Ridge Regression

Simge Cinar

2023-11-12

```
df2 <- read.csv("/Users/simgecinar/Desktop/ML_lab1/parkinsons.csv")

# Shuffle the data
set.seed(123)
df2 <- df2[sample(nrow(df2)), ]
```

Question 1,2)

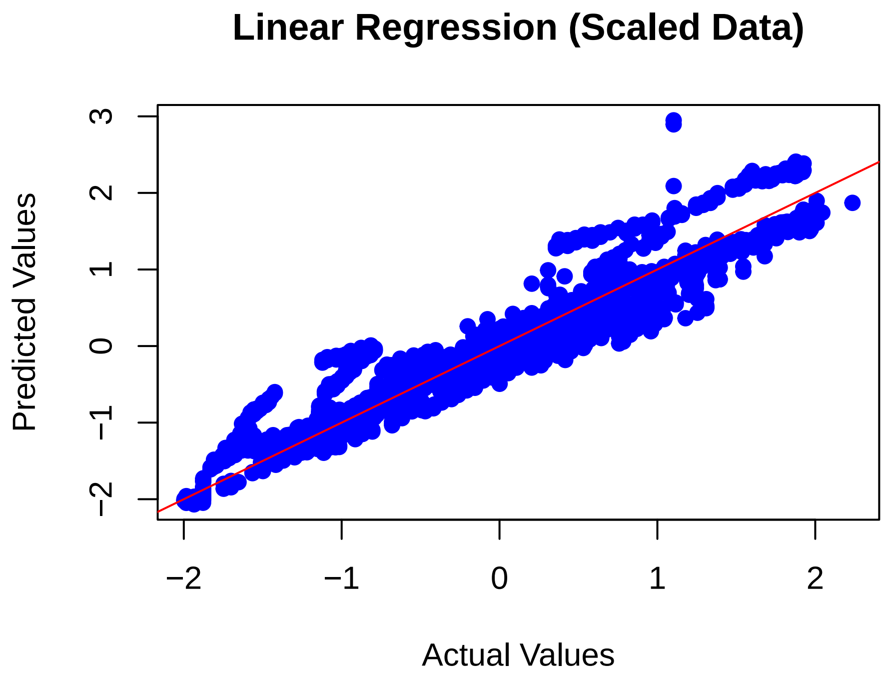
```
set.seed(123)
# Split train and test
train_indices <- createDataPartition(df2$motor_UPDRS, p = 0.6, list = FALSE)
train_data <- df2[train_indices, ]
test_data <- df2[-train_indices, ]

predictor_cols <- setdiff(names(train_data), "motor_UPDRS")
scaler <- preProcess(train_data)
trainS <- predict(scaler, train_data)
testS <- predict(scaler, test_data)

#train_sd <- apply(train_data, 2, sd)

# Linear regression model
lm_model <- lm(motor_UPDRS ~ ., data = trainS)

# Predictions on the test data
trainS_x <- trainS[, predictor_cols]
testS_x <- testS[, predictor_cols]

predS_train <- predict(lm_model, newdata = trainS_x)
predS_test <- predict(lm_model, newdata = testS_x)

mse_train <- mean((trainS$motor_UPDRS - predS_train)^2)
mse_test <- mean((testS$motor_UPDRS - predS_test)^2)

cat("Mean Squared Error (MSE) on the training data:", mse_train, "\n")
```

```
## Mean Squared Error (MSE) on the training data: 0.09478013
```

```
cat("Mean Squared Error (MSE) on the test data:", mse_test, "\n")
```

```
## Mean Squared Error (MSE) on the test data: 0.09364679
```

```r
plot(testS$motor_UPDRS, predS_test, main = "Linear Regression (Scaled Data)",
     xlab = "Actual Values", ylab = "Predicted Values", pch = 19, col = "blue")
abline(a = 0, b = 1, col = "red")
```



