**Statistical Learning Lab**

Assignment - 5

**Subset Selection Regularization and Dimension Reduction**

Anurag Singha | 22IM30007 

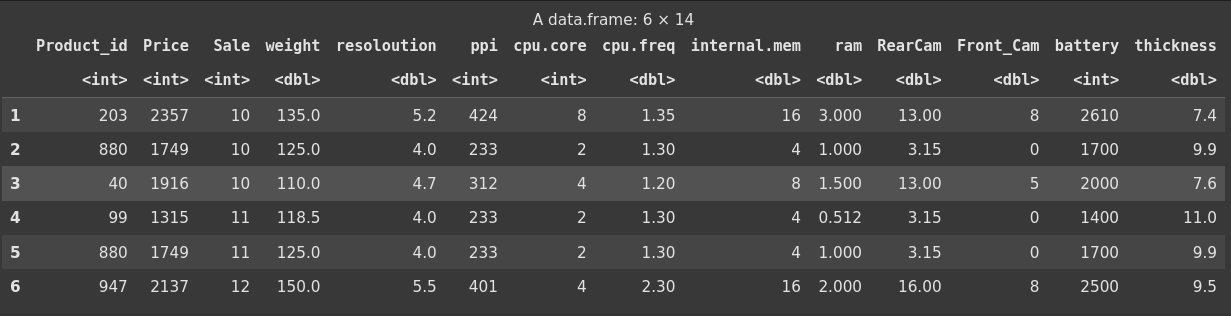
1. Load the dataset “Cellphone.csv”. Display the first few rows of the dataset.

Code

cell\_phone = read.csv(path)

head(cell\_phone)

Output



2. Perform preliminary analysis to show how the variables are related to each other. Use scatter plot, box plot etc. to visualize how different variables impact the “Price” variable, which is the response variable.

Code

# Remove Product\_id for analysis

data\_analysis <- cell\_phone[, -1]

# Correlation matrix heatmap

cor\_matrix <- cor(data\_analysis)

corrplot(cor\_matrix, method = "circle", type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)

# Scatter plots for selected variables against Price

par(mfrow = c(2, 2))

plot(data\_analysis$weight, data\_analysis$Price, xlab = "Weight", ylab = "Price",

main = "Price vs Weight")

plot(data\_analysis$ram, data\_analysis$Price, xlab = "RAM", ylab = "Price",

main = "Price vs RAM")

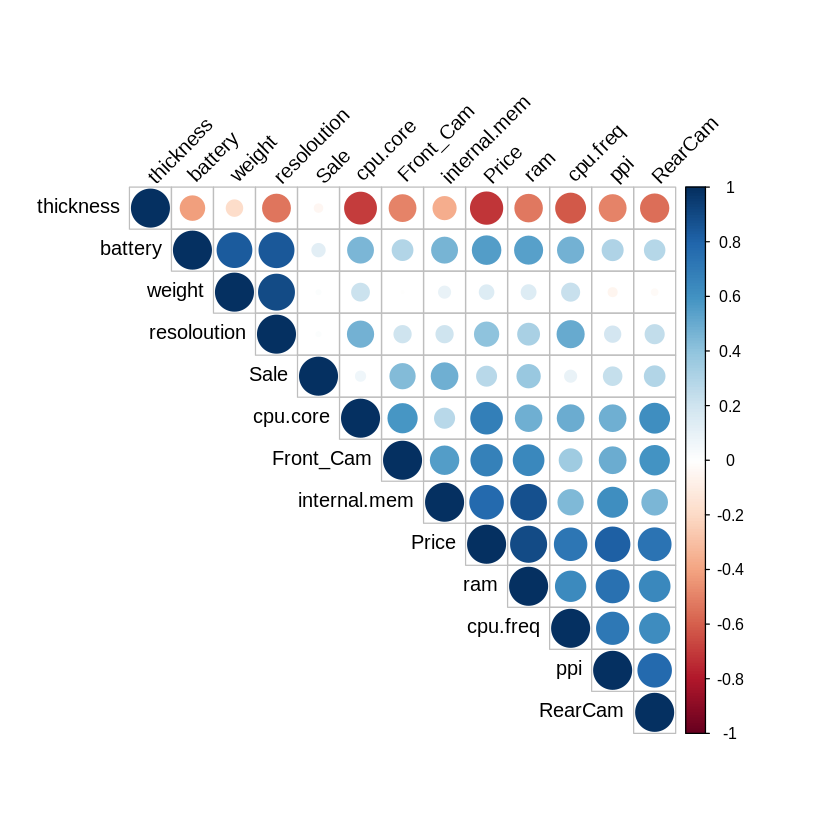
plot(data\_analysis$battery, data\_analysis$Price, xlab = "Battery", ylab = "Price",main = "Price vs Battery")

