PREDICTING BETA-CAROTENE CONTENT IN PLASMA

Project 2

Zaide Montes Dmytro Perepolkin

Group 1 Lund University
Lund, Sweden
https://lu.se

```r
#install.packages("corrgram")
library(tidyverse)
library(broom)
library(corrgram) # visualisation of correlations
library(lmtest)  # more linear regression tools
library(hrbrthemes) #ggplot styling
library(GGally) # Pair plot
library(ggplot2) # ggplot 2
library(hrbrthemes) # theme for ggplot
library(kableExtra)
library(leaps)
library(glmnet)
library(patchwork)
library(plotmo)
ggplot2::theme_set(hrbrthemes::theme_ipsum_rc())

extrafont::loadfonts(device = "all", quiet = TRUE)

knitr::opts_chunk$set(dev = 'png')
options(device = function(file, width, height) {
  png(tempfile(), width = width, height = height)
})
```

1 Introduction

1.1 Data description

The dataset is from `gamlss.data` package [@harrell2002PlasmaRetinolBetaCarotene]. It is a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations.

An original data frame has 315 observations of the following variables

Variable	Description
`age`	age(years)
`sex`	sex, 1=male, 2=female
`smokstat`	smoking status 1=never, 2=former, 3=current Smoker
`bmi`	body mass index weight/(height^2)
`vituse`	vitamin use 1=yes, fairly often, 2=yes, not often, 3=no
`calories`	number of calories consumed per day
`fat`	grams of fat consumed per day
`fiber`	grams of fiber consumed per day
`alcohol`	number of alcoholic drinks consumed per week
`cholesterol`	cholesterol consumed (mg per day)
`betadiet`	dietary beta-carotene consumed (mcg per day)
`retdiet`[1]	dietary retinol consumed (mcg per day)
`betaplasma`	plasma beta-carotene (ng/ml)
`retplasma`[2]	plasma retinol (ng/ml)

We import the data

```r
plasma_df <- read_csv("data/plasma.txt", show_col_types = FALSE)
```

> Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer… We designed

[1] Not present in the current version of the dataset

[2] Not present in the current version of the dataset

a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. [@harrell2002PlasmaRetinolBetaCarotene]

2 Multivariate regression

glmnet requires a model-matrix as the model specification. A model matrix is a matrix with a column for each predictor (=x-variable) in the model (including the extra columns for polynomials, dummy variables etc), and each row is for a unique observation. Sometimes the model-matrix includes a column for the intercept, in glmnet it should (typically) not.

```
#head(plasma_df)
x <- model.matrix(log(betaplasma) ~ ., plasma_df)[, -1]
#head(x)
y <- log(plasma_df$betaplasma)
#head(y)
```

```
# Grid of lambdas
grid <- 10^seq(5, -2, length = 100) #grid

# Run ridge regression (alpha =0) for each lambda in grid
ridge.mod <- glmnet(x, y, alpha = 0, lambda = grid)
```

The `ridge.mod` contains 13 rows and 100 columns corresponding to the number of λ values we created earlier.

```
# Prepare for cross validation by splitting the data in training and validation set
set.seed(42)
train <- sample(1:nrow(x), nrow(x) / 2)
test <- (-train)
y.test <- y[test]
```

```
# Fit ridge regression on the training data, for our grid of lambda
ridge.mod <- glmnet(x[train, ], y[train], alpha = 0,
    lambda = grid, thresh = 1e-12)
# Find predictions for our testdata for a specific lambda (here lambda =4)
ridge.pred <- predict(ridge.mod, s = 4, newx = x[test, ])
```

The MSE for the $\lambda = 4$ is 0.4227268. Figure 2.1 shows the ridge model coefficients are approaching to zero.

```
#plot(ridge.mod,xvar="lambda",lwd=1.5)
plot_glmnet(ridge.mod,xvar="lambda",lwd=1.5)
```

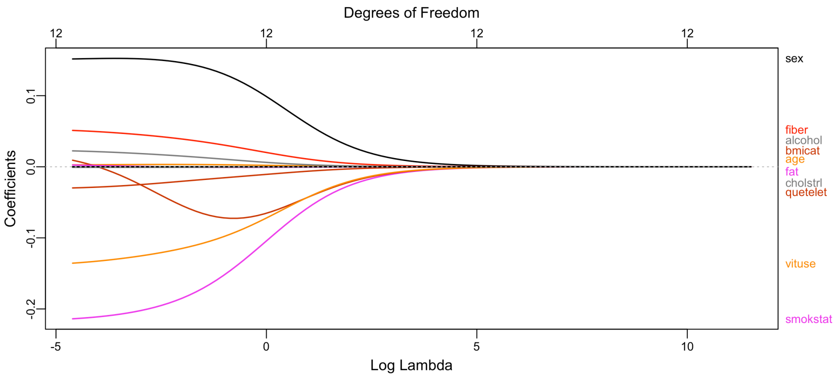

Figure 2.1: The ridge regression coefficients are displayed for the Plasma data set, as a function of .

```
set.seed(1)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 0)
plot(cv.out)
```

