Compulsory exercise: Team SuperGreat

MA8701 Advanced Statistical Learning V2023

Nora Aasen, Elias Angelsen, Jonas Nordstrom

23 mars, 2023

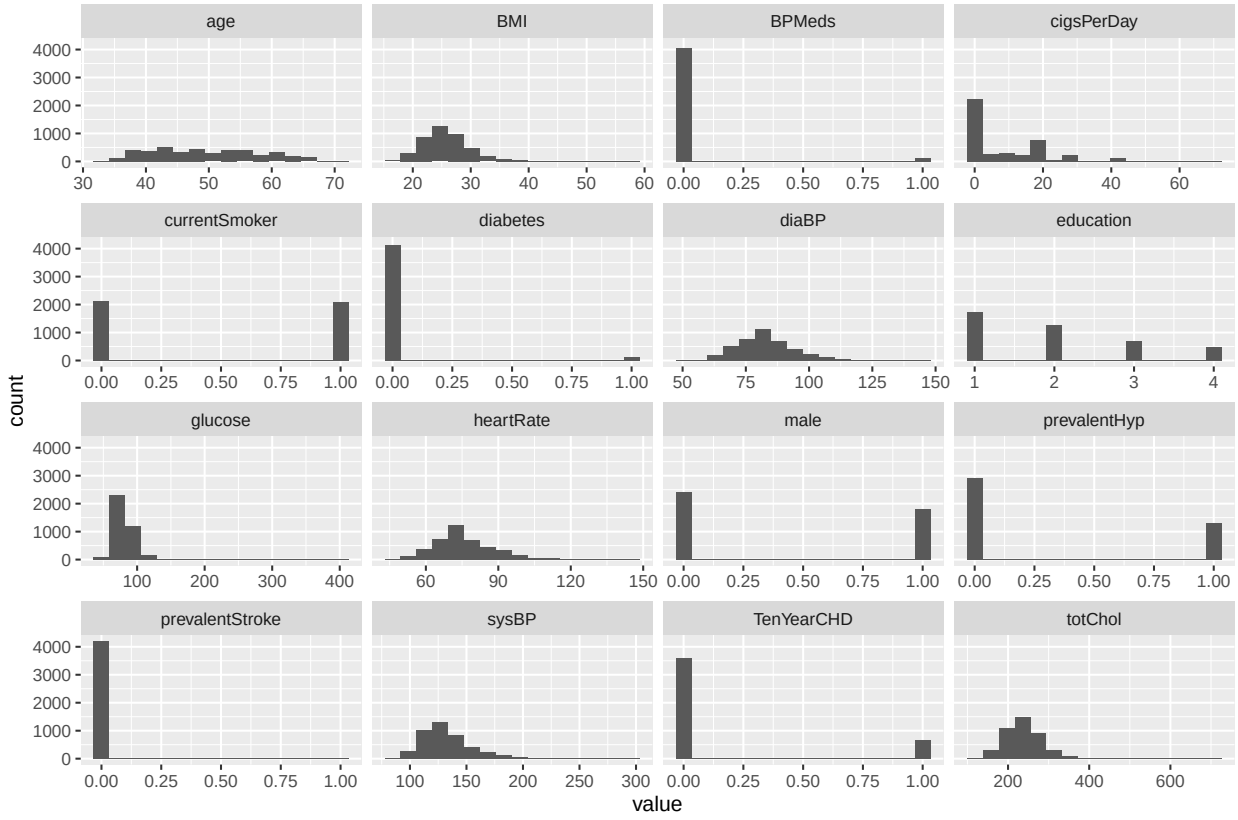
Introduction

In this project we have studied the Framingham Coronary Heart Disease Dataset. This dataset contains patient information for inhabitants in Framingham, Massachusetts, and is typically used to predict the chance of getting coronary heart disease (CHD) within the next 10 years. For this project, however, we intend to use lasso to find the most important risk factors. A big part of the task is to handle missing data. We will do single regression imputation manually and through the `mice` package, and investigate a bit what the Lasso is doing on the imputed data sets, compared to the complete case.

Exploratory Analysis

We start by examining the data set.

```
# Load and look at data
# data <- read.csv("C:\\Users\\noraa\\Student\\5thYear\\MA8701\\data_analysis_proj\\Heart_disease_analy
data <- read.csv("framingham.csv")
data_dim = dim(data)
pos_response = sum(data$TenYearCHD==1)

library(ggplot2)
library(tidyr) # for function 'gather'

# We visualize the data
ggplot(gather(data), aes(value)) +
    geom_histogram(bins = 16) +
    facet_wrap(~key, scales = 'free_x')
```

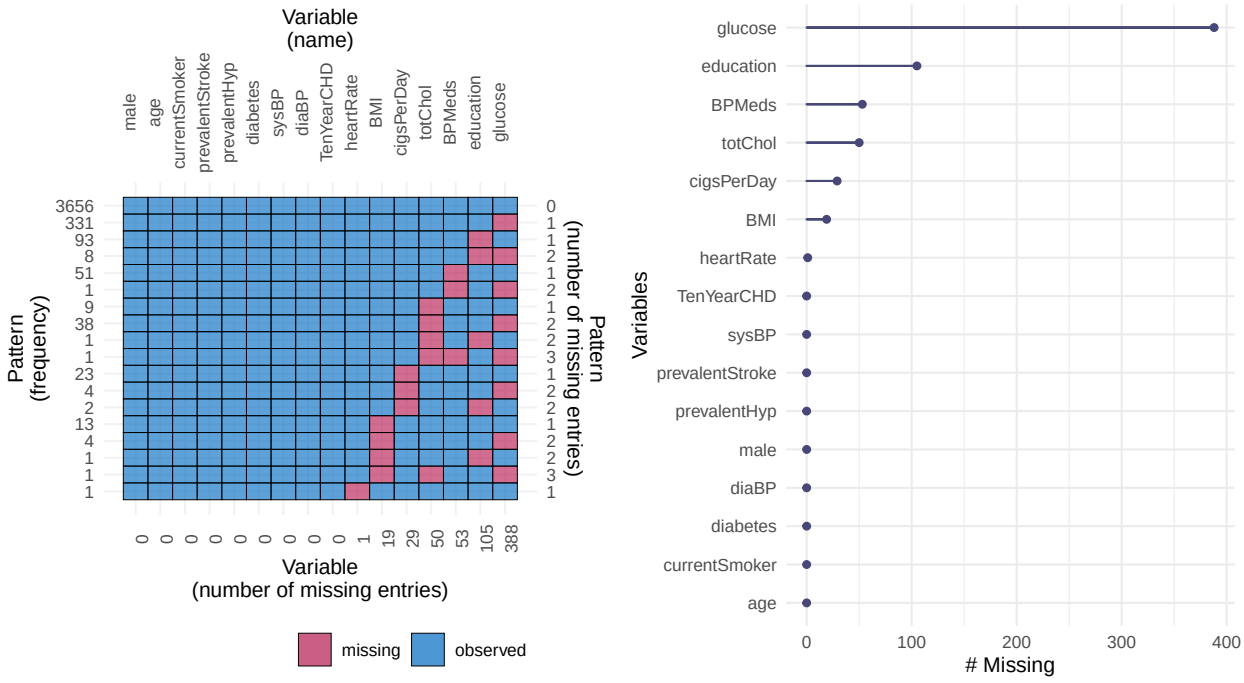



```
# Code education as a factor variable instead of 1-2-3-4.
data$education = factor(data$education, labels = c("none","hs","college","post-grad"))
```

