Part II: Generalized Linear Models

Chapter II.3

Models for Binary Response

Logistic Regression: Goodness of Fit - Infinite Estimates

Goodness of Fit for Logistic Regression

II.3.20 Remark (analysis of deviance)

For GLMs, the analysis of variance procedure of LMs generalizes to the *analysis of deviance*.
For binomial responses $m_i Y_i \sim \mathcal{B}(m_i,\pi_i)$, it holds $a(\phi;i) = \phi/w_i$ with $\phi = 1$ and $w_i = m_i$, $i = 1,\ldots,n$. If all $w_i = 1$, then the data are ungrouped. In this case, if $\ell_{\mathcal{M}}(\hat{\boldsymbol{\mu}};\boldsymbol{y})$ and ℓ_{sat} are the maximized log likelihood for the model under consideration $\mathcal{M}$ and the saturated model $\mathcal{M}_{sat}$, then the deviance (s. Definition II.2.35):

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = -2\left[\ell(\hat{\boldsymbol{\mu}};\boldsymbol{y}) - \ell_{sat}\right] = \sum_{i=1}^{n} w_i \left(y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\right),$$

with $\hat{\theta}_i$ being the MLE of θ_i under $\mathcal{M}$ and $\tilde{\theta}_i = y_i$ (perfect fit under $\mathcal{M}_{\text{sat}}$), equals the LRS G^2 for testing model $\mathcal{M}$ (s. also Remark II.2.37(2)):

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = D(\boldsymbol{y};\hat{\boldsymbol{\pi}}) = G^2(\mathcal{M}) = 2\sum_{i=1}^{n} m_i y_i \log\left(\frac{m_i y_i}{m_i \hat{\pi}_i}\right) + 2\sum_{i=1}^{n} m_i(1-y_i)\log\left(\frac{m_i(1-y_i)}{m_i(1-\hat{\pi}_i)}\right).$$

☞ $m_i y_i$ $[m_i(1-y_i)]$ is the number of observed successes [failures] in the i-group ($i = 1,\ldots,n$).

Binary Logistic Regression Goodness of Fit (GoF) for Grouped Data

II.3.21 LRS G^2 and Pearson's X^2

For Binomial GLMs with grouped data ($\mathbf{x}$: categorical with n settings), the deviance (i.e LRS) G^2 and the Pearson's X^2 of a model $\mathcal{M}$ are Goodness of Fit (GoF) statistics for testing that $\mathcal{M}$ truly holds (H_0). They are expressed by

$$G^2 = 2\sum_i \sum_{k=1}^{2} n_{ik} \log\left(\frac{n_{ik}}{\hat{m}_{ik}}\right) \quad \text{and} \quad X^2 = \sum_i \sum_{k=1}^{2} \frac{(n_{ik} - \hat{m}_{ik})^2}{\hat{m}_{ik}} \ ,$$

where for the i-th group (i-th setting of $\mathbf{x}$) of size m_i ($w_i = m_i$), $i = 1, \ldots, n$, we denote

- n_{i1}: observed number of 'successes' ($Y = 1$) at $\mathbf{x}_i$,
- n_{i2}: observed number of 'failures' ($Y = 0$) at $\mathbf{x}_i$
- ☞ $n_{i+} = n_{i1} + n_{i2} = m_i$ (size of this group),
- $\hat{\pi}_i$: estimated success probability $P(Y = 1)$ at $\mathbf{x}_i$,
- $\hat{m}_{i1} = m_i \hat{\pi}_i$: predicted number of 'successes' for this group,
- $\hat{m}_{i2} = m_i(1 - \hat{\pi}_i)$: predicted number of 'failures' for this group ☞ $\hat{m}_{i1} + \hat{m}_{i2} = m_i$

Binary Logistic Regression Goodness of Fit (GoF) for Grouped Data

II.3.22 Remark (asymptotic equivalence of G^2 and X^2)

- For grouped data with fixed number of settings (n) for the explanatory variables

$$X^2,\ G^2 \overset{as.}{\sim} \chi^2_{df}\ ,$$

  as $n_{tot} = \sum_{i=1}^{n} m_i$ increases, provided that most $\{m_{ik}\}$ are large (practical rule: at least 80% of them ≥ 5). The degrees of freedom are $df = \text{dim}(\mathcal{M}_{sat}) - \text{dim}(\mathcal{M})$, where the dimension of the parameter space under $\mathcal{M}$ equals the number of parameters of this model ($\text{dim}(\mathcal{M}) = p$).

