NBA Salary Prediction Part 2

Eric Cheung

```
library(tidyverse)
library(glmnet)
library(scales)
```

```
nba_data <- read_csv("nba_2022-23_all_stats_with_salary.csv")
```

```
New names:
Rows: 467 Columns: 52
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(3): Player Name, Position, Team dbl (49): ...1, Salary, Age, GP, GS, MP, FG,
FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA...
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...1`
```

```
nba_data <- nba_data |> select(-1, -2, -4, -6)
nba_data <- nba_data |>
  mutate(across(everything(), ~ replace_na(.x, 0)))
nba_data$Salary <- log(nba_data$Salary)
```

```
set.seed(2003125)
n <- nrow(nba_data)
n_fold <- ceiling(n / 10)
order_ids <- rep(1:10, times = n_fold)
order_ids <- order_ids[1:n]
shuffle <- sample.int(n)
shuffled_ids <- order_ids[shuffle]
```

```r
data <- nba_data
data$fold <- shuffled_ids
```

```r
get_MSPE <- function(Y, Y_hat) {
  residuals <- Y - Y_hat
  resid_sq <- residuals^2
  SSPE <- sum(resid_sq)
  MSPE <- SSPE / length(Y)
  return(MSPE)
}
```

```r
CV_MSPEs <- array(0, dim = c(10, 4))
colnames(CV_MSPEs) <- c("Linear", "Linear-all","LASSO-min", "LASSO-1SE")

for (i in 1:10) {
  data_train <- filter(data, fold != i)
  data_valid <- filter(data, fold == i)
  y_train <- data_train$Salary
  y_valid <- data_valid$Salary
  n_train <- nrow(data_train)

  linear_all <- lm(Salary ~ ., data = data_train)
  linear_mod <- lm(Salary ~ PTS + TRB + AST + Age + MP, data = data_train)

  y <- data_train$Salary
  x <- as.matrix(data_train[, 2:48])
  lasso_mod <- cv.glmnet(y = y, x = x, family = "gaussian")

  # Predict

  pred_lin <- predict(linear_mod, data_valid)
  pred_lin_all <- predict(linear_all, data_valid)

  x_pred <- as.matrix(data_valid[, 2:48])
  pred_lasso_min <- predict(lasso_mod, newx = x_pred, s = lasso_mod$lambda.min)
  pred_lasso_1se <- predict(lasso_mod, newx = x_pred, s = lasso_mod$lambda.1se)

  CV_MSPEs[i, "Linear"] <- get_MSPE(data_valid$Salary, pred_lin)
  CV_MSPEs[i, "Linear-all"] <- get_MSPE(data_valid$Salary, pred_lin_all)
  CV_MSPEs[i, "LASSO-min"] <- get_MSPE(data_valid$Salary, pred_lasso_min)
  CV_MSPEs[i, "LASSO-1SE"] <- get_MSPE(data_valid$Salary, pred_lasso_1se)
```

```
}
```

```
colMeans(CV_MSPEs)
```

```
   Linear Linear-all  LASSO-min  LASSO-1SE
1.2353022  1.1149152  0.9838478  1.0661254
```

```
lasso_coefs_min <- as.data.frame(as.matrix(coef(lasso_mod, s = lasso_mod$lambda.min)))
lasso_coefs_min <- lasso_coefs_min |>
  rownames_to_column(var = "Variable")
colnames(lasso_coefs_min) <- c("Variable", "Coefficients")

lasso_coefs_min |>
  filter(Coefficients != 0) |>
  arrange(desc(Coefficients))
```

```
      Variable Coefficients
1  (Intercept) 10.500667507
2          STL  0.377241909
3          Age  0.103270433
4          2PA  0.078877874
5          DRB  0.040281866
6          FGA  0.029196685
7           GP  0.021958361
8          FTA  0.010035547
9           GS  0.001196419
10         FTr -0.206820892
```

```
# lasso_coefs_1se$Variable <- rownames(lasso_coefs_1se)
# colnames(lasso_coefs_1se) <- c("Coefficient", "Variable")
# lasso_coefs_1se |>
#   filter(Coefficient != 0) |>
#   select(-2)
```

```
lasso_coefs_1se <- as.data.frame(as.matrix(coef(lasso_mod, s = lasso_mod$lambda.1se)))
lasso_coefs_1se <- lasso_coefs_1se |>
  rownames_to_column(var = "Variable")
colnames(lasso_coefs_1se) <- c("Variable", "Coefficients")
```

```r
lasso_coefs_1se |>
  filter(Coefficients != 0) |>
  arrange(desc(Coefficients))
```

```
       Variable Coefficients
1   (Intercept) 1.136872e+01
2           STL 2.392721e-01
3           Age 7.990820e-02
4           2PA 6.026900e-02
5           FGA 3.058510e-02
6           DRB 2.433442e-02
7            GP 1.638091e-02
8 Total Minutes 1.767068e-04
```