Topic 13

Logistic Regression

Sergey Mastitsky ©
Klaipeda, 28-30 September 2011

13. Logistic regression

13.1. Logistic regression as a generalized linear model

Generalized Linear Models (GLMs)

- Thus far, we assumed that the response variable y was normally distributed and had constant variance irrespective of x
- In many situations, however, response variables are inherently non-normal and demonstrate positive relationship between variance and mean:
  - count data expressed as proportions
  - count data that are not proportions
  - binary response variables
  - data on time to death

Generalized Linear Models (GLMs)

- Generalized Linear Models – class of models designed to deal with the abovementioned non-normal response variables
- These models are characterized by:
  - an *error distribution* giving the distribution of the response around its mean (e.g., binomial, Poisson, Gamma)
  - a *link function*, *g*, which transfers the mean values of response to a scale in which the relation to predictors becomes linear and additive
  - the *variance function*

Common link functions in GLMs

- The link function linearizes the response:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k$$

- Common link functions:
  - `identity` -> normal errors (e.g., linear regression, ANOVA)
  - `poisson` -> Poisson errors (for counts)
  - `logit` -> binomial errors (binomial responses, counts as proportions)

Author: Sergey Mastitsky

Calculation of GLMs

- GLMs are estimated by the *method of maximum likelihood* (finds a set of parameters that optimizes a goodness-of-fit criterion)
- The measure of fit is expressed as *deviance*, which estimates how closely the model-based fitted values of the response approximate the observed values
- Two models can be compared with a likelihood-ratio test, which produces a χ^2-distributed statistic

Logistic regression

- Logistic regression is designed for binary response variables and proportions
- Probabilities of binary outcomes cannot be correctly analyzed with regression models (predicted values can become negative or >1)
- With the *logit* link, probabilities are transformed to a log scale, where they demonstrate linearity:

$$\operatorname{logit} p = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k$$

logit p = log of the odds in favor of an event of interest

$$\text{logit}\ p = \log[p/(1-p)]$$

13. Logistic regression

13.2. Logistic regression with tabulated data

Author: Sergey Mastitsky

Infection of *Dreissena polymorpha* with *Echinoparyphium recurvatum* in Lake Naroch

- From May to October 2006, *D. polymorpha* were collected monthly from depths of 0.8 m and 4 m in Lake Naroch, Belarus
- 15 molluscs were dissected at each sampling date from each depth to estimate the prevalence of infection (% infected) with the trematode *E. recurvatum*
- Did the prevalence change significantly over the period of study, and was there a difference between depths?

Loading *Dreissena* infection data

- Use the command

```
> setwd("~/Introductory R
+ Course/R_Course_Datasets")
```

- Or in **RStudio** do

Tools -> Set Working Directory -> Choose Directory -> …your Desktop -> folder "Introductory R Course" -> folder "R_Course_Datasets"

Loading *Dreissena* infection data

```
> infection <- read.table(
file = "dreissena_infection.txt",
header = TRUE,
sep = "\t")
```

Examine the data:

```
> infection
> summary(infection)
```


Author: Sergey Mastitsky

Fitting logistic regression to tabular data in R

- R can fit logistic regression to tabular data in two different ways:
  - Response is specified as a matrix where one column is the number of “diseased” and the other is the number of “healthy” individuals
  - Response is specified as proportions of “diseased” from total

Fitting logistic regression to tabular data in R

```
# Fitting response as a matrix:
> inf.tbl <-
  cbind(infection$Infected,
        infection$Noninfected)

> M1 <- glm(inf.tbl ~ Day + Depth,
 family = binomial(link = "logit"),
 data = infection)
```

Results of the logistic regression analysis

```
> summary(M1)
Call:
glm(formula = inf.tbl ~ Day + Depth, family = binomial(link = "logit"),
    data = infection)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.1129  -0.9595  -0.1563   0.7182   2.0214

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.836627   0.651911  -5.885 3.98e-09 ***
Day          0.011039   0.004623   2.388  0.01695 *
Depth4m      1.543597   0.537679   2.871  0.00409 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.394  on 11  degrees of freedom
Residual deviance: 18.146  on  9  degrees of freedom
AIC: 45.338

Number of Fisher Scoring iterations: 5
```


Author: Sergey Mastitsky

Results of the logistic regression analysis

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1129  -0.9595  -0.1563   0.7182   2.0214
```

- The *deviance* corresponds to the sum of squares in linear normal models
- *Deviance Residuals* indicate contribution of each cell of the table to the deviance of the model

Results of the logistic regression analysis

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.836627   0.651911  -5.885 3.98e-09 ***
Day          0.011039   0.004623   2.388  0.01695 *
Depth4m      1.543597   0.537679   2.871  0.00409 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

- Estimates of the regression coefficients and their significance (interpretation is identical to the linear regression output)

Results of the logistic regression analysis

```
    Null deviance: 34.394  on 11  degrees of freedom
Residual deviance: 18.146  on  9  degrees of freedom
AIC: 45.338
```

- *Null deviance* – deviance of the “empty” model
- *Residual deviance* – the deviance which is left unexplained after incorporating `Month` and `Depth` into the model
- *AIC* – measure of goodness-of-fit that takes the number of fitted parameters into account

Results of the logistic regression analysis

```
Number of Fisher Scoring iterations: 5
```

- Purely technical term
- Indicates how many iterations were performed before satisfactory estimations of the model coefficient were found
- Don’t pay too much attention to it. However, if the number of iterations is large, the model is likely to be to complex

The analysis of deviance table

Similar to ANOVA tables in multiple regression analysis:

```
> anova(M1, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: inf.tbl

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     11     34.394
Day    1   6.4053        10     27.989  0.011378 *
Depth  1   9.8424         9     18.146  0.001705 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Be careful with interpretation of the P-values!

Fitting logistic regression for tabular data in R

```
# Fitting responses as proportions from total:
> n.total <- infection$Infected + infection$Noninfected
> prop.inf <- infection$Infected/n.total

> M2 <- glm(prop.inf ~ Day + Depth, weights = n.total, family = binomial(link = "logit"), data = infection)
```


Author: Sergey Mastitsky

13. Logistic regression

13.3. Logistic regression with raw data

Raw data on *Dreissena* infection

```
> inf.raw <- read.table(
file =
"dreissena_infection_raw_data.txt",
header = TRUE,
sep = "\t")

> head(inf.raw)
```

Fitting logistic regression to raw binary data in R

```
> M3 <- glm(EchinoPresence ~
  Length + Day + Depth,
  family = binomial(link =
  "logit"), data = inf.raw)

> summary(M3)
```

Coefficients of M3

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.302906   1.038326  -3.181  0.00147 **
Length      -0.043238   0.054376  -0.795  0.42652
Day          0.010781   0.004829   2.233  0.02556 *
Depth4m      1.569001   0.621190   2.526  0.01154 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.20  on 181  degrees of freedom
Residual deviance: 116.72  on 178  degrees of freedom
AIC: 124.72
```

Reducing M3

```
> M4 <- glm(EchinoPresence ~
  Day + Depth,
  family = binomial(link =
  "logit"), data = inf.raw)
```

Comparing M3 and M4

```
> anova(M3, M4, test = "Chisq")

Analysis of Deviance Table

Model 1: EchinoPresence ~ Length + Day + Depth
Model 2: EchinoPresence ~ Day + Depth
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       178     116.72
2       179     117.36 -1 -0.64642    0.4214


> AIC(M3, M4)
   df      AIC
M3  4 124.7151
M4  3 123.3615
```