

# STAT 545: Categorical Data Analysis (Part II)

Liang Li

Department of Biostatistics  
The University of Texas MD Anderson Cancer Center

[LLi15@mdanderson.org](mailto:LLi15@mdanderson.org)

Fall 2015 at Rice University

# Overview of Part II of this class

Oct 19, 2015 to December 2, 2015. There will be homework/projects and a final exam

- Regression model for binary data
- Regression model for ordinal data
- Regression model for counts data
- Extensions of standard regression models for categorical data
- Marginal models for longitudinal categorical data
- Conditional models for longitudinal categorical data

## Logistic Regression

# Logistic Regression Model

$$\begin{aligned}\pi(x) &= P(Y = 1|X = x) = 1 - P(Y = 0|X = x) \\ &= \text{expit}(\alpha + \beta x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \in (0, 1)\end{aligned}$$

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x \in (-\infty, \infty)$$

- ① Binary outcome; for binomial outcome, the model is similar
- ② Interpretation of  $\beta$  (log odds ratio)
- ③ Simple visual model checking by grouping (§ 5.1.2)
- ④ Logistic regression with retrospective studies (§ 5.1.4)
- ⑤ Model fitting through maximum likelihood estimation (§ 5.5)
- ⑥ Inference about model parameters and probabilities (§ 5.2.1)
- ⑦ Checking goodness of fit (§ 5.2.5)

# The (log) odds ratio and its interpretation

$$\text{logit} [\pi(\mathbf{x})] = \alpha + \beta \mathbf{x}$$



$$\text{logit} [\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p$$

# Simple visual model checking by grouping

- 1 Group the (continuous) covariate into 10 categories by cutoffs at the quantiles, with  $n_i$  subjects in each group ( $i = 1, 2, \dots, 10$ )
- 2 Calculate the average covariate within each group ( $\bar{x}_i$ )
- 3 Calculate the proportion of  $Y = 1$  within each group ( $\bar{y}_i$ )
- 4 Plot logit of  $\bar{y}_i$  vs.  $\bar{x}_i$ . It should be approximately a straight line
- 5 Note: may need correction when  $\bar{y}_i = 0$  or 1.

$$\log \frac{\bar{y}_i}{n_i - \bar{y}_i} \Rightarrow \log \frac{\bar{y}_i + 0.5}{n_i - \bar{y}_i + 0.5}$$

- 6 Only work with a single covariate

## Logistic regression with retrospective studies (§ 5.1.4)



# Model fitting through maximum likelihood estimation (§ 5.5)





# Inference on parameters and probabilities (§ 5.2.1)

Test  $H_0 : \beta = 0$  in logistic model  $\text{logit} [\pi(x)] = \alpha + \beta x$

- ① Wald, Likelihood ratio, and Score tests are applicable (§ 1.3.3)
- ② The predicted probability and **its confidence interval**



## Checking goodness of fit (§ 5.2.3)

$$\text{logit} [\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- ① Visual checking through grouping (works best with a single covariate)
- ② Adding interactions, quadratic terms, etc., and testing for significance or looking at AIC/BIC: problematic but widely used
- ③ Making the model more flexible by using splines
- ④ Global goodness of fit checking by *Hosmer & Lemeshow test*

$$\sum_{i=1}^g \frac{\left( \sum_j y_{ij} - \sum_j \hat{\pi}_{ij} \right)^2}{n_i \left( \sum_j \hat{\pi}_{ij} / n_i \right) \left[ 1 - \left( \sum_j \hat{\pi}_{ij} \right) / n_i \right]} \sim \chi_{g-2}^2$$

- A large value of any global fit statistic merely indicates *some* lack of fit but provides no insight about its nature

# Logistic models with categorical predictors (§ 5.3)

- When there is a single categorical predictor, the data can be arranged in an  $I \times 2$  contingency table (e.g., Table 5.3)
- When the categories are unordered (e.g., nominal data), the (saturated) model is  $\text{logit}(\pi_i) = \beta_i$  ( $i = 1, 2, \dots, I$ ), with  $I$  unknown parameters.
- We may write the model as  $\text{logit}(\pi_i) = \alpha + \beta_i$  with set-to-zero constraint  $\beta_1 = 0$  or sum-to-zero constraint  $\sum_i \beta_i = 0$
- The model for subject  $j$  ( $j = 1, 2, \dots, n$ ) is  $\text{logit}(\pi_j) = \alpha + \sum_{i=1}^I \beta_i 1\{j \in \text{group } i\}$
- When the categories are ordered (e.g., ordinal data), we may assume that  $\text{logit}(\pi_i) = \alpha + \beta x_i$ 
  - The number of parameters reduced with the linear assumption.
  - Be careful about coding  $x_i$  ( $i=1,2,\dots,I$ ): (1,2,3) or (1,4,9)?
  - Treat the  $x_i$  like a continuous variable.

