PM 592
Regression Analysis for Public Health Data Science

Week 8

Logistic Regression I

1

1

Logistic Regression I

Introduction to Binary Outcomes

The Odds Ratio

The Logit

Logistic Regression: Applied

Maximum Likelihood

2

Lecture Objectives

- Given a 2x2 table, compute the odds and odds ratio of an event.
- Given an equation for the logit, be able to provide the odds ratio associated with a change in X values.
- Given an equation for the logit, be able to provide the predicted probability associated with a set of X values.

3

1. Review

4

✓ Polynomial Terms, Splines

✓ Overfitting and adjusted R-squared

4

2. Binary Outcome Variables: Introduction

5

Type of Binary Variables

There are several reasons why we may want to examine a binary outcome variable:

- The outcome is **inherently binary** (male/female, disease/nondisease)
- The outcome is continuous but categorized into some **meaningful cutpoint** (scale score > a particular value, indicating depression vs. no depression)
- The outcome is continuous but **does not display a linear relationship** with the predictors
- The outcome is dichotomized on some **arbitrary cutpoint** (median, mean)

5

2. Binary Outcome Variables: Introduction

6

A Primer for Binary Outcomes

	Continuous	Binary
Population Mean/Proportion	μ	π
Sample Mean/Proportion	$\bar{Y} = \frac{\Sigma(Y_i)}{N}$	$p = \frac{\sum Y_i}{N}, Y_i \in (0,1)$
Confidence Limits	$\bar{Y} \pm z_{1-\frac{\alpha}{2}} \sigma_{\bar{Y}}$	$p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{N}}$
Test Statistic	$z = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}}$	$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{N}}}$

6

Example

The national prevalence of asthma in adolescents is 10%. Using the data from the CHS, determine whether there is evidence that the asthma percentage for adolescents in Southern California is different from the national average.

```
> z.test.prop(chs$asthma, .1)
# A tibble: 1 x 7
      p     n  ci.l  ci.u   pi0     z         pval
  <dbl> <int> <dbl> <dbl> <dbl> <dbl>        <dbl>
1 0.146  1200 0.126 0.166   0.1  5.34 0.0000000910
```

The proportion of participants in the CHS with asthma is 14.6% (95% CI = 12.6, 16.6), which is statistically significantly different from 10% (p<.001).

7

2 x 2 Contingency Tables

A **contingency table** is a way to summarize the relationship between two categorical variables.

A contingency table has **i rows** and **j columns**, with **n_{ij} counts** in each cell.

	Y=1	Y=2	Total
X=1	n_{11}	n_{12}	$n_{1\bullet}$
X=2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

8

There are three ways of estimating associations in 2x2 tables:

- Risk Difference
- Relative Risk
- Odds Ratio

Example

Is there an association between regular aspirin use and experiencing a heart attack (MI = myocardial infarction)?

	MI	No MI	Total
Aspirin	104	10933	11037
No Aspirin	189	10845	11034
Total	293	21778	22071

9

Risk Difference

The **risk difference** for group 2 vs. group 1 is the difference in proportions (π_2-π_1).

If the variables are **independent**, the difference is 0.

The **range** of the difference is -1 to +1.

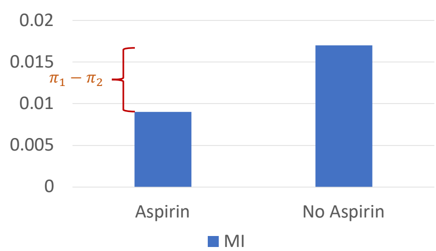
	MI	No MI	Total
Aspirin	104 0.9%	10933	11037
No Aspirin	189 1.7%	10845	11034
Total	293	21778	22071

P_1 = 104/11037 = 0.9%

P_2 = 189/11034 = 1.7%

The difference in proportions is P_2-P_1 = 0.77%

10



11

Relative Risk

The **relative risk** is the ratio of proportions $\frac{\pi_1}{\pi_2}$.

If the variables are **independent**, the ratio is 1.

The **range** of the relative risk is 0 to ∞.

	MI	No MI	Total
Aspirin	104 0.9%	10933	11037
No Aspirin	189 1.7%	10845	11034
Total	293	21778	22071

P1 = 104/11037 = 0.9%

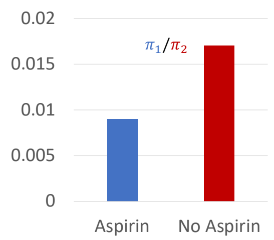
P2 = 189/11034 = 1.7%

Relative risk is P1/P2 = 0.55

12

The risk of heart attack among those taking aspirin is 0.55 times the risk among those who did not take aspirin.

(Taking aspirin lowers the risk of heart attack by 45%.)



13

Odds Ratio

The **odds ratio** is the ratio of odds $\frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)}$.

If the variables are **independent**, the ratio is 1.

The **range** of the odds ratio is 0 to ∞.