Question 3)

```r
# Define the functions
Loglikelihood <- function(theta, std){
  n <- nrow(trainS_x)
  prediction <- as.matrix(trainS_x) %*% as.matrix(theta)
  actual <- trainS$motor_UPDRS
  res <- actual-prediction
  likelihood <- (-(n/2) * log(2*pi*std^2) - (1/(2*std^2)) * sum(res^2))
  return(likelihood)
}

Ridge <- function(theta, std, lambda){
  likelihood_ridge <- -Loglikelihood(theta, std) + (lambda/2)*sum(theta^2)
  return(likelihood_ridge)
}

#optim() function minimizes
RidgeOpt <- function(lambda){
  # Define a new function to optimize
  my_fnc <- function(parameters){
    theta <- parameters[1:(length(parameters)-1)]
    std <- parameters[length(parameters)]
    return(Ridge(theta, std, lambda))
  }
  initial_values <- c(rep(0, ncol(trainS_x)), 1)

  optimal_values <- optim(par = initial_values, fn = my_fnc, method = "BFGS")$par
```

```r
  optimal_theta <- optimal_values[1:length(predictor_cols)]
  optimal_std <- optimal_values[length(predictor_cols) + 1]
  optimal_lambda <- optimal_values[length(optimal_values)]

  result_list <- list(theta = optimal_theta, std = optimal_std, lambda = optimal_lambda)
  return(result_list)
}
```

Formula for the degree of freedom in ridge regression is as follows:

$$df(\lambda) = trace((X^T X + \lambda I)^{-1} X^T X)$$

```r
library(psych)
DF <- function(lambda){
  X <- as.matrix(trainS_x)
  dof <- tr((solve(t(X) %*% X + lambda*diag(ncol(trainS_x)))) %*% t(X) %*% X)
  return(dof)
}
```

Question 4)

```r
lambda <- 1
result_ridge_1 <- RidgeOpt(lambda)

optimal_theta_1 <- result_ridge_1$theta
optimal_std_1 <- result_ridge_1$std
optimal_lambda_1 <- result_ridge_1$lambda

ridge_pred_train_1 <- as.matrix(trainS_x) %*% as.matrix(optimal_theta_1)
ridge_pred_test_1 <- as.matrix(testS_x) %*% as.matrix(optimal_theta_1)

mse_ridge_train_1 <- mean((trainS$motor_UPDRS - ridge_pred_train_1)^2)
mse_ridge_test_1 <- mean((testS$motor_UPDRS - ridge_pred_test_1)^2)

cat("Lambda:", lambda, "\n")
```

```
## Lambda: 1
```

```r
cat("Mean Squared Error (MSE) ridge regression on training data:", mse_ridge_train_1, "\n")
```

```
## Mean Squared Error (MSE) ridge regression on training data: 0.09481838
```

