101C Final EDA

Andy Shen

11/30/2020

```
rm(list = ls())
library(tidyverse)
library(BASS)
library(TBASS)
```

```
training <- read.csv("/Users/andyshen/Desktop/Git/101C-Final/training.csv")
test <- read.csv("/Users/andyshen/Desktop/Git/101C-Final/test.csv")
```

```
ind <- sample(1:nrow(training), size = floor(0.8 * nrow(training)), replace = F)
train <- training[ind,]
train_resp <- train$growth_2_6
train_preds <- train %>% select(-growth_2_6, -PublishedDate)
ttest <- training[-ind, ]
```

```
b <- bass(train_preds, train_resp)
```

```
## MCMC Start #-- Nov 30 13:31:24 --# nbasis: 0
## MCMC iteration 1000 #-- Nov 30 13:31:27 --# nbasis: 20
## MCMC iteration 2000 #-- Nov 30 13:31:30 --# nbasis: 17
## MCMC iteration 3000 #-- Nov 30 13:31:33 --# nbasis: 20
## MCMC iteration 4000 #-- Nov 30 13:31:36 --# nbasis: 22
## MCMC iteration 5000 #-- Nov 30 13:31:39 --# nbasis: 24
## MCMC iteration 6000 #-- Nov 30 13:31:42 --# nbasis: 23
## MCMC iteration 7000 #-- Nov 30 13:31:45 --# nbasis: 23
## MCMC iteration 8000 #-- Nov 30 13:31:47 --# nbasis: 25
## MCMC iteration 9000 #-- Nov 30 13:31:50 --# nbasis: 22
## MCMC iteration 10000 #-- Nov 30 13:31:54 --# nbasis: 23
```

```
plot(b)
```

Training Fit

Posterior mean residuals
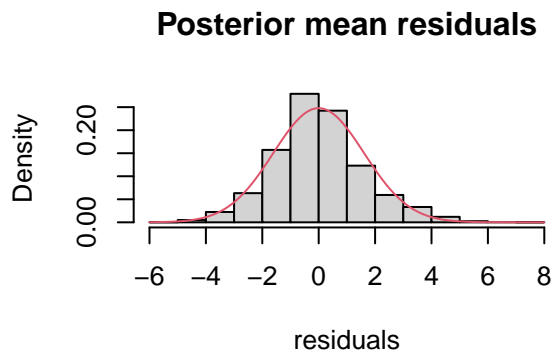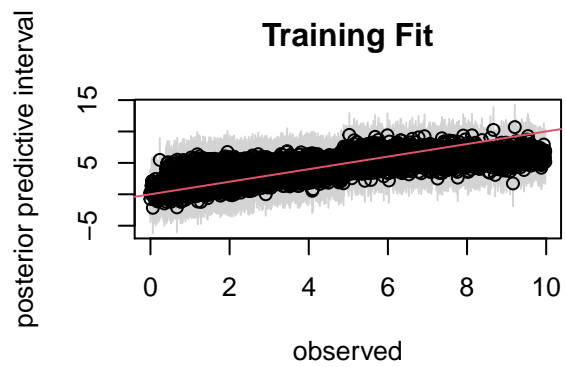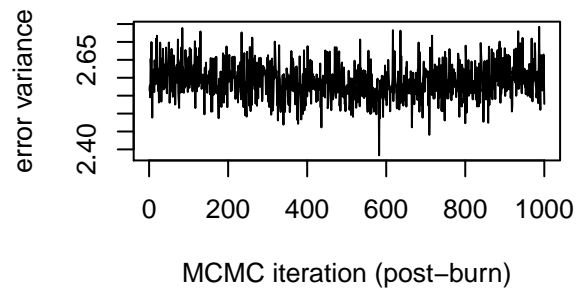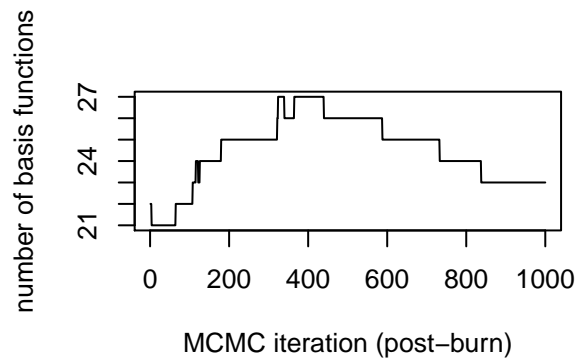
```r
testp <- ttest %>% select(-growth_2_6, -PublishedDate)
p <- predict(b, testp)
print(
  RMSE <- sqrt(mean(sum(ttest$growth_2_6 - p)^2))
)
```

```
## [1] 15486.6
```