

BIOSTAT 651
Notes #1: Introduction to GLM

- Lecture Topics:
 - Class outline
 - Linear regression
 - Motivation: more general approaches
 - Generalized Linear Models (GLM)
 - Examples

Class outline

- Generalized Linear Model (GLM)
- First half: general framework
 - GLM Model: systematic and random components
 - Parameter estimation
 - Hypothesis test
- Second half: applications
 - Binary data
 - Multinomial data
 - Count data
 - Over-dispersion

Linear regression

- Linear regression: based on the assumption that error terms are *continuous* and *normally distributed*
- Linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + e_i$$

where

$$e_i \sim N(0, \sigma^2).$$

- Relating p predictors for subject i to a response Y_i .
- Assumptions can be summarized by:

$$Y_i \sim N(\mu_i, \sigma^2),$$

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\mathbf{x}_i^T = (1, X_{i1}, \dots, X_{ip})$ and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$.

Linear regression

- Assumptions:
 - Systematic component: predictor effect through linear regression on the mean (*linearity assumption*)

$$E[Y_i|\mathbf{x}_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

- Random component: at each level of the predictor, variation in the response is characterized as

$$N(0, \sigma^2)$$

- Independence (between subjects)

Generalizing the linear model

- In many applications, the distribution of a *continuous* response may be *non-normal*.
- In addition, the response may be *discrete*, e.g.,
 - binary ($Y_i = 1, Y_i = 0$)
 - unordered categorical or nominal ($Y_i \in \{1, \dots, C\}$, with the ordering unimportant)
 - ordered categorical ($Y_i \in \{1, \dots, C\}$, with the ordering of the index important)
 - count ($Y_i \in \{0, 1, \dots, \infty\}$)
- In addition, a *non-linear* regression model relating the predictors to the mean may be needed.

Types of Responses

- Numeric response:
 - continuous
eg., weight, blood pressure
 - discrete
e.g., number of deaths, cancer cases, etc
- Categorical response:
 - nominal
e.g., blood type, gender, state
 - ordinal
e.g., low/medium/high; age group; calendar period

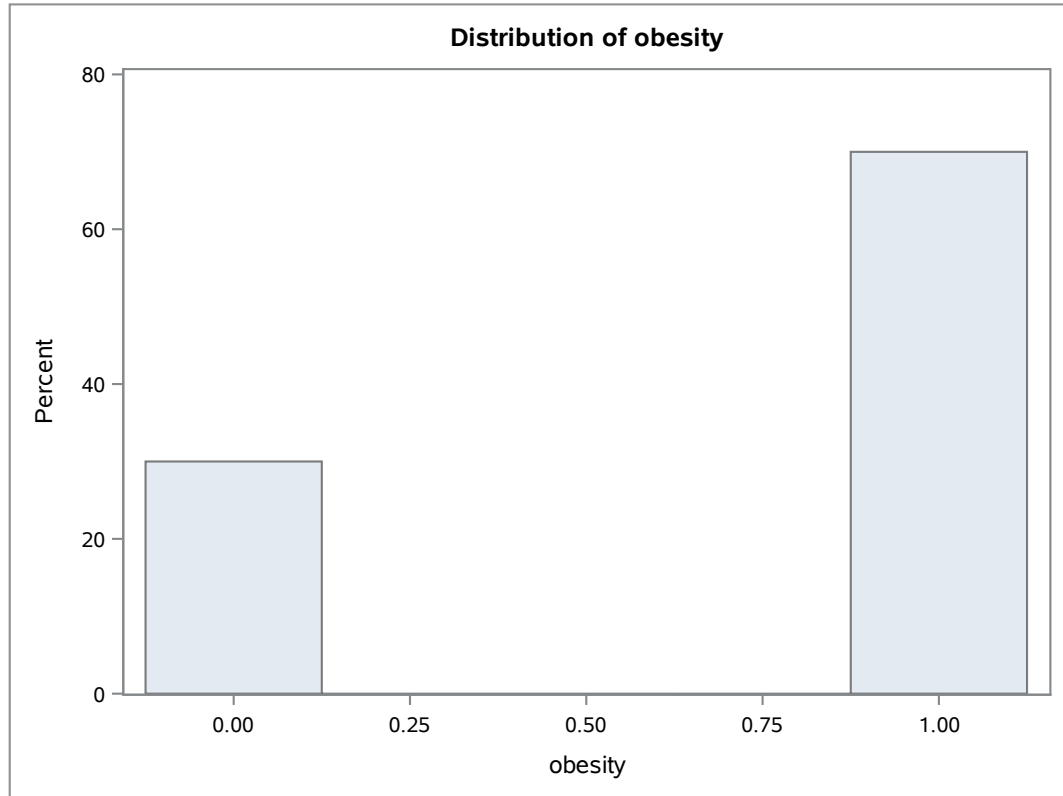
Example: Binary Response

- Childhood obesity data:
 - Response: obesity ($Y_i = 1$ if obese; $Y_i = 0$ otherwise)
 - Predictors: Age (in years) and Smoking Status
- Fit a linear regression model with normality assumption
 - Imagine a histogram of Y
 - What assumptions of the linear model are clearly violated for binary responses?

Histogram of Y

Monday, October 19, 2015 04:45:45 PM 3

The UNIVARIATE Procedure



Logistic and Linear Regression

- Recall our assumptions in linear regression:
 $Y_i \sim \text{Normal}$
 $V(Y_i) = \sigma^2$, constant variance
- Suppose Y_i takes on one of only two possible values (0 or 1), as in our example
 - e.g., $Y_i=0$ (alive) or 1 (dead)
 - e.g., $Y_i=0$ (no lung cancer) or 1 (lung cancer present)
- Clearly, Y_i does not follow a normal distribution
- In fact, $Y_i \sim \text{Bernoulli}(\pi_i)$, where
 $\pi_i = \pi(\mathbf{x}_i) \equiv P(Y_i = 1|\mathbf{x}_i)$

Bernoulli Distribution

- we've set $\pi_i = P(Y_i = 1|\mathbf{x}_i)$
- recall that for binary random variables:

$$E[Y_i] = \pi_i$$

$$V[Y_i] = \pi_i(1 - \pi_i)$$

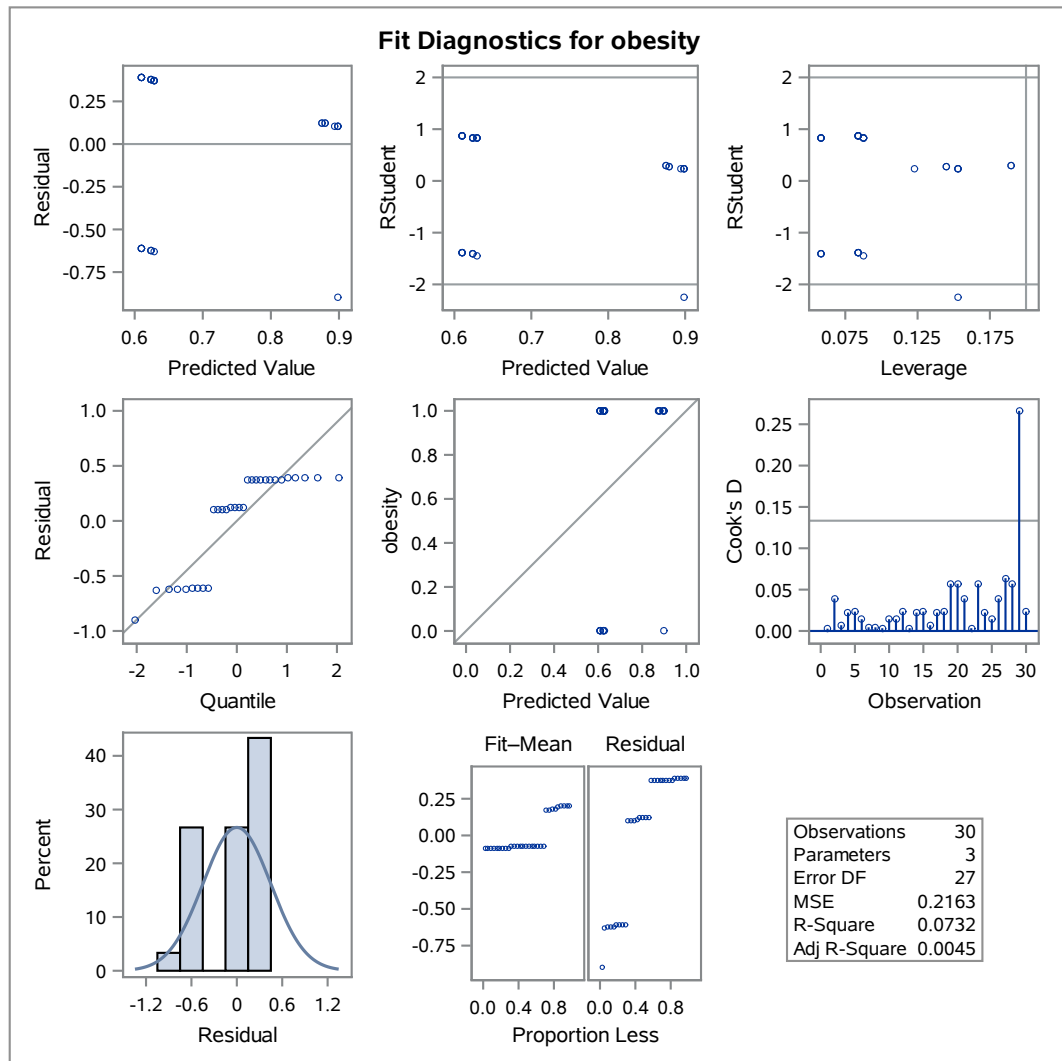
$$\text{note: } \pi_i = \pi(\mathbf{x}_i)$$

- hence, constant variance assumption is inherently violated, as variance is a function of the mean
- therefore, linear regression is invalid: normality and constant variance assumptions blatantly violated

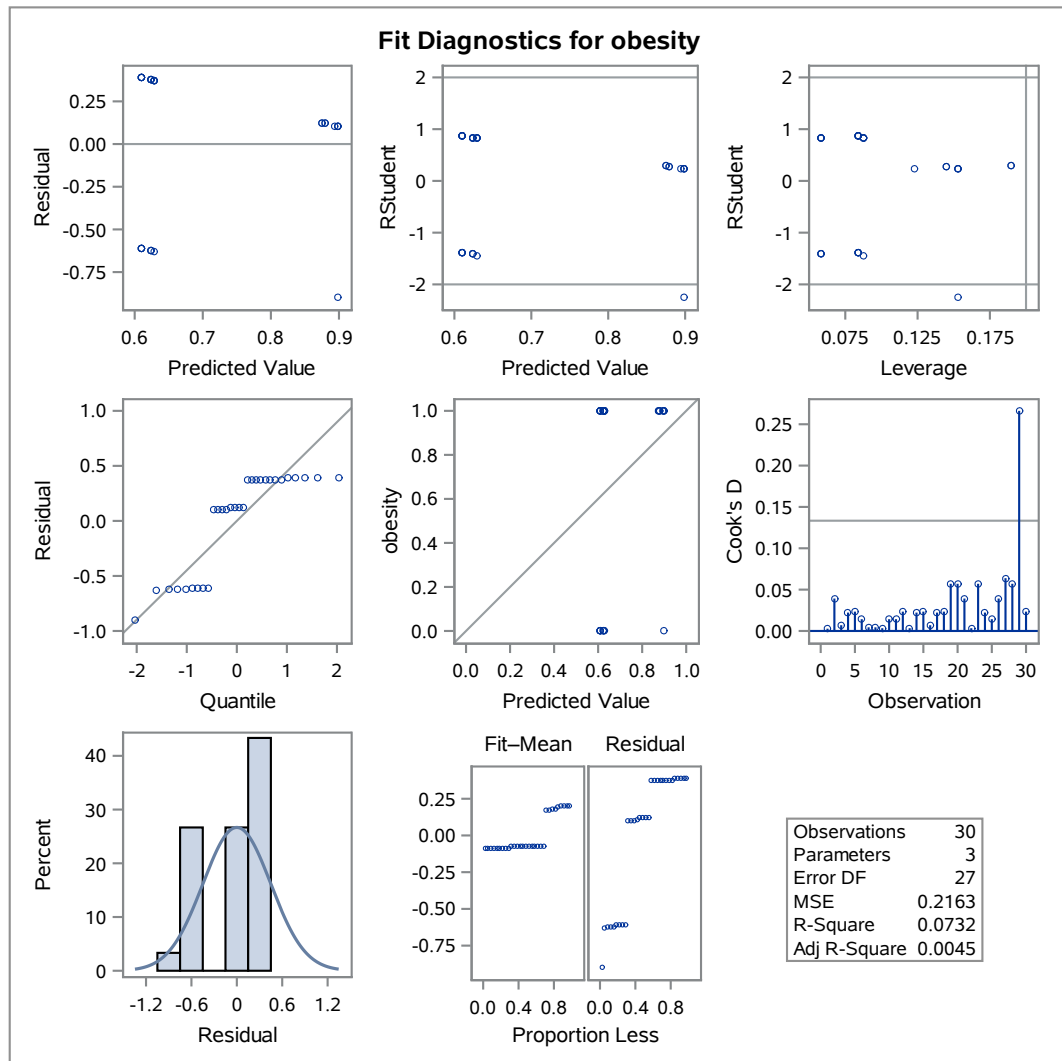
Example: Linear regression

```
DATA Weights;  
INPUT id wt age smoke obesity;  
datalines;  
1 22509.41 7 0 1  
2 33452.27 7 1 0  
3 13380.91 3 0 1  
4 24947.45 8 1 1  
5 15875.65 4 1 1  
.  
.  
.  
  
proc reg;  
model obesity = age smoke;  
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: obesity



The REG Procedure
Model: MODEL1
Dependent Variable: obesity



Linear Regression, Binary Data

- suppose we ignore the binary nature of Y_i , and fit the following linear regression model:

$$E[Y_i] = \mathbf{x}_i^T \boldsymbol{\beta}$$

- we fit the linear model, obtaining $\hat{\boldsymbol{\beta}}$ and the estimated means:

$$\hat{Y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$$

- note: $0 \leq \hat{Y}_i \leq 1$ need not hold, which is a major limitation since $E[Y_i]$ is a probability
- thus, we need to model some function of $E[Y_i]$, as opposed to $E[Y_i]$ itself

Logistic Function

- we need to find a transformation of $E[Y_i]$ to model as a linear function of covariates
- define the inverse-logit function:

$$\frac{e^x}{1 + e^x}$$

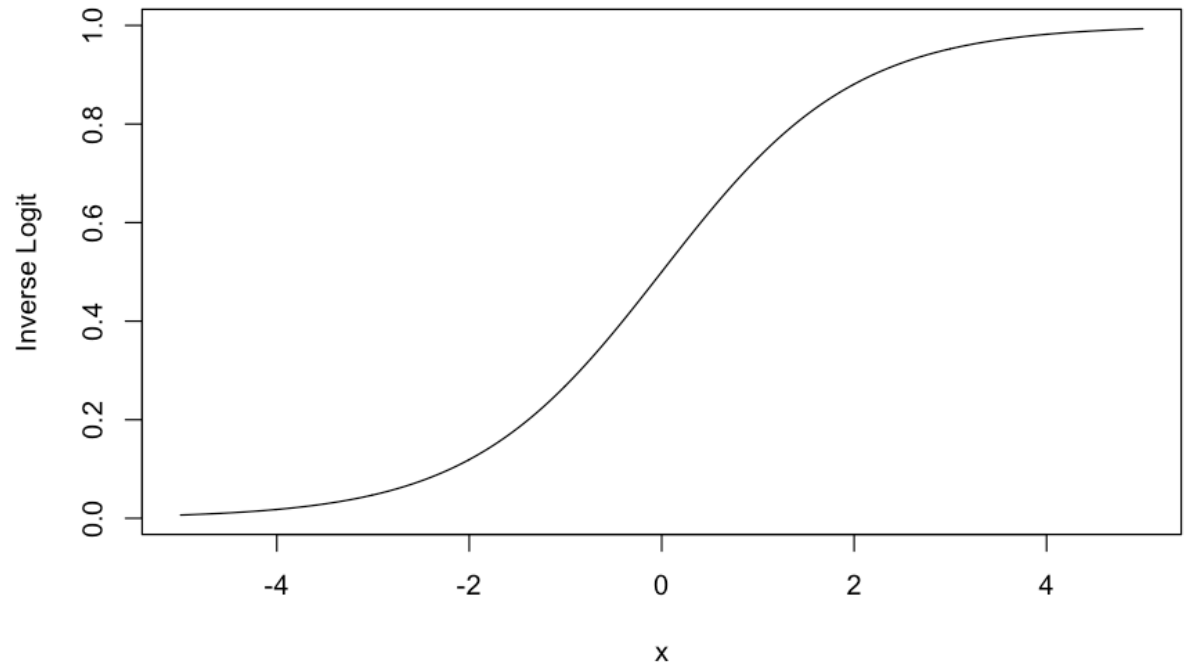
- clearly,

$$0 \leq \frac{e^x}{1 + e^x} \leq 1, \text{ for all } x$$

- this motivates the model:

$$E[Y_i] = \mu_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}, \text{ or}$$
$$\log \frac{\mu_i}{1 - \mu_i} = \mathbf{x}_i^T \boldsymbol{\beta}$$

- Inverse-logit function:



Example: Logistic regression

