BIOSTAT 651
Notes #9: Logistic Regression

- Lecture Topics:
  - Logistic model
  - Parameter estimation & Inference
  - Saturated model
  - Goodness of fit
- Text (Dobson & Barnett, 2nd Ed.): Chapter 7

Logistic Regression: Set-Up

- Assume that we have the following set-up:
  - response, Y_i can take values from 0 to n_i
  - observed data: $(\mathbf{x}_i, Y_i)$ for $i = 1, \ldots, n$
  - pairs $(\mathbf{x}_i, Y_i)$ are independent
- GLM
  - Systematic component:

$$g(\pi_i) = \log\left\{\frac{\pi_i}{1-\pi_i}\right\} \quad = \quad \mathbf{x}_i^T\boldsymbol{\beta}$$

  - Random component:

$$Y_i \quad \sim \quad Binomial(n_i, \pi_i)$$

Logistic Regression: Set-Up

- Group level
  - Group level covariates (ex. categorical covariates)
  - ex. 2x2 table (treatment and placebo groups)
  - Y_i: number of subjects with events in each group

    $$Y_i = 0, \ldots, n_i$$

  - $n_i \geq 1$
- Individual level
  - Individuals can have different patterns of covariates (ex. continuous covariates).
  - Y_i: indicator of event for each subject.

    $$Y_i = 0, 1$$

  - $n_i = 1$

Logistic Regression: Set-Up

- Group level: Pneumonia data

Pneumonia (y_i)	n_i	Dust exposure (year)
1	98	5.8
1	54	15.0
3	43	21.5
⋮	⋮	⋮

Logistic Regression: Set-Up

- Individual level: Low Birth Weight data

LBW (y_i)	Mother age	race
0	19	black
0	20	white
1	25	other
$\vdots$	$\vdots$	$\vdots$

Logistic Regression as a GLM

- The logistic model is a special case of a GLM
  - link function:
  - mean function:
  - variance function:

Logistic Regression: Measures

- Disease frequency measures (recall):
  - risk:
  - odds:
  - logit:
- Logistic regression is referred to as *log-odds* model

Interpretation of Parameters

- Consider a *simple logistic regression model*,

$$logit(\pi_i) = \beta_0 + \beta_1 X_i,$$

  where (for now) X_i is continuous

- Interpretation of β_0:

Interpretation of Parameters (continued)

- Interpretation of β_1
- Difference in logit:

  $\beta_1 =$

- Exponentiate:

  $\exp\{\beta_1\} =$

Logistic Regression: Multiple Covariates

- Interpretations are as before, but with *all other covariates held constant*
- e.g., Suppose the covariates are

  M_i =Male indicator

  A_i =Age

  W_i =Weight

- Model:

$$logit(\pi_i) = \beta_0 + \beta_1 M_i + \beta_2 A_i + \beta_3 W_i$$

  - $\exp\{\beta_1\} =$

  - $\exp\{\beta_2\} =$

Parameter estimation: Saturated model

- A *saturated model* contains as many parameters as there are ...

- For grouped data with the saturated model, we can estimate β analytically.

Example: 2x2 table

- Example: A study of childhood asthma sought to determine the role of gender in asthma incidence. Children enrolled in the study (n=100) were followed prospectively in order to determine whether or not they were hospitalized for asthma between birth and the attainment of age 4.

Parameter estimation: Saturated model (continued)

The observed data:

	Y_i=0	Y_i=1	total
F_i=0	24	36	60
F_i=1	21	19	40
total	45	55	100

- The model is given by:

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \beta_1 F_i$$

- We have two samples, so the saturated model has two parameters.

$$\widehat{\beta}_0 =$$

$$\widehat{\beta}_0 + \widehat{\beta}_1 =$$

$$\widehat{\beta}_1 =$$

Odds Ratio as a Cross-Product

- Note that the MLE of the odds ratio equals that obtained through the standard cross-product calculation
  - i.e., based on previous calculations: $\exp\{\widehat{\beta}_1\} = \exp\{-0.5056\} = 0.603$
  - and, based on cross-product:

$$\widehat{OR}_F = \frac{24 \cdot 19}{36 \cdot 21} = 0.603$$

Saturated Model Example: Reparametrization

- Suppose we re-parameterized the model as follows:

