#8-2017F

Logistic Regression I

Summary of the last lecture

- Types of residuals
- Residual plots
- Over-dispersion

Key terms of this lecture

- Logistic regression
  - Terminology
  - Ungrouped data/Grouped data

Reading

- McCullagh and Nelder (1989) Chapter 7
- Dobson and Bartnett (2008) Chapter 4

Binomial Distribution

- Example: Budworm data. This example comes from Venables and Ripley's *Modern Applied Statistics with S*, Springer 4th edition, 2002. [See the attached pages for detail.]
  - Batches of 20 moths subjected to increasing doses of a poison, "success"=death.
  - Data is grouped: for each of 6 doses (1.0, 2.0, 4.0, 8.0, 16.0, 32.0 mg) and each of male and female, we have 20 moths. in each group.
  - There are 12 covariate patterns. or there are $6 \times 2 = 12$ groups based on different values of X's.

dose	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

The number of moths that died in each group is shown in the table. Data suggests that the probability of dying increases with dose.

Terminology

- Odds for a single moth to die

$$\text{odds}(p) = \frac{p}{1-p}.$$

- Log-odds (a.k.a logit)

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Terminology (cont'd)

- Between two groups, e.g., exposed vs. unexposed, Trt1 vs. Trt2)

  – In budworm example, consider the male budworms receiving doses 1 and 2 at which $y_1 = 1$ and $y_2 = 4$ moths died out of $m = 20$.

  - Risk Difference: $RD = p_1 - p_2$ (e.g., $0.05 - 0.2 = -0.15$) $\frac{1}{20} - \frac{4}{20} = 0.05 - 0.2$
  - Relative Risk: $RR = p_1/p_2$ (e.g., $0.05/0.2 = 0.25$)
  - Odds of dying: $\text{odds}(p_1) = p_1/(1-p_1)$ and $\text{odds}(p_2) = p_2/(1-p_2)$
  - Odds ratio: $OR = \text{odds}(p_1)/\text{odds}(p_2)$ (e.g., $(0.05/0.95)/(0.2/0.8) = 0.21$) odds of dying at dose 1 related to dose 2
  - Log-Odds ratio: Log-$OR = \log(OR)$ (e.g., $\log(0.21) = -1.56$)

Terminology (cont'd)

- Interpretation:
  - RD: hard to explain a difference of -0.15 in probability
  - RR and OR: A ratio of 0.25 in probability is easier to interpret.
    - $RR = 0.25$: The risk of dying from dose 1 is only 1/4 of the risk of dying from dose 2.
    - Thus, having $RR < 1$ means that dose 1 is less poisonous than dose 2.
    - Note: 0.25 does not say anything about the risk of dying from dose 2.
  - Log-OR $OR \in (0, \infty)$; Log-$OR \in \ -\infty, \infty)$. Useful in interpretation of parameters in logistic regression.

Grouped vs. Ungrouped Data

- Grouped data (budworm example, seed example)
  - Data are presented by distinct covariate values (e.g., dose-gender seed-extract combination)
  - After counting number of responses in each category, we have a binomial r.v. (i.e. Response is the number of success)
- Ungrouped data
  - Each record represents one individual. Response is binary (0 or 1)
- Example. Seed data
  - Both results are exactly the same.
  - Differences:
    - DFs are different, so is GOF
    - Overdispersion be found in grouped data.

Grouped data vs. ungrouped data.
Estimates and standard errors are the same, but the measures of goodness of fit differ. see example: Senility and Waise page 140, 3rd edi
Further grouping: the Hosmer–Lemeshow ~~Statistic~~ χ^2_{HL}. By grouping values of $\hat{\pi}_i$, so ~~that the total number~~ of observations per group are approximately equal

Budworm Example

- Goal: describe the dependence of the mortality of budworms on sex S and the applied dose D.
  - i.e. explain $E[Y]$ as a function of the covariates.
- Data:
  - D: logarithm of dose (1,2,4,8,16,32 (μg)) of cypermethrin (continuous covariates)
  - S: sex (male, female) (class or factor covariates)
- Distribution assumption: $y_{ij} \sim$ Binomial(n_{ij}, p_{ij}), for $i = 1, 2$, $j = 1, \quad , 6,$ with $E[Y_{ij}] = n_{ij}p_{ij}$ where $n_{ij} = 20$ known in this example.
- Thus, the question is the relationship between p_{ij} and covariates.

