LNL_HW_week9

Patrick Kelly

Thursday, March 12, 2015

The data in the file HIV_Protease.csv show part of the analysis of HIV enzyme called Protease. Protease plays important role in spreading the virus and is often targeted by drugs trying to contain the virus in patient.

Because HIV virus mutates a lot it develops resistance to drugs when codons in certain positions in the sequence enzyme get replaced by other codons.

Read the data.

```
HIV.Protease.Data<-read.csv(file="C:/Users/Patrick/Documents/R/UChicago/Linear_NonLinear/HIV_Protease.cs
HIV.Protease.Data<-as.data.frame(HIV.Protease.Data)

head(HIV.Protease.Data)
```

```
##   Resp  VL CODON_2
## 1    0 4.3     caa
## 2    0 3.6     caa
## 3    0 3.2     caa
## 4    0 5.7     caa
## 5    0 3.5     caa
## 6    0 3.9     caa
```

The first column shows if the effect of the treatment on number of patients: 0 means no effect (resistance to drug) and 1 means successful treatment.

The second column shows the virus load, i.e. some measure of home many viruses are in the patient's body before the treatment. The last column shows the symbol in the second position (CODON_2) of HIV Protease which each patient has.

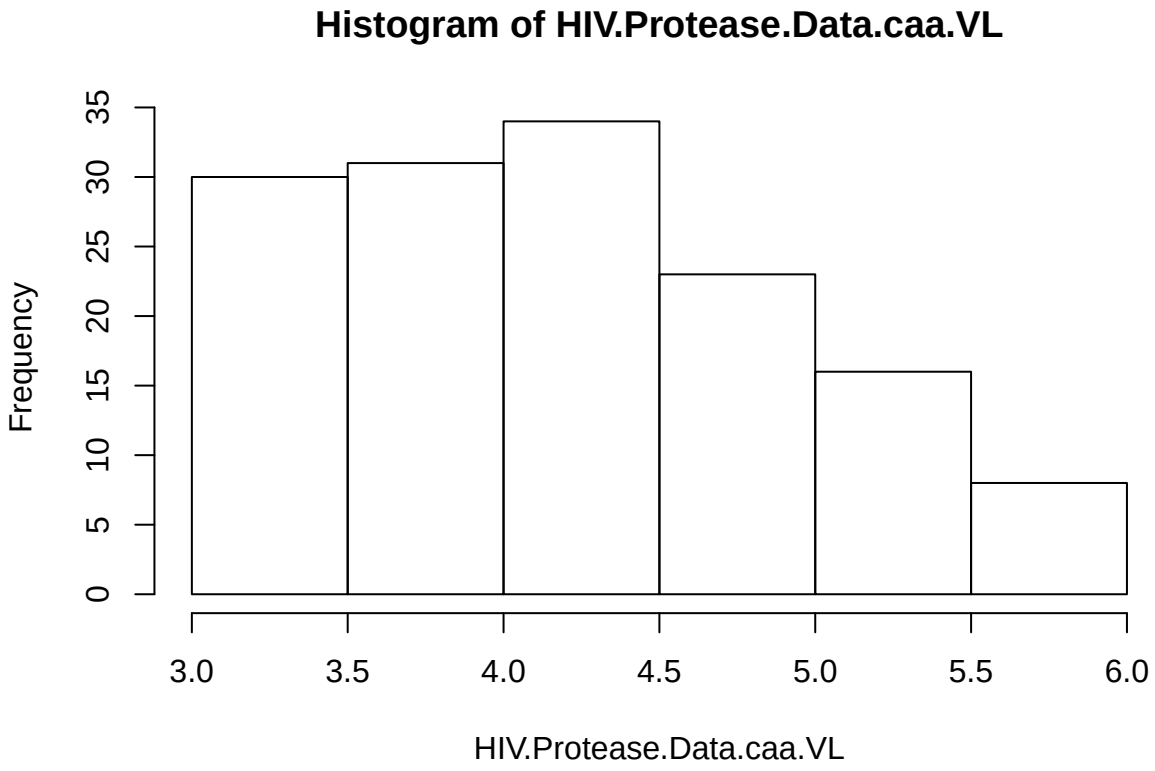
Show all possible expressions of CODON_2 and how many times they appear in the last column.

```
table(HIV.Protease.Data$CODON_2)
```

```
##
## caa cag cak car maa
## 142 691   1  67   1
```