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} \quad = \quad \beta_M(1-F_i) + \beta_F F_i$$

  - $\widehat{\beta}$:

$$\begin{aligned} \widehat{\beta}_M &= \\ \widehat{\beta}_F &= \end{aligned}$$

Example: Likelihood Calculations

- Note that the previously listed parameter estimates can always be obtained through standard likelihood calculations
- e.g., if we work with the re-parameterized model,

	Y_i=0	Y_i=1	total
F_i=0	24	36	60
F_i=1	21	19	40
total	45	55	100

with cell probabilities:

Example: Likelihood Calculations (continued)

- Likelihood,

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \left\{\frac{1}{1+e^{\beta_M}}\right\}^{24}\left\{\frac{e^{\beta_M}}{1+e^{\beta_M}}\right\}^{36} \\
&\quad \times \left\{\frac{1}{1+e^{\beta_F}}\right\}^{21}\left\{\frac{e^{\beta_F}}{1+e^{\beta_F}}\right\}^{19} \\
&= e^{36\beta_M}(1+e^{\beta_M})^{-60}e^{19\beta_F}(1+e^{\beta_F})^{-40}
\end{aligned}
$$

- Log likelihood,

$$\ell(\boldsymbol{\beta}) = 36\beta_M - 60\log(1+e^{\beta_M}) + 19\beta_F - 40\log(1+e^{\beta_F})$$

- Score function,

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta_M} &= 36 - 60\frac{e^{\beta_M}}{1+e^{\beta_M}} \\
\frac{\partial \ell}{\partial \beta_F} &= 19 - 40\frac{e^{\beta_F}}{1+e^{\beta_F}}
\end{aligned}
$$

- Solving score equation,

$$\begin{aligned} e^{\beta_M} &= \frac{36}{24} \\ e^{\beta_F} &= \frac{19}{21} \end{aligned}$$

- Computing MLEs,

$$\begin{aligned} \widehat{\beta}_M &= 0.4055 \\ \widehat{\beta}_F &= -0.1001 \end{aligned}$$

- i.e., the same estimates obtained by exploiting the *saturated* property of the model

GLM: Maximum Likelihood

- We already derived the score and information functions for the special case where:
  - $Y_i \sim$ exponential family
  - GLM is assumed
  - canonical link
- GLM: Score and Fisher information

$$U(\boldsymbol{\beta}) =$$

$$J(\boldsymbol{\beta}) =$$

Logistic Regression: MLE Methods

- Applying these general results to the case where $Y_i \sim$ Binomial (n_i, π_i) with

$$\pi_i = \pi(\mathbf{x}_i) \quad = \quad \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

  - Link function:

$\eta_i =$

$U(\boldsymbol{\beta}) =$

$J(\boldsymbol{\beta}) =$

Logistic Model: MLE

- Naturally, $U(\boldsymbol{\beta})$ and $J(\boldsymbol{\beta})$ can always be derived from likelihood function
  - Likelihood, log likelihood:

$$\begin{aligned} L_i(\boldsymbol{\beta}) &= \pi_i^{Y_i}\,(1-\pi_i)^{n_i - Y_i} \\ \ell_i(\boldsymbol{\beta}) &= Y_i \log \pi_i + (n_i - Y_i)\log(1-\pi_i) \end{aligned}$$

  - Score function,

$$\begin{aligned} U_i(\boldsymbol{\beta}) &= \frac{\partial \ell_i}{\partial \pi_i}\,\frac{\partial \pi_i}{\partial \boldsymbol{\beta}} \\ &= \left\{\frac{Y_i}{\pi_i} - \frac{n_i - Y_i}{1-\pi_i}\right\} \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i^T \boldsymbol{\beta}})^2}\mathbf{x}_i \\ &= \{Y_i(1-\pi_i) - (n_i - Y_i)\pi_i\}\mathbf{x}_i \\ &= (Y_i - n_i\pi_i)\mathbf{x}_i \end{aligned}$$