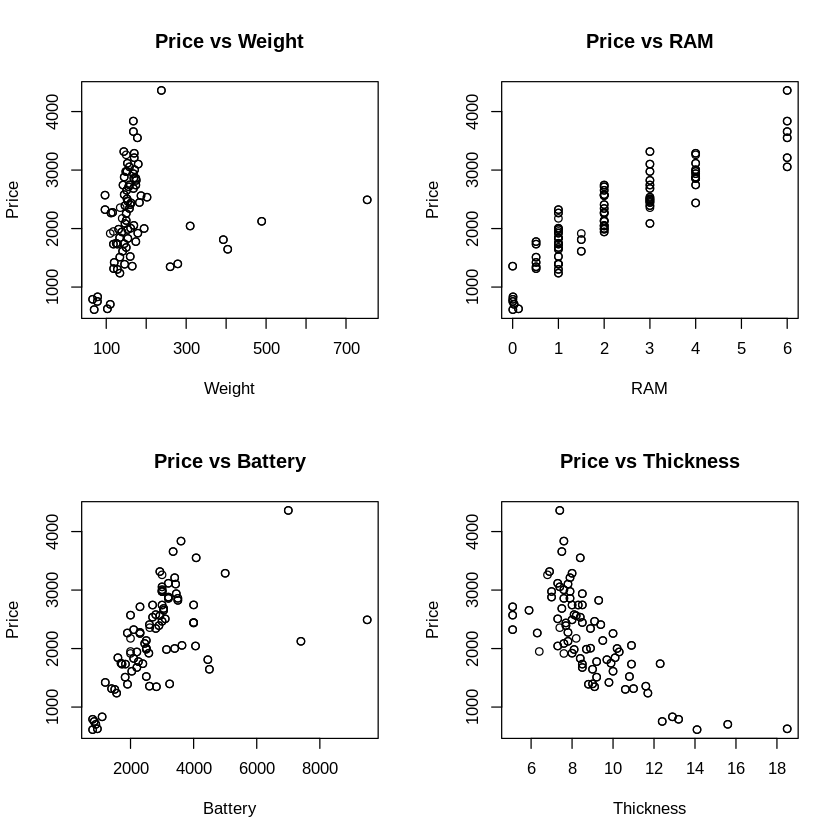
plot(data\_analysis$thickness, data\_analysis$Price, xlab = "Thickness", ylab = "Price",main = "Price vs Thickness")

