Homework 6

Gabriel Lahman

3/1/2019

```
data("faithful")
```

Question 1

```
FIT_glm = function(data, train_ind, form, outcome_name) {
  n = dim(data)[1]
  data_train = subset(data, c(1:n)%in%train_ind)
  model = glm(formula = form,
              data = data_train,
              family = Gamma(link = "log"))
  return (model)
}

CVM_glm = function(fit, data, train_ind, form, outcome_name) {
  n = dim(data)[1]
  data_test = subset(data, !(c(1:n)%in%train_ind))
  pred = predict(fit,
                 newdata=data_test,
                 type = "response")
  actual = data_test[,outcome_name]
  mean_abs_error = mean(abs(pred - actual))
  return (mean_abs_error)
}
```

Question 2

```
SPLIT_Kfold = function(n, K) {
  ind = c(1:n)
  out = list()
  for (i in 1:(K - 1)) {
    size = length(ind)* 1/(K - i + 1)
    split = sort(sample(ind, size))
    ind = ind[ !( ind %in% split) ]
    out[[i]] = split
  }
  out[[K]] = ind
  return(out)
}

CV_Kfold = function(data,K_SPLIT,FIT,CVM,...){
  n = dim(data)[1]
  if(class(K_SPLIT)=="list"){
    split_ind = K_SPLIT
    K = length(K_SPLIT)
  }else{
    split_ind = SPLIT_Kfold(n,K_SPLIT)
    K = K_SPLIT
  }
  cvm = rep(NA,K)
  ind = c(1:n)
  for(i in 1:K){
    train_ind = ind[ -split_ind[[i]] ]
    fit = FIT(data=data,train_ind=train_ind,...)
    cvm[i] = CVM(fit=fit,data=data,train_ind=train_ind,...)
  }
  return(list(cvm = cvm, avg_cvm = mean(cvm), split_ind = split_ind))
}

outcome_name = "eruptions"
form_1 = eruptions ~ waiting
form_5 = eruptions ~ poly(waiting,5)
K_SPLIT = dim(faithful)[1]
cv_poly_1 = CV_Kfold(faithful, K_SPLIT = K_SPLIT, FIT_glm, CVM_glm, form = form_1, ou
tcome_name = outcome_name)
cv_poly_5 = CV_Kfold(faithful, K_SPLIT = cv_poly_1$split_ind, FIT_glm, CVM_glm, form=
form_5,
                     outcome_name=outcome_name)
```

```
## [1] 0.4409814 0.3114608
```

Question 3

```
max_D = 15
K_SPLIT = SPLIT_Kfold(dim(faithful)[1], 10)
avg_cvm = rep(0, max_D)
for (d in 1:max_D) {
  form = eruptions ~ poly(waiting,d)
  cv = CV_Kfold(faithful, K_SPLIT, FIT_glm, CVM_glm, form=form, outcome_name=outcome_
name)
  avg_cvm[d] = cv$avg_cvm
}
```

```
##  [1] 0.4412438 0.3901640 0.3436156 0.3158453 0.3130698 0.2986254 0.3069136
##  [8] 0.2916358 0.3005023 0.2894087 0.3210205 0.3526475 0.3666202 0.3699356
## [15] 1.1113055
```

```
## [1] "Best Model = 10"
```

```
K_SPLIT = SPLIT_Kfold(dim(faithful)[1], 10)
avg_cvm = rep(0, max_D)
for (d in 1:max_D) {
  form = eruptions ~ poly(waiting,d)
  cv = CV_Kfold(faithful, K_SPLIT, FIT_glm, CVM_glm, form=form, outcome_name=outcome_
name)
  avg_cvm[d] = cv$avg_cvm
}
```

```
##  [1] 0.4422847 0.3914960 0.3430150 0.3157466 0.3102410 0.2965756 0.3045267
##  [8] 0.2885321 0.2969940 0.2897857 0.3124269 0.3537591 0.3718020 0.5081988
## [15] 4.0527239
```

```
## [1] "Best Model = 8"
```

Yes, the model which uses an 8th degree polynomial was the best model regardless of split.