	MI	No MI	Total
Aspirin	A	B	A+B
No Aspirin	C	D	C+D
Total			

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{(1-p_1)p_2}$$

$$= \left(\frac{a}{a+b}\right)\left(\frac{a+b}{b}\right)\left(\frac{d}{c+d}\right)\left(\frac{c+d}{c}\right) = \frac{ad}{bc}$$

14

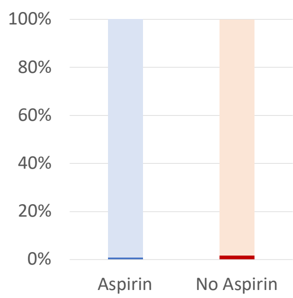
	MI	No MI	Total
Aspirin	104	10933	11037
No Aspirin	189	10845	11034
Total	293	21778	22071

$$OR = \frac{(104)(10845)}{(10933)(189)} = 0.55$$

15

2. Binary Outcome Variables: Introduction 16

The odds of heart attack among those taking aspirin is 0.55 times that of not taking aspirin.



16

2. Binary Outcome Variables: Introduction 17

Recap

- There are many reasons why we may want to examine dichotomous outcome variables.
- Some common measures of association for binary outcomes are the risk difference, risk ratio, and odds ratio.

17

2. Binary Outcome Variables: Introduction 18

Recap

➢Define the relationship between a binary X and Y variable by using the risk difference, risk ratio, and odds ratio.

18

3. The Odds Ratio

19

The odds ratio has some nice properties, so we will explore it more as a way to examine associations between an exposure and binary outcome.

19

3. The Odds Ratio

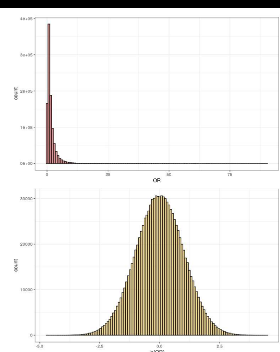
20

What is the range of odds ratio (OR) sizes if:

1) X reduces the odds of Y? (0, 1)
2) X increases the odds of Y? (1, ∞)

Therefore, the sampling distribution of the OR is highly skewed.

However, the distribution for the **log odds** (ln(OR)) is not!



20

3. The Odds Ratio

21

If…	Then…
OR = 1	ln(OR) = 0
OR < 1	ln(OR) < 0
OR > 1	ln(OR) > 0

We can link the ln(OR) to the OR.

An OR=1 means there is **no association.**

An OR < 1 means there is a **protective effect** of the exposure.

An OR > 1 means there is a **risk effect** of the exposure.

21

3. The Odds Ratio

22

The ln(OR) has an approximately symmetric sampling distribution.

Additionally, its **standard error** is easily calculated:

$$SE(\ln(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

So a 95% CI for the ln(OR) is straightforward:

$$\ln(OR) \pm 1.96\, SE(\ln(OR))$$

22

3. The Odds Ratio

23

For our example:

$$95\%\mathrm{CI}_{\ln(\mathrm{OR})} = \ln(0.55) \pm 1.96\sqrt{\frac{1}{189} + \frac{1}{10933} + \frac{1}{104} + \frac{1}{10845}} = (-0.84, -0.36)$$

	MI	No MI	Total
Aspirin	104	10933	11037
No Aspirin	189	10845	11034
Total	293	21778	22071

23

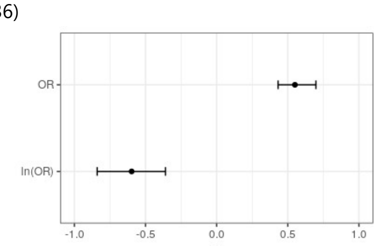
3. The Odds Ratio

24

Exponentiate the endpoints to get the 95% CI for the OR

95% CI [ln(OR)] = (-0.84, -0.36)

95% CI [OR] = (0.43, 0.70)



24

3. The Odds Ratio 25

Some additional **nice properties** of the OR:

- It will not change if we exchange rows and columns in our contingency table
- It will not change if the cell frequencies within a row or column are multiplied by a non-zero constant (multiplicative invariance property) (the CI will change, however)
- It is related to the Risk Ratio: $OR = \frac{p_1}{p_2}\left(\frac{1-p_2}{1-p_1}\right) = RR\left(\frac{1-p_2}{1-p_1}\right)$

25

3. The Odds Ratio 26

Implications

- When computing an OR for an association between two variables, it is not necessarily important which one is classified as the "exposure" and which is the "outcome".
- The odds ratio shouldn't change depending on sampling frequencies for exposure or outcome (the RR does)
- If p_1 and p_2 are close to 0 (i.e., rare disease assumption) the OR approximates the RR. This means that in case-control studies where the disease is rare, the OR can be used to approximate the RR.

26

3. The Odds Ratio 27

- We do not need to classify variables are response vs. predictor/explanatory

	MI	No MI
Aspirin	104	10933
No Aspirin	189	10845

	Aspirin	No Aspirin
MI	104	189
No MI	10933	10845

$$OR = \frac{(104)(10845)}{(10933)(189)} = 0.55$$

$$OR = \frac{(104)(10845)}{(189)(10933)} = 0.55$$

27

3. The Odds Ratio 28

- We may sample prospectively or retrospectively, and may use different sampling fractions for different categories

I know we can't have 0.5 people… this is just for illustrative purposes.

