Weekly Exercise - Week 5

Marjolein Fokkema

Read in data:

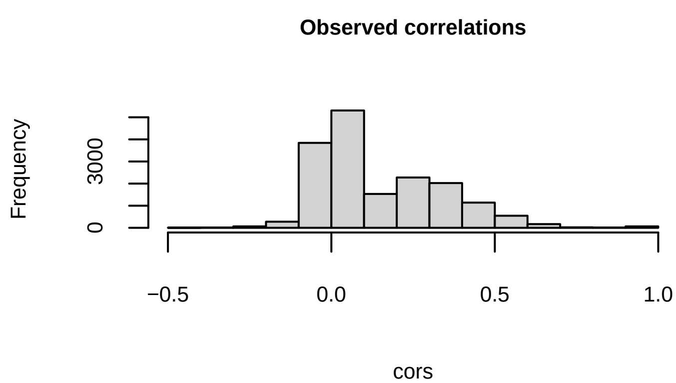
```
train <- readRDS("masq_train.Rda")
test <- readRDS("masq_test.Rda")
```

Prepare predictors and response in training and test set for analyses with `cv.glmnet`:

```
library("glmnet")
x <- model.matrix(D_DEPDYS ~ ., data = train)
y <- train$D_DEPDYS
x_test <- model.matrix(D_DEPDYS ~ . -1, data = test)
y_test <- test$D_DEPDYS
```

a) Likely, many of the MASQ items are substantially correlated, because they are all measures of psychopathology symptoms and such symptoms are often correlated.

```
cors <- cor(x)
diag(cors) <- NA
hist(cors, main = "Observed correlations", cex.lab = .7, cex.axis = .7,
     cex.main = .7)
```



There is a bump of correlations between 0.2 and 0.6, which seem beyond what would be expected for independent predictors. With such multicollinearity, the 'true' solution is unlikely to be sparse, thus performance of the lasso may likely benefit from inclusion of a ridge penalty.

Note that some predictors have correlation perfectly equal to 1 (even after removal of the diagonal of the correlation matrix). This seems mostly for the 'missing' categories of the demographic indicators, which are likely not good predictors anyway so this does not seem worrying.

b) There is no single best answer to this question, but this is my motivation:

(Relaxed) lasso would be useful, because:

- For interpretation and application, a sparse model with only few predictors would likely be useful. For example, if only a subset of items is relevant for classifying depressed versus not, we could administer only a subset of items to future patients, reducing assessment burden. Thus, the (relaxed) lasso would be useful in light of practical applications. The relaxed lasso could further reduce the number of predictors retained, without damaging predictive accuracy too much, because it eases shrinkage on large coefficients with $\gamma < 1$.
- In light of the multicollinearity, some ridge penalization might be beneficial, so I will also try a ridge model and an elastic net (which combines the ridge and lasso penalties) with α of .25.

c)

```
## ridge
set.seed(42)
r_mod <- cv.glmnet(x = x, y = y, family = "binomial", alpha = 0)
r_mod
```

```
##
## Call:  cv.glmnet(x = x, y = y, family = "binomial", alpha = 0)
##
## Measure: Binomial Deviance
##
##     Lambda Index Measure      SE Nonzero
## min 0.1439    81   1.033 0.02579     132
## 1se 0.8428    62   1.057 0.01893     132
```

```
## elastic net alpha .25
set.seed(42)
en_mod <- cv.glmnet(x = x, y = y, family = "binomial", alpha = 0.25)
en_mod
```