Logistic Model: MLE (continued)

- Information matrix,

$$
\begin{aligned}
J_i(\boldsymbol{\beta}) &= -\frac{\partial U_i}{\partial \pi_i} \, \frac{\partial \pi_i}{\partial \boldsymbol{\beta}^T} \\
&= \mathbf{x}_i n_i \pi_i (1 - \pi_i) \mathbf{x}_i^T
\end{aligned}
$$

Hypothesis Testing: Logistic Regression

- Suppose that the (full) model is given by:

$$\begin{aligned} \log\left\{\frac{\pi_i}{1-\pi_i}\right\} &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq} \\ &= \mathbf{x}_i^T \boldsymbol{\beta} \\ \boldsymbol{\beta} &= (\beta_0, \beta_1, \ldots, \beta_q)^T \end{aligned}$$

- Wald test: General form,
  - H_0 :
  - Test statistic:

- Special case of Wald test: $H_0 : \beta_j = 0$
  - set $\mathbf{C} =$
  - test statistic reduces to:

  - such tests are given by PROCs LOGISTIC and GENMOD for $j = 0, \ldots, q$

Likelihood Ratio Test

- Likelihood ratio test:

$$2\{\ell(\widehat{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}}^0)\}$$

  - can be carried out by fitting model twice

  - also available through difference of Deviances:

$$D_0 - D_1$$

Goodness of Fit

- Deviance and Pearson χ^2 for the binomial data.

$$
\begin{aligned}
D &= 2\sum_{j=1}^{n}\left[Y_i \log\left(\frac{Y_i}{n_i\widehat{\pi}_i}\right)\right. \\
&\qquad \left. + (n_i - Y_i) \log\left(\frac{n_i - Y_i}{n_i - n_i\widehat{\pi}_i}\right)\right] \\
X_p^2 &= \sum_{i=1}^{n}\frac{(Y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)}
\end{aligned}
$$

- Both deviance and Pearson χ^2 approximately follow χ^2_{n-q}

Goodness of Fit

- Deviance and Pearson χ^2 work well when the expected number of events (and non-events) > 5
- When n_i is small, they don't work well.
- SAS does not provide Deviance (and Pearson χ^2) when $n_i = 1$

Goodness of Fit: Hosmer and Lemeshow test

- Group subjects based on fitted risk values.
- Based on groups, carry out Pearson χ^2 test.
- HL test statistic

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)}$$

  - O_g: number of observed event in the gth risk group
  - E_g: number of expected event
  - N_g: number of observations
  - π_g: predicted risk
- H asymptotically follows a χ^2 distribution with $G - 2$ degrees of freedom.

R^2

- R^2 = Explained variation / Total variation.
- Intercept only model ($\widehat{\pi}_i^{intercept}$):

$$logit(\pi_i) = \beta_0$$

- Pseudo R^2 (Cox & Snell)

$$R^2 = 1 - \left\{ \frac{L(\widehat{\pi}_i^{intercept})}{L(\widehat{\pi})} \right\}^{2/N}$$

  - Improvement from the intercept only model to fitted model.
  - In linear regression, Pseudo R^2 yields the classical R^2.

- Max adjusted R^2 (Nagelkerke)
  - Maximum of the Cox & Snell R^2 can be smaller than 1.
  - NagelKerke proposed a max adjusted Cox & Snell R^2.

$$\text{max-adjusted } R^2 = \frac{R^2}{maxR^2}$$

- There are many different versions of pseudo R^2. In the book, McFadden R^2 is introduced.
- Cox & Snell R^2 and Max adjusted Cox & Snell R^2 are implemented in SAS.

Residuals

- Pearson residuals

$$\widehat{r}_i^P = \frac{Y_i - n_i\widehat{\pi}_i}{\sqrt{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)}}$$

- Deviance residuals

$$\widehat{r}_i^D = sign(Y_i - n_i\widehat{\pi}_i)\sqrt{|D_i|}$$