भारतीय प्रबन्धन संस्थान तिरुचिरापल्ली

Indian Institute of Management Tiruchirappalli

Post Graduate Program in Management

Batch: PGPM 2023-25

Subject

Business Data Analytics with Categorical and Censored Outcomes

Submitted To:

Prof. Gulasekaran Rajaguru

Submitted By:

Dia Majumder

Roll no.: 2301021

Abstract:

This study uses a logistic regression model to predict the likelihood of developing chronic heart disease (CHD) within ten years. The analysis is based on a dataset containing medical and demographic information of individuals aged 32 to 70, with variables such as age, gender, education level, smoking habits, blood pressure, and glucose levels. After feature selection, significant factors like age, gender, daily cigarette consumption, systolic blood pressure, and glucose levels were identified and used in the model.

The results conclude that older individuals, men, smokers, and those with elevated blood pressure or glucose levels have a higher risk of CHD. Marginal effects give explanations of each factor’s individual impact. To evaluate the model's reliability for CHD risk prediction, the accuracy, precision, and recall are calculated. These findings may have significant implications for healthcare policy and decision making, companies introducing new products in the market, etc.

Introduction:

Econometrics is a branch of economics that uses statistical and mathematical models to analyse different kinds of data and their relationships. It uses statistical models and tools like regression analysis, hypothesis testing and time series analysis to analyse economic models and therefore approve or reject economic theories.

Limited dependent variables (LDVs) are outcome variables in econometrics whose values are restricted or contained. There are various types of constraints, leading to different kinds of limited dependent variables. Some of them include binary or dichotomous variables (can only take two values), categorical, ordinal (categories with natural ordering), censored (all outcomes are not visible to everyone), etc.

Examples of such data include:

- Binary: whether a customer buys a product (1) or not (0).
- Categorical: the type of transport chosen by a person like bus, train, or car.
- Ordinal: the position of a horse in a race like first, second, third, etc.
- Censored: Income of a person reported as >10 lakhs per annum, and not the exact figure.
- Truncated: for the purpose of a study, only people with education higher than high school level are considered.

Different kinds of limited dependent variables require different models to study them. If appropriate models aren’t selected based on the data type, the prediction accuracy might be low. The proper models chosen will ensure that the relationship between dependent and independent variables is predicted correctly. In statistical analysis, linear or logistic regression is used to predict binary variables, ordered logit or probit models are used for ordinal data, and multinomial models are used for multiple class classifications.

The purpose of this study is to develop a model to predict the chance of chronic heart disease in a person ten years later. The dataset consists of information about people’s age, gender, whether they are a smokers or not and how many cigarettes they smoke per day, whether they have hypertension or stroke, people’s blood pressure and glucose levels, etc. Using this data, the model has to predict whether the person will develop chronic heart disease after 10 years.

A logistic regression model is used to predict the outcome in this case. After running an initial model with the variable TenYearCHD as the dependent variable and other columns as the independent variables, it was found that not all other variables are significant at a 95% confidence interval. The independent variables male, age, cigsPerDay, sysBP, and glucose were found to be significant, and therefore used in the model.

A correlation matrix was calculated with the significant variables, and none of the variables showed strong correlation with each other, which implies that there is no multicollinearity.

This report analyses the logistic regression model used to predict chronic heart disease, and explains how the correct statistical model can be used to study relationships within the data and help healthcare professionals offer specialised care if and when required.

Literature Review:

A large amount of research has been done on modelling limited dependent variables, and some of those were studied to understand the problem at hand better. Ordinary Least Squares (OLS) regression models are a type of model that is used in econometrics to estimate the relationship between an outcome variable and one or more independent variables. Simple and Mulitple Linear regression are types of OLS models. Some assumptions which must hold true while using an OLS model are:

- Linear relationship: The relationship between the dependent and independent variables is linear.
- Observations are independent of each other.
- No multicollinearity: independent variables are not strongly correlated.
- Homoskedasticity: the variance of errors is constant across all independent variables.
- Normal distribution of errors.

These assumptions might not hold true for LDVs, therefore other models are required to predict them.

In case of LDV models, marginal effects must be computed to understand the relationship between independent variables and the dependent variable. Marginal effects give the change that occurs in a dependent variable for a unit change in an independent variable, the other independent variables remaining constant.

Feature selection is the process of identifying and selecting certain features (independent variables) from a large dataset to improve model performance and reduce complexity of the model. If done properly, feature selection provides many advantages while working with a large dataset:

- The selected features can be used to build a more compact model that accurately predicts the data.
- The selected features can reflect the characteristics of the original data.
- The chosen features can help the decision maker pick valuable information from a large volume of noisy data.

