Lesson 7

**Divide in group: can you bootstrapped 1000 sample with method 1 the correlation between poor and mort?**

##check the mean and sd

boot <- 1000

cor\_boot <- matrix(NA, boot, 1)

for (i in 1:boot){

s\_boot <- sample(c(1:dim(data)[1]),dim(data)[1], replace=T)

data\_boot <- data[s\_boot,]

cor\_boot[i,] <- cor(data\_boot$poor, data\_boot$mort)

}

hist(cor\_boot)

mean(cor\_boot)

sd(cor\_boot)

**Divide in group: can you bootstrapped 1000 sample with method 2 the mean of the poor variable?**

##check the mean and sd of your bootstrapped samples

fc\_mean <- function(d, i){

d2 <- data[i,]

return(mean(d2$poor))

}

bootmean <- boot(data, fc\_mean, R=1000)

mean(bootmean$t)

sd(bootmean$t)

hist(bootmean$t)

**Divide in group: Do the same cross validation with 80% for training and 20% for the test, including in the model the variable poor. What model do you prefer in term of prediction?**

# Build the model

modelP <- lm(mort~so2+educ+nonw+poor, data = train.data)

# Make predictions and compute the R2, RMSE and MAE

predictionsP <- modelP %>% predict(test.data)

data.frame( R2 = R2(predictionsP, test.data$mort),

RMSE = RMSE(predictionsP, test.data$mort),

MAE = MAE(predictionsP, test.data$mort))

We prefer the model with poor since the error for the two measures is better!

**Divide in group: Do the LOOCV for the mdoel that contains the poverty as predictors. What model do you prefer in term of prediction?**

modelP <- train(mort~so2+educ+nonw+poor, data = data, method = "lm",trControl = train.control)

# Summarize the results

print(modelP)

We still prefer the model wit poor since the error for the two measures is better!

**Divide in group: Do the k-fold cross-validation for the model that contains the poverty as covariate. What model do you prefer in term of prediction?**

modelP <- train(mort~so2+educ+nonw+poor, data = data, method = "lm",trControl = train.control)

print(modelP)

The RMSE and MAE are lower for the model with poverty but the R^2 is higher for the model without poverty. We have to find a balance, I will choose. Amodel without poor if you have to just explain the data, I will choose a model with poor if your aim is to predict.