## Cochran-Armitage Trend Test (§ 5.3.5)


- Developed by Armitage (1955) and Cochran (1954) for  $I \times 2$  tables with ordered rows
- They used a linear probability model  $\pi = \alpha + \beta x_i$
- It is a chi-square test of the independence between rows and columns under the linear assumption.  $H_0 : \beta = 0$ .
- This test is equivalent to the score statistic for testing  $H_0 : \beta = 0$  in the linear logit model.
- Using directed models can improve inferential power
  - If the trend is indeed linear, making use of the linear trend (as in Cochran-Armitage test) is more powerful than not making use of the linear trend (as in  $\text{logit}(p_i) = \beta_i$ )

# Model Selection (§ 6.1)

The data set is  $\{Y_i, X_{1i}, X_{2i}, \dots, X_{pi}; i = 1, 2, \dots, n\}$ . The logistic regression model is


$$\pi(\mathbf{X}_i) = \text{expit}(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi})$$

The  $p$  covariates include interactions, quadratic terms, etc. We want to retain only the predictive covariates in the model.

- Model selection is both science and art
- The same principles that you learned in linear model class still apply
- Two goals: (1) complex enough to fit the data well; (2) relatively simple to interpret (avoid overfitting) 
- Confirmatory studies vs. exploratory studies

# How many covariates can be included in the model?

$Y_i \sim \text{Bernoulli}$  with  $\pi(\mathbf{X}_i) = \text{expit}(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi})$

- The effective sample size of a logistic regression is either  $\sum_i Y_i$  or  $n - \sum_i Y_i$ , whichever is smaller
- **The rule of thumb:** no more than the effective sample size divided by 10 (or, 10 events per covariate)
- Including too many covariates may cause non-convergence 
- Avoid multicollinearity, as in linear regression (📖 Page 209, Table 6.1)
  - The overall test is highly significant ( $p < 0.0001$ )
  - The individual covariates are, in general, not very significant due to the multicollinearity between the horseshoe crab's width and weight ( $r = 0.887$ )


# Forward, backward, and stepwise model selection

$Y_i \sim \text{Bernoulli with } \pi(\mathbf{X}_i) = \text{expit}(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi})$

- Forward procedure: (1) start with just the intercept (2) at each step, add the covariate with the smallest p-value in likelihood ratio or Wald test (3) stop when no more significant covariate is available (However, it can stop prematurely due to lack of power)
- Stepwise procedure: at each step, retest the significance of the terms added at previous stages
- Backward procedure: (1) start with full model (2) at each step, remove the covariate with the largest p-value (3) stop when all remaining covariates are significant. (However, full model may not be stable)
- The dummy variables for a single categorical covariate should be added or removed together (likelihood ratio test); do not place an interaction in the model without the main effect terms
- SAS PROC LOGISTIC offers additional entry and exit p-value criteria

# Further comment on forward, backward, and stepwise model selection

$Y_i \sim \text{Bernoulli}$  with  $\pi(\mathbf{X}_i) = \text{expit}(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi})$

-  Page 211, Table 6.2 illustrates
  - three-way interaction is usually not significant (e.g., lack of power) and not desirable (hard to interpret)
  - dropping multiple covariates at once using likelihood ratio test (LRT) or dropping them one at a time (Wald or LRT)
- All these procedures are not rigorously justified (*ad hoc*); use with caution!
- Modern approaches are available (LASSO, bagging, etc.)
- Philosophically, there is no such thing as “the correct model” or “the true model”: ALL MODELS ARE WRONG, SOME ARE USEFUL — George Box






# Akaike Information Criterion (AIC)

Select the model with smaller AIC or BIC ( $L$ : maximized log likelihood;  $m$ : number of parameters in the model;  $n$ : sample size)

$$AIC = -2L + 2m$$


$$BIC = -2L + \log(n)m$$

- **Rationale:** Including more covariates will always include the log likelihood, but may cause overfitting; so we put a “penalty” by adjusting for the size of the model. There are mathematical reasons why the penalty must take this form.
- Other penalties are available: HQ, DIC, etc.
- BIC puts more penalty on larger model, and therefore tends to select the simpler model  Page 213
- Like scatter plot smoothing, the “desired” amount of penalty is a somewhat subjective choice 
- Need a comprehensive assessment of AIC/BIC, significance, residuals, scientific rationale, parsimony and interpretability, etc. 

