Binary Response:

Logistic Regression

STAT 526
Professor Olga Vitek

March 29, 2011

Model Specification and Interpretation

Probability Distribution of a Binary Outcome Y

- In many situations, the response variable has only two possible outcomes
  - Disease ($Y = 1$) vs Not diseased ($Y = 0$)
  - Employed ($Y = 1$) vs Unemployed ($Y = 0$)
- Response is *binary or dichotomous*
- Can model response using Bernoulli dist

Y_i	Probability
1	$\Pr\{Y_1 = 1\} = \pi_i$
0	$\Pr\{Y_1 = 0\} = 1 - \pi_i$

- $E\{Y_i\} = \pi_i$
- $Var\{Y_i\} = \pi_i(1 - \pi_i)$

Goal: express $E\{Y\}$ as function of a covariate X

- The simple regression is not appropriate

$$E\{Y_i\} = \beta_0 + \beta_1 X_i$$

It violates several assumptions:

(1) Does not enforce the constraint
$0 \leq E\{Y_i\} \leq 1$ is

(2) Non-normal (binary) distribution of $\varepsilon \mid X$:

$$\begin{aligned} \text{When } Y_i = 0 &: \varepsilon_i = 0 - \beta_0 - \beta_1 X_i \\ \text{When } Y_i = 1 &: \varepsilon_i = 1 - \beta_0 - \beta_1 X_i \end{aligned}$$

(3) Non-constant variance

$$\begin{aligned} Var\{Y_i\} &= \pi_i(1-\pi_i) \\ &= (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i) \end{aligned}$$

Solution: a Generalized Linear Model

- A generalized linear model is

$$
\begin{aligned}
E\{Y_i\} &= g(\beta_0 + \beta_1 X_i), \text{ or} \\
g^{-1}(E\{Y_i\}) &= \beta_0 + \beta_1 X_i
\end{aligned}
$$

  where g is a sigmoid function in (0,1).

  - g is called the *mean response function*
  - g^{-1} is called the *link function*

- A choice of g produces different models
  - $g(t) = \text{Identity}$
    $\rightarrow$ linear regression
  - $g(t) = \Phi(t) = \text{standard Normal CDF}$
    $\rightarrow$ probit regresison
  - $g(t) = \frac{\exp(t)}{1+exp(t)} = \text{CDF of the logistic distrib.}$
    $\rightarrow$ logistic regresison

Motivation for Probit Regression: Latent Variable

- Assume that the binary response is guided by a non-observed continuous variable

- Example: linear model for blood pressure:

$$\text{bp} = \beta_0 + \beta_1 \text{age} + \varepsilon$$

Only observe

$$Y = \begin{cases} 1 \text{ (disease)}, & \text{if blood pressure} > c \\ 0 \text{ (healthy)}, & \text{if blood pressure} \leq c \end{cases}$$

$$\begin{aligned} & \text{Pr}\{Y = 1\} \\ = & \quad \text{Pr}\{\text{bp} > c\} = \text{Pr}\{\beta_0 + \beta_1 \text{age} + \varepsilon > c\} \\ = & \quad \text{Pr}\{\varepsilon < \beta_0 + \beta_1 \text{age} - c\} \\ = & \quad \text{Pr}\{\frac{\varepsilon}{\sigma} < \frac{\beta_0 - c}{\sigma} + \frac{\beta_1}{\sigma}\text{age}\} \\ = & \quad \text{Pr}\{z < \beta_0' + \beta_1' \text{age}\} \\ = & \quad \Phi(\beta_0' + \beta_1' \text{age}) \end{aligned}$$

Logistic Response Function and Logistic Regression

- A sigmoidal response function

$$E\{Y_i\} = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i))}$$

  - A monotonic increasing/decreasing function
  - Explicit functional form
  - Restricts $0 \leq \mathrm{E}(Y_i) \leq 1$
  - Example of a **nonlinear** model

- **Logit** link function

$$\log\left(\frac{E\{Y_i\}}{1 - E\{Y_i\}}\right) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_i$$

Probability Distribution of Y in Logistic Regression

- Y_i are independent but not identically distributed Bernoulli random variables

$$
\begin{aligned}
Y_i &\overset{ind}{\sim} \text{Bernoulli}(\pi_i) \text{ where} \\
\pi_i &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}
\end{aligned}
$$

  - note no more error term!

