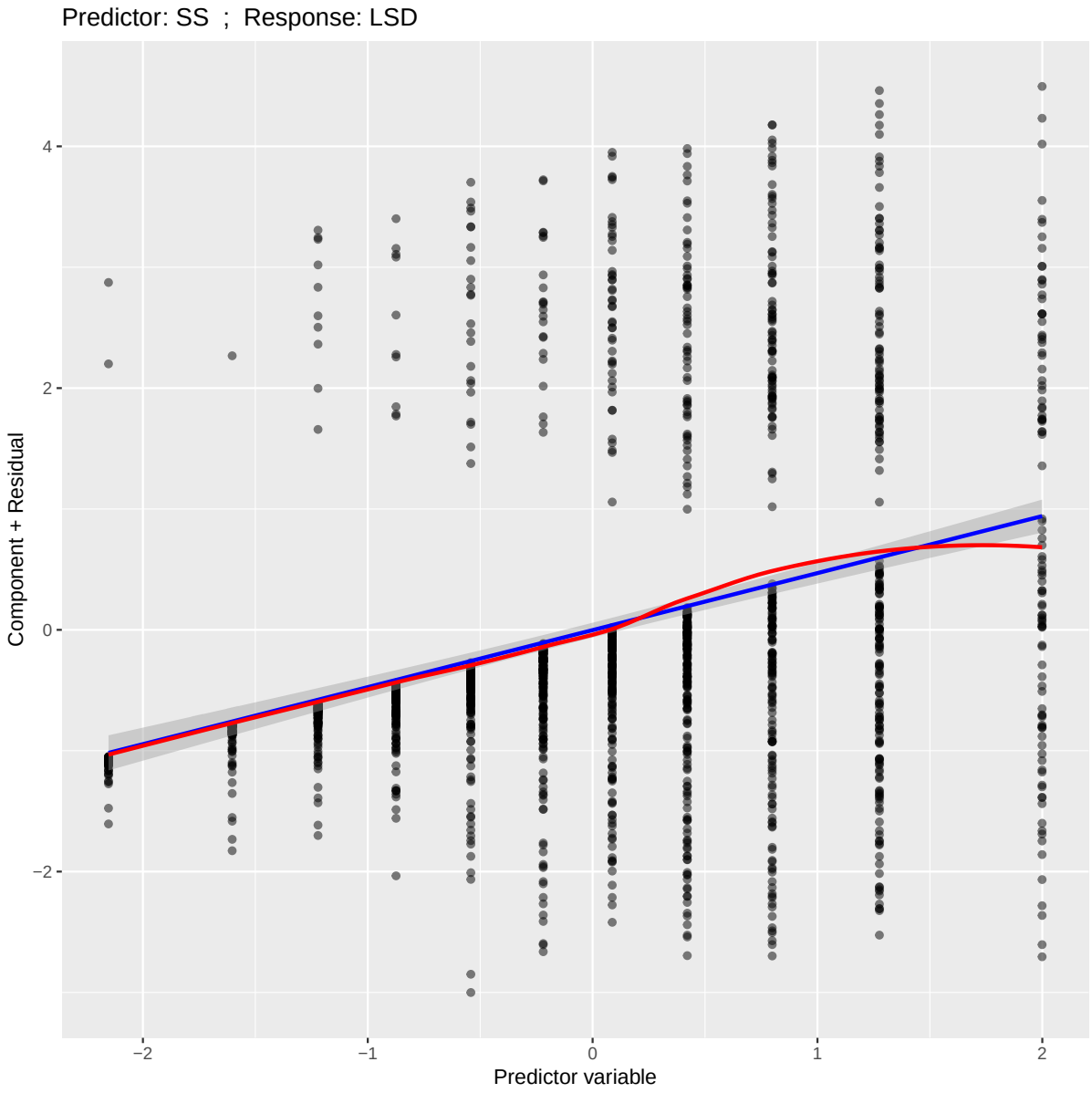

Predictor: SS ; Response: Ca
Component + Residual
2.5
0.0
−2.5
−2
−1
0
1
2
Predictor variable

