AMA-1 Technical Report: Concentration of protein in blood

Danyu Zhang, Daniel Alonso

February 10th, 2021

Introduction

Our objective is to determine which SNPs are related to concentration. SNPs are specific sections of genetic code transcribed to string format (i.e. *AGTGCTGATCGA*).

Here we have a preview of the dataset:

Table 1: Genetic data

id	conc	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8
id1	33.49489	2	3	3	1	2	1	3	3
id2	30.53090	1	2	3	2	1	2	2	2
id3	31.60567	2	1	2	3	1	3	3	3
id4	20.90570	1	1	3	3	3	3	2	2
id5	32.03528	2	3	1	1	2	2	2	2
id6	29.79114	3	1	2	1	2	2	3	1

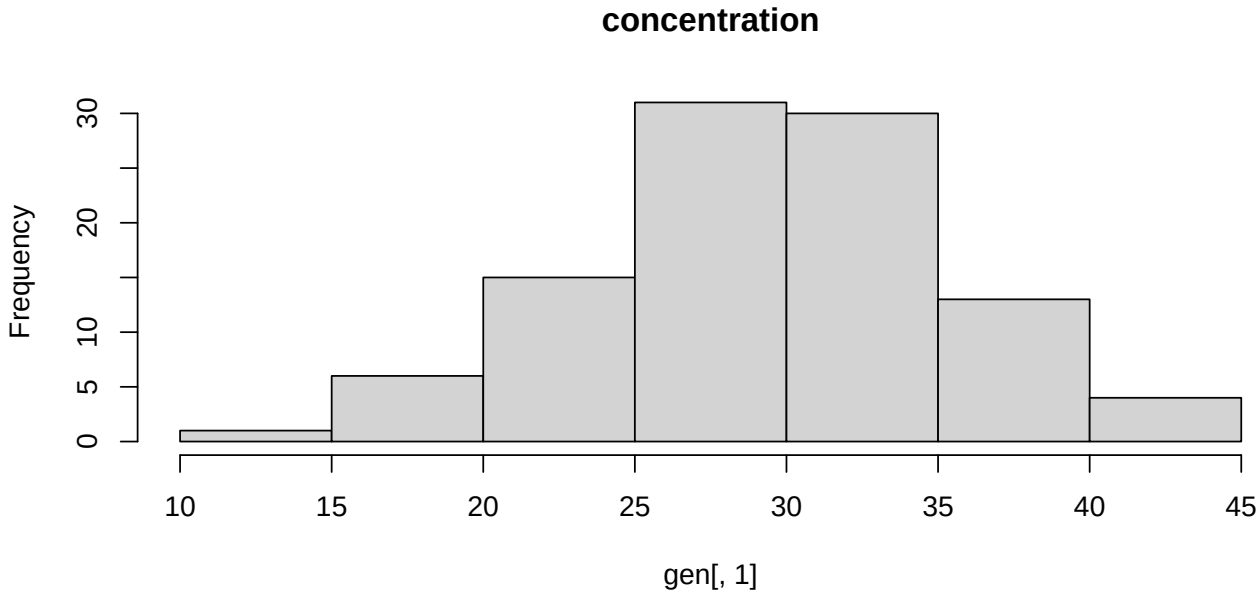
The dataset contains the following variables:

- **id**: refers to the ID of the individual who's genetic sequence was analyzed
- **conc**: concentration
- **snp1** through **snp8**: specific sections of the genetic code analyzed (SNPs)

