BITS PILANI, K.K. BIRLA GOA CAMPUS

SEMESTER I- 2024-25

Assignment-I

Course No.: ECON F215

Course Title: Computational Economics

Course Instructor In-Charge: Dr. Sandip Sarkar

Submission deadline: 15/02/2025

Date of Presentation: 19/02/2025

Suvid Singhal

2023B3A70972P

Problem 1

Problem Statement: Develop a story based on the given dataset and run a linear regression. Interpret your findings carefully.

How Do Land Owned, Expenses on education, medical, clothing, bedding etc. Influence Household Expenditure?

In a diverse and rapidly developing country like India, household consumption patterns are influenced by a variety of socioeconomic factors. The aim of this study is to find how different factors affect the monthly household expenditure of individuals. Household expenditure is a crucial measure of spending power of the people in an economy. Estimating it is very useful for governments in formulating policies for its citizens.

All the dependent and independent variables chosen are continuous in nature.

The dependent variable chosen from the dataset is Total Household Expenditure which is taken as Y in our linear regression model.

The independent variables chosen from the dataset are: Land held (in acres), Consumption of clothing, bedding etc. during last 30 days value (Rs.) and Expenditure on education and medical goods and services during last 30 days representing X_1, X_2 and X_3 respectively.

Reasons for choosing the independent variables:

1) **Land held**: Land held by the family may influence the spending and consumption decisions. The families with more land may spend more even if their income is slightly less.
2) **Consumption of clothing, bedding etc. during last 30 days**: This is a part of their monthly spending so it should directly affect their spending.
3) **Expenditure on education and medical goods and services**: This is a part of their occasional spending so it should also affect their spending.

The source code for this problem can be found here:
https://pastebin.com/DfubrMc7

Also had to merge different dataframes based on the HHID key for the analysis.

Linear regression is being used for this problem.

The results of the regression are as follows:

```
> # Display Results
> summary(model)

Call:
lm(formula = MPCE_MRP ~ Land_owned + Last_30days_Value + Expenditure_in_Rs_last_30_days,
    data = final_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2166256   -98457   -48943    35966  9083619

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.857e+05  3.651e+02  508.55   <2e-16 ***
Land_owned                   3.106e+00  1.255e-01   24.75   <2e-16 ***
Last_30days_Value            8.383e+01  3.991e-01  210.04   <2e-16 ***
Expenditure_in_Rs_last_30_days 3.910e+00  7.725e-02   50.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 240000 on 647608 degrees of freedom
Multiple R-squared:  0.07095,   Adjusted R-squared:  0.07095
F-statistic: 1.649e+04 on 3 and 647608 DF,  p-value: < 2.2e-16
```

The adjusted R^2 value of approximately 0.07 indicates that the model explains only 7% of the variability in MPCE_MRP. This suggests that the selected predictors (land ownership, last 30 days' expenditure, and education & medical expenditure) have limited explanatory power for household consumption. Other unaccounted factors may play a significant role in determining MPCE_MRP.

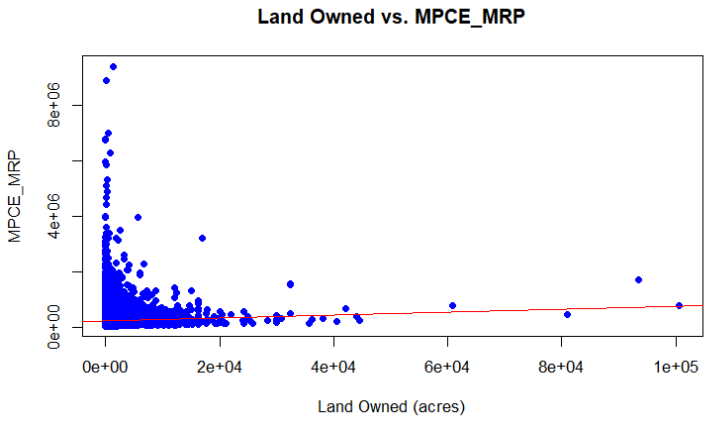
So, to improve it further, natural log was taken on each of the terms and the linear regression was run again.