This data set contains 4238 observations, 15 covariates and a binary response variable `TenYearCHD`. We will try to fit a logistic regression model. The response variable has 644 observations that are 1, which equals about 15.2% of the total observations. Most of our covariates are either binary, or numeric. However, we notice that the variable education is most likely a categorical covariate. We could not find any further elaboration for which four categories the numbers represent, so based on the frequency of each value and qualified guessing, we changed it to a factor variable and defined the four categories as none, hs, college, post-grad.

The next thing we looked at was the number of missing data in our data set.

```
# Look at the missing data
library(mice)
library(ggmice)
library(gridExtra)
plot1 <- plot_pattern(data, rotate = T)
plot2 <- gg_miss_var(data)
grid.arrange(plot1, plot2, ncol=2)
```



As we can see there are six covariates that has missing data: `glucose`, `education`, `BPMeds`, `totChol`, `cigsPerDay`, and `BMI`. We cannot use the rows that contain missing values as is. The easiest solution is to remove all rows that contains `NA`'s. This is the *complete case* solution. We split the data into training and test sets, as well as copying the complete case data set for later solutions.

```
# Split into training and test first to avoid data leakage
set.seed(8701)
tr = 7/10 # train ratio
r = dim(data)[1]
size = round(r*tr)
train = sample(1:r,size = size)

d.train = data[train,]
d.test = data[-train,]

# Make a dataset containing only the complete cases
d.complete <- data[complete.cases(data), ]
d.train.complete <- d.train[complete.cases(d.train), ]
d.test.complete <- d.test[complete.cases(d.test), ]

pos_response_c = sum(data[complete.cases(data),]$TenYearCHD==1)
```

The complete data set contains 3656 observations and the response variable has 557 observations that are 1, which equals about 15.2% of the total observations. As we can see, the proportion of positive observations in the response is the same, which is a good indicator that our data is missing at random (MAR), as we will discuss later.

Missing Data

We start by recalling that there are several types of mechanisms for missing data. Let $Z = (X, y)$ denote the full collection of covariates and responses, respectively, and we let a subscript mis/obs indicate whether we are restricting Z (or X) to the missing or observed parts, respectively. We may form an indicator (0-1) matrix R indicating missing (0) and observed (1) covariates. Assume ψ is short for the parameters in the distribution of R.

The missing data may be characterized by the conditioning in the distribution of R. We define the data to be:

- missing completely at random (MCAR) if $P(R|Z, \psi) = P(R|\psi)$,
- missing at random (MAR) if $P(R|Z, \psi) = P(R|Z_{obs}, \psi)$,
- missing not at random (MNAR) if $P(R|Z, \psi) = P(R|Z, \psi)$ (i.e. we don't have MCAR or MAR).

```
x = which(colSums(is.na(data)) > 0)

M = matrix(nrow=2, ncol = length(x)+1)
M[1,] = c(colSums(is.na(data))[x], NA)

for (i in 1:length(x)){
  r = is.na(data[,x[i]])
  df = data[r,]
  M[2,i] = round(sum(df$TenYearCHD)/length(df$TenYearCHD),3)
}

M[2,length(x)+1] = round(sum(data$TenYearCHD)/length(data$TenYearCHD),3)


colnames(M) = c(colnames(data)[colSums(is.na(data)) > 0],"full")
rownames(M) = c("# miss", "response freq")
as.table(M)
```

```
##               education cigsPerDay  BPMeds totChol     BMI heartRate glucose
## # miss          105.000     29.000  53.000  50.000  19.000     1.000 388.000
## response freq     0.152      0.069   0.208   0.180   0.526     1.000   0.129
##                  full
## # miss
## response freq   0.152
```

By exploring the missing pattern of for example the variable `cigsPerDay`, we see that our missing mechanism is not MCAR. No non-smoker has failed to answer the question ''How may cigarettes do you smoke a day?'', which is a question only aimed at smokers. The simple explanation may be that the survey automatically fills in 0 for `cigsPerDay` if you claim to be a non-smoker. In more mathematical terms, the missingness of `cigsPerDay` depends on the observed answer to''Do you smoke?" (found in variable `currentSmoker`), indicating that we do not work with MCAR data. Luckily, most methods are applicable if our missingness is at least MAR.

We will assume that the missing mechanism is MAR for all our missing observations, as there is no clear reason to suspect it to be MNAR.

To treat the missing data, we will use single imputation, as multiple imputation may cause difficulties with the resulting inference, as Rubin's rules needs to be combined with the Lasso, bootstrap and concluding inference. Multiple imputation was therefore not considered because it is beyond the scope of this project.

To perform single imputation we will use regression imputation, where we adapt our regression technique depending on the type of variable imputed. For continuous variables, we use a linear regression model, for binary variables, we use logistic regression, and for the variable ''education", which is a four-class variable, we have utilized kNN for multiclass imputation. This is implemented manually, but this could have been done using the `MICE` package and the function `mice`.

To avoid encountering observations with more than one missing value, and hence problems with regressing, we remove all rows containing more than one `NA`.

Our data is split into training and test sets, with the test-to-training ratio being 3:7. In order to avoid data leakage in our imputation of the test set, we fit the imputation models on the training set. The main idea is that the test set should be viewed as several independent observations. Using the test set to impute itself will use information not present at the time of training and will yield unintended correlation.

Mice uses polyreg for the factor variable and predictive mean matching (pmm) for all other.