- Probability density of Y_i

$$f(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$$

- Least Squares Estimates are inappropriate

  - use maximum likelihood for parameter estimation

Simple Logistic Regression: Interpretation of b_1

- Fitted value for the individual i
  - $\widehat{\pi}_i = \frac{e^{b_0+b_1X_i}}{1+e^{b_0+b_1X_i}}$

- Fitted logistic response (i.e. fitted log odds)
  - $\log_e\left(\widehat{\text{odds}}(X_i)\right) = \log_e \frac{\widehat{\pi}_i(X_i)}{1-\widehat{\pi}_i(X_i)}$
    $= b_0 + b_1X_i$
  - b_1 is the slope of the fitted logistic response

- b_1 is interpreted as log(odds ratio)
  - $$\begin{aligned} b_1 &= \log_e\left(\frac{\widehat{\pi}_i(X_i+1)}{1-\widehat{\pi}_i(X_i+1)}\right) - \log_e\left(\frac{\widehat{\pi}_i(X_i)}{1-\widehat{\pi}_i(X_i)}\right) \\ &= \log_e\left(\frac{\text{odds}(X_i+1)}{\text{odds}(X_i)}\right) \end{aligned}$$
  - $\widehat{\text{odds ratio}}(X_i) = \exp(b_1)$

Estimation by Maximum Likelihood

- $Y_i \overset{ind}{\sim}$ Bernoulli(π_i) where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

- Log likelihood: $\log_e(L) =$

$$\begin{aligned}
&= \log\left\{\prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}\right\} \\
&= \sum_{i=1}^{n} Y_i \log(\pi_i) + \sum_{i=1}^{n} (1-Y_i)\log(1-\pi_i) \\
&= \sum_{i=1}^{n} Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \log(1-\pi_i) \\
&= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log(1 + \exp(\beta_0 + \beta_1 X_i))
\end{aligned}$$

- MLEs do not have closed forms

Equivalent specification: Binomial distribution

- Change in notation
  - Data: (Y_{ij}, n_i, X_i), $i = 1, 2, \cdots, c$
  - X_i : predictor for observation i
  - n_i : # of Bernoulli trials in observation i
  - $Y_i' := \sum_{j=1}^{n_i} Y_{ij}$
  - Model:

$$\begin{aligned} Y_i' &\overset{ind}{\sim} Binomial(n_i, \pi_i), \text{ where} \\ \pi_i &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \end{aligned}$$

- Log-Likelihood: $\log_e(L) =$

$$\begin{aligned} &= \log \prod_{i=1}^{c} \left\{ \binom{n_i}{Y_i'} \pi^{Y_i'} (1 - \pi_i)^{n_i - Y_i'} \right\} \\ &= \sum_{i=1}^{c} \left\{ Y_i' \log(\pi_i) + (n_i - Y_i') \log(1 - \pi_i) + \log \binom{n_i}{Y_i'} \right\} \\ &= \sum_{i=1}^{c} \left\{ Y_i' \log \frac{\pi_i}{1 - \pi_i} + n_i \log(1 - \pi_i) + \log \binom{n_i}{Y_i'} \right\} \end{aligned}$$

Equivalence of Bernouilli and Binomial Models

- Binomial Log-Likelihood equals Bernouilli Log-Likelihood, up to a constant:

$$\log_e(L)^{Binomial} =$$

$$\begin{aligned}
&= \sum_{i=1}^{c} \left\{ Y_i' \log(\pi_i) + (n_i - Y_i') \log(1-\pi_i) \right\} + constant \\
&= \sum_{i=1}^{c} \left\{ \sum_{j=1}^{n_i} Y_{ij} \log(\pi_i) + (n_i - \sum_{j=1}^{n_i} Y_{ij}) \log(1-\pi_i) \right\} + constant \\
&= \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\{ Y_{ij} \log(\frac{\pi_i}{1-\pi_i}) + \log(1-\pi_i) \right\} + constant
\end{aligned}$$

$$= \log_e(L)^{Bernouilli} + constant$$

- Both models lead to same parameter estimates and inferences, but have different deviances.

Prospective and Retrospective Studies

- Prospective study: fix predictors, observe the outcome
  - Recruit patients with 2 genotypes, compare occurrence of disease.
  - Expensive; large variance for rare diseases

- Retrospective (=case-control) study: fix outcome, observe predictors
  - Recruit patients with and without the disease, compare the genotypes.
  - Cheaper

- Log odds ratio based on logistic regression is the same for both studies.
  - Not true for other link functions

Prospective Model With Retrospective Data: Formulation

- Assume a prospective model:

$$\pi(X_i) = P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}$$

- However, the subjects are selected retrospectively.

- The random variable is $Z_i = 1_{\{\text{including subject } i\}}$, conditional on Y.
  - Denote
    $\theta_0 = P(Z_i = 1|Y_i = 0)$ and
    $\theta_1 = P(Z_1 = 1|Y_i = 1)$
  - Note that θ_0 and θ_1 are indenpent of X_i, otherwise introduce selection bias.