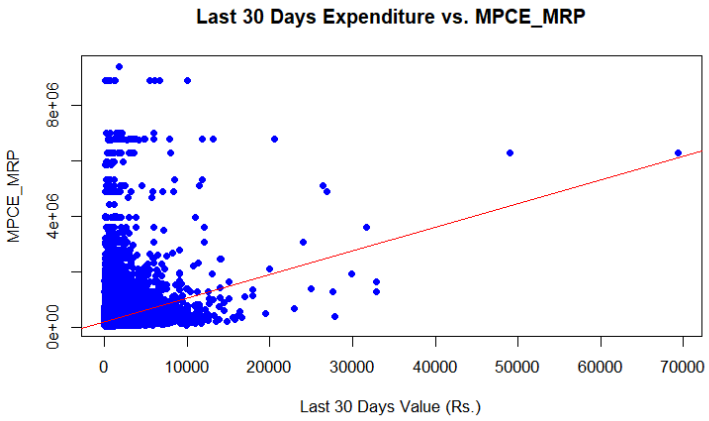
The results were as follows:

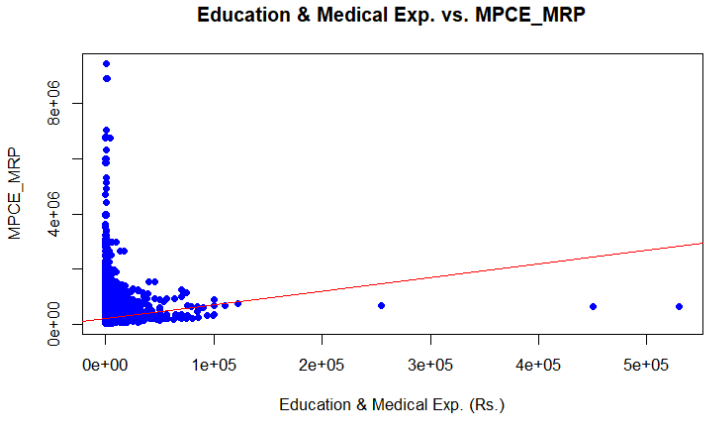
```
> summary(model)

Call:
lm(formula = log(MPCE_MRP) ~ log(Land_owned + 1) + log(Last_30days_Value +
    1) + log(Expenditure_in_Rs_last_30_days + 1), data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2308 -0.3850 -0.0494  0.3296  4.0192

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             10.6781086  0.0044890 2378.71   <2e-16 ***
log(Land_owned + 1)                     -0.0117386  0.0003035  -38.68   <2e-16 ***
log(Last_30days_Value + 1)               0.1686054  0.0006866  245.58   <2e-16 ***
log(Expenditure_in_Rs_last_30_days + 1)  0.1075963  0.0004971  216.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5602 on 647608 degrees of freedom
Multiple R-squared:  0.1683,    Adjusted R-squared:  0.1683
F-statistic: 4.368e+04 on 3 and 647608 DF,  p-value: < 2.2e-16
```

As we can see, the Adjusted R^2 improved from 0.07 to 0.168 which is a significant improvement.

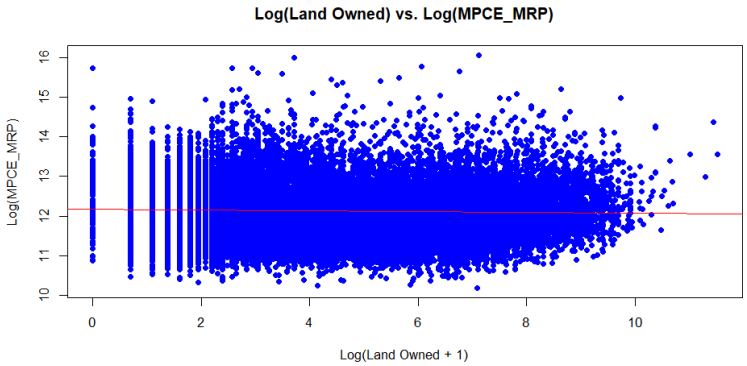
The regression plots for the original model are as follows:

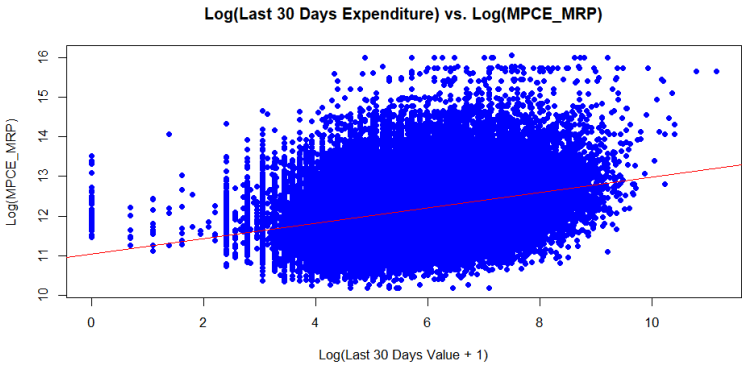


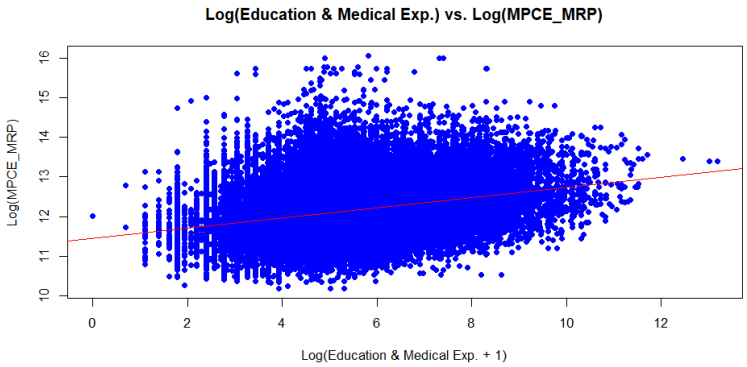




The regression plots for new model are as follows:







Problem 2

Problem Statement: Find a binary variable and run a logistic regression. The model also has to be justifiable. Predict the probability of success for all observations and compute the mismatch with actual dependent variable. You may use a threshold of 0.5 for this problem. Repeat the problem, leaving 10% of the sample as testing data and illustrating the error rates using the testing data.

How Do Household Size, Household Type and Occupation Influence Land Ownership?

The aim of this study is to find how different factors affect the land ownership in different households. Land ownership is a crucial measure wealth and prosperity of the people in an economy. Here, we are predicting if a family owns land based on its size, type and principal occupation.

We use binary logistic regression for this problem.

Our dependent variable is binary in nature: Yes or No. The independent variables are a mix of both categorical and discrete variables. Household size is a discrete variable whereas Household type and Principal Occupation is a categorical variable.

The variables are:

1) **whether_Land_owned**: This is the dependent variable. Its value can be Yes or No.
2) **HH_Size**: Its value is a natural number.
3) **HH_Type**: It is a categorical variable. Its classification is as follows:

   household type *: for rural areas: self-employed in: agriculture -1, non-agriculture - 2; regular wage/salary earning - 3, casual labour in: agriculture - 4, non-agriculture -5; others-9*
   for urban areas: self-employed-1, regular wage/salary earning-2, casual labour-3, others-9

   Source: Survey form used for collecting the data
4) **NCO_2004**: It is also a categorical variable whose detailed list of values can be found here - https://microdata.gov.in/nada43/index.php/catalog/110/download/1129

The source code for this problem can be found here: https://pastebin.com/zj12zjrE

Logistic regression was done for part 1 of this problem. The results were as follows:

```
NCO_200492x  11.796614 289.750222   0.041 0.967525
NCO_2004931  -0.442411   0.123849  -3.572 0.000354 ***
NCO_2004932  -0.865373   0.143621  -6.025 1.69e-09 ***
NCO_2004933  -0.471155   0.132060  -3.568 0.000360 ***
NCO_200493x   9.895735 374.897827   0.026 0.978942
NCO_2004X00  -0.255750   1.182605  -0.216 0.828785
NCO_2004X10   0.749511   0.384366   1.950 0.051177 .
NCO_2004X99   0.118531   0.370329   0.320 0.748916
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80852  on 101661  degrees of freedom
Residual deviance: 68199  on 101522  degrees of freedom
AIC: 68479

Number of Fisher Scoring iterations: 12
```

Other relevant metrics of the model including the confusion matrix are as follows:

```
> # Compute Confusion Matrix
> conf_matrix <- table(Actual = data$land_ownership, Predicted = data$predicted_class)
> conf_matrix
      Predicted
Actual     0     1
     0   869 12958
     1   436 87399
>
> # Calculate accuracy
> accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
> mismatch_rate <- 1 - accuracy
>
> # Print results
> print(paste("Model Accuracy:", round(accuracy, 3)))
[1] "Model Accuracy: 0.868"
> print(paste("Error Rate:", round(mismatch_rate, 3)))
[1] "Error Rate: 0.132"
>
```

For the part 2 of the problem, logistic regression was performed again, this time on 90% data as 10% data was reserved for testing purposes.