# Residuals: Pearson, Deviance, Standardized

Let  $y_i$  denote the binomial outcome for  $n_i$  trials at setting  $i$  of the explanatory variables,  $i = 1, 2, \dots, N$ . Let  $\hat{\pi}_i$  denote the model estimate of  $P(Y = 1)$ .



- Pearson residual is like the residual for linear regression, but with standardization
- Deviance residual is motivated from the likelihood and deviance (which resembles the sum of squares in linear regression)
- Standardized residual has an approximate  $N(0, 1)$  distribution and is the one that we usually use, BUT:
  - use it with grouped data (binomial instead of binary).  Page 217, Table 6.5

# Influence diagnosis for logistic regression

- A single observation can have a much more exorbitant influence in linear regression than in logistic regression, since linear regression has no bound on the distance of  $y_i$  from the expected value.
- Points that have extreme predictor values need not have high leverage. In fact, the leverage can be relatively small if  $\hat{\pi}_i$  is close to 0 or 1.

# Predictive power of a logistic regression model: pseudo $R^2$

- For linear regression  $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i$ , the  $R^2$  is

$$R^2 = 1 - \frac{\sum_i (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}})^2}{\sum_i (Y_i - \bar{Y})^2}$$

- For logistic regression, the analog

$$1 - \frac{\sum_i (Y_i - \hat{\pi}_i)^2}{\sum_i (Y_i - \bar{Y})^2}$$

may not be nondecreasing as the model gets more complex  
(undesirable)




# Predictive power of a logistic regression model: pseudo $R^2$

For logistic regression, a more widely used measure is the pseudo  $R^2$  of McFadden (1974):  $\frac{L_M - L_0}{L_S - L_0} = 1 - \frac{L_M}{L_0}$

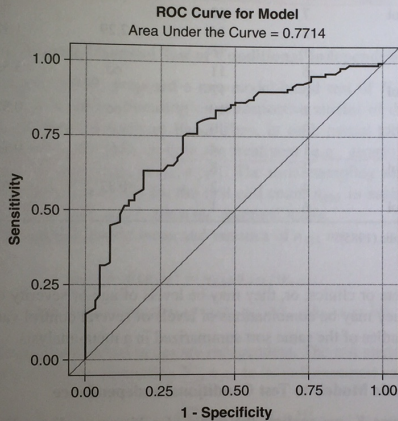
$$L = \log \prod_{i=1}^N [\pi_i^{y_i} (1 - \hat{\pi}_i)^{1-y_i}] = \sum_{i=1}^N [y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)]$$

- $L_M$  is the log likelihood evaluated at the MLE  $\hat{\pi}_i = \text{expit}(\mathbf{X}_i^T \hat{\beta})$
- $L_0$  is the log likelihood evaluated under the MLE of the null model:  
 $\hat{\pi}_i = N^{-1} \sum_i y_i$
- $L_S$  is the log likelihood evaluated under the saturated model with  
 $\hat{\pi}_i = y_i$ .  $L_S = 0$

# Receiver Operative Characteristics (ROC) curve

- $Y_i = 0$  (non disease) or 1 (disease). Estimated viral load  $\hat{\pi}_i \in (0, 1)$ . We classify the subject as a case ( $Y = 1$ ) when  $\hat{\pi} > c$  and control ( $Y = 0$ ) when  $\hat{\pi} \leq c$ .
- Sensitivity  $P(\hat{\pi} > c | Y = 1) \leftarrow \frac{\sum_i 1\{\hat{\pi}_i > c\}}{\sum_i Y_i}$  
- Specificity  $P(\hat{\pi} \leq c | Y = 0) \leftarrow \frac{\sum_i 1\{\hat{\pi}_i \leq c\}}{\sum_i (1 - Y_i)}$  
- ROC curve  p225
- The area under the ROC curve (AUC) is reported as c-statistic in SAS PROC LOGISTIC. It is a number between 0 and 1.  $AUC = 0.5$  is like flipping a coin. So  $AUC < 0.5$  is unlikely. Good classification requires  $AUC > 0.80$  (excellent,  $> 0.9$ ).

# ROC Curve



**Figure 6.4** ROC curve (from SAS PROC LOGISTIC) for logistic regression model estimating the probability a crab has satellites, using width and color predictors.

# Cochran-Mantel-Haenszel Test (§ 6.4)

- Study the association between a treatment variable (e.g., binary) and a disease outcome (e.g., binary) after adjusting for a possibly confounding variable (e.g., categorical or continuous but grouped) that might influence that association
- Example in Table 6.9: multicenter randomized clinical trial comparing treatment vs. placebo on a binary outcome (cured vs. not)
- The logistic regression approach ( $i = 1, 2; k = 1, 2, \dots, K; x_i = 1$  or  $2$ ):

$$\pi_{ik} = P(Y = 1 | X = i, Z = k)$$

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z$$

- Test  $H_0: \beta = 0$  using Wald or likelihood ratio test
- What if there is interaction between  $X$  and  $Z$ , i.e.,  $\beta$  depends on  $Z$ ?



