# Logistic regression

*Key points from hybrid webinar simple, multiple, multinomial and ordinal logistic regression*

by Nur Aazifah Ilham[1] and Lim Poh Ying[2]
[1]Clinical Research Unit, HSAAS; [2]Department of Community Health, FMHS UPM

## What is Logistic Regression?

- Estimate the relationship between the categorical dependent variable with one or more independent variable/covariate
- Common in medical studies
- Goal: To establish a model that

❑ Best fit

❑ Parsimonious

❑ Biologically sound/Biological plausibility

## How parameters will be estimated in logistic regression?

There are many methods for parameter estimation in logistic regression, but commonly in medical health research, we use maximum likelihood.

## What is maximise likelihood estimation?

Estimation method to find the value of model parameters that make the observed data most probable under the model

## Terminology:

### Parsimonious

A parsimonious model is a model that achieves a desired level of goodness of fit using as few explanatory variables as possible.

### Parameter

Parameters do not relate to actual measurements or contribute to individuals but will quantify the theoretical model.

### Variables

are quantities that vary among individuals.

### Estimation

The process of calculating statistics from sample data as an approximation of a parameter of the population

- Two types of estimation:

– Point: a single numerical value used as an estimation of a parameter value.

– Interval: two numerical values presented as a range that includes the parameter value, confidence interval.

# What are the types of logistic regression?

| Independent variables/ Predictor | Dependent Variables/ Outcome measure | Example | Logistic Regression |
|---|---|---|---|
| Single variable | Binary | Cholesterol level ~ CAD+ or CAD-ve | Simple |
| Multiple | Binary | Age+Cholesterol level+Gender ~ CAD+ or CAD - | Multiple |
| Single/ Multiple | Polytomous(>2) | Parents education level~Food choice by children (Fast food, Vegan, Balance diet) | Multinomial |
| Single/Multiple | Ordinal | Years of smoking ~Stage of cancer | Ordinal |

## Equation in logistic regression:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

$$OR = e^{bx}$$

Odd ratio is the product of exponential of beta coefficient

What is an Odd ratio?

- Odds=Chance
- Odds of an event is the ratio between the number of events occurring vs the number of events not occurring
- Odd ratio is calculated by dividing two odds
- OR> 1, OR =1, OR <1

Variables in the Equation

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) Lower | Upper |
|---|---|---|---|---|---|---|---|---|---|
| Step 1[a] | dbp | .050 | .003 | 212.621 | 1 | .000 | 1.051 | 1.044 | 1.058 |
| | chol | .137 | .035 | 15.663 | 1 | .000 | 1.146 | 1.071 | 1.227 |
| | gender(1) | .398 | .092 | 18.552 | 1 | .000 | 1.488 | 1.242 | 1.783 |
| | Constant | -7.242 | .349 | 429.940 | 1 | .000 | .001 | | |

a. Variable(s) entered on step 1: dbp, chol, gender.

# Steps in data analysis :

1)Data exploration and cleaning

Checking data set, looking at measurement, level of data, missing data and outliers.

2)Univariable analysis

– Simple logistic regression – open enter

3)Variable selection – Applicable to multivariable analysis of multiple, multinomial and ordinal logistic regression.

Univariate analysis p-value < 0.25 put in the final model

There are many methods (in SPSS) such as backward selection, forward selection) but bear in mind the aim is to produce a model of:

- Best fit
- Parsimonious
- Biologically sound/Biological plausibility

4)Checking multicollinearity/interaction

5)Checking model adequacy(model fit) and assumption

- Hosmer-Lemeshow Test : p.value Chi-Square >0.05
- Area under the curve : >0.7
- Correctly Classified Percentage : >70%

6)Assumptions

a)Random sample

b)Independent sample- error term should be independent

c)Dependent variable- binary/dichotomous variable

d)No multicollinearity

e)Linearity- There is a linear relationship between continuous x and logit y.

7)Interpretation

## Terminology:

**Multicollinearity**
Correlation between the independent variable

**Interaction**
Situation in which two or more predictor/risk factors modify the effect of each other or outcome.

It can be an additional or multiplication interaction.

## ✓ Checking model adequacy(model fit) and assumption

Model Fit/Model Adequacy/Goodness of Fit

Tests to determine whether a set of actual observed values match with the predicted by the model.

