**Nour Mohamed Eletreby**

**Task 6**

**Logistic Regression Analysis**

**Introduction**

Logistic regression analysis is a statistical method used for modeling and analyzing the relationship between a binary or categorical dependent variable (the outcome or response variable) and one or more independent variables (predictors or features). Unlike simple linear regression, which is used for continuous outcome variables, logistic regression is designed specifically for situations where the dependent variable is categorical and has two possible outcomes, typically coded as 0 and 1, or "success" and "failure."

The primary goal of logistic regression is to estimate the probability that the binary outcome occurs as a function of the independent variables. It models the log-odds (logit) of the probability, which can take any real value, and then transforms this log-odds value into a probability between 0 and 1 using the logistic or sigmoid function. This allows logistic regression to provide predictions and insights about the likelihood of an event happening based on the values of the predictor variables.

In essence, logistic regression helps answer questions like whether a customer will make a purchase (yes/no), whether a patient has a disease (presence/absence), whether an email is spam or not (spam/ham), and many other binary classification problems across various fields, such as healthcare, marketing, finance, and social sciences.

**Example to illustrate logistic regression analysis**

**Problem:** Predict whether a student will be admitted to a university based on their score on a single entrance exam. The outcome variable is binary: admitted (1) or not admitted (0).

**Data:** We have a small dataset of 5 students with their exam scores and admission outcomes:

| **Exam Score (X)** | **Admitted (Y)** |
| --- | --- |
| 45 | 0 |
| 60 | 1 |
| 55 | 0 |
| 75 | 1 |
| 80 | 1 |

**Logistic Regression Analysis:**

1. **Data Preparation:** We have the data without any missing values or outliers.
2. **Model Building:** We'll build a logistic regression model to predict admission based on exam scores. The logistic regression equation will be:

Log(odds of admission) = β0 + β1 \* Exam Score

* + β0 represents the intercept (the log-odds of admission when the exam score is 0).
  + β1 is the coefficient for the "Exam Score" variable, indicating how much the log-odds change for a one-unit increase in the exam score.

1. **Model Training:** We estimate the coefficients (β0 and β1) using a statistical method like maximum likelihood estimation.
2. **Model Interpretation:** After training, we get the following coefficients:
   * β0 (Intercept) = -4.234
   * β1 (Exam Score) = 0.116

Now, we can use these coefficients to make predictions. For example, if a student has an exam score of 70:

Log(odds of admission) = -4.234 + 0.116 \* 70 = -4.234 + 8.12 = 3.886

To get the probability of admission, we apply the sigmoid function:

Probability(Admitted) = 1 / (1 + e^(-3.886)) ≈ 0.979

So, a student with an exam score of 70 has an estimated probability of approximately 97.9% of being admitted to the university.

1. **Model Evaluation:** We can assess the model's performance using various metrics and techniques, but in this simple example, we've focused on prediction based on the logistic regression model.

**Types of logistic regression analysis**

Logistic regression analysis comes in several variations, each tailored to different types of problems and data. Here are some common types of logistic regression analysis:

1. Binary Logistic Regression: This is the most common type, and it's used when the dependent variable is binary, having only two possible outcomes (e.g., yes/no, pass/fail, 1/0). It's employed for classification tasks where the goal is to predict one of two classes based on predictor variables.
2. Multinomial Logistic Regression: this type is used when the dependent variable has more than two unordered categories. It's applied when you have a categorical outcome variable with three or more non-ordinal levels (e.g., colors like red, green, blue).
3. Ordinal Logistic Regression: When the dependent variable is ordinal (i.e., it has ordered categories but not necessarily evenly spaced), ordinal logistic regression is used. This is suitable for situations where the outcome variable has multiple ordered categories, like customer satisfaction levels (e.g., very dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, very satisfied).