# Table 6.9

226

BUILDING, CHECKING, AND APPLYING LOGISTIC REGRESSION MODELS

**Table 6.9 Clinical Trial Relating Treatment to Response for Eight Centers, with Expected Value and Variance (of Success Count for Drug) Under Conditional Independence**

Center	Treatment	Response		Odds Ratio	$\mu_{11k}$	$\text{var}(n_{11k})$
		Success	Failure			
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	$\infty$	3.52	1.20
	Control	0	12			
6	Drug	1	10	$\infty$	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

often medical centers or clinics; or they may be levels of age or severity of the condition.

# Cochran-Mantel-Haenszel Test (§ 6.4)

Data from Center  $k$  ( $k = 1, 2, \dots, K$ )

	cured	not	Total
Treatment	$n_{11k}$	$n_{12k}$	$n_{1+k}$
placebo	$n_{21k}$	$n_{22k}$	$n_{2+k}$
Total	$n_{+1k}$	$n_{+2k}$	$n_{++k}$


- Test  $H_0$ : Treatment and outcome independent conditional on center
- Both the treatment (row) and outcome (column) totals fixed,  $n_{11k} \sim$  hypergeometric distribution
- Under the null, the hypergeometric mean and variance of  $n_{11k}$  are

$$\mu_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k}$$

$$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/[n_{++k}^2(n_{++k} - 1)]$$

- The CMH statistic is  $CMH = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k \text{var}(n_{11k})}$ , which has a large sample chi-squared null distribution with  $df = 1$ .

# Cochran-Mantel-Haenszel Test vs. Logistic Regression

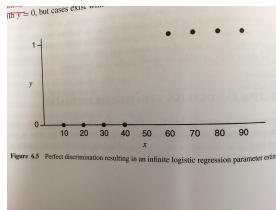
- When the sample size per center (also called strata) is moderately large, the two produce similar results (CMH is a score test of the logistic model)
- When the number of strata is large (like matched pairs data), the logistic regression does not apply but CMH still applies
- A point estimator of the overall odds ratio is available from CMH 

$$\hat{\theta}_{CMH} = \frac{\sum_k (n_{11k}n_{22k}/n_{++k})}{\sum_k (n_{12k}n_{21k}/n_{++k})} = \frac{\sum_k n_{++k}p_{11|k}p_{22|k}}{\sum_k n_{++k}p_{12|k}p_{21|k}}$$

- CMH is the standard method for stratified analysis of categorical data, applicable to  $I \times J \times K$  contingency table
- If the treatment effect differs across strata (interaction), use logistic model with interaction

# Quasi-complete Separation in Logistic Regression

- Be careful about excessively large (or small) odds ratios or excessively large standard errors
  - multi-collinearity; remove one of the correlated covariates
  - complete or quasi-complete separation
- Complete separation: there exists a vector  $\mathbf{b}$  such that  $\mathbf{b}^T \mathbf{x}_i > 0$  whenever  $y_i = 1$  and  $\mathbf{b}^T \mathbf{x}_i < 0$  whenever  $y_i = 0$ . (more likely with continuous covariates)
- Quasi-complete separation:  $\mathbf{b}^T \mathbf{x}_i \geq 0$  whenever  $y_i = 1$  and  $\mathbf{b}^T \mathbf{x}_i \leq 0$  whenever  $y_i = 0$ . (more likely with categorical covariates)
- Not a problem with linear regression



## Regression Models for Multinomial Data

# Models for Multinomial Responses (§ 8)

- Nominal data vs. ordinal data: presence or absence of intrinsic order
- For nominal outcome variable  $Y = 1, 2, \dots, J$ . We need to model  $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$  under constraint  $\sum_j \pi_j(\mathbf{x}) = 1$ .
- $Y$  follows a multinomial distribution with probabilities  $\{\pi_1(\mathbf{x}), \dots, \pi_J(\mathbf{x})\}$ .
- Baseline-category logit model (e.g., pick  $J$  as the baseline/reference category)

$$\log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j + \beta_j^T \mathbf{x} \quad , \quad j = 1, 2, \dots, J - 1$$


- These  $J - 1$  equations determine parameters for logits with other pairs of response categories, as well as the response probabilities:

$$\log \frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})} = \log \frac{\pi_a(\mathbf{x})}{\pi_J(\mathbf{x})} - \log \frac{\pi_b(\mathbf{x})}{\pi_J(\mathbf{x})}$$