```
proc logistic;  
model obesity (event='1') = age smoke;  
run;
```

Logistic regression for binary responses**The LOGISTIC Procedure**

| Model Information | |
|---------------------------|------------------|
| Data Set | WORK.WEIGHTS |
| Response Variable | obesity |
| Number of Response Levels | 2 |
| Model | binary logit |
| Optimization Technique | Fisher's scoring |

| | |
|-----------------------------|----|
| Number of Observations Read | 30 |
| Number of Observations Used | 30 |

| Response Profile | | |
|------------------|---------|-----------------|
| Ordered Value | obesity | Total Frequency |
| 1 | 0 | 9 |
| 2 | 1 | 21 |

Probability modeled is obesity=1.

| Model Convergence Status |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

| Model Fit Statistics | | |
|----------------------|----------------|--------------------------|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 38.652 | 40.176 |
| SC | 40.053 | 44.379 |
| -2 Log L | 36.652 | 34.176 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|--|------------|----|------------|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 2.4762 | 2 | 0.2899 |
| Score | 2.1960 | 2 | 0.3335 |
| Wald | 1.9249 | 2 | 0.3819 |

Logistic regression for binary responses

The LOGISTIC Procedure

| Analysis of Maximum Likelihood Estimates | | | | | |
|--|----|----------|----------------|-----------------|------------|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 1.9242 | 1.6964 | 1.2867 | 0.2567 |
| age | 1 | 0.0266 | 0.2286 | 0.0135 | 0.9074 |
| smoke | 1 | -1.5967 | 1.1526 | 1.9190 | 0.1660 |

| Odds Ratio Estimates | | | |
|----------------------|----------------|----------------------------|-------|
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| age | 1.027 | 0.656 | 1.607 |
| smoke | 0.203 | 0.021 | 1.939 |

| Association of Predicted Probabilities and Observed Responses | | | |
|---|------|-----------|-------|
| Percent Concordant | 57.1 | Somers' D | 0.349 |
| Percent Discordant | 22.2 | Gamma | 0.440 |
| Percent Tied | 20.6 | Tau-a | 0.152 |
| Pairs | 189 | c | 0.675 |

Example: Ordered Response

- Example: The University of Regensburg conducted an investigation on senior psychology students regarding future job prospects. One of the key questions was whether they expected to find adequate employment after obtaining their degree.
- Response: Ordered categorical 1-3:
 1. Don't expect to find adequate employment
 2. Not sure
 3. Will obtain adequate employment immediately
 - Predictor: Age in years
- Scale of response: ordered categorical

Example: Ordered Response (continued)

| Age in years | Response | | |
|-----------------|----------|----|---|
| | 1 | 2 | 3 |
| 19 | 1 | 2 | 0 |
| 20 | 5 | 18 | 2 |
| 21 | 6 | 19 | 2 |
| 22 | 1 | 6 | 3 |
| 23 | 2 | 7 | 3 |
| 24 | 1 | 7 | 5 |
| 25 | 0 | 0 | 3 |
| 26 | 0 | 1 | 0 |
| 27 | 0 | 2 | 1 |
| 29 | 1 | 0 | 0 |
| 30 | 0 | 0 | 2 |
| 31 | 0 | 1 | 0 |
| 34 | 0 | 1 | 0 |

- Multinomial logistic regression:
 - Systematic component: logit function
 - Random component: multinomial distribution

Example: Count Response

- Cellular Differentiation (Piegorsch, Weinberg & Margolin, 1988):
 - Interest in the effect of two agents of immuno-activating ability that may introduce cell differentiation.
 - Do the agents TNF (tumor necrosis factor) and IFN (interferon) simulate cell differentiation independently or is there a synergetic effect?
- Response: number of cells that exhibited markers after exposure was recorded.
- Covariates:
 - TNF
 - IFN

| Number of cells differentiating | Dose of TNF (U/ml) | Dose of IFN (U/ml) |
|------------------------------------|-----------------------|-----------------------|
| 11 | 0 | 0 |
| 18 | 0 | 4 |
| 20 | 0 | 20 |
| 39 | 0 | 100 |
| 22 | 1 | 0 |
| 38 | 1 | 4 |
| 52 | 1 | 20 |
| 69 | 1 | 100 |
| 31 | 10 | 0 |
| 68 | 10 | 4 |
| 69 | 10 | 20 |
| 128 | 10 | 100 |
| 102 | 100 | 0 |
| 171 | 100 | 4 |
| 180 | 100 | 20 |
| 193 | 100 | 100 |

Example: Count Response (continued)

- Response: count
 - Poisson distribution often used for modeling counts
 - regression version of Poisson model:
$$E[Y_i] = \mathbf{x}_i^T \boldsymbol{\beta}$$
- Q: What issues arise if linear regression is used?
consider properties of Poisson variate ...

Notation: Counts, Rates, Person-Time

- Before deciding on a regression method, we set up some notation:
 - Y_i = event count, cell i
 - Y_i can take values 0, 1, 2, 3...
 - λ_i = event rate, cell i

$$\lambda_i = E[Y_i] \tag{1}$$

- covariates: $\mathbf{x}_i^T = (1, X_{i1}, \dots, X_{ip})$

Linear Regression for Count Data

- We return to our question of an appropriate modeling strategy
- Q: Could we model Y_i using linear regression?
 - Y_i is discrete and non-negative; violation of Normality assumption
 - $\lambda_i = E[Y_i]$ is an event rate; should be positive.
 - usually, when Y_i is a count, $V[Y_i]$ is related to $E[Y_i]$, in violation of the constant variance assumption

Poisson Regression: Deriving the Model

- We now set up our model equation...
- Begin by writing the rates as a linear function:

$$\lambda_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

- If we were to fit this model, there is no guarantee that $\mathbf{x}_i^T \hat{\boldsymbol{\beta}} > 0$
- How did we handle a parallel issue when deriving our model for binary data?

Poisson Regression Model

- Solution: since $e^x \geq 0$ for all x ,

$$\lambda_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$$

$$\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

(2)

- Distribution assumption

$$Y_i \sim \text{Poisson}(\lambda_i)$$

- We fit this model through the method of maximum likelihood

Summary

- Linear regression is inappropriate for each of these examples.
 - need a more general regression framework accounting for response data having a variety of measurement scales.
 - methods for model fitting and inference under this framework.
- Ideally, some elements of linear regression should carry over.
- Generalizations to more complex settings (correlated data, censored observations, etc) will be necessary in many applications (BIOSTAT 653, BIOSTAT 675)

Comment



All models are wrong, but some are useful.

BIOSTAT 651

Notes #2: Linear regression review

- Lecture Topics:
 - Review of linear regression
 - Weighted least squares

Linear regression

- response: Y_i
- covariate: $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$
- i : index of subject
- n : total number of subjects in the data

Model

- Linearly relate predictors to the mean response (assume X is deterministic)
- For i -th subject,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \end{aligned}$$

where $\epsilon_i \sim N(0, \sigma^2)$.

- Matrix form:
 - set $\mathbf{Y} = (Y_1, \dots, Y_n)^T$
 - design matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

- Model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim MVN_n(\mathbf{0}, \sigma^2 I)$$

i.e.

$$\mathbf{Y} \sim MVN_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I).$$

Model

- Assumptions:
 - Systematic component: predictor effect through linear regression on the mean (*linearity assumption*)

$$E[Y_i|\mathbf{x}_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

- Random component: at each level of the predictor, variation in the response is characterized as

$$N(0, \sigma^2)$$

- Independence (between subjects)

Interpretation

- In simple linear regression:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i,$$

- β_1 : change in mean response per unit increase of x_1 .

$$\beta_1 = E[Y_i | x_{i1} = a + 1] - E[Y_i | x_{i1} = a]$$

- Multiple linear regression: need to adjust for other covariates (or holding them constant)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

- β_1 : change in mean response per unit increase of x_1 , adjusting for x_2 (holding x_2 constant)

$$\begin{aligned} \beta_1 = & E[Y_i | x_{i1} = a + 1, x_{i2} = c] \\ & - E[Y_i | x_{i1} = a, x_{i2} = c] \end{aligned}$$

Parameter Estimation

- Least Squares Estimation (LSE): minimize the sum of squared errors

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Estimator:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},$$

where \mathbf{X} is of full rank.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},$$

- $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

- Variance of $\hat{\boldsymbol{\beta}}$:

$$V(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

- σ^2 estimator:

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} SSE = \frac{1}{n - p - 1} \mathbf{Y}^T (\mathbf{I} - \mathbf{H}) \mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

- Residual:

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Analysis of Variance

- ANOVA

- SST : total variation of \mathbf{Y} around mean

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \mathbf{Y}^T (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n) \mathbf{Y}$$

- SSR : variation of \mathbf{Y} explained by regression

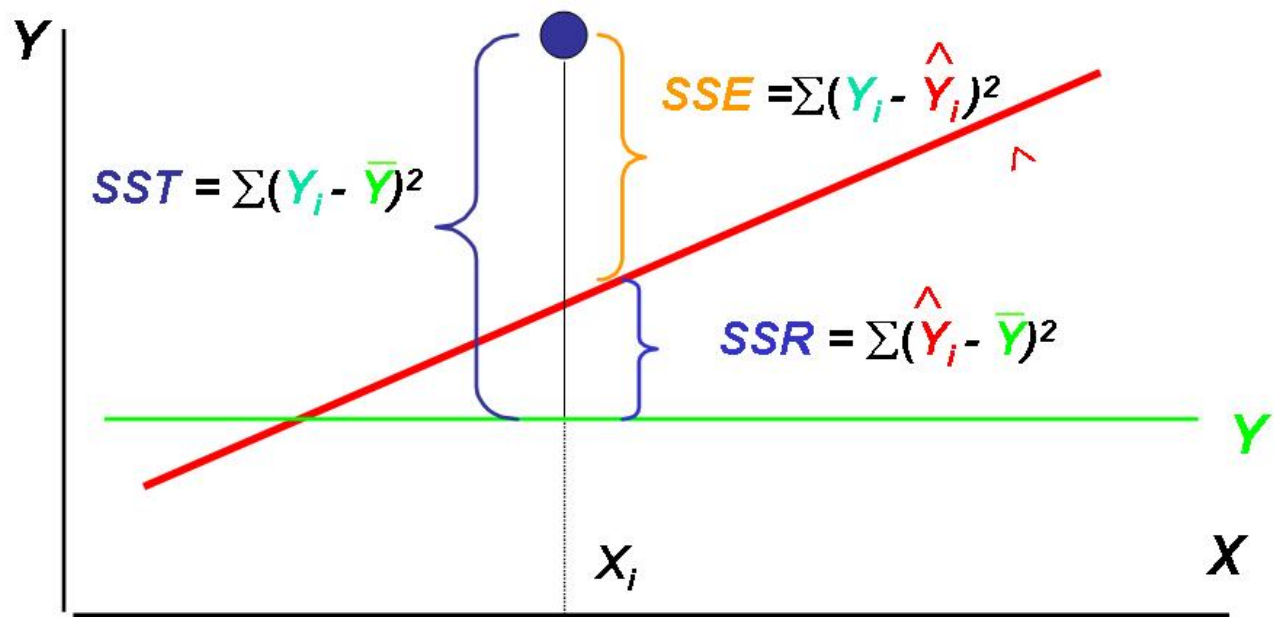
$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \mathbf{Y}^T (\mathbf{H} - \mathbf{1}\mathbf{1}^T/n) \mathbf{Y}$$

- SSE : variation of \mathbf{Y} unexplained by regression

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \mathbf{Y}^T (\mathbf{I} - \mathbf{H}) \mathbf{Y}$$

- $SST = SSR + SSE$

- Sum of squares



[<http://www.trizsigma.com/regression.html>]

- ANOVA table

| Source | SS | DF | MS | F |
|------------|-----|-------|-------------|---|
| Regression | SSR | p | SSR/p | |
| Error | SSE | n-p-1 | SSE/(n-p-1) | |
| Total | SST | n-1 | SST/(n-1) | |

- R^2 : explained sum of squares over total sum of square

$$R^2 = SSR/SST$$

Example: Child obesity data

- Example: A study on childhood obesity examined the relationship between a child's weight, age and exposure to pre-natal smoke.
 - Response: weight (kg)
 - Predictors: age, pre-natal smoke
- Linear regression model

$$\begin{aligned} Y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\ &= \beta_0 + \beta_1 A_i + \beta_2 S_i + \beta_3 A_i \times S_i + \epsilon_i \end{aligned}$$

where

- A_i : Age in years
- S_i : =1 if exposed; =0 if unexposed
- $A_i \times S_i$: interaction term

Example: Child obesity data

```
DATA Weights;
INPUT id wt age smoke obesity;
wt_kg = wt / 1000;
A_S = age * smoke;
datalines;
1 22509.41 7 0 1
2 33452.27 7 1 0
3 13380.91 3 0 1
4 24947.45 8 1 1
5 15875.65 4 1 1
. . .

