Assignment 2 - Part 1B

Generalised additive models

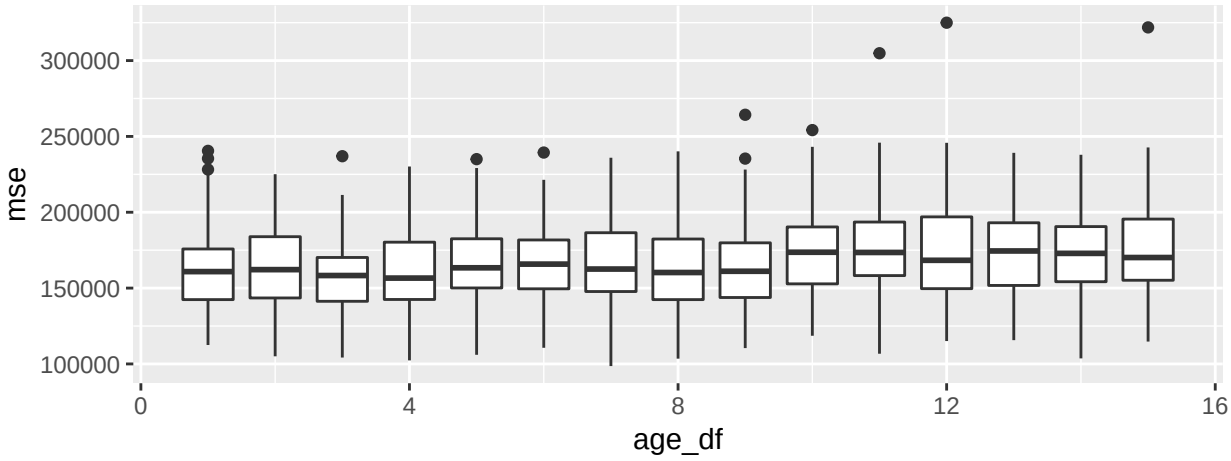
Fernando Cagua

1 Mean squared error vs. age

```
Credit <- read_csv("Credit.csv")
set.seed(2017)
ex <- resample_partition(Credit, p = c(train = 0.5, test = 0.5))
gam_age <- function(age_df, n){
  # monte carlo cross validation of the train data
  cv_ex <- crossv_mc(as.data.frame(ex$train), n)
  models <- map(cv_ex$train,
                ~ gam(balance ~ ns(income, df = 4) + ns(age, df = age_df) + student,
                      data = .))
  errors <- map2_dbl(models, cv_ex$test, rmse)
  errors^2
}
# range of degrees of freedom to evaluate
dfs <- 1:15
# calculate MSEs
cv_age_df <- dfs %>%
  plyr::ldply(function(x) gam_age(x, n = 100)) %>%
  mutate(age_df = dfs) %>%
  reshape2::melt("age_df", value.name = "mse")
```

First I varied the degrees of freedom of age and compared it with the MSE, and could't find a strong pattern. Tweaking the seed used to split between the train and test data revealed high variability. K fold cross validation using 10 folds had a similar problem. So so I decided to use monte carlo cross validaton with 100 partitions.

```
cv_age_df %>%
  ggplot(aes(x = age_df, y = mse)) +
  geom_boxplot(aes(group = age_df))
```



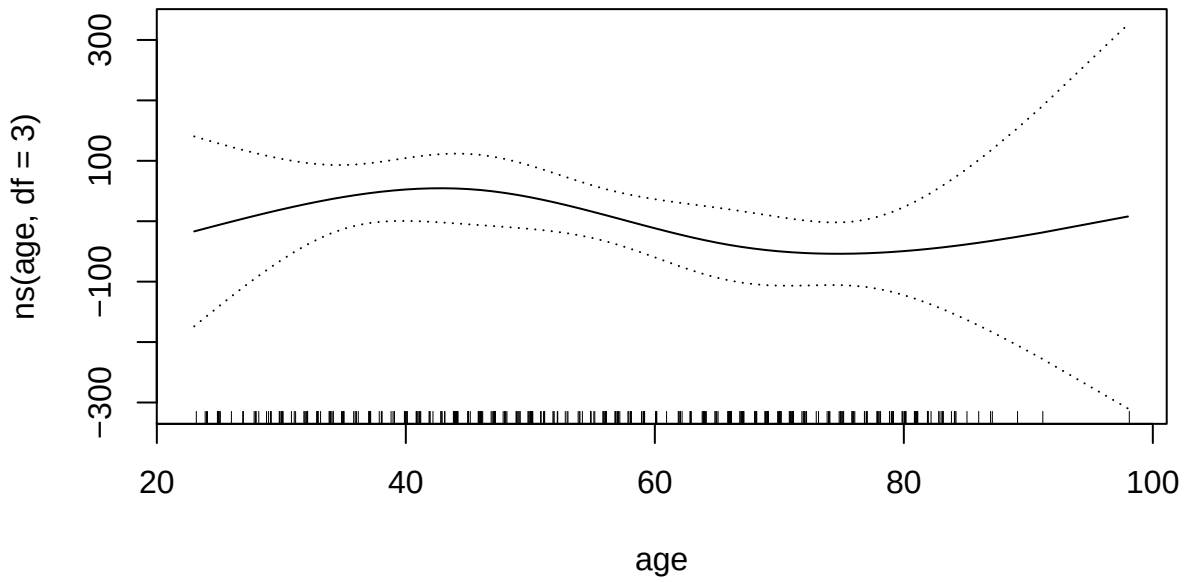
The analysis reveals that there is little difference between the different degrees of freedom although it tends to increasse more substantially after the degrees of freedom are larger than 10.

2 Chose a model

```
# calculate mean MSEs per degree of freedom
mean_mse <- cv_age_df %>% group_by(age_df) %>% summarise(mse = mean(mse)) %>% arrange(mse)
```

We chose the model with the smallest mean MSE, which here corresponds to the one that models age with a natural spline with 3 degrees of freedom.

```
gam(balance ~ ns(income, df = 4) + ns(age, df = 3) + student,
                      data = Credit) %>%
  plot(se = TRUE, terms = "ns(age, df = 3)")
```



The balance seems to increasse between ~20 to ~40 years and then decreasses slightly until ~70 years after which it starts to increasse again. However, the size of the effect vs the standard error suggest a relatively unimportant effect of age to balance.

3 Comparing RMSE with response

The mean RMSE of the choosen model is approximately 399. On the other hand balance ranges between 0 and 1999 with a median of 459.5.

This seems to indicate that the model is performing poorly. Being ~400 dollars off when the typical account has ~460 dollars doesn