# Models for Multinomial Responses: baseline-category logit model

- The constraint leads to:  $\pi_J(\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x})}$
- The probability for the  $j$ -th category ( $j = 1, 2, \dots, J-1$ ):

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \beta_j^T \mathbf{x})}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j^T \mathbf{x})}$$

- With more than two response categories ( $J > 2$ ), the probability of a given category need not continuously increase or decrease (e.g.,  $\pi_j(\mathbf{x})$  may not be a monotone function of  $\mathbf{x}$ ) 
- The model is fit by maximum likelihood. Let  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iJ})$ , where  $y_{ij} = 1$  when the response is in category  $j$  and 0 otherwise, so that  $\sum_j y_{ij} = 1$ . The log likelihood is:

$$\sum_{i=1}^n \log \left[ \prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right]$$

## Models for Ordinal Responses (§ 8.2)

- $Y = 1, 2, \dots, J$ . We need to model  $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$  under constraint  $\sum_j \pi_j(\mathbf{x}) = 1$ .
- Due to the intrinsic ordering of the response categories, we model the cumulative probabilities

$$P(Y \leq j|\mathbf{x}) = \pi_1(\mathbf{x}) + \dots + \pi_j(\mathbf{x})$$

- The cumulative logits are defined as:

$$\text{logit}[P(Y \leq j|\mathbf{x})] = \log \frac{P(Y \leq j|\mathbf{x})}{1 - P(Y \leq j|\mathbf{x})} = \log \frac{\pi_1(\mathbf{x}) + \dots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \dots + \pi_J(\mathbf{x})}$$

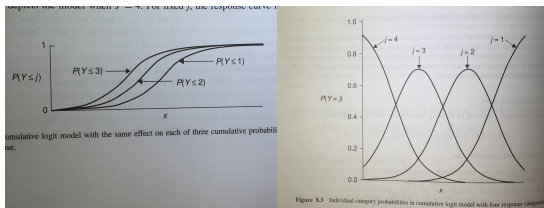
- Cumulative logit model

$$\text{logit}[P(Y \leq j|\mathbf{x})] = \alpha_j + \beta^T \mathbf{x} \quad , \quad j = 1, 2, \dots, J - 1$$

- The cumulative logit is monotone in  $\mathbf{x}$ ; this feature not available in baseline-category logit model



# Proportional Odds Model



- Proportional odds model  $\text{logit}[P(Y \leq j|\mathbf{x})] = \alpha_j + \beta^T \mathbf{x}$ , where  $j = 1, 2, \dots, J - 1$
- Need to assume the same  $\beta$  for each logit. Therefore, *cumulative logit model* is also called the *proportional odds model*

$$\text{logit}[P(Y \leq j|\mathbf{x}_1)] - \text{logit}[P(Y \leq j|\mathbf{x}_2)] = \beta^T (\mathbf{x}_1 - \mathbf{x}_2)$$

- Cumulative odds ratio  $\exp(\beta)$  does not depend on  $j$ , i.e.,  $\beta_j \equiv \beta, \forall j$
- We cannot make the model more generalizable by letting replacing  $\beta$  with  $\beta_j$ : the different cumulative probabilities may cross, which is

impossible

# Proportional Odds Model

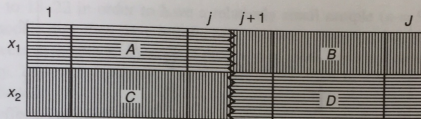


Figure 8.4 Uniform odds ratios  $AD/BC$  whenever  $x_1 - x_2 = 1$ , for all binary collapsings of the response in cumulative logit model of proportional odds form.

likelihood function is

$$\begin{aligned} \prod_{i=1}^n \left[ \prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \prod_{i=1}^n \left\{ \prod_{j=1}^J [P(Y \leq j | \mathbf{x}_i) - P(Y \leq j-1 | \mathbf{x}_i)]^{y_{ij}} \right\} \\ &= \prod_{i=1}^n \left\{ \prod_{j=1}^J \left[ \frac{\exp(\alpha_j + \beta^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \beta^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \beta^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \beta^T \mathbf{x}_i)} \right]^{y_{ij}} \right\}, \end{aligned} \quad (8.6)$$

viewed as a function of  $(\{\alpha_j\}, \beta)$ . This can be maximized to obtain the ML estimates using the Newton-Raphson algorithm (see, e.g., McCullagh and Nelder, 1989) or the Newton-Raphson algorithm.