The results of logistic regression are as follows:

```
NCO_200492x  11.831455 289.733234   0.041 0.967427
NCO_2004931  -0.402196   0.132247  -3.041 0.002356 **
NCO_2004932  -0.824740   0.153080  -5.388 7.14e-08 ***
NCO_2004933  -0.413042   0.140825  -2.933 0.003357 **
NCO_200493x   9.925442 374.928478   0.026 0.978880
NCO_2004X00  -0.266423   1.181485  -0.225 0.821591
NCO_2004X10   0.878732   0.409861   2.144 0.032035 *
NCO_2004X99   0.160384   0.412298   0.389 0.697276
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 72766  on 91495  degrees of freedom
Residual deviance: 61418  on 91356  degrees of freedom
AIC: 61698

Number of Fisher Scoring iterations: 12
```

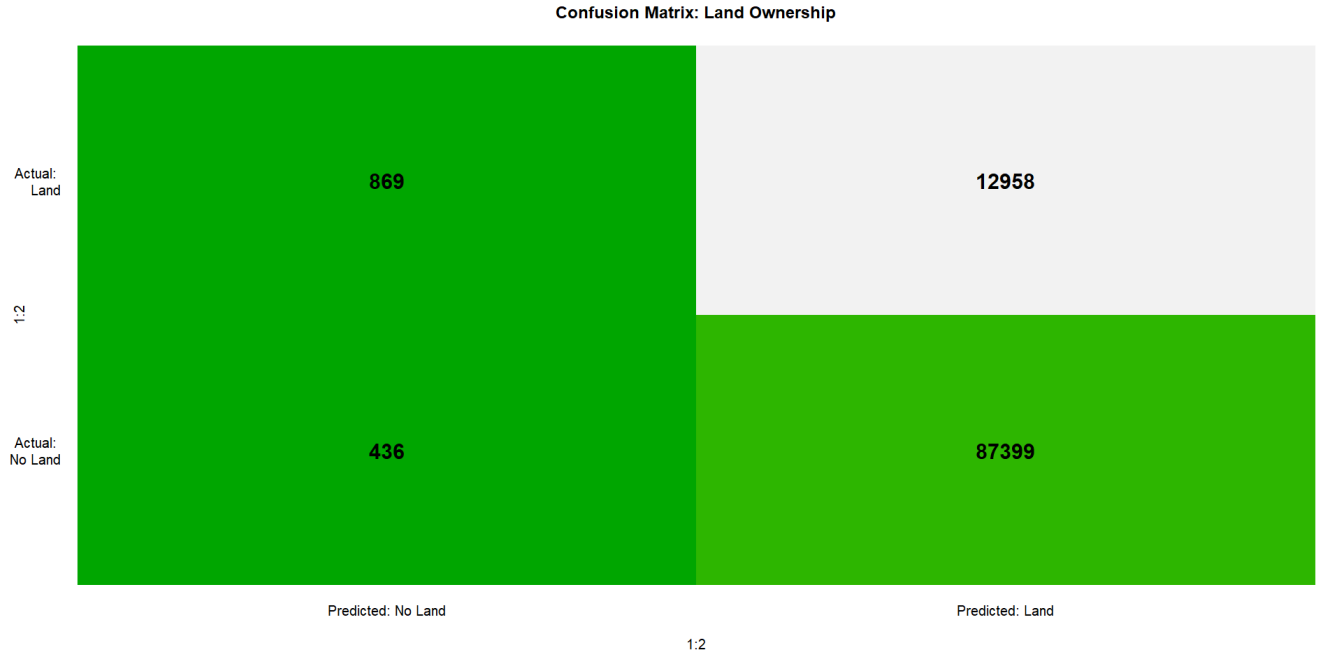
Other relevant metrics of the model including the confusion matrix are as follows:

```
> # Compute mismatch (Error Rate)
> conf_matrix_test <- table(Actual = test_data$land_ownership, Predicted = test_data$predicted_c
s)
> conf_matrix_test
      Predicted
Actual    0    1
     0   77 1306
     1   46 8737
>
> # Calculate accuracy
> accuracy_test <- sum(diag(conf_matrix_test)) / sum(conf_matrix_test)
> error_rate_test <- 1 - accuracy_test
>
> # Print results
> print(paste("Test Model Accuracy:", round(accuracy_test, 3)))
[1] "Test Model Accuracy: 0.867"
> print(paste("Test Error Rate:", round(error_rate_test, 3)))
[1] "Test Error Rate: 0.133"
```