We can use logreg for BPMeds - since it is binary -, polyreg for education and linear regression for all other columns with missing

```
set.seed(8701)
meth = mice(data, maxit = 0)$method[-16]  #Remove response
meth[which(meth == "pmm")] = "norm.predict"
meth["BPMeds"] <- "logreg"

# Important that we don't use the response as predictor when imputing: Only if the goal is prediction.
#y.train <- d.train$TenYearCHD
#y.test <- d.test$TenYearCHD

#d.train["TenYearCHD"] <- NULL
#d.test["TenYearCHD"] <- NULL

imp_mod.train = mice(d.train[-16], m=1, printFlag = F, method = meth)
d.train.imp = data.frame(complete(imp_mod.train), "TenYearCHD" = d.train[16])

# Use the imputation model trained on the training set on the test set, to ensure no data leakage.
imp_mod.test = mice.reuse(imp_mod.train, d.test[-16], printFlag = F)
d.test.imp = data.frame(imp_mod.test$"1", "TenYearCHD" = d.test[16])
```

Model

In the model section we will consider the two data sets; the complete case and imputed case. Both data sets are further divided into a train and test set. Since we want to do Lasso, we must standardize the data. The problem with this is data leakage. If we want to standardize the test data, we should standardize it using the mean and the standard deviation of the training data. Most importantly, using the test data to scale the test data will introduce correlation between the independent observations of the test set. Since the scaling information from the test set is "not available" to us at the time of training, we cannot expect the coefficients in the Lasso to be appropriately scaled compared to the test data. We solve this by scaling the training data, and then using the attributes of the training data to scale the test data accordingly.

```
# Make the training data ready for lasso by scaling.

set.seed(8701)

# Scale data for lasso
```

```
x.train.complete = scale(model.matrix(TenYearCHD ~ . -1, data = d.train.complete, family = binomial()))
train.complete.mean = attr(x.train.complete, "scaled:center")
train.complete.sd = attr(x.train.complete, "scaled:scale")
y.train.complete = d.train.complete$TenYearCHD

x.train.imp = scale(model.matrix(TenYearCHD ~ . -1, data = d.train.imp, family = binomial()))
train.imp.mean = attr(x.train.imp, "scaled:center")
train.imp.sd = attr(x.train.imp, "scaled:scale")
y.train.imp = d.train.imp$TenYearCHD

# Same for test data, but with training set attributes to avoid data leak: When do we use this?
x.test.imp = scale(model.matrix(TenYearCHD ~ . -1, data = d.test.imp, family = binomial()), center = tra
```

Given the binary response it is natural to consider fitting a logistic regression model to our data. Although we intend to use lasso, it is nice to start by fitting a regular logistic regression model on the complete case data to get an indication of which covariates that are most present, and for later comparison. We obtain the regression coefficients, a confusion matrix, a ROC-curve and the ROC-AUC.

```
# Fit a logistic model
set.seed(8701)
mod0 <- glm(TenYearCHD ~ ., data = d.train.complete, family = binomial())
round(summary(mod0)$coefficients,7)
```

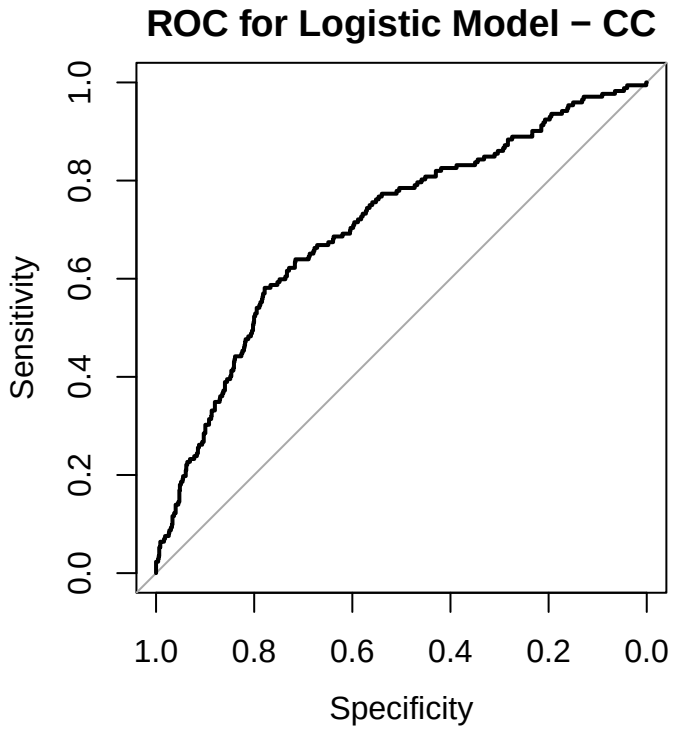
```
##                      Estimate Std. Error    z value  Pr(>|z|)
## (Intercept)        -8.1909517  0.8648439 -9.4710177 0.0000000
## male                0.6102028  0.1348701  4.5243748 0.0000061
## age                 0.0644352  0.0082290  7.8302194 0.0000000
## educationhs        -0.3167564  0.1532992 -2.0662624 0.0388037
## educationcollege   -0.2142688  0.1813577 -1.1814710 0.2374157
## educationpost-grad -0.1121211  0.1954944 -0.5735261 0.5662885
## currentSmoker      -0.0122323  0.1911538 -0.0639921 0.9489765
## cigsPerDay          0.0209125  0.0076803  2.7228564 0.0064720
## BPMeds              0.3232638  0.2692250  1.2007202 0.2298598
## prevalentStroke    -0.0717723  0.7060112 -0.1016589 0.9190275
## prevalentHyp        0.3578473  0.1679542  2.1306245 0.0331201
## diabetes           -0.1037130  0.3737342 -0.2775047 0.7813926
## totChol             0.0025062  0.0013865  1.8075291 0.0706798
## sysBP               0.0155640  0.0046170  3.3710343 0.0007489
## diaBP              -0.0051849  0.0078536 -0.6601879 0.5091332
## BMI                -0.0004882  0.0155704 -0.0313554 0.9749861
## heartRate          -0.0041223  0.0051729 -0.7968906 0.4255146
## glucose             0.0074510  0.0027470  2.7123748 0.0066803
```

```
mod0_preds <- predict(mod0, newdata = d.test.complete, type = "response")

predicted_value <- factor(round(mod0_preds))
expected_value <- factor(d.test.complete$TenYearCHD)

conf.mat.cc = confusionMatrix(data=predicted_value, reference = expected_value)$table
conf.mat.cc
```

```
##           Reference
```

```
## Prediction   0   1
##          0 917 161
##          1   9  11
```

```
roc_obj_cc <- roc(d.test.complete$TenYearCHD, mod0_preds, levels = c(0,1), direction = "<")
plot(roc_obj_cc, main = "ROC for Logistic Model - CC", cex = 0.5)
```




```
auc_cc = auc(roc_obj_cc)
auc_cc
```

```
## Area under the curve: 0.7054
```

