CBio-Project 2

Samuel Higgins

4/27/2020

Introduction

```r
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ----------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
liverd <- read.csv("https://github.com/zhamuelh/samprojects2/raw/master/Data%20sets/liverdrecords.csv")

liverd <- liverd %>% rename(Liver_Disease = Dataset) %>% rename(Total_Proteins = Total_Protiens) %>%
  mutate(Liver_Disease = ifelse(Liver_Disease == "1", 1, 0)) %>% na.omit

head(liverd)
```

```
##   Age Gender Total_Bilirubin Direct_Bilirubin Alkaline_Phosphotase
## 1  65 Female             0.7              0.1                  187
## 2  62   Male            10.9              5.5                  699
## 3  62   Male             7.3              4.1                  490
## 4  58   Male             1.0              0.4                  182
## 5  72   Male             3.9              2.0                  195
## 6  46   Male             1.8              0.7                  208
##   Alamine_Aminotransferase Aspartate_Aminotransferase Total_Proteins Albumin
## 1                       16                         18            6.8     3.3
## 2                       64                        100            7.5     3.2
## 3                       60                         68            7.0     3.3
## 4                       14                         20            6.8     3.4
## 5                       27                         59            7.3     2.4
## 6                       19                         14            7.6     4.4
##   Albumin_and_Globulin_Ratio Liver_Disease
## 1                       0.90             1
## 2                       0.74             1
## 3                       0.89             1
## 4                       1.00             1
## 5                       0.40             1
## 6                       1.30             1
```

This data set contains 583 observations with 416 liver disease patients and 167 non-afflicted patients. Each numeric variable (except for age) is a measurement relating to a liver protein, enzyme, etc. Categorical variables include liver disease status and sex. Liver patient records were collected from North East of Andhra Pradesh, India. The data was obtained here, however.

Hypothesis Testing

```
man1 <- manova(cbind(Total_Bilirubin, Direct_Bilirubin, Alkaline_Phosphotase,
                     Alamine_Aminotransferase, Total_Proteins, Albumin,
                     Albumin_and_Globulin_Ratio, Age) ~ Liver_Disease, data = liverd)
summary(man1)
```

```
##                 Df  Pillai approx F num Df den Df    Pr(>F)
## Liver_Disease    1 0.11601   9.3503      8    570 3.647e-12 ***
## Residuals      577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.aov(man1)
```

```
##  Response Total_Bilirubin :
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## Liver_Disease    1  1087.2 1087.15  29.408 8.633e-08 ***
## Residuals      577 21330.3   36.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Direct_Bilirubin :
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## Liver_Disease    1  278.1 278.087  37.255 1.903e-09 ***
## Residuals      577 4307.0   7.464
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Alkaline_Phosphotase :
##                 Df   Sum Sq Mean Sq F value    Pr(>F)
## Liver_Disease    1  1152846 1152846  20.075 8.982e-06 ***
## Residuals      577 33135491   57427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Alamine_Aminotransferase :
##                 Df   Sum Sq Mean Sq F value    Pr(>F)
## Liver_Disease    1   516055  516055  15.772 8.049e-05 ***
## Residuals      577 18879287   32720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Total_Proteins :
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Liver_Disease    1   0.77 0.76831  0.6527 0.4195
## Residuals      577 679.22 1.17715
##
##  Response Albumin :
```

```
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## Liver_Disease   1   9.31  9.3118  15.114 0.0001129 ***
## Residuals     577 355.48  0.6161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Albumin_and_Globulin_Ratio :
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## Liver_Disease   1  1.571 1.57107  15.775 8.037e-05 ***
## Residuals     577 57.465 0.09959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Age :
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## Liver_Disease   1   2697 2697.09  10.416 0.00132 **
## Residuals     577 149401  258.93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(liverd$Total_Bilirubin, liverd$Liver_Disease, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  liverd$Total_Bilirubin and liverd$Liver_Disease
##
##   0
## 1 8.6e-08
##
## P value adjustment method: none
```

```
pairwise.t.test(liverd$Direct_Bilirubin, liverd$Liver_Disease, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  liverd$Direct_Bilirubin and liverd$Liver_Disease
##
##   0
## 1 1.9e-09
##
## P value adjustment method: none
```

```
pairwise.t.test(liverd$Alkaline_Phosphotase, liverd$Liver_Disease, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  liverd$Alkaline_Phosphotase and liverd$Liver_Disease
##
##   0
## 1 9e-06
##
## P value adjustment method: none
```

```r
pairwise.t.test(liverd$Alamine_Aminotransferase, liverd$Liver_Disease, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  liverd$Alamine_Aminotransferase and liverd$Liver_Disease
##
##   0
## 1 8e-05
##
## P value adjustment method: none
```

```r
pairwise.t.test(liverd$Albumin, liverd$Liver_Disease, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  liverd$Albumin and liverd$Liver_Disease
##
##   0
## 1 0.00011
##
## P value adjustment method: none
```

```r
pairwise.t.test(liverd$Albumin_and_Globulin_Ratio, liverd$Liver_Disease, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  liverd$Albumin_and_Globulin_Ratio and liverd$Liver_Disease
##
##   0
## 1 8e-05
##
## P value adjustment method: none
```

```r
pairwise.t.test(liverd$Age, liverd$Liver_Disease, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  liverd$Age and liverd$Liver_Disease
##
##   0
## 1 0.0013
##
## P value adjustment method: none
```

```r
.05/16 #Bonferroni correction
```

```
## [1] 0.003125
```