	MI	No MI
Aspirin	104	10933
No Aspirin	189	10845

	MI	No MI
Aspirin	104	10933
No Aspirin	189/2 = 94.5	10845/2 = 5422.5

$$OR = \frac{(104)(10845)}{(10933)(189)} = 0.55 \qquad OR = \frac{(104)(5422.5)}{(10933)(94.5)} = 0.55$$

28

3. The Odds Ratio 29

Test Yourself

Suppose the study on Aspirin and MI was instead a case-control study. We sampled all 293 subjects with MI and randomly sampled 293 subjects without MI. Assume the 293 subjects without MI had the same exposure probability that we previously calculated.

- ❑ Set up the 2x2 table.
- ❑ What is the odds of not using aspirin for those who had an MI?
- ❑ What is the odds of not using aspirin for those who did not have an MI?
- ❑ Calculate the OR of not using aspirin for those with MI vs. those without MI.

29

3. The Odds Ratio 30

❑ Set up the 2x2 table.

	MI	No MI	Total
Aspirin	104	147	251
No Aspirin	189	146	335
Total	293	293	586

	MI	No MI	Total
No Aspirin	189	146	335
Aspirin	104	147	251
Total	293	293	586

❑ What is the odds of not using aspirin for those who had an MI?

Odds = (189/293)/(104/293) = 1.82

❑ What is the odds of not using aspirin for those who did not have an MI?

Odds = (146/293)/(147/293) = 0.99

❑ Calculate the OR of not using aspirin for those with MI vs. those without MI.

OR = 1.82/0.99 = 1.83 Note: this is (189*147)/(104*146)

30

3. The Odds Ratio

Also Note:

	MI	No MI	Total
Aspirin	104	147	251
No Aspirin	189	146	335
Total	293	293	586

	MI	No MI	Total
No Aspirin	189	146	335
Aspirin	104	147	251
Total	293	293	586

The OR of MI for aspirin compared to no aspirin is (104*146)/(189*147) = 0.55

The OR of MI for no aspirin compared to aspirin is (189*147)/(104*146) = 1.83

31

3. The Odds Ratio

Is the relationship **statistically significant**?

The Chi-Square test statistic is most commonly used, and there are two varieties:

- **Pearson Chi-Square Statistic**
  - Compares the observed cell frequencies to the frequencies that would be expected if X and Y were not related.
- **Likelihood Ratio Chi-Square Statistic**
  - Based on the multinomial likelihood.

32

3. The Odds Ratio

Pearson Chi-Square Statistic

Here, n_{ij} is the **observed** frequencies in each cell, and μ_{ij} is the expected cell frequencies.

$$Pearson\ \chi^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

If the null hypothesis is true, we **expect** each cell value μ_{ij} to be equal to the product of its row and column marginal values.

$$\mu_{ij} = n\pi_{i}.\pi._{j}$$

Under H_0: $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ for all i, j

	Y=1	Y=2	Total
X=1	n_{11}	n_{12}	$n_{1\bullet}$
X=2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

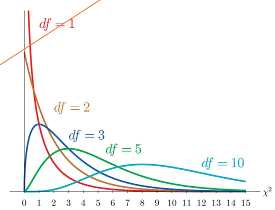
33

3. The Odds Ratio

Under the null hypothesis (independence), when expected frequencies are large, both statistics follow a Chi-square distribution.

The chi-square distribution has one parameter: df (degrees of freedom).

For this distribution, the mean = df, var = 2df



If expected frequencies are not large (<5 in any cell), then different methods such as the Exact test may need to be used instead.

34

3. The Odds Ratio

How do we determine the degrees of freedom?

The df is the difference between the number of parameters under H_A and H_0.

Under H_0, the marginal probabilities determine the cell probabilities. There are I-1 and J-1 nonredundant marginal probabilities.

Under H_A, the cell probabilities are unrestricted; they must simply sum to 1. There are IJ-1 nonredundant cell probabilities.

df = [IJ-1] – [(I-1) + (J-1)] = (I-1)(J-1)

35

3. The Odds Ratio

Example

$$Pearson\ \chi^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$
$$= \frac{(189 - 146.5)^2}{146.5} + \frac{(10845 - 10887.5)^2}{10887.5} + \frac{(104 - 146.5)^2}{146.5} + \frac{(10933 - 10890.5)^2}{10890.5}$$
$$= 25.0$$

	MI	No MI	Total
Aspirin	104	10933	11037
No Aspirin	189	10845	11034
Total	293	21778	22071

```
> tibble(
+   aspirin = c(1, 1, 0, 0),
+   mi = c(1, 0, 1, 0),
+   freq = c(104, 10933, 189, 10845)
+ ) %>%
+   xtabs(freq ~ aspirin + mi, data = .) %>%
+   chisq.test(correct = F)

        Pearson's Chi-squared test

data:  .
X-squared = 25.014, df = 1, p-value = 5.692e-07
```

We performed a chi-square test for the relationship between aspirin use and heart attack. We found that aspirin use is statistically significantly associated with heart attack (χ^2_1=25.0, p<.001).