- Of interest in retrospective study is $P(Y_i = 1|X_i, Z_i = 1)$

Prospective Model With Retrospective Data: Estimation of log(OR)

- With retrospective sampling, we model:

$$
\begin{aligned}
&P(Y_i = 1|X_i, Z_i = 1) \\
&\quad = \frac{P(Y_i = 1, Z_i = 1|X_i)}{P(Z_i = 1|X_i)} \\
&\quad = \frac{P(Y_i = 1, Z_i = 1|X_i)}{P(Y_i = 0, Z_i = 1|X_i) + P(Y_i = 1, Z_i = 1|X_i)} \\
&\quad = \frac{\theta_1 \times \pi(X_i)}{\theta_0 \times \{1 - \pi(X_i)\} + \theta_1 \times \pi(X_i)} \\
&\quad = \frac{\theta_1 e^{\beta_0+\beta_1 X_i}}{\theta_0 + \theta_1 e^{\beta_0+\beta_1 X_i}} = \frac{e^{\log(\theta_1/\theta_0)+\beta_0+\beta_1 X_i}}{1 + e^{\log(\theta_1/\theta_0)+\beta_0+\beta_1 X_i}}
\end{aligned}
$$

- Uncover the same β_1 as in the prospective study

$$
\Longrightarrow log \left\{ \frac{P(Y_i = 1|X_i, Z_i)}{P(Y_i = 0|X_i, Z_i)} \right\} = [\log(\theta_1/\theta_0) + \beta_0] + \beta_1 X_i
$$

Multiple Logistic Regression

- Extension of the response function:

$$E\{Y_i\} = \frac{\exp(\mathbf{X}_i'\beta)}{1 + \exp(\mathbf{X}_i'\beta)}$$

- Extension of the logit link function:

$$\log_e\left(\frac{E\{Y_i\}}{1 - E\{Y_i\}}\right) = \log_e\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{X}_i'\beta$$

- The multivariate likelihood function:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\mathbf{X}_i'\beta) - \sum_{i=1}^{n} \log_e[1 + \exp(\mathbf{X}_i'\beta)]$$

- Interpretation of b_j:

$$log_e\left(\frac{\text{odds}(X_j + 1)}{\text{odds}(X_j)}\right)$$

while other predictors are held fixed

Testing

Asymptotic Properties of $\widehat{\beta}$

- Asymptotic existence and uniqueness:
  - $P\{\widehat{\beta}$ exists and is unique $\to 1\}$ as $n \to \infty$

- Consistency:
  - $\widehat{\beta} \to \beta$ as $n \to \infty$

- Asymptotic Normality:
  - $\widehat{\beta} \overset{Ass.}{\sim} \mathcal{N}(\beta, I(\widehat{\beta})^{-1})$ as $n \to \infty$

- Asymptotic efficiency:
  - The MLE has asymptotically smaller variance than many other estimators

Inference About Individual β_j: Wald Test

- Test $H_0: \ \beta_j = 0$ versus $H_a: \ \beta_j \neq 0$.

- Test statistic $z^* = \frac{b_j - 0}{s\{b_j\}}$

- **Approximate** variance $s^2\{\mathbf{b}\}$

$$s^2\{\mathbf{b}\} = \left(\left[- \frac{\partial^2 \log_e L(\beta)}{\partial \beta_j \partial \beta_{j'}} \right]_{\beta = \mathbf{b}} \right)^{-1}$$

- **Approximate** distribution of z
  - $z^* \sim \mathcal{N}(0, 1)$. Alternatively, $(z^*)^2 \sim \chi^2_1$
  - reject H_0 if $|z^*| > z^{1-\alpha/2}$
  - CI for β_j: $b_j \pm z^{1-\alpha/2} s\{b_j\}$

Simultaneous Inference About Several $\beta_j = 0$: Likelihood Ratio Test

- Multivariate logistic regression

  $$\log\left(\frac{E\{Y_i\}}{1-E\{Y_i\}}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

- Test $H_0: \ \beta_1 = \beta_2 = \cdots = \beta_q = 0$
  versus H_a: not all $\beta_1, \ \beta_2, \ \cdots, \ \beta_q = 0$

- Test statistic

  $$\begin{aligned} G^2 &= -2\log_e\left[\frac{L(\text{reduced model})}{L(\text{full model})}\right] \\ &= -2\left[\log_e L(\text{reduced model}) - \log_e L(\text{full model})\right] \end{aligned}$$

- **Approximate** distr of G^2 for large n
  - Reject H_0 if $G^2 > \chi^2(1-\alpha, q)$

Comments: Wald Test Versus Likelihood Ratio Test

- Both tests are approximate, for large n

- Likelihood Ratio test: $\beta_j = 0$, or several $\beta_j = 0$ simultaneously

  - no other values of β_j

  - no one-sided tests

- Wald test: $\beta_j = {\beta_j}^{of\ \ interest}$, or linear combinations of β_j

- When testing a single $H_0: \ \beta_j = 0$, the tests may lead to different conclusions

  - due to the approximate nature of the tests

  - unlike in linear regression

Quality of Fit: Replicated data

Lack of Fit for Replicated Data: Pearson χ^2

$$H_0: \quad \log\left(\frac{E\{Y_{ij}\}}{1-E\{Y_{ij}\}}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} \text{ vs}$$

$$H_a: \quad \log\left(\frac{E\{Y_{ij}\}}{1-E\{Y_{ij}\}}\right) \neq \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