proc reg;
model wt_kg = age smoke A_S;
run;
```

Linear regression for continuous responses

The REG Procedure
Model: MODEL1
Dependent Variable: wt_kg

| | |
|-----------------------------|----|
| Number of Observations Read | 30 |
| Number of Observations Used | 30 |

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 3 | 973.37997 | 324.45999 | 6.74 | 0.0016 |
| Error | 26 | 1250.94969 | 48.11345 | | |
| Corrected Total | 29 | 2224.32967 | | | |

| | | | |
|----------------|----------|----------|--------|
| Root MSE | 6.93639 | R-Square | 0.4376 |
| Dependent Mean | 23.22381 | Adj R-Sq | 0.3727 |
| Coeff Var | 29.86756 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 4.95865 | 6.65202 | 0.75 | 0.4627 |
| age | 1 | 2.87548 | 1.05916 | 2.71 | 0.0116 |
| smoke | 1 | 0.52160 | 8.58019 | 0.06 | 0.9520 |
| A_S | 1 | 0.20133 | 1.37335 | 0.15 | 0.8846 |

Hypothesis Test

- Test whether β s or linear combinations of β s have specific values:
 - $H_0 : \beta_1 = 0$
 - $H_0 : \beta_1 = \beta_2 = 0$
 - $H_0 : \beta_1 - \beta_2 = 1$
- Can be written as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b}$$

(most often $\mathbf{b} = \mathbf{0}$), where \mathbf{C} is of rank r .

- Test statistics:

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{b})^T \{\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T\}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{b}) / r}{\hat{\sigma}^2} \sim F_{r, n-p-1}.$$

- For $\mathbf{b} = \mathbf{0}$:

- If $\mathbf{C} = (0, \dots, 0, 1, 0, \dots)$, a single vector with $c_j = 1$ and $c_k = 0$ for all $k \neq j$. Then it is the same as the t -test for $\beta_j = 0$,

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim t_{n-p-1},$$

and $t^2 = F$ where $F \sim F_{1, n-p-1}$.

- If $\mathbf{C} = \text{diag}(0, 1, 1, \dots, 1)$. Then it is an overall F test (i.e. $\beta_1 = \dots = \beta_p = 0$).

- Hypothesis test using full and reduced model:

$$\begin{aligned}
 F &= \frac{SSR(\text{full}) - SSR(\text{reduced})/\Delta df}{SSE(\text{full})/(n - p - 1)} \\
 &\sim F_{\Delta df, n-p-1}
 \end{aligned}$$

or

$$\begin{aligned}
 F &= \frac{SSE(\text{reduced}) - SSE(\text{full})/\Delta df}{SSE(\text{full})/(n - p - 1)} \\
 &\sim F_{\Delta df, n-p-1}
 \end{aligned}$$

- Exactly same result as previous!

Example: Child obesity data

- Test for the main and interaction effect of Smoke.
 - $H_0: \beta_2 = \beta_3 = 0$
- Use the contrast matrix

$$C =$$

- Use full and reduced models
 - Full Model:
 - Reduced Model:

Example: Child obesity data

- Use the contrast matrix:

```
proc reg;  
model wt_kg = age smoke A_S;  
age: test smoke=0, A_S=0;  
run;
```

- Fit the full and the reduced models:

```
proc reg;  
model wt_kg = age smoke A_S;  
run;  
proc reg;  
model wt_kg = age ;  
run;
```

Use the contrast matrix

Friday, January 8, 2016 12:54:50 PM 7

**The REG Procedure
Model: MODEL1**

| Test age Results for Dependent Variable wt_kg | | | | |
|---|----|-------------|---------|--------|
| Source | DF | Mean Square | F Value | Pr > F |
| Numerator | 2 | 9.75141 | 0.20 | 0.8178 |
| Denominator | 26 | 48.11345 | | |

Fit the full and the reduced models

Friday, January 8, 2016 12:54:50 PM 8

The REG Procedure Model: MODEL1 Dependent Variable: wt_kg

| | |
|-----------------------------|----|
| Number of Observations Read | 30 |
| Number of Observations Used | 30 |

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 3 | 973.37997 | 324.45999 | 6.74 | 0.0016 |
| Error | 26 | 1250.94969 | 48.11345 | | |
| Corrected Total | 29 | 2224.32967 | | | |

| | | | |
|----------------|----------|----------|--------|
| Root MSE | 6.93639 | R-Square | 0.4376 |
| Dependent Mean | 23.22381 | Adj R-Sq | 0.3727 |
| Coeff Var | 29.86756 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 4.95865 | 6.65202 | 0.75 | 0.4627 |
| age | 1 | 2.87548 | 1.05916 | 2.71 | 0.0116 |
| smoke | 1 | 0.52160 | 8.58019 | 0.06 | 0.9520 |
| A_S | 1 | 0.20133 | 1.37335 | 0.15 | 0.8846 |

Fit the full and the reduced models

Friday, January 8, 2016 12:54:50 PM 11

The REG Procedure
Model: MODEL1
Dependent Variable: wt_kg

| | |
|-----------------------------|----|
| Number of Observations Read | 30 |
| Number of Observations Used | 30 |

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 953.87715 | 953.87715 | 21.02 | <.0001 |
| Error | 28 | 1270.45252 | 45.37330 | | |
| Corrected Total | 29 | 2224.32967 | | | |

| | | | |
|----------------|----------|----------|--------|
| Root MSE | 6.73597 | R-Square | 0.4288 |
| Dependent Mean | 23.22381 | Adj R-Sq | 0.4084 |
| Coeff Var | 29.00459 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 5.41375 | 4.07439 | 1.33 | 0.1947 |
| age | 1 | 3.00170 | 0.65467 | 4.59 | <.0001 |

Example: Child obesity data

- Use the contrast matrix:
 - Test statistics:
 - Null distribution:
- Use the full and the reduced model:
 - Test statistics:
 - Null distribution:

Diagnostics: Violations of Assumptions

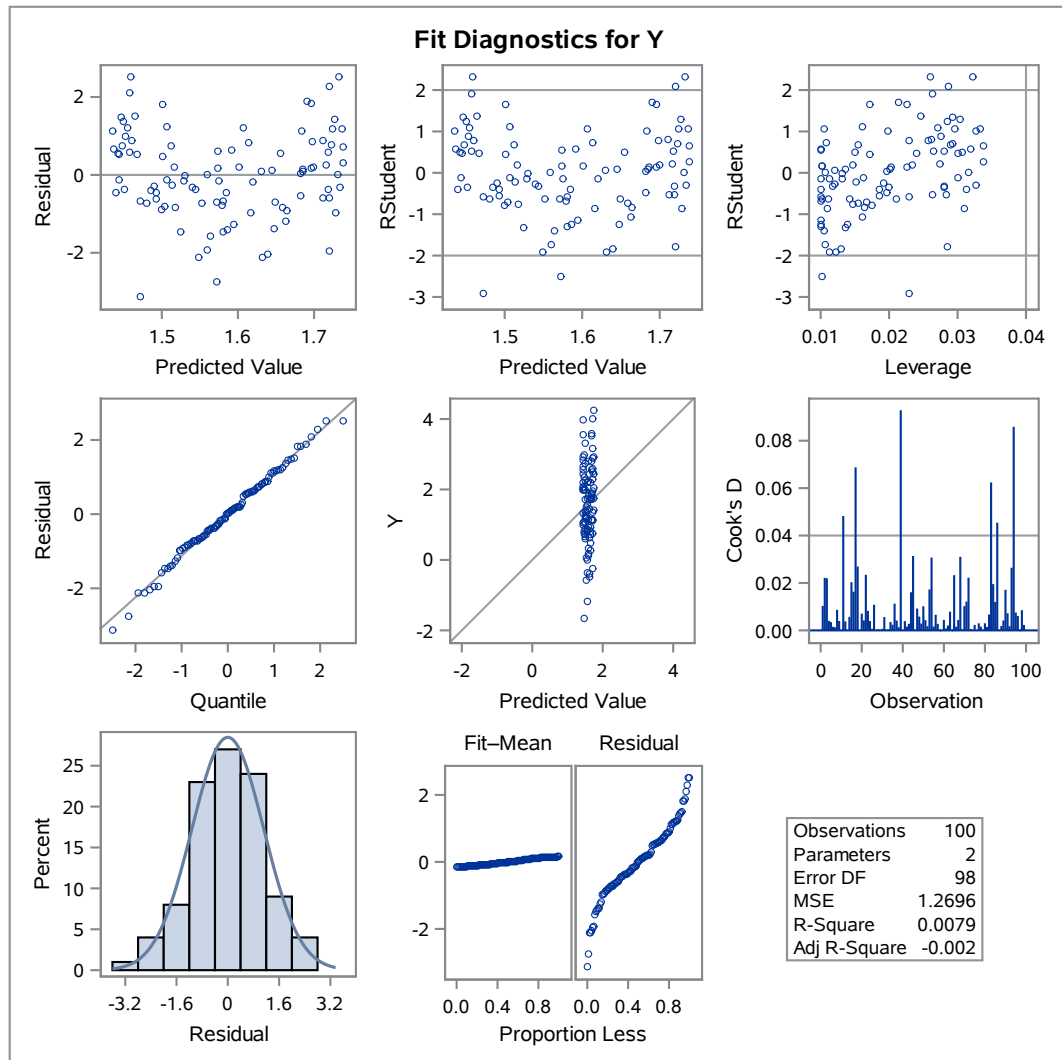
- Assumptions:
 - Linearity: $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$
 - Normality
 - Equal variance (homoscedasticity)
 - Independence

Diagnostics: Violations of Assumptions

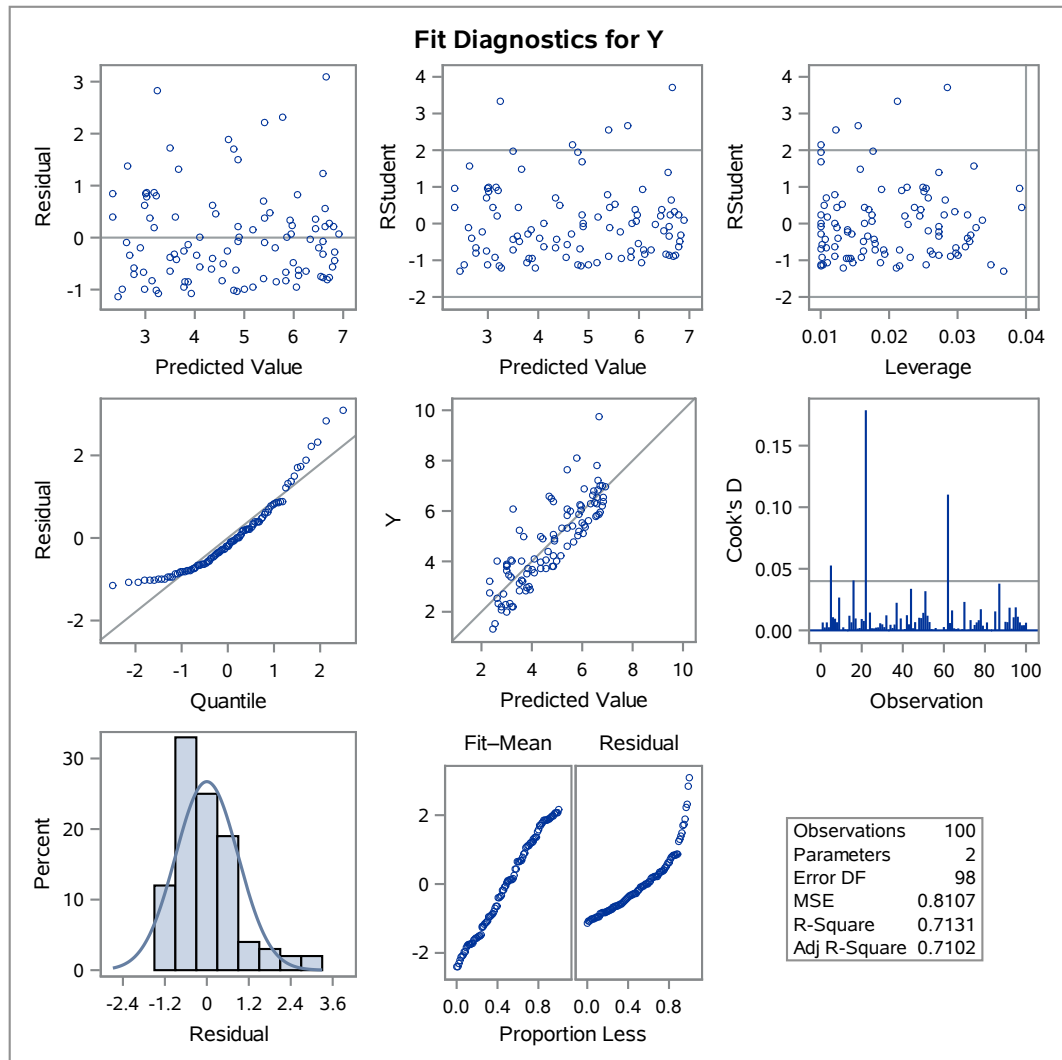
- Linearity:
 - Check: Partial regression plot, residual plot.
 - Remedy: Transformation, Add another regressor (ex. add x^2)
- Normality:
 - Check: Normal quantile plot, Statistical tests for normality (ex. Shapiro-Wilk test)
 - Remedy: Transformation, GLM
- Equal variance:
 - Check: Residual plot
 - Remedy: Transformation (ex. log), Weighted Least Square, GLM
- Independence:
 - Check: Done by intuition (e.g., repeatedly measured..), Residual plots
 - Remedy: Longitudinal, Time series

- Violation of which assumption?

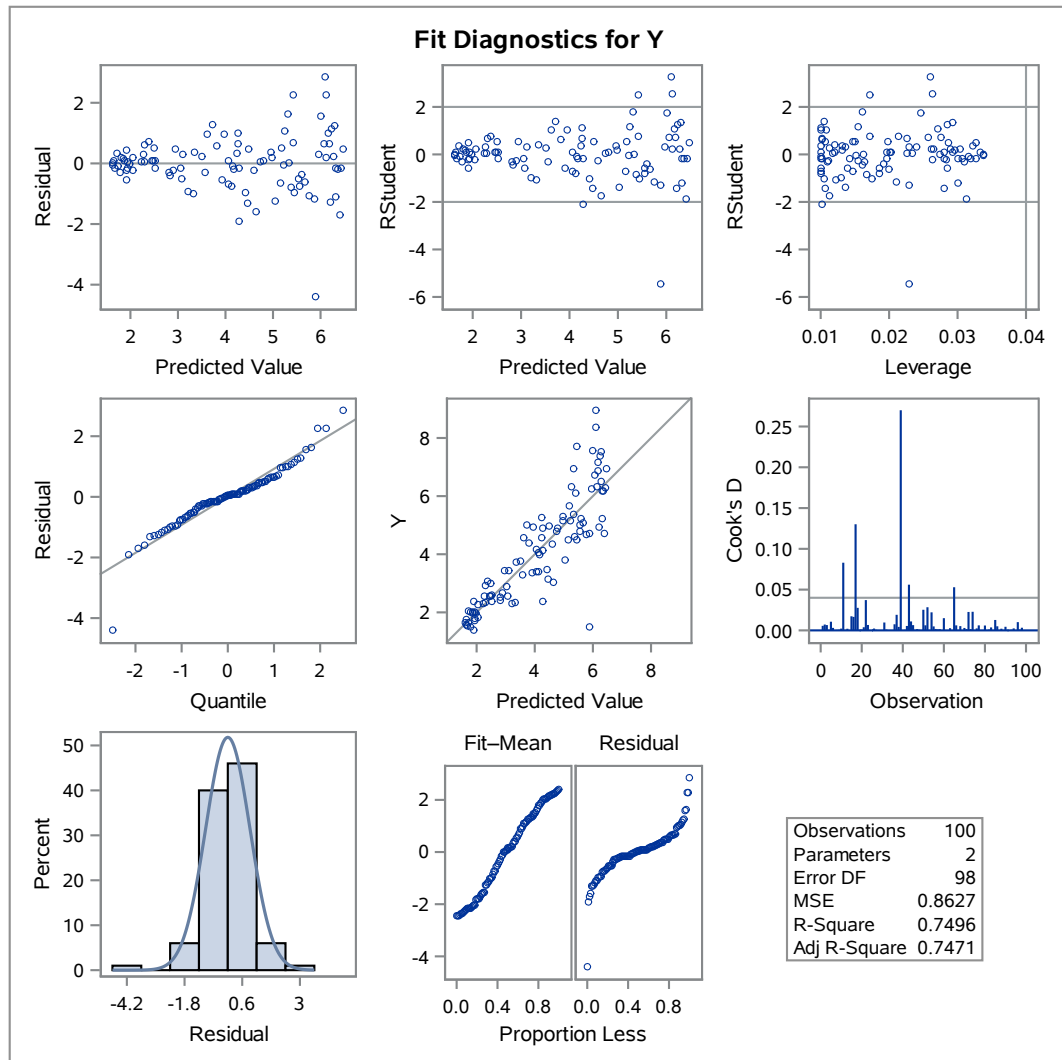
The REG Procedure
Model: MODEL1
Dependent Variable: Y



The REG Procedure
Model: MODEL1
Dependent Variable: Y



The REG Procedure
Model: MODEL1
Dependent Variable: Y



Weighted Least Squares

- Suppose that each observation has difference variance (heteroscedasticity):

$$\sigma_1 \neq \sigma_2 \neq \cdots \neq \sigma_n$$

$$V = \text{diag}[\sigma_1^2, \dots, \sigma_n^2]$$

- Grouped (Aggregate) data: Y_i is an average of n_i observations:

$$Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} E_{ij}, \quad \text{Var}(E_{ij}) = \sigma^2,$$

and then

$$\text{Var}(Y_i) = \sigma_i^2 = \sigma^2/n_i$$

- Count data: variance increases as the mean increases:

$$\sigma_i^2 \approx \mu_i \sigma^2$$

- Original model:

$$Y = X\beta + \epsilon$$

- Transformed model:

$$\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon}$$

where $\tilde{Y} = V^{-1/2}Y$, $\tilde{X} = V^{-1/2}X$ and $\tilde{\epsilon} = V^{-1/2}\epsilon$

- Satisfy the equal variance assumption

$$\begin{aligned} Var(\tilde{\epsilon}) &= V^{-1/2}Var(\epsilon)V^{-1/2} \\ &= V^{-1/2}VV^{-1/2} = I \end{aligned}$$

- Weighted Least Squares (WLS) estimator

$$\begin{aligned} \beta_{wls} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y} \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad (1) \end{aligned}$$

Example: Apple shots

- Researchers recorded the average number of stem shots in apple trees in each day. Varying numbers of trees n_i are observed in each day.
- Model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2/n_i)$$

- Response: $Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} E_{ij}$
- E_{ij} : number of stem shoots from the j th tree on the i th day of the growing season.
- x_i : number of days since dormancy.

Example: Apple shots