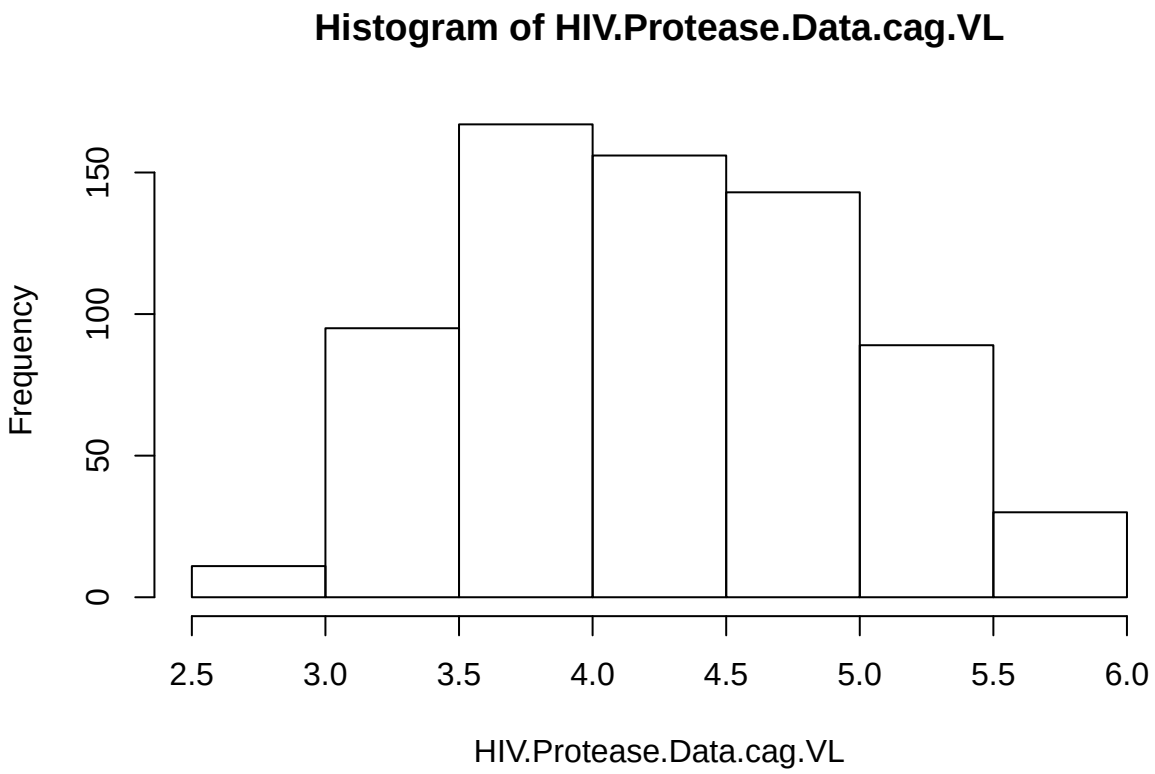
Show the levels of VL that appear simultaneously with expression "caa" and look at the histogram of VL values.

```
HIV.Protease.Data.caa.VL<-subset(HIV.Protease.Data$VL,HIV.Protease.Data$CODON_2=="caa")
hist(HIV.Protease.Data.caa.VL)
```



Do the same for the expression “cag”

```
HIV.Protease.Data.cag.VL<-subset(HIV.Protease.Data$VL,HIV.Protease.Data$CODON_2=="cag")
hist(HIV.Protease.Data.cag.VL)
```



Compare the means.

```
c(mean(HIV.Protease.Data.cag.VL),mean(HIV.Protease.Data.caa.VL))
```

```
## [1] 4.314616 4.248239
```

Fit linear model

```
lm.CODON_2.VL <- lm(VL ~ CODON_2, data=HIV.Protease.Data)

AIC(lm.CODON_2.VL)
```

```
## [1] 1931.876
```

```
anova(lm.CODON_2.VL)
```

```
## Analysis of Variance Table
##
## Response: VL
##            Df Sum Sq Mean Sq F value Pr(>F)
## CODON_2     4   2.44 0.60906  1.2312 0.2959
## Residuals 897 443.72 0.49468
```

The means do not seem to be different. Explain why.

