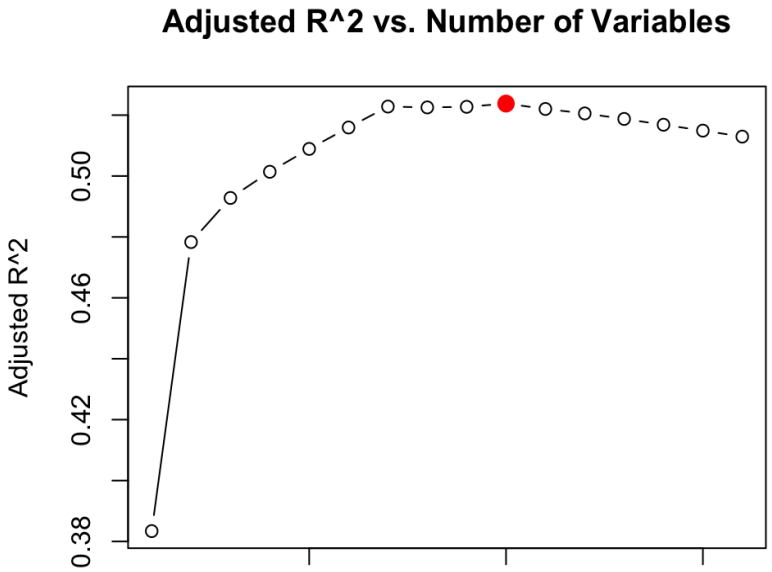
Probability and Statistics - AUC - Baseball salaries

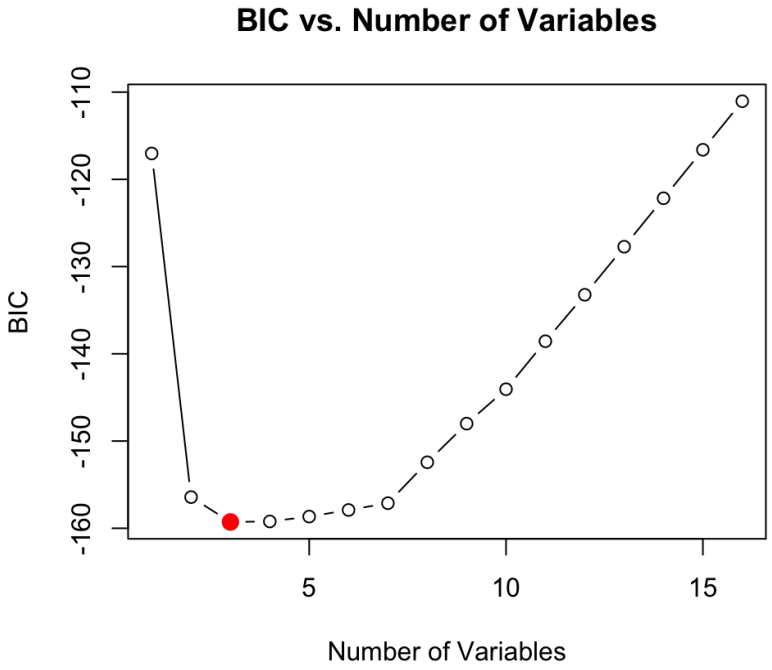
Group 4 - Maya Ozbayoglu and Mafalda Candal

========== Question 1 ==========

(a)



Optimal model size based on Adjusted R^2: 10.



Optimal model size based on BIC: 3.

(b)

Coefficients of the Best Model Based on Adjusted R^2:

- Intercept: 4.2436600051
- AtBat: -0.0026777520
- Hits: 0.0131059039
- HmRun: 0.0127882359
- Walks: 0.0096372429
- Years: 0.0624748890
- CAtBat: 0.000991275
- CRBI: -0.0006153286
- CWalks: -0.0011104349
- PutOuts: 0.0003262487
- Assists: 0.0007131006

Coefficients of the Best Model Based on BIC:

- Intercept: 3.979087983
- Hits: 0.010304524
- Years: 0.098132155
- PutOuts: 0.000330865

Similarities: Both models include the following predictors: Hits, Years, and PutOuts.