```
data apple;  
input day ni Y;  
cards;  
0 5 10.2  
3 5 10.4  
7 5 10.6  
13 6 12.5  
18 5 12.0  
24 4 15.0  
25 6 15.17  
32 5 17.0  
38 7 18.71  
.
```


Example: Apple shots

```
proc reg data=apple;  
model Y = day;  
weight ni;  
run;
```

- SAS will construct a weight matrix equals to

$$V^{-1} = \begin{pmatrix} w_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n \end{pmatrix}$$

where $w_i = n_i$ in this example.

Example: Apple shots

- Use IML to estimate β

```
proc iml;
  use apple;
  read all var {Y} into Y;
  read all var {day} into X_1;
  read all var {ni} into W;

  n=nrow(Y);
  one_n=j(n,1,1);
  X=one_n||X_1;
  V_inv=DIAG(W);

  beta=inv(t(X)*V_inv*X)*t(X)*V_inv * Y;

  print beta;

quit;
```

PROC REG

Wednesday, January 6,

The REG Procedure
Model: MODEL1
Dependent Variable: Y

| | |
|-----------------------------|----|
| Number of Observations Read | 22 |
| Number of Observations Used | 22 |

Weight: ni

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 6164.27627 | 6164.27627 | 1657.24 | <.0001 |
| Error | 20 | 74.39209 | 3.71960 | | |
| Corrected Total | 21 | 6238.66835 | | | |

| | | | |
|----------------|----------|----------|--------|
| Root MSE | 1.92863 | R-Square | 0.9881 |
| Dependent Mean | 21.42212 | Adj R-Sq | 0.9875 |
| Coeff Var | 9.00297 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 9.97375 | 0.31427 | 31.74 | <.0001 |
| day | 1 | 0.21733 | 0.00534 | 40.71 | <.0001 |

IML

| beta |
|-----------|
| 9.9737537 |
| 0.2173303 |

Weighted Least Squares

- Important technique in linear regression to address for heteroscedasticity.
- GLM model: WLS is used to estimate parameters
 - Iteratively Reweighted Least Squares (IRWLS)

BIOSTAT 651
Notes #3: Maximum Likelihood

- Lecture Topics:
 - Maximum likelihood estimation (MLE)
 - Hypothesis testing

Data Structure

- The general set-up is described as follows:
 - sample size: n subjects (independent)
 - response: Y_i
 - covariate $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots)$
 - model parameters: $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^T$
 - set $\mathbf{Y} = (Y_1, \dots, Y_n)^T$
 - design matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

- e.g., linear regression: $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$

Parameter Estimation

- Consider a model of Y_i based on the parameter θ
 - observed data: (\mathbf{x}_i^T, Y_i) for $i = 1, \dots, n$
 - fitted values: \hat{Y}_i
- Different choices of $\hat{\theta}$ will yield different $\hat{\mathbf{Y}}$

Q: How to select the “best” $\hat{\theta}$?
- In linear regression we used LSE, which minimize the following function:

$$S_2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- Many other possible criteria exist:

$$S_1 = \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

$$S_\infty = \max_{i=1, \dots, n} |Y_i - \hat{Y}_i|$$

- Another well-known method: Maximum Likelihood

Likelihood

- density: $f(Y_i; \boldsymbol{\theta})$
- joint density: $f(\mathbf{Y}; \boldsymbol{\theta}) = \prod_{i=1}^n f(Y_i; \boldsymbol{\theta})$
 - calculation based on various \mathbf{Y} values, for fixed $\boldsymbol{\theta}$
- likelihood function: $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Y})$
 - often abbreviated to $L(\boldsymbol{\theta})$, or even L
 - viewed as a function of $\boldsymbol{\theta}$, with (\mathbf{X}, \mathbf{Y}) held constant (at their realized values)
- likelihood is proportional to the joint density,

$$L(\boldsymbol{\theta}) \propto f(\mathbf{Y}; \boldsymbol{\theta})$$

- derived by setting $L(\boldsymbol{\theta}) = f(\mathbf{Y}; \boldsymbol{\theta})$, then deleting multiples that are *not* functions of $\boldsymbol{\theta}$

Likelihood Principles (continued)

- Assuming that Y_1, \dots, Y_n are independent,

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^n f_i(Y_i; \boldsymbol{\theta}),$$

- if, in addition, the Y_i 's are identically distributed,

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^n f(Y_i; \boldsymbol{\theta}),$$

- in most cases we consider, the (\mathbf{x}_i^T, Y_i) will be independent and identically distributed

Maximum Likelihood Estimators (MLE)

- A *Maximum Likelihood Estimator* (MLE) is a maximizer of the likelihood function $L(\boldsymbol{\theta}|\mathbf{Y})$, denoted as $\hat{\boldsymbol{\theta}}$, i.e.

$$L(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

where Θ is the parameter space

- Note: MLE is also an maximizer of the log-likelihood, $\ell(\boldsymbol{\theta})$.
- For a given parametric model, maximum likelihood identifies the parameter values which make the realized data “most likely”

MLE: Functions

- For convenience, we often maximize the log-likelihood function:

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$$

- score function,

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta})$$

- *observed* information,

$$J(\boldsymbol{\theta}) = \frac{-\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \ell(\boldsymbol{\theta})$$

- *expected* information,

$$I(\boldsymbol{\theta}) = E[J(\boldsymbol{\theta})] = -E \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \ell(\boldsymbol{\theta}) \right]$$

- $J(\boldsymbol{\theta})$ may be easier to calculate than $I(\boldsymbol{\theta})$
- In the book, \mathfrak{J} represents the expected information.

Score Function

- When $\ell(\boldsymbol{\theta})$ is differentiable w.r.t. $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ can typically be obtained as the solution to the score equation, $U(\boldsymbol{\theta}) = \mathbf{0}$, where

$$U(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_q} \end{bmatrix}$$

- This will work if $J(\boldsymbol{\theta})$ is positive-definite, where

$$J(\boldsymbol{\theta}) = - \begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_q \partial \theta_1} & \cdots & \frac{\partial^2 \ell}{\partial \theta_q \partial \theta_q} \end{bmatrix}$$

Information Matrix

- Expected information is calculated as:

$$I(\boldsymbol{\theta}) = -E \begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_q \partial \theta_1} & \cdots & \frac{\partial^2 \ell}{\partial \theta_q \partial \theta_q} \end{bmatrix}$$

- We then have

$$J(\boldsymbol{\theta}) = -\frac{\partial U^T}{\partial \boldsymbol{\theta}}$$
$$I(\boldsymbol{\theta}) = -E \left[\frac{\partial U^T}{\partial \boldsymbol{\theta}} \right]$$

MLE: Functions (cont'd)

- In the *iid* setting, we can write,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n L_i(\boldsymbol{\theta})$$

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$$

$$U(\boldsymbol{\theta}) = \sum_{i=1}^n U_i(\boldsymbol{\theta})$$

$$I(\boldsymbol{\theta}) = \sum_{i=1}^n I_i(\boldsymbol{\theta})$$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^n J_i(\boldsymbol{\theta})$$

MLE: Score and Information

- It can be shown that,

$$\begin{aligned}E[U(\boldsymbol{\theta}_0)] &= \mathbf{0} \\V[U(\boldsymbol{\theta}_0)] &= E[U(\boldsymbol{\theta}_0)^{\otimes 2}] = I(\boldsymbol{\theta}_0),\end{aligned}$$

where $\boldsymbol{\theta}_0$ is the true underlying value of $\boldsymbol{\theta}$ and $\mathbf{z}^{\otimes 2} = \mathbf{z}\mathbf{z}^T$

- Note:

$$\begin{aligned}V[U(\boldsymbol{\theta}_0)] &= E[U(\boldsymbol{\theta}_0)^{\otimes 2}] \\&= E\left[\sum_{i=1}^n U_i(\boldsymbol{\theta}_0) \sum_{j=1}^n U_j(\boldsymbol{\theta}_0)^T\right] \\&= E\left[\sum_{i=1}^n U_i(\boldsymbol{\theta}_0)^{\otimes 2}\right] \\&= nE[U_1(\boldsymbol{\theta}_0)^{\otimes 2}] \\&= nI_1(\boldsymbol{\theta}_0)\end{aligned}$$

Maximum Likelihood Estimation

- Maximum likelihood estimator, $\hat{\boldsymbol{\theta}}$, computed by solving the score equation,

$$U(\boldsymbol{\theta}) = \mathbf{0}$$

- Note: maximizer may lie on the boundary of $\boldsymbol{\Theta}$, in which case the MLE is ill-behaved.
 - in BIOSSTAT 651, we assume that $\ell(\boldsymbol{\theta})$ is *concave*, with information matrix assumed to be positive-definite

MLE Example: Normal

- Example: Suppose that $Y_i \sim N(\mu, \sigma^2)$ with σ^2 known. Determine the MLE of μ .

$$f(Y_i; \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(Y_i - \mu)^2 / (2\sigma^2)}$$

$$L_i(\mu) = e^{-(Y_i - \mu)^2 / (2\sigma^2)}$$