Budworm Example (cont'd)

- Exploratory Plots: see R and SAS code, `budworm.R` and `budworm.SAS`
  - Observed proportions

$$\hat{p}_{ij} = \frac{y_{ij}}{n_{ij}}.$$

  - Empirical logits

$$\text{elogit}_{ij} = \log\left(\frac{y_{ij} + 0.5}{n_{ij} - y_{ij} + 0.5}\right)$$

- What do plots suggest? [See the attached plots.]

Budworm Example (cont'd)

- Components of GLM logistic regression
  - Random component

$$y_{ij} \sim Bin(n_{ij}, p_{ij}), \quad n_{ij} = 20$$

$$\text{or } y_i \sim Bin(n_i, p_i), \quad n_i = 20$$

  - Linear predictor (Systematic component)

$$\eta_i = \beta_0 + \beta_1 \, Sex + \beta_2 \, logdose$$

  - Link function

$$\eta_i = \log \frac{p_i}{1-p_i} \quad \text{or} \quad \log \frac{p_i}{1-p_i} = \eta_i$$

- Thus, the model would be,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \, Sex + logdose$$

Budworm Example (cont'd)

- Parameter estimation
  - MLE of β using IWLS.
- Interpretation of Parameters
  - β_0: log-odds when Sex $=0$ (Female) and logdose $=0$ (dose $=1$)
  - β_1. log-OR when Sex is changed from Female ($=0$) to male ($=1$), holding logdose Constant
  - $\beta_0 + \beta_1$ log-odds when Sex $=1$ (male) and logdose $=0$ (dose $=1$)
  - β_2: log-OR when logdose is increased by one unit while holding Sex unchanged.

- Revisit:
  - Odds

$$\text{odds}(p_{ij}) = \frac{p_{ij}}{1-p_{ij}} = \exp(\text{logit}(p_{ij})) = \exp(\eta_{ij}),$$

Thus

$$p_{ij} = \frac{\exp(\text{logit}(p_{ij}))}{1+\exp(\text{logit}(p_{ij}))}.$$

$$= \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}$$

The fitted probabilities: $\hat{p}_{ij} = \frac{\exp(\hat{\eta}_{ij})}{1+\exp(\hat{\eta}_{ij})}$, $\hat{\eta}_{ij} = X_{ij}^T\hat{\beta}$.

$\hat{\eta}_{ij}$ vs logdose —— linear

$\hat{p}_{ij}$ vs logdose —— non-linear.

Budworm Example (cont'd): Interpretation

- Interpretation of a continuous covariate
  - Using GLM for Binomial distribution, let $D = \log(\text{dose})$,

$$\begin{aligned}\text{logit}(p_{i,d+1}) &= \beta_0 + \beta_1 S + \beta_2(D+1)\\ \text{logit}(p_{i,d}) &= \beta_0 + \beta_1 S + \beta_2 D\end{aligned}$$

thus, β_2 is a change in logits per unit increase of log-dose. i.e.

$$\begin{aligned}\text{odds}(p_{i,D+1}) &= \exp(\text{logit}(p_i, D+1))\\ &= \exp(\beta_0 + \beta_1 S + \beta_2(D+1))\\ &= \exp(\beta_2)\exp(\beta_0+\beta_1 S+\beta_2 D)\\ &= \exp(\beta_2)\,\text{odds}(p_i, D)\end{aligned}$$

$$\Rightarrow \exp(\beta_2) = \frac{\text{odds}(p_i, D+1)}{\text{odds}(p_i, D)} = OR$$

$\beta_2 = \log OR$

thus, in example, increasing the log-dose by one unit increases the odds of death by a factor of $\exp(\beta_2)$

- Confidence interval for β_2 (Wald CI or LR CI)

e.g., Wald CI for β_2: $\hat{\beta}_2 \pm 1.96 SE(\hat{\beta}_2)$

- Confidence interval for the odds ratio (OR) is.

e.g., Wald CI for OR exp(CI for β_2)

$= \exp(\hat{\beta}_2 \pm 1.96\, SE(\hat{\beta}_2))$

This CI Excluding unity (1) indicates β_2 is significantly different from zero.

