AMS 597: Statistical Computing

Pei Fen Kuan
Generalized linear models

- Logistic and Poisson regressions belong to the class of generalized linear models.
- These models are characterized by their response distribution (here the binomial distribution) and a link function, which characterizes the relationship between the linear predictor and the mean of the distribution function, i.e.

  \[ E(Y) = \mu = g^{-1}(X\beta) \]

- where \( g \) is the link function
Generalized linear models

- In a logistic regression analysis, the link function is the logit function \( x\beta = \ln\left(\frac{\mu}{1-\mu}\right) \).

- In R generalized linear models are handled by the `glm` function. This function is very similar to `lm`.

- The two functions use essentially the same model formulas and extractor functions (summary, etc.), but `glm` also needs to have specified which generalized linear model is desired.
This is done via the family argument. To specify a binomial model with logit link (i.e., logistic regression analysis), you write
family=binomial("logit").

The parameters of the model can be estimated by the method of maximum likelihood via the iteratively reweighted least squares algorithm. This is a quite general technique, similar to the least-squares method in that it finds a set of parameters that optimizes a goodness-of-fit criterion.
Generalized linear models

- In fact, the least-squares method itself is a slightly modified maximum-likelihood procedure.
- The likelihood function \( L(\beta) \) is the probability of the entire observed data set for varying parameters.
- The deviance is
  \[
  D = 2[\log L(\hat{\beta}) - \log L(\hat{\beta})]
  \]
- where \( \hat{\psi} \) and \( \hat{\beta} \) are the MLEs of the saturated and proposed model, respectively.
Generalized linear models

- Changes in deviance caused by a model reduction will be approximately chi-squared distributed with degrees of freedom equal to the change in the number of parameters, i.e.,

\[ D \sim \chi^2_{K-p} \]

- where \( K \) and \( p \) are the number of parameters in the saturated and proposed models, respectively.
Logistic regression

- Sometimes you wish to model binary outcomes, variables that can have only two possible values: diseased or non-diseased, and so forth.

- A linear model for transformed probabilities can be set up as

\[ \text{logit } p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \]

in which \( \text{logit } p = \log[p/(1–p)] \) is the log odds.
A constant additive effect on the logit scale corresponds to a constant odds ratio.

Logistic slope coefficients can be interpreted as the effect of a unit of change in the X variable on the predicted logits with the other variables in the model held constant, i.e., how a one unit change in X effects the log of the odds when the other variables in the model held constant.
Logistic regression

- Odds ratios in logistic regression can be interpreted as the effect of a one unit of change in $X$ in the predicted odds ratio with the other variables in the model held constant.

$$\frac{\text{odds}(\text{if the corresponding variable is incremented by 1})}{\text{odds}(\text{if variable not incremented})}$$

$$\frac{P(\text{event} \mid x + 1) / (1 - P(\text{event} \mid x + 1))}{P(\text{event} \mid x) / (1 - P(\text{event} \mid x))}$$
Logistic regression

- For instance, say you estimate the following logistic regression model:

  \[ -13.70837 + 0.1685 x_1 + 0.0039 x_2 \]

- The effect of the odds of a 1-unit increase in \( x_1 \) is \( \exp(0.1685) = 1.18 \)

- Meaning incrementing \( x_1 \) by 1 unit increases the odds by 18% regardless of the value of \( x_2 \)
The choice of the logit function is not the only one possible, but it has some mathematically convenient properties.

Other choices do exist; the probit function (the quantile function of the normal distribution) or \( \log(-\log p) \), which has a connection to survival analysis models.
One thing to notice about the logistic model is that there is no error term as in linear models. We are modelling the probability of an event directly, and that in itself will determine the variability of the binary outcome. There is no variance parameter as in the normal distribution.
Logistic regression

```r
## Construct a dataset
no.yes <- c("No","Yes")
smoking <- gl(2,1,8,no.yes)
## The function ‘gl’ is to generate levels. The first three
## arguments to gl are, respectively, the number of levels,
## the repeat count of each level, and the total length of the
## vector.
obesity <- gl(2,2,8,no.yes)
snoring <- gl(2,4,8,no.yes)
n.tot <- c(60,17,8,2,187,85,51,23)
n.hyp <- c(5,2,1,0,35,13,15,8)
data.frame(smoking,obesity,snoring,n.tot,n.hyp)

  smoking obesity snoring n.tot n.hyp
  1     No      No    No    60    5
  2    Yes      No    No    17    2
  3      No      Yes    No     8    1
  4    Yes      Yes    No     2    0
  5     No      No     Yes   187   35
  6    Yes      No     Yes    85   13
  7      No      Yes     Yes    51   15
  8    Yes      Yes     Yes    23    8
```
R is able to fit logistic regression analyses for tabular data in two different ways.

You can specify the response as a matrix, where one column is the number of “diseased” and the other is the number of “healthy” (or “success” and “failure”, depending on context).