Differences: The model based on Adjusted R^2 includes additional predictors such as AtBat, HmRun, Walks, CAtBat, CRBI, CWalks, and Assists, which are not present in the model based on BIC.

Explanation:

The differences arise due to the different criteria for model selection. Adjusted R^2 aims to maximize the explanatory power while penalizing the number of predictors, thus allowing more predictors if they significantly increase the model's explanatory power. In contrast, BIC incorporates a stronger penalty for model complexity, leading to simpler models with fewer predictors. BIC prioritizes parsimony to avoid overfitting, while Adjusted R^2 focuses on maximizing the fit of the model.

(c)

The determination coefficient for the model based on Adjusted R^2 is **0.626**, which means this model explains 62.6% of the variance in Salary.
The determination coefficient for the model based on BIC is **0.585**, meaning this model explains 58.5% of the variance in Salary.

The Adjusted R^2 model includes 10 predictors, while the BIC model includes 3 predictors. This means the increase of 4% in explanatory power comes at the expense of 7 more predictors.
Given the small increase, it might not justify selecting the more complex model, especially considering the risks of overfitting and the benefits of model simplicity and interpretability. Thus, unless the additional complexity addresses a specific need or significantly improves prediction in a practical sense, the simpler model chosen by BIC would be preferable.

(d)

Comparison with Models from (b)

Model Based on Adjusted R^2:

- **Forward Selection:** Includes `AtBat`, `Hits`, `HmRun`, `Walks`, `Years`, `CAtBat`, `CRBI`, `CWalks`, `PutOuts`, `Assists`.
- **Best Subset:** Same predictors as forward selection.

Model Based on BIC:

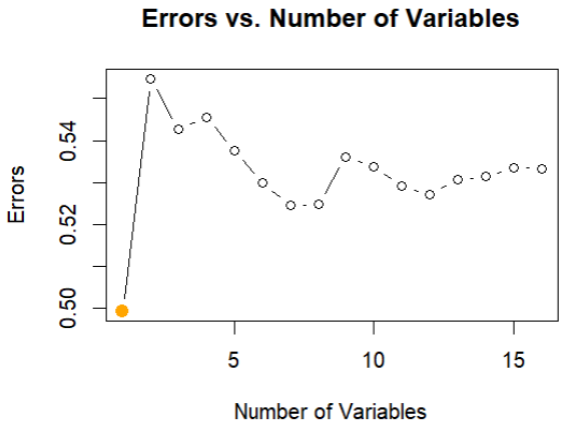
- **Forward Selection:** Includes `Hits`, `Years`, `PutOuts`.
- **Best Subset:** Same predictors as forward selection.

There are no differences between the models obtained through forward selection and those obtained through best subset selection for both Adjusted R^2 and BIC criteria.

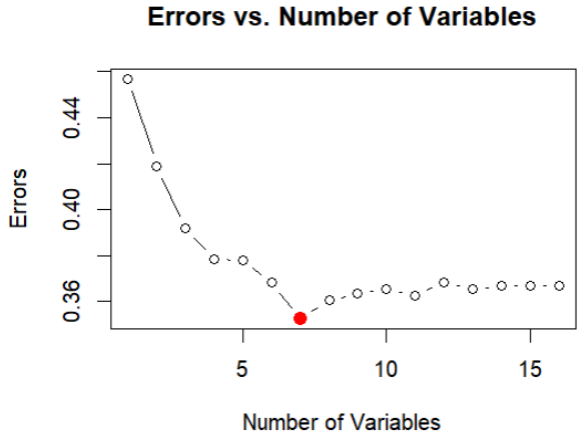
The forward selection method yielded the same optimal models as the best subset method for both Adjusted R^2 and BIC criteria, indicating consistency in the model selection process across these methods.