In total 16 tests were conducted: 1 MANOVA, 8 ANOVAs, and 7 post-hoc t-tests. After a bonferroni adjustment, the probability of making a type I error is .0031. A one-way MANOVA was conducted to determine the effect of liver disease status on all of our numeric variables. Significant differences were found for liver disease status for at least one of our dependent

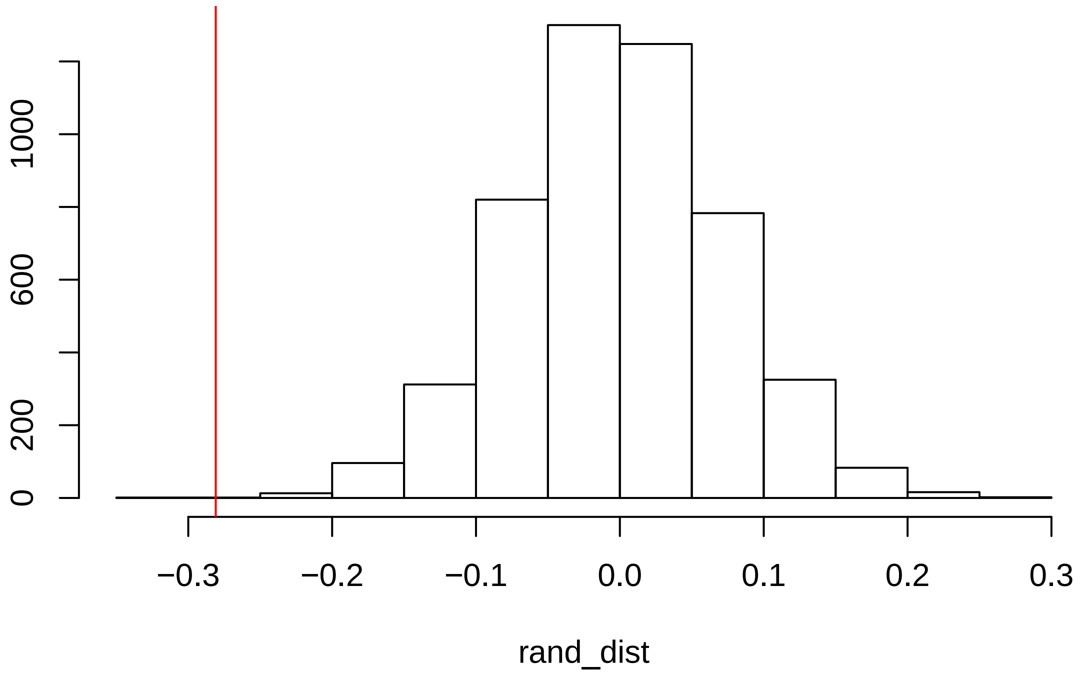
variables, F = 9.350, p < .0001. After running univariate ANOVAs for each of our dependent variables, only "Total Proteins" was found to not be significant (F = 0.652, p = 0.419). Post-hoc t-tests were calculated to determine if liver disease status differed across our variables. Liver disease onset and absence were found to differ from each other significantly in regards to all the variables that were tested.

Randomization Test

```
rand_dist <- vector()
for(i in 1:5000){
  new <- data.frame(albumin = sample(liverd$Albumin), liver_disease = liverd$Liver_Disease)
  rand_dist[i] <- mean(new[new$liver_disease == "1" ,]$albumin) -
    mean(new[new$liver_disease == "0" ,]$albumin)
}

liverd %>% group_by(Liver_Disease) %>% summarise(ldmean = mean(Albumin)) %>%
  summarise(diff_mean = diff(ldmean))
```

```
## # A tibble: 1 x 1
##   diff_mean
##       <dbl>
## 1    -0.281
```

```
hist(rand_dist, main = NULL, ylab = NULL) ; abline(v = -0.2809, col = "red")
```



```
mean(rand_dist > .2809 | rand_dist < -.2809)
```

```
## [1] 4e-04
```

```
t.test(Albumin ~ Liver_Disease, data = liverd)
```

```
##
##  Welch Two Sample t-test
##
## data:  Albumin by Liver_Disease
## t = 3.9067, df = 304.84, p-value = 0.0001153
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1394315 0.4224481
## sample estimates:
## mean in group 0 mean in group 1
##        3.339394        3.058454
```

Albumin was chosen because low levels of the protein could indicate the onset of liver disease (more info can be found here). The null hypothesis is that there is no difference between the means of albumin and liver disease status. Likewise, the alternative hypothesis is that there is a difference between the means of albumin and liver disease status. After conducting a randomization test, a p-value of 0 was obtained, leading to a rejection of the null hypothesis and further conclude that there is a significant difference between the true means of albumin and liver disease status (t = 3.907, p = 0).

Linear Regression Model

```
library(lmtest)
```

```
## Loading required package: zoo
```