- c covariate configurations, each with n_j cases
- Observed counts
  - O_{i0}: observed # of 0's in configuration i
  - O_{i1}: observed # of 1's in configuration i
- Expected counts
  - $E_{i0} = n_i(1-\hat{\pi}_i)$: expected # of 0's in conf. i
  - $E_{i1} = n_i(\hat{\pi}_i)$: expected # of 1's in conf. i
- Test statistic: reject H_0 if

$$X^2 = \sum_{i=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} > \chi^2(1-\alpha, c-p)$$

Lack of Fit for Replicated Data: Deviance

$$H_0: \quad \log\left(\frac{E\{Y_{ij}\}}{1-E\{Y_{ij}\}}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} \text{ vs}$$
$$H_a: \quad \log\left(\frac{E\{Y_{ij}\}}{1-E\{Y_{ij}\}}\right) \neq \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

- c covariate configurations, each with n_i cases
  - H_0: model of interest; H_a: saturated model
- Model of interest
  - $E\{Y_{ij}\} = \pi_i$; $\widehat{E\{Y_{ij}\}} = \widehat{\pi}_i$
- Saturated model
  - $E\{Y_{ij}\} = p_i$; $\widehat{E\{Y_{ij}\}} = \widehat{p}_i = \frac{\sum_j Y_{ij}}{n_i}$
- Test statistic: LR, also called deviance
- Deviance of a saturated model always = 0

Lack of Fit for Replicated Data: Deviance

$$
\begin{aligned}
G^2 &= DEV(X_0, X_1, \ldots, X_{p-1}) \\
&= -2\,[\ \log_e L(\text{current model}) - \log_e L(\text{saturated model})\ ] \\
&= -2\sum_{i=1}^{c}\left[\ \sum_{j=1}^{n_i} Y_{ij}\log_e \widehat{\pi}_i + (n_i - \sum_{j=1}^{n_i} Y_{ij})\log_e(1-\widehat{\pi}_i)\ \right] \\
&\quad +2\sum_{i=1}^{c}\left[\ \sum_{j=1}^{n_i} Y_{ij}\log_e \widehat{p}_i + (n_i - \sum_{j=1}^{n_i} Y_{ij})\log_e(1-\widehat{p}_i)\ \right] \\
&= -2\sum_{i=1}^{c}\left[\ \sum_{j=1}^{n_i} Y_{ij}\log_e\left(\frac{\widehat{\pi}_i}{\widehat{p}_i}\right) + (n_i - \sum_{j=1}^{n_i} Y_{ij})\log_e\left(\frac{1-\widehat{\pi}_i}{1-\widehat{p}_i}\right)\ \right]
\end{aligned}
$$

- Reject H_0 if $G^2 > \chi^2(1-\alpha, c-p)$
- Approximation $\chi^2(1-\alpha, c-p)$ can be poor
  - The closer the distribution to Gaussian, and the closer the link to identity, the better the approximation
  - Unlike with the LR test, the quality of approximation does not improve with the sample size

Note: Can Use Deviance for LR Test of Nested Models

$$\log\left(\frac{E\{Y_{ij}\}}{1-E\{Y_{ij}\}}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$$

- Test $H_0: \ \beta_1 = \beta_2 = \cdots = \beta_q = 0$
  versus H_a: not all $\beta_1, \ \beta_2, \ \cdots, \ \beta_q = 0$

- Likelihood Ratio test statistic $G^2 =$

$$\begin{aligned}
&= -2\log_e\left[\frac{L(\text{reduced model})}{L(\text{full model})}\right] \\
&= -2\left[\log_e L(\text{reduced model}) - \log_e L(\text{full model})\right] \\
&= -2\left[\log_e L(\text{reduced model}) - \log_e L(\text{saturated model})\right] \\
&\quad +2\left[\log_e L(\text{full model}) - \log_e L(\text{saturated model})\right] \\
&= Deviance(\text{reduced model}) - Deviance(\text{full model})
\end{aligned}$$

- Approximate distr of G^2 for large n

- Reject H_0 if $G^2 > \chi^2(1-\alpha, q)$

Quality of Fit:
Individual Observations

Non-Replicated Data: Hosmer-Lemeshow Goodness of Fit

$$
\begin{aligned}
H_0: &\quad \log\left(\frac{E\{Y_i\}}{1-E\{Y_i\}}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} \text{ vs} \\
H_a: &\quad \log\left(\frac{E\{Y_i\}}{1-E\{Y_i\}}\right) \neq \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}
\end{aligned}
$$

- Group cases based on values of estimated probabilities $\widehat{\pi}_i$ into c groups
  - E.g., find $c = 9$ groups based on percentiles

- Apply Pearson χ^2 test to the groups

- Reject H_0 if $X^2 > \chi^2(1-\alpha, c-2)$
  - showed by simulation that this distribution is appropriate

Can Write Pearson χ^2 (But Not Use for Tests)

- Test statistic:

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{c} \sum_{k=0}^{1} \frac{(O_{ik} - E_{ik})^2}{E_{ik}} \\
&= \sum_{i=1}^{c} \frac{(O_{i0} - E_{i0})^2}{E_{i0}} + \sum_{i=1}^{c} \frac{(O_{i1} - E_{i1})^2}{E_{i1}}
\end{aligned}
$$

- The corresponding quantities (KNNL p. 591):
  - $c = n$, $n_i = 1$
  - $O_{i0} = 1 - Y_i$, $O_{i1} = Y_i$
  - $E_{i0} = 1 - \widehat{\pi}_i$, $E_{i1} = \widehat{\pi}_i$

- Test statistic:

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{n} \frac{[(1 - Y_i) - (1 - \widehat{\pi}_i)]^2}{1 - \widehat{\pi}_i} + \sum_{i=1}^{n} \frac{(Y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i} \\
&= \sum_{i=1}^{n} \frac{(Y_i - \widehat{\pi}_i)^2}{1 - \widehat{\pi}_i} + \sum_{i=1}^{n} \frac{(Y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i} = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i(1 - \widehat{\pi}_i)}
\end{aligned}
$$

Can Write Deviance (But Not Use for Tests)

- Test statistic $G^2 = DEV(X_0, X_1, \ldots, X_{p-1})$:

$$= \; -2\sum_{i=1}^{c}\left[\; \sum_{j=1}^{n_i} Y_{ij}\log\left(\frac{\widehat{\pi}_i}{\widehat{p}_i}\right) + (n_i - \sum_{j=1}^{n_i} Y_{ij})\log\left(\frac{1-\widehat{\pi}_i}{1-\widehat{p}_i}\right)\;\right]$$

- The corresponding quantities (KNNL p. 592):
  - $c = n$, $n_i = 1$, $\sum_{j=1}^{n_i} Y_{ij} = Y_i$, $\widehat{p}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i = Y_i$

- Test statistic:

$$\begin{aligned}
G^2 &= -2\sum_{i=1}^{n}\left[\; Y_i\log\left(\frac{\widehat{\pi}_i}{Y_i}\right) + (1-Y_i)\log\left(\frac{1-\widehat{\pi}_i}{1-Y_i}\right)\;\right] \\
&= -2\sum_{i=1}^{n}[\; Y_i\log\widehat{\pi}_i + (1-Y_i)\log(1-\widehat{\pi}_i) \\
&\quad -Y_i\log Y_i - (1-Y_i)\log(1-Y_i)\;] \\
&= -2\sum_{i=1}^{n}[\; Y_i\log\widehat{\pi}_i + (1-Y_i)\log(1-\widehat{\pi}_i)\;]
\end{aligned}$$

Diagnostics: Residuals

- Logistic Regression Residuals

$$e_i = \begin{cases} 1 - \widehat{\pi}_i, & \text{if } Y_i = 1 \\ -\widehat{\pi}_i, & \text{if } Y_i = 0 \end{cases}$$

- Pearson residual

$$r_{P_i} = \frac{Y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1 - \widehat{\pi}_i)}}$$

  - e_i divided by the standard error of Y_i
  - $\sum_{i=1}^{n} r_{P_i}^2$ equals **non-replicated** Pearson X^2

- Studentized Pearson Residual

$$r_{P_i} = \frac{Y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1 - \widehat{\pi}_i) \cdot (1 - h_{ii})}}$$

  - e_i divided by the SE of $e_i \rightarrow$ unit variance
  - h_{ii} is the diagonal element of the hat matrix

  $\mathbf{H} = \widehat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{\frac{1}{2}}$, where

  $\widehat{\mathbf{W}} = \text{diag}\,(\widehat{\pi}_i(1 - \widehat{\pi}_i))$

Diagnostics: Residuals

- Deviance Residuals

$$
\begin{aligned}
d_i &= \text{sign}(Y_i - \widehat{\pi}_i)\sqrt{-2\left[Y_i \log \frac{\widehat{\pi}_i}{Y_i} + (1 - Y_i)\log \frac{1-\widehat{\pi}_i}{1-Y_i}\right]} \\
&= \text{sign}(Y_i - \widehat{\pi}_i)\sqrt{-2\left[Y_i \log \widehat{\pi}_i + (1 - Y_i)\log(1-\widehat{\pi}_i)\right]}
\end{aligned}
$$

  - the signed square root of the contribution of Y_i to the model deviance

- Analysis of residuals

  - unknown distribution of residuals under true model

  - plot residual by predicted value. A flat lowess smooth to this plot suggests good model

- Other summaries as in linear regression

  - DFFITS, DFBETAS
  - $\Delta\chi^2$, Δdev, Cook's distance (see KNNL p. 598)

Graphical Check of the Fit

- Partition the observations into groups of covariate patterns x_i.

- Haldane (1956) recommended to plot
$$\widehat{\eta}_i = \log \frac{y_i + 0.5}{n_i - y_i + 0.5}$$
against covariate patterns $\mathbf{x}_i$.