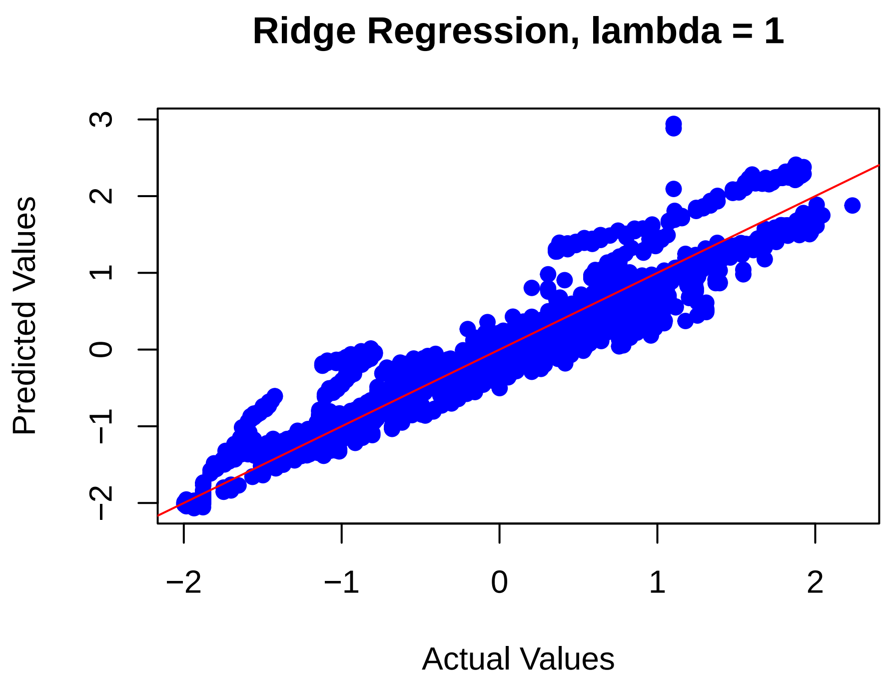
```r
cat("Mean Squared Error (MSE) ridge regression on test data:", mse_ridge_test_1, "\n")
```

```
## Mean Squared Error (MSE) ridge regression on test data: 0.09360224
```

```r
cat("Degree of freedom:", DF(lambda), "\n")
```

```
## Degree of freedom: 18.84537
```

```r
plot(testS$motor_UPDRS, ridge_pred_test_1, main = "Ridge Regression, lambda = 1",
     xlab = "Actual Values", ylab = "Predicted Values", pch = 19, col = "blue")
abline(a = 0, b = 1, col = "red")
```




```r
lambda <- 100
result_ridge_100 <- RidgeOpt(lambda)

optimal_theta_100 <- result_ridge_100$theta
optimal_std_100 <- result_ridge_100$std
optimal_lambda_100 <- result_ridge_100$lambda

ridge_pred_train_100 <- as.matrix(trainS_x) %*% as.matrix(optimal_theta_100)
ridge_pred_test_100 <- as.matrix(testS_x) %*% as.matrix(optimal_theta_100)

mse_ridge_train_100 <- mean((trainS$motor_UPDRS - ridge_pred_train_100)^2)
mse_ridge_test_100 <- mean((testS$motor_UPDRS - ridge_pred_test_100)^2)

cat("Lambda:", lambda, "\n")
```

```
## Lambda: 100
```

```r
cat("Mean Squared Error (MSE) ridge regression on training data:", mse_ridge_train_100, "\n")
```

```
## Mean Squared Error (MSE) ridge regression on training data: 0.09487141
```

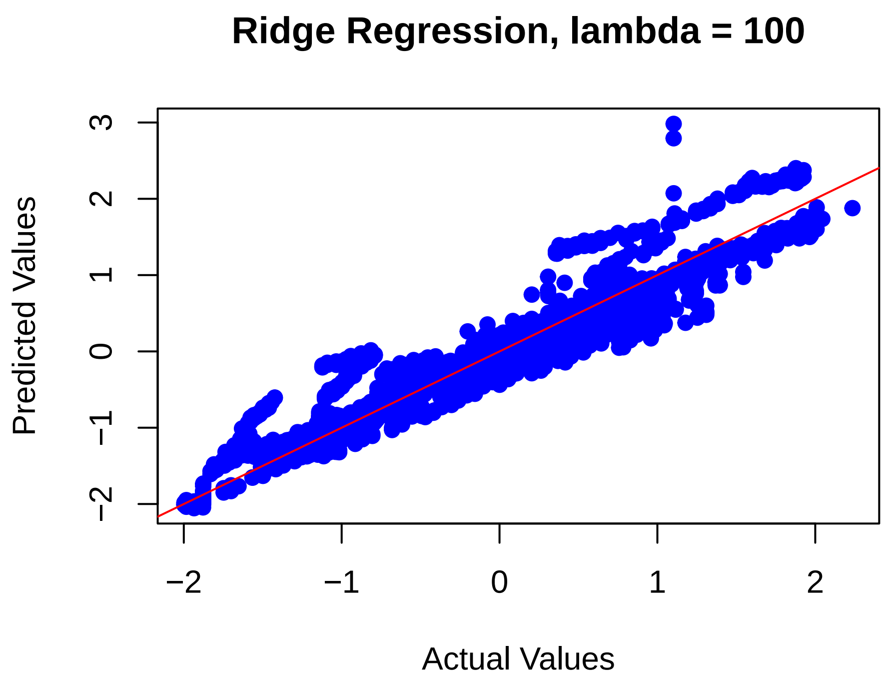
```r
cat("Mean Squared Error (MSE) ridge regression on test data:", mse_ridge_test_100, "\n")
```

```
## Mean Squared Error (MSE) ridge regression on test data: 0.09361211
```

```r
cat("Degree of freedom:", DF(lambda), "\n")
```

```
## Degree of freedom: 14.61243
```

```r
plot(testS$motor_UPDRS, ridge_pred_test_100, main = "Ridge Regression, lambda = 100",
     xlab = "Actual Values", ylab = "Predicted Values", pch = 19, col = "blue")
abline(a = 0, b = 1, col = "red")
```




```r
lambda <- 1000
result_ridge_1000 <- RidgeOpt(lambda)

optimal_theta_1000 <- result_ridge_1000$theta
optimal_std_1000 <- result_ridge_1000$std
optimal_lambda_1000 <- result_ridge_1000$lambda

ridge_pred_train_1000 <- as.matrix(trainS_x) %*% as.matrix(optimal_theta_1000)
ridge_pred_test_1000 <- as.matrix(testS_x) %*% as.matrix(optimal_theta_1000)

mse_ridge_train_1000 <- mean((trainS$motor_UPDRS - ridge_pred_train_1000)^2)
mse_ridge_test_1000 <- mean((testS$motor_UPDRS - ridge_pred_test_1000)^2)

cat("Lambda:", lambda, "\n")
```

```
## Lambda: 1000
```

```r
cat("Mean Squared Error (MSE) ridge regression on training data:", mse_ridge_train_1000, "\n")
```

```
## Mean Squared Error (MSE) ridge regression on training data: 0.09621063
```

