STAT1201 - Summer Semester 2022

Lecture 12 -Module 11: Chi-square test and Logistic Regression

Dr. Wasanthi Thenuwara

Learning Objectiives

- Chi-square test
- Logistic Regression

Chi-square Test - An overview

- Use for hypothesis testing
- Commonly use to test the relationships between two categorical variables
- We will learn two main uses of Chi-square test in this lecture
  - *A Chi-square test for independence* - to test the independence of two categorical variables (we compare two variables in a contingency table to see if they are related or independent)
  - *A Chi-square goodness of fit test* - to evaluate the goodness of fit of a data set to a specific probability distribution

Chi-square test for independence of two categorical variables

e.g. Gender and level of education (Below greade 12, Grade 12, Bachelor's degree, Postgraduate degree)

H_0 : The two categorical variables are independent (i.e. there is no relationship between gender and education level)

H_1 : The two categorical variables are related (i.e. there is a relationship between gender and education level)

Chi-square test for independence of two categorical variables - Example

The following table records the data from a double blind experiment on the effectiveness of oral nicotine inhalers in reducing smoking (Bolliger et. al. 2000). 400 subjects who had tried to reduce their smoking and but failed to do so were considered in the experiment. Nicotine inhalers were randomly assigned to half of the subjects while the other half received placebo inhaler. After 4 months the researchers recorded which subjects had sustained in smoking reduction (Smoking Reduction: Yes/No).

Contingency table for observed frequencies

	Nicotine	Placebo	Sum
No	148	182	330
Yes	52	18	70
Sum	200	200	400

Chi-square test for independence of two categorical variables - Example cont...

We can use Chi-square test to test whether inhaler content and smoking reduction is related.

H_0 : Inhaler content and smoking reduction are independent

H_1 : Inhaler content and smoking reduction are related

The Chi-square ($\chi 2$) test statistic is:

$\chi 2_{stat} = \sum_{allcells} \frac{(f_o - f_e)^2}{f_e}$

f_o - Observed frequency and f_e - Expected frequency

Expected frequencies can be found using the following formula.

$\frac{RowTotal * ColumnTotal}{GrandTotal}$

Chi-square test for independence - Example cont. . .

Observed frequencies

	Nicotine	Placebo	Sum
No	148	182	330
Yes	52	18	70
Sum	200	200	400

Expected frequencies

	Nicotine	Placebo
No	165	165
Yes	35	35

$$\chi 2_{stat} = \frac{(148-165)^2}{165} + \frac{(182-165)^2}{165} + \frac{(52-35)^2}{35} + \frac{(18-35)^2}{35}$$

$$\chi 2_{stat} = 20.02$$

This $\chi 2_{stat}$ has a $\chi 2$ distribution with df = (rows -1)*(columns -1)

Chi-square test for independence - Example cont. . .

$\chi 2_{stat} = 20.02$

p - value = P($\chi 2_{(1)} > 20.02$)

Using R:

p - value = 1 - pchisq(20.02, df=1)

p - value = 0.000000766

p - value < 0.01, there is strong evidence to conclude that inhaler content and smoking reduction outcome are related.

Chi-square test for independence - Example cont. . .

Poll Question 1

Suppose that you test whether gender and level of education are related using Chi-square test. You considered four levels of education (Below_12, Grade_12, Bachelor's degree, Post graduate degree)) and Gender (Male, Female). The degrees of freedom used to find the p-value is:

a) 1

b) 2

c) 3

d) 4

Chi-square test for independence - Example using R

```
inhaler = read.csv("M11Inhaler.csv")

## to obtain observed frequencies
addmargins(table(inhaler$Reduction, inhaler$Inhaler))
```

```
     Nicotine Placebo Sum
  No       148     182 330
  Yes       52      18  70
  Sum      200     200 400
```

```
chisq=chisq.test(table(inhaler$Reduction, inhaler$Inhaler))
chisq
```

```
    Pearson's Chi-squared test with Yates' continuity corre

data:  table(inhaler$Reduction, inhaler$Inhaler)
X-squared = 18.857, df = 1, p-value = 1.409e-05
```