- The test statistics G^2 and X^2 are asymptotically equivalent under H_0 (model $\mathcal{M}$ holds). As n_{tot} increases, X^2 converges faster to χ^2-distribution than G^2 and behaves "better" when some m_{ik} are small (< 5). [a]

[a] For the proof of this equivalence and a detailed discussion s. Agresti (2013, Categorical Data Analysis (CDA), 3rd ed., Wiley, p. 597).

Binary Logistic Regression Goodness of Fit

II.3.23 Remark

- When some or all of the components of $\mathbf{x}$ are continuous, X^2 and G^2 are no more asymptotically χ^2 distributed, but they are still useful in comparing models applied on the same data set.

- X^2 and LRS (Deviance)
  In order to proceed to asymptotic inference with ungrouped data, *grouping* of the data is required.

  The grouping is based on the number of distinct values/levels of the explanatory variables. If this number of groups n is too large (common for continuous explanatory variables), then the expected values are to low (< 5) and hence we cannot compute p-values based on the χ^2-approximation.

  ☞ In such cases, if X^2/df (G^2/df) is close to 1, this is an indication of good fit.

II.3.24 The Test of Hosmer and Lemeshow

For ungrouped data, the total number of observations is $n_{tot} = n$. These n observations are grouped in g groups (usually, $g = 10$).

The first group consists of the n_1 observations that correspond to the n/g smallest $\hat{\pi}$, the 2nd group consists of the n_2 observations with the next n/g smallest $\hat{\pi}$, etc.

Let y_i^* be the number of observed successes in group i $(i = 1, \ldots, g)$ and $\overline{\pi}_i$ the average of $\hat{\pi}_i$ for the observations in group i.

The test is based on the statistical function

$$\hat{C} = \sum_{i=1}^{g} \frac{(y_i^* - n_i\overline{\pi}_i)^2}{n_i\overline{\pi}_i(1 - \overline{\pi}_i)} \ .$$

Hosmer and Lemeshow [a] proved via simulations that, provided that all the levels of the explanatory variables are different and the logistic regression model holds, the distribution of $\hat{C}$ is approximated by X^2_{g-2} (independently from the number of explanatory variables).

[a] Hosmer and Lemeshow (1980). A goodness-of-fit test for multiple logistic regression model. Commun. Stat. A 9: 1043-1069.

II.3.25 Example (horseshoe crabs data)

For a detailed desciption/analysis, see Agresti (2013, CDA, Sec. 4.3.2, Sec. 6.1)
Ungrouped data set: sample of $n = 173$ female crabs,

Variables:
C = color (4 categories)
S = spine condition (3 categories)
W = width of carapace shell (cm) $\rightarrow$ ($m = 66$ different levels)
SAT = number of satellites
WT = weight of crab (kg)
Y = whether a female horseshoe crab has 'satellites' (1 = yes, 0=no)

- Data in file '`crabs.dat`'.

```
> setwd("C:/.../ADA_II (for R)");  fungal <- read.table("crabs.dat", header=T)
> crabs[1:2,] # try:  head(crabs)
     C   S     W     SAT    WT    Y
 1   3   3   28.3    8    3050   1
 2   4   3   22.5    0    1550   0
 3   2   1   26.0    9    2300   1
> cor(crabs$W,crabs$WT)  # high correlated weight and width
[1] 0.8868715
> crabs$S <- factor(crabs$S)
> # color has values from 2 (light) to 5 (dark) --> recode to 1-4:
> C4 <- factor(crabs$C-1); C2 <- crabs$C
> C2[which(crabs$C < 5)] <- 2    # -> merge color to binary:
> C2[which(crabs$C == 5)] <- 1    # 1:dark, 2:other
> C2 <- factor(C2)
> fit <- glm(Y ~ C2*W, family=binomial, data =crabs)
> anova(fit)
Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                   172     225.76
C2    1  10.9656       171     214.79
W     1  26.8351       170     187.96
C2:W  1   1.1715       169     186.79
```

```
> fit2 <- glm(Y ~ C2+W, family=binomial, data =crabs)
> summary(fit2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.9795     2.7272  -4.759 1.94e-06 ***
C22           1.3005     0.5259   2.473   0.0134 *
W             0.4782     0.1041   4.592 4.39e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom
AIC: 193.96

Number of Fisher Scoring iterations: 4
```

Infinite Estimates in Logistic Regression

II.3.26 Remark (on ML estimation of logistic model parameters)

At least one parameter estimate is infinite if we can separate with a plane the x values where $y = 1$ and where $y = 0$.