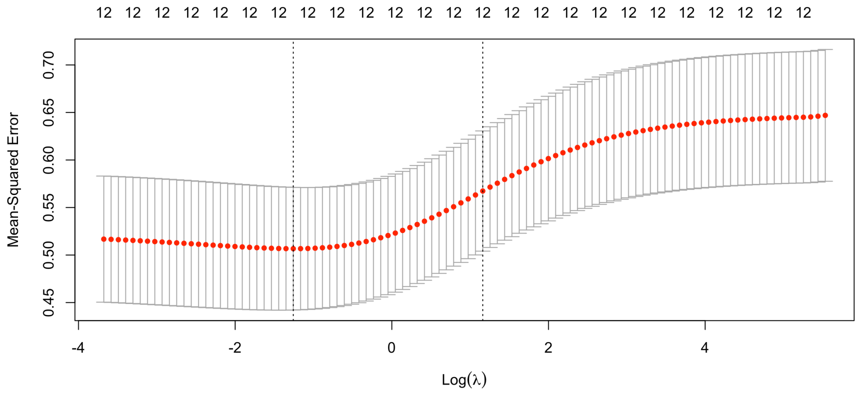

Figure 2.2: A graph of the ridge model coefficients for different lambdas. The dashed lines are the log values corresponding to the min (left dashed line) and 1 (right dashed line).

```
# Choose the best:
bestlam <- cv.out$lambda.min
```

The best lambda is $\lambda = 0.2841768$.

```
# MSE associated with this lambda
ridge.pred <- predict(ridge.mod, s = bestlam,
    newx = x[test, ])
mean((ridge.pred - y.test)^2)
```

```
[1] 0.4320415
```

The R-squared turns out to be 0.4319724. That is, the best model which explains 34.19% of the variation in the response values of the training data.

```
# Fit lasso model: alpha=1
# Use only the training data
lasso.mod <- glmnet(x[train, ], y[train], alpha = 1,
    lambda = grid)
```

```
plot_glmnet(lasso.mod,label=TRUE,xvar="lambda",lwd=1.5)
```

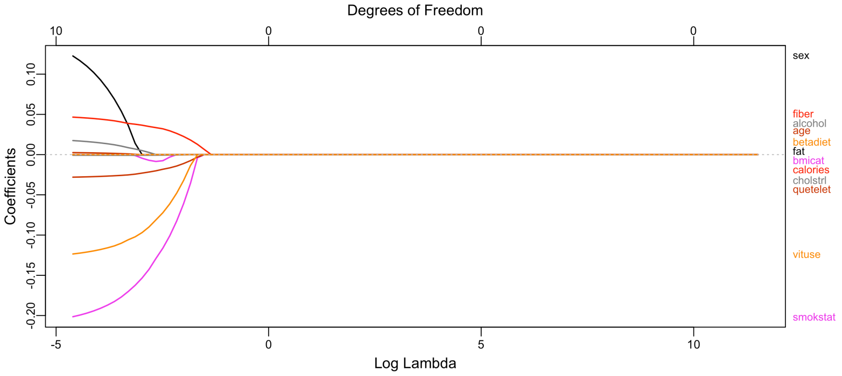

Figure 2.3: The Lasso regression coefficients are displayed for the Plasma data set, as a function of .

```
plot_glmnet(lasso.mod,label=TRUE,lwd=1.5)
```