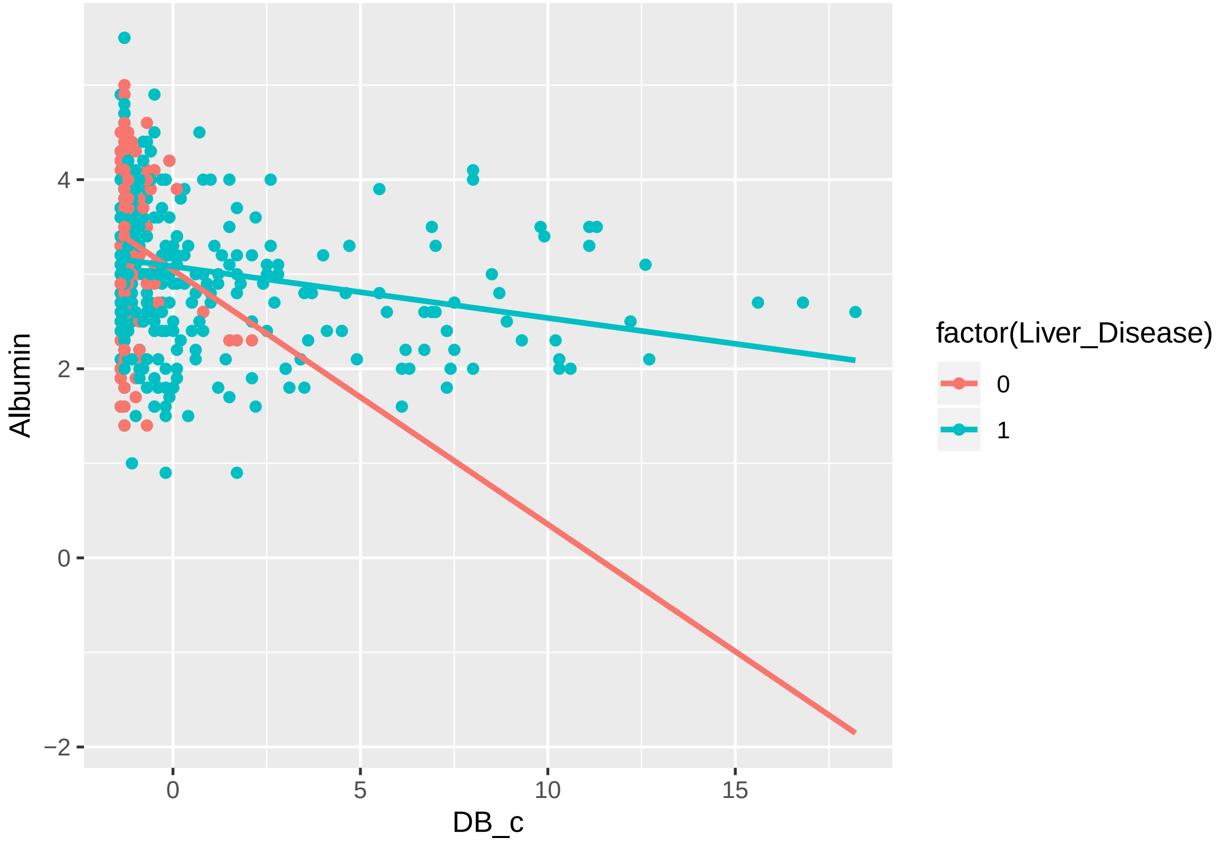
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

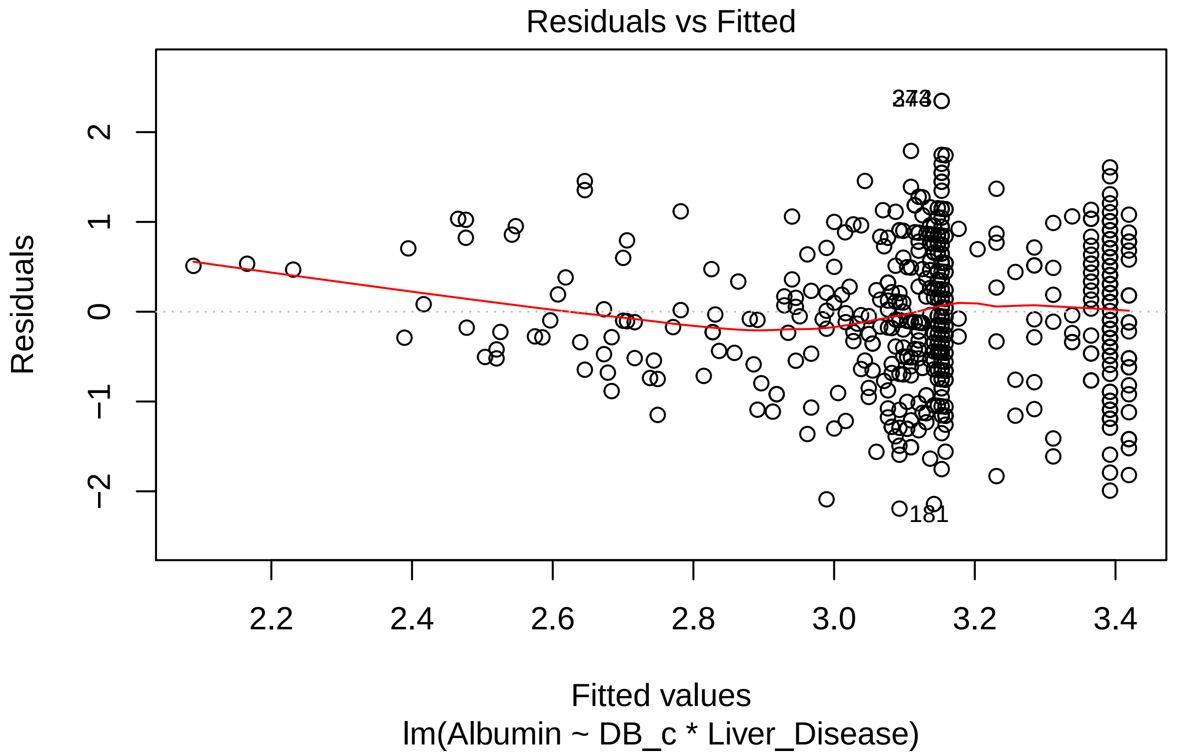
```
library(sandwich)

liverd$DB_c <- liverd$Direct_Bilirubin - mean(liverd$Direct_Bilirubin)

ld_fit <- lm(Albumin ~ DB_c * Liver_Disease, data = liverd)
summary(ld_fit)
```

