Data Analytics Laboratory

(CSPC-322)

B.Tech VIth Semester (January – June 2024)

Submitted by

Kirtan Gohil (21103073) Group-G3

Submitted to

Ms. Sukhwinder Kaur

Department of Computer Science & Engineering
Dr. B.R. Ambedkar National Institute of Technology
Jalandhar -144008, Punjab, India

Lab-09

Implementation of Lasso, Ridge, and Logistic types of Regressions in R

Aim: Implementation of Lasso, Ridge, and Logistic type of regressions in R.

Theory:

1. **Lasso Regression:** Lasso regression (Least Absolute Shrinkage and Selection Operator) is another regularization technique used for variable selection and regularization in linear regression models. It is particularly useful when dealing with high-dimensional datasets with many predictors, as it tends to shrink coefficients to zero, effectively performing variable selection. Lasso regression adds a penalty term to the least squares objective function, which is the sum of the absolute values of the coefficients multiplied by a regularization parameter (lambda or alpha). Performs variable selection by shrinking some coefficients to exactly zero, handles multicollinearity, and reduces model complexity. May not perform well when there are highly correlated predictors, and tends to select only one variable from a group of correlated variables.
2. **Ridge Regression:** Ridge regression is a regularization technique used to prevent overfitting in linear regression models by penalizing large coefficients. It is commonly applied in situations where multicollinearity exists among predictor variables, as it can shrink the coefficients towards zero without excluding any variables. Ridge regression adds a penalty term to the least squares objective function, which is the sum of squared coefficients multiplied by a regularization parameter (lambda or alpha). Helps to reduce overfitting, stabilizes coefficient estimates, and performs well when there are many correlated predictors. Does not perform variable selection, only shrinks coefficients toward zero without eliminating them.
3. **Logistic Regression:** Logistic regression is a statistical method used for predicting the probability of a binary outcome based on one or more predictor variables. It is widely used in various fields such as healthcare (predicting disease risk), marketing (customer churn prediction), and finance (credit risk assessment). Logistic regression models the relationship between the dependent variable (binary outcome) and independent variables using the logistic function, which ensures that predicted probabilities lie between 0 and 1. Simple and interpretable model provides probabilities of class membership, and handles non-linear relationships between predictors and the outcome. Assumes a linear relationship between predictors and the log odds of the outcome, sensitive to outliers.

Procedure:

Installing packages and attaching library:

```
> #install.packages("glmnet")
> install.packages("glmnet")

package 'iterators' successfully unpacked and MD5 sums checked
package 'foreach' successfully unpacked and MD5 sums checked
package 'shape' successfully unpacked and MD5 sums checked
package 'glmnet' successfully unpacked and MD5 sums checked


> library(glmnet)
```

Importing Dataset:

```
> data <- read.csv("C:\\Users\\Akshita\\Desktop\\income.data.csv", header = TRUE)
```

Preparing data and target variable:

```
> X <- as.matrix(data[, -1])
> y <- data[, 1]
> set.seed(123)
```

Splitting data into training and testing sets:

```
> train_idx <- sample(1:nrow(data), 0.7 * nrow(data))
> X_train <- X[train_idx, ]
> y_train <- y[train_idx]
> X_test <- X[-train_idx, ]
> y_test <- y[-train_idx]
```

Lasso Regression:

```
> #Lasso regression
> lasso_model <- glmnet(X_train, y_train, alpha = 1)
> lasso_predictions <- predict(lasso_model, s = 0.01, newx = X_test)
>
```

Ridge Regression:

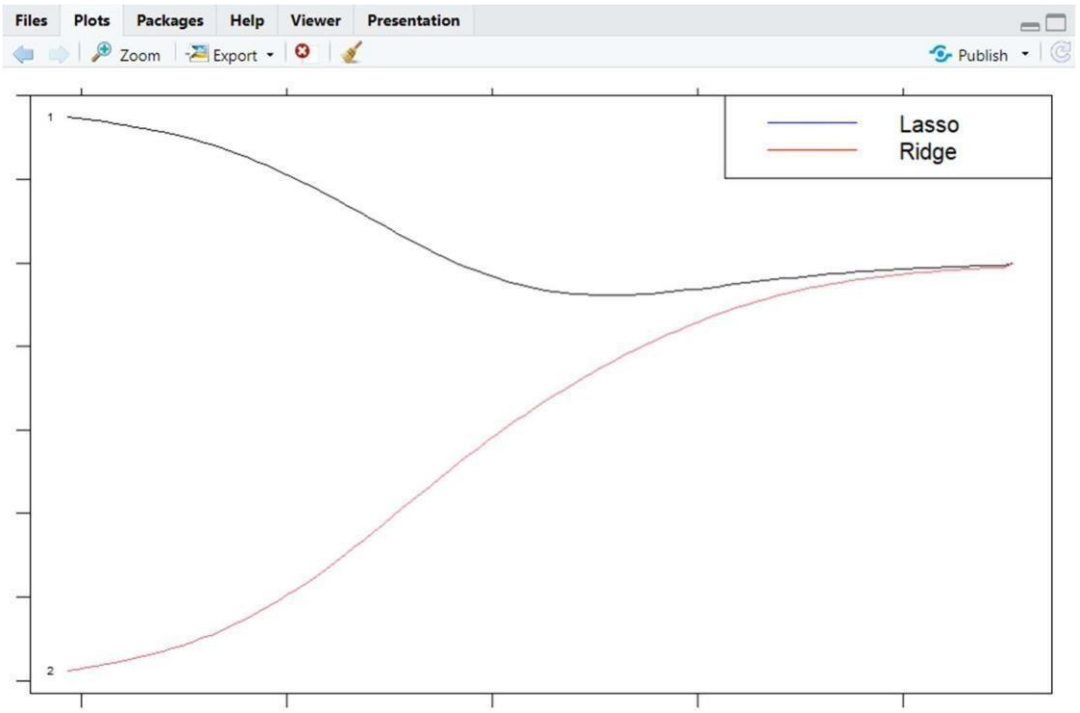
```
> #Ridge regression
> ridge_model <- glmnet(X_train, y_train, alpha = 0)
> ridge_predictions <- predict(ridge_model, s = 0.01, newx = X_test)
>
```

Root Mean Square Errors for Lasso and Ridge:

```
> lasso_rmse <- sqrt(mean((lasso_predictions - y_test)^2))
> ridge_rmse <- sqrt(mean((ridge_predictions - y_test)^2))
>
> print(paste("Lasso RMSE:", lasso_rmse))
[1] "Lasso RMSE: 156.055452402162"
> print(paste("Ridge RMSE:", ridge_rmse))
[1] "Ridge RMSE: 156.020156706722"
```

Plotting models for Lasso and Ridge:

```
> plot(lasso_model, xvar = "lambda", label = TRUE)
> plot(ridge_model, xvar = "lambda", label = TRUE)
>
> legend("topright", legend = c("Lasso", "Ridge"), col = c("blue", "red"), lty = 1)
>
```



Logistic Regression:

Installing packages and attaching library:

```
> #logistic regression
> install.packages("car")
```

```
package 'car' successfully unpacked and MD5 sums checked
```

```
> library(carData)
> library(car)

> library(ggplot2)
```

Importing Dataset:

```
> data <- read.csv("C:\\Users\\Akshita\\Desktop\\heart.data.csv", header = TRUE)
```

Preparing target variable:

```
> data$heart.disease.binary <- ifelse(data$heart.disease > 0.5, 1, 0)
> table(data$heart.disease.binary)

  1
498
```

Logistic Regression Model:

```
> model <- glm(heart.disease.binary ~ biking + smoking, data = data, family = binomial, maxit = 1000)
```

Checking for multicollinearity:

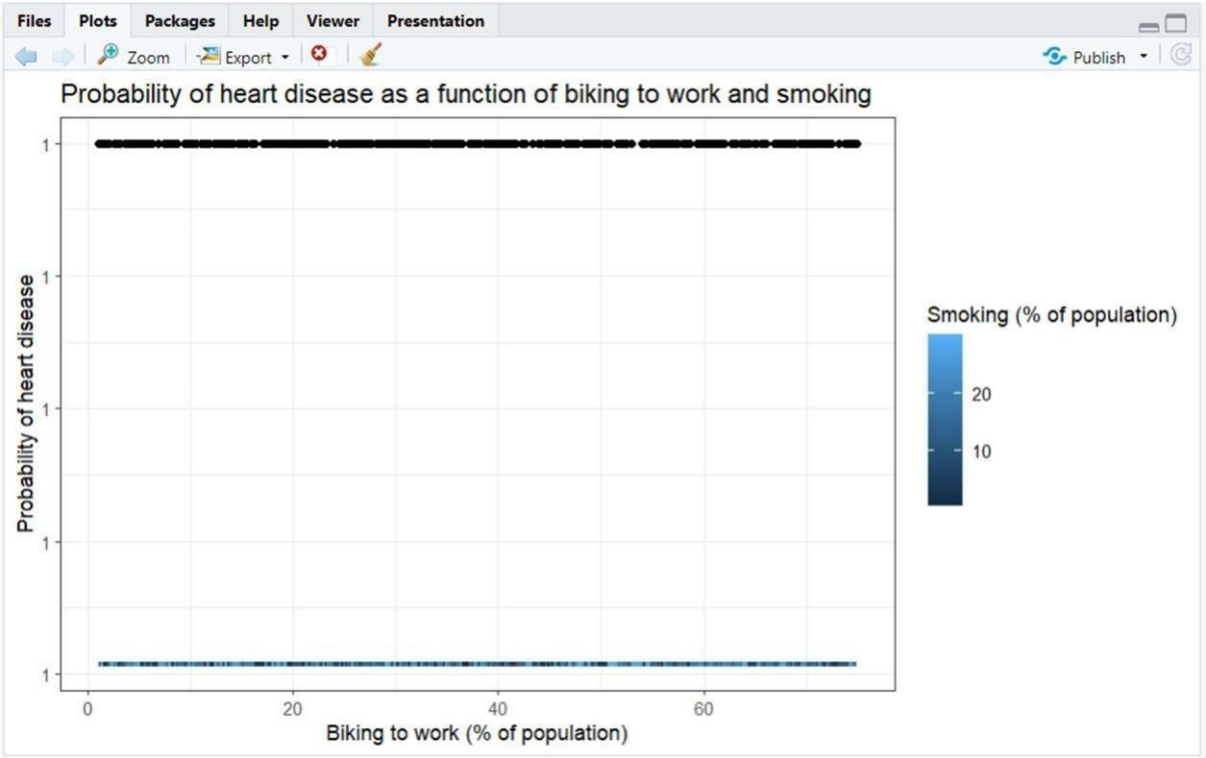
```
> vif_values <- vif(model)
> print(vif_values)
  biking  smoking
1.000229 1.000229
```

Summary:

```
> summary(model)

Call:
glm(formula = heart.disease.binary ~ biking + smoking, family = binomial,
    data = data, maxit = 1000)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.857e+01  1.186e+05       0        1
biking      -1.388e-08  2.022e+03       0        1
smoking     -2.008e-08  5.239e+03       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.00e+00  on 497  degrees of freedom
Residual deviance: 3.91e-10  on 495  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 27
```

Plotting logistic model line:

```
> data.plot.logistic <- ggplot(data, aes(x = biking, y = heart.disease.binary)) +
+     geom_point() +
+     geom_line(aes(y = prob, color = smoking), linewidth = 1.25) +
+     theme_bw() +
+     labs(title = "Probability of heart disease as a function of biking to work and smoking",
+          x = "Biking to work (% of population)",
+          y = "Probability of heart disease",
+          color = "Smoking (% of population)")
> data.plot.logistic
```



Result: Through this experiment, we have successfully implemented Lasso, Ridge, and Logistics types of regressions models in R.