Logistic Regression

Analysis of Categorical outcome data

Greg W Fegan

Tuesday, February 17, 2015

Logistic Regression

Objectives

1. Use a logistic model to compare the log odds of disease (or any binary outcome variable) in two groups
2. Use a logistic model to compare the odds of an outcome for a categorical exposure with 2 or more levels and to estimate crude odds ratios associated with each level.
3. Understand statistical tests of the null hypothesis - there is no association between the exposure and outcome
    1. using the Wald test
    2. using the Likelihood Ratio Test
4. Models with more than one explanatory variable
5. Interaction/Effect Modification using logistic regression models

Introduction

Definition

Logistic regression - a regression modelling technique for producing Odds Ratios (ORs); models the log odds of a binary "outcome"

Examples

1. Effect of T.B infection on death in HIV positive patients crude(unadjusted) OR; 95% CI and hypothesis tests
2. Effect of mothers education on childs' measles immunisation status
3. Effect of ethnicity on risk of death from breast cancer
4. Effect of gender on being a high wage earner

Why model the log odds of disease?

The reason for modelling the log odds rather than risk or odds is

- that the log odds can take any value, positive or negative,
- whereas risks are constrained to lie between 0 and 1.
- When using statistical models it is easier to model a quantity which is unconstrained than one which is constrained.

This avoids the possibility of predicting impossible values (like risks which are negative or greater than 1) from the model.

Modelling log odds is referred to as **logistic regression**, and the models are referred to as **logistic models**.

A reminder of Odds and Odds Ratios (OR).

$$Odds = \frac{Number\ with\ the\ disease\ (D)}{Number\ without\ the\ disease\ (H)}$$

$$Odds\ Ratio(OR) = \frac{Odds\ in\ exposed\ group\ (\frac{D_1}{H_1})}{Odds\ in\ unexposed\ group\ (\frac{D_0}{H_0})}$$

Odds in exposed group = (Odds in unexposed group) x (Odds ratio)

Log (odds in exposed group) = Log (odds in unexposed) + Log (odds ratio)

Log odds = Baseline + Exposure.

A logistic model with a single binary exposure variable

Estimating the odds, log odds and the odds ratio "by hand"

OR = (odds in exposed group)/(odds in baseline)

Therefore: odds in exposed = (odds in baseline) x OR

log (odds in exposed) = log (odds in baseline) + log OR

Microfilariae Infection	Savannah	Forest	Total
Negative	267	213	480
Positive	281	541	822
Total	548	754	1302

Using the Microfilariae by Area data

Substituting the areas in our example gives:

odds in forest = (odds in savannah) x OR (forest compared to savannah)

log (odds in forest) = log (odds in savannah) + log OR(forest vs savannah)

We can write the second of these two expressions as the logistic regression model: log odds = Baseline + Area

where Baseline = log (odds in savannah), Area = log OR for individuals in the forest and 0 individuals in the savannah.

EXERCISE 1

Calculate the prevalence, odds and log odds of Microfilariae infection in the forest and savannah areas

Microfilariae Infection	Savannah	Forest	Total
Positive	267	213	480
Negative	281	541	822
Total	548	754	1302

	savannah	forest	overall
Risk/prevalence			63.1
Odds			1.712
Log odds			0.538

What are the odds ratio and log odds ratio?

Exercise 1 solution

	savannah	forest	overall
Risk/prevalence(%)	51.3	71.8	63.1
Odds	1.052	2.540	1.712
Log odds	0.052	0.932	0.538

The Odds ratio $= 2.41$ Whilst the log odds ratio? $= 0.881$

area	odds	log odds of disease
0=savannah	1.052	0.051
1=forest	$1.052 \times 2.41 = 2.536$	$0.051 + 0.881 = 0.932$

Summarise results in a model

log odds = Baseline + Area

where Baseline = log (odds in savannah) = $(0.051 + 0.881 \times 0) = 0.051$

Area = log odds for individuals in the forest and 0 individuals in the savannah

= $(0.051 + 0.881 \times 1) = 0.932$

R Code and Output for Logistic Model

```
onch <-read.csv("onchall.csv") # Read in CSV data
m1 <- glm(mf~area, data=onch,family=binomial)# Run model
summary(m1) # Show model
##
## Call:
## glm(formula = mf ~ area, family = binomial, data = onch)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5900  -1.1992   0.8148   0.8148   1.1558
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05111    0.08546   0.598     0.55
## area         0.88102    0.11767   7.487 7.05e-14 ***
## ---
```

Getting the ORs and Confidence intervals using 

```
exp(coef(m1))      # transform the coeffs into ORs #
## (Intercept)         area
##    1.052434     2.413363
exp(confint(m1))    # and show their CIs
## Waiting for profiling to be done...
##                   2.5 %   97.5 %
## (Intercept) 0.8901384 1.244620
## area        1.9176644 3.042055
```

NOTE: the output tells of the ratio of odds between the levels of the factor, and does NOT tell us how frequent the disease is - ie. does not tell us the prevalence or odds of the infection in either level.