The two papers provided by Prof Gulasekaran Rajguru, titled "Do Jockeys "Look to" or "Rest on" Their Laurels After a Sequence of Winning Rides?" and "Do (Australian) Jockeys Have Hot Hands?" were studied in order to understand more about the selection of the independent variables for the selected problem of predicting chronic heart disease.

From all the secondary research it was obvious that the features need to be selected very carefully for the analysis. The predictive model for this problem is a logistic regression model that tries to predict the occurrence of chronic heart disease by considering factors like age, gender, smoking habits, blood pressure, glucose levels, etc.

Data and Methodology:

The chosen dataset consists of 4000+ data points related to medical and demographic information of people from the ages of 32-70 years. It consists of information like age, gender, level of education, current smoking status of the person, etc.

A brief description of the data is as follows:

male: a binary variable that takes the value 1 if the person is a male and 0 if they are a female
age: a continuous variable that takes integer values between 32 and 70.
education: indicates the education level of the person, takes integer values between 1-4
currentSmoker: a binary variable that indicates whether the person is a smoker
cigsPerDay: number of cigarettes the person smokes in a day
BPMeds: a binary variable that indicates whether the person takes medicine for blood pressure
prevalentStroke: a binary variable that indicates whether the person has had a stroke
prevalentHyp: a binary variable that indicates whether the person has hypertension
diabetes: a binary variable that indicates whether the person has diabetes
totChol: the total cholesterol level of the person
sysBP: systolic blood pressure of the person
diaBP: diastolic blood pressure of the person
BMI: BMI of the person, continuous variable
heartRate: the heart rate of a person
glucose: the blood glucose level of the person
TenYearCHD: the dependent variable, a binary variable that takes the value 1 if the person is predicted to have CHD after ten years and 0 if not.

A logistic regression model was developed that determined the likelihood of a person getting chronic heart disease after 10 years of the initial study. To predict the outcome, variables like age, gender, number of cigarettes smoked in a day, and glucose level of a person were considered. Education level, cholesterol level and other medical data was also considered initially, but it was found that these variables are not significant at a 95% confidence interval to predict the likelihood of chronic heart disease in 10 years.

The marginal effects for all variables at the mean were also calculated. This was done to study the effect of each variable on the outcome of a model by measuring the expected change in the outcome variable as a result of a change in the explanatory variable.

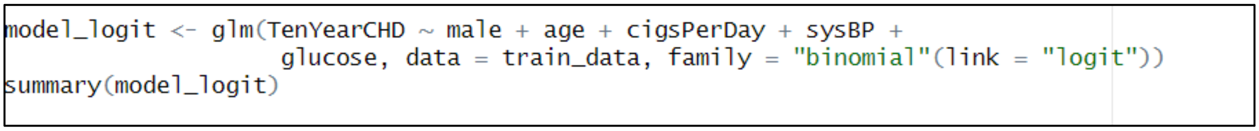
Steps followed:

- Required packages were installed and loaded into the environment.
- The dataset was split into training (80%) and test data (20%), stratified by the TenYearCHD variable, as that is our outcome variable in this case. Using separate training and testing data helps calculate the accuracy of the model and check for overfitting.
- Logistic regression model is created, with dependent variable *TenYearCHD*, and independent variables *male, age, cigsPerDay, sysBP, glucose.*

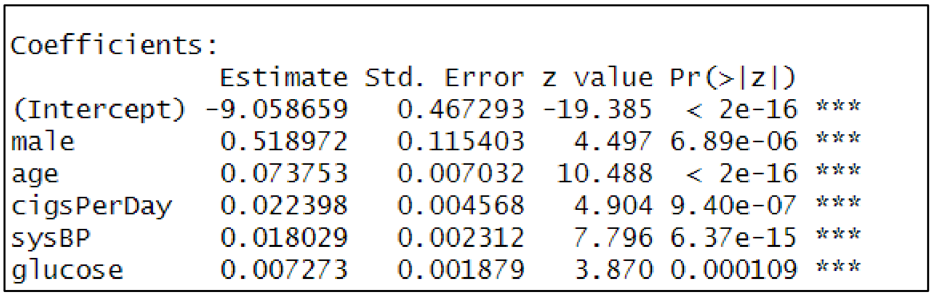
  The glm() function in R was used as it is flexible and can be used with various types of output variables. It is also customizable and can be used with various link functions.
- A correlation matrix is created to check whether mulitcollinearity exists. If there is multicollinearity in the model, some variables might need to be excluded.
- Marginal effects of the independent variables are measured about the mean. These values give us the effect that a specific variable has on the outcome when it is changed by one unit.