The means do not appear to be different because although the distributions of are different between caa and cag (as observed above), the centers (averages) of the VL values across the different `CODON_2` values are similar (also see above). Observe the pvalue of the ANOVA above, which indicates that we cannot reject the null hypothesis that the means are different.

Fit the model with random effect.

```
library(lme4)
```

```
## Loading required package: Matrix
## Loading required package: Rcpp
```

```
lm.CODON_2.VL.Mixed <- lmer(VL ~ (1|CODON_2), data=HIV.Protease.Data)

AIC(lm.CODON_2.VL.Mixed)
```

```
## [1] 1936.485
```

Compare the fit with model with fixed effect.

The linear model with the fixed effect is marginally better than the random effect model based on the AIC values (1931.9 vs 1936.5).

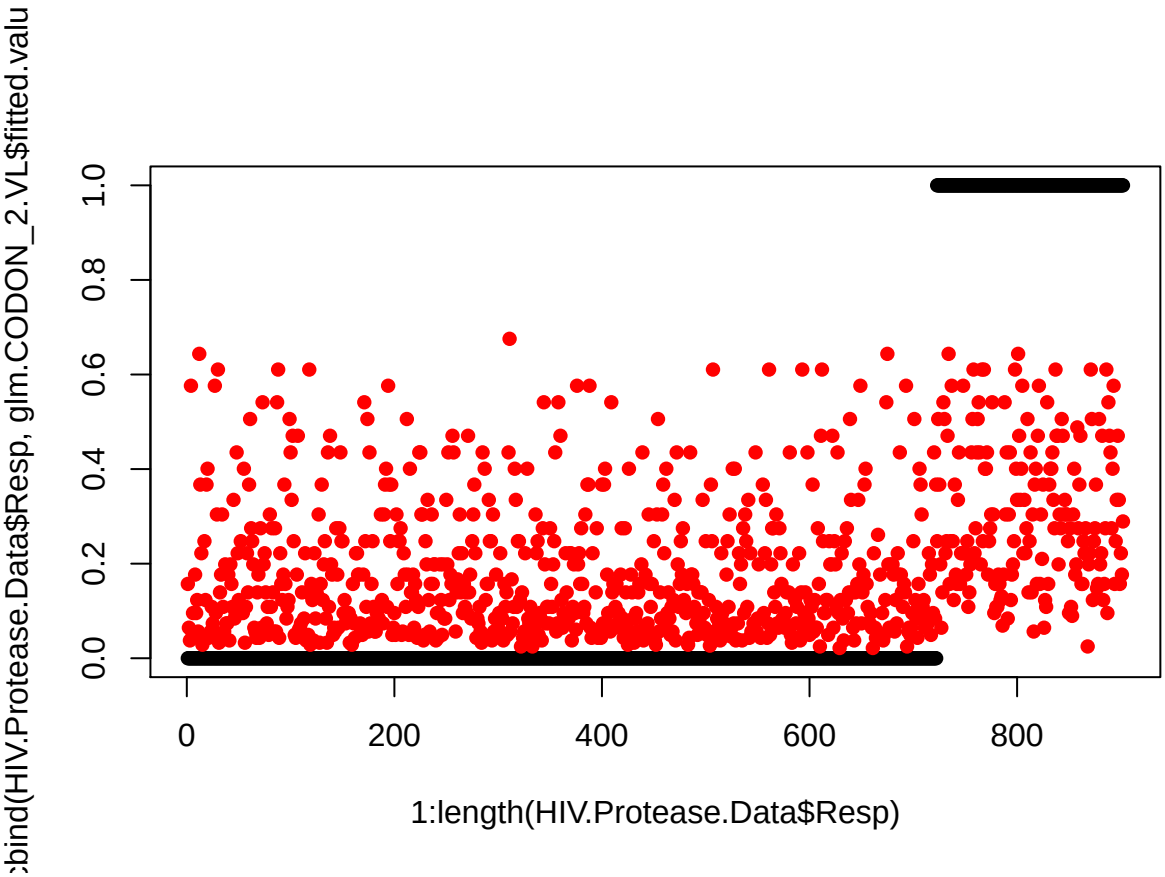
Fit logistic regression `Resp~VL`

```
glm.CODON_2.VL <- glm(Resp ~ VL, family = "binomial", data=HIV.Protease.Data)

AIC(glm.CODON_2.VL)
```

```
## [1] 782.068
```

```
matplot(1:length(HIV.Protease.Data$Resp),cbind(HIV.Protease.Data$Resp,glm.CODON_2.VL$fitted.values),pch=
```




```
head(cbind(HIV.Protease.Data$Resp,glm.CODON_2.VL$fitted.values))
```

```
##   [,1]       [,2]
## 1    0 0.15714185
## 2    0 0.06457383
## 3    0 0.03765410
## 4    0 0.57625950
## 5    0 0.05651215
## 6    0 0.09557494
```