- The model parameters are estimated by maximum likelihood

# Latent Variable Motivation for Proportional Odds Model

- A continuous latent variable  $y^*$  with  $y^* = \tilde{\beta}^T \mathbf{x} + \epsilon$  and the distribution function of  $\epsilon$  is  $G(\cdot)$  (not mean zero)
- Thresholds  $-\infty = \tilde{\alpha}_0 < \tilde{\alpha}_1 < \dots < \tilde{\alpha}_J = \infty$
- The observed response  $y$  satisfies  $y = j$  if  $\tilde{\alpha}_{j-1} < y^* \leq \tilde{\alpha}_j$

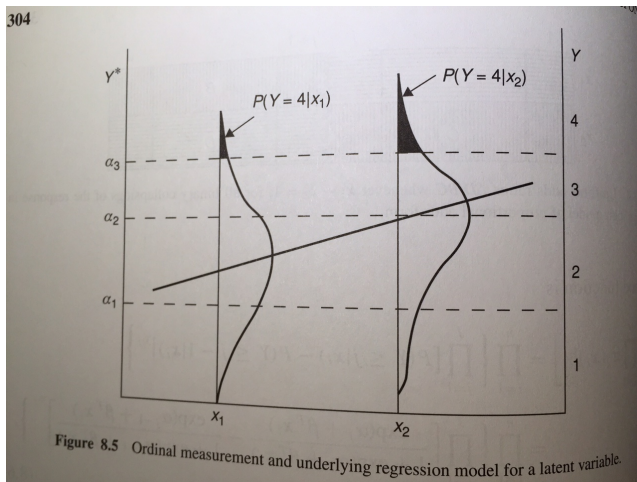
$$P(Y \leq j | \mathbf{x}) = P(y^* \leq \tilde{\alpha}_j | \mathbf{x}) = G(\tilde{\alpha}_j - \tilde{\beta}^T \mathbf{x})$$

- The proportional odds model

$$P(Y \leq j | \mathbf{x}) = \text{expit}(\alpha_j + \beta^T \mathbf{x})$$

- $G(\cdot) \sim \text{expit}(\cdot)$ ,  $G^{-1}(\cdot) \sim \text{logit}(\cdot)$ ,  $\tilde{\alpha}_j = \alpha_j$ ,  $\tilde{\beta} = -\beta$

# Proportional Odds Model: Interpretation using latent variable



# Probit and Logit: Latent Variable Motivation for Binary Outcome Model

- A continuous latent variable  $y^*$  with  $y^* = \tilde{\beta}^T \mathbf{x} + \epsilon$  (no intercept) and the distribution function of  $\epsilon$  is  $G(\cdot)$
- Threshold  $-\infty < \tilde{\alpha} < \infty$ . The observed response  $y$  satisfies  $y = 1$  if  $y^* \leq \tilde{\alpha}$  and  $y = 0$  otherwise

$$P(Y = 1|\mathbf{x}) = P(y^* \leq \tilde{\alpha}|\mathbf{x}) = G(\tilde{\alpha} - \tilde{\beta}^T \mathbf{x})$$




- The logistic regression model:  $P(Y = 1|\mathbf{x}) = \text{expit}(\alpha + \beta^T \mathbf{x})$ .  
 $G(\cdot) \sim \text{expit}(\cdot)$ ,  $G^{-1}(\cdot) \sim \text{logit}(\cdot)$ ,  $\tilde{\alpha} = \alpha$ ,  $\tilde{\beta} = -\beta$
- The probit regression model:  $P(Y = 1|\mathbf{x}) = \Phi(\alpha + \beta^T \mathbf{x})$ .  
 $G(\cdot) \sim \Phi(\cdot)$ ,  $G^{-1}(\cdot) \sim \Phi^{-1}(\cdot)$ ,  $\tilde{\alpha} = \alpha$ ,  $\tilde{\beta} = -\beta$

# Check the Proportional Odds Assumption

$$P(Y \leq j | \mathbf{x}) = \text{expit} \left( \alpha_j + \beta^T \mathbf{x} \right)$$

- Proportional odds model is parsimonious and easy to interpret
- Replacing  $\beta$  with  $\beta_j$  may cause the cumulative probabilities to cross
- A score test of proportional odds model is available (SAS PROC LOGISTIC)
- Retain proportional odds model unless there is strong deviation from this assumption
- What to do when the proportional odds assumption is violated:
  - Adding additional terms, such as interaction
  - alternative ordinal model (next slides)
  - partial proportional odds model (SAS PROC LOGISTIC)
  - baseline category logit model