Budworm Example (cont'd): Interpretation

- Interpretation of a factor
  - Same as one for a continuous covariate

$$\begin{aligned}\text{logit}(\overset{\text{Male}}{\overset{\uparrow}{p}}_{M,D} - \text{logit}(\overset{\text{Female}}{\overset{\uparrow}{p}}_{F,D} &= \beta_0 + \beta_1 \overset{\text{Male}}{\overset{\downarrow}{(1)}} + \beta_2 D - (\beta_0 + \beta_1 \overset{\text{Female}}{\overset{\downarrow}{(0)}} + \beta_2 D) \\ &= \beta_1\end{aligned}$$

thus, the OR of dying as a male moth relative to a female moth is

$\exp(\beta_1)$

i.e. the odds of dying for males is about $\underline{\exp(\beta_1)}$ times the odds for females.

- Confidence interval for β_1 e.g., Wald CI: $\hat\beta_1 \pm 1.96\, SE(\hat\beta_1)$
- Confidence interval for the odds ratio (OR) is.. $\exp(\hat\beta_1 \pm 1.96\, SE(\hat\beta_1))$

Budworm Example (cont'd): Probability

- Estimation (or prediction) of the probability of dying at certain levels of the covariates

$$\text{logit}(p_{ij} = \eta = \beta_0 + \beta_1 S + \beta_2 D.$$

(1) Estimate of log-odds at female moth to die at dose 27 (i.e. $\log(27) = 3.3$).

$$\hat\eta = \hat\beta_0 + \hat\beta_1(S = 0) + \hat\beta_2(D = 3.3).$$

- Confidence interval for η (log-odds, here)

$$\hat\eta = \mathbf{C}\hat\beta,$$

where $\mathbf{C} = (1, S, D) = (1, 0, 3.3)$. Using Wald Statistic,

$$(\mathbf{C}\hat\beta - \mathbf{C}\beta)^T\{\hat V(\mathbf{C}\hat\beta)\}^{-1}(\mathbf{C}\hat\beta - \mathbf{C}\beta) \sim \chi^2_1$$

Budworm Example (cont'd): Model Comparison

- Consider an interaction model Testing all parameters are equal

(1) Test $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3$

$$\Leftrightarrow \mathbb{C}\beta = 0, \quad \mathbb{C} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \hat{V}(\hat\beta) = \begin{pmatrix} 0.30548 & & & * \\ & 0.60577 & & \\ & & 0.05812 & \\ * & & & 0.15772 \end{pmatrix}$$

Then $(\mathbb{C}\hat\beta - \mathbb{C}\beta)^T \hat{V}(\hat\beta)^{-1} (\mathbb{C}\hat\beta - \mathbb{C}\beta) = (\mathbb{C}\hat\beta - 0)^T \hat{V}(\hat\beta)^{-1} (\mathbb{C}\hat\beta - 0) \sim \chi^2(3)$

(2) Test $H_0 \quad \beta_3 = 0$.

To compare the main effects model and the interaction model and test importance of covariates.

Wald test: $\hat\beta_3 = 0.5091$, $SE(\hat\beta_3) = 0.3895$, 95% CI: $(-0.2543, 1.2726)$

H_0: $\beta_3 = 0$, $\chi^2_{calc}(1) = 1.71$, p-Value $= 0.1912$, non-significant

Deviance : $D_2 - D_1 = 6.7571 - 4.9937 = 1.7634 \approx LR$ (In fact, they are equivalent).