Discussion of Results:

The summary of the model is given below:

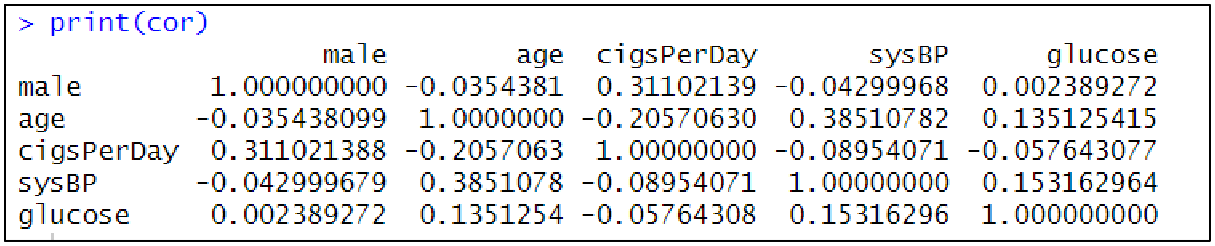


```
model_logit <- glm(TenYearCHD ~ male + age + cigsPerDay + sysBP +
                     glucose, data = train_data, family = "binomial"(link = "logit"))
summary(model_logit)
```



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -9.058659   0.467293 -19.385  < 2e-16 ***
male         0.518972   0.115403   4.497 6.89e-06 ***
age          0.073753   0.007032  10.488  < 2e-16 ***
cigsPerDay   0.022398   0.004568   4.904 9.40e-07 ***
sysBP        0.018029   0.002312   7.796 6.37e-15 ***
glucose      0.007273   0.001879   3.870 0.000109 ***
```

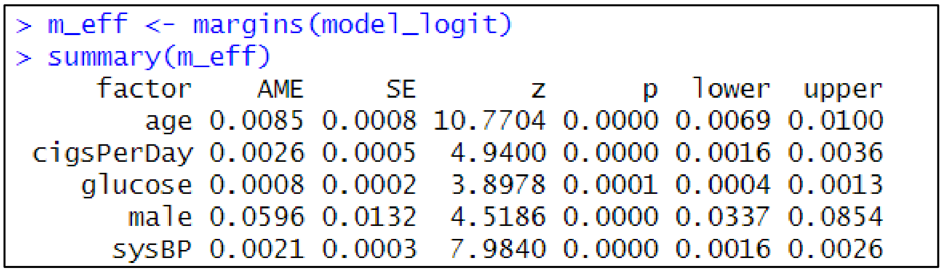
Correlation matrix of the independent variables is as follows:



```
> print(cor)
                    male        age   cigsPerDay       sysBP      glucose
male         1.000000000 -0.0354381  0.31102139 -0.04299968  0.002389272
age         -0.035438099  1.0000000 -0.20570630  0.38510782  0.135125415
cigsPerDay   0.311021388 -0.2057063  1.00000000 -0.08954071 -0.057643077
sysBP       -0.042999679  0.3851078 -0.08954071  1.00000000  0.153162964
glucose      0.002389272  0.1351254 -0.05764308  0.15316296  1.000000000
```

As all the values are below 0.5, it shows that there is no multicollinearity.

The summary of the marginal effects is given below:



```
> m_eff <- margins(model_logit)
> summary(m_eff)
     factor    AME     SE       z      p  lower  upper
        age 0.0085 0.0008 10.7704 0.0000 0.0069 0.0100
 cigsPerDay 0.0026 0.0005  4.9400 0.0000 0.0016 0.0036
    glucose 0.0008 0.0002  3.8978 0.0001 0.0004 0.0013
       male 0.0596 0.0132  4.5186 0.0000 0.0337 0.0854
      sysBP 0.0021 0.0003  7.9840 0.0000 0.0016 0.0026
```

Interpretation of the marginal effects:

- Age: AME = 0.0085

  p-value = 0.0000 (significant)

  This means that one unit change in age increases the chance of CHD in 10 years by 0.85%, when the other factors are constant. Older people are more likely to develop CHD in 10 years.

- cigsPerDay: AME = 0.0026

  p-value = 0.0000 (significant)

  This implies that one unit change in the value of cigsPerDay increases the likelihood of the outcome by 0.26%, other things remaining constant. A higher number of cigarettes per day results in a higher probability of developing CHD in the future.

- glucose: AME = 0.0008

  p-value = 0.0001 (significant)

  This implies that one unit change in the glucose level increases the likelihood of the outcome by 0.08%, other things remaining constant. People with higher glucose levels are more likely to develop CHD in the future.

- male: AME = 0.0596

  p-value = 0.0000 (significant)

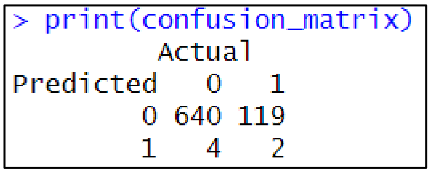
  This implies that one unit change in the value of this variable increases the likelihood of the outcome by 5.96 %, other things remaining constant. It can also be said that men are 5.96% more likely to develop CHD in the future.

- sysBP: AME = 0.0021

  p-value = 0.0000 (significant)

  This implies that one unit change in the value of this variable increases the likelihood of the outcome by 0.21 %, other things remaining constant. The higher a person's systolic blood pressure, the more likely they are to develop CHD in the future.

The model is trained on the subset of the data called train_data and then the model is run on test_data to predict the values. The confusion matrix is given below, which shows us the counts of true positives and negatives predicted by the model:



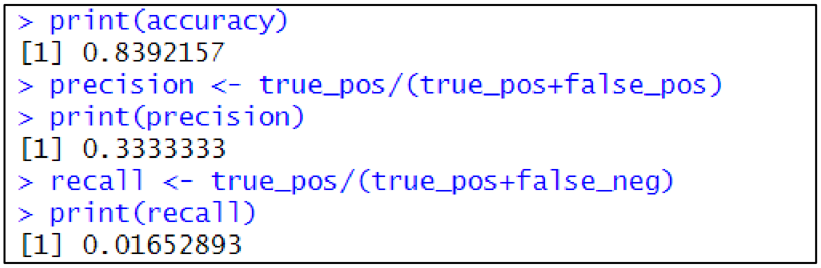
```
> print(confusion_matrix)
         Actual
Predicted   0   1
        0 640 119
        1   4   2
```

The model performance is evaluated based on 3 metrics – accuracy, precision, and recall.

Accuracy refers to the percentage of correct predictions among all the predictions made.

Precision refers to the percentage of the true positives among all the predicted positives, while recall measures the percentage of true positives among the true positives and false negatives.

The metrics for the model are:



```
> print(accuracy)
[1] 0.8392157
> precision <- true_pos/(true_pos+false_pos)
> print(precision)
[1] 0.3333333
> recall <- true_pos/(true_pos+false_neg)
> print(recall)
[1] 0.01652893
```

Conclusion:

This study demonstrates the importance of selecting and using appropriate statistical models to analyze limited dependent variables (LDVs). By focusing on predicting the likelihood of chronic heart disease (CHD) after ten years using logistic regression, it highlights how proper model selection ensures reliable predictions and actionable insights.

Key findings include the significance of independent variables like age, gender, daily cigarette consumption, systolic blood pressure, and glucose levels in determining CHD risk. The study confirms that marginal effects provide an understanding of how changes in these variables influence outcomes. Moreover, the absence of multicollinearity strengthens the reliability of the model.

This study is a minor example of applying econometric techniques to healthcare data, which might result in the development of predictive tools that enable personalized care and early interventions. Overall, this analysis reaffirms that careful feature selection and rigorous model evaluation are crucial for impactful and accurate decision-making in data-driven fields.

Such models may be of use in certain policies and healthcare related decision-making. Insurance companies can use these models and encourage better lives by tailoring health insurance premiums depending on individual risk variables such as blood pressure, glucose levels, or smoking behaviours. Pharma companies can develop products that target and deal with specific factors like blood pressure medication and advanced glucose measuring devices. Other companies can also take the help of such models to develop medical devices and wearables that help predict the probability of getting chronic heart disease.

References:

1. Rod Falvey, Gulasekaran Rajaguru, and Robert Wrathall. “Do Jockeys “Look to” or “Rest on” Their Laurels After a Sequence of Winning Rides?”
2. Robert Wrathall, Rod Falvey and Gulasekaran Rajaguru. “Do (Australian) jockeys have hot hands?”
3. Wiersema, Margarethe F., and Harry P. Bowen. "The use of limited dependent variable techniques in strategy research: Issues and methods." *Strategic management journal* 30.6 (2009): 679-692. https://doi.org/10.1002/smj.758