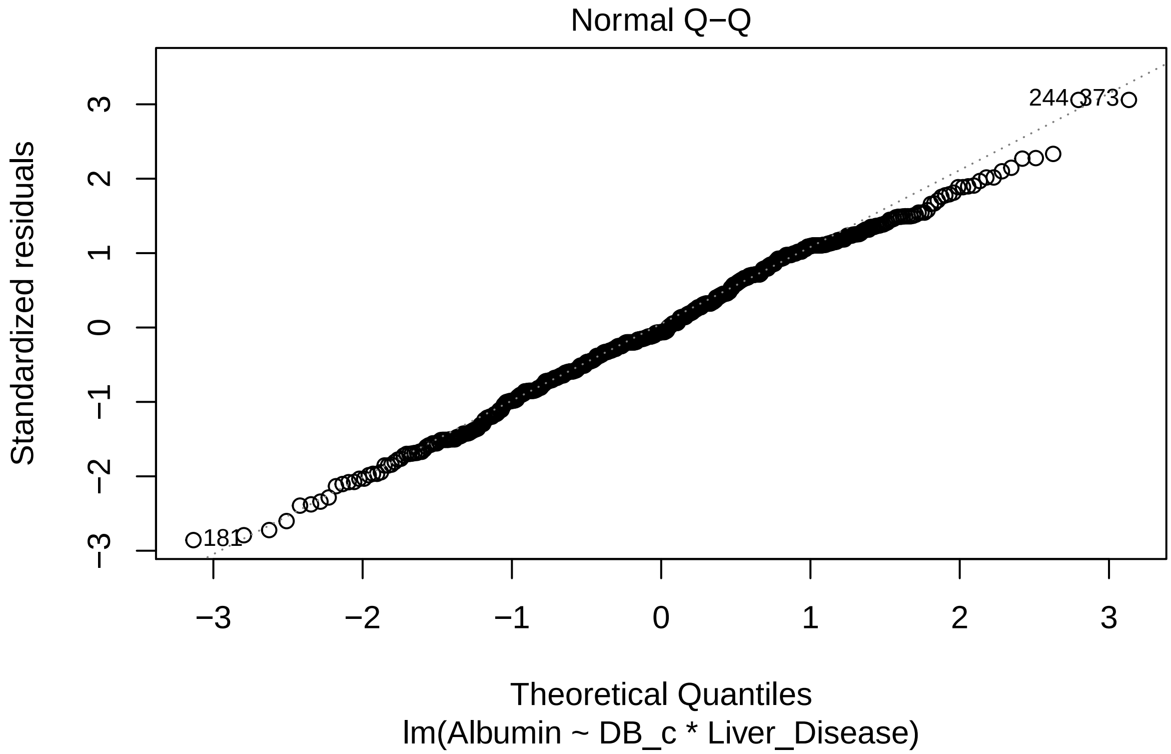
```
##
## Call:
## lm(formula = Albumin ~ DB_c * Liver_Disease, data = liverd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19291 -0.49528 -0.04746  0.57218  2.34709
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.04414    0.13967  21.795   <2e-16 ***
## DB_c                -0.26896    0.11497  -2.339   0.0197 *
## Liver_Disease       0.03818    0.14478   0.264   0.7921
## DB_c:Liver_Disease  0.21441    0.11557   1.855   0.0641 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7685 on 575 degrees of freedom
## Multiple R-squared:  0.06915,    Adjusted R-squared:  0.06429
## F-statistic: 14.24 on 3 and 575 DF,  p-value: 5.819e-09
```

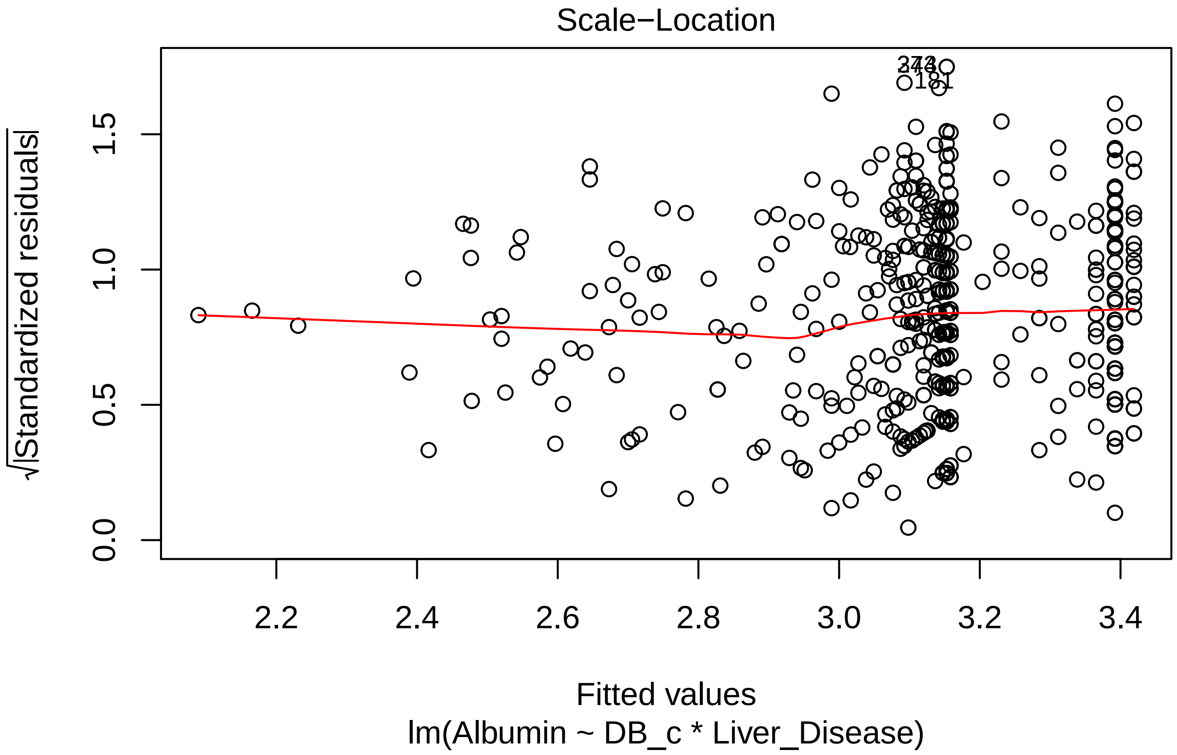
```r
liverd %>%
  ggplot(aes(x = DB_c, y = Albumin, color = factor(Liver_Disease))) +
  geom_point() +
  stat_smooth(method = "lm", se = F, fullrange = T)
```