- The plot should be roughly linear if the model is appropriate for the data

- When all $n_i = 1$ or all n_i are small, one can group the data with nearby x values to make the plot

Overdispersion

Overdispersion

$$Y \overset{ind}{\sim} Binomial(n, \pi), \ \pi = \frac{\exp(\mathbf{X}'\beta)}{1 + \exp(\mathbf{X}'\beta)}$$

- Implies $E\{Y\} = \pi$, $Var\{Y\} = n\pi(1 - \pi)$
  - Overdispersion: $Var\{Y\} > n\pi(1 - \pi)$
  - Underdispersion: $Var\{Y\} < n\pi(1 - \pi)$

- Mechanisms of overdispersion:
  - Suppose $Y_1, Y_2, \ldots, Y_n$ are Bernoulli r.v., $E\{Y_i\} = \pi$.
  - define $Y = \sum_{i=1}^{n} Y_i$
  - Can think of at least two situations when Y does not have a Binomial distribution (and therefore a different variance)

Overdispersion from correlation

- Suppose $Y_1, Y_2, \ldots, Y_n$ are not independent
- Suppose all pairs (Y_i, Y_j) have a same correlation ρ

$$
\begin{aligned}
Var(Y) &= Cov\left(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i\right) \\
&= \sum_{i=1}^{n} Var(Y_i) + \sum_{i \neq j} Corr(Y_i, Y_j)\sqrt{Var(Y_i)}\sqrt{Var(Y_j)} \\
&= n\pi(1-\pi) + n(n-1)\rho\pi(1-\pi) \\
&> n\pi(1-\pi)
\end{aligned}
$$

- The variance exceeds the variance of the Binomial distribution

Overdispersion from clustered data

- Suppose $Y = \sum_{i=1}^{n} Y_i \mid \pi \sim Binomial(\pi)$

- Suppose π is a random variable, $E\{\pi\} = p$, $Var\{\pi\} = p(1-p)$

  – Special case: $\pi \sim Beta(\alpha, \beta)$

$$
\begin{aligned}
E\{Y\} &= E\{E\{Y \mid \pi\}\} = p \\
Var\{Y\} &= Var\{E\{Y|\pi\}\} + E\{Var\{Y|\pi\}\} \\
&= Var\{n\pi\} + E\{n\pi(1-\pi)\} \\
&= n^2 Var\{\pi\} + np - n[Var\{\pi\} + p^2] \\
&= np(1-p) + n(n-1)Var\{\pi\} \\
&> np(1-p)
\end{aligned}
$$

- The variance exceeds the variance of the Binomial distribution

Modeling Overdispersion

- Introduce additional parameter

$$E\{Y_i\} = n_i\pi_i,\ Var\{Y_i\} = \phi n_i\pi_i\{1-\pi_i\}$$

  - When $\phi \neq 1$, Y_i follows a *quasi-binomial distribution*. The distribution is characterized by its expectation and variance. The probability distribution function is unspecified.

- $\widehat{\phi} = \chi^2/(n-p)$
  - χ^2 is the Pearson lack of fit statistic (= sum of squared Pearson residuals with **non-replicated** data)
  - p is the number of parameters in the model
  - $\widehat{\phi} >> 1$ indicates evidence of overdispersion.

- Since ϕ does not affect $E\{Y_i\}$, modeling overdispersion does not change $\widehat{\beta}$.

- $SE\{\widehat{\beta}\}$ is multiplied by $\sqrt{\widehat{\phi}}$.

Comparing Nested Models in Presence of Overdispersion

- Regular likelihood-based approaches (e.g. LRT, AIC) are not applicable.

- F test approximates deviance-based LR test

$$F = \frac{D_{reduced} - D_{full}}{df_{reduced} - df_{full}} / \widehat{\phi} \overset{ass., \; H_0}{\sim} F_{df_{reduced} - df_{full}, df_{full}}$$

- Assumes roughly equal covariate classes

- Modeling strategy:
  - fit the full model (with all predictors)
  - estimate $\widehat{\phi}$
  - compare nested models with F test to reduce the number of predictors

Prediction
and
Classification

Prediction of the Mean

- Point estimate for the link $\widehat{\pi}'_h$:

$$\widehat{\pi}'_h = \mathbf{X}'_h \mathbf{b}$$

- Point estimate for the response $\widehat{\pi}_h$:

$$\widehat{\pi}_h = \frac{1}{1+\exp(-\widehat{\pi}'_h)} = \frac{1}{1+\exp(-\mathbf{X}'_h \mathbf{b})}$$

- Interval estimate for the link $\widehat{\pi}'_h$:

$$s^2\{\widehat{\pi}'_h\} \;=\; \mathbf{X}'_h s^2\{\mathbf{b}\}\mathbf{X}'_h$$

$$(1-\alpha)\%\text{ CI for } \widehat{\pi}'_h \quad : \quad \widehat{\pi}'_h \pm z^{1-\alpha/2} s\{\widehat{\pi}'\} \;= (L,\ U)$$