Testing for associations

1. Wald Test
2. Likelihood Ratio Test

Let's start with the Wald test...

Wald test (1)

- The null hypothesis for this test is that the true parameter (logOR) value is 0.
- The test statistic (z) is obtained by dividing the parameter estimate by its SE and comparing it with a Standard Normal distribution.
- The Wald test for area assesses the H_0 that the true log OR=0 (i.e. that the true OR is 1) versus the alternative that the true log OR is not 0.

Wald test (2)

```
summary(m1)
##
## Call:
## glm(formula = mf ~ area, family = binomial, data = onch)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5900  -1.1992   0.8148   0.8148   1.1558
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05111    0.08546   0.598     0.55
## area         0.88102    0.11767   7.487 7.05e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
```

Wald test (3)

The Wald test for the association between microfilarial infection and area is given by: $z = \log(OR)/SE(\log OR) = 0.881/0.118 = 7.487$

The corresponding p-value is small ($p << 0.001$), indicating strong evidence against the null hypothesis of no association between microfilarial infection and area.

Practical 1

Use the dataset ond15p.csv

- We wish to investigate the association between microfilarial infection and optic nerve disease Variables of interest include
  1. ond (optic nerve disease)
  2. mfpos (microfilarial positve/negative)
  3. sex (male/female)
- Tabulate ond and mfpos - calculate the chi test
- Compute the odds ratio of optic nerve disease in microfilarial positive patients
- Comment on the wald test and the 95% CI
- Compute the odds ratio of optic nerve disease in females
- Comment on the odds ratio, wald test and the 95% CI

Comparison of more than two levels in a group

Consider

	Ages[1]				
Microfil. Inf.	5-9	10-19	20-39	40+	Total
Negative	156	119	125	80	480
Positive	46	99	299	378	822
Total	202	218	424	458	1302

Age Group Variable values coded as 0,1,2,3 respectively

[1] in years

Exercise 2

Calculate the missing values in the table below

	5 -9	10 – 19	20 -39	>=40
Odds	0.29	0.83	2.392	
Odds ratio	1.00	2.82		16.03
Log odds	-1.221		0.872	
Log OR	0	1.037		

NB We have used the first age group (ie 5-9 years) as the Reference group –> OR=1

Solutions to Exercise 2

Microfil. Inf.	5-9	10-19	20-39	40+	Total
Negative	156	119	125	80	480
Positive	46	99	299	378	822

	5 -9	10 – 19	20 -39	>=40
Odds	0.29	0.83	2.392	4.725
Odds ratio	1.00	2.82	8.11	16.03
Log odds	-1.221	-0.184	0.872	1.553
Log OR	0	1.037	2.093	2.774

Note that you can calculate the log odds ratio for age group 10-19 compared to age group 5-9 either as: $1.037 = \log(2.82) =$ (the log OR), or $1.037 =$ (-0.184) - (-1.221) (the difference in the log odds).

Comparison of more than two levels using 

The association between age group and mf infection using the logistic model:

```
log odds = Baseline + Agegrp
```

- Baseline is the log odds in the lowest age group (age group 5-9)
- Agegrp is the logOR for each level of age group relative to age group 5-9 (three non-zero logORs)

```
onch <-read.csv("onchall.csv") # Read in CSV data
m2 <- glm(mf ~ as.factor(agegrp), data=onch, family=binomial)
```

NB We use the function **as.factor** as we are not using the values of the age groups ie 0-3 as these are categorical indicators called *factors* in R

Prediction of being infected with MF according to age

```
m2
##
## Call:  glm(formula = mf ~ as.factor(agegrp), family = binom
##
## Coefficients:
##         (Intercept)  as.factor(agegrp)1  as.factor(agegrp)2
##              -1.221                1.037                2.093
## as.factor(agegrp)3
##               2.774
##
## Degrees of Freedom: 1301 Total (i.e. Null);  1298 Residual
## Null Deviance:       1714
## Residual Deviance: 1456  AIC: 1464
```