**Assumptions of logistic regression analysis**

1. **Binary or Ordinal Outcome:** Logistic regression assumes that the dependent variable is binary or ordinal in nature. In the case of binary logistic regression, there are only two categories (e.g., yes/no, pass/fail), while ordinal logistic regression deals with ordered categories (e.g., low/medium/high).
2. **Linearity of Log-Odds:** Logistic regression assumes a linear relationship between the log-odds of the outcome variable and the predictor variables. This means that a one-unit change in a predictor variable results in a constant change in the log-odds of the outcome, while holding other variables constant. You can check this assumption by plotting the log-odds against the predictor variables to see if the relationship is roughly linear.
3. **Independence of Observations:** Logistic regression assumes that the observations are independent of each other. This means that the outcome of one observation should not depend on the outcome of another observation. Violations of this assumption can occur in clustered or time-series data, which may require specialized techniques like hierarchical logistic regression.
4. **No Multicollinearity:** Multicollinearity occurs when two or more predictor variables in the model are highly correlated with each other. This can make it difficult to determine the individual effects of these variables on the outcome. It's important to check for multicollinearity and consider addressing it through variable selection or regularization techniques.
5. **No Outliers:** Logistic regression assumes that there are no extreme outliers in the data that could unduly influence the model's results. Outliers can lead to biased coefficient estimates and should be carefully examined and, if necessary, addressed.
6. **No Endogeneity:** Logistic regression assumes that there is no endogeneity, meaning that the predictor variables are not influenced by the outcome variable. In some situations, especially in causal modeling, this assumption may not hold, and more advanced techniques like instrumental variable regression may be necessary.

**Advantages of logistic analysis**

* Simple to understand, easy to implement, and efficient to train
* Performs well when the dataset is linearly separable
* It offers the direction of association (positive or negative)
* Useful to find relationships between features

**Limitations**

1. **Linearity Assumption:** Logistic regression assumes a linear relationship between the log-odds of the outcome and the predictor variables. If the relationship is highly non-linear, logistic regression may not model the data effectively. In such cases, more flexible modeling techniques like decision trees or neural networks may be more appropriate.
2. **Assumptions of Independence:** Logistic regression assumes that observations are independent of each other. In practice, this assumption may be violated when dealing with clustered or time-series data, which requires more advanced modeling techniques like hierarchical logistic regression or time series analysis.
3. **Limited Outcome Types:** Logistic regression is primarily designed for binary or ordinal outcomes. It may not be suitable for situations where the outcome has more complex structures, such as nominal outcomes with multiple unordered categories.

**Conclusion**

In summary, logistic regression is like a helpful tool in the world of data analysis. It's great for predicting things that have only two choices, like "yes" or "no," and it's good at telling us how different factors affect those predictions.

So, in the end, logistic regression is a handy tool that helps us make predictions about binary choices, but we should always use it with care and consider other options when needed.

**Logistic Regression Report on Bank Marketing**

**Introduction**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Logistic regression is an important statistical technique in the analysis of bank marketing data, especially when the marketing campaigns are based on phone calls and involve predicting binary outcomes like whether a client will subscribe to a bank term deposit ('yes') or not ('no').

**Input variables:** # bank client data: 1 - age (numeric)

2- marital: marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)

3 - education (categorical: "unknown", "secondary", "primary", "tertiary")

**Output variable (desired target):** 4 - y - has the client subscribed a term deposit? (binary: "yes", "no")

**The data consists of 45211 units**

**I will be using SPSS for this logistic regression**

**Descriptive Statistics**

First the discrete variables (input data):

I will start with marital status variable

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **marital** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | divorced | 5207 | 11.5 | 11.5 | 11.5 |
| married | 27214 | 60.2 | 60.2 | 71.7 |
| single | 12790 | 28.3 | 28.3 | 100.0 |
| Total | 45211 | 100.0 | 100.0 |  |

Figure 1: descriptive statistics on marital variable

I can see that 12790 of the bank clients are single with 28.3% and 27214 of the clients are married with 60.2% and 5207 are divorced or widowed with 11.5%