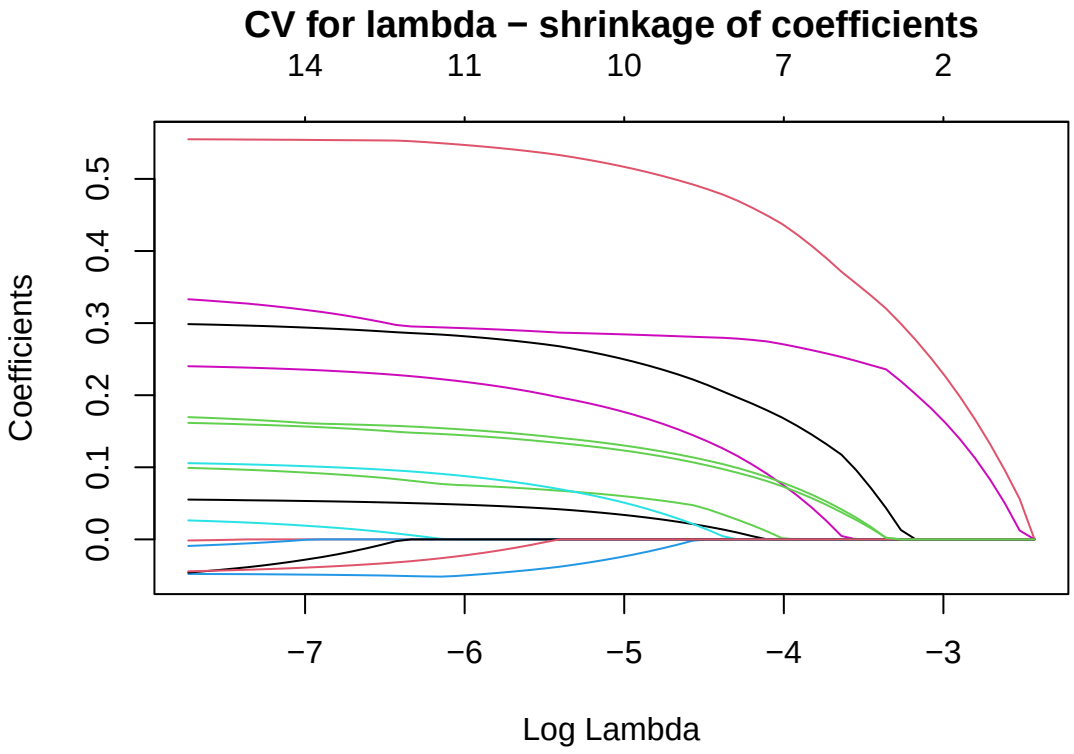
The logistic regression model chooses (Intercept), male, age, cigsPerDay, sysBP, glucose as the significant covariates, where the p-value cutoff is 0.01. It classifies very few positives correctly, which is very problematic if the model would be used to predict hearth disease.

Should there be 0.01 or 0.05 above?

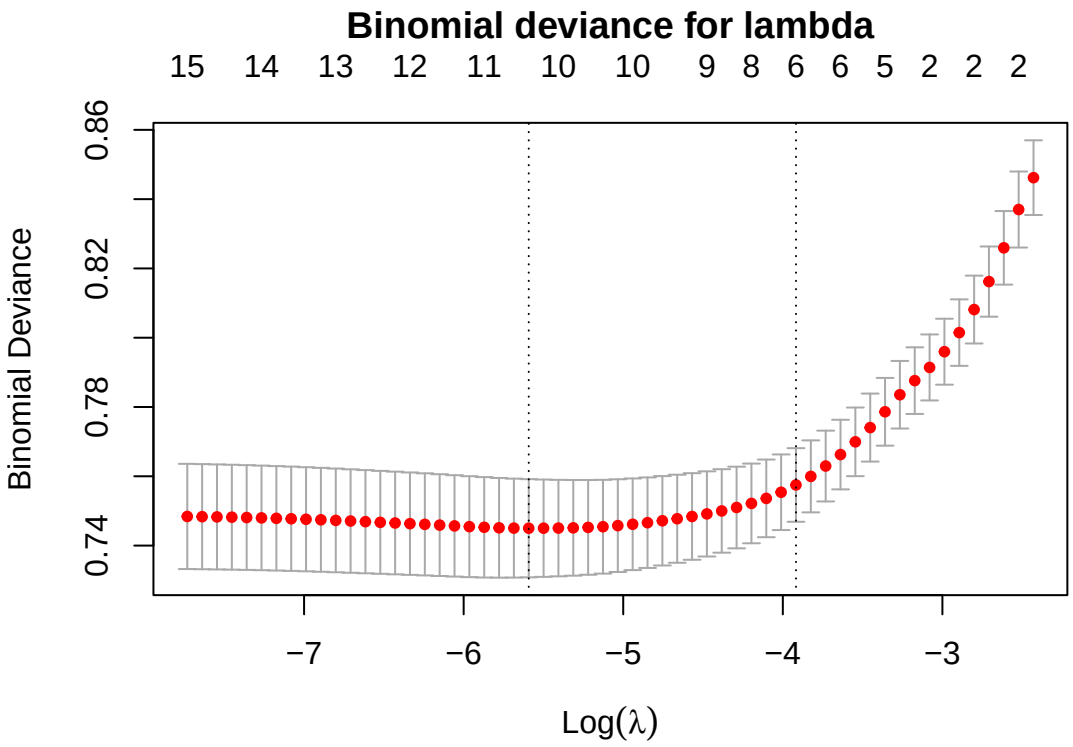
Lasso on Complete Case

We continue to do the Lasso on the complete case data. To do this, we use cross-validation to find λ_{min} and use the highest λ with deviance within one standard deviation of λ_{min}. We cross-validate for λ and plot the shrinkage and binomial deviance.

```r
#, fig.height = 4.3 , fig.width=6
set.seed(8701)

# Use cross-validation to find lambda (Should not standardize = T/F give the same answer?)
cv.out = cv.glmnet(x.train.complete, y.train.complete, family = "binomial",intercept = T, standardize=T
plot(cv.out$glmnet.fit, "lambda", label=F, main = c("CV for lambda - shrinkage of coefficients",""))
```



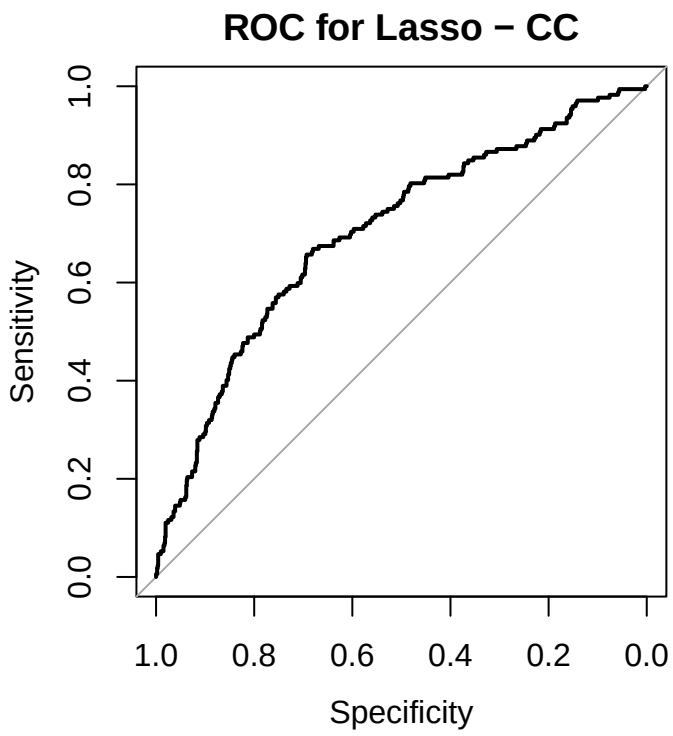

```r
plot(cv.out, main = c("Binomial deviance for lambda",""))
```




```
# Fit the Lasso model

lasso_mod_cc = glmnet(x.train.complete, y.train.complete, family = "binomial",intercept = T, standardiz

lasso_coef_cc <- coef(lasso_mod_cc,s=cv.out$lambda.1se)
```