Chi-square test for independence - Example using R

```
## To obtain the expected frequencies

chisq$expected
```

```
     Nicotine Placebo
No        165     165
Yes        35      35
```

Chi-square goodness of fit test

- Use to determine how well a set of data matches a specific probability distribution.
- Compare the observed frequencies in a category with the theoretically expected frequencies, if the data follow a specific probability distribution.
- Use the rule of thumb that all expected frequencies should be at least one (1) and 80% of them should be at least five (5)

Chi-square goodness of fit test - Example

An inspector working for the office of Gambling Regulation wants to test a die from the local casino to ensure that it is a fair die. The die is thrown 1200 times and recorded the number of times each number is shown.

Outcome_of_throw	Observed_Frequency
1	185
2	190
3	210
4	205
5	195
6	215

H_0 : The die is fair (observations follow hypothesed distribution)

H_1 : The die is not fair (observations do not follow hypothesed distribution)

Chi-square goodness of fit test - Example cont. . .

If the die is fair we expect that each of the faces has the same probability of landing facing up. That is, P(1)=1/6, P(2)=1/6,. . . , P(6)=1/6. Since n=1200, we can find expected frequencies using the formula $f_e = n * p$

Outcome	Observed_Freq	Expected_Freq	stat
1	185	200	1.125
2	190	200	0.500
3	210	200	0.500
4	205	200	0.125
5	195	200	0.125
6	215	200	1.125

$\chi 2_{stat} = 3.5$

degrees of freedom = (no. of categries - k -1)

where k is the number of parameters estimated using the data.

This $\chi 2_{stat}$ has a $\chi 2$ distribution with dgrees of freedom equals 5.

Chi-Square goodness of fit test - Example cont. . .

In this example no parameters were estimated using data as a fair die has a uniform distribution.

Using R, we can calculate p-value

pchisq(3.5, df=5)

This gives 0.3766

p-value = 1-0.3766 = 0.6234

Conclusion: We do not have evidence to conclude that the die is not fair.

Chi-Square goodness of fit test - Example cont...

** Poll Question 2**

Instead of a six sided die now suppose that you are inspecting a four sided die. What is the degrees of freedom of the Chi-square test to test whether it is a fair die.

a) 4

b) 3

c) 5

d) 2

Chi-square goodness of fit test - Example

A research conducted in 2015 found that the level of social media usage in Country A is as follows.

20% - Low, 70% - Medium, 10% - High

A similar research conducted in 2020 using 400 people found that the number of social media users in the three categories are 100, 255 and 45 respectively. Does the distribution of social media usage in 2020 follow the same distribution as in 2015?

.

.

.

.

.

.

Logistic Regression

Introduction

- In simple and multiple linear regression models (Module 8), the dependent (or response) variable was quantitative.
- In Logistic Regression, the dependent (or response) variable is dichotomous (binary).
- The log odds of the outcome (dependent) variable is modelled as a linear combination of independent (or explanatory) variables.

Logistic Regression

Examples

- What are the risk factors for lung cancer?

  – Response (or dependent variable) - Got a lung cancer (Yes = 1, No = 0)

  – Independent (or explanatory) variables - No. of cigarettes smoke per day, Body weight

- Is there evidence that obese is related to weight?

  – Response (or dependent variable) - Is obese (Yes = 1, No = 0)

  – Independent (or explanatory variable) - Body weight

Odds and Odds Ratio (OR) Review

First define Odds and Odds Ratio.

$$\text{Odds} = \frac{Probability-of-event}{Probability-of-non-event}$$

$$\text{Odds} = \frac{p}{1-p}$$

Consider the following contingency table for a study of reducing smoking for two groups, nicotine inhalers and placebo.