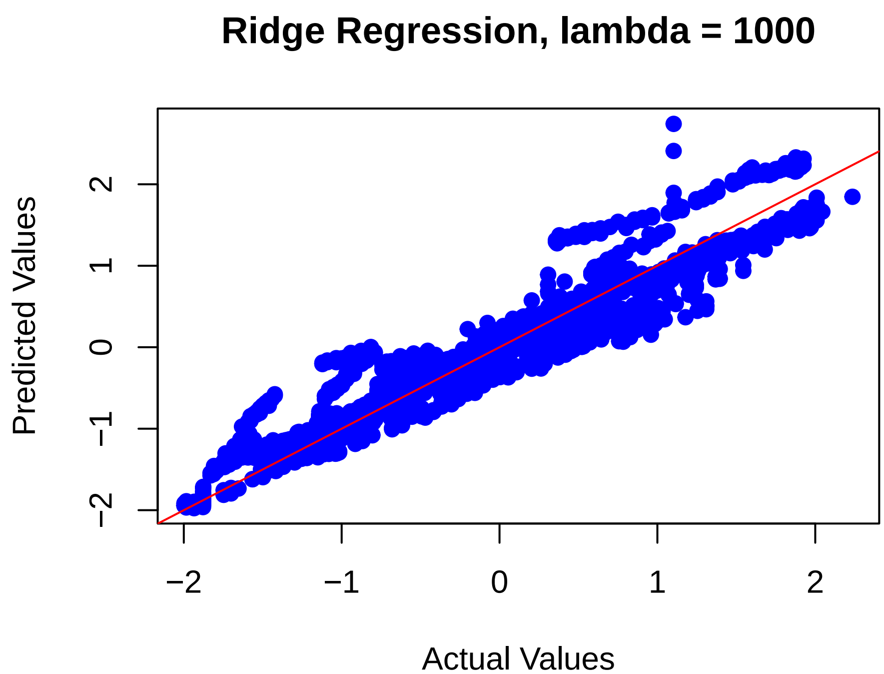
```r
cat("Mean Squared Error (MSE) ridge regression on test data:", mse_ridge_test_1000, "\n")
```

```
## Mean Squared Error (MSE) ridge regression on test data: 0.09423032
```

```r
cat("Degree of freedom:", DF(lambda), "\n")
```

```
## Degree of freedom: 9.222909
```

```r
plot(testS$motor_UPDRS, ridge_pred_test_1000, main = "Ridge Regression, lambda = 1000",
     xlab = "Actual Values", ylab = "Predicted Values", pch = 19, col = "blue")
abline(a = 0, b = 1, col = "red")
```



As lambda increases, the degrees of freedom decreases and the MSE increases