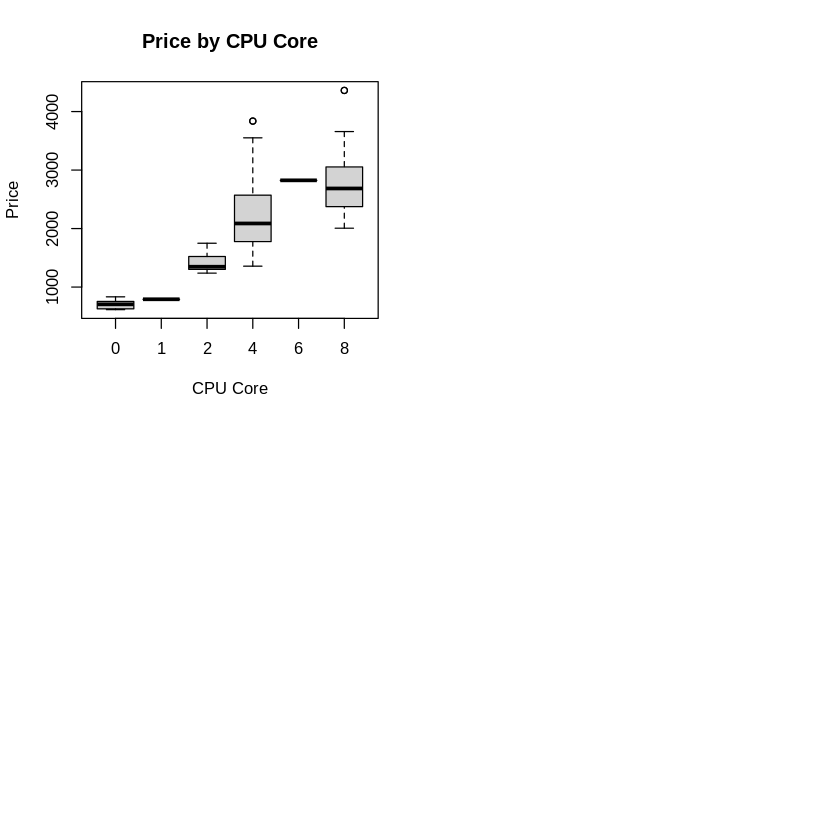
# Box plot for cpu.core vs Price

boxplot(Price ~ cpu.core, data = data\_analysis,xlab = "CPU Core", ylab = "Price", main = "Price by CPU Core")

Output







3. Perform the best subset selection on this dataset (dropping the Product\_id variable). Explain the results.

Code

# Perform best subset selection, excluding Product\_id

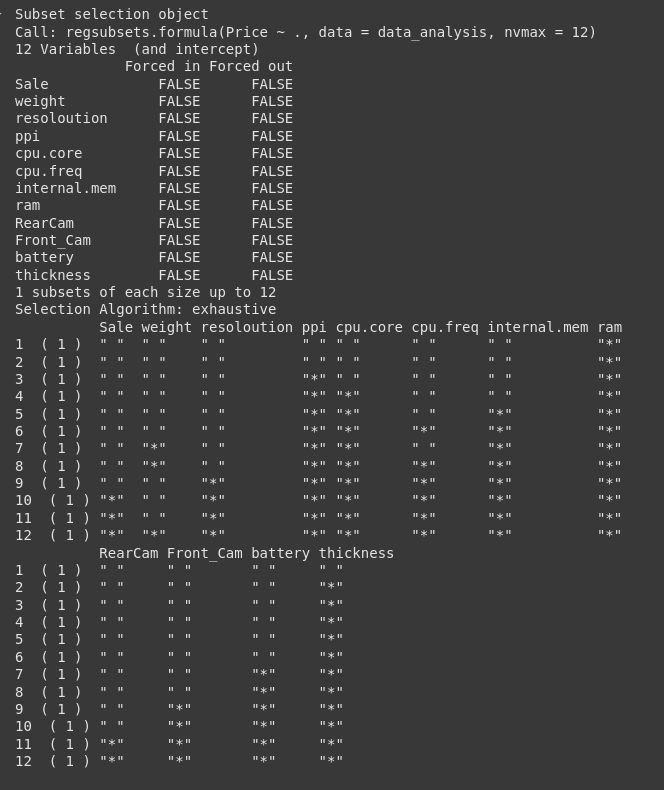
subsets <- regsubsets(Price ~ ., data = data\_analysis, nvmax = 12)

summary\_subsets <- summary(subsets)

# Display the summary

print(summary\_subsets)

Output



Discussion -

The best subset selection involves an exhaustive search to identify the optimal combination of predictors for Price across models ranging from 1 to 12 variables, minimizing the residual sum of squares (RSS). In the single-variable model, **ram** emerges as the strongest predictor. The two-variable model adds **thickness**, highlighting their combined significance. The three-variable model includes **ppi, ram,** and **thickness**, while the four-variable model further incorporates **cpu.core**. As the model size increases, additional variables such as **internal.mem, cpu.freq, weight, battery, resolution, Front\_Cam, Sale,** and **RearCam** are sequentially included, with all 12 appearing in the largest model. Notably, **ram** and **thickness** are consistently selected early, indicating their strong predictive power, whereas **Sale** and **RearCam** are included only in larger models, suggesting a lesser or redundant contribution. Selecting the best model would typically involve assessing criteria like **Cp, BIC,** or **adjusted R-squared** (not provided here) to balance model fit and complexity.