~~Pearson χ^2. $\chi^2_{calc}(1) = 5.3060 - 3.5047 = 1.8013$, non-sig~~

$LR = 2[(-105.7388) - (-106.6204)] = 2(0.8816) = 1.7632$ non-significant

Here, when $\phi = 1$, $D^*(Y, \hat\mu) = D(Y, \hat\mu)/\phi = D(Y, \hat\mu)$, so $LR = Dev.$

Budworm Example (cont'd): GOF

- Using Deviance, For model with interaction

$$D(\mathbf{Y}, \hat\mu) = 4.9937, \quad df = 12 - 4 = 8$$

Compare this model with the saturated model with the n parameters, the test is not significant, indicating a good fit

- Using Pearson's χ^2

$$X^2 = \sum_{i=1}^{n} \frac{(\tilde{Y}_i - \hat\mu_i)^2}{V(\hat\mu_i)} = 3.5047, \quad df = 8,$$

Conclusion is the same

where $\tilde{Y}_i = Y_i/m$ is the response variable for Binomial proportion.

For binomial distribution, $V(\hat\mu_i) = \hat\mu_i(1-\hat\mu_i)/m_i$,

$$\chi^2 = \sum_i \frac{(y_i - m_i\hat\mu_i)^2}{m_i(1-\hat\mu_i)\hat\mu_i}$$

Thus, the CI for η is

$$\hat{\eta} - \sqrt{\chi^2_{0.05,1}\hat{V}(\mathbf{C}\hat{\beta})} \le \eta \le \hat{\eta} + \sqrt{\chi^2_{0.05,1}\hat{V}(\mathbf{C}\hat{\beta})}.$$

(2) Estimation of p

$$\hat{p} = \frac{\exp(\hat{\eta})}{1+\exp(\hat{\eta})}$$

and the CI for p is

$\because$ $\eta = C\beta$, $Var(\hat{\eta}) = \hat{V}(C\hat{\beta}) = C\,\hat{V}(\hat{\beta})\,C^T$. $\hat{V}(\hat{\beta})$ can be obtained from SAS budworm.sas, by model ndead/ntotal=Sex logdose/ dist=binomial Covb;

i.e., $$\hat{V}(\hat{\beta}) = \begin{pmatrix} 0.21951 & -0.09875 & -0.07575 \\ -0.09875 & 0.12601 & 0.01875 \\ -0.07575 & 0.01875 & 0.03576 \end{pmatrix}.$$

$\therefore$ 95% CI for $\eta = C\beta$ is $\hat{\eta} \pm \sqrt{\chi^2_{0.05,1}\,\hat{V}(C\hat{\beta})} = \hat{\eta} \pm 1.96\sqrt{C\,\hat{V}(\hat{\beta})\,C^T}$

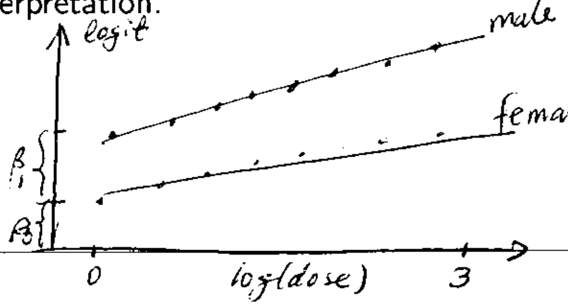
and 95% CI for p is $$\frac{\exp(\hat{\eta} - 1.96\sqrt{C\,\hat{V}(\hat{\beta})\,C^T})}{1+\exp(\hat{\eta} - 1.96\sqrt{C\,\hat{V}(\hat{\beta})\,C^T})} < p < \frac{\exp(\hat{\eta} + 1.96\sqrt{C\,\hat{V}(\hat{\beta})\,C^T})}{1+\exp(\hat{\eta} + 1.96\sqrt{C\,\hat{V}(\hat{\beta})\,C^T})}$$

Budworm Example (cont'd): Interaction

- Consider a model with interaction

$$\text{logit}(p) = \eta = \beta_0 + \beta_1 S + \beta_2 D + \beta_3 S \times D.$$

- Interpretations of parameters.
  - β_0: log-odds at $S = 0$ (Female) and $D = 0$ (dose = 1)
  - β_1 log-OR of dying as a male ($S=1$) moth relative to a female.
  - β_3: Interaction effect; log-OR of dying as a male moth relative to a female moth at a fixed D level when $D=1$
- Graphical interpretation.



$\beta_3 \neq 0$ implies two lines are not parallel