$$\ell_i(\mu) = -(Y_i - \mu)^2 / (2\sigma^2)$$

$$U_i(\mu) = (Y_i - \mu) / \sigma^2$$

$$U(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \mu)$$

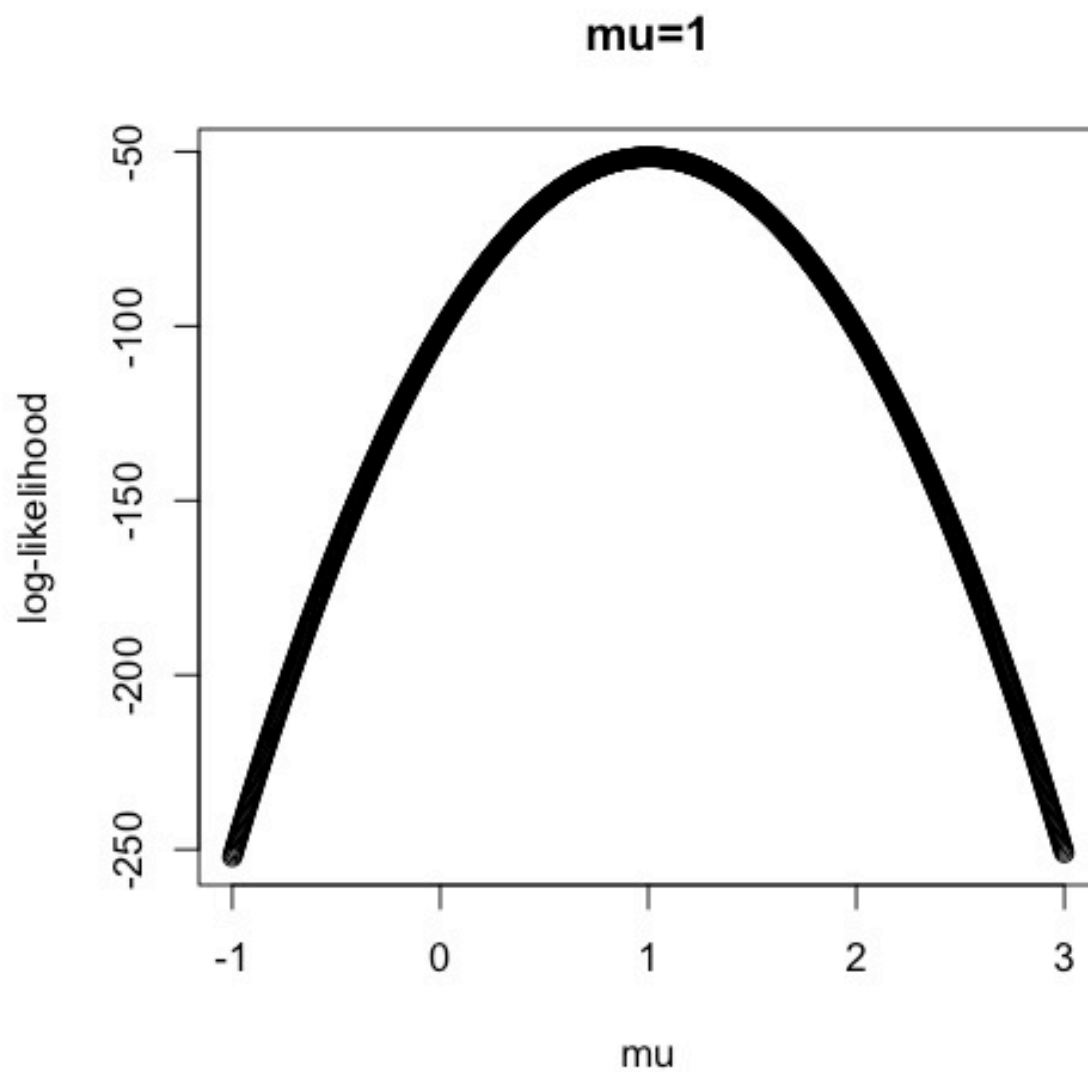
$$\hat{\mu} = \bar{Y}$$

- Note:

$$J(\mu) = I(\mu) = -\frac{\partial U}{\partial \mu} = \frac{n}{\sigma^2}$$

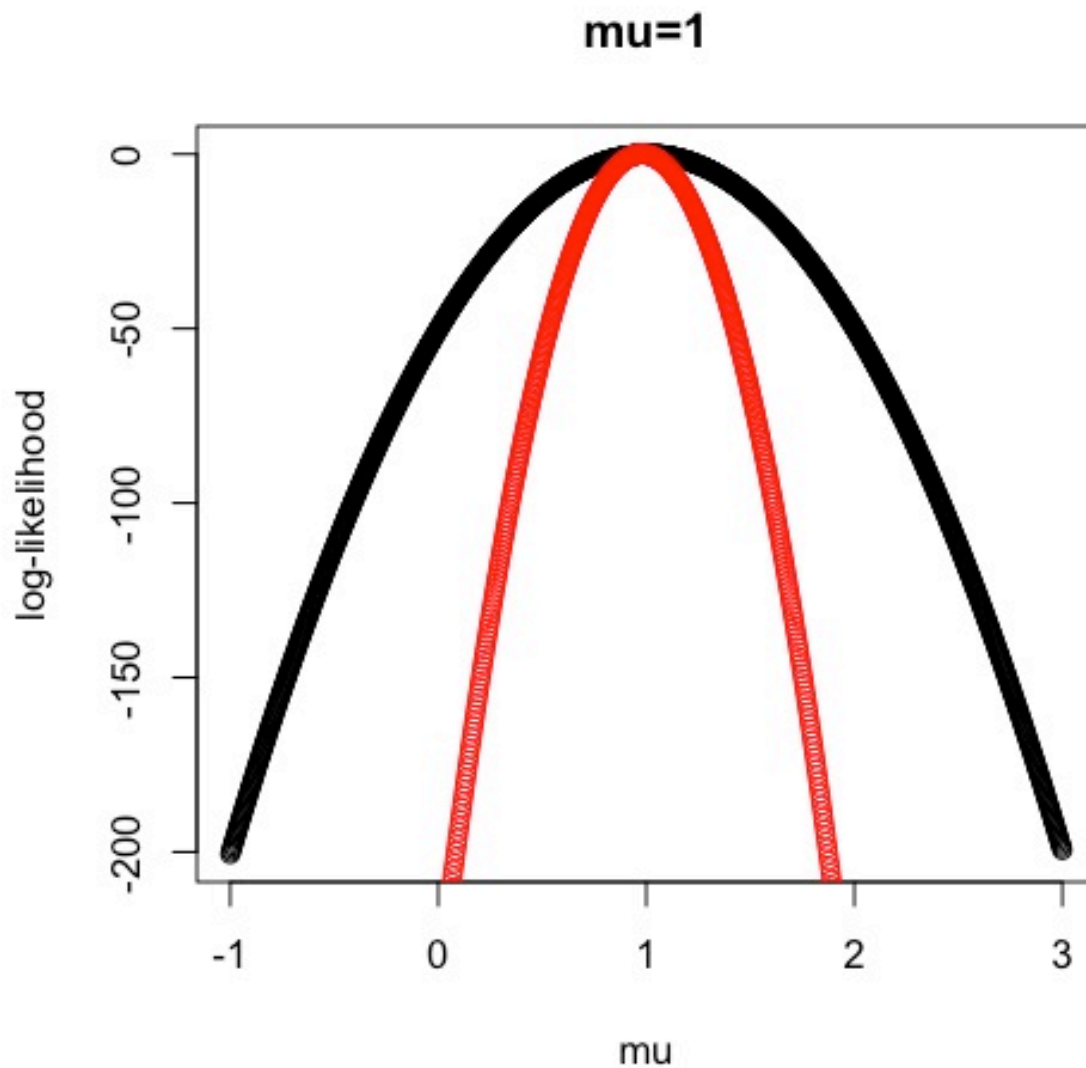
MLE Example: Normal

- Log likelihood function ($n=100$)



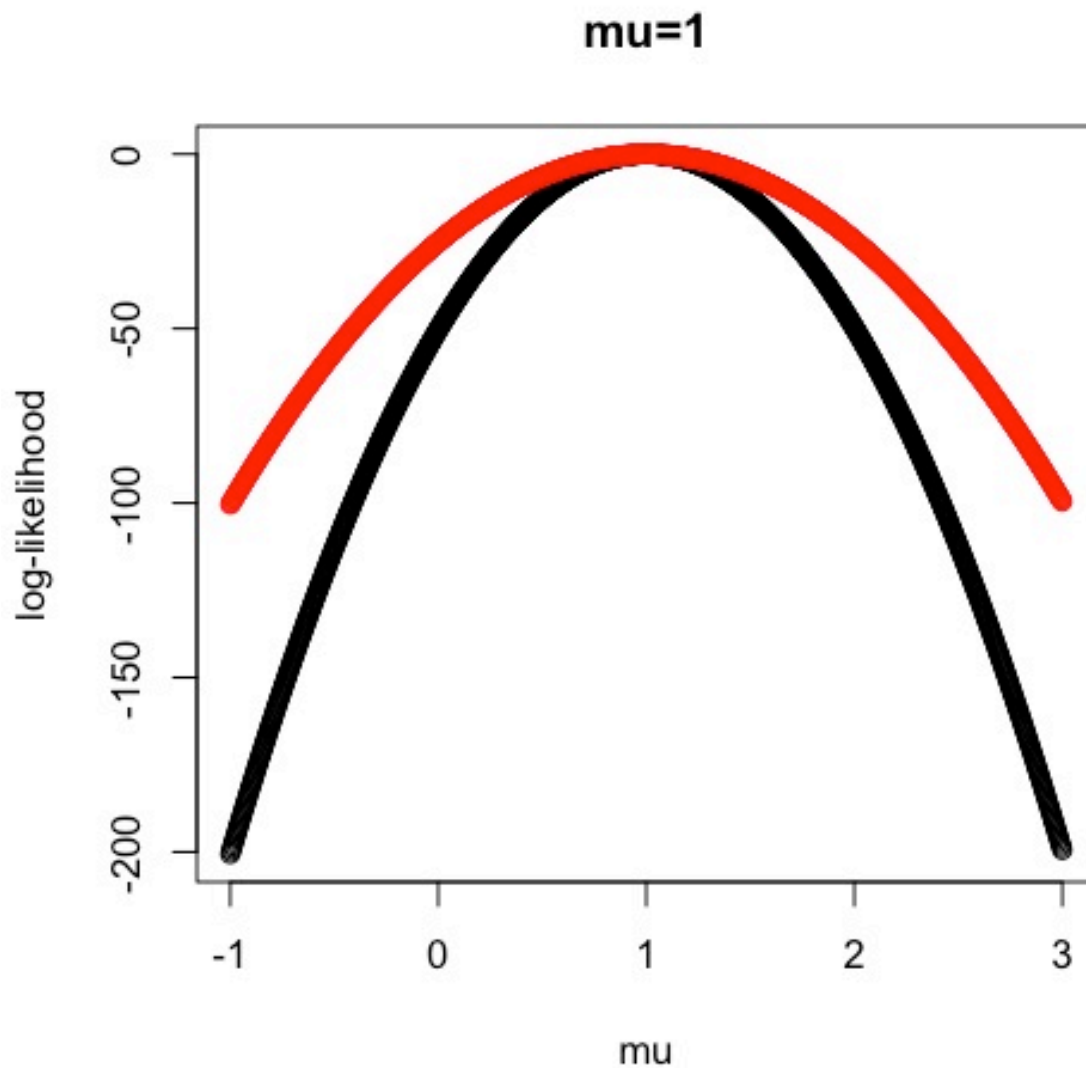
MLE Example: Normal - Fisher Information

- $J(\mu) = I(\mu) = -\frac{\partial U}{\partial \mu} = \frac{n}{\sigma^2}$
- $n = 100$ (black) vs. $n = 500$ (red)



MLE Example: Normal - Fisher Information

- $J(\mu) = I(\mu) = -\frac{\partial U}{\partial \mu} = \frac{n}{\sigma^2}$
- $\sigma^2 = 1$ (black) vs. $\sigma^2 = 2$ (red)



MLE Example: Binomial

- Example: Suppose that $Y_{\bullet} = Y_1 + \dots + Y_n$ follows a Binomial distribution with parameter π . Compute the MLE of π .

$$p(Y; \pi) = \binom{n}{Y} \pi^Y (1 - \pi)^{n-Y}$$

$$L(\pi) = \pi^Y (1 - \pi)^{n-Y}$$

$$\ell(\pi) = Y \log(\pi) + (n - Y) \log(1 - \pi)$$

$$U(\pi) = \frac{Y}{\pi} - \frac{n - Y}{1 - \pi}$$

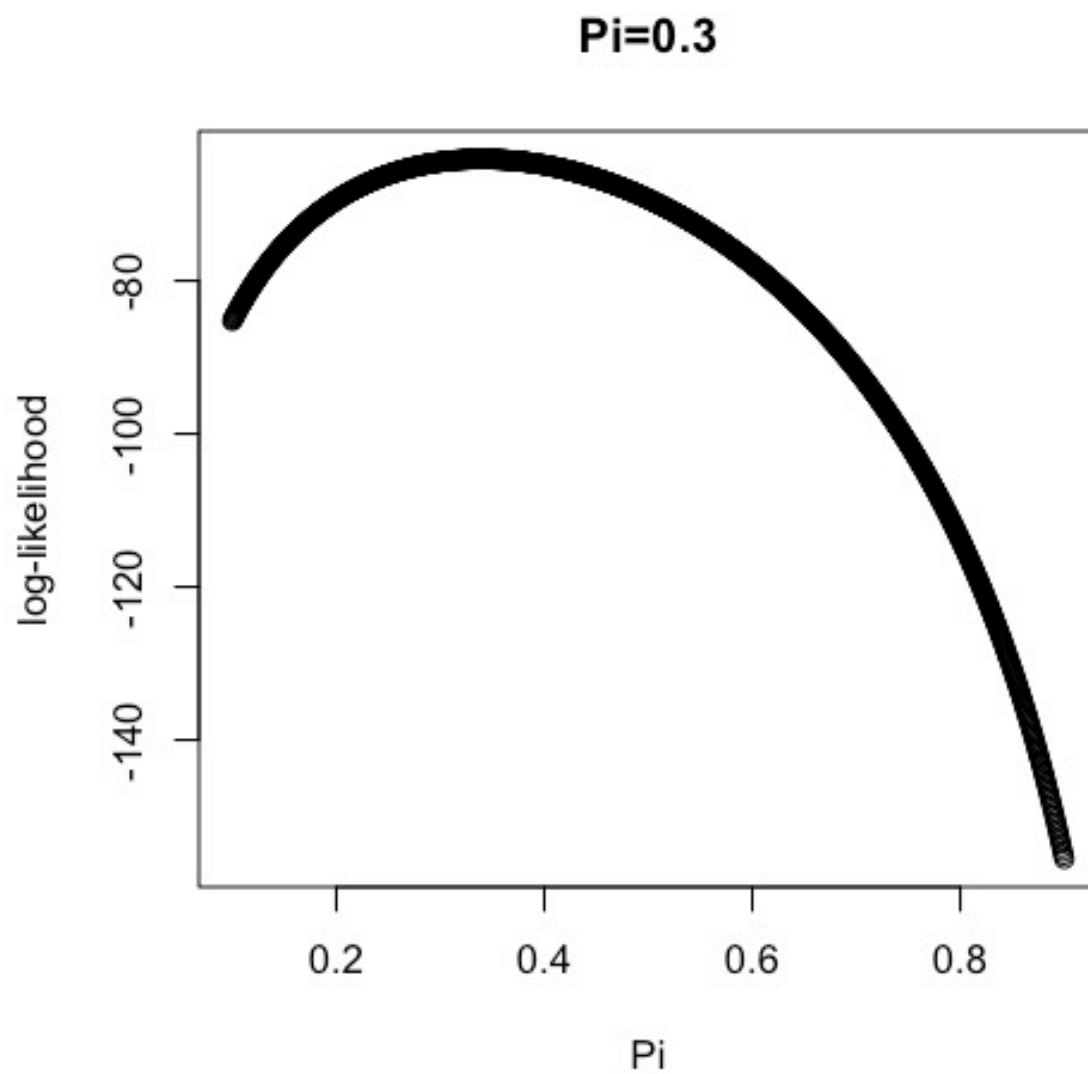
$$\hat{\pi} = \bar{Y}$$

- Second derivative,

$$\frac{\partial U}{\partial \pi} = \frac{-Y}{\pi^2} - \frac{n - Y}{(1 - \pi)^2}$$

MLE Example: Binomial

- Log likelihood function ($n=100$)



MLE Example: Poisson Case

- Example: Suppose that Y_i is distributed as $\text{Poisson}(\theta)$ for $i = 1, \dots, n$. Determine the maximum likelihood estimator of θ .

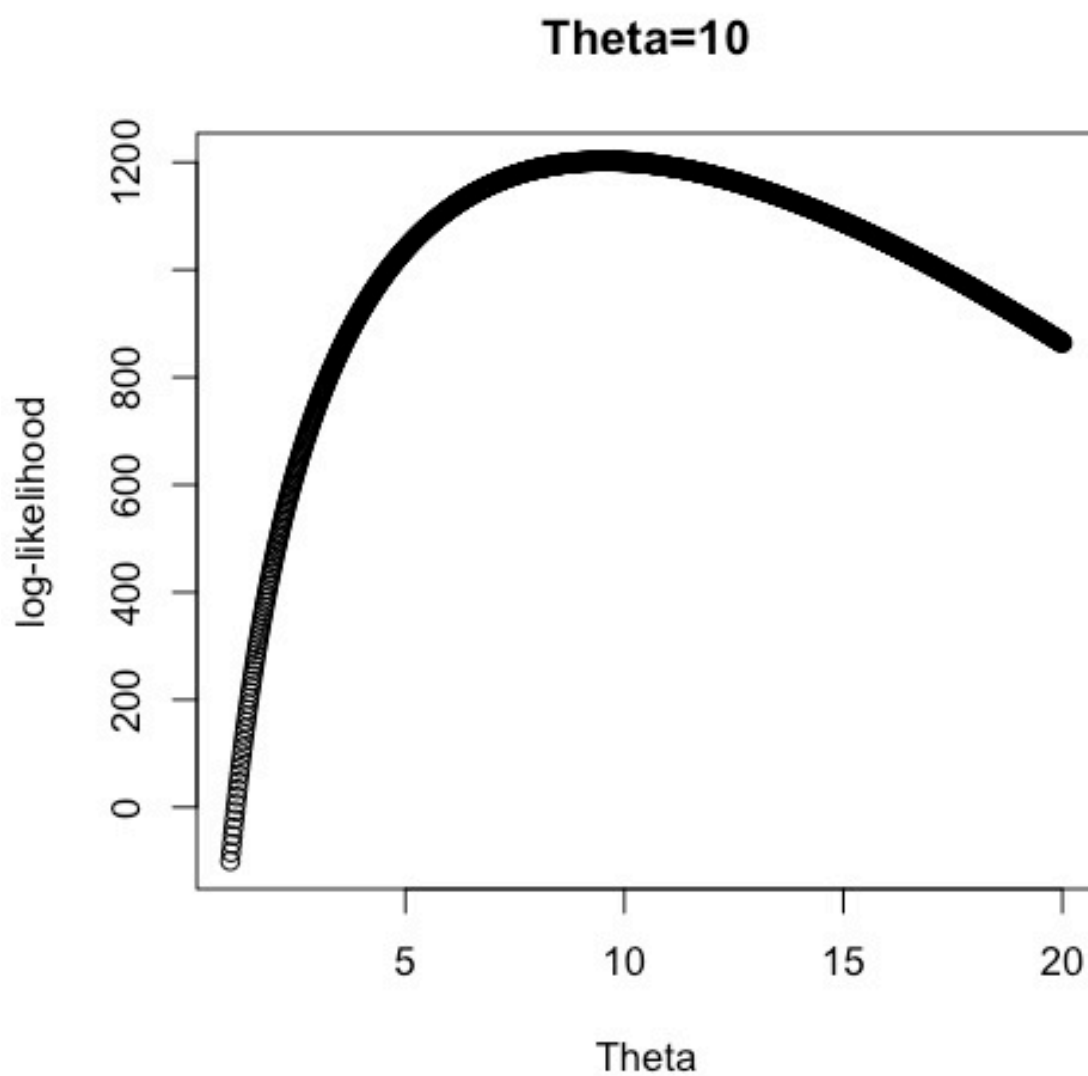
$$\begin{aligned}f(Y_i; \theta) &= \frac{e^{-\theta} \theta^{Y_i}}{Y_i!} \\L_i(\theta) &= e^{-\theta} \theta^{Y_i} \\\ell_i(\theta) &= -\theta + Y_i \log \theta \\U_i(\theta) &= -1 + \frac{Y_i}{\theta} \\J_i(\theta) &= \frac{Y_i}{\theta^2} \\U(\theta) &= \dots \\\hat{\theta} &= \dots\end{aligned}$$

- Expected and observed information:

$$\begin{aligned}J(\theta) &= \\I(\theta) &= \end{aligned}$$

MLE Example: Poisson

- Log likelihood function ($n=100$)



Maximum Likelihood Estimation

- Usually, a closed-form solution for $\hat{\boldsymbol{\theta}}$ is not available
 - ex) Logistic regression
- Need to solve $U(\boldsymbol{\theta}) = \mathbf{0}$ through iterative methods

e.g., Newton-Raphson ...

Newton-Raphson Procedure

- Pre-specify tolerance, ξ ; start with an initial “estimate”, $\hat{\boldsymbol{\theta}}_{(0)}$, and
 - e.g., $\hat{\boldsymbol{\theta}}_{(0)} = \mathbf{0}$, with $\xi = 10^{-4}$

- Update the estimate,

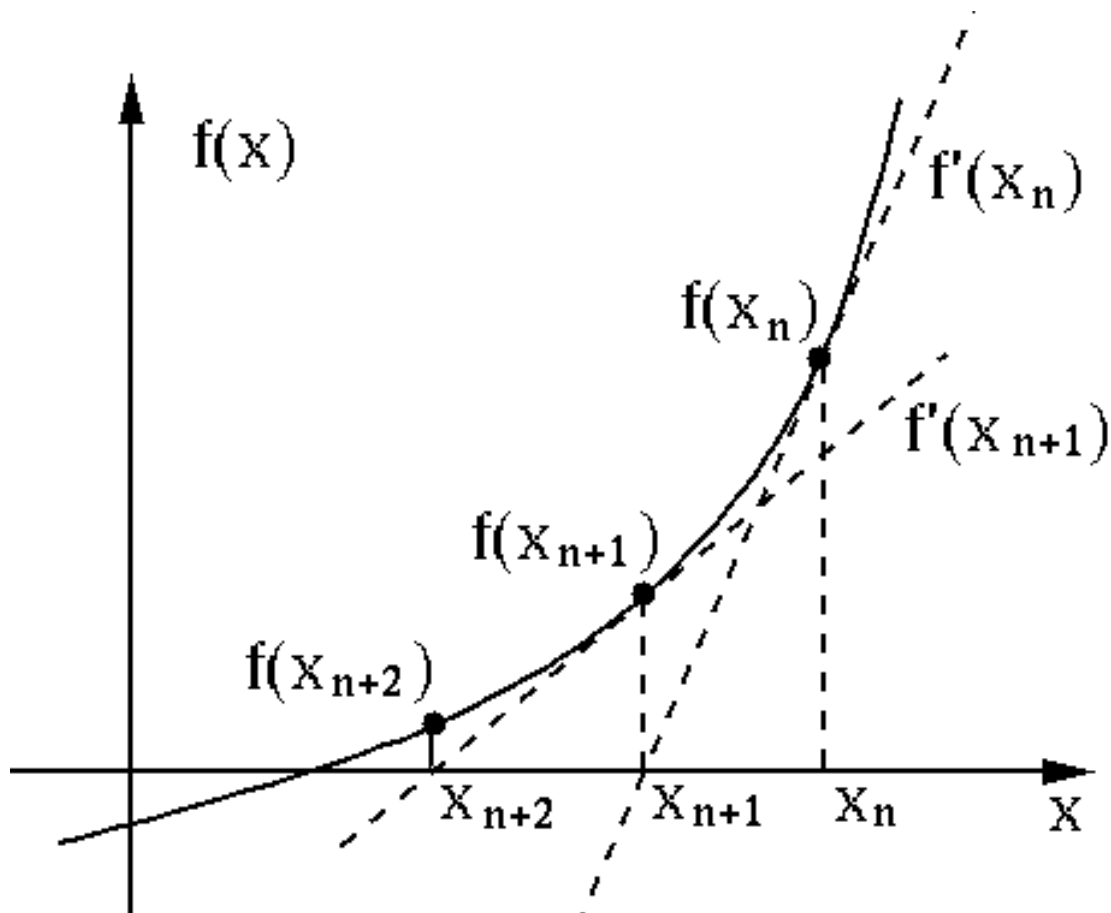
$$\hat{\boldsymbol{\theta}}_{(j+1)} = \hat{\boldsymbol{\theta}}_{(j)} + J^{-1}(\hat{\boldsymbol{\theta}}_{(j)})U(\hat{\boldsymbol{\theta}}_{(j)})$$

- Continue until convergence is attained; e.g.,

$$\begin{aligned} \|\hat{\boldsymbol{\theta}}_{(j+1)} - \hat{\boldsymbol{\theta}}_{(j)}\| &< \xi \\ \|U(\hat{\boldsymbol{\theta}}_{(j)})\| &< \xi \end{aligned}$$

where $\|\mathbf{z}\| = (\mathbf{z}^T \mathbf{z})^{1/2}$

Newton-Raphson Procedure



From

fourier.eng.hmc.edu/e176/lectures/NM/node20.html

Properties of MLEs

Properties of MLEs

- Under certain regularity conditions:

- $\hat{\boldsymbol{\theta}}$ is the unique maximizer of $\ell(\boldsymbol{\theta})$
- $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$

$$\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$$

- $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to a mean zero Normal with a covariance $I_1(\boldsymbol{\theta}_0)^{-1}$

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Rightarrow N(0, I_1(\boldsymbol{\theta}_0)^{-1})$$

- $n^{-1}J(\hat{\boldsymbol{\theta}}) \xrightarrow{p} I_1(\boldsymbol{\theta}_0)$

Invariance Property

- If $\hat{\boldsymbol{\theta}}$ is the MLE for $\boldsymbol{\theta}_0$, then $g(\hat{\boldsymbol{\theta}})$ will be the MLE for $g(\boldsymbol{\theta}_0)$
 - i.e, assuming $g(\cdot)$ is a well-behaved function
 - e.g., continuous (differentiable)
- Application: depending on the specifics of the likelihood, it may be more convenient to maximize $L(g(\boldsymbol{\theta}))$ than $L(\boldsymbol{\theta})$
 - obtain $\widehat{g(\boldsymbol{\theta})}$,
then obtain $\hat{\boldsymbol{\theta}} = g^{-1}\{\widehat{g(\boldsymbol{\theta})}\}$
- e.g., set $g(\boldsymbol{\theta}) = \log \boldsymbol{\theta}$

MLE: Interval Estimation

- Given the afore-listed large-sample properties of MLEs, interval estimators can be computed using the Normal approximation ...
 - e.g., 95% confidence interval for θ_0 :
 - e.g., 95% confidence interval for $g(\theta_0)$:
- Note: could use *Delta Method* to compute interval estimate of a function of θ_0
 - e.g., 95% confidence interval for $g(\theta_0)$:

Example: CI, Normal

- Example: We return to the case where $Y_i \sim N(\mu, \sigma^2)$ with σ^2 known. Compute a 95% confidence interval for μ .

Recall that

$$\begin{aligned}\hat{\mu} &= \bar{Y} \\ \frac{\partial U}{\partial \mu} &= \frac{-n}{\sigma^2}\end{aligned}$$

We then have

$$\begin{aligned}J(\mu) &= \\ V(\hat{\mu}) &= \end{aligned}$$

such that a 95% CI is then given by:

CI Example: Binomial

- Example: We revisit the setting in which $Y_{\bullet} = Y_1 + \dots + Y_n$ follows a Binomial distribution with parameter π . Compute an interval estimate of π .

Recall that:

$$\begin{aligned}\hat{\pi} &= \bar{Y} \\ \frac{\partial U}{\partial \pi} &= \frac{-Y}{\pi^2} - \frac{n - Y}{(1 - \pi)^2}\end{aligned}$$

such that

$$\begin{aligned}J(\pi) &= \frac{Y}{\pi^2} + \frac{n - Y}{(1 - \pi)^2} \\ I(\pi) &= \frac{n}{\pi} + \frac{n}{(1 - \pi)} = \frac{n}{\pi(1 - \pi)}\end{aligned}$$

- Therefore, the CI is given by:

CI Example: Binomial (continued)

- Q1: What is one limitation of the CI just derived?
- Q2: How to remedy?

CI Example: Poisson Case

- Example: Determine an interval estimator for the case where Y_i is distributed as $\text{Poisson}(\theta)$ for $i = 1, \dots, n$.

- Based on previous calculations,

$$\begin{aligned}\hat{\theta} &= \bar{Y} \\ I(\theta) &= \frac{n}{\theta}\end{aligned}$$

such that we obtain the 95% CI as

- Q1: Problem with this estimator?
- Q2: Solution?

Hypothesis Testing

- For the next few slides, we consider the following setting:
 - let $\boldsymbol{\theta}$ be a $q \times 1$ parameter, partitioned as, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are of dimension $q_1 \times 1$ and $q_2 \times 1$, respectively
 - we wish to test $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1H}$ versus $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{1H}$
 - estimation is based on ML
 - let $\hat{\boldsymbol{\theta}}_H$ be the MLE, constrained by H_0 ; i.e., $\hat{\boldsymbol{\theta}}_H = (\boldsymbol{\theta}_{1H}^T, \hat{\boldsymbol{\theta}}_{2H}^T)^T$,
- Three most commonly used tests: Score, Wald and Likelihood ratio

Score Test

- Score test makes use of asymptotic result that $U(\boldsymbol{\theta}_0) \sim N(\mathbf{0}, I(\boldsymbol{\theta}_0))$
- Score test statistic:

$$U(\hat{\boldsymbol{\theta}}_H)^T J(\hat{\boldsymbol{\theta}}_H)^{-1} U(\hat{\boldsymbol{\theta}}_H) \sim \chi_{q_1}^2$$

- Properties:
 - only the restricted (H_0) MLE is computed, not the unrestricted (H_1)
 - computationally very fast

Wald Test

- The Wald test exploits the result that $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_0, I(\boldsymbol{\theta}_0)^{-1})$

- Wald statistic

$$(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1H})^T V(\hat{\boldsymbol{\theta}}_1)^{-1} (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1H}) \sim \chi_{q_1}^2$$

- Properties:
 - most intuitive
 - only the unrestricted (or “full model”) MLE is computed
- Most frequently used test, especially when $q_1 = 1$

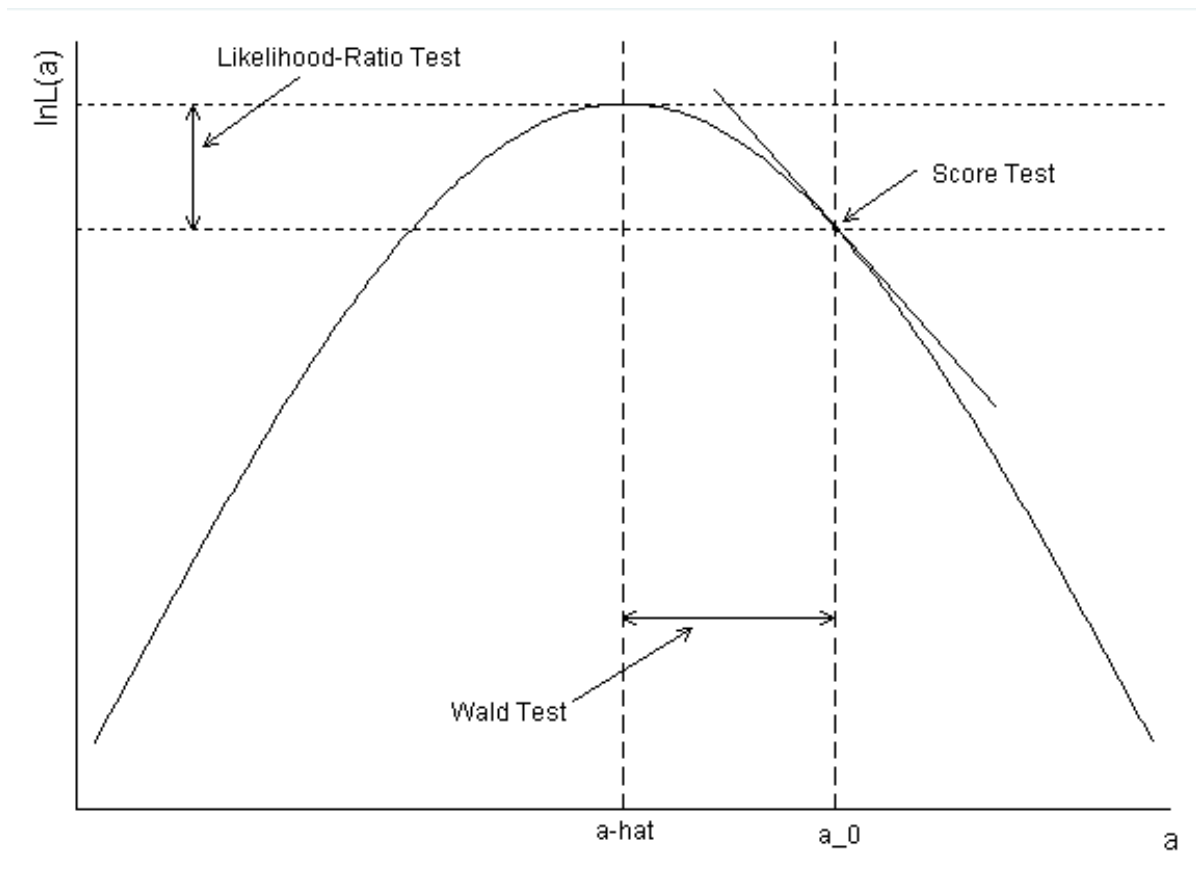
Likelihood Ratio Test

- LR Statistic:

$$\log \left\{ \left[\frac{L(\hat{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}}_H)} \right]^2 \right\} \sim \chi_{q_1}^2$$

- Properties of LRT:
 - requires computation of both full and restricted MLEs
 - of the three tests, the LRT is considered the best
- Often written as $-2 \times \{\ell(\hat{\boldsymbol{\theta}}_H) - \ell(\hat{\boldsymbol{\theta}})\}$

Three tests



www.ats.ucla.edu

MLE Example: Exponential Model