========== Question 2 ==========

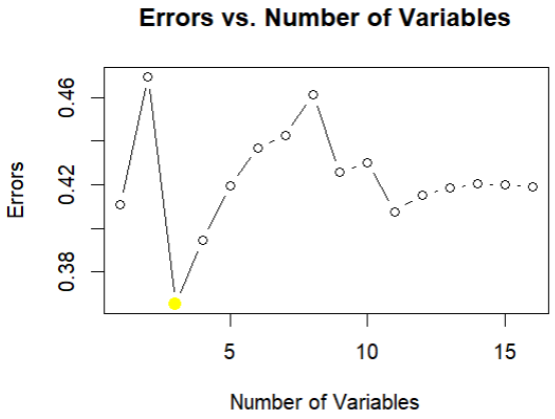
(a)



(b)



Seed 2

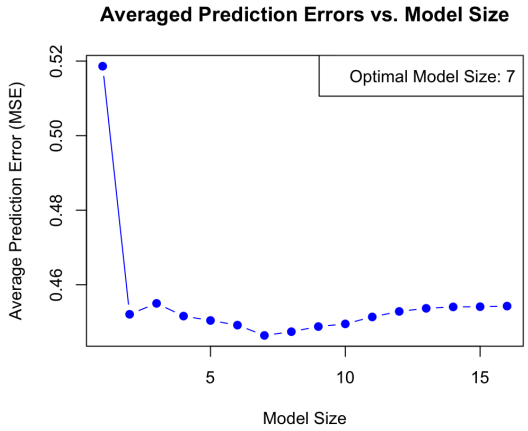


Seed

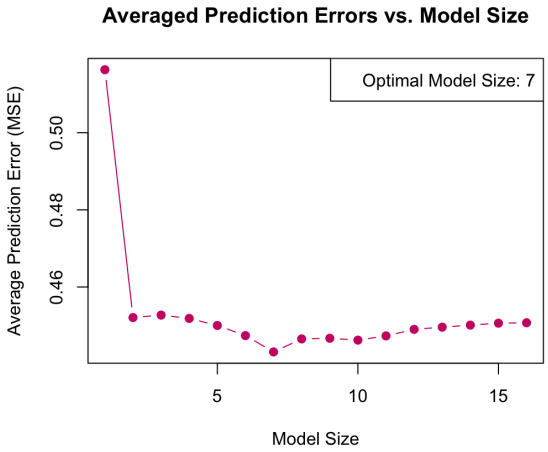
(c)

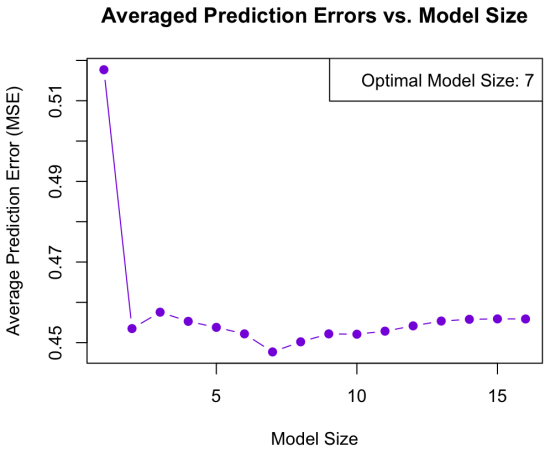
The large variability in the optimal number of predictors can be explained by the randomness in the data splitting process, which leads to different training and validation sets each time. This variability results in different subsets of predictors being selected as optimal due to fluctuations in model performance across different splits. Other contributing factors include sample size, presence of noise and outliers, and multicollinearity among predictors. These elements can cause certain predictors to appear more or less important in different splits, thus affecting the model selection process. Using multiple cross-validation runs and averaging the results can help reduce this variability and lead to a more consistent determination of the optimal number of predictors.