36

3. The Odds Ratio 37

Example

Is there a relationship between gender and presence of asthma for adolescents in Southern California?

```
> chs %>%
+   with(.,
+        table(asthma, male)) %>%
+   chisq.test()

        Pearson's Chi-squared test with Yates' continuity correction

data:  .
X-squared = 5.9088, df = 1, p-value = 0.01507
```

By default, chisq.test() uses the Yates' continuity correction. In 2x2 tables, the continuous chi-square distribution is used, which tends to be more liberal, especially when expected cell counts are low (<40). The Yates' correction makes the test more conservative, but some argue too conservative.

There is a statistically significant relationship between asthma and gender (χ_1^2=5.9, p=.015).

37

3. The Odds Ratio 38

Recap

- The properties of the odds ratio make it a versatile measure of effect size.
- Pearson's χ^2 test is the most common measure of significance for contingency tables (and odds ratios).

38

3. The Odds Ratio 39

Recap

- ➢Given a 2x2 table, compute the odds ratio for association between X and Y.
- ➢Explain how the odds ratio is affected by changes to the 2x2 table.
- ➢Calculate Pearson's χ^2 test for the relationship between two binary variables.

39

4. The Logit Link 40

The Generalized Linear Model

Up to this point in the course we have focused on the linear model *lm()*, which has:

- A random component
- A systematic component
- A link function

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

40

4. The Logit Link 41

Random Component

Recall one of the assumptions of linear regression: normality

We assume our observed Y follow some random distribution, with E(Y) = $\hat{Y}$ = μ.

For each observation i (i = 1, …, n) the Y_i are IID.

This means the Y are uncorrelated and all follow the same probability distribution.

For linear regression, the **random component** (probability distribution) is $Y_i \sim \text{Normal}(\mu, \sigma^2)$

41

4. The Logit Link 42

Systematic Component

This component specifies the form for the independent variables.

This is specified as a linear function: $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$

η_i is called the linear predictor

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

This means:

Flexibility – We can easily include x variables that are a combination of other variables (e.g., interactions: $x_3 = x_1 * x_2$) or transformations (e.g., $x_2 = x_1^2$)

Interpretability – The linear predictor has a range of –∞ to +∞

42

4. The Logit Link 43

In OLS regression this makes sense.

We have dealt with variables where:

1) The association between predictor and outcome is linear and
2) The prediction can take on values between $-\infty$ to $+\infty$

So, we can **directly use the predictions** given by the model.

43

4. The Logit Link 44

Are there any special cases where these assumptions would be violated?

❑ Where it wouldn't make sense for the predicted value to have a range of $-\infty$ to $+\infty$?

❑ Where our outcome may not follow a normal distribution?

44

4. The Logit Link 45

Are there any special cases where these assumptions would be violated?

❑ Where it wouldn't make sense for the predicted value to have a range of $-\infty$ to $+\infty$?

When the outcome is bounded – e.g. probabilities are bounded by [0,1], count variables have a lower limit of 0, etc.

❑ Where our outcome may not follow a normal distribution?

Outcomes may follow several other distributions (binomial, multinomial, Poisson, etc.)

45

Link Function

The **link function** allows us to use a **linear systematic component** to predict an outcome that isn't necessarily normally distributed.

We have to ask: what **transformation of Y** would express the systematic component as a linear function of the covariates?

The link function is given as g(μ).

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots = \sum_{j=1}^{p} \beta_j x_{ij}$$

It is a **transformation** we make on the expected value so it conforms to a different distribution.

46

Let's look at something we're already familiar with.

Linear regression

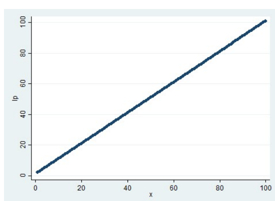
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

We don't need to perform any transformation on this outcome.

We can use the predicted Y values as-is.

So, we use a very simple link.

The **identity link** is g(μ) = μ.



47

Linear models use the **identity link.**

Models that are **nonlinear** use a **different link**.

Intercept. For any GLM, the intercept β_0 is the expected value of the link function $g(\mu)$ when all values of the independent variables x_k are zero.

Slope. For any GLM, the slope parameters $(\beta_1 - \beta_k)$ are interpreted as the change in the link function, g(μ), per unit of x_k.

This is essentially the linear regression we are accustomed to!

The only difference is now we are thinking about interpretation with regard to the link function instead.

48

4. The Logit Link

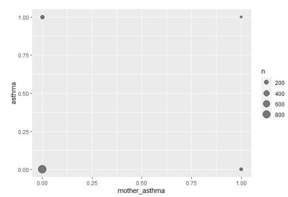
49

Example

Suppose we want to examine the heritability of asthma, so we examine a model to predict child's asthma from the mother and father.

This poses a bit of a problem, as the outcome (and predictor) are dichotomous.

Why is this such a problem?



49

4. The Logit Link

50

Problem 1

We can get a result, but there is no guarantee that it makes sense.