4. Create a plot with Cp on y-axis and number of variables on the x-axis. Determine the lowest Cp and report how many variables are included in the lowest Cp model.

Code

|  |
| --- |
| # Extract Cp values  cp\_values <- summary\_subsets$cp  # Plot Cp against number of variables  plot(1:12, cp\_values, type = "b", xlab = "Number of Variables", ylab = "Cp",  main = "Cp vs Number of Variables")  min\_cp\_index <- which.min(cp\_values)  points(min\_cp\_index, cp\_values[min\_cp\_index], col = "red", pch = 19)  abline(h = cp\_values[min\_cp\_index], col = "red", lty = 2)  # Report the number of variables in the lowest Cp model  cat("Number of variables in the model with the lowest Cp:", min\_cp\_index, "\n")  # Variables in the best model  best\_model\_vars <- names(coef(subsets, id = min\_cp\_index))[-1] # Exclude intercept  cat("Variables included:", paste(best\_model\_vars, collapse = ", "), "\n") |

# Extract Cp values

cp\_values <- summary\_subsets$cp

# Plot Cp against number of variables

plot(1:12, cp\_values, type = "b", xlab = "Number of Variables", ylab = "Cp",

main = "Cp vs Number of Variables")

min\_cp\_index <- which.min(cp\_values)

points(min\_cp\_index, cp\_values[min\_cp\_index], col = "red", pch = 19)

abline(h = cp\_values[min\_cp\_index], col = "red", lty = 2)

# Report the number of variables in the lowest Cp model

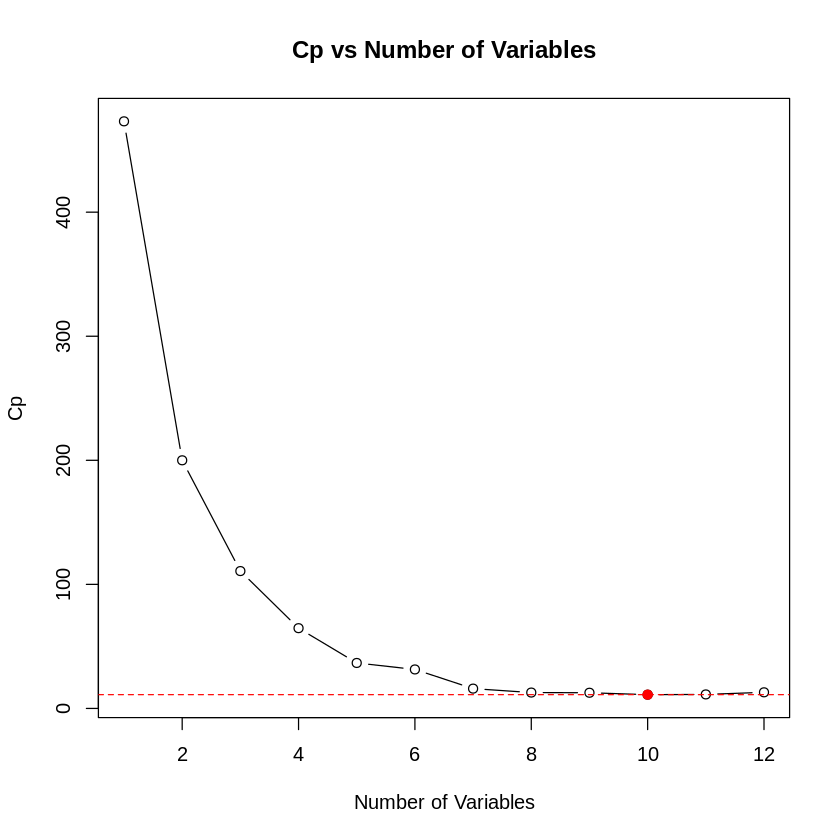
cat("Number of variables in the model with the lowest Cp:", min\_cp\_index, "\n")