(d)



(e)





Repeating the cross-validation with 500 random splits and adding the results to the plot consistently identifies the optimal number of predictors as 7. This consistency can be explained by the central limit theorem, which indicates that increasing the number of splits reduces the variability and noise in the prediction errors. Averaging the errors over a large number of splits provides a more stable and reliable estimate of the optimal model size, ensuring that the results are robust regardless of the specific set of 500 splits used.

This approach confirms that the optimal model size of 7 is a reliable choice, as it consistently emerges as the best model size across different sets of splits, validating the robustness of the cross-validation process.

```
#install.packages("tidyverse")
#install.packages("leaps")
#install.packages("ISLR")

library(tidyverse)
library(leaps)
library(ISLR)

data <- Hitters
str(data)
data<-na.omit(data)

# Goal: You are going to use different methods to select a linear regression
# model for explaining the observed variation in salary across players.

data$League<-NULL
data$Division<-NULL
data$NewLeague<-NULL

data$Salary<-log(data$Salary)

# PART I: Best subset method
best_subset <- regsubsets(Salary ~ ., data = data, nvmax = 16)
summary(best_subset)

results<-summary(best_subset)
results$adjr2
results$bic

# a)

# Plot Adjusted R^2
plot(results$adjr2, type = "b", xlab = "Number of Variables", ylab = "Adjusted R^2",
    main = "Adjusted R^2 vs. Number of Variables")
points(which.max(results$adjr2), max(results$adjr2), col = "red", cex = 2, pch = 20)

# Plot BIC
plot(results$bic, type = "b", xlab = "Number of Variables", ylab = "BIC",
    main = "BIC vs. Number of Variables")
points(which.min(results$bic), min(results$bic), col = "red", cex = 2, pch = 20)

# Optimal model size based on adjusted R^2
optimal_adjr2_size <- which.max(results$adjr2)
cat("Optimal model size based on adjusted R^2:", optimal_adjr2_size, "\n")
```

```
# Optimal model size based on BIC
optimal_bic_size <- which.min(results$bic)
cat("Optimal model size based on BIC:", optimal_bic_size, "\n")

# b)

# 5. Extract Coefficients for Optimal Models**:
# Coefficients of the model with the highest adjusted R^2
best_adjr2_model <- coef(best_subset, optimal_adjr2_size)
print("Coefficients of the best model based on adjusted R^2:")
print(best_adjr2_model)

# Coefficients of the model with the lowest BIC
best_bic_model <- coef(best_subset, optimal_bic_size)
print("Coefficients of the best model based on BIC:")
print(best_bic_model)

# c)
print("Determination coefficient of the optimal adjusted R^2:")
print((summary(best_subset)$rsq[optimal_adjr2_size]))

print("Determination coefficient of the optimal BIC")
print((summary(best_subset)$rsq[optimal_bic_size]))

# d)
# Forward Selection
forward_selection <- regsubsets(Salary ~ ., data = data, nvmax = 16, method = "forward")
forward_summary <- summary(forward_selection)

# Optimal models from forward selection
optimal_forward_adjr2_size <- which.max(forward_summary$adjr2)
optimal_forward_bic_size <- which.min(forward_summary$bic)

best_forward_adjr2_model <- coef(forward_selection, optimal_forward_adjr2_size)
best_forward_bic_model <- coef(forward_selection, optimal_forward_bic_size)

cat("Optimal model size based on adjusted R^2 (Forward Selection):",
optimal_forward_adjr2_size, "\n")
cat("Optimal model size based on BIC (Forward Selection):", optimal_forward_bic_size,
"\n")
print("Coefficients of the best model based on adjusted R^2 (Forward Selection):")