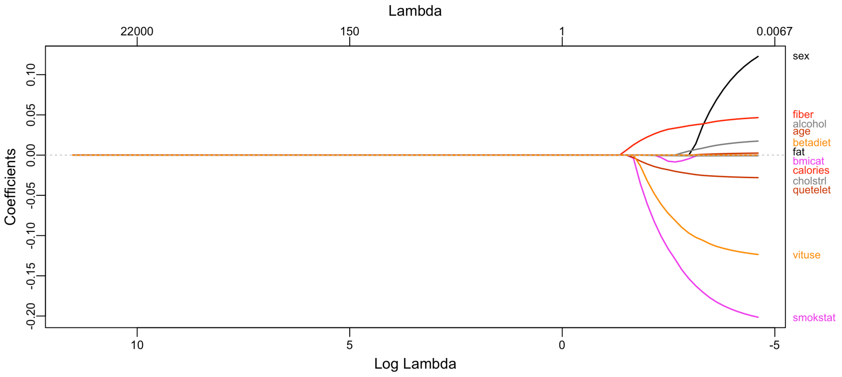

Figure 2.4: The Lasso regression coefficients are displayed for the Plasma data set, as a function of ˆ L 1/ ˆ 1.

```
set.seed(1)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 1)
plot(cv.out)
```

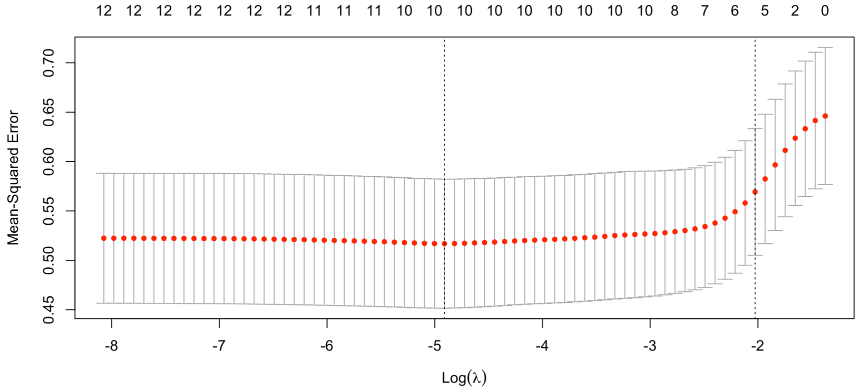

Figure 2.5: The figure illustrates the cross validation process for picking the best lambda in lasso regression. The dashed lines are the log values corresponding to the min (left dashed line) and 1 (right dashed line).

```
bestlam2 <- cv.out$lambda.min
```

The best lambda value that minimizes the test MSE turns out to be $\lambda = 0.0073745$.

```
# compute the test error for this choice of lambda or MSE associated with this lambda
lasso.pred <- predict(lasso.mod, s = bestlam2,
    newx = x[test, ])
mean((lasso.pred - y.test)^2)
```

```
[1] 0.4807286
```

The R-squared turns out to be 0.4319724. That is, the best model which explains 34.19% of the variation in the response values of the training data.

```
# Now fit on full data using the lambda selected by cross-validation
out <- glmnet(x, y, alpha = 1, lambda = grid)

lasso.coef <- predict(out, type = "coefficients",
    s = bestlam2)[1:13, ]
lasso.coef
```

```
  (Intercept)           age           sex      smokstat      quetelet 
 5.357977e+00  4.948625e-03  2.062469e-01 -1.192238e-01 -2.530994e-02 
       bmicat        vituse      calories           fat         fiber 
-4.068202e-02 -1.332267e-01 -4.699483e-05 -6.083977e-04  2.355829e-02 
      alcohol   cholesterol      betadiet 
 8.192186e-04 -3.090258e-04  4.443765e-05 
```

```
lasso.coef[lasso.coef != 0]
```

```
  (Intercept)           age           sex      smokstat      quetelet 
 5.357977e+00  4.948625e-03  2.062469e-01 -1.192238e-01 -2.530994e-02 
       bmicat        vituse      calories           fat         fiber 
-4.068202e-02 -1.332267e-01 -4.699483e-05 -6.083977e-04  2.355829e-02 
      alcohol   cholesterol      betadiet 
 8.192186e-04 -3.090258e-04  4.443765e-05 
```

The variance of ridge regression is slightly lower than the variance of the lasso. Consequently, the minimum MSE of ridge regression is slightly smaller than that of the lasso. In addition, the models generated from the lasso were much easier to interpret than those produced by ridge regression. The lasso yields sparse models that involve only a subset of the variables. Hence, depending on the value of , the lasso can produce a model involving any number of variables. In contrast, ridge regression will always include all of the variables in the model, although the magnitude of the coefficient estimates will depend on . In the lasso regression we observed that the variables that influenced more in the model are the alcohol, fiber and fat (Figurex). To sum up, the lasso should perform better in a setting where a relatively small number of predictors have significant coefficients, and the remaining predictors have very small coefficients or that equal zero. In contrast, the ridge regression will perform better when the response is a function of many predictors, all with roughly equal-sized coefficients.

```
#Principal components regression
```

3 References