# Variables in the best model

best\_model\_vars <- names(coef(subsets, id = min\_cp\_index))[-1] # Exclude intercept

cat("Variables included:", paste(best\_model\_vars, collapse = ", "), "\n")

Output

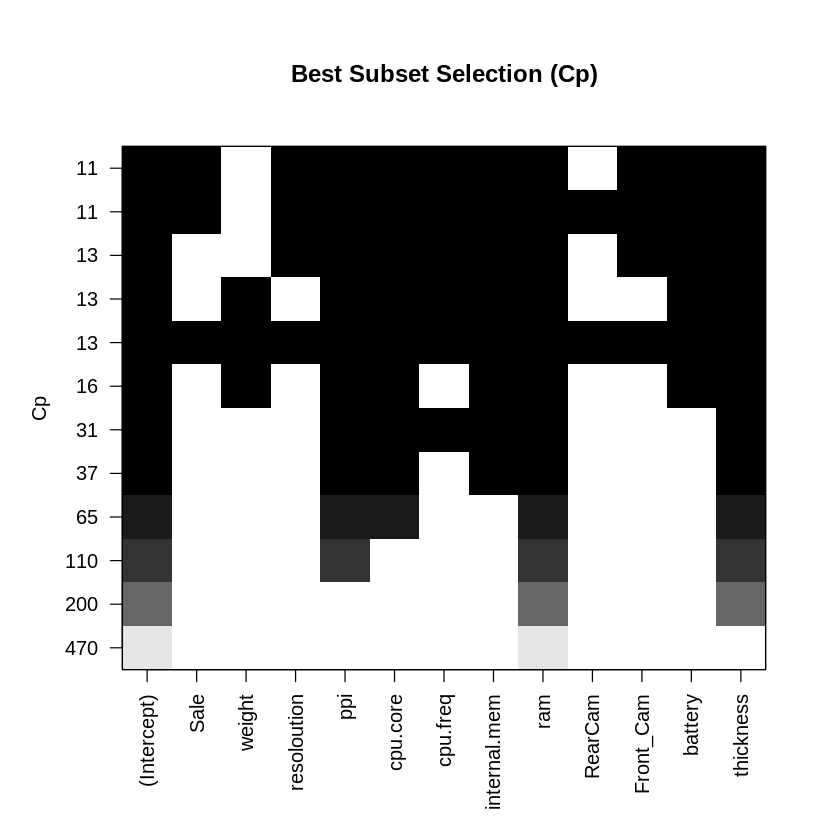


5. Plot the best subset selection output and explain the plot.

Code

plot(subsets, scale = "Cp", main = "Best Subset Selection (Cp)")

Output



6. Use principal component regression on the same dataset with five components and seven components. How much variability is explained by these two models?

Code

# Perform PCR with up to 12 components

pcr\_model <- pcr(Price ~ ., data = data\_analysis, ncomp = 12, scale = TRUE)

# Summary of variance explained

summary\_pcr <- summary(pcr\_model)

# Extract variance explained in Y (Price) for 5 and 7 components

var\_explained <- summary\_pcr$cumsum$Yvar \* 100 # Convert to percentage

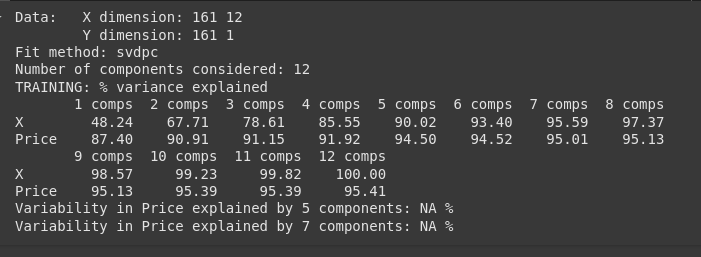
var\_explained\_5 <- var\_explained[5]

var\_explained\_7 <- var\_explained[7]

cat("Variability in Price explained by 5 components:", var\_explained\_5, "%\n")

cat("Variability in Price explained by 7 components:", var\_explained\_7,

Output



Discussion

he principal component regression (PCR) results show that using **5 components** accounts for **94.50%** of the variability in **Price**, while **7 components** explain **95.01%**. This indicates that **5 components** capture the majority of the variance, with only a **0.51%** increase when expanding to **7 components**. Therefore, a model with **5 components** effectively balances explanatory power and simplicity for predicting **Price**