1. Example: The following $n = 10$ failure times are observed and assumed to arise from an Exponential(θ) distribution.

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|----|----|---|---|---|---|----|---|---|----|
| Y_i | 10 | 12 | 8 | 7 | 2 | 4 | 15 | 6 | 5 | 19 |

- Summary statistic: $S \equiv \sum_{i=1}^n Y_i = 88$

Example: (a) Computing MLE

(a) Estimate θ using maximum likelihood.

$$f(Y_i; \theta) = \theta e^{-\theta Y_i}$$

$$L(\theta) = \theta^n \exp \left\{ -\theta \sum_{i=1}^n Y_i \right\}$$

$$\ell(\theta) = n \log \theta - S\theta$$

$$U(\theta) = \frac{n}{\theta} - S$$

$$\hat{\theta} = \frac{n}{S} = \frac{10}{88} = 0.114$$

Example: (b) Asymptotic CI

(b) Derive a 95% confidence interval for θ by referring to the asymptotic properties of MLE's.

- Recall: $U(\theta) = n/\theta - S$

$$\frac{\partial U}{\partial \theta} = \frac{-n}{\theta^2}$$

$$I(\theta) = -E \left[\frac{-n}{\theta^2} \right] = \frac{n}{\theta^2}$$

$$I(\hat{\theta}) = \frac{10}{(0.114)^2} = 769.47$$

$$\widehat{SE}(\hat{\theta}) = (769.47)^{-1/2} = 0.036$$

$$CI(\theta) = 0.114 \pm (1.96)(0.036) = (0.043, 0.185)$$

Example: Hypothesis Testing

- A previous study, conducted under similar conditions but in a different university, estimated $\hat{\theta} = 0.15$.
- (c) Conduct a Wald test of whether or not the results of the current investigation are consistent with those of the previous study.

$$H_0 : \theta = 0.15 \text{ vs. } H_1 : \theta \neq 0.15$$

$$\text{from (b), } \widehat{SE}(\hat{\theta}) = 0.036$$

$$\begin{aligned} X_W^2 &= \left\{ \frac{\hat{\theta} - \theta_H}{\widehat{SE}(\hat{\theta})} \right\}^2 \\ &= \left\{ \frac{0.114 - 0.15}{0.036} \right\}^2 \\ &= 1.00 \\ &< \chi_{0.95}^2 = 3.84 \end{aligned}$$

fail to reject $H_0 : \theta = 0.15$.

Example: (d) Score Test

(d) Test $H_0 : \theta = 0.15$ vs. $H_1 : \theta \neq 0.15$ using the score test.

$$\begin{aligned}U(\theta_H) &= U(0.15) \\&= \frac{10}{0.15} - S = -21.33 \\I(0.15) &= \frac{10}{0.15^2} = 444.44 \\X_S^2 &= (-21.33)(444.44)^{-1}(-21.33) = 1.02 \\&< 3.84\end{aligned}$$

- fail to reject $H_0 : \theta = 0.15$

Example: (e) Likelihood Ratio Test

(e) Test the same hypothesis using the likelihood ratio test.

- computing the maximized and restricted log likelihoods,

$$\begin{aligned}\ell(\hat{\theta}) &= 10 \log(0.114) - (0.114)(88) \\ &= -31.75\end{aligned}$$

$$\begin{aligned}\ell(\theta_H) &= 10 \log(0.15) - (0.15)(88) \\ &= -32.17\end{aligned}$$

$$2\{\ell(\hat{\theta}) - \ell(\theta_H)\} = 2(32.17 - 31.75) = 0.84$$

- fail to reject H_0

Likelihood: Additional Comments

- Exact inference only available for select (and really simple) cases
 - asymptotic results usually employed
 - if applicability of large-sample results is in question (e.g., low n), re-sampling algorithm could be used
 - bootstrap
 - jackknife
- LR, score and Wald tests are asymptotically equivalent and usually yield similar results for even moderate size n

Example: Maximum Likelihood Estimation

A more flexible alternative to the exponential distribution is the *Weibull* model, with density

$$f(y) = (\lambda\gamma)y^{\gamma-1} \exp\{-\lambda y^\gamma\}$$

- We analyze the data set used in the previous example.
- Since it is required that $\lambda > 0$ and $\gamma > 0$, we reparameterize: $\lambda = e^\beta$, $\gamma = e^\alpha$, then set $\boldsymbol{\theta} = (\beta, \alpha)^T$
- We will derive $L_i(\boldsymbol{\theta})$, $\ell_i(\boldsymbol{\theta})$, $U_i(\boldsymbol{\theta})$ and $J_i(\boldsymbol{\theta})$, then program Newton-Raphson using IML, then
 - solve for $\hat{\boldsymbol{\theta}}$
 - estimate $SE(\hat{\theta}_j)$ for $j = 1, 2$

Likelihood and Related Functions: Weibull Case

- Under the reparameterized model,

$$\begin{aligned} f_i &= e^{\beta+\alpha} Y_i^{\exp\{\alpha\}-1} \exp \left\{ -e^{\beta} Y_i^{\exp\{\alpha\}} \right\} \\ &= L_i \end{aligned}$$

- We then obtain

$$\ell_i = \beta + \alpha + (e^{\alpha} - 1) \log Y_i - e^{\beta} Y_i^{\exp\{\alpha\}}$$

- Setting up the score equation,

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta} &= 1 - e^{\beta} Y_i^{\exp\{\alpha\}} \\ \frac{\partial \ell_i}{\partial \alpha} &= 1 + e^{\alpha} \log Y_i - e^{\beta+\alpha} Y_i^{\exp\{\alpha\}} \log Y_i \end{aligned}$$

- Information matrix,

$$\frac{\partial^2 \ell_i}{\partial \beta^2} = -e^\beta Y_i^{\exp\{\alpha\}}$$

$$\begin{aligned} \frac{\partial^2 \ell_i}{\partial \alpha^2} &= e^\alpha \log Y_i \\ &\quad - e^{\beta+\alpha} \log Y_i Y_i^{\exp\{\alpha\}} \{1 + e^\alpha \log Y_i\} \end{aligned}$$

$$\frac{\partial^2 \ell_i}{\partial \alpha \partial \beta} = -Y_i^{\exp\{\alpha\}} e^{\beta+\alpha} \log Y_i$$

- Computing MLE through Newton-Raphson:
Refer to SAS code ...

BIOSTAT 651
Notes #4: Generalized Linear Models

- Topics:
 - Introduction to GLM
 - Exponential families
- Text (Dobson & Barnett, 3rd Ed.): Chapter 3

From Linear Regression to GLM

- Linear regression model:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i$$

$$E[Y_i | \mathbf{x}_i] = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$V(Y_i | \mathbf{x}_i) = \sigma^2$$

$$Y_i \sim \text{Normal}$$

- The *generalization* part of GLM refers to:
 - dropping the Normality requirement
 - relaxing the constant variance assumption
 - allowing for some function of $E[Y_i]$ to be linear in the parameters
- In GLM, the focus is on the *exponential family*
 - members include: Exponential, Poisson, Binomial, Gamma, Normal

Exponential Family

Exponential Family

- If a distribution is an exponential family, then its probability/density function can be written as:

$$f(Y; \theta, \phi) = \exp \left\{ \frac{t(Y)\theta - b(\theta)}{a(\phi)} + c(Y, \phi) \right\}$$

- typically, θ is the parameter of interest
relates to the mean function
 - in contrast, ϕ (dispersion) is treated as a
nuisance parameter related to the variance
- In GLM, we attempt to separate the mean and variance components
- If $t(Y) = Y$, the family is in *canonical form*, in which case θ is referred to as the canonical (*natural*) parameter

Exponential Family (continued)

- Note:
 - for now, we have one θ indexing any Y
 - in the regression setting (later), we replace θ with θ_i

Exponential Family: Binomial

- Suppose $Y \sim \text{Binomial}(n, \pi)$

$$p(Y; \pi) = \binom{n}{Y} \pi^Y (1 - \pi)^{n-Y}$$

$$= \exp \left\{ Y \log \left(\frac{\pi}{1 - \pi} \right) + n \log(1 - \pi) + \log \left(\binom{n}{Y} \right) \right\}$$

- Therefore,

$$t(Y) =$$

$$a(\phi) =$$

$$\theta =$$

$$b(\theta) =$$

$$c(Y, \phi) =$$

Exponential Family: Poisson Case

- Suppose $Y \sim \text{Poisson}(\lambda)$,

$$\begin{aligned} p(Y; \lambda) &= \frac{e^{-\lambda} \lambda^Y}{Y!} \\ &= \exp \{Y \log(\lambda) - \lambda - \log(Y!)\} \end{aligned}$$

- Therefore,

$$t(Y) =$$

$$a(\phi) =$$

$$\theta =$$

$$b(\theta) =$$

$$c(Y, \phi) =$$

Exponential Family: Normal

- $Y \sim \text{Normal}(\mu, \sigma^2)$, with σ^2 known

$$\begin{aligned} f(Y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(Y - \mu)^2}{2\sigma^2} \right\} \\ &= \exp \left\{ -\frac{(Y - \mu)^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) \right\} \\ &= \exp \left\{ \frac{2\mu Y - \mu^2 - Y^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) \right\} \\ &= \exp \left\{ \frac{\mu Y - (1/2)\mu^2}{\sigma^2} - \frac{Y^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}) \right\} \end{aligned}$$

such that

$$t(Y) =$$

$$\theta =$$

$$a(\phi) =$$

$$b(\theta) =$$

Exponential Family: Likelihood

- For a single data point

$$L_i(\theta) \propto f(Y_i; \theta, \phi)$$

$$\ell_i(\theta) = \log L_i(\theta)$$

- Referring to the previous set-up (canonical form),

$$\ell_i(\theta) = \frac{Y_i\theta - b(\theta)}{a(\phi)}$$

taking derivatives w.r.t θ ,

$$U_i(\theta) = \frac{\partial \ell_i}{\partial \theta} = \frac{Y_i - b'(\theta)}{a(\phi)}$$

$$J_i(\theta) = \frac{-\partial^2 \ell_i}{\partial \theta^2} = \frac{b''(\theta)}{a(\phi)}$$

$$I_i(\theta) = E[J_i(\theta)] = \frac{b''(\theta)}{a(\phi)}$$

Exponential Family: Likelihood (continued)

- Properties of the likelihood function:

$$\begin{aligned}E[U_i(\theta)] &= 0 \\V[U_i(\theta)] &= I_i(\theta)\end{aligned}$$

- Combining these results,

$$E[Y_i] \equiv \mu = b'(\theta)$$