- Approximate interval estimate for the response $\widehat{\pi}'_h$:

$$(1-\alpha)\%\text{ CI for } \widehat{\pi}_h \quad : \quad \left(\frac{1}{1+\exp(-L)},\ \frac{1}{1+\exp(-U)}\right)$$

- Use Bonferroni if multiple $\mathbf{X}$ are of interest

Measures of Agreement

- Have N observations
- Consider all pairs of distinct responses
  - In this example $t = 16 \times 14 = 224$
- Compare predicted probabilities
  - Concordant if $\widehat{\pi}_{Y=1} > \widehat{\pi}_{Y=0}$
  - Discordant if $\widehat{\pi}_{Y=1} < \widehat{\pi}_{Y=0}$
  - Tie if $\widehat{\pi}_{Y=1} = \widehat{\pi}_{Y=0}$
- Measures of agreement
  - Somers' D : (#C - #D)/t
  - Goodman-Kruskal Gamma : (#C - #D)/(#C + #D)
  - Kendall's Tau-a : (#C - #D)/(.5N(N-1))
  - c : (#C +.5(t-#C-#D))/t

Prediction of a New Observation (i.e. Classification)

- Choose a cutoff $c \in (0, 1)$

$$\widehat{Y}_h = \begin{cases} 1, & \text{if } \widehat{\pi}_h > c \\ 0, & \text{if } \widehat{\pi}_h \leq c \end{cases}$$

  – Sensitivity:

$$\frac{\text{\# predicted '1' \& true '1'}}{\text{\# true '1'}} = \frac{\sum_{i=1}^{n} \widehat{Y}_i \cdot Y_i}{\sum_{i=1}^{n} Y_i}$$

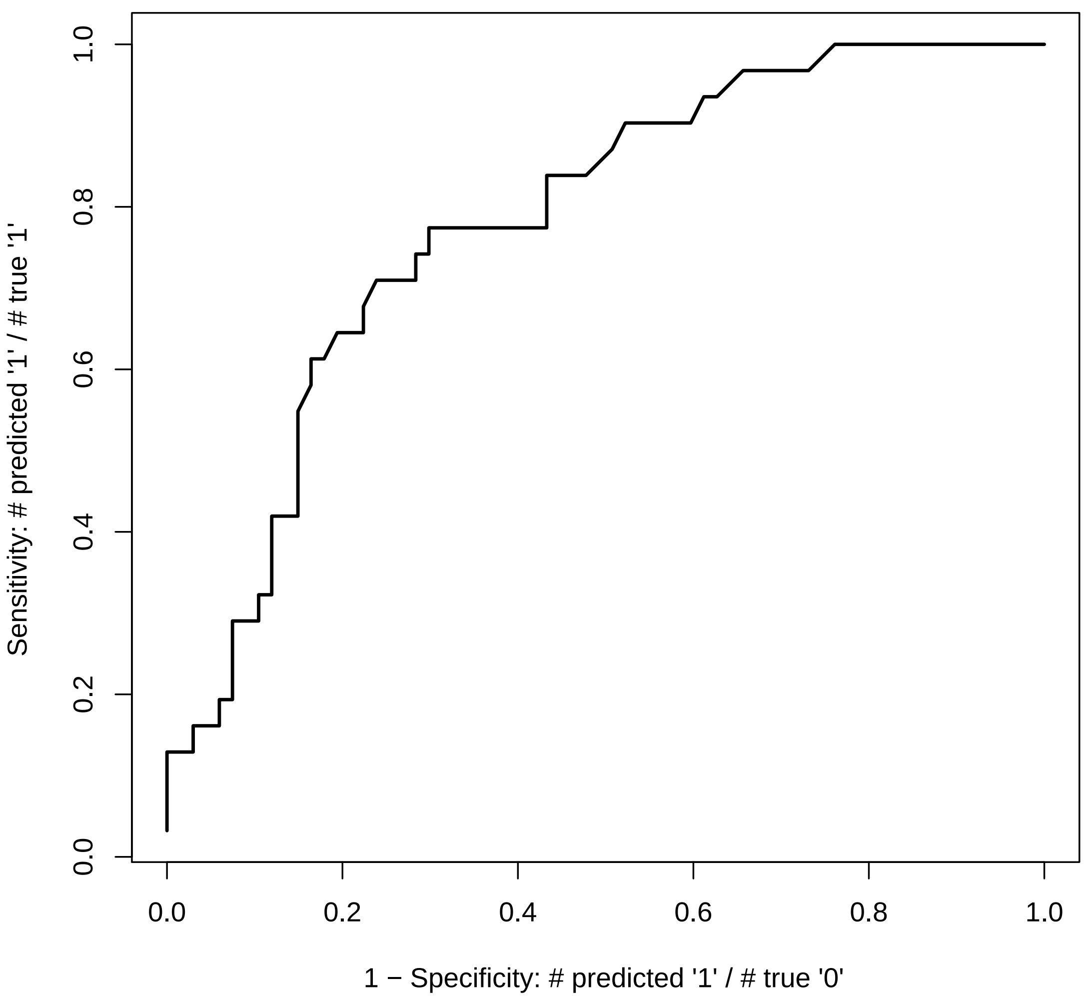
  – Specificity:

$$\frac{\text{\# predicted '0' \& true '0'}}{\text{\# true '0'}} = \frac{\sum_{i=1}^{n} (1-\widehat{Y}_i) \cdot (1-Y_i)}{\sum_{i=1}^{n} 1-Y_i}$$