# Alternative Models for Ordinal Data

- Cumulative link model:  $G^{-1} [P(Y \leq j|\mathbf{x})] = \alpha_j + \beta^T \mathbf{x}$
- Cumulative probit model:  $\Phi^{-1} [P(Y \leq j|\mathbf{x})] = \alpha_j + \beta^T \mathbf{x}$
- Cumulative complementary log-log model:  
 $\log \{ -\log [1 - P(Y \leq j|\mathbf{x})] \} = \alpha_j + \beta^T \mathbf{x}$ 
  - Equivalent to a latent variable model with extreme value distribution for the residuals
  - Equivalent to proportional hazard model in for discrete survival data analysis (e.g., year of death at 1, 2, 3, ... ) 
  - Rare event logistic regression
- Adjacent category logit model 
- Continuation ratio logit model 

# Regression Models for Counts Data



# Poisson Regression: introduction

- Applicable to counts data 0, 1, 2, ..... Popularized in the 1970s and 1980s.
- Suppose  $Y$  has Poisson distribution with mean  $\mu$

$$P(Y = y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots \quad (1)$$

and  $E(Y) = \text{var}(Y) = \mu$

- Suppose we want to relate  $Y$  to a covariate  $X$ , we assume  $g(\mu) = \alpha + \beta X$ . What choices do we have for the link function  $g(\cdot)$ ? Any monotone mapping from  $(0, \infty)$  to  $(-\infty, \infty)$ ?
- Log link (log linear model)  $\mu = \exp(\alpha + \beta X)$ : the multiplicative effect of  $X$  on  $\mu$ . Interpret  $\exp(\beta)$  as a rate ratio.
- Example: A study of 400 patients with malignant melanoma.  $Y$ : tumor counts;  $X$ : tumor type (Hutchinson's, Superficial, Nodular, Indeterminate) and tumor site (Head & Neck, Trunk, Extremities)

# Poisson Regression: Offset

- Frequently, discrete counts represent information collected over time (days, years) or in space (volume for bacteria counts) and interest lies in modeling rates. Denote the exposure time or volume by  $N$ , then the rate is  $Y/N$ , which has expectation  $\mu/N$ .
- Modeling this rate with a log linear model as  $\log \frac{\mu}{N} = \alpha + \beta X$ . That is  $\log \mu = \alpha + \beta X + \log(N)$  or  $\mu = N \exp(\alpha + \beta X)$ .
- $Y \sim \text{Poisson}(\mu)$ .




- The log-likelihood is

$$L = \log \prod_{i=1}^n \mu_i^{Y_i} \{ \exp(-\mu_i) \} / Y_i! \quad , \quad \mu_i = N_i \exp(\alpha + \beta X_i)$$

# Poisson Regression: Example with Offset

- Table 12.2 of “Categorical Data Analysis with the SAS System”.

Region	Age Group	Cases	Total
North	< 35	61	2,880,262
South	35 – 44	76	564,535
...	...	...	...



- Counts:** number of new cases in 1961-1971; **offset:** size of the populations at risk; **covariates:** age group and region
- Can this example be analyzed with binomial regression? 

```
proc genmod data = melanoma order = data ;  
class age region ;  
model cases = age region / dist = poisson link = log offset = ltotal ;  
run ;
```

# Poisson Regression: Overdispersion

- Real data often have more variation than expected from a Poisson distribution (due to unobserved heterogeneity not captured by the covariates);  $\text{var}(Y) > E(Y)$ .
- Discover overdispersion through overdispersion parameter (next slide)
- Three ways to correct for overdispersion
  - Overdispersion model
  - Negative binomial regression
  - Poisson regression with random effect (later)


# Overdispersion Model

- $\text{var}(Y) = \phi E(Y)$ .  $\phi > 1$  is the dispersion parameter. ( $0 < \phi < 1$  is called under dispersion)
- $\mu_i = E(Y_i | \mathbf{X}_i) = N_i \exp(\mathbf{X}_i^T \theta)$  and  $\text{var}(Y_i | \mathbf{X}_i) = \phi \mu_i$ . In a Poisson regression model, we model the mean as a function of covariates, but not  $\phi$
- The likelihood equation  $0 = \sum_{i=1}^n (Y_i - N_i \exp(\mathbf{X}_i^T \theta)) \mathbf{X}_i$ , which uses only the mean function. So the point estimator is always correct even when overdispersion is ignored; but variance estimation and p-values will be incorrect.
- Estimate  $\phi$  , and then multiply the ordinary standard error estimates by  $\sqrt{\hat{\phi}}$ . Example in page 150 
- An approximate method motivated from quasi-likelihood ideas
- Maybe wrong if the mean function is misspecified (major weakness)

# Overdispersion with Binomial Model

- $y_i$  is the sample proportion from  $n_i$  Bernoulli trials with parameter  $\pi_i$ ,  $i = 1, 2, \dots, n$ .
- According to binomial model  $E(Y_i) = \pi_i$  and  $\text{var}(Y_i) = \pi_i(1 - \pi_i)/n_i$ .
- $v(\pi_i) = \phi\pi_i(1 - \pi_i)/n_i$
- Estimate  $\phi$  by  $X^2/(n - p)$  and multiply the standard error by  $\sqrt{(\hat{\phi})}$
- Not applicable when some  $n_i = 1$  (and some not) because  $\phi$  can only be 1 in this case.