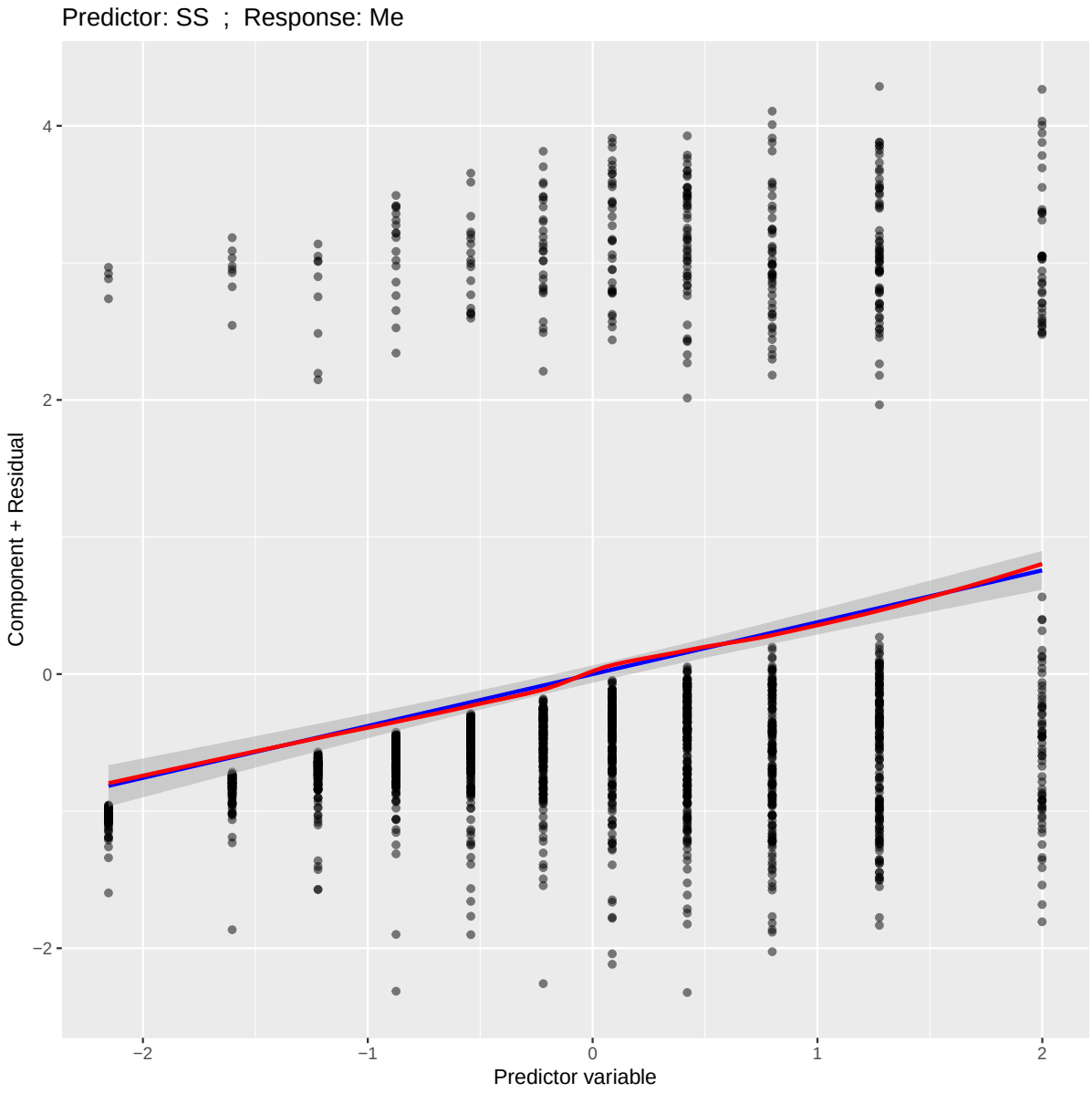

Predictor: SS ; Response: Co
Component + Residual
4
2
0
−2
−2
−1
0
1
2
Predictor variable