Alternatively, the response can be a binary vector.
```r
> hyp.tbl <- cbind(n.hyp, n.tot-n.hyp)
> glm.hyp <- glm(hyp.tbl ~ smoking + obesity + snoring, family = binomial("logit"))
> summary(glm.hyp)

Call:
glm(formula = hyp.tbl ~ smoking + obesity + snoring, family = binomial("logit"))

Deviance Residuals:
     1       2       3       4       5       6
  -0.04344  0.54145 -0.25476 -0.80051  0.19759 -0.46602
  7       8
 -0.21262  0.56231

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.37766  0.38018  -6.254   4e-10 ***
smokingYes   -0.06777  0.27812  -0.244    0.8075
obesityYes   0.69531  0.28509   2.439    0.0147 *
snoringYes   0.87194  0.39757   2.193    0.0283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537
```
Logistic regression

> anova(glm.hyp, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: hyp.tbl

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>7</td>
<td>14.1259</td>
<td>0.962724</td>
</tr>
<tr>
<td>smoking</td>
<td>1</td>
<td>0.0022</td>
<td>7.2963</td>
</tr>
<tr>
<td>obesity</td>
<td>1</td>
<td>6.8274</td>
<td>0.008977  **</td>
</tr>
<tr>
<td>snoring</td>
<td>1</td>
<td>5.6779</td>
<td>0.017179  *</td>
</tr>
</tbody>
</table>

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Logistic regression

- Notice that the Deviance column gives differences between models as variables are added to the model in turn (alt: recall Anova() in library(car) for Type II and III errors)
- The deviances are approximately chi-squared distributed with the stated degrees of freedom. It is necessary to add the test="chisq" argument to get the approximate chi-square tests.
- From the above, you can read that smoking is removable.
Logistic regression

```r
> glm.hyp <- glm(hyp.tbl~obesity+snoring,binomial)
> anova(glm.hyp, test="Chisq")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: hyp.tbl

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>7</td>
<td>14.1259</td>
<td>7</td>
<td>14.1259</td>
<td></td>
</tr>
<tr>
<td>obesity</td>
<td>1</td>
<td>6.8260</td>
<td>6</td>
<td>7.2999</td>
<td>0.008984 **</td>
</tr>
<tr>
<td>snoring</td>
<td>1</td>
<td>5.6218</td>
<td>5</td>
<td>1.6781</td>
<td>0.017738 *</td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Logistic regression

- Example 2:
Many studies often involve the calculation of rates, typically rates of death or incidence rates of a chronic or acute disease.

This is based upon counts of events occurring within a certain amount of time.

The Poisson regression method is often employed for the statistical analysis of such data.
However, data that are not actually counts of events but rather measurements of time until an event (or non-event) can be analyzed by a technique which is formally equivalent.

The data that we wish to analyze can be in one of two forms.

They can be in aggregate form as an observed count $x$ based on a number of person-years $T$ (years of observation time per person).
We may also have individual-level data, in which for each subject we have a time under observation $T_i$ and a 0/1 indicator $x_i$ of whether the subject has had an event.

The class of generalized linear models also includes the Poisson distribution, which by default uses a log link function.

This is the mathematically convenient option and also a quite natural choice since it allows the linear predictor to span the entire real line.
Poisson regression

- We can use this to formulate models for the log rates of the form

\[ \log \rho = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \]

- The following example was used by Erling B. Andersen in 1977.

- It involves the rates of lung cancer by age in four Danish cities and may be found as eba1977 in the ISwR package.
Poisson regression

> names(eba1977)
[1] "city" "age" "pop" "cases"
> attach(eba1977)
> fit <- glm(cases~city+age, family=poisson)
> summary(fit)

Call:
glm(formula = cases ~ city + age, family = poisson)

Deviance Residuals:
          Min          1Q       Median          3Q          Max
-2.54853    -0.57942    -0.02872     0.49797     1.68933
Poisson regression

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 2.24374  | 0.20363    | 11.019  | <2e-16   *** |
| cityHorsens    | -0.09844 | 0.18129    | -0.543  | 0.587    |
| cityKolding    | -0.22706 | 0.18770    | -1.210  | 0.226    |
| cityVejle      | -0.22706 | 0.18770    | -1.210  | 0.226    |
| age55-59       | -0.03077 | 0.24810    | -0.124  | 0.901    |
| age60-64       | 0.26469  | 0.23143    | 1.144   | 0.253    |
| age65-69       | 0.31015  | 0.22918    | 1.353   | 0.176    |
| age70-74       | 0.19237  | 0.23517    | 0.818   | 0.413    |
| age75+         | -0.06252 | 0.25012    | -0.250  | 0.803    |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 27.704  on 23  degrees of freedom
Residual deviance: 20.673  on 15  degrees of freedom
AIC: 135.06

Number of Fisher Scoring iterations: 5
Poisson regression

- Example 2:
  - [https://onlinecourses.science.psu.edu/stat504/node/169](https://onlinecourses.science.psu.edu/stat504/node/169)