and, in addition,

$$\begin{aligned}\frac{b''(\theta)}{a(\phi)} &= \frac{V(Y_i)}{a(\phi)^2} \\V(Y_i) &= b''(\theta)a(\phi)\end{aligned}$$

Mean and Variance Functions

- Note that $E[Y_i]$ depends only on the natural parameter, θ
 - although $V(Y_i)$ is a function of both θ and ϕ
- The variance is often expressed as

$$V(Y_i) = v(\mu)a(\phi)$$

where $v(\mu)$ is written in terms of only μ

- Since we have already derived $V(Y_i) = b''(\theta)a(\phi)$, it follows that

$$v(\mu) = b''(\theta)$$

Exponential Family: Mean and Variance

- e.g., Applying these ideas to the binomial case:

$$b(\theta) = n \log(1 + e^\theta)$$

$$b'(\theta) = n \frac{e^\theta}{(1 + e^\theta)}$$

$$b''(\theta) = n \frac{e^\theta}{(1 + e^\theta)^2}$$

such that

$$E[Y] = b'(\theta) = n \frac{e^\theta}{(1 + e^\theta)}$$

$$V(Y) = b''(\theta)a(\phi) = n \frac{e^\theta}{(1 + e^\theta)^2}$$

Mean and Variance (continued)

- e.g., applying to the Normal case:

$$b(\theta) = \frac{\theta^2}{2}$$

$$b'(\theta) = \theta$$

$$b''(\theta) = 1$$

such that

$$E[Y] = \theta$$

$$V(Y) = \sigma^2$$

General k-Parameter Exponential Family

- Set $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$
- A distribution is a k -parameter exponential family if its probability/density function can be expressed in the following form:

$$f(Y; \boldsymbol{\theta}) = \exp \left\{ \sum_{j=1}^k t_j(Y) \theta_j - b(\boldsymbol{\theta}) + c(Y) \right\}$$

- In this setting, all k parameters are of interest
- e.g., Normal (σ^2 unknown)

Regression Modeling Using GLM

Generalized Linear Models

- Initially developed by Nelder & Wedderburn (1972, *JRSSA*)
 - assume that a known function of $\mu_i = E[Y_i]$ is related linearly to \mathbf{x}_i

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

- $g(\cdot)$ is referred to as the *link* function
- Still assume independence of Y_1, \dots, Y_n
- Linearity assumption now applies to $g(\mu_i)$, which need not equal $E[Y_i]$

Components of the GLM

- In setting up a GLM, the following are specified:

1. Distribution (random component)

- Y_i assumed to follow a (canonical) exponential family:

$$f(Y_i; \theta_i, \phi) = \exp \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

2. Systematic component

- linear predictor: $\eta_i \equiv \mathbf{x}_i^T \boldsymbol{\beta}$

3. Link function

- connects \mathbf{x}_i and μ_i
- $g(\mu_i) = \eta_i$
- required that g be monotone, differentiable function

$$g^{-1}(\eta_i) = \mu_i$$

Link Functions

- Commonly chosen link functions include

$$\text{log} \quad \eta_i = \log(\mu_i)$$

$$\text{logit} \quad \eta_i = \log \left\{ \frac{\mu_i}{1 - \mu_i} \right\}$$

$$\text{probit} \quad \eta_i = \Phi^{-1}(\mu_i)$$

complementary

$$\text{log-log} \quad \eta_i = \log\{-\log(1 - \mu_i)\}$$

where $\Phi(\cdot)$ is the CDF for a $N(0,1)$ variate

Canonical Link

- We observe (Y_i, \mathbf{x}_i) for $i = 1, \dots, n$, where the distribution of $(Y_i | \mathbf{x}_i)$ is assumed to be of the form

$$f(Y_i; \theta_i, \phi) = \exp \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

- Using previously described properties of exponential families:

$$\begin{aligned} E[Y_i] = \mu_i &= b'(\theta_i) \\ V(Y_i) &= b''(\theta_i) a(\phi) = v(\mu_i) a(\phi) \end{aligned}$$

- Link function, $g(\cdot)$, is canonical if $\eta_i = \theta_i$
- Note: the canonical link is usually preferred due to some desirable statistical and computational properties.

Range Restrictions

- In linear regression, $\mu_i \in (-\infty, \infty)$ and $\mathbf{x}_i^T \boldsymbol{\beta} \in (-\infty, \infty)$
 - in fact, $g(\mu_i) = \mu_i$ (identity link) is typically chosen when $Y_i \sim \text{Normal}$
- For links other than the identity, range restrictions should be accommodated
 - e.g., for $Y_i \sim \text{Poisson}$, $\mu_i > 0$
select $\mu_i = e^{\eta_i} > 0$
 - e.g., for $Y_i \sim \text{Bernoulli}$, $\mu_i \in (0, 1)$
select $\mu_i = e^{\eta_i} / \{1 + e^{\eta_i}\} \in (0, 1)$
 - in both cases, canonical link

Deriving Canonical Link

- Examples: deriving the canonical link:
 - e.g., $Y_i \sim \text{Normal}$
 - e.g., $Y_i \sim \text{Bernoulli}$
 - e.g., $Y_i \sim \text{Poisson}$

Choice of Link Function

- It is possible to use links that are not canonical
- e.g., possible that $Y_i \sim \text{Normal}$, but that covariate effects are multiplicative
 - implies $\mu_i = e^{\eta_i}$
- e.g., $Y_i \sim \text{Poisson}$, but with additive covariate effects
 - implies $\mu_i = \eta_i$
 - preferably, $\hat{\mu}_i < 0$ never, or rarely
- Some would argue that the link function should be chosen in accordance with the investigator's objectives

BIOSTAT 651
Notes #5: GLM: Estimation

- Lecture Topics:
 - Parameter estimation
 - Iterative methods
- Text (Dobson & Barnett, 3rd Ed.): Chapter 4

Exponential Family: Recap

- Suppose that Y_i arises from an exponential family with parameters θ_i and ϕ , where ϕ is known

- density:

$$f(Y_i; \theta_i, \phi) = \exp \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

- link function:

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \qquad \eta_i = g(\mu_i)$$

- moments:

$$E[Y_i] \equiv \mu_i = b'(\theta_i)$$

$$V(Y_i) = b''(\theta_i)a(\phi) = \frac{\partial \mu_i}{\partial \theta_i} a(\phi) = v(\mu_i)a(\phi)$$

$$v(\mu_i) \equiv \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

GLM: Canonical Link

- The function $g(\cdot)$ is a *canonical* link if $\theta_i = \eta_i$
- For canonical link,
 - $g(\mu_i) = \theta_i$
 - note: we already showed that $\mu_i = b'(\theta_i)$
 - therefore, $g(\cdot)$ and $b'(\cdot)$ are inverse functions

$$g(b'(x)) = b'(g(x)) = x$$

$$g^{-1}(x) = b'(x)$$

$$b'^{-1}(x) = g(x)$$

where the -1 refers to *inverse* as opposed to reciprocal

- Note that $g(\cdot)$ and $b'(\cdot)$ are one-to-one in the settings of our interest

GLM: Variance Function

- Calculating the variance function:
 - recall that $\mu_i = b'(\theta_i)$
 - $v(\mu_i) = b''(\theta_i)$
- Under the canonical link function:
 $v(\mu_i) = 1/g'(\mu_i)$

$$\begin{aligned} v(\mu_i) &= \frac{\partial \mu_i}{\partial \theta_i} \\ &= \left\{ \frac{\partial \theta_i}{\partial \mu_i} \right\}^{-1} \\ &= \left\{ \frac{\partial \eta_i}{\partial \mu_i} \right\}^{-1} \\ &= \frac{1}{g'(\mu_i)} \end{aligned}$$

Examples: Variance Function

- e.g., $Y_i \sim \text{Normal}$:

$$\begin{aligned}g(\mu_i) &= \mu_i \\g'(\mu_i) &= 1 \\v(\mu_i) &= 1\end{aligned}$$

- e.g., Logistic:

$$\begin{aligned}g(\mu_i) &= \log \left\{ \frac{\mu_i}{1 - \mu_i} \right\} \\g'(\mu_i) &= \frac{1}{\mu_i(1 - \mu_i)} \\v(\mu_i) &= \mu_i(1 - \mu_i)\end{aligned}$$

- e.g., Poisson:

$$\begin{aligned}g(\mu_i) &= \log(\mu_i) \\g'(\mu_i) &= \frac{1}{\mu_i} \\v(\mu_i) &= \mu_i\end{aligned}$$

GLMs with Canonical Link

| Response | Distribution | η_i | $v(\mu_i)$ |
|--------------|--------------|---|--------------------|
| continuous | Normal | μ_i | 1 |
| 0, 1 | Bernoulli | $\log \left\{ \frac{\mu_i}{1 - \mu_i} \right\}$ | $\mu_i(1 - \mu_i)$ |
| 0, 1, 2, ... | Poisson | $\log(\mu_i)$ | μ_i |

Maximum Likelihood: GLM

- Likelihood:

$$L_i = \exp \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} \right\}$$

- Log likelihood:

$$\ell_i = \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)}$$

- Score function:

- with ϕ treated as a *nuisance parameter*, the focus is on $\boldsymbol{\beta}$
- therefore, work with

$$U_i(\boldsymbol{\beta}) = \frac{\partial \ell_i}{\partial \boldsymbol{\beta}}$$

- although we could derive U_i from first principles, it is often easier to employ the *chain rule* ...

Maximum Likelihood: GLM (continued)

- Recall: Chain Rule for differentiation:

$$\frac{\partial}{\partial x} f(g(x)) = \frac{\partial f}{\partial g} \frac{\partial g}{\partial x}$$

- Applied to our setting,

$$U_i(\boldsymbol{\beta}) = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}$$

- Computing each of the partial derivatives,

$$\begin{aligned} \frac{\partial \ell_i}{\partial \theta_i} &= \frac{Y_i - b'(\theta_i)}{a(\phi)} \\ \frac{\partial \theta_i}{\partial \mu_i} &= \frac{1}{b''(\theta_i)} = \frac{1}{v(\mu_i)} \\ \frac{\partial \mu_i}{\partial \eta_i} &= \left\{ \frac{\partial \eta_i}{\partial \mu_i} \right\}^{-1} = \frac{1}{g'(\mu_i)} \\ \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} &= \mathbf{x}_i \end{aligned}$$