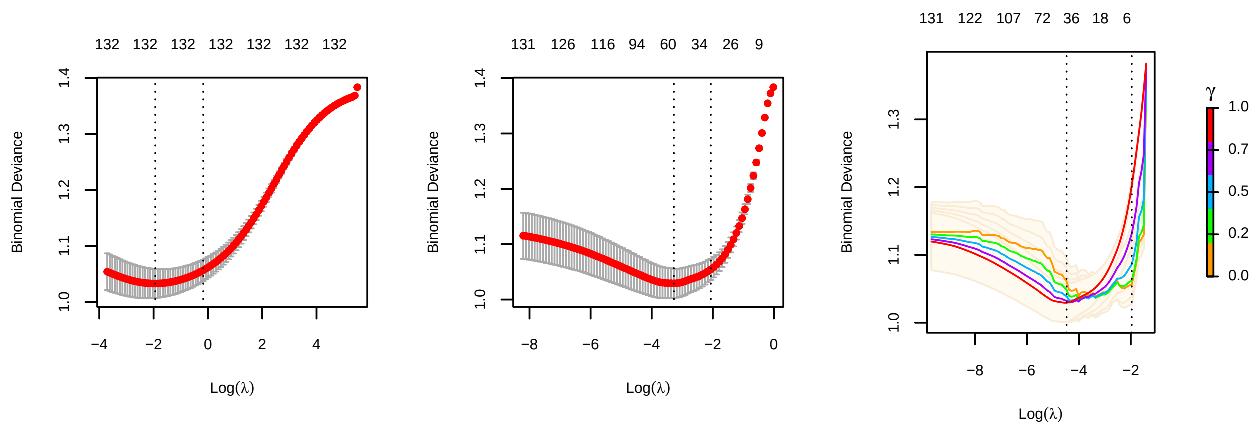
```
##
## Call:  cv.glmnet(x = x, y = y, family = "binomial", alpha = 0.25)
##
## Measure: Binomial Deviance
##
##      Lambda Index Measure      SE Nonzero
## min 0.03787    36   1.029 0.02650      54
## 1se 0.12692    23   1.054 0.01926      32
```

```
## relaxed lasso (which includes the original lasso)
set.seed(42)
rl_mod <- cv.glmnet(x = x, y = y, relax = TRUE, family = "binomial")
rl_mod
```

```
##
## Call:  cv.glmnet(x = x, y = y, relax = TRUE, family = "binomial")
```

```
##
## Measure: Binomial Deviance
##
##     Gamma Index Lambda Index Measure      SE Nonzero
## min     1     5 0.0114    34   1.029 0.02803      37
## 1se     0     1 0.1406     7   1.055 0.02820       4
```

```
par(mfrow = c(1, 3))
plot(r_mod); plot(en_mod); plot(rl_mod)
```



The curves for cross-validated performance are convex, from which we can conclude that there are relevant predictors in the dataset and some penalization is beneficial.

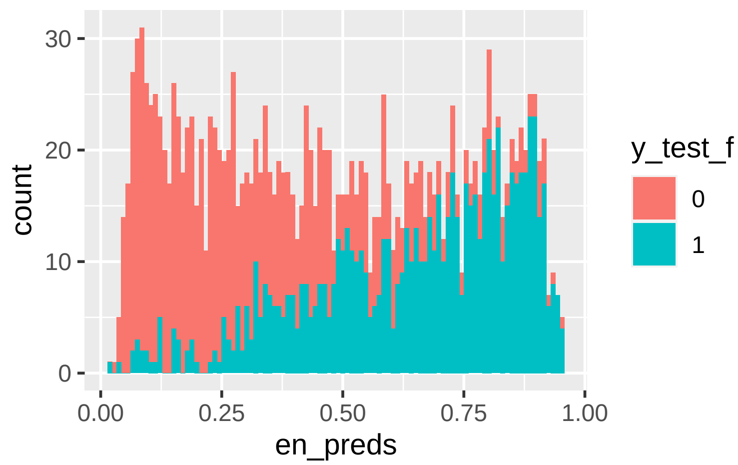
Optimal predictive accuracy (i.e., lowest cross-validated deviance of 1.029) is obtained for the elastic net and relaxed lasso, when using the `lambda.min` criterion. For the relaxed lasso, this criterion yields a model with 37 predictors, and a γ of 1 should be employed (i.e., the original lasso fit).

Note that the lowest standard errors are obtained for ridge regression, illustrating how it tends to have better stability than the lasso.

For optimal sparsity, we might prefer the `lambda.1se` criterion (the default in `glmnet`). Then the minimum deviance is obtained with elastic net, but the relaxed lasso seems only slightly less accurate while it is very much sparser.

d) We evaluate the best-fitting model(s) on the test data:

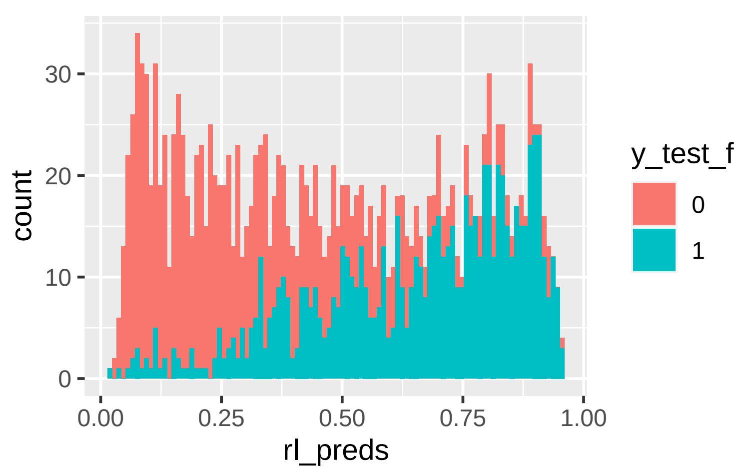
```
y_test <- test$D_DEPDYS
library("ggplot2")

## Elastic net (alpha .25)
df <- data.frame(y_test, y_test_f = factor(y_test))
df$en_preds <- predict(en_mod, newx = x_test, s = "lambda.min", type = "response")
ggplot() + geom_histogram(data = df, aes(x=en_preds, fill=y_test_f), bins = 100)
```




```
1 - sum(diag(table(df$en_preds > 0.5, y_test)))/nrow(train) ## MCR
```

```
## [1] 0.2296997
```

```
## Relaxed lasso
df$rl_preds <- predict(rl_mod, newx = x_test, s = "lambda.min", type = "response")
ggplot() + geom_histogram(data = df, aes(x=rl_preds, fill=y_test_f), bins = 100)
```




```
1 - sum(diag(table(df$rl_preds > 0.5, y_test)))/nrow(train) ## MCR
```

```
## [1] 0.2263626
```