This is because the range of linear regression is (-∞, ∞). Since the outcome is a probability, anything outside of [0, 1] is nonsensical.

```
> lm(asthma ~ mother_asthma + father_asthma, data = chs)
%>% summary()

Call:
lm(formula = asthma ~ mother_asthma + father_asthma,
data = chs)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5007 -0.1084 -0.1084 -0.1084  0.8916

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.10843    0.01160   9.347  < 2e-16 ***
mother_asthma  0.20079    0.03527   5.693 1.62e-08 ***
father_asthma  0.19150    0.03818   5.015 6.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.3429 on 1049 degrees of
freedom
  (148 observations deleted due to missingness)
Multiple R-squared:  0.05158,	Adjusted R-squared:
0.04978
F-statistic: 28.53 on 2 and 1049 DF,  p-value: 8.63e-13
```

50

4. The Logit Link

51

In this example we see that the predicted probability of asthma for someone with no parent asthma, no wheeze, hayfever, smoke, or allergy is -2.7%.

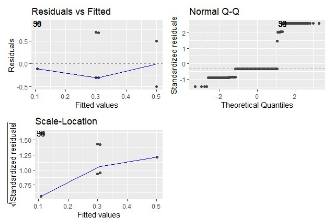
```
> lm(asthma ~ mother_asthma + father_asthma + wheeze + hayfever + smoke + allergy, data = chs) %>% summary()

Call:
lm(formula = asthma ~ mother_asthma + father_asthma + wheeze +
    hayfever + smoke + allergy, data = chs)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72093 -0.09834  0.02751  0.02751  1.03009

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.027512   0.013932  -1.975  0.04862 *
mother_asthma  0.105672   0.033702   3.135  0.00177 **
father_asthma  0.110335   0.035205   3.134  0.00178 **
wheeze         0.332960   0.022410  14.857  < 2e-16 ***
hayfever       0.073627   0.028170   2.614  0.00911 **
smoke         -0.002575   0.026666  -0.097  0.92308
allergy        0.125852   0.024109   5.220 2.23e-07 ***

Residual standard error: 0.293 on 882 degrees of freedom
  (311 observations deleted due to missingness)
Multiple R-squared:  0.3344,	Adjusted R-squared:  0.3299
F-statistic: 73.86 on 6 and 882 DF,  p-value: < 2.2e-16
```

51

Problem 2

The residuals are horrible!

Performing an ordinary linear regression on a binary outcome will ensure the residuals are NOT normally distributed (though this is less of a problem with large sample size).



52

Solution?

The logit is one function that will yield predictions constrained between 0 and 1.

$$\text{logit}(Y_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

The "logit" is another way of saying the "log odds."

In extreme cases:

Note: we could also use μ_i instead of π_i.

Y = 0: $\text{logit}(0) = \ln\left(\frac{0}{1-0}\right) = -\infty$

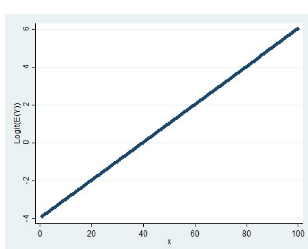
Y = 1: $\text{logit}(1) = \ln\left(\frac{1}{1-1}\right) = +\infty$

53

Suppose we use the linear systematic component to model the logit (instead of Y directly).

Using the logit link:

$$\text{logit}(\pi_i) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$$



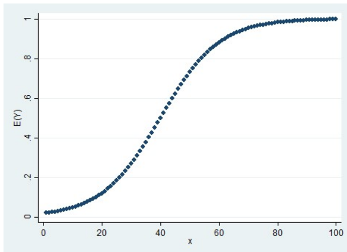
This model is "linear in the logit."

54

4. The Logit Link 55

What happens when the equation is transformed in terms of Y?

$$\mu_i =$$
$$\pi_i =$$
$$\frac{e^{\beta_0+\beta_1 x_1+..+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+..+\beta_k x_k}}$$

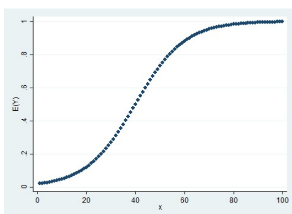


55

4. The Logit Link 56

Some properties:

- The predicted probability asymptotically approaches 0 and 1.
- A change in x has less impact on E(Y) at the extremes.
- A change in x has more impact on the expected probability in the mid-range of π.



56

4. The Logit Link 57

Interpreting Parameters

$$\text{logit}(\hat{\pi}) = \beta_0 + \beta_1 X_1$$

The intercept is the log odds that Y=1 when X=0.

When X=0, $\text{logit}(\hat{\pi}) = \beta_0$.

We can convert this to a "baseline probability":

$$P(Y=1|X=0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

57

4. The Logit Link

58

Interpreting Parameters

$$\text{logit}(\hat{\pi}) = \beta_0 + \beta_1 X_1$$

The slope β_1 is the change in log odds (logit) for a 1-unit increase in X.

58

4. The Logit Link

59

What if we looked at this in terms of predicted probability, and not in terms of predicted logit?