	Nicotine	Placebo	Sum
No	148	182	330
Yes	52	18	70
Sum	200	200	400

Define p = probability of reduction in smoking for subjects who were given nicotine inhaler

$$\text{p} = \frac{52}{200} = 0.26$$

Odds and Odds Ratio (OR) Review

	Nicotine	Placebo	Sum
No	148	182	330
Yes	52	18	70
Sum	200	200	400

$p = \frac{52}{200} = 0.26$

Odds of reduction in smoking for subjects who were given nicotine inhaler:

$\frac{p}{1-p} = \frac{0.26}{1-0.26} = 0.3514$

That is, odds for a reduction in smoking for subjects who were given nicotine inhaler is 0.3514 to 1 or (1 to 2.845).

Similarly, the odds for a reduction in smoking for subjects who were given placebo inhaler is:

$\frac{18/200}{1-(18/200)} = \frac{0.09}{0.91} = 0.0989$

Odds and Odds Ratio (OR) Review

Odds Ratio (OR) is the relative measure of an effect. This allows to compare the effect of an intervention group of study relative to the placebo group. In our smoking example OR is defined as below.

$$\text{OR} = \frac{Odds-of-reduction-in-smoking-for-subjects-given-nicotine-inhaler}{Odds-of-reduction-in-smoking-for-subjects-given-placebo-inhaler}$$

$$\text{OR} = \frac{0.3514}{0.0989} = 3.55$$

That is, based on sample data the effect of inhaler on reduction in smoking is 3.55 times higher if someone is using a nicotine inhaler than using placebo inhaler. This implies that nicotine inhalers are beneficial in assisting reduction of smoking based on this sample data. However, we need to test whether this is due to sampling variability or it is a statistically significant outcome.

Odds $= \frac{p}{1-p}$. Since p is a probability, $0 <$ odds $< \infty$

$\ln(\frac{p}{1-p})$ is used in Logistic regression as $-\infty < \ln(\frac{p}{1-p}) < \infty$

For Odds < 1, $\ln(\frac{p}{1-p})$ is negative.

The Logistic Regression Model - Inferences

The population logistic regression model takes the following form.

$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X$

β_0 and β_1 are population intercept and slope parameters. These population parameters are unknown and should be estimated.

However, we cannot use least square method as we did in linear regression models to estimate the parameters.

We use maximum likelihood estimation method, but the details of this method will not be discussed at this level.

The estimated model is written as:

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

The Logistic Regression Model - Inferences - Example

Suppose a medical researcher is interested to see whether lung cancer is related to the number of cigarettes smoke per day.

The population logistic regression equation is:

$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X$

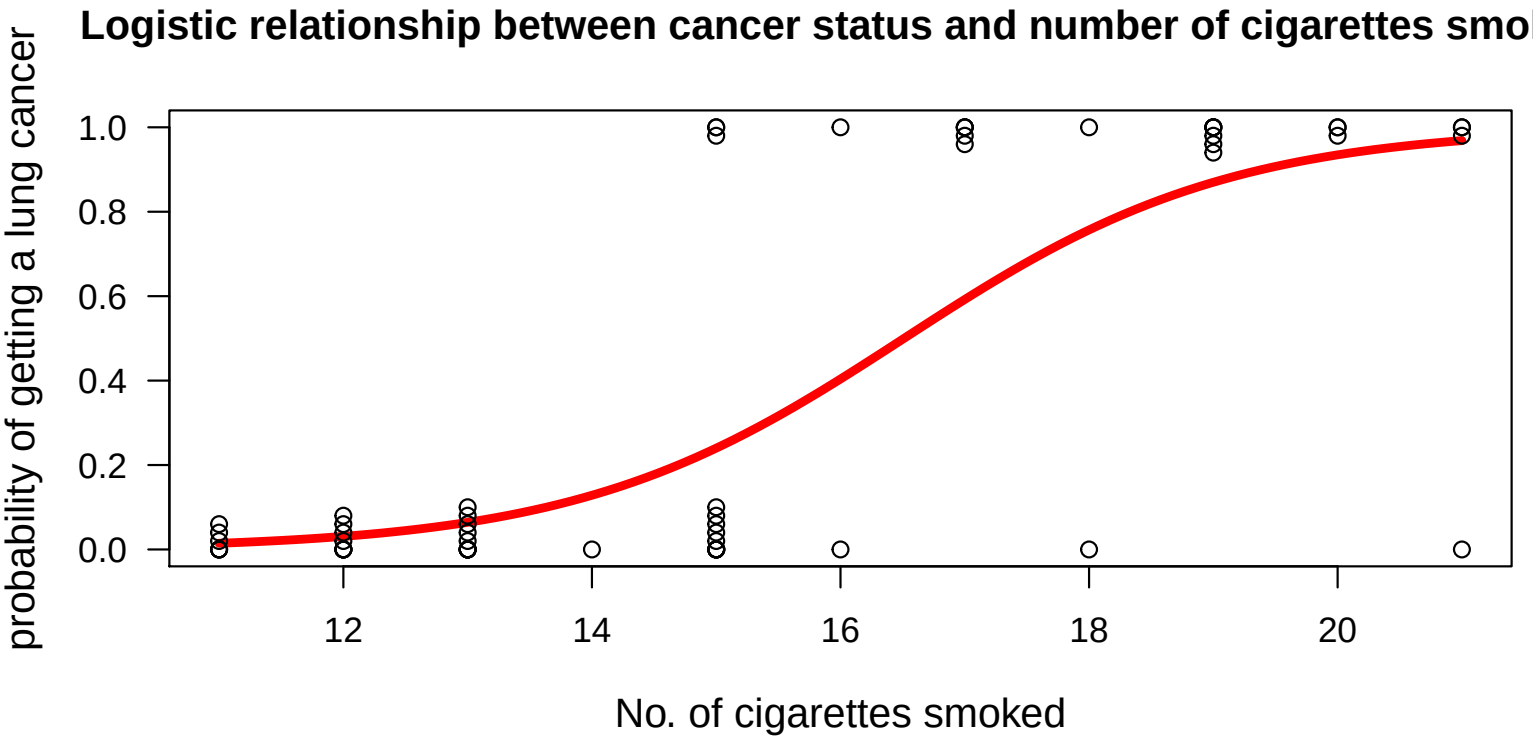
where p = probability of getting a lung cancer and X = number of cigarettes smoke per day

The Logistic Regression Model - Inferences - Example

R codes to get the plot to show the Logistic relationship between cancer status and number of cigarettes smoked

```
lungs = read.csv("M11Lungs.csv")
lungs$Cancer = ifelse(lungs$Cancer == "Yes", 1, 0)
library(popbio)
# Logistic relationship between cancer status and
# number of cigarettes smoked
logi.hist.plot(lungs$Cigarettes, lungs$Cancer, boxp=FALSE,
ylabel="probability of getting a lung cancer",
xlabel="No. of cigarettes smoked", main = "Logistic
relationship
between cancer status and number of cigarettes smoked")
```

The Logistic Regression Model - Inferences - Example



Logistic regression fits an "S" shaped logistic function. The curve goes from 0 to 1. The curve tells us the probability that a person getting a lung cancer based on the number of cigarettes he had per day.