print(best_forward_adjr2_model)
print("Coefficients of the best model based on BIC (Forward Selection):")
```

```
print(best_forward_bic_model)

# Comparison of models
print("Comparison of Best Subset and Forward Selection Models based on Adjusted R^2:")
cat("Best Subset Adjusted R^2 Model Size:", optimal_adjr2_size, "\n")
cat("Forward Selection Adjusted R^2 Model Size:", optimal_forward_adjr2_size, "\n")
print("Best Subset Adjusted R^2 Model Coefficients:")
print(best_adjr2_model)
print("Forward Selection Adjusted R^2 Model Coefficients:")
print(best_forward_adjr2_model)

print("Comparison of Best Subset and Forward Selection Models based on BIC:")
cat("Best Subset BIC Model Size:", optimal_bic_size, "\n")
cat("Forward Selection BIC Model Size:", optimal_forward_bic_size, "\n")
print("Best Subset BIC Model Coefficients:")
print(best_bic_model)
print("Forward Selection BIC Model Coefficients:")
print(best_forward_bic_model)



# PART II: Cross-validation

# a)
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(data), replace = TRUE, prob = c(0.6, 0.4))
train <- data[sample, ] # training set
valid <- data[!sample, ] # validation set

best_subset <- regsubsets(Salary ~ ., train, nvmax = 19)
X_valid <- model.matrix(Salary ~ ., data = valid)

# Initialization of a vector that will contain the errors
errors <- rep(0, 16)

# Loop over all model sizes from 1 to 16
for (i in 1:16) {
  # Extract coefficients for model with i predictors
  coef_x <- coef(best_subset, i)
  # Predict response variable (Salary)
  pred_x <- X_valid[, names(coef_x)] %*% coef_x
  # Compute mean squared prediction error
  errors[i] <- mean((valid$Salary - pred_x)^2)
}
```

```
# Define model sizes
model_sizes <- 1:16

# Plot Errors
plot(model_sizes, errors, type = "b", xlab = "Number of Variables", ylab = "Errors",
     main = "Errors vs. Number of Variables")
points(which.min(errors), min(errors), col = "yellow", cex = 2, pch = 20)



# b) Repeat cross-validation for two different seed values and plot the results
# We only need to run the code snippet for 2(a) changing the set.seed(1) for
# 2 and 3.



# d)
set.seed(1)
n_splits <- 500
max_model_size <- 16
errors_matrix <- matrix(0, nrow = n_splits, ncol = max_model_size)

for (split in 1:n_splits) {
  # Split the data into training and validation sets
  sample <- sample(c(TRUE, FALSE), nrow(data), replace = TRUE, prob = c(0.6, 0.4))
  train <- data[sample, ]
  valid <- data[!sample, ]

  # Fit the best subset selection model
  best_subset <- regsubsets(Salary ~ ., train, nvmax = max_model_size)
  X_valid <- model.matrix(Salary ~ ., data = valid)

  for (i in 1:max_model_size) {
    # Extract coefficients for the model with i predictors
    coef_x <- coef(best_subset, id = i)

    # Predict response variable (Salary)
    pred_x <- X_valid[, names(coef_x)] %*% coef_x

    # Compute mean squared prediction error
    errors_matrix[split, i] <- mean((valid$Salary - pred_x)^2)
  }
}
```

```
# Compute average errors over all splits
average_errors <- colMeans(errors_matrix)

# Plot the averaged errors as a function of model size
model_sizes <- 1:max_model_size
plot(model_sizes, average_errors, type = "b", pch = 19, col = "deeppink3",
     xlab = "Model Size", ylab = "Average Prediction Error (MSE)",
     main = "Averaged Prediction Errors vs. Model Size")

# Identify the optimal model size
optimal_model_size <- which.min(average_errors)

# Add a legend
legend("topright", legend = paste("Optimal Model Size:", optimal_model_size))


#(e)
#Repeat the 2(a) changing the set.seed(1) for 2 and 3.
```