## Negative Binomial Regression §4.3.4

- If  $Y$  has a Poisson distribution with both mean and variance equal to  $\lambda$ , and  $\lambda$  has a gamma distribution (extra variation; overdispersion) with mean  $\mu$  and variance  $\mu^2/k$ , then marginally  $Y$  has a negative binomial distribution with mean  $\mu$  and variance  $\mu + \gamma\mu^2$  ( $\gamma = 1/k > 0$ ) 
- When  $\gamma \rightarrow 0$  (random effect variance approaches zero), the NB distribution approaches Poisson
- Like the overdispersion model, we let  $\gamma$  to be a overdispersion constant and model the mean  $\mu = N \exp(\mathbf{X}^T \theta)$ .
- Standard error will become larger after applying negative binomial regression
- Preferred over overdispersion model
- Available in SAS PROC GENMOD


# Conditional Logistic Regression for Binary Matched Pairs



# Conditional Logistic Regression

- (Positively) correlated continuous data 
- (Positively) correlated binary data 
- For the example in Table 11.1  , let  $(Y_{i1}, Y_{i2})$  denote the data pair,  $i = 1, 2, \dots, n$ . The conditional logistic regression model is ( $x_t = 0$  or  $1$ )

$$\text{logit}[P(Y_{ij} = 1)] = \mu + \beta x_t + \alpha_i, \quad t = 1, 2$$

- $\alpha_i$  is the random intercept, analogous to the linear regression case. It has a distribution, but the advantage of the methodology here is that it works without making any assumption on the form of that distribution
- Note:  $\beta$  is NOT the marginal odds ratio (, SAS)

## Conditional Logistic Regression Example

**Table 11.1 Presidential Votes in 2004 and in 2008, for Males Sampled in 2010 by the General Social Survey**

2004 Election	2008 Election		Total
	Democrat	Republican	
Democrat	175	16	191
Republican	54	188	242
Total	229	204	433

## COMPARING DEPENDENT PROPORTIONS

subject or matched pair randomly selected from the population of interest, let

- 2004 Odds ( $D/R$ ) = 191/243 ; 2008 Odds ( $D/R$ ) = 229/204 ; log OR (2008 vs. 2004) = 0.35
- Conditional logistic regression code below gives  $\beta = 1.22$ ,  $p < 0.0001$  (actually,  $\exp(\beta) = 54/16$ )
- McNemar test has  $p < 0.0001$ .

# SAS code

```
proc logistic ;  
strata pair ;  
model Vote(event = 'D') = Year ; run ;
```

```
proc freq ;  
tables Y2004 * Y2008 / agree ;  
run ;
```

# Fit Conditional Logistic Regression


$$\text{logit}[P(Y_{ij} = 1)] = \alpha_i + \beta^T X_{it} \quad , \quad t = 1, 2$$

- Maximize the likelihood by treating  $\alpha_i$  as fixed parameters?
- Maximize the likelihood by treating  $\alpha_i$  as random parameters (effects)?
- Eliminate all the  $\alpha_i$ 's by using a conditional likelihood.



# A Few Notes on Fitting Conditional Logistic Regression

$$\text{logit}[P(Y_{ij} = 1)] = \alpha_i + \beta^T X_{it} \quad , \quad t = 1, 2$$

- If  $X_{it}$  is the same for each subject, then  $\beta$  cannot be estimated with this method. Example:  $Y_{it}$  is the outcome at week 1 and week 4 after surgery, and  $X_{it}$  is the intra-operative treatment.
- When  $X_{i1} = 0$  and  $X_{i2} = 1$ , the test of  $\beta = 0$  is equivalent to McNemar test. Also, it can be shown that  $\exp(\beta)$  is estimated by  $n_{12}/n_{21}$  where  $n_{12}$  and  $n_{21}$  are off-diagonal counts in the McNemar test table 
- Since the likelihood conditioned on the discordant pairs ( $Y_{i1} \neq Y_{i2}$ ), the concordant pairs ( $Y_{i1} = Y_{i2}$ ) are non-informative and thus can be ignored (as in McNemar test). What is the intuition behind it?
- An analog in the case of a continuous outcome and linear model with random intercept. 