#1

> table(test$Class, PredC)

PredC

0 1

0 18 4

1 1 42

Good prediction. Only 5 misclassification out of 65.

> print(m1)

Call: glm(formula = Class ~ ., family = binomial, data = combined)

Coefficients:

(Intercept) RI Na Mg Al Si K

8080.723 -4129.884 -13.395 -4.662 -36.192 -20.826 -8.160

Ca Ba Fe

-4.467 -18.878 27.583

Degrees of Freedom: 148 Total (i.e. Null); 139 Residual

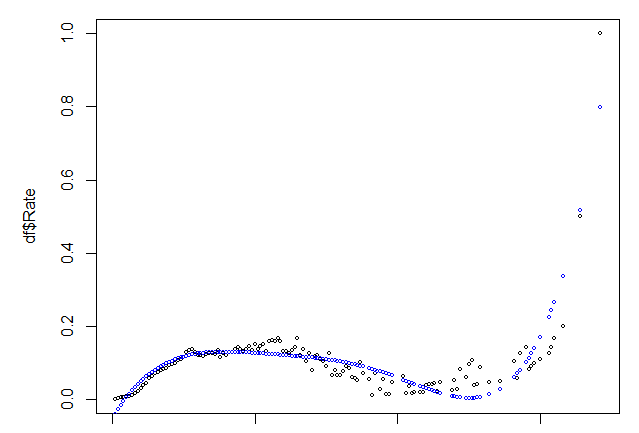
Null Deviance: 146.9

Residual Deviance: 14.74 AIC: 34.74

Probabilistic model:

Decision boundary:

#2



Model seems to be underfitted (biased). The problem is either with placement of the knot or the amount of knots.

> summary(m2)

Call:

lm(formula = Y ~ ., data = df1)

Residuals:

Min 1Q Median 3Q Max

-0.136799 -0.016559 -0.002466 0.016241 0.201196

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.262e-02 1.923e-02 -2.737 0.007063 \*\*

X1 1.319e-02 2.251e-03 5.859 3.62e-08 \*\*\*

X2 -3.420e-04 7.719e-05 -4.431 1.98e-05 \*\*\*

X3 3.804e-06 9.839e-07 3.866 0.000174 \*\*\*

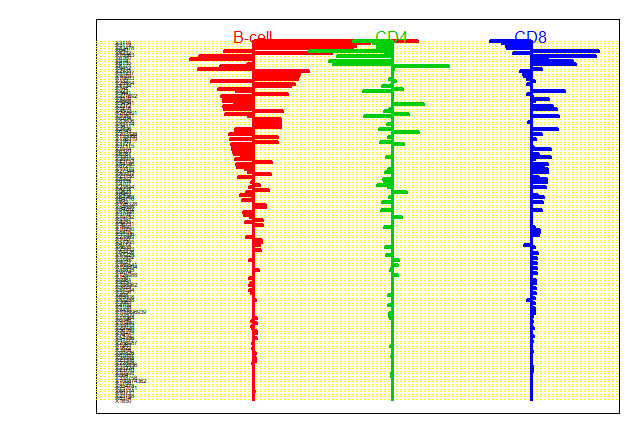
X4 -1.607e-08 4.179e-09 -3.844 0.000188 \*\*\*

X5 3.898e-08 6.604e-09 5.903 2.92e-08 \*\*\*

All terms are significant. Degrees of freedom is 6 – the amount of parameters in the linear model

The model is parametric since whatever data we take, the degrees of freedom will be 6 and therefore does not change with size/complexity of data.

#3



X3119

X3112

X83478

X640

X6232 are most important genes because they have largest deviations overall (differentiating B-cells).

Positive values mean positive deviation of the gene expression to the overall mean

There can not be all positive deviations because of this.

Total number of genes: > print(nrow(a))

[1] 143