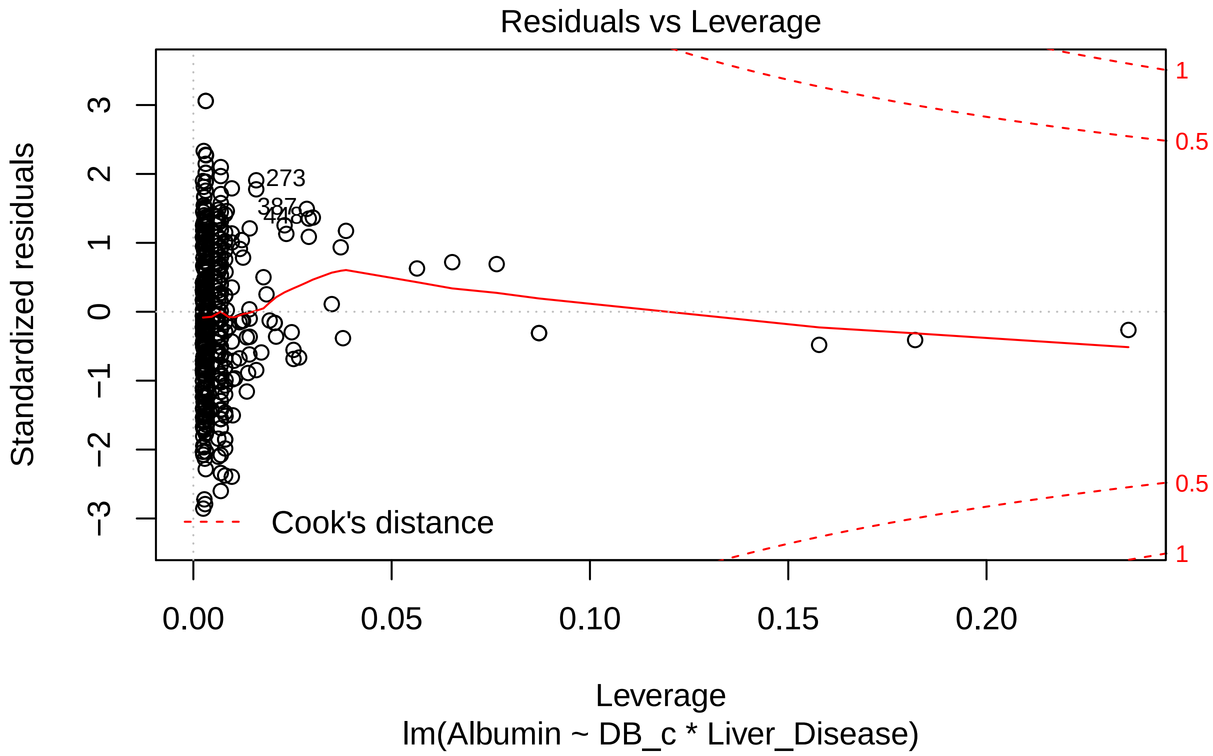


```r
plot(ld_fit) #Assumptions Check
```


Residuals vs Fitted
Residuals
Fitted values
lm(Albumin ~ DB_c * Liver_Disease)
−2
−1
0
1
2
2.2
2.4
2.6
2.8
3.0
3.2
3.4
374
243
181


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
244
373
181
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Albumin ~ DB_c * Liver_Disease)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.2
2.4
2.6
2.8
3.0
3.2
3.4
274
343
181
Fitted values
lm(Albumin ~ DB_c * Liver_Disease)



```r
coeftest(ld_fit, vcov. = vcovHC(ld_fit))
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          3.044144   0.092244 33.0008 < 2.2e-16 ***
## DB_c                -0.268956   0.075238 -3.5747 0.0003801 ***
## Liver_Disease        0.038175   0.099979  0.3818 0.7027252
## DB_c:Liver_Disease   0.214409   0.075949  2.8231 0.0049212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The predicted albumin level for a non-afflicted patient with an average direct bilirubin level is 3.044 g/dL. Controlling for liver disease status, for every 1 mg/dL increase in direct bilirubin level, albumin decreases by 0.269 on average. Controlling for direct bilirubin, a patient with liver disease shows a 0.038 g/dL increase in albumin. The slope for direct bilirubin on albumin is 0.214 greater for liver disease afflicted patients than non-afflicted patients.After recomputing the regression with robust standard errors, the interaction between DB and liver disease status become significant, p = 0.0049. Average DB also becomes "more" significant, p = 0.00038 compared to p = 0.0197. Average DB and the interaction between average DB and liver disease status show significant variation in albumin (t = -3.57, p = 0.0003 & t = 2.82, p = 0.0049 respectively).

Bootstrapping