7. Perform Lasso on the model and explain the results.

Code

# Perform PCR with up to 12 components

pcr\_model <- pcr(Price ~ ., data = data\_analysis, ncomp = 12, scale = TRUE)

# Summary of variance explained

summary\_pcr <- summary(pcr\_model)

# Extract variance explained in Y (Price) for 5 and 7 components

var\_explained <- summary\_pcr$cumsum$Yvar \* 100 # Convert to percentage

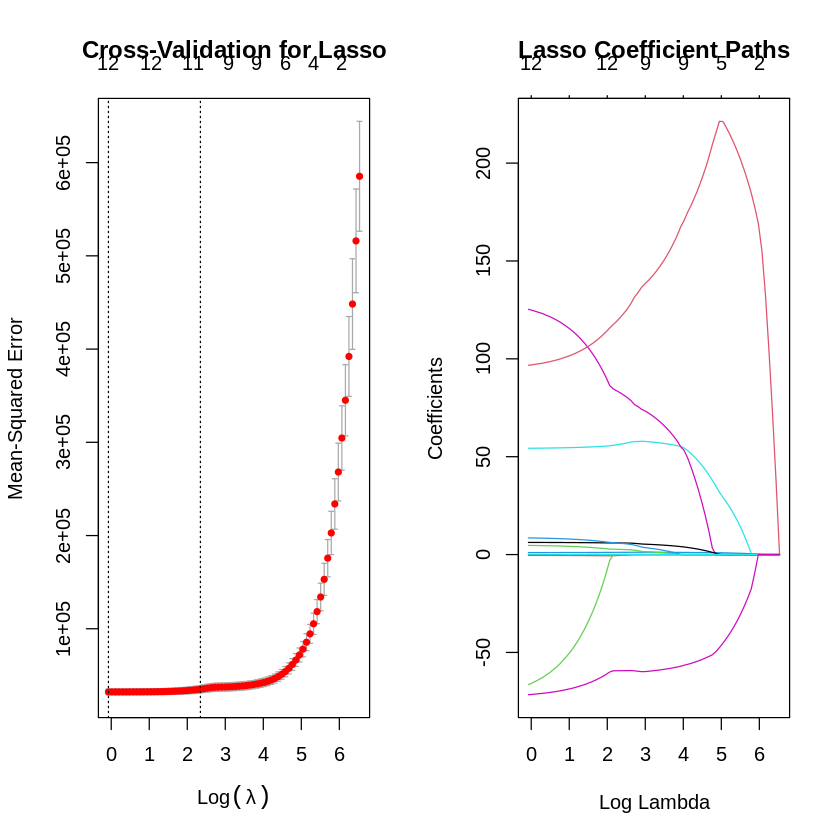
var\_explained\_5 <- var\_explained[5]

var\_explained\_7 <- var\_explained[7]

cat("Variability in Price explained by 5 components:", var\_explained\_5, "%\n")

cat("Variability in Price explained by 7 components:", var\_explained\_7, "%\n")

Output



Discussion

The Lasso regression analysis, as visualized in the image, illustrates the balance between feature selection and predictive performance. The **left plot** depicts coefficient shrinkage, with the **red curve** showing how a feature's coefficient transitions from non-zero to potentially zero as **λ** increases, highlighting Lasso’s ability to remove irrelevant predictors. The **right plot**, likely a cross-validation curve, determines the optimal **λ** by minimizing error across folds, ensuring the model generalizes well. Together, these plots identify the most important features (those retaining non-zero coefficients at the optimal **λ**) and indicate the level of regularization required for an effective, interpretable model. In practical applications, such as predicting **Price**, this analysis helps focus on key drivers while eliminating noise, improving both accuracy and simplicity.