The confusion table, ROC and ROC-AUC is given below.

```
##            Reference
## Prediction   0   1
##          0 924 170
##          1   2   2
```



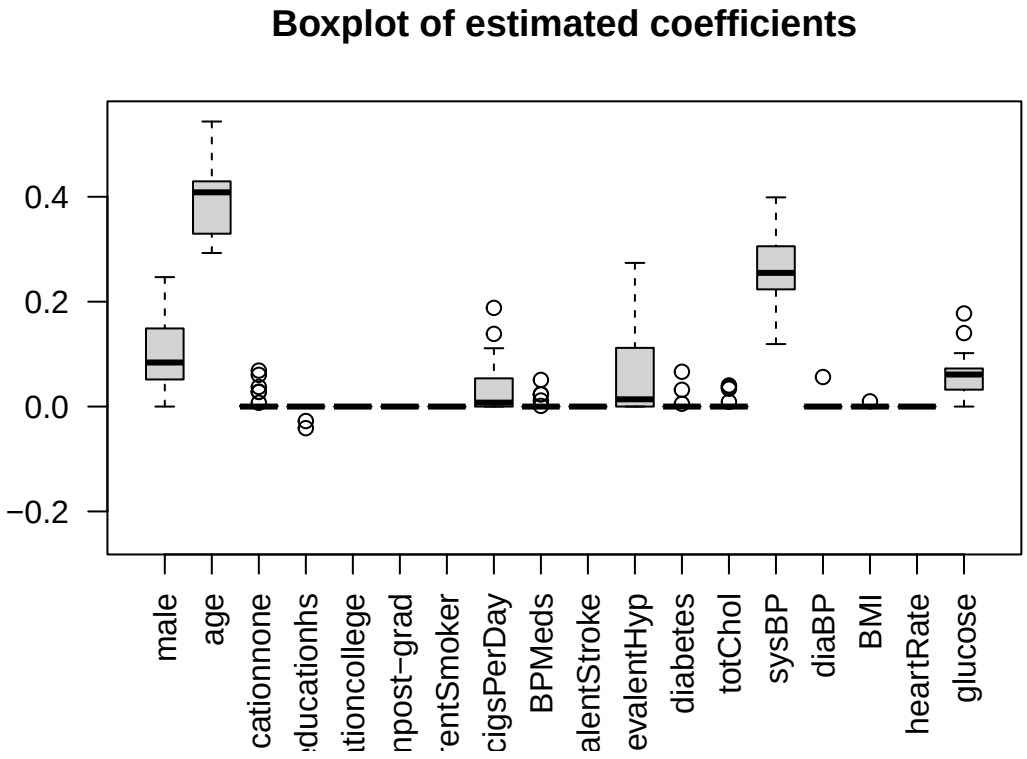

```
## Area under the curve: 0.7014
```

The Lasso on the complete case data chooses male, age, cigsPerDay, prevalentHyp, sysBP, glucose as the significant covariates. Initially, this is the same lasso model is much better at classifying positives than the full logistic model and has only slightly worse AUC, at 0.7013882 for the Lasso versus 0.705441 for the logistic model. We see that the Lasso includes `prevalentHyp`, while the logistic model finds it almost significant, with p-value 0.0331201. This agrees with the reasoning we made earlier, saying it could be an important parameter, but not as important as (either) `sysBP` or `diaBP`.

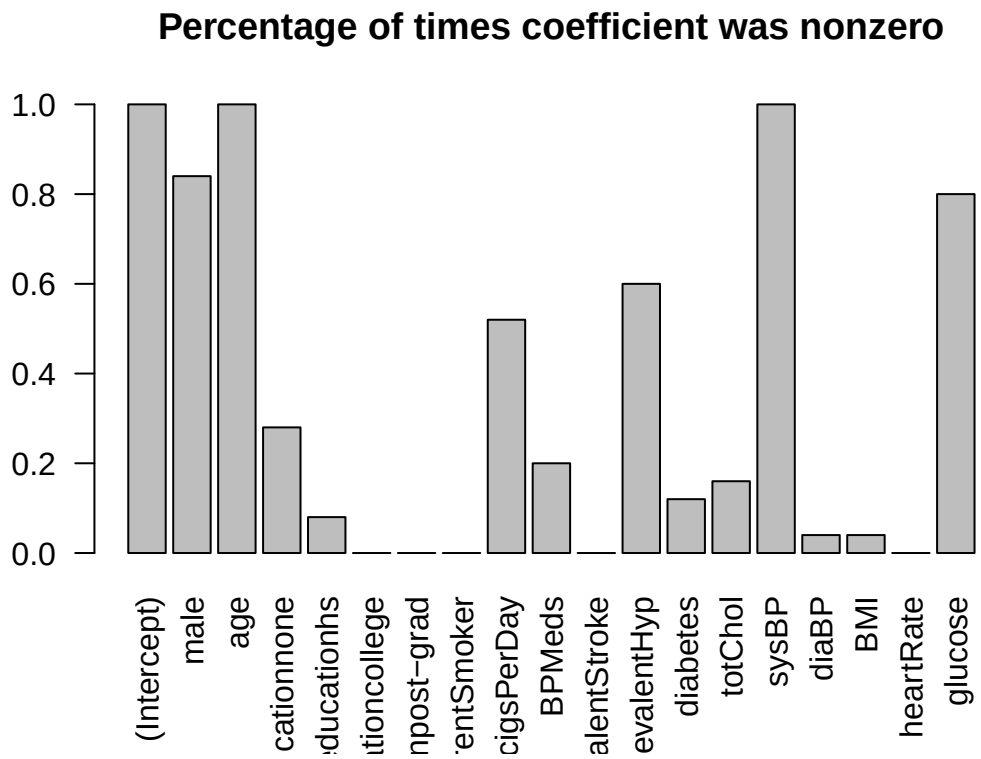
To obtain a better understanding of these coefficients, we bootstrap from the training data to fit Lasso models and store their coefficients.

```
set.seed(8701)

B = 25
boot_size_cc = dim(d.train.complete)[1]

B_coef_cc = matrix(NA, nrow = B, ncol = length(lasso_coef_cc[,1]))
for (i in 1:B){
  data_b = sample(1:boot_size_cc, size = boot_size_cc, replace = TRUE)
  x = x.train.complete[data_b,]
  y = y.train.complete[data_b]
  cv.out = cv.glmnet(x, y, family = "binomial", alpha = 1)
  lasso_mod = glmnet(x, y, family = "binomial", alpha = 1, intercept = T, lasso = cv.out$lambda.1se)

  B_coef_cc[i,] <- coef(lasso_mod,s=cv.out$lambda.1se)[,1]
}
```

```
colnames(B_coef_cc) = names(coef(lasso_mod,s=cv.out$lambda.1se)[,1])
boxplot.matrix(B_coef_cc[,-1], ylim = c(-0.25,0.55), las = 2, main = "Boxplot of estimated coefficients
```




```
B_coef_count_cc = ifelse(B_coef_cc == 0,0,1)
barplot(apply(B_coef_count_cc, 2, sum)/B, las = 2, main = "Percentage of times coefficient was nonzero"
```




```
names_lasso_cc_70 <- names(apply(B_coef_count_cc, 2, sum)/B)[apply(B_coef_count_cc, 2, sum)/B > 0.7]
names_lasso_cc_50 <- names(apply(B_coef_count_cc, 2, sum)/B)[apply(B_coef_count_cc, 2, sum)/B > 0.5]
```