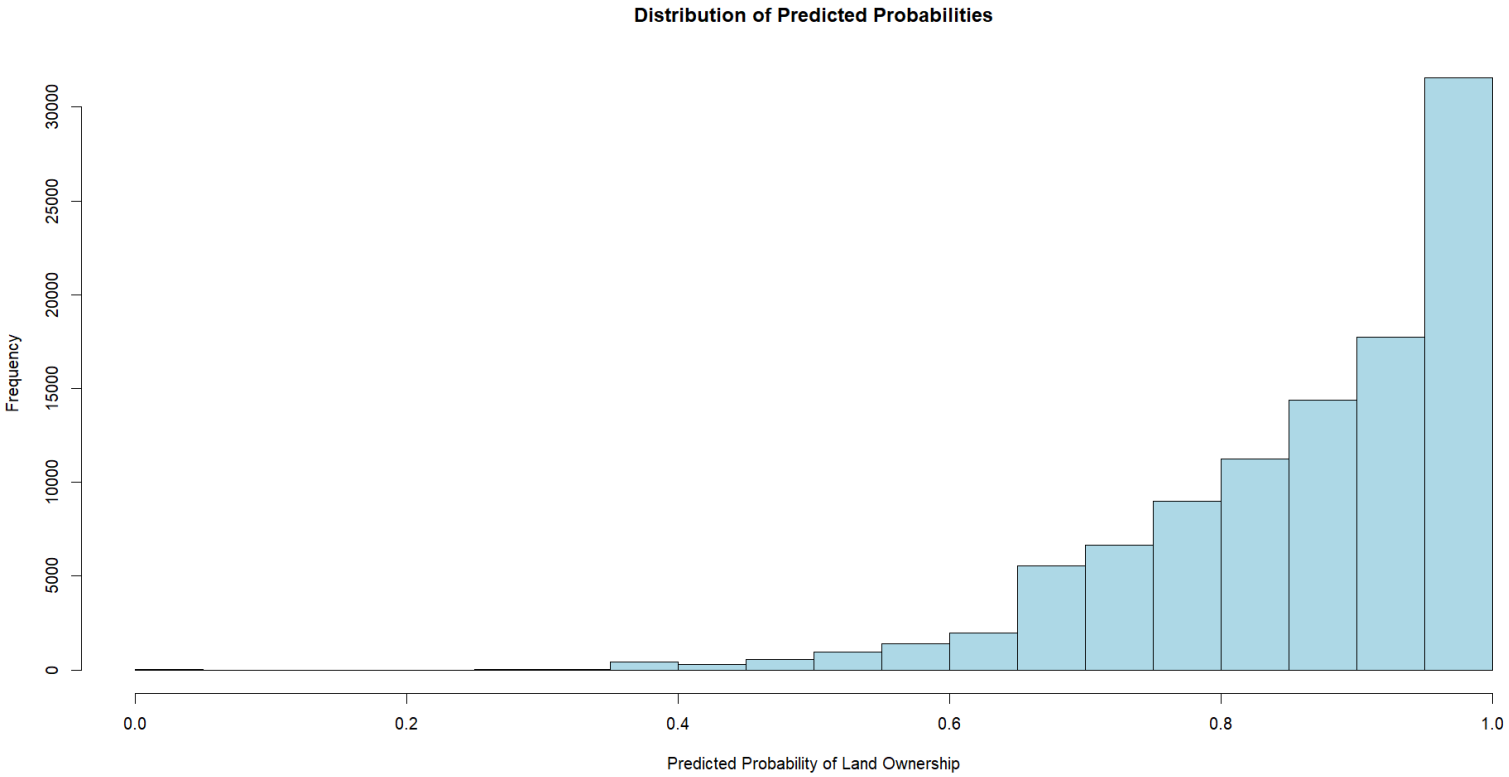
The first model obviously outperforms the second one by a little bit because the more data was provided as training data as compared to the second model.

The graphs and data visualizations obtained from the second model are as follows:

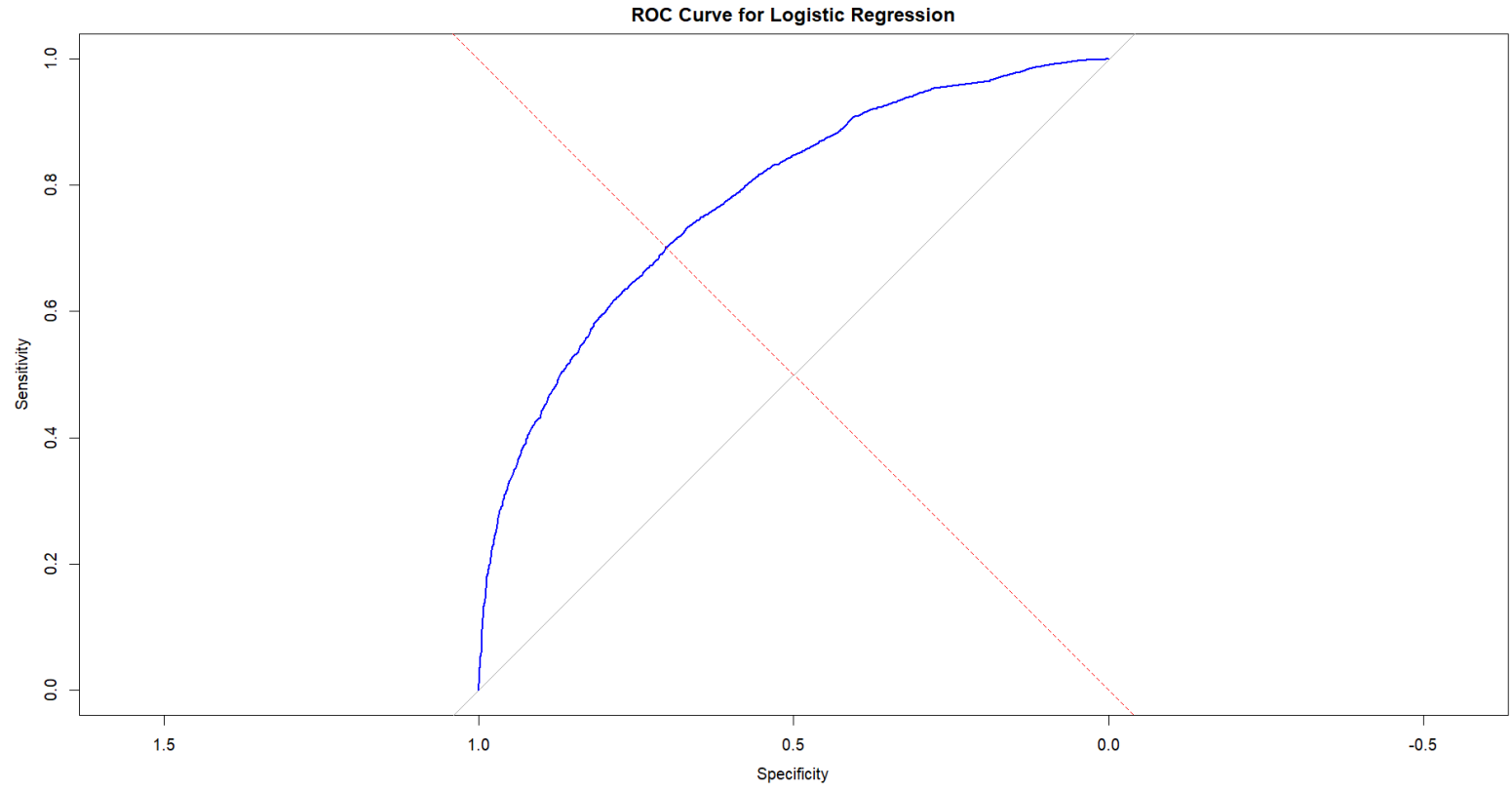
The confusion matrix



The probability distribution of predicted probabilities



ROC Curve for the model



Problem 3

Problem Statement: Now find a nominal variable with K categories (K>2). Classify all the observations in K categories, using Linear and Quadratic Discriminant analysis. Determine the false positive rate and false negative rates using different thresholds in [0.5,1]. Find the error rates in training and testing datasets (keep 10% of the sample as testing data).

How Do Lighting Fuel choice and Expenditure Influence the Cooking Fuel choice?

The aim of this study is to find how different factors affect the cooking fuel choice in different households. Cooking fuel being used is a crucial measure for the government as based on this data, government can run targeted schemes and provide subsidies to people in order to ensure sustainable development. Here, we are predicting the cooking fuel used in a household based on its lighting fuel choice and monthly expenditure.

We use Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) for this problem.

Our dependent variable is a categorical variable and is split into 3 categories. The independent variables are a mix of both categorical and continuous variables. Lighting Fuel type is a categorical variable whereas Monthly Expenditure is a continuous variable.

The variables are:

1) **Cooking_Code**: This is the dependent variable. It is a categorical variable. Its classification is as follows:

   primary source of energy for cooking *: coke, coal-01, firewood and chips-02, LPG-03, gobar gas-04, dung cake-05, charcoal-06, kerosene-07, electricity-08, others-09, no cooking arrangement-10*

   Source: Survey form used for collecting the data
2) **MPCE_MRP**: It's a continuous variable. It is the monthly per capital expenditure in maximum retail prices.
3) **Lighting_Code**: It is a categorical variable. Its classification is as follows:

   primary source of energy for lighting *: kerosene-1, other oil -2, gas-3, candle-4, electricity-5, others-9, no lighting arrangement-6*

   Source: Survey form used for collecting the data

The source code for this problem can be found here:
https://pastebin.com/Tgz4kCDt

Linear Discriminant Analysis and Quadratic Discriminant Analysis was performed for this problem.