Predictions as ORs with Confidence intervals

```
exp(coef(m2))      # transform the coeffs into ORs #
##         (Intercept) as.factor(agegrp)1 as.factor(agegrp)2
##           0.2948718           2.8213372           8.1120000
## as.factor(agegrp)3
##          16.0239130
exp(confint(m2))    # and show their CIs
## Waiting for profiling to be done...
##                          2.5 %     97.5 %
## (Intercept)          0.2099851  0.4059964
## as.factor(agegrp)1   1.8566452  4.3348741
## as.factor(agegrp)2   5.5353487 12.0770820
## as.factor(agegrp)3 10.7442422 24.3134387
```

Note separate Wald tests

```
summary(m2)
##
## Call:
## glm(formula = mf ~ as.factor(agegrp), family = binomial, da
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8681  -0.7189   0.6196   0.8358   1.7203
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.2212     0.1678  -7.279 3.37e-13 ***
## as.factor(agegrp)1   1.0372     0.2160   4.802 1.57e-06 ***
## as.factor(agegrp)2   2.0933     0.1987  10.534  < 2e-16 ***
## as.factor(agegrp)3   2.7741     0.2081  13.332  < 2e-16 ***
## ---
```

Testing for association (2)

1. Wald Test
2. **Likelihood Ratio Test**

The Likelihood ratio test (1)

- H_0 - the model without the term for age group is adequate, and we do not need the extra term for age group in our model.
  - odds of microfilarial infection are the same in all the age groups ie OR_i=1 (the logOR=0).
- The Likelihood Ratio Test (LRT) is based on the Likelihood Ratio Statistic (LRS): LRS=2(L_1-L_0); where
  - L_1 is the maximised log likelihood under the alternative hypothesis, ie different odds of disease in each group
  - L_0 is the log likelihood under the null hypothesis ie one with no age effect included

Performing a likelihood ratio test

1. **Obtain the value of** L_1 by fitting a model with the term for age group (i.e fit a model with mf and agegroup

```
# Fit the model with age groups
m1 <- glm(mf ~ as.factor(agegrp), data=onch, family=binomial)
```

2. **Obtain the value of** L_0 This requires us to fit a model without the term for age group (i.e. Fit a model with mf alone)

```
# Fit the empty model
m0<- glm(mf ~ 1, data=onch, family=binomial)
```

3. **Compare** L_1 *and* L_0

```
anova(m0,m1,test="LRT") # Compare the two LLs using anova
```