We ran the bootstrap using 25 iterations. The variables that have nonzero coefficients in the Lasso models at least 70% of the time, are (Intercept), male, age, sysBP, glucose. Similarly, by those to those who are included at least 50% of the time, we obtain (Intercept), male, age, cigsPerDay, prevalentHyp, sysBP, glucose. This is indeed similar to the ones we picked out earlier.

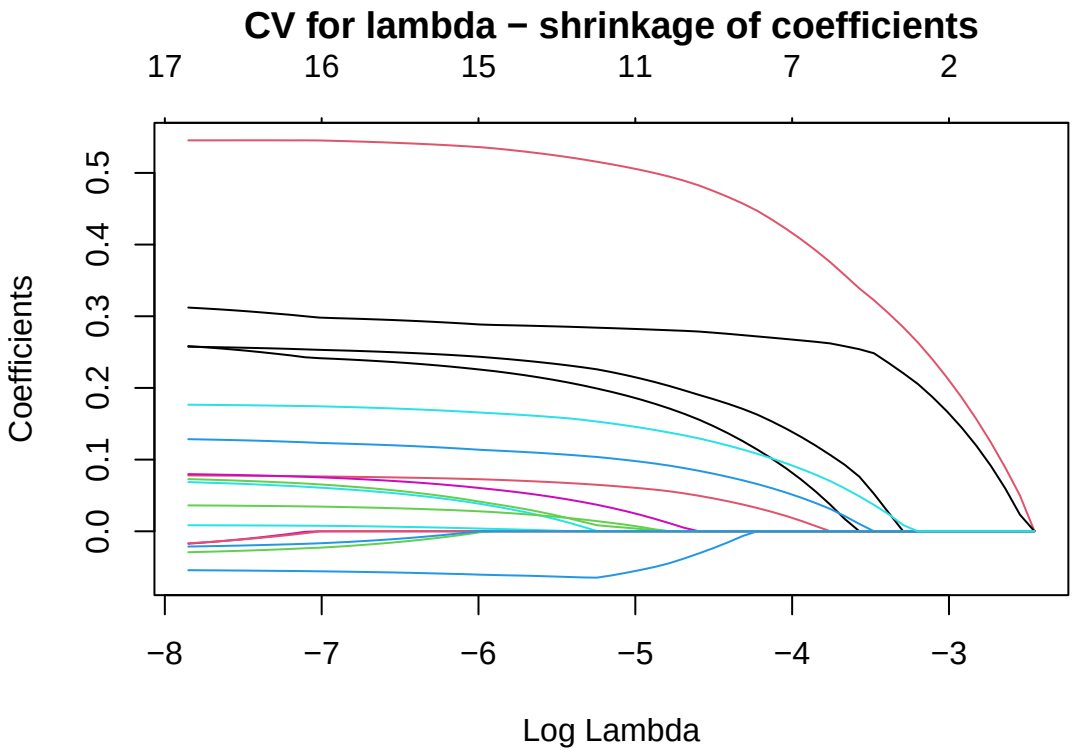
Lasso on Imputed Data

We now do the same thing, just using the imputed data instead of the complete case. Even though the imputed data includes more samples, the data quality is going down when we impute.

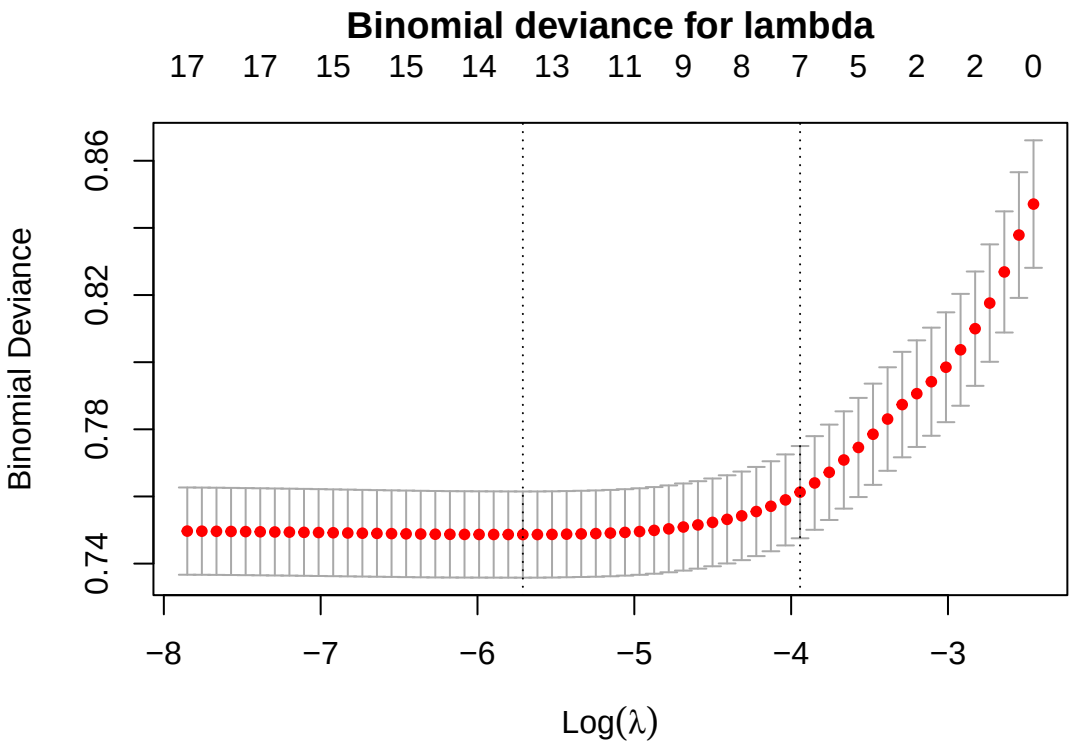
To get hands-on experience with the `MICE` package, we also construct an imputed data set using `mice` and `mice.reuse`, for comparison. `mice` can be used on the entire training data, without needing to remove those samples with two or more covariates missing. The imputation model from the training set is emplyed to impute the test set as well, to avoid data leakage.