```
ld_dist <- replicate(5000, {
  boot_ld <- liverd[sample(nrow(liverd), replace = T),]
  fit <- lm(Albumin ~ DB_c * Liver_Disease, data = boot_ld)
  coef(fit)
})

ld_dist %>% t %>% as.data.frame() %>% summarise_all(sd)
```

```
##   (Intercept)       DB_c Liver_Disease DB_c:Liver_Disease
## 1   0.1077164 0.08702241     0.1142434           0.087528
```

After bootstrapping standard errors, there is an increase in the SE values compared to the robust SEs that were calculated prior. However, compared to the original SEs, the values of the boot SEs are lower.

Logistic Regression and Cross Validation

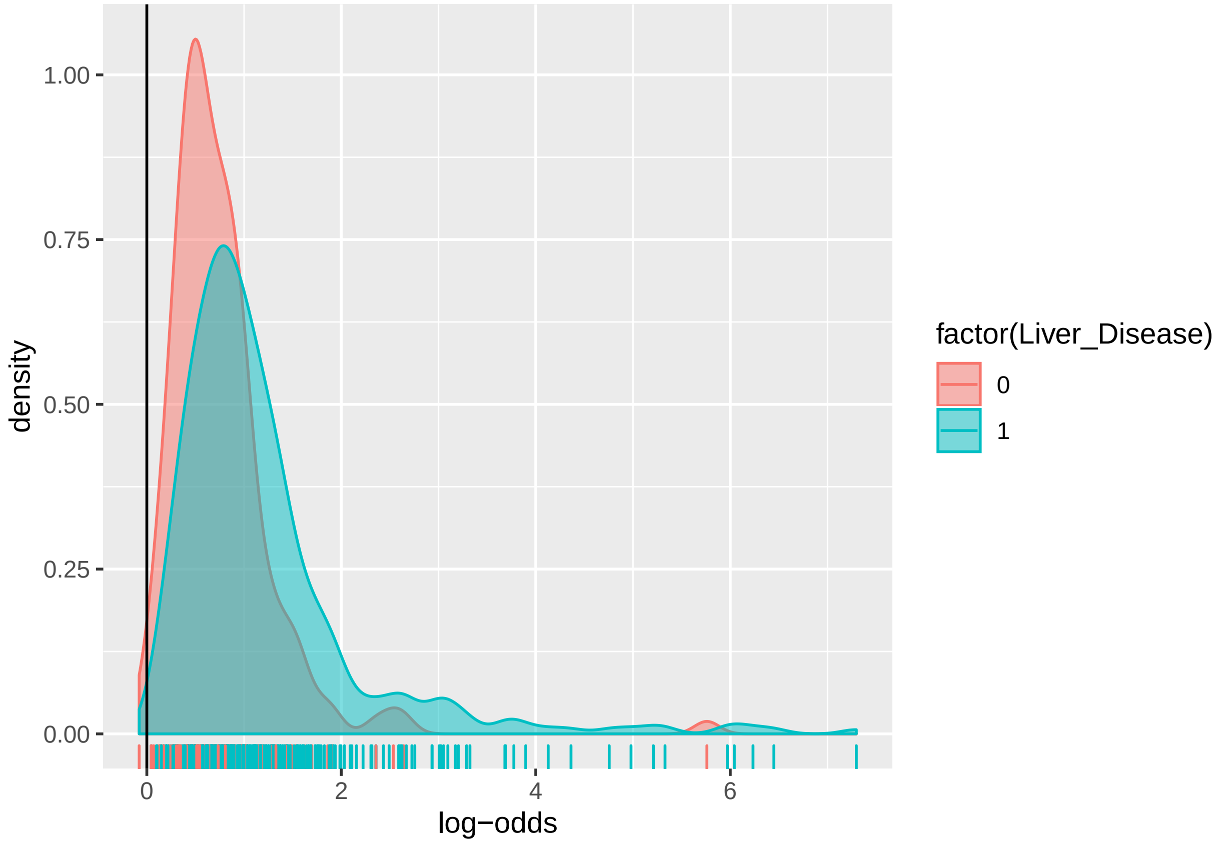
```
library(plotROC)

ld_fit2 <- glm(Liver_Disease ~ Albumin + Alkaline_Phosphotase, data = liverd,family = "binomial")
summary(ld_fit2)
```