Which results in

```
# Fit the model with age groups
m1 <- glm(mf ~ as.factor(agegrp), data=onch, family=binomial)
# Fit the empty model
m0<- glm(mf ~ 1, data=onch, family=binomial)
anova(m0,m1,test="LRT") # Compare the two LLs using anova
```

```
## Analysis of Deviance Table
##
## Model 1: mf ~ 1
## Model 2: mf ~ as.factor(agegrp)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1301     1714.1
## 2      1298     1455.7  3    258.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Practical 2

- Use the dataset ond15p.csv
  - Explore the association between microfilarial infection and optic nerve disease by modelling Ond (optic nerve disease) with Mfpos (microfilarial positve/negative)
    - add sex (male/female) to mfpos; and then agegrp to mfpos (separately)
- Tabulate ond and agegrp- calculate the chi test
- Compute the odds ratio of optic nerve disease in the various age groups
- Comment on the Wald test and the 95% CI
- Test whether adding agegroup into the model with *mfpos* and *sex* already in it improves model fit and comment on your findings

Interaction (or Effect modification)

Definition

". there is an interaction between the effects of two exposures if the effect of one exposure varies according to the level of the other exposure." p322 Kirkwood and Sterne, Essential Medical Statistics 2nd Ed, 2003 Blackwell

Example

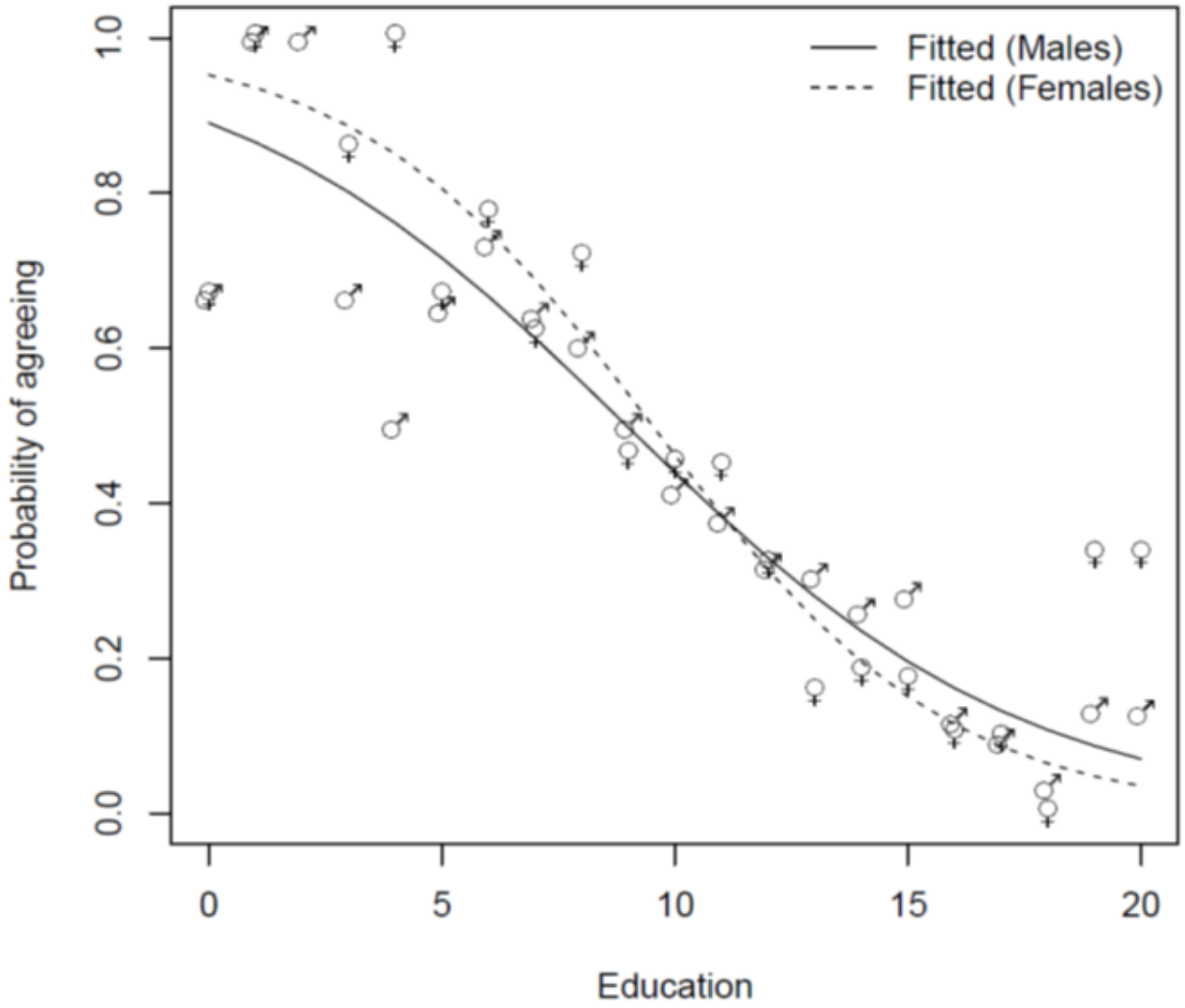
"... the protective effect of breastfeeding against infectious diseases in early infancy is more pronounced among infants living in poor environmental conditions than among those living in areas with adequate water supply and sanitation facilities" Kirkwood & Sterne *ibid*

Intreaction example

```
data("womensrole", package = "HSAUR2")
fm1 <- cbind(agree, disagree) ~ gender + education
wrole1 <- glm(fm1, data = womensrole, family = binomial())
coef(wrole1)
```