First we try the Lasso on the imputed dataset where we imputed with our manual technique. We show the shrinkage and the binomial deviance over λ, which is a part of the cross-validation

```
# We need to standardize the matrix so that we can drop intercepts.
cv.out = cv.glmnet(x.train.imp, y.train.imp, family = "binomial", intercept = T, standardize=TRUE, alpha
plot(cv.out$glmnet.fit, "lambda", label=F, main = c("CV for lambda - shrinkage of coefficients",""))
```



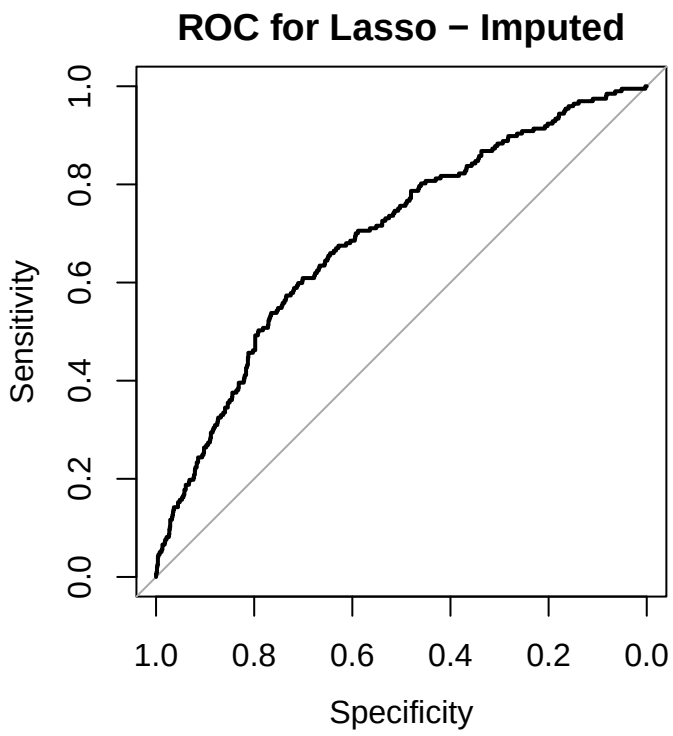

```
plot(cv.out, main = c("Binomial deviance for lambda",""))
```




```
lasso_mod_imp = glmnet(x.train.imp, y.train.imp, family = "binomial", alpha = 1, intercept = F, standar

lasso_coef_imp <- coef(lasso_mod,s=cv.out$lambda.1se)
```

We also give the confusion matrix, the ROC and the ROC-AUC.

```
lasso_preds_imp <- predict(lasso_mod_imp, newx = x.test.imp, type = "response",  s = cv.out$lambda.1se)

predicted_value_imp <- factor(round(lasso_preds_imp))
expected_value_imp <- factor(d.test.imp$TenYearCHD)

conf_mat_imp = confusionMatrix(data=predicted_value_imp, reference = expected_value_imp)$table


roc_obj_imp <- roc(d.test.imp$TenYearCHD, lasso_preds_imp, levels = c(0,1), direction = "<")
plot(roc_obj_imp, main = "ROC for Lasso - Imputed", cex = 0.5)
```




```
auc(roc_obj_imp)
```

```
## Area under the curve: 0.6915
```

```
conf_mat_imp
```

```
##           Reference
## Prediction   0   1
##          0 613  58
##          1 461 139
```

```
sens_imp = sensitivity(conf_mat_imp)
spes_imp = specificity(conf_mat_imp)
names_lasso_coef_imp = names(which(lasso_coef_imp[,1] > 0))
```

The performance of this model is quite similar to that of the complete case, which is to be expected. The Lasso on the manually imputed data chooses male, age, educationnone, cigsPerDay, diabetes, sysBP, glucose as the significant covariates.

Recall that the sensitivity can be measured by the true positive rate (i.e. the number of true positives over all positives) and the specificity can be measured by the true negative rate (i.e. true negative over all negative). Comparing the Lasso on the complete case data with the imputed data, we note that we have sensitivity 0.9978402 in the complete case and sensitivity 0.5707635 in the imputed case. The specificity of the complete case Lasso is 0.0116279, while for the imputed case it is 0.7055838.

Although the difference is not huge, it may resemble that the data quality in the imputed data is slightly lower, although for selecting covariates, we obtained the same answer. In our data rich situation, this

is neither clear enough to be rendered true, nor actually a problem, but for data poor situations, this is something to keep in mind.

We try to do the same thing, using the `MICE`-imputed data set, first plotting the shrinkage and binomial deviance over λ, and then give the confusion matrix, ROC and ROC-AUC.

The predictive performance has not changed a lot, in the eyes of AUC. The performance of this model is quite similar to that of the complete case, which is to be expected.

We suspect that the sensitivity and specificity is even lower for the `MICE`-imputed data, as it also imputes those with more than one missing value.

Inference

In order to do inference we simply fit a logistic regression model using the `glm` function in `R`, and extract the inference from there. However, we will keep the coefficients chosen by the lasso-bootstrapping iterations in the earlier models, and now use the test data to fit a logistic model to avoid overfitting. Note that we have only used the test data to predict and observe different measures, such as ROC-AUC, sensitivity/specificity, and so on. The test data is therefore suitable for inference, as it has not been perturbed in the procedure of fitting the models.

We start out with the complete case models and fit a logistic model with the most important variables, namely (Intercept), male, age, sysBP, glucose. We first state the coefficients of the new regression model and their confidence intervals.

```
# Fit the model on the complete case training data.

mod <- glm(TenYearCHD ~ age + male + sysBP + glucose, data = d.complete[-train,], family = binomial())

# Confidence intervals.

CI_mod = confint(mod)

CI_mod0 = confint(mod0)

# Look at the coefficients

summary(mod)$coefficients
```

```
##                Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept) -8.42663685 0.727151909 -11.588551 4.710346e-31
## age          0.04891710 0.011168331   4.379982 1.186889e-05
## male         0.85267388 0.181851299   4.688852 2.747416e-06
## sysBP        0.01998341 0.003974812   5.027511 4.968880e-07
## glucose      0.01273926 0.003233364   3.939941 8.150151e-05
```