Elastic net yields the lowest MCR, but the difference in performance with the relaxed lasso is small, especially in light of the smaller number of predictors selected. Note that the predicted probabilities seem more effective at distinguishing the depressed than the non-depressed patients.

Similar conclusions would apply if we would have opted for the default lambda-1se criterion.

d) We inspect which items were selected for prediction:

```r
## elastic net
en_coefs <- as.matrix(coef(en_mod, s = "lambda.min"))
round(sort(en_coefs[en_coefs != 0, ][-1]), digits = 3)
```

```
##   DEMOG26   DEMOG53 DEMOG3NA    MASQ02    MASQ59    MASQ03  DEMOG34 DEMOG6NA
##    -0.370    -0.174   -0.141    -0.063    -0.050    -0.041   -0.036   -0.026
## DEMOG7NA DEMOG8NA   DEMOG42    MASQ35    MASQ55    MASQ32   MASQ07    MASQ57
##    -0.025    -0.023   -0.009    -0.007    -0.006     0.003    0.005     0.006
##    MASQ83    MASQ11 Leeftijd    MASQ39    MASQ80    MASQ50   MASQ04    MASQ05
##     0.006     0.007    0.007     0.007     0.007     0.012    0.014     0.015
##    MASQ29    MASQ44    MASQ33    MASQ70    MASQ54    MASQ24   MASQ38    MASQ21
##     0.016     0.016    0.021     0.022     0.024     0.033    0.036     0.036
##    MASQ14    MASQ31    MASQ60    MASQ18    MASQ76    MASQ78   MASQ13    MASQ43
##     0.043     0.043    0.050     0.052     0.056     0.057    0.059     0.062
##   GENDERv   DEMOG32    MASQ90    MASQ62   DEMOG72    MASQ22   MASQ37    MASQ30
##     0.067     0.072    0.073     0.074     0.083     0.094    0.098     0.101
##    MASQ41   DEMOG55    MASQ89    MASQ01   DEMOG62    MASQ16
##     0.123     0.134    0.139     0.152     0.181     0.257
```

```r
## relaxed lasso
rl_coefs <- as.matrix(coef(rl_mod, s = "lambda.min"))
round(sort(rl_coefs[rl_coefs != 0, ][-1]), digits = 3)
```

```
##   DEMOG26 DEMOG3NA   DEMOG53    MASQ02    MASQ59    MASQ03  DEMOG34    MASQ44
##    -0.318   -0.204    -0.186    -0.077    -0.053    -0.031   -0.002     0.001
##    MASQ83 Leeftijd    MASQ05    MASQ54    MASQ70    MASQ14   MASQ38    MASQ24
##     0.004    0.007     0.008     0.012     0.014     0.022    0.025     0.031
##    MASQ18   MASQ21    MASQ31    MASQ60    MASQ78   GENDERv   MASQ13    MASQ43
##     0.032    0.033     0.045     0.049     0.050     0.050    0.056     0.058
##    MASQ76  DEMOG32    MASQ62    MASQ90    MASQ22    MASQ37  DEMOG55    MASQ30
##     0.060    0.068     0.074     0.076     0.103     0.106    0.117     0.122
##    MASQ41   MASQ89    MASQ01   DEMOG62    MASQ16
##     0.142    0.148     0.184     0.230     0.345
```

- Anhedonic Depression: Items 1, 14, 18, 21, 23, 26, 27, 30, 33, 35, 36, 39, 40, 44, 49, 53, 58, 66, 72, 78, 86 and 89.
- Anxious Arousal: Items 3, 19, 25, 45, 48, 52, 55, 57, 61, 67, 69, 73, 75, 79, 85, 87 and 88.
- General Distress Depression: Items 6, 8, 10, 13, 16, 22, 24, 42, 47, 56, 64 and 74.
- General Distress Anxiety: Items 2, 9, 12, 15, 20, 59, 63, 65, 77, 81 and 82.
- General Distress Mixed: Items 4, 5, 17, 29, 31, 34, 37, 50, 51, 70, 76, 80, 83, 84 and 90.

The most important item (16) belongs to the General Distress Depression subscale, but most of the selected items were from the Anhedonic Depression subscale. This makes a lot of sense, because the prediction target is depression. For other disorder types (e.g., anxiety), items from other subscales may be more informative.