Fit model with CODON_2 expression as predictor.

Compare the results with other models. Compare the predictions with the response variable. Plot predicted values and interpret them.

```
library(MASS)

glm.CODON_2.LV_2<-glmmPQL(Resp~VL,random=~1|CODON_2,family=binomial,data=HIV.Protease.Data)
```

```
## Loading required package: nlme
##
## Attaching package: 'nlme'
##
## The following object is masked from 'package:lme4':
##
```

```
##     lmList
##
## iteration 1
## iteration 2
## iteration 3
## iteration 4
## iteration 5
```

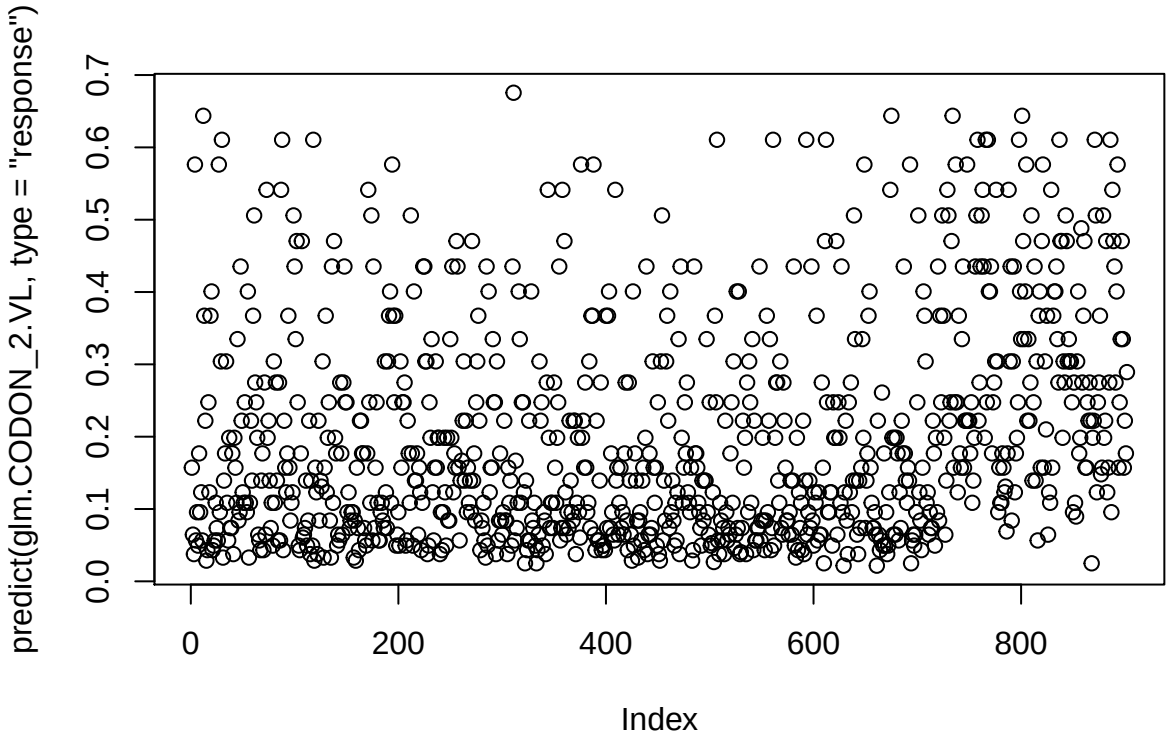
```
AIC(glm.CODON_2.LV_2)
```

```
## [1] NA
```