```
##
## Call:
## glm(formula = Liver_Disease ~ Albumin + Alkaline_Phosphotase,
##     family = "binomial", data = liverd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3950  -1.3050   0.6972   0.8746   1.1590
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.2578707  0.4720395   2.665  0.00770 **
## Albumin              -0.3738801  0.1229782  -3.040  0.00236 **
## Alkaline_Phosphotase  0.0033936  0.0008688   3.906 9.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 692.01  on 578  degrees of freedom
## Residual deviance: 651.99  on 576  degrees of freedom
## AIC: 657.99
##
## Number of Fisher Scoring iterations: 5
```

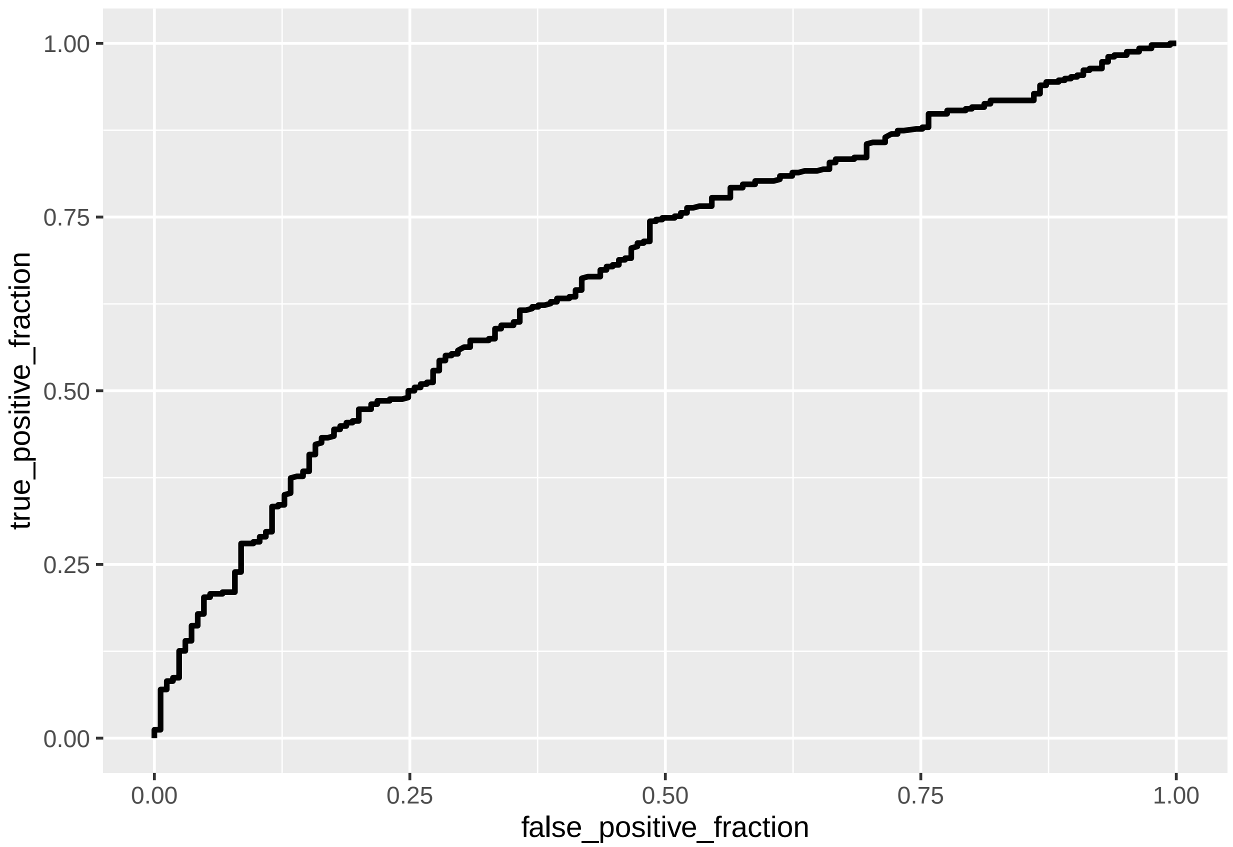
```
#Confusion Matrix
ld_prob <- predict(ld_fit2, type = "response")
table(predict = as.numeric(ld_prob > .5), truth = liverd$Liver_Disease) %>% addmargins
```

```
##        truth
## predict   0   1 Sum
##     0     1   0   1
##     1   164 414 578
##     Sum 165 414 579
```

```
liverd$logit <- predict(ld_fit2, type = "link")

#AUC plot
liverd %>%
  ggplot() +
  geom_density(aes(logit, color = factor(Liver_Disease), fill = factor(Liver_Disease)), alpha = 0.5) +
  geom_vline(xintercept = 0) +
  xlab("log-odds") +
  geom_rug(aes(logit, color = factor(Liver_Disease)))
```



```
#ROC curve
ld_ROC <- liverd %>%
  ggplot() +
  geom_roc(aes(d = Liver_Disease, m = ld_prob), n.cuts = 0)
ld_ROC
```



```
calc_auc(ld_ROC)
```

```
##   PANEL group       AUC
## 1     1    -1 0.6737374
```

```
#10-Fold Cross Validation
k = 10

ld_cv <- liverd[sample(nrow(liverd)),]
folds <- cut(seq(1:nrow(liverd)), breaks = k, labels = F)

diags <- NULL
for(i in 1:k){
  train <- ld_cv[folds != i,]
  test <- ld_cv[folds == i,]
  truth <- test$Liver_Disease

  cvfit <- glm(Liver_Disease ~ Albumin + Alkaline_Phosphotase, data = train, family = "binomial")
  probs <- predict(cvfit, newdata = test, type = "response")
  diags <- rbind(diags, class_diag(probs, truth)) #class_diag for convenience
}

summarise_all(diags, mean)
```

```
##         acc      sens       spec       ppv       auc
## 1 0.7166969 0.9974359 0.01081871 0.7167985 0.6678924
```

```
yhat <- predict(cvfit)
mean((liverd$Liver_Disease - yhat)^2)
```

```
## Warning in liverd$Liver_Disease - yhat: longer object length is not a multiple
## of shorter object length
```

```
## [1] 1.152784
```

Controlling for alkaline phosphotase, albumin has a significant negative impact on the odds of liver disease onset. Controlling for albumin, alkaline phosphotase has a significant positive impact on the odds of liver disease onset. After computing a confusion matrix, the sensitivity for the model is 0.716 and the specificity is a value of 1. Calculating the AUC gives a value of 0.674, which tells us that the model is a poor at classifying patients with liver disease and those without. By performing 10-fold cross validation on the model, there is a very miniscule increase in auc (=0.676).

LASSO