```
##   (Intercept) genderFemale    education
##    2.50937187  -0.01144685  -0.27062085
```

Main Effects Model

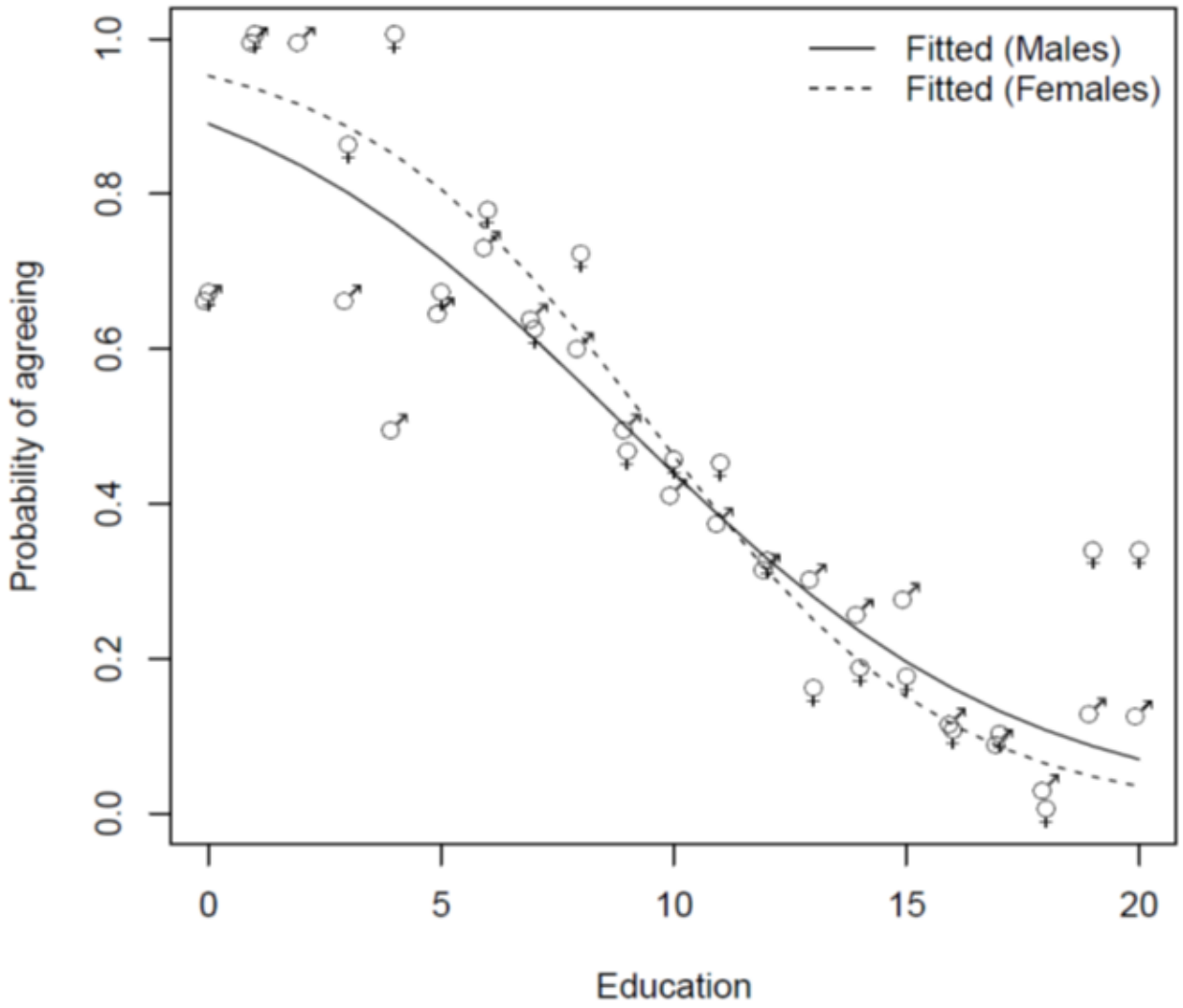


Fitting and testing a model with an interaction

```
# Fit the model with gender and education interacting
wrole2 <- glm(cbind(agree, disagree) ~ gender * education,
              data =  womensrole, family = binomial())
# Test this model against the main effect model
anova(wrole2,wrole1,test="Chisq") ## NB Chisq <==> LRT
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(agree, disagree) ~ gender * education
## Model 2: cbind(agree, disagree) ~ gender + education
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        37     57.103
## 2        38     64.007 -1  -6.9039 0.008601 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

With interaction (effect modification accounted for)



Summary

1. Obtain log odds of outcome
2. Obtain OR and 95% CI
3. Wald Test (null hypothesis: OR=1)
    - Assess null hypothesis for each level/group
4. Likelihood Ratio Test (null hypothesis: OR=1)
    - Assess null hypothesis for addition of an extra term/variable
5. Application of LRT to check for effect modification in logistic regression

Practical 3

- Use the dataset onchall.csv
  - Fit a model predicting microfilarae infection (mf) with both area and agegrp as main effects
  - Fit a model of mf with the interaction between the two explanatory variables area and agegrp
  - Compute a likelihood ratio test of the more complex model compared to the simpler model
- Which model should we use? The simpler one without the interaction or the more complicated with the interaction?