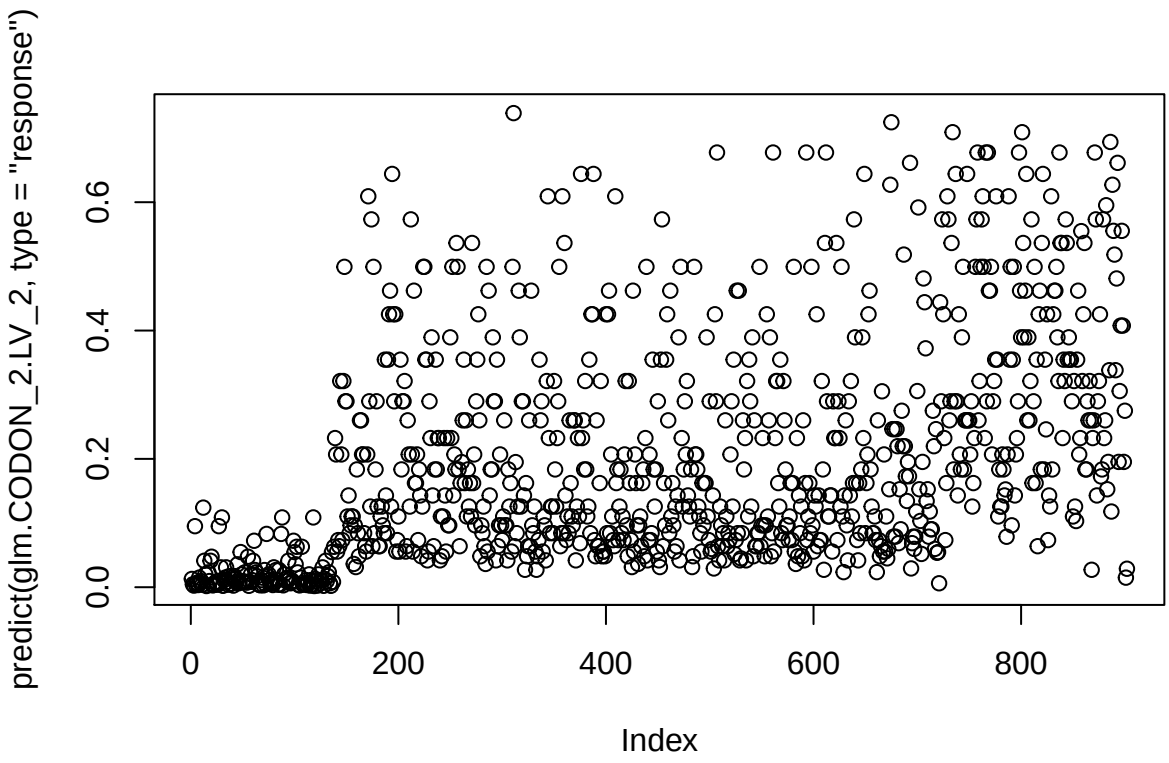
```
names(glm.CODON_2.LV_2)
```

```
##  [1] "modelStruct"  "dims"         "contrasts"    "coefficients"
##  [5] "varFix"       "sigma"        "apVar"        "logLik"
##  [9] "numIter"      "groups"       "call"         "terms"
## [13] "method"       "fitted"       "residuals"    "fixDF"
## [17] "na.action"    "data"         "family"
```

```
plot(predict(glm.CODON_2.VL,type="response"))
```



```
plot(predict(glm.CODON_2.LV_2,type="response"))
```




```
head(cbind(HIV.Protease.Data$Resp,predict(glm.CODON_2.LV_2,type="response")))
```

```
##     [,1]        [,2]
## caa    0 0.012893407
## caa    0 0.004582969
## caa    0 0.002530831
## caa    0 0.095129182
## caa    0 0.003951190
## caa    0 0.007146782
```

Although the mixed model using CODON_2 as a random variable does not produce an AIC value, by observing the plot of the responses from the two models above we a can see that they are distinctly different and the latter plot (with CODON_2 as a mixed variable) produces better results.