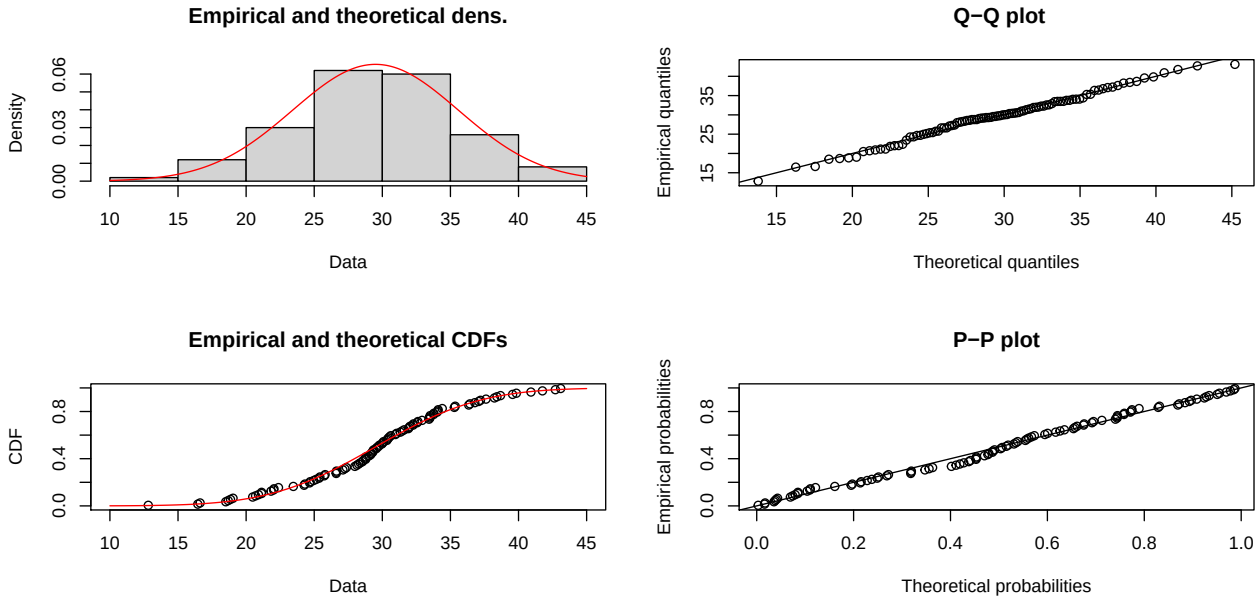
Basic exploratory data anaysis

We will take a look at the data graphically in a very simple manner:

Distribution of *concentration*

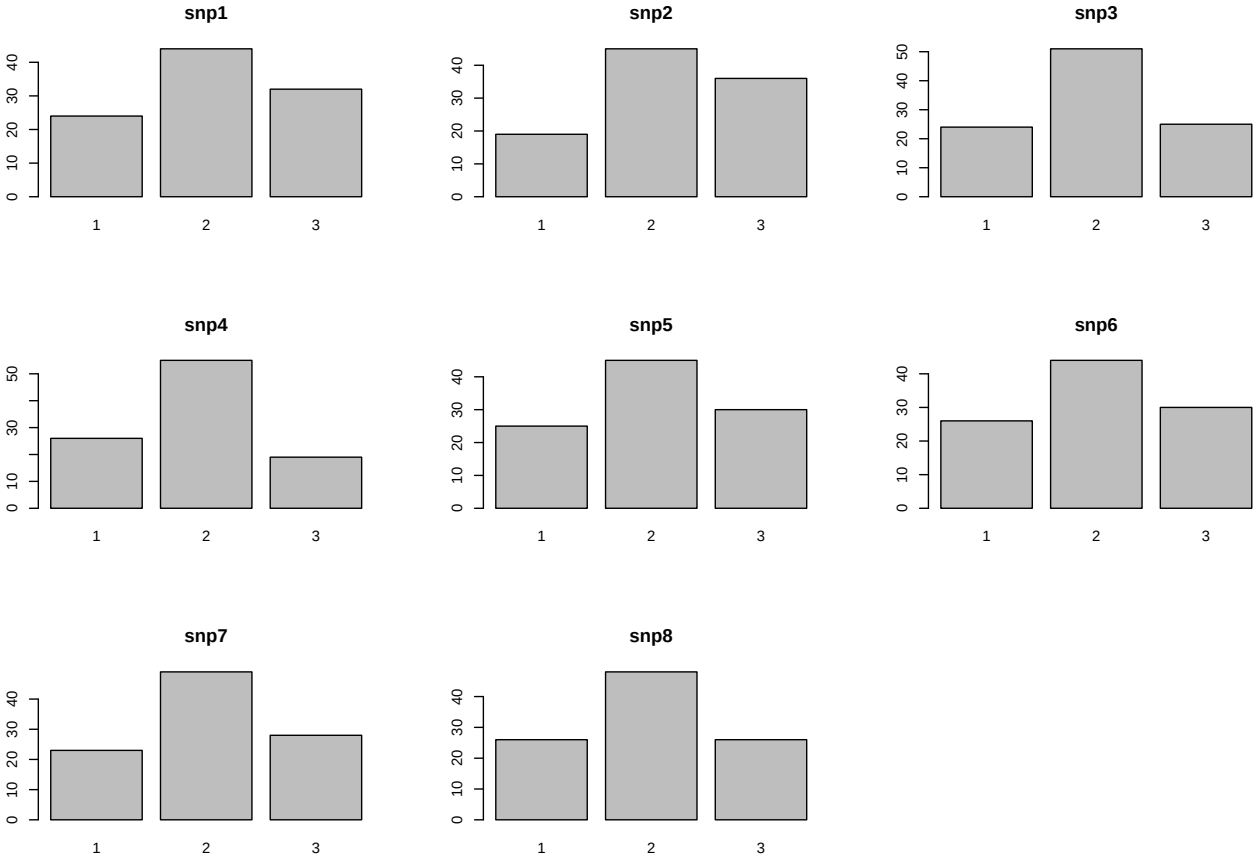


We can see that concentration approaches a normal distribution. It would be reasonable to attempt to fit it to a normal distribution:



And we can see that yes, we can somewhat comfortably say that the concentration comes from a normal distribution.

