Linear Regression

Sriram Vivek

2025-02-19

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(MASS)
library(leaps)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Using the random seed 123 to divide the data into 75% training and 25% testing.

```
ames_data <- read.csv("/Users/sriram/Desktop/SEMESTER 2/AMS 580/Linear Regression/Ames_Housing_Data.csv


ames_data <- na.omit(ames_data)


set.seed(123)


index <- createDataPartition(ames_data$SalePrice, p = 0.75, list = FALSE)
train_data <- ames_data[index, ]
test_data <- ames_data[-index, ]
```

Finding the best model using the stepwise variable selection method (based on the BIC criterion) using the training data and displaying necessary information.

```
full_model <- lm(SalePrice ~ ., data = train_data)


step_model <- stepAIC(full_model, direction = "both", k = log(nrow(train_data)), trace = FALSE)


cat("Stepwise Model Coefficients:\n")
```

```
## Stepwise Model Coefficients:
```

```
print(coef(step_model))
```

```
##   (Intercept)       LotArea   OverallQual   OverallCond     YearBuilt
## -1.017464e+06  7.339645e-01  1.667461e+04  5.943478e+03  4.812917e+02
##     X1stFlrSF     X2ndFlrSF  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd
##  1.013470e+02  6.477953e+01 -1.488137e+04 -3.342536e+04  4.361293e+03
##    GarageArea
##  3.559750e+01
```

```
step_predictions <- predict(step_model, newdata = test_data)


step_rmse <- sqrt(mean((test_data$SalePrice - step_predictions)^2))
step_r_squared <- cor(test_data$SalePrice, step_predictions)^2

cat("\nStepwise Model Performance:\n")
```

```
##
## Stepwise Model Performance:
```

```
cat("RMSE:", step_rmse, "\n")
```

```
## RMSE: 48727.68
```

```
cat("R^2:", step_r_squared, "\n")
```

```
## R^2: 0.6601152
```

Finding the best model using the best subset variable selection method (based on the SSE criterion) and siplaying the necessary information.

```
subset_model <- regsubsets(SalePrice ~ ., data = train_data, nvmax = 20)

best_subset_index <- which.min(summary(subset_model)$bic)
best_subset_vars <- names(coef(subset_model, id = best_subset_index))[-1]  # Exclude intercept

best_subset_formula <- as.formula(paste("SalePrice ~", paste(best_subset_vars, collapse = " + ")))
best_subset_model <- lm(best_subset_formula, data = train_data)

cat("\nBest Subset Model Coefficients:\n")
```

```
##
## Best Subset Model Coefficients:
```

```
print(coef(best_subset_model))
```

```
##   (Intercept)       LotArea   OverallQual   OverallCond     YearBuilt
## -1.017464e+06  7.339645e-01  1.667461e+04  5.943478e+03  4.812917e+02
##     X1stFlrSF     X2ndFlrSF  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd
##  1.013470e+02  6.477953e+01 -1.488137e+04 -3.342536e+04  4.361293e+03
##    GarageArea
##  3.559750e+01
```

```
best_subset_predictions <- predict(best_subset_model, newdata = test_data)

best_subset_rmse <- sqrt(mean((test_data$SalePrice - best_subset_predictions)^2))
best_subset_r_squared <- cor(test_data$SalePrice, best_subset_predictions)^2

cat("\nBest Subset Model Performance:\n")
```

```
##
## Best Subset Model Performance:
```

```
cat("RMSE:", best_subset_rmse, "\n")
```

```
## RMSE: 48727.68
```

```
cat("R^2:", best_subset_r_squared, "\n")
```

```
## R^2: 0.6601152
```

Comparison of the above two models

```
step_bic <- BIC(step_model)
best_subset_bic <- BIC(best_subset_model)

cat("\nModel Comparison (BIC):\n")
```

```
##
## Model Comparison (BIC):
```

```
cat("Stepwise Model BIC:", step_bic, "\n")
```

```
## Stepwise Model BIC: 26016.31
```

```
cat("Best Subset Model BIC:", best_subset_bic, "\n")
```

```
## Best Subset Model BIC: 26016.31
```

```r
cat("\nModel Comparison (Test Data):\n")
```

```
##
## Model Comparison (Test Data):
```

```r
cat("Stepwise Model RMSE:", step_rmse, "\n")
```

```
## Stepwise Model RMSE: 48727.68
```

```r
cat("Best Subset Model RMSE:", best_subset_rmse, "\n")
```

```
## Best Subset Model RMSE: 48727.68
```

```r
cat("Stepwise Model R^2:", step_r_squared, "\n")
```

```
## Stepwise Model R^2: 0.6601152
```

```r
cat("Best Subset Model R^2:", best_subset_r_squared, "\n")
```

```
## Best Subset Model R^2: 0.6601152
```