The education variable

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **education** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | primary | 6851 | 15.2 | 15.2 | 15.2 |
| secondary | 23202 | 51.3 | 51.3 | 66.5 |
| tertiary | 13301 | 29.4 | 29.4 | 95.9 |
| unknown | 1857 | 4.1 | 4.1 | 100.0 |
| Total | 45211 | 100.0 | 100.0 |  |

Figure (2): descriptive statistics for education variable

I can see that 6851 of the bank clients are in the primary level with 15.2% while 23202 of the clients are in the secondary level with 51.3% and 13301 of the clients are in the tertiary level with 29.4% and 1857 of the clients are unknown for us with 4.1%

Second the discrete variable (the output variable y):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Outcome Variable** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | no | 39922 | 88.3 | 88.3 | 88.3 |
| yes | 5289 | 11.7 | 11.7 | 100.0 |
| Total | 45211 | 100.0 | 100.0 |  |

Figure (3): descriptive statistics for the outcome variables

I can see that 5289 of the bank clients subscribed a term deposit with 11.7% (YES) while 39922 of the clients did not describe a term deposit 88.3% (NO)

Third the continuous input variable (the age)

Figure (4): descriptive statistics for age

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | |
|  | N | Minimum | Maximum | Mean | Std. Deviation |
| age | 45211 | 18.0 | 95.0 | 40.936 | 10.6188 |
| Valid N (listwise) | 45211 |  |  |  |  |

I can see that the minimum value for age 18 and the maximum 95 with mean 40.936 and standard deviation 10.6188

**Logistic regression**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | age | .020 | .002 | 173.147 | 1 | .000 | 1.020 |
| marital |  |  | 258.850 | 2 | .000 |  |
| marital(1) | -.454 | .053 | 72.701 | 1 | .000 | .635 |
| marital(2) | -.576 | .036 | 258.172 | 1 | .000 | .562 |
| education |  |  | 205.438 | 3 | .000 |  |
| education(1) | -.451 | .081 | 31.165 | 1 | .000 | .637 |
| education(2) | -.200 | .072 | 7.793 | 1 | .005 | .819 |
| education(3) | .170 | .073 | 5.457 | 1 | .019 | 1.185 |
| Constant | -2.369 | .093 | 654.971 | 1 | .000 | .094 |
| * + 1. Variable(s) entered on step 1: age, marital, education.   Figure (5): logistic regression analysis for our model | | | | | | | |

I can see that all variables in our significant (highlighted in yellow) so this is the model I will be working on (all variables including dummies are significant) we have 2 dummies for marital and 3 dummies for education (following k-1 rule to avoid multicollinearity)

Now I will interpret on odds ratio (highlighted in blue):

-The odds ratio for age is 1.020, meaning that the for the odds of a person in a bank subscribed a bank deposit (y=1) change by a factor of 1.020 with every unit increase on age holding other variables constant.

-The base category for marital is single, divorced clients is approximately 37% relatively less likely to be subscribed in a bank deposit than single clients (odds ratio=0.635) holding other variables constant.

-The base category for marital is single, married clients is approximately 44% relatively less likely to be subscribed in a bank deposit than single clients (odds ratio=0.562) holding other variables constant.

-The base category for education is unknown, primary clients is approximately 36% relatively less likely to be subscribed in a bank deposit that clients who chose unknown (odds ratio=0.637) holding other variables constant.

-The base category for education is unknown, secondary clients is approximately 18.1% relatively less likely to be subscribed in a bank deposit than clients who chose unknown (odds ratio=0.819) holding other variables constant.

-The base category for education is unknown, tertiary clients is approximately 18.5% relatively more likely to be subscribed in a bank deposit than clients who chose unknown (odds ratio=1.185) holding other variables constant.

**Conclusion**

In summary, logistic regression is a valuable tool for analyzing bank marketing data involving phone call campaigns because it is well-suited for binary classification problems, provides interpretable results, handles multiple contacts, and offers probability estimates for better decision-making in marketing strategies. All variables in the model is significant so I will keep all of them.