The predicted probability…

$$\pi_{Y=1} = P[Y = 1|x] = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

$$\pi_{Y=0} = P[Y = 0|x] = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

59

4. The Logit Link

60

We can use this information to get the **odds** of event.

$$\pi_{Y=1}(1) = P[Y = 1|X = 1] = \frac{\exp(\beta_0 + \beta_1(1))}{1 + \exp(\beta_0 + \beta_1(1))}$$

$$\pi_{Y=0}(1) = P[Y = 0|X = 1] = \frac{1}{1 + \exp(\beta_0 + \beta_1(1))}$$

$$\frac{\pi_{Y=1}(1)}{\pi_{Y=0}(1)} = ODDS[Y = 1|X = 1] = \exp(\beta_0 + \beta_1)$$

60

4. The Logit Link

61

We can use this information to get the **odds** of event.

$$\pi_{Y=1}(0) = P[Y = 1|X = 0] = \frac{\exp(\beta_0 + \beta_1(0))}{1 + \exp(\beta_0 + \beta_1(0))}$$

$$\pi_{Y=0}(0) = P[Y = 0|X = 0] = \frac{1}{1 + \exp(\beta_0 + \beta_1(0))}$$

$$\frac{\pi_{Y=1}(0)}{\pi_{Y=0}(0)} = ODDS[Y = 1|X = 0] = \exp(\beta_0)$$

61

4. The Logit Link

62

We can then take the **odds ratio (OR)** as:

(odds of event | x=1) / (odds of event | x=0)

$$OR = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

This means if we exponentiate the beta coefficients in our model output, we directly obtain the odds ratio for a 1-unit increase in X!

62

4. The Logit Link

63

A brief FAQ:

- Why are we examining the odds ratio?

  Because it has some nice properties and allows us to directly interpret the parameter estimates from the model output.

- Couldn't we have made sense of the β coefficients some other way?

  The odds ratio is the easiest way to directly interpret these coefficients.

- What is the interpretation of the odds ratio?

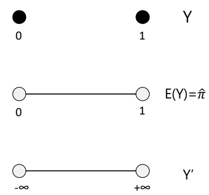
  The (multiplicative) change in odds of outcome associated with a one-unit change in x.

63

4. The Logit Link 64

Summary

- When an outcome Y is binary, we can only *observe* two possible values: 0 and 1
- The expected value of Y ("π", or "μ"), also known as the expected probability of Y, falls between (0, 1).
- To use linear regression, we model Y′ – a transformation of Y – that is unbounded (e.g., the logit).



64

4. The Logit Link 65

Summary

The logit link is $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = logit\big(\pi(x)\big) = \beta_0 + \beta_1 x_1 + \cdots$

The logit can range from -∞ to +∞.

We can transform the logit to get a predicted probability of outcome.

This predicted probability is: $\pi(x) = \frac{\exp(\beta_0+\beta_1 x_1+\cdots)}{1+\exp(\beta_0+\beta_1 x_1+\cdots)}$

The predicted probability ranges from 0 to 1.

65

4. The Logit Link 66

Recap

- The logit link allows us to model binary outcomes in a way that guarantees a predicted outcome in (0, 1) and treats the residuals correctly.
- There are other possible transformations instead of the logit link, but the logit is popular and quite interpretable.

66

4. The Logit Link

67

Recap

- Explain the 3 components of a generalized linear model
- Describe the logit transformation
- Interpret slope parameters from a logit model as odds ratios
- Compute a predicted probability, given a logit model

67

5. Logistic Regression Applied

68

Example

Is the prevalence of asthma related to parents' asthma?

(Create a variable that indicates whether either parent has asthma.)

Approach 1: Contingency Table

```
> chs %>%
+   with(., table(asthma, asthma_parent))
       asthma_parent
asthma   0   1
     0 773 137
     1  94  62

> (62*773)/(137*94)
[1] 3.721541

> chs %>%
+   with(., table(asthma, asthma_parent)) %>%
+   chisq.test(correct = F)

        Pearson's Chi-squared test

X-squared = 53.462, df = 1, p-value = 2.636e-13
```

The contingency table.

The odds ratio AD/BC. Children are 3.72 times as likely to have asthma if they have a parent who also has asthma.

This association is statistically significant (χ^2_1=53.4, p<.001).

68

5. Logistic Regression Applied

69

Approach 2: Logistic Regression

```
> glm(asthma ~ asthma_parent, data = chs, family = binomial) %>% summary()

Call:
glm(formula = asthma ~ asthma_parent, family = binomial, data = chs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8641  -0.4791  -0.4791  -0.4791   2.1080

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.1070     0.1092 -19.289  < 2e-16 ***
asthma_parent   1.3141     0.1880   6.989 2.78e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 887.57  on 1065  degrees of freedom
Residual deviance: 842.00  on 1064  degrees of freedom
  (134 observations deleted due to missingness)
AIC: 846

Number of Fisher Scoring iterations: 4
```

$logit(\hat{\pi}) = -2.71 + 1.31X$

The odds ratio e^{β_1} = exp(1.314) = 3.72. Children are 3.72 times as likely to have asthma if they have a parent who also has asthma.