The Logistic Regression Model - Inferences - Example

```
summary(glm(Cancer ~ Cigarettes, data = lungs,
            family = "binomial"))
```

```
Call:
glm(formula = Cancer ~ Cigarettes, family = "binomial", dat

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6295  -0.4043  -0.2510   0.5282   1.6895

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.5974     3.6584  -3.443 0.000575 ***
Cigarettes    0.7630     0.2263   3.371 0.000749 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

(Dispersion parameter for binomial family taken to be 1)
```

The Logistic Regression Model - Inferences - Example

The estimated logistic regression model:

$\ln(\frac{\hat{p}}{1-\hat{p}}) =$ -12.5974 + 0.7630X

where $\hat{p}$ is the estimated probability of getting a lung cancer

Rearranging to isolate $\hat{p}$:

$\frac{\hat{p}}{1-\hat{p}} = e^{-12.5974+0.7630X}$

$\hat{p} = e^{-12.5974+0.7630X}$ - $e^{-12.5974+0.7630X} * \hat{p}$

$\hat{p} = \frac{e^{-12.5974+0.7630X}}{1+e^{-12.5974+0.7630X}}$

The Logistic Regression Model - Inferences - Example

$$\hat{p} = \frac{e^{-12.5974+0.7630X}}{1+e^{-12.5974+0.7630X}} \text{ — (1)}$$

If X = 15, what is $\hat{p}$?

Substitute X = 15, in equation (1):

$$\hat{p} = \frac{e^{-12.5974+0.7630*15}}{1+e^{-12.5974+0.7630*15}}$$

$$\hat{p} = \frac{e^{-1.1524}}{1+e^{-1.1524}}$$

$$\hat{p} = \frac{0.3159}{1+0.3159}$$

$$\hat{p} = 0.24$$

That is, the estimated probability for getting a lung cancer for a person who smokes 15 cigarettes per day is 0.24.

$$\text{Odds} = \frac{0.24}{0.76} = 0.3159$$

Odds for getting a lung cancer for a person who smokes 15 cigarettes is 0.3159 to 1 (or 1 to 3.165).

The Logistic Regression Model - Inferences - Example

Poll Question 3

Based on the estimated model for lung cancer, the lower limit of a 95% CI for the population slope parameter is:

a) 0.7630 - 0.2263*qnorm(0.975)

b) 0.2263*qnorm(0.975)

c) 0.7630 + 0.2263*qnorm(0.95)

d) 0.7630 - 0.2263

Logistic Regression to compare Odds between groups

Consider nicotine inhaler study again.

Define an indicator variable X:

X = 1, if in the nicotine group; 0 in the placebo group

The population logistic regression equation is:

$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X$

where p = probability of reduction in smoking

Logistic Regression to compare Odds between groups

```r
inhaler = read.csv("M11Inhaler.csv")
inhaler$Inhaler = ifelse(inhaler$Inhaler == "Nicotine",
                         1, 0)
summary(glm(Reduction ~ Inhaler, data = inhaler,
            family="binomial"))
```

```
Call:
glm(formula = Reduction ~ Inhaler, family = "binomial", dat

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7760  -0.7760  -0.4343  -0.4343   2.1945

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3136     0.2471  -9.364  < 2e-16 ***
Inhaler       1.2677     0.2950   4.297 1.73e-05 ***
```

Logistic Regression to compare Odds between groups

The estimated equation is:

$\ln(\frac{\hat{p}}{1-\hat{p}})$ = -2.3136 + 1.2677X

For people in the Nicotine group:

$\ln(\frac{\hat{p}}{1-\hat{p}})$ = -2.3136 + 1.2677 — (3). Here, 1.2677 indicates the increase in ln(Odds) that a person in Nicotine group of reduction in smoking

For people in the Placebo group:

$\ln(\frac{\hat{p}}{1-\hat{p}})$ = -2.3136 — (4)

ln(OR) is given by (3) - (4)

ln(OR) = 1.2677

OR = $e^{1.2677}$ = 3.55

That is, the effect of Nicotine inhaler on reduction in smoking is 3.55 times higher for a person in Nicotine group.

Next

** Reminders**

Quizzes 7 and 8 will be closing on Monday, 16 Jan 2023 at 3:00 pm

Lecture 13 (Module 12) - Non-parametric Methods - Tuesday, 17 Jan 2023

Lecture 14 (Revision) - Thursday, 19 Jan 2023