### Hosmer-Lemeshow test

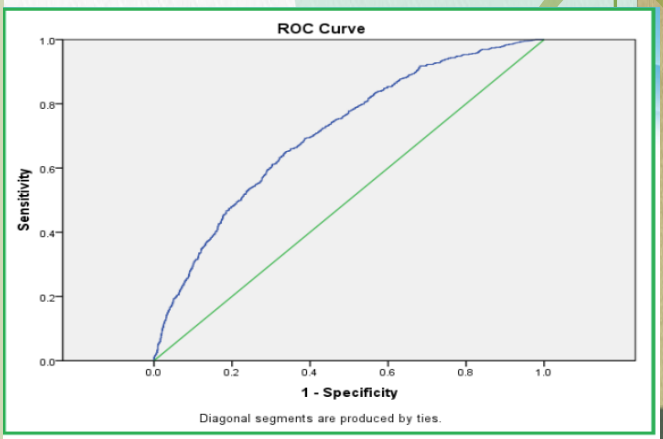
- Based on grouping cases into deciles of risk
- It compares the observed probability with the expected probability within each centile
- P-value > 0.05 meaning there is no significant difference between the observed probability and the expected probability

**Hosmer and Lemeshow Test**

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 10.790 | 8 | .214 |

### ROC

- Range from 0-1
- A value of 0.5 means the model cannot be used for discrimination
- The recommended area for model fitting is at least 0.7




**Area Under the Curve**

Test Result Variable(s):Predicted probability

| Area | Std. Error[a] | Asymptotic Sig.[b] | Asymptotic 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower Bound | Upper Bound |
| .709 | .011 | .000 | .688 | .730 |

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

### Classification table

- Classified percentage >70% is consider good.
- Can calculate sensitivity and specificity

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | coronary artery disease | | |
| | | | no cad | cad | Percentage Correct |
| Step 1 | coronary artery disease | no cad | 4036 | 33 | 99.2 |
| | | cad | 604 | 17 | 2.7 |
| | Overall Percentage | | | | 86.4 |

a. The cut value is .500

## Simple Logistic Regression

- To estimate the relationship between a single IV/predictor to a binary outcome.
- Use as a preliminary step in selecting variables for multiple logistic regression.
- Rarely being done a lot as it did not cater for the confounding/third variable in the model.
- Logistic regression is similar to ordinary least squares (OLS) regression, but it uses a different function to model the relationship between the dependent variable and the independent variables.
- The logistic function is used to transform the predicted probabilities from the model into probabilities that are between 0 and 1.

# Multiple logistic regression

Multiple logistic regression is a statistical method used to analyze binary dependent variables (i.e., variables that have only two possible outcomes, such as yes/no, true/false, or heads/tails) with more than one independent variable. It is used to identify/predict the factors that are associated with the outcome of the dependent variable. Multiple logistic regression, incorporating all relevant variables simultaneously helps uncover the collective influence or confounding on the outcome.

For instance, in research objective is to investigate whether diabetes (binary outcome: yes and no) is affected by a sedentary lifestyle, treatment A, soft drink intake and intervention program.

Variable selection methods, including forward selection and backward elimination, compute the best model. Evaluation of the model considers factors such as parsimony, choosing models with a balance of simplicity and explanatory power, biological plausibility, and adherence to assumptions like random sampling, independence of samples, absence of multicollinearity, linearity, and absence of outliers.

In the end, a comprehensive analysis and rigorous variable selection process lead to a well-validated model, providing valuable insights into the complex interplay between various factors and the likelihood of diabetes. This model, meeting the assumptions and criteria outlined, can serve as a reliable tool for understanding and predicting diabetes risks in a given population.

The results of a multiple logistic regression model are typically reported in the form of coefficients and odds ratios. The coefficients represent the change in the log odds of the dependent variable for a one-unit increase in the independent variable. The odds ratios represent the ratio of the odds of the dependent variable occurring for a given value of the independent variable to the odds of the dependent variable occurring for the reference value of the independent variable.

**Analysis Steps:**

- Start with a single independent variable (Simple logistic regression)
- Multiple logistic regression: Expand to include all relevant variables.
- Check for multicollinearity and interaction(preliminary final mode).
- Validate assumptions.
- Finalize the model for presentation.