- Vary the cut-off $c \in (0, 1)$, and choose c to optimize sensitivity and specificity

ROC curve for classification

Vary c, and plot sensitivity vs 1-specificity.
Evaluate models by area under the curve.



Evaluation of the Predictive Ability of the Model

- Area under ROC can be used to compare models
  - Area = 1 $\rightarrow$ perfect classification
  - Area = .5 $\rightarrow$ random classification
- Classification on the training set is overly optimistic
- Use cross-validation to construct a more accurate ROC curve

Variable Selection

Automatic Variable Selection

- Exhaustive search. Minimize:

$$
\begin{aligned}
& -2\log_e L(\mathbf{b}) \\
AIC_p &= -2\log_e L(\mathbf{b}) + 2p \\
BIC_p &= -2\log_e L(\mathbf{b}) + p\log_e(n)
\end{aligned}
$$

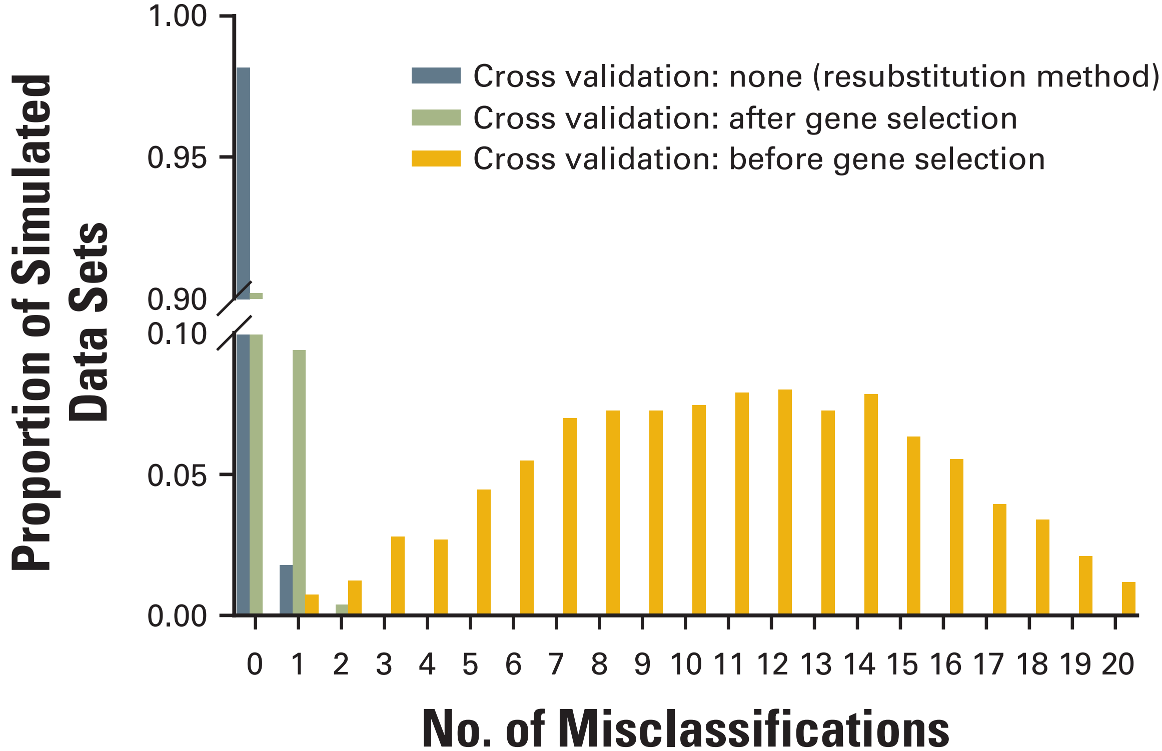
- Heuristic search
  - forward selection; backward elimination; stepwise selection
  - based on Wald statistic and Normal distribution

Variable Selection Should be Done as Part of Cross-Validation

- Example from Simon *et al.*, JNCI, 2003.
- Simulated data with no structure
  - 20 observations with random labels
  - 6,000 possible but unrelated predictors
  - Repeated 200 times
- Estimated predictive accuracy using
  - no cross-validation
  - selecting features on full dataset,
    then using cross-validation
  - selecting features at each step of cross-validation

Variable Selection Should be Done as Part of Cross-Validation

Example from Simon *et al.*, JNCI, 2003.



- Conclusion
  - Incorporating selection of predictors within the cross-validation procedure is key