Assignment 2 - Part 1C

High-dimensional data

Fernando Cagua

```
parkinsons <- read.csv("parkinsons.csv") %>%
  rename(patient = X) %>%
  mutate_at(vars(contains("X")), scale)
X <- model.matrix(UPDRS ~ ., select(parkinsons, - patient))
y <- parkinsons$UPDRS
set.seed(987654312)
train <-sample(1:nrow(X),30)
test <- -train
```

1 Linear model

```
linear <- lm(y[train] ~ X[train, -1])
sum(residuals(linear))
```

```
## [1] 0
```

The linear model has a perfect fit and it's not useful because its too complex and overfits the train data. This means that the predictive performance on the test data is likely to be extremely poor.

2 Using lasso

```
grid <- 10^seq(3,-1,length.out = 100)
lasso <- cv.glmnet(X[train, -1], y[train], alpha =1, lambda = grid, nfolds = 30,
                   thresh = 1e-10)
# optimal value
lasso$lambda.min
```

```
## [1] 0.6428073
```

```
# test error
mean((predict(lasso, X[test, -1], s = lasso$lambda.min) - y[test])^2)
```

```
## [1] 4.637133
```

3 Final model

```
predictors <- coef(lasso, s = lasso$lambda.min) %>% as.matrix()
predictors <- predictors[predictors[, 1] != 0,]
# make it a string
model <- paste(round(predictors,4),"*",names(predictors)) %>%
  paste(collapse = " + ") %>%
  gsub("\\* \\(Intercept\\) ", "", .)
```

The final model is $UPDRS = 26.6124 + 0.0328 * X9 + 0.6082 * X10 + -0.238 * X55 + 0.6761 * X83 + 0.2706 * X95 + 9.1399 * X97$. The lasso algorithm selected 6 features. X97 had indeed the largest effect. All other features are relatively unimportant and had effects that are around one order of magnitude smaller than X97.

4 Different random split

```
set.seed(1)
train <-sample(1:nrow(X),30)
test <- -train
lasso <- cv.glmnet(X[train, -1], y[train], alpha =1, lambda = grid, nfolds = 30,
                   thresh = 1e-10)
predictors <- coef(lasso, s = lasso$lambda.min) %>% as.matrix()
predictors <- predictors[predictors[, 1] != 0,]
# make it a string
model <- paste(round(predictors,4),"*",names(predictors)) %>%
  paste(collapse = " + ") %>%
  gsub("\\* \\(Intercept\\) ", "", .)
```

Using a different seed, now the final model is $UPDRS = 27.2835 + -0.1389 * X2 + 0.8655 * X9 + -0.6139 * X55 + 0.9645 * X83 + 9.1724 * X97$, which cointains 5 features. Four or them are in both models, however the coefficients for those predictors seem to be considerably different.