Variable selection method

Forward selection: Variables are sequentially entered into the model, from most significant first

Backward selection: All variable are entered in the model and then sequentially removed, from the least significant first)

Want to get a deeper understanding of the variable selection in the model and how to check the multicollinearity and interaction?

Contact cru at cru_hsaas@upm.edu.my if you are interested in watching the recording hybrid webinar and SPSS demonstration as well as getting the slides and the training dataset.

# Multinomial Logistic Regression: An Introduction

Multinomial logistic regression is a statistical method to analyse categorical data with more than two categories. This type of regression is often used to study factors associated with choice, preference and decisions.

Example :

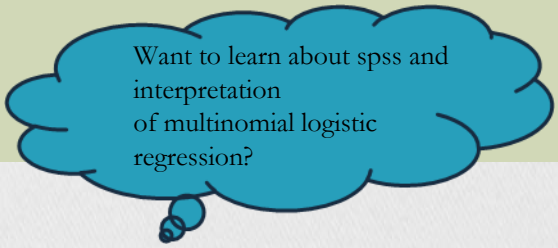
Let's say we want to study the factors associated with the type of noodles preferred by Malaysian consumers. We could use multinomial logistic regression to analyse the data. The DV would be the type of noodles preferred (laksa, curry mee, or mee goreng) and the IVs could be age, gender, income group, etc.

The multinomial logistic regression would model the probabilities of each of the three categories of the DV. For example, the model would estimate the probability of a consumer preferring laksa, the probability of a consumer preferring curry mee, and the probability of a consumer preferring mee goreng.

The results of the multinomial logistic regression would tell us which factors are associated with a higher or lower probability of a consumer preferring a particular type of noodles. For example, we might find that younger consumers are more likely to prefer laksa, or that consumers with a higher income are more likely to prefer curry mee.

Conclusion

Multinomial logistic regression is a powerful tool for analysing categorical data with more than two categories and understanding the factors associated with the choices, preferences, and decisions of individuals. It is often used in marketing research, consumer behaviour research, and other fields.




Contact cru at cru_hsaas@upm.edu.my, if you are interested to watch the hybrid webinar and SPSS demonstration as well as getting the slides and the training dataset.

# Ordinal Logistic Regression: An Introduction

Ordinal logistic regression is a statistical method used to dependent variable ordinal data, which is data that can be ranked into categories but does not have equal intervals between categories. This type of regression is often used to study factors associated with attitudes, opinions, and beliefs.

Ordinal logistic regression differs from simple and multiple logistic regression in that it uses cumulative probabilities of categories in its equation. An example of ordinal logistic regression is to study the factors associated with higher/lower obesity groups.

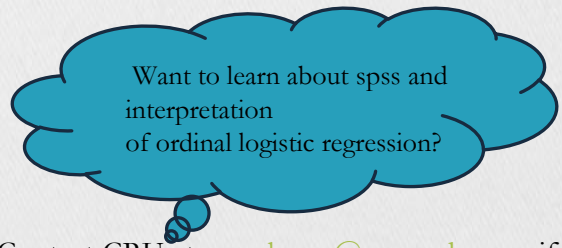
•The assumptions for ordinal logistic regression are:

- •The dependent variable must have ordinal data with more than two categories.
- •The independent variable can be continuous or categorical.
- •The DV must have mutually exclusive categories.
- •There must be linearity between the continuous IV and DV.
- •There must be no multicollinearity for continuous IVs.

Example

Let's say we want to study the factors associated with the Likert scale of attitude towards a new product. We could use ordinal logistic regression to analyze the data. The DV would be the Likert scale (strongly disagree, disagree, neutral, agree, strongly agree) and the IVs could be age, gender, income group, etc. The ordinal logistic regression would model the cumulative probabilities of the DV categories. For example, the model would estimate the probability of a respondent strongly agreeing with the new product, the probability of a respondent agreeing with the new product, and so on.

The results of the ordinal logistic regression would tell us which factors are associated with a higher or lower probability of a respondent having a positive attitude towards the new product. For example, we might find that younger respondents are more likely to have a positive attitude towards the new product, or that respondents with a higher income are more likely to have a positive attitude towards the new product.




Contact CRU at cru_hsaas@upm.edu.my , if you are interested to watch the hybrid webinar and SPSS demonstration as well as getting the slides and the training dataset.