Logistic example

Ken Mwai

12/6/2019

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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1714.1  on 1301  degrees of freedom
## Residual deviance: 1657.0  on 1300  degrees of freedom
## AIC: 1661
##
## Number of Fisher Scoring iterations: 4
```

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.

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

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 ~ area + as.factor(agegrp), family = binomial,
##     data = onch)
##
## Coefficients:
##        (Intercept)                area  as.factor(agegrp)1
##            -1.9148              1.1260              0.9552
## as.factor(agegrp)2  as.factor(agegrp)3
##             2.2788              2.8703
##
## Degrees of Freedom: 1301 Total (i.e. Null);  1297 Residual
## Null Deviance:       1714
## Residual Deviance: 1385  AIC: 1395
```

Predictions as ORs with Confidence intervals

```
exp(coef(m2))     # transform the coeffs into ORs #
##        (Intercept)                area as.factor(agegrp)1
##          0.1473754           3.0832239           2.5991319
## as.factor(agegrp)2 as.factor(agegrp)3
##          9.7654104          17.6415839
exp(confint(m2))   # and show their CIs
## Waiting for profiling to be done...
##                            2.5 %     97.5 %
## (Intercept)          0.09918879  0.2145632
## area                 2.35981704  4.0492517
## as.factor(agegrp)1   1.68960336  4.0396884
## as.factor(agegrp)2   6.54431561 14.8185868
## as.factor(agegrp)3 11.65289300 27.1890105
```

Note separate Wald tests

```
summary(m2)
##
## Call:
## glm(formula = mf ~ area + as.factor(agegrp), family = binomial,
##     data = onch)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0972  -0.8656   0.4849   0.8068   2.0260
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.9148     0.1966  -9.741  < 2e-16 ***
## area                 1.1260     0.1376   8.181 2.82e-16 ***
## as.factor(agegrp)1   0.9552     0.2221   4.301 1.70e-05 ***
## as.factor(agegrp)2   2.2788     0.2082  10.944  < 2e-16 ***
## as.factor(agegrp)3   2.8703     0.2159  13.295  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1714.1  on 1301  degrees of freedom
## Residual deviance: 1384.8  on 1297  degrees of freedom
## AIC: 1394.8
##
## Number of Fisher Scoring iterations: 4
```

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

```
## Analysis of Deviance Table
```

```
##
## Model 1: mf ~ area
## Model 2: mf ~ area + as.factor(agegrp)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1300     1657.0
## 2      1297     1384.8  3   272.22 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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*

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?