The results of which are as follows:

```
> print(results)
  Threshold   LDA_FPR   LDA_FNR LDA_Error      QDA_FPR   QDA_FNR  QDA_Error
1       0.5 0.4772009 0.9919338 0.4912159 0.0000986972 1.0000000 0.01796289
2       0.6 0.7818792 0.9869982 0.7895776 0.5904066325 0.9919759 0.60353336
3       0.7 0.9627912 0.9852383 0.9664430 0.6854520332 0.9906407 0.69690091
4       0.8 0.9980261 0.9822983 0.9982235 0.8558033952 0.9900819 0.86517963
5       0.9 0.9985195 0.9821093 0.9985195 0.9087050928 0.9872923 0.91502171
6       1.0 1.0000000 0.0000000 1.0000000 1.0000000000 0.0000000 1.00000000
```

The testing error rates for both LDA and QDA are as follows:

```
> # Compute Testing Error Rate
> lda_test_preds <- predict(lda_model, test_data)$class
> qda_test_preds <- predict(qda_model, test_data)$class
>
> lda_test_error <- mean(lda_test_preds != test_data$Cooking_Code, na.rm = TRUE)
> qda_test_error <- mean(qda_test_preds != test_data$Cooking_Code, na.rm = TRUE)
>
> print(paste("LDA Testing Error:", round(lda_test_error, 3)))
[1] "LDA Testing Error: 0.352"
> print(paste("QDA Testing Error:", round(qda_test_error, 3)))
[1] "QDA Testing Error: 0.49"
```

Now, the thresholds were varied from 0.5 to 1 by taking steps of 0.05.

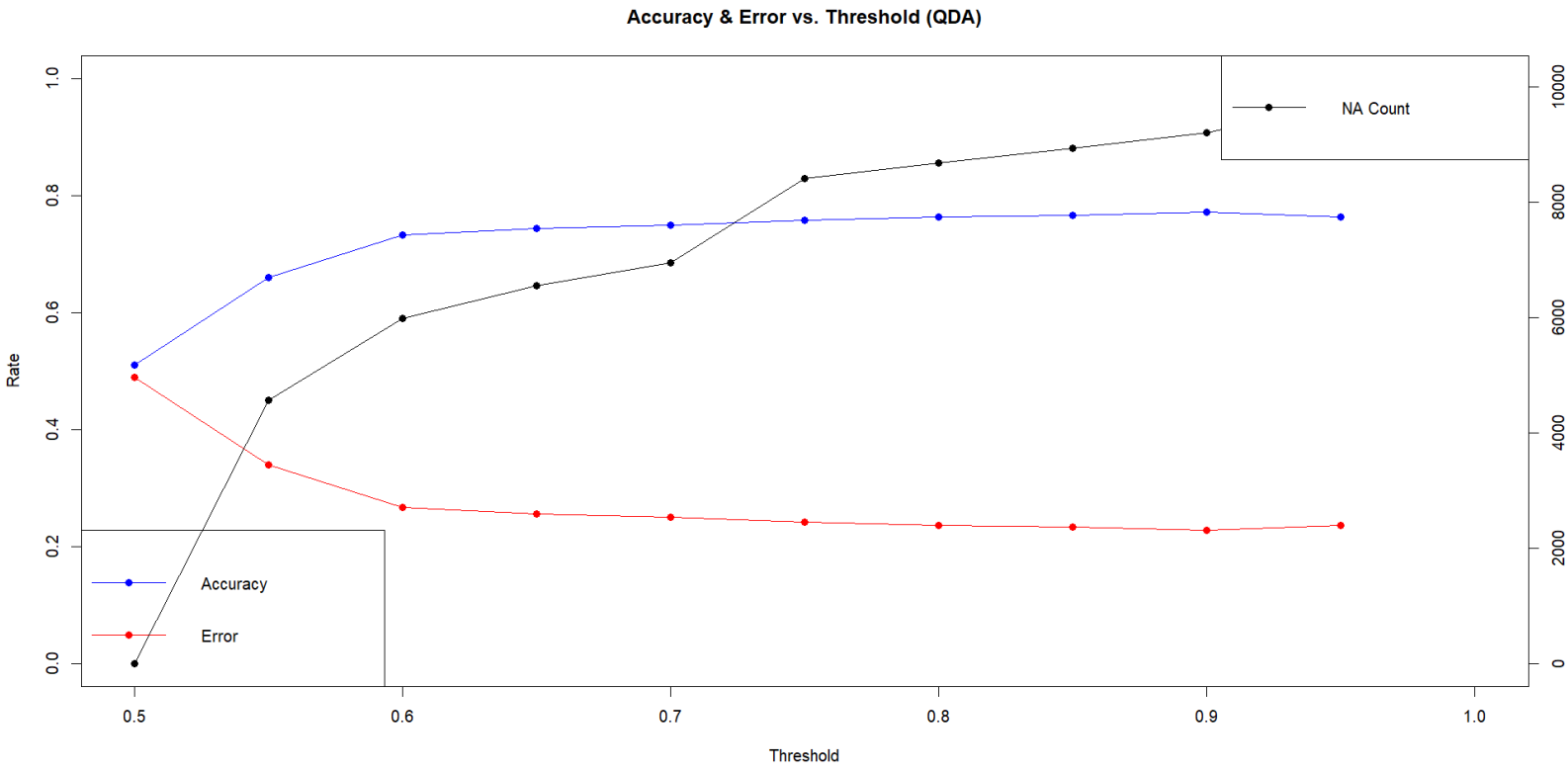
The result of which are as follows:

```
> print("Error Rates for Different Thresholds:")
[1] "Error Rates for Different Thresholds:"
> print(error_results)
   Threshold LDA_Error QDA_Error
1       0.50 0.6175484 0.4898342
2       0.55 0.7862219 0.6370904
3       0.60 0.8345835 0.6995657
4       0.65 0.9486775 0.7365772
5       0.70 0.9715752 0.7641137
6       0.75 0.9906238 0.8707067
7       0.80 0.9998026 0.8897552
8       0.85 0.9999013 0.9088038
9       0.90 1.0000000 0.9294315
10      0.95 1.0000000 0.9670351
11      1.00 1.0000000 1.0000000
```