- Complete separation: No observations on that plane
- Quasi-complete separation: On the plane boundary, both outcomes occur (common in contingency tables, i.e. logit models)

☞ Most software does not adequately detect this.

Data with infinite MLE in R

II.3.27 Example

```
> x <- c(10, 20, 30, 40, 60, 70, 80, 90)
> y <- c(0, 0, 0, 0, 1, 1, 1, 1)    # complete separation
> fit <- glm(y ~ x, family=binomial)
> summary(fit)
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-118.158	296046.187	0	1
x	2.363	5805.939	0	1

```
Residual deviance: 2.1827e-10 on 6 degrees of freedom
```

Remark: If we add two observations at $x = 50$, one with $y = 1$ and one with $y = 0$, then quasi-complete separation occurs.

☞ No warning for complete or quasi-complete separation!!
Sign: Huge values for the Std. Errors of the parameters' MLEs.

II.3.28 Example (grouped data: quasi-complete separation)

Consider the following fungal infection data (Agresti, CDA 2013, Section 6.5.2):

		Response	
Center (C)	Group	Success (S)	Failure (F)
1	Treatment (T)	0	5
	Placebo	0	9
2	Treatment	1	12
	Placebo	0	10
3	Treatment	0	7
	Placebo	0	5
4	Treatment	6	3
	Placebo	2	6
5	Treatment	5	9
	Placebo	2	12

- Data in file '`fungal.dat`'.

```
> setwd("C:/.../ADA_II (for R)");  fungal <- read.table("fungal.dat", header=T)
> fungal[1:2,] # try:  head(fungal)
     center   treatment   my   m
 1      1          1       0    5
 2      1          0       0    9
> fit <- glm(my/m ~ treatment + factor(center), weights=m, family=binomial, data=fungal)
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.459e+01  2.330e+04  -0.001   0.9992
treatment         1.546e+00  7.017e-01   2.203   0.0276 *
factor(center)2   2.039e+01  2.330e+04   0.001   0.9993
factor(center)3   4.809e-03  3.172e+04   0.000   1.0000
factor(center)4   2.363e+01  2.330e+04   0.001   0.9992
factor(center)5   2.257e+01  2.330e+04   0.001   0.9992
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.53202  on 9  degrees of freedom
Residual deviance:  0.50214  on 4  degrees of freedom
AIC: 24.859

Number of Fisher Scoring iterations: 21
```

```
> summary(fit)
```

Model: $\text{logit}[P(S)] = \log\left(\frac{P(S)}{1-P(S)}\right) = \beta_0 + \beta^T x + \beta_j^C$,
with $\beta_1^C = 0$ (for identifiability) and $x = 1$ for drug and 0 for control.

Equivalently we can fit a model without intercept: $\text{logit}[P(S)] = \alpha_j^C + \beta^T x$,
for which no identifiability constraint is required (☞ localizes the problem in centers 1 and 3).

```
> fit2 <- glm(my/m ~ -1+treatment + factor(center), weights=m, family=binomial, data=fungal)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
treatment             1.5460     0.7017   2.203 0.027569 *
factor(center)1     -24.5922 23296.3959  -0.001 0.999158
factor(center)2      -4.2025     1.1891  -3.534 0.000409 ***
factor(center)3     -24.5874 21523.6453  -0.001 0.999089
factor(center)4      -0.9592     0.6548  -1.465 0.142956
factor(center)5      -2.0223     0.6700  -3.019 0.002540 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 73.07369  on 10  degrees of freedom
Residual deviance:  0.50214  on  4  degrees of freedom
AIC: 24.859

Number of Fisher Scoring iterations: 21
> summary(fit2)
```

- Zero margins for centers 1,3 $\longrightarrow \hat{\alpha}_1^C = \hat{\alpha}_3^C = \infty$
- Strategies to estimate treatment effect β:
  - remove 'uninformative' centers 1,3
  - or add very small constant (e.g. 10^{-8}) to zero cells, so that all estimates exist
  - or combine some centers
- Cochran-Mantel-Haenszel test or exact test about treatment effect (not considered here) ignore centers 1,3

II.3.29 Remark

In Example II.3.28, quasi-complete separation affects $\{\hat{\alpha}_j^C\}$, but not $\hat{\beta}$.

Alternative Approaches for Infinite Estimates?

II.3.30 Remark

- Bayesian approach:
  Influence of prior distribution smooths data and results in finite posterior mean estimates.
- Penalized likelihood approach (Firth, Biometrika 1993):
  Add a penalty term to the likelihood function. Maximizing the penalized likelihood results in shrinking estimates toward 0. (In `R`: package `logistf`)