The confidence intervals of these variables for the new regression model (left) and naive logistic model fit (right) on the complete data set is given by the following.

```
# Add the response values that we took out for the mice data.

d.test.imp["TenYearCHD"] <- d.test$TenYearCHD

# Fit the model.
```

```
mod <- glm(TenYearCHD ~ age + male + sysBP + cigsPerDay + glucose, data = d.test.imp, family = binomial

summary(mod)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ age + male + sysBP + cigsPerDay +
##     glucose, family = binomial(), data = d.test.imp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4654  -0.6065  -0.4698  -0.3371   2.6692
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.516048   0.689519 -10.900  < 2e-16 ***
## age          0.056879   0.010537   5.398 6.74e-08 ***
## male         0.382425   0.172078   2.222  0.02626 *
## sysBP        0.014085   0.003616   3.896 9.80e-05 ***
## cigsPerDay   0.022402   0.006722   3.333  0.00086 ***
## glucose      0.007291   0.002816   2.589  0.00963 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1096.3  on 1270  degrees of freedom
## Residual deviance: 1009.2  on 1265  degrees of freedom
## AIC: 1021.2
##
## Number of Fisher Scoring iterations: 5
```

```
# Consider confidence interval and coefficients
CI_mice = confint(mod)
cbind(as.matrix(round(summary(mod)$coefficients,5)), as.matrix(round(CI_mice,5)))
```

```
##             Estimate Std. Error   z value Pr(>|z|)    2.5 %   97.5 %
## (Intercept) -7.51605    0.68952 -10.90042  0.00000 -8.89560 -6.18929
## age          0.05688    0.01054   5.39786  0.00000  0.03637  0.07773
## male         0.38242    0.17208   2.22240  0.02626  0.04497  0.72034
## sysBP        0.01408    0.00362   3.89559  0.00010  0.00700  0.02121
## cigsPerDay   0.02240    0.00672   3.33275  0.00086  0.00916  0.03556
## glucose      0.00729    0.00282   2.58882  0.00963  0.00174  0.01297
```

Again, we can observe that all the coefficients in the new model are significant!

We take a brief look at the complete case model with Lasso-selected variables to the model on the imputed data with Lasso-selected variables. The intervals are for the model on the imputed data (left) and the model on the complete data (right).

```
cbind(as.matrix(round(CI_mice[c(1,2,3,4,6,5),],4)), rbind(as.matrix(round(CI_mod,4)),c("*","*")))
```

```
##             2.5 %     97.5 %    2.5 %     97.5 %
## (Intercept) "-8.8956" "-6.1893" "-9.8907" "-7.0368"
## age         "0.0364"  "0.0777"  "0.0272"  "0.071"
## male        "0.045"   "0.7203"  "0.4997"  "1.2136"
## sysBP       "0.007"   "0.0212"  "0.0123"  "0.0279"
## glucose     "0.0017"  "0.013"   "0.0067"  "0.0195"
## cigsPerDay  "0.0092"  "0.0356"  "*"       "*"
```

```
diff1 = CI_mod[3,2] - CI_mod[3,1]
diff2 = CI_mice[3,2] - CI_mice[3,1]
```

We can see, for example by considering `sysBP`, see that the width of the interval has gone further down by working on the imputed data, as the confidence interval width for `sysBP` was 0.7139159 for the complete case model and 0.6753695. This may simply be because we use more data to fit the model, but it may also be a nonsensical question, as we are in essence fitting two different models. The inclusion of `cigsPerDay` in the model on imputed data is probably a key reason why we see such a difference. More or less, we obtain the same results, as it is the same covariates that come back time and time again.

Discussion

When we compare the chosen coefficients from the complete case data compared to the imputed data, we see that they correspond. Thus, in a data-rich situation imputation may only introduce unnecessary variance, without having an immediate effect on the quality of the model or inference.

Another interesting thing we discovered, was that even though we obtain good results with the imputed data, the data quality seems to go down, if only barely. The sensitivity and specificity of the model went down slightly when using the imputed data, but as the Lasso selects the same variables, it does not matter a lot for our purposes. The decrease in data quality might have been more visible if the percentage of imputed values were higher, or if we were in a data-poor situation.

What we can probably conclude, is that the variables (Intercept), male, age, sysBP, glucose are the most significant, and that other important variables are `cigsPerDay` and `prevalentHyp`.

References

Bates, Douglas, Martin Maechler, and Mikael Jagan. 2022. *Matrix: Sparse and Dense Matrix Classes and Methods*. https://CRAN.R-project.org/package=Matrix.

Binder, Martin, Florian Pfisterer, Michel Lang, Lennart Schneider, Lars Kotthoff, and Bernd Bischl. 2021. "mlr3pipelines - Flexible Machine Learning Pipelines in r." *Journal of Machine Learning Research* 22 (184): 1–7. https://jmlr.org/papers/v22/21-0281.html.

Binder, Martin, Florian Pfisterer, Lennart Schneider, Bernd Bischl, Michel Lang, and Susanne Dandl. 2022. *Mlr3pipelines: Preprocessing Operators and Pipelines for Mlr3*. https://CRAN.R-project.org/package=mlr3pipelines.

Borowski, Jan, and Piotr Fic. 2022. *NADIA: NA Data Imputation Algorithms*. https://CRAN.R-project.org/package=NADIA.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Friedman, Jerome, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay, Noah Simon, and James Yang. 2022. *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. https://CRAN.R-project.org/package=glmnet.

Kassambara, Alboukadel. 2022. *Ggcorrplot: Visualization of a Correlation Matrix Using Ggplot2.* http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2.

Kuhn, Max. 2022. *Caret: Classification and Regression Training.* https://github.com/topepo/caret/.

Lang, Michel, Martin Binder, Jakob Richter, Patrick Schratz, Florian Pfisterer, Stefan Coors, Quay Au, Giuseppe Casalicchio, Lars Kotthoff, and Bernd Bischl. 2019. "mlr3: A Modern Object-Oriented Machine Learning Framework in R." *Journal of Open Source Software*, December. https://doi.org/10.21105/joss.01903.

Lang, Michel, Bernd Bischl, Jakob Richter, Patrick Schratz, Martin Binder, Florian Pfisterer, Raphael Sonabend, and Marc Becker. 2022. *Mlr3: Machine Learning in r - Next Generation.* https://CRAN.R-project.org/package=mlr3.

Lang, Michel, Bernd Bischl, Jakob Richter, Xudong Sun, and Martin Binder. 2022. *Paradox: Define and Work with Parameter Spaces for Complex Algorithms.* https://CRAN.R-project.org/package=paradox.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2022. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* https://CRAN.R-project.org/package=e1071.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for r and s+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12: 77.

———. 2021. *pROC: Display and Analyze ROC Curves.* http://expasy.org/tools/pROC/.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with r.* New York: Springer. http://lmdvr.r-forge.r-project.org.

———. 2021. *Lattice: Trellis Graphics for r.* http://lattice.r-forge.r-project.org/.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. https://doi.org/10.18637/jss.v039.i05.

Tierney, Nicholas, and Dianne Cook. 2023. "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." *Journal of Statistical Software* 105 (7): 1–31. https://doi.org/10.18637/jss.v105.i07.

Tierney, Nicholas, Di Cook, Miles McBain, and Colin Fay. 2023. *Naniar: Data Structures, Summaries, and Visualisations for Missing Data.* https://github.com/njtierney/naniar.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

———. 2022. *Mice: Multivariate Imputation by Chained Equations.* https://CRAN.R-project.org/package=mice.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2022. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Maximilian Girlich. 2022. *Tidyr: Tidy Messy Data.* https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr.* 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown.* Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/bookdown.

———. 2022a. *Bookdown: Authoring Books and Technical Documents with r Markdown.* https://CRAN.R-project.org/package=bookdown.

———. 2022b. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.