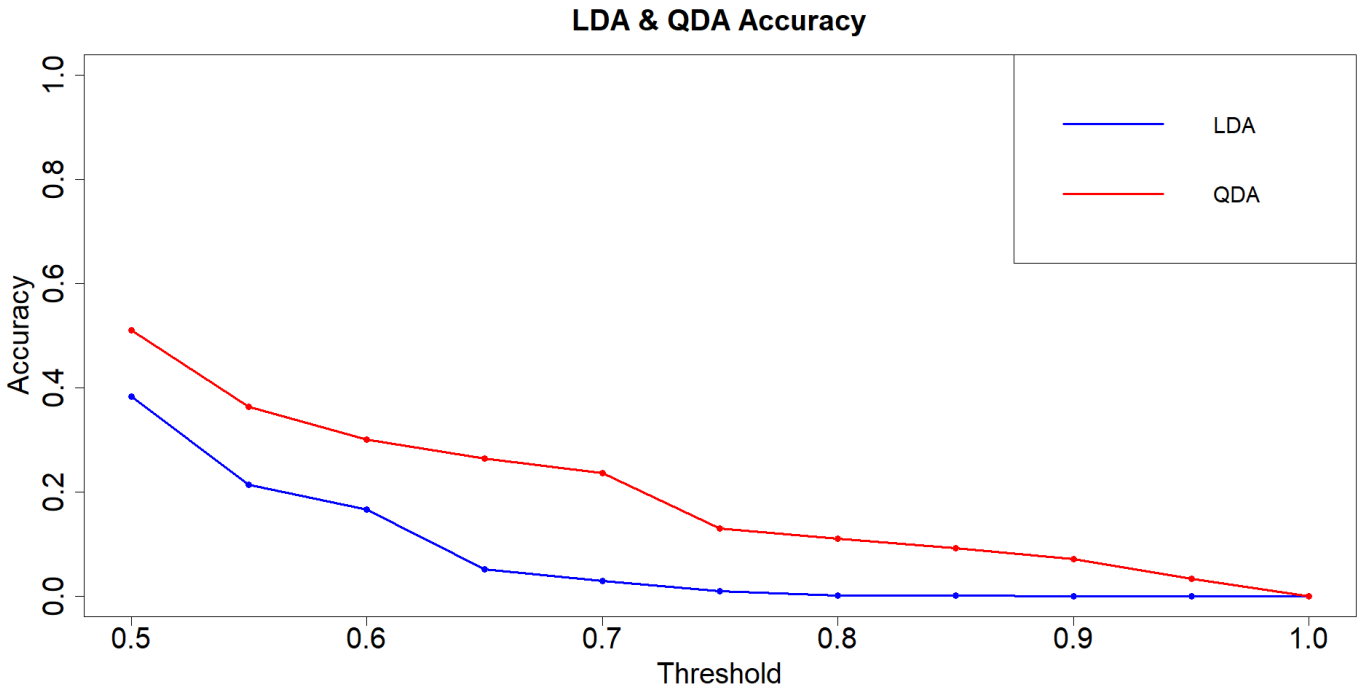
Both LDA and QDA errors seem to increase as the threshold is made stricter on each run which is the expected trend.

The graphs and data visualizations obtained from the model are as follows:

Accuracy and Error v/s Threshold for QDA



LDA & QDA Accuracy v/s Threshold



LDA & QDA Error v/s Threshold