This association is statistically significant (z=6.99, p<.001)

69

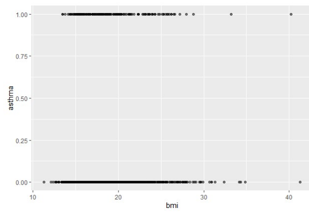
5. Logistic Regression Applied

70

Example

Now, let's look at a continuous predictor. We can no longer examine a contingency table because of this. Is asthma related to BMI?

Here we see observed values which are not particularly helpful (maybe more individuals with Y=0 at lower BMI?).



70

5. Logistic Regression Applied

71

We may want to assess linearity first, but we'll look at that more next lecture. In the meantime, let's jump into the output:

```
> glm(asthma ~ bmi, data = chs, family = binomial) %>% summary()

Call:
glm(formula = asthma ~ bmi, family = binomial, data = chs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9458  -0.5729  -0.5382  -0.5149   2.0767

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.73943    0.40592  -6.749 1.49e-11 ***
bmi          0.05251    0.02085   2.518   0.0118 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 907.68  on 1084  degrees of freedom
Residual deviance: 901.65  on 1083  degrees of freedom
  (115 observations deleted due to missingness)
AIC: 905.65
```

$logit(\hat{\pi}) = -2.74 + 0.05X$

The odds ratio e^{β_1} = exp(0.0525) = 1.05. A one-unit increase in BMI is associated with 1.05 times the odds of asthma.

This association is statistically significant (z=2.52, p=.012)

71

5. Logistic Regression Applied

72

General principle for interpreting slope coefficients in logistic regression

1. Find the logit (log odds) for each level of x

   ln[odds(Y=1)| x] = $\beta_0+\beta_1 x_1$

   ln[odds(asthma=1)|BMI=1] = -2.74+0.05

   ln[odds(asthma=1)|BMI=0] = -2.74

2. Take the difference between the logits

   (-2.74+0.05) – (-2.74) = 0.05

3. Exponentiate

   exp(0.05) = 1.05 = OR (of asthma, for a 1-unit increase in BMI).

72

5. Logistic Regression Applied

73

What if we wanted to examine the effect on asthma for a 10-unit increase in BMI?

1. Find the logit for each level of x

   $\ln[\text{odds}(Y=1)|\ x] = \beta_0+\beta_1 x_1$

   $\ln[\text{odds}(\text{asthma}=1)|\text{BMI}=10] = -2.74+0.052(10)$

   $\ln[\text{odds}(\text{asthma}=1)|\text{BMI}=0] = -2.74$

2. Take the difference between the logits

   $(-2.74+0.52) - (-2.74) = 0.52$

3. Exponentiate

   exp(0.52) = 1.69 = OR (of asthma, for a 10-unit increase in BMI).

73

5. Logistic Regression Applied

74

True or False?

- ❑ A 10-unit increase in x is associated with a 10β increase in the logit.
- ❑ A 10-unit increase in x is associated with a 10exp(β) change in the OR.
- ❑ The p-value of the effect of x will be larger when examining a 10-unit change in x.
- ❑ The OR for a 10-unit increase in x is exp(10β).
- ❑ The OR for a 10-unit increase in x is the same regardless of the specific x values being compared.

74

5. Logistic Regression Applied

75

True or False?

- ❑ A 10-unit increase in x is associated with a 10β increase in the logit.
- ❑ A 10-unit increase in x is associated with a 10exp(β) change in the OR.
  If the OR for a 1-unit increase in x is exp(β), then the OR for a 10-unit increase in x is exp(10β).
- ❑ The p-value of the effect of x will be larger when examining a 10-unit change in x.
  Scaling a variable by a constant does not change the significance of association between x and y.
- ❑ The OR for a 10-unit increase in x is exp(10β).
- ❑ The OR for a 10-unit increase in x is the same regardless of the specific x values being compared.
  This is one of the benefits of using the OR as a measure of effect.

75

5. Logistic Regression Applied

Recap

- The interpretation of slope coefficients in logistic regression is similar to in linear regression.
- When comparing two observations, take the difference in logits first and *then* exponentiate to get the multiplicative difference in odds.

76

5. Logistic Regression Applied

Recap

- ➢Interpret the output from a logistic regression model
- ➢Use slope parameters from logistic regression models to compare odds

77

6. Maximum Likelihood Estimation

Note some additional differences in the output:

```
> glm(asthma ~ bmi, data = chs, family = binomial) %>% summary()

Call:
glm(formula = asthma ~ bmi, family = binomial, data = chs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9458  -0.5729  -0.5382  -0.5149   2.0767

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.73943    0.40592  -6.749 1.49e-11 ***
bmi          0.05251    0.02085   2.518   0.0118 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 907.68  on 1084  degrees of freedom
Residual deviance: 901.65  on 1083  degrees of freedom
  (115 observations deleted due to missingness)
AIC: 905.65
```

The residuals are now called the “deviance residuals.” In logistic regression the conventional residual doesn't exist, so the deviance residuals are used to assess which observations contribute most to the lack of fit.