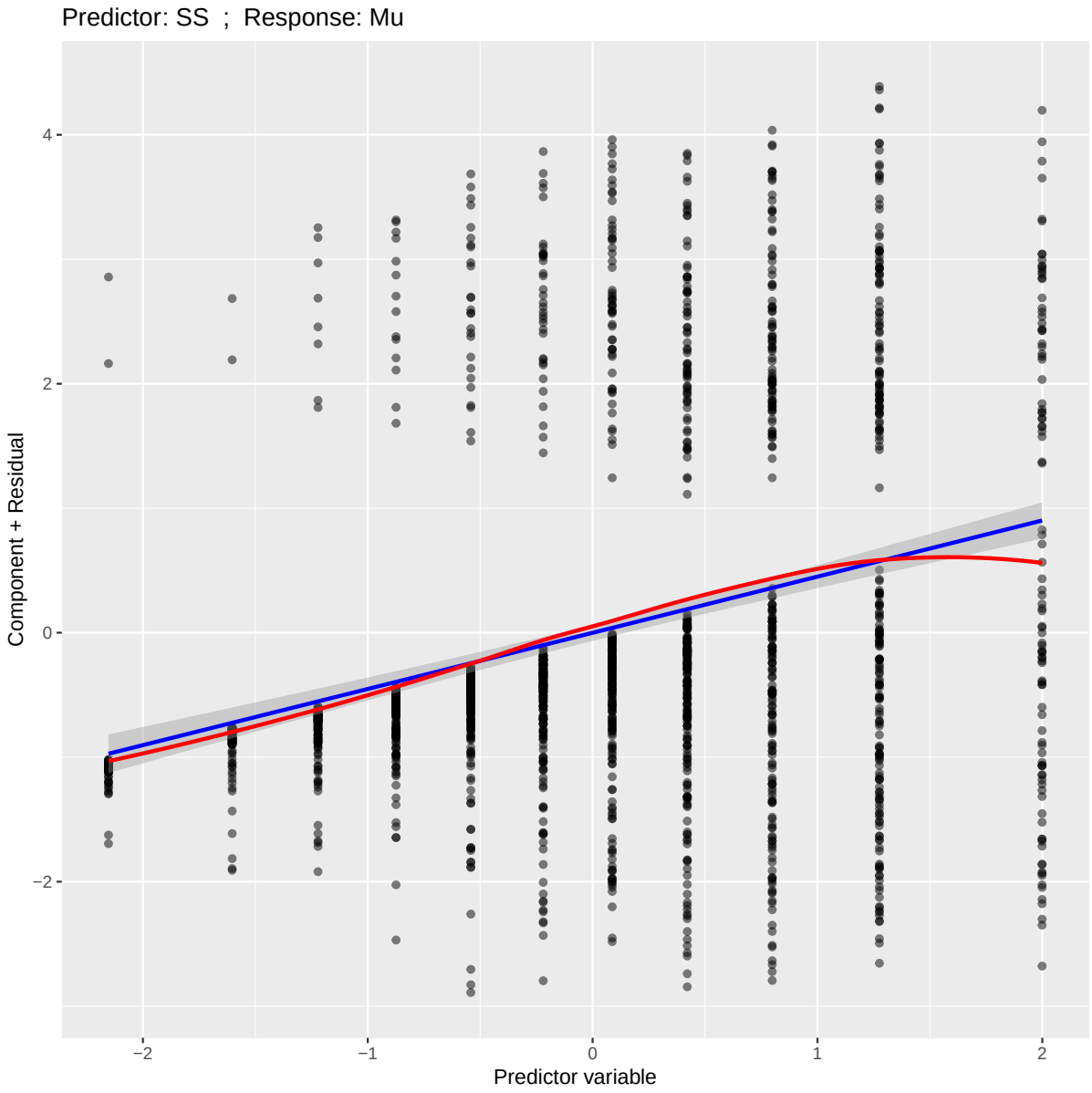

Predictor: SS ; Response: Ex
Component + Residual
4
2
0
−2
−2
−1
0
1
2
Predictor variable


Predictor: SS ; Response: Ke
Component + Residual
4
2
0
−2
−1
0
1
2
Predictor variable


Predictor: SS ; Response: Le
Component + Residual
4
2
0
−2
−2
−1
0
1
2
Predictor variable


Predictor: SS ; Response: LSD
Component + Residual
4
2
0
−2
−2
−1
0
1
2
Predictor variable


Predictor: SS ; Response: Me
Component + Residual
Predictor variable
4
2
0
−2
−2
−1
0
1
2


Predictor: SS ; Response: Mu
Component + Residual
4
2
0
−2
−2
−1
0
1
2
Predictor variable