Most dominant SNP values



As for the SNPs, we can see that the most dominant value for the SNPs is **2**, while for *snp1* and *snp2* the value **3** seems to be slightly more common than the value **1**.

For the rest of the SNPs, **2** is the most common value, while **1** and **3** seem quite even.

Method

We are going to use multiple regression using as predictors all other variables minus *conc* in order to predict our target variable which is *conc*, we expect to obtain numerical results as our problem is to predict concentration.

We can see that only the first and the second variables are significant by checking the p-values, additionaly, the mean squared error is around 26, which is acceptable.

```
#>
#> Call:
#> lm(formula = conc ~ ., data = trainset)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -12.5477  -4.1447   0.6243   4.3540  11.8014
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 31.88895    6.16232   5.175 2.69e-06 ***
#> snp1         3.22962    0.97608   3.309  0.00158 **
#> snp2         1.94927    1.02505   1.902  0.06194 .
#> snp3        -1.42215    0.97374  -1.461  0.14928
#> snp4        -1.90791    1.04548  -1.825  0.07291 .
#> snp5        -1.93077    1.04101  -1.855  0.06847 .
#> snp6         0.08094    0.93123   0.087  0.93102
#> snp7        -1.61372    1.10379  -1.462  0.14888
#> snp8         0.33321    0.95093   0.350  0.72724
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 5.678 on 61 degrees of freedom
#> Multiple R-squared:  0.3416, Adjusted R-squared:  0.2552
#> F-statistic: 3.955 on 8 and 61 DF,  p-value: 0.0008014
#> [1] 26.80507
```

Results

Our most important variables to forecast concentration of protein in blood of corresponding individual are the genetic profiles “snp1” and “snp2”, and all the other genetic profiles are useless when used to predict this concentration.

Additionaly, genetic profiles “snp1” and “snp2” cause positive effects on the concentration of protein of the individual. Which means that, increasing genetic profiles “snp1” by 1 unit causes an increment on the concentration by 25 units; and 1 unit of increment on “snp2” produces 3.27 of increment on the concentration of protein in the blood.

The model works relatively well as it has mean squared error around 26. So we can use this model to predict the concentration of protein in the blood of corresponding individual.

Table 2: Variable importance for each SNP

	Overall
snp1	3.3087686
snp2	1.9016321
snp3	1.4605048
snp4	1.8249087
snp5	1.8547113
snp6	0.0869160
snp7	1.4619827
snp8	0.3504041

Bibliography

Data obtained from Stefano Cabras

Appendix: Code

Here goes the *literate programming* part.

```
# importing libraries
library(dplyr)
library(MLmetrics)
library(caret)
library(MASS)
library(stringr)
library(ggplot2)
library(fitdistrplus)

# importing data
gen <- read.csv("./data/gendata.csv", header=TRUE, sep=";")
knitr::kable(
    head(gen),
    booktabs=TRUE,
    caption="Genetic data",
)
gen <- gen[2:length(gen)]
cols <- length(names(gen))
gen <- sapply(gen, as.numeric)

# concentration histogram
par(mfrow=c(1,1))
hist(gen[,1], main="concentration")

# SNP plots
par(mfrow=c(3,3))
for (i in 2:cols) {
    barplot(table(gen[,i]), main=str_interp("snp${i-1}"))
}

# converting gen to dataframe
gen <- as.data.frame(gen)

# Splitting dataset into train and test
n=nrow(gen)
set.seed(7)
trainset=(1:n)%in%sample(n,floor(n*0.7))
testset=!trainset
trainset=gen[trainset,]
testset=gen[testset,]

# Modelling
model1 <- lm(conc ~ ., data = trainset)
summary(model1)

# Prediction and Metrics
pred1 = predict(model1, testset)
MSE(pred1, testset$conc)

# variable importance plot
impo <- varImp(model1, scale=FALSE)
knitr::kable(
    impo,
    booktabs=TRUE,
    caption="Variable importance for each SNP",
)
```