The deviance residual measures each observation's contribution to the overall deviance.

The deviance is a measure of lack of fit. We see that the model with BMI in it <u>reduces</u> the value of the deviance (“residual deviance”) compared to the null model (with no predictors).

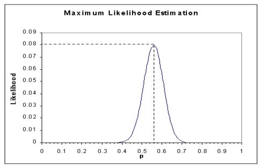
78

6. Maximum Likelihood Estimation

Because we don’t have the traditional sums of squares, logistic regression relies on **maximum likelihood estimation**.

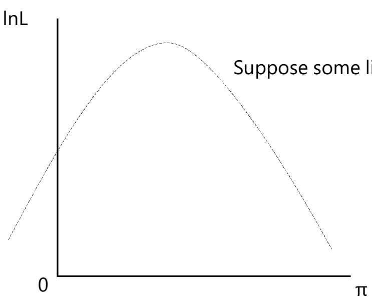
The basic idea:

- We have some parameter(s) we want to fit (e.g., β_0, β_1)
- We create a function that links the values of these parameters to the likelihood of our observed data
- We examine the possible values of our parameters find those that maximize the likelihood function



79

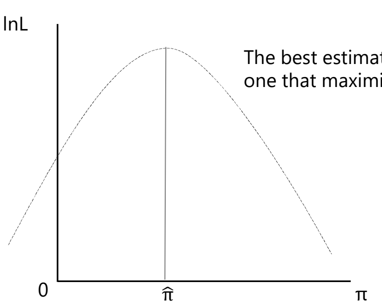
6. Maximum Likelihood Estimation



Suppose some likelihood function

80

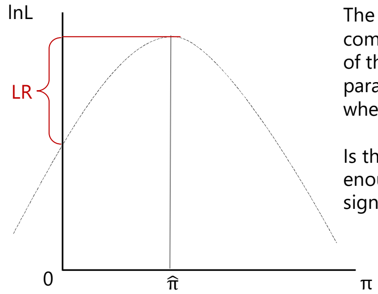
6. Maximum Likelihood Estimation



The best estimate of our parameter is the one that maximizes the likelihood

81

6. Maximum Likelihood Estimation



The likelihood ratio test compares the log likelihoods (LL) of the model where the parameter is freely estimated, vs. when it is set to zero (or π_0).

Is the difference in LL large enough to be statistically significant?

82

6. Maximum Likelihood Estimation

To perform the **likelihood ratio** test, we need the value of the LL under the null and alternative hypotheses.

The test statistic (for one parameter) is:

$$G = -2 \ln(L_0/L_1)$$
$$= -2 [\ln(L_0) - \ln(L_1)]$$
$$\sim \chi^2_{(1)}$$

The likelihood under H_0

The likelihood at an alternate value, typically evaluated at the MLE

83

6. Maximum Likelihood Estimation

R provides the value of the **deviance**

$$D = -2(LL)$$

If we plug in the deviance, the likelihood ratio test becomes:

$$G = D_0 - D_1 \sim \chi^2_{(1)}$$

```
> glm(asthma ~ bmi, data = chs, family = binomial) %>% anova(test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: asthma

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                  1084     907.68
bmi   1   6.0378      1083     901.65    0.014 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is analogous to the Extra Sums of Squares test, but for logistic regression – it tells us if a full model is statistically better than the reduced model.

This p-value is for the Likelihood Ratio Test. The p-value associated with BMI in the output is for the Wald test. These tests are slightly different but in general agree quite well with each other.

84

Recap

- Linear models use the "residual sums of squares" as a measure of how poorly a model fits, while generalized linear models use the "deviance."
- As with the *extra sums of squares test*, with generalized linear models we can use the *likelihood ratio test* to compare the fit of two nested models.

85

Recap

➢Explain what the null and residual deviance are measuring in glm() output

➢Compute the likelihood ratio test to compare two nested models

86

- **Logistic Regression** uses a linear regression equation, but has the logit transformation applied to it. We can still model in the same way as before re: examining confounding/interaction, polynomial terms, dummy coding, etc.
- **The odds ratio.** The main outcome in logistic regression is the odds ratio: the multiplicative change in odds for a one-unit increase in X. The odds ratio for a 1-unit increase in X is computed as e^{β_1}.
- **Likelihood.** The likelihood is a measure of how well a logistic regression model fits the data. This leads to a couple differences:
  - **Residuals** are computed differently and have a slightly different meaning
  - **The R^2** cannot be computed the same way, but there is a pseudo R^2
  - **Likelihood Ratio Test** compares the fit of a full model to a reduced model

87

7. Recap 88

Additional Reading

- More on the Yates' correction
  https://www.jstor.org/stable/2285661
- More on Fisher's exact test when cell sizes are really small
  https://www.sciencedirect.com/topics/medicine-and-dentistry/fisher-exact-test
- Discussion on pseudo R-squared values
  https://statisticalhorizons.com/r2logistic

88

7. Recap 89

Packages and Functions

- `glm()`

89