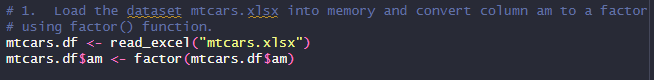
Seif Kungulio  
11/13/2024  
Project 3  
DATA 640  
Section: 01W  
Instructor: Chris Shannon  
File Name: Project3\_Kungulio\_Seif.docx

# PART I.

1. Load the dataset mtcars.xlsx into memory and convert column am to a factor using factor() function.

mtcars.df <- read\_excel("mtcars.xlsx")

mtcars.df$am <- factor(mtcars.df$am)



1. Split the data into training set and test set. The training set contains the first 35 observations, the test set containing the remaining observations.

train\_set <- mtcars.df[1:35, ]

test\_set <- mtcars.df[-(1:35), ]

A blue screen with white text

Description automatically generated

1. Build a logistic regression model with the response is am and the predictors are mpg, cyl, hp, and wt using glm() function

model <- glm(am ~ mpg + cyl + hp + wt, data = train\_set, family = binomial)

A blue and white text

Description automatically generated

1. Compute the test error on the test data set using a confusion matrix. Is it a good model based on test error?

test\_predictions <- predict(model, newdata = test\_set, type = "response")

test\_pred\_class <- ifelse(test\_predictions > 0.5, 1, 0)

# Create the confusion matrix

conf\_matrix <- confusionMatrix(factor(test\_pred\_class), test\_set$am)

print(conf\_matrix)

# Print test error rate

test\_error\_rate <- 1 - sum(diag(conf\_matrix$table)) / sum(conf\_matrix$table)

print(paste("Test Error Rate:", round(test\_error\_rate, 3)))

A computer code with white text

Description automatically generated with medium confidence

# Assess model quality based on test error rate

if (test\_error\_rate < 0.2) {

print("The model is reasonably accurate.")

} else {

print("The model has room for improvement.")

}

A screenshot of a computer error

Description automatically generated

**Explanation**

* **Test Error Rate Interpretation:** A test error rate of 16.7% indicates that 83.3% of the predictions made by the model are correct, while 16.7% are incorrect.
* **Model Evaluation:** Whether this is a "good" model depends on the context and the acceptable level of accuracy for the problem at hand. Generally, a test error rate below 20% could be considered reasonable, depending on the application, but it's not particularly low.

# Part II.

1. Build a linear model to forecast number of total rentals (count) using potential predictors, season, holiday, workingday, weather, atemp, and registered. Note here the linear model is not ideal for predicting count. We can work around this drawback by rounding up or rounding down the predictions. Please read the attached paper for the classical regression models for count data in R.

# Load the dataset

bike.df <- read.csv("Bike.csv")

# Convert datetime to Date format if needed

bike.df$datetime <- as.POSIXct(bike.df$datetime, format = "%Y-%m-%d %H:%M:%S")

# Convert categorical variables to factors

bike.df$season = factor(bike.df$season,

levels = c(1, 2, 3, 4),

labels = c("Spring", "Summer", "Fall", "Winter")

)

bike.df$holiday <- factor(bike.df$holiday,

levels = c(0,1),

labels = c("No", "Yes")

)

bike.df$workingday <- factor(bike.df$workingday,

levels = c(0,1),

labels = c("No", "Yes")

)

bike.df$weather <- factor(bike.df$weather,

levels = c(1, 2, 3, 4),

labels = c("Clear", "Misty\_cloudy",

"Light\_snow", "Heavy\_rain")

)

# Linear model for count prediction

linear\_model <- lm(count ~ season + holiday + workingday +

weather + atemp + registered,

data = bike.df)

# Display the statistical summary of the model

summary(linear\_model)

# Generate predictions using the model and round the predictions

predictions <- predict(linear\_model, bike.df)

rounded\_predictions <- round(predictions)

# Show first few rounded predictions

head(rounded\_predictions)

1. Perform best subset selection using bestglm() function based on BIC. What’s the best model based on BIC?

# Prepare data for bestglm (needs to be a dataframe with only predictors

# and response)

model\_data <- model.matrix(~ season + holiday + workingday + weather +

atemp + registered + count, data = bike.df)

# Remove the first column from the model\_data dataset

model\_data <- model\_data[,-1]

# Display the first few rows

head(model\_data)

# Convert model\_data to dataframe

model\_data.df <- data.frame(model\_data)

# Check to see if model\_data.df is a dataframe

class(model\_data.df)

# Find the best model based on the BIC.

best\_bic\_model <- bestglm(model\_data.df, IC = "BIC",

family = gaussian)

# Display the best model based on the BIC

best\_bic\_model

# Display the statistical summary of the best model of the best\_bic\_model

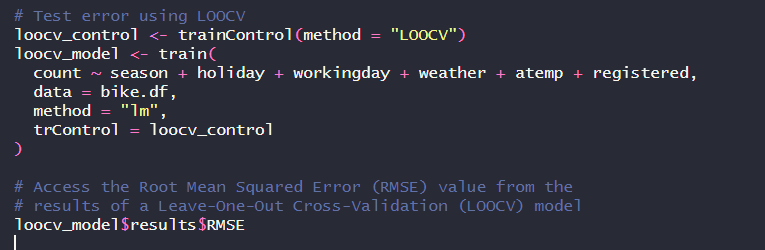
summary(best\_bic\_model$BestModel)

A screenshot of a computer program

Description automatically generated

This model suggests strong predictive performance, with seasonal, weather, holiday, and temperature-related factors all significantly influencing the outcome variable. The model’s high R2 and significance levels indicate it is likely capturing the key drivers in the data effectively.

1. Compute the test error of the best model based on BIC using LOOCV.

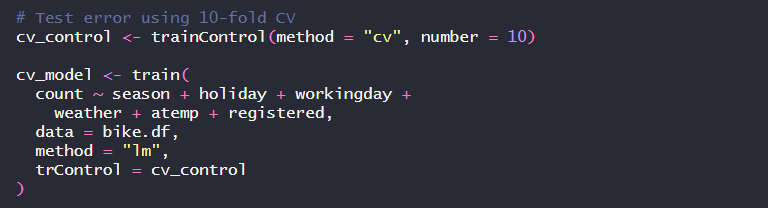


The test error is 35.09127



An RMSE value of 35.09127 indicates that, on average, the model's predictions deviate from the actual values by approximately 35 units. Lower RMSE values generally indicate better model performance.

1. Calculate the test error of the best model based on BIC using 10-fold CV.



The test error is 35.06165



1. Perform best subset selection using bestglm() function based on CV. What’s the best model based on CV?

A computer screen shot of text

Description automatically generated

Statistical summary of the best model of the best\_cv\_model:

A screenshot of a computer program

Description automatically generated

The model’s high R-squared and significant predictors suggest a well-fitting model with meaningful relationships between the predictors and the outcome. However, some predictors, particularly workingdayYes and seasonFall, have large effects that might need careful interpretation to ensure they align with domain expectations or are not due to multicollinearity. Additionally, the wide range in residuals suggests that while the model is overall accurate, there may be outliers or variations that are not captured by these predictors.

1. Perform the backward stepwise selection using stepAIC() function. What’s the best model?

# Backward stepwise selection using stepAIC

stepwise\_model <- stepAIC(linear\_model, direction = "backward")

# Display the statistical summary of the best model

summary(stepwise\_model)

A screenshot of a computer program

Description automatically generated

The model shows that bike rentals are strongly influenced by season, holidays, working days, weather, temperature, and the number of registered users, with some predictors (like "registered users") having a particularly substantial impact. However, occasional outliers suggest there may be other factors not captured by this model.