```
library(glmnet)
```

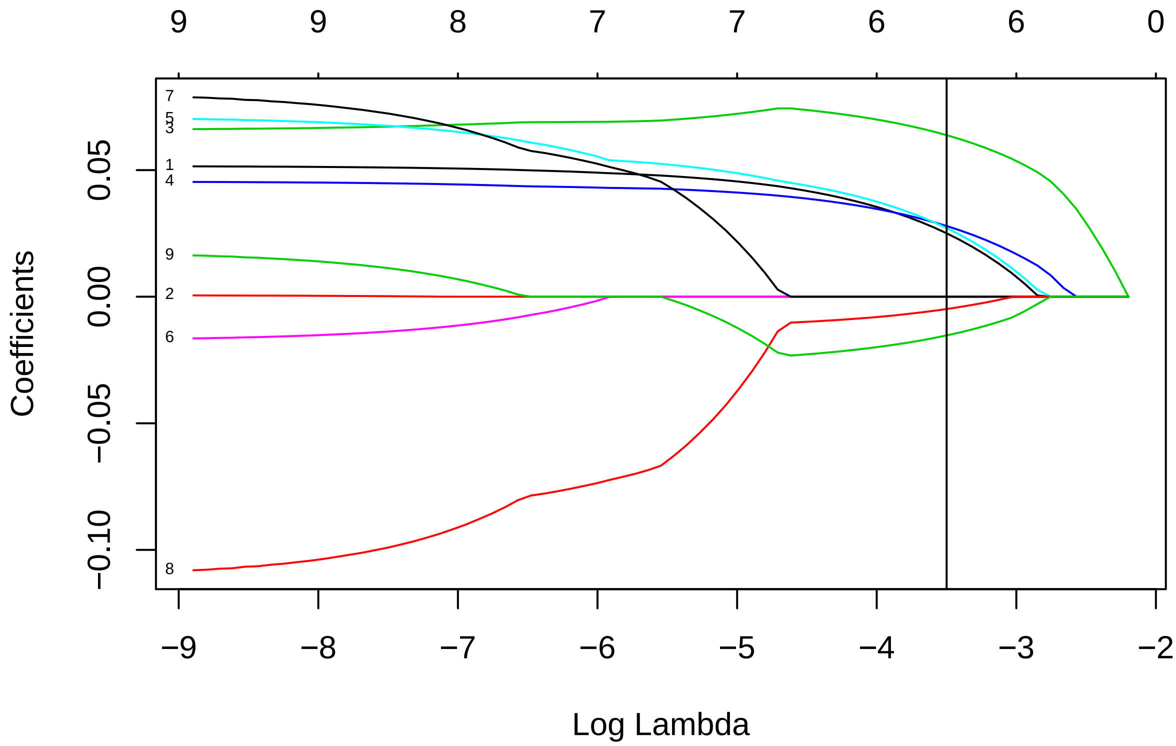
```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 3.0-2
```

```
liverd$LD_n <- liverd$Liver_Disease %>% as.numeric #code for orignial LD_n was lost, here for knit

y <- as.matrix(liverd$Liver_Disease)
x <- liverd %>% dplyr::select(-Liver_Disease, -Gender, -DB_c, -LD_n, -logit) %>% mutate_all(scale) %>% a

cv <- cv.glmnet(x,y)
plot(cv$glmnet.fit, "lambda", label = T) ; abline(v = log(cv$lambda.1se)) #Plot looks cool
```



```
lasso1 <- glmnet(x, y, lambda = cv$lambda.1se)
coef(lasso1)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                                      s0
## (Intercept)                 0.715025907
## Age                         0.025052404
## Total_Bilirubin             .
## Direct_Bilirubin            0.063810116
## Alkaline_Phosphotase        0.027924118
## Alamine_Aminotransferase    0.027105360
## Aspartate_Aminotransferase  .
## Total_Proteins              .
## Albumin                    -0.004874222
## Albumin_and_Globulin_Ratio -0.015297130
```

```
#LASSO Assisted 10-Fold CV
k = 10

ld_cv2 <- liverd[sample(nrow(liverd)),]
folds2 <- cut(seq(1:nrow(liverd)), breaks = k, labels = F)

diags2 <- NULL
for(i in 1:k){
  train2 <- ld_cv[folds != i,]
  test2 <- ld_cv[folds == i,]
  truth2 <- test2$Liver_Disease
  cvfit2 <- glm(Liver_Disease ~ Albumin + Alkaline_Phosphotase + Age +
                  Direct_Bilirubin + Alamine_Aminotransferase +
                  Albumin_and_Globulin_Ratio, data = train2, family = "binomial")
  probs2 <- predict(cvfit, newdata = test2, type = "response")
  diags2 <- rbind(diags, class_diag(probs2, truth2)) #class_diag for convenience
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summarise_all(diags2, mean)
```

```
##         acc     sens        spec       ppv       auc
## 1 0.7189408 0.997669 0.009835194 0.7190331 0.6674991
```

```
yhat2 <- predict(cvfit2)
mean((liverd$Liver_Disease - yhat2)^2)
```

```
## Warning in liverd$Liver_Disease - yhat2: longer object length is not a multiple
## of shorter object length

## [1] 8.71098
```

After performing a LASSO on the data, the variables age, direct bilirubin, alkaline phosphotase, alamine aminotransferase, albumin, and albumin/globulin ratio are retained. The mean-squared error is a value of 12.107, which is larger than the mean-squared error that was obtained prior (1.123).