Score Function: GLM

- Combining these results,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ \frac{Y_i - \mu_i}{a(\phi)} \right\} \frac{1}{v(\mu_i)g'(\mu_i)} \mathbf{x}_i$$

- A more compact representation,

$$U(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \mathbf{X}^T \mathbf{V}^{-1} \boldsymbol{\Delta}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

where we have

$$\begin{aligned} \mathbf{V} &= \text{diag}\{v(\mu_1), \dots, v(\mu_n)\} \\ \boldsymbol{\Delta} &= \text{diag}\{g'(\mu_1), \dots, g'(\mu_n)\} \end{aligned}$$

- If the canonical link is used, then

$$U(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu})$$

Connection to Moment Estimator

- Therefore, under the canonical link, we could compute $\hat{\beta}$ as the solution to

$$\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0}$$

- Note: $\hat{\beta}$ could also be viewed as a Method-of-Moments (MoM) estimator
 - i.e., equate the sufficient statistic for β , namely $\mathbf{X}^T \mathbf{Y}$, with its mean:

$$\mathbf{X}^T \mathbf{Y} = E[\mathbf{X}^T \mathbf{Y}] = \mathbf{X}^T \boldsymbol{\mu}$$

Score Function: Normal Response

- Note: we've now derived the general form of the score function for any exponential family (with canonical link function)
- Now, suppose $Y_i \sim \text{Normal}$ with constant variance,

$$\begin{aligned}\theta_i &= \eta_i = \mu_i \\ \mu_i &= \mathbf{x}_i^T \boldsymbol{\beta} \\ a(\phi) &= \sigma^2\end{aligned}$$

- Score function could then be written as

$$U(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Score Functions: Canonical Link

- To obtain the score function for other distributions, we just need expressions for μ and $a(\phi)$
- e.g., Logistic regression:

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \log \left\{ \frac{\mu_i}{1 - \mu_i} \right\}$$

$$\mu_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}_i \left(Y_i - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)$$

- e.g., Poisson regression:

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \log(\mu_i)$$

$$\mu_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$$

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}_i \left(Y_i - e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right)$$

Information Matrix: Canonical Link

- We need to compute the information matrix for *inference*, and even for *parameter estimation* itself
- Observed information (canonical link):

$$J(\boldsymbol{\beta}) = \frac{-\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = - \sum_{i=1}^n \frac{\partial U_i}{\partial \boldsymbol{\beta}^T}$$

- Applying the chain rule again,

$$\frac{\partial U_i}{\partial \boldsymbol{\beta}^T} = \frac{\partial U_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T}$$

with each of the partial derivatives given by

$$\begin{aligned} \frac{\partial U_i}{\partial \mu_i} &= -\frac{1}{a(\phi)} \mathbf{x}_i \\ \frac{\partial \mu_i}{\partial \eta_i} &= v(\mu_i) \\ \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T} &= \mathbf{x}_i^T \end{aligned}$$

Information Matrix: Can. Link (continued)

- Combining results on the preceding slide,

$$\begin{aligned} J(\boldsymbol{\beta}) &= \frac{1}{a(\phi)} \sum_{i=1}^n \mathbf{x}_i v(\mu_i) \mathbf{x}_i^T \\ &= \frac{1}{a(\phi)} \mathbf{X}^T \mathbf{V} \mathbf{X}, \end{aligned}$$

where $\mathbf{V} = \text{diag}\{v(\mu_1), \dots, v(\mu_n)\}$

- e.g., $Y_i \sim \text{Bernoulli}$, $v(\mu_i) = \mu_i(1 - \mu_i)$,

$$J(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \mu_i(1 - \mu_i)$$

- e.g., $Y_i \sim \text{Poisson}$, $v(\mu_i) = \mu_i$,

$$J(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \mu_i$$

Information Matrix: Non-Canonical Link (Added)

- Recall: Score function (with non-canonical link)

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ \frac{Y_i - \mu_i}{a(\phi)} \right\} \frac{1}{v(\mu_i)g'(\mu_i)} \mathbf{x}_i$$

- Convenient to use the expected information,

$$\begin{aligned} I_i &= V(U_i) \\ &= E \left[(Y_i - \mu_i)^2 \right] \frac{1}{a(\phi)^2 g'(\mu_i)^2 v(\mu_i)^2} \mathbf{x}_i \mathbf{x}_i^T \\ &= \frac{1}{a(\phi) v(\mu_i) g'(\mu_i)^2} \mathbf{x}_i \mathbf{x}_i^T \end{aligned}$$

Information Matrix: Non-Canonical Link (Added)

- Expected Information

$$I_i = \frac{1}{a(\phi)v(\mu_i)g'(\mu_i)^2} \mathbf{x}_i \mathbf{x}_i^T$$

- We then obtain

$$\begin{aligned} I(\boldsymbol{\beta}) &= \sum_{i=1}^n I_i \\ &= \mathbf{X}^T \{a(\phi) \boldsymbol{\Delta} \mathbf{V} \boldsymbol{\Delta}\}^{-1} \mathbf{X} \end{aligned}$$

where we have

$$\begin{aligned} \mathbf{V} &= \text{diag}\{v(\mu_1), \dots, v(\mu_n)\} \\ \boldsymbol{\Delta} &= \text{diag}\{g'(\mu_1), \dots, g'(\mu_n)\} \end{aligned}$$

Computing MLEs

- Closed-form version of $\hat{\beta}$ generally only for the Normal model with identity link
in all other cases, iterative methods are required ...
- We will now study:
 - Newton-Raphson method
 - Fisher scoring
 - role of WLS

Newton-Raphson: MLE

- Applying Newton-Raphson to solve the score equation,
 - initial value: $\hat{\beta}_0$; often set to $\mathbf{0}$
 - update step:
$$\hat{\beta}_{(j+1)} = \hat{\beta}_j + J^{-1}(\hat{\beta}_j)U(\hat{\beta}_j)$$
 - stopping criterion: $\|\hat{\beta}_{(j+1)} - \hat{\beta}_j\| < \xi$
- Procedure is somewhat sensitive to the choice of starting value

Fisher Scoring

- Same general idea as Newton-Raphson, but replace $J(\boldsymbol{\beta})$ with its expectation, $I(\boldsymbol{\beta})$
 - update step:

$$\hat{\boldsymbol{\beta}}_{(j+1)} = \hat{\boldsymbol{\beta}}_j + I^{-1}(\hat{\boldsymbol{\beta}}_j)U(\hat{\boldsymbol{\beta}}_j)$$

- Lacks optimality properties of N-R method
 - generally takes longer to converge
 - more robust to poor choice of $\hat{\boldsymbol{\beta}}_{(0)}$
- Note: for GLM with canonical link, Newton-Raphson and Fisher Scoring are equivalent

IRWLS in GLM: Canonical Link

- Consider the $(j + 1)$ th Fisher Scoring iterate,

$$\hat{\boldsymbol{\beta}}_{(j+1)} = \hat{\boldsymbol{\beta}}_j + I^{-1}(\hat{\boldsymbol{\beta}}_j)U(\hat{\boldsymbol{\beta}}_j)$$

- Multiply both sides by $I(\cdot)$,

$$I(\hat{\boldsymbol{\beta}}_j)\hat{\boldsymbol{\beta}}_{(j+1)} = I(\hat{\boldsymbol{\beta}}_j)\hat{\boldsymbol{\beta}}_j + U(\hat{\boldsymbol{\beta}}_j)$$

- Written more in terms of the observed data,

$$\mathbf{X}^T \mathbf{V}_j \mathbf{X} \hat{\boldsymbol{\beta}}_{(j+1)} = \mathbf{X}^T \mathbf{V}_j \left\{ \mathbf{X} \hat{\boldsymbol{\beta}}_j + \mathbf{V}_j^{-1} (\mathbf{Y} - \boldsymbol{\mu}_j) \right\}$$

- Setting $\boldsymbol{\eta}_j = \mathbf{X} \boldsymbol{\beta}_j$,

$$\mathbf{X}^T \mathbf{V}_j \mathbf{X} \hat{\boldsymbol{\beta}}_{(j+1)} = \mathbf{X}^T \mathbf{V}_j \left\{ \boldsymbol{\eta}_j + \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\mu}_j) \right\}$$

- Set $\mathbf{Z}_j = \boldsymbol{\eta}_j + \mathbf{V}_j^{-1} (\mathbf{Y} - \boldsymbol{\mu}_j)$, then solving,

$$\hat{\boldsymbol{\beta}}^{(j+1)} = (\mathbf{X}^T \mathbf{V}_j \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}_j \mathbf{Z}_j$$

- amounts to WLS estimator, with covariate \mathbf{X} , weight matrix \mathbf{V}_j and response vector \mathbf{Z}_j

IRWLS

- Algorithm is known as *Iteratively Reweighted Least Squares* (IRWLS)

- need initial estimate: e.g., $\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$
- then, compute $\hat{\mu}_{i,0}$, $V(\hat{\mu}_{i,0})$, $\hat{\eta}_{i,0} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_0$
- update weight matrix, \mathbf{V}_0 , and response, $\mathbf{Z}_0 = \hat{\boldsymbol{\eta}}_0 + \{\mathbf{V}_0\}^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)$
- finally, update parameter estimate:

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}^T \mathbf{V}_0 \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}_0 \mathbf{Z}_0$$

- iterate until convergence obtained

IRWLS in GLM: Non-canonical Link (Added)

- Use the same algorithm with

$$U(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \mathbf{X}^T \mathbf{V}^{-1} \boldsymbol{\Delta}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

$$I(\boldsymbol{\beta}) = \mathbf{X}^T \{a(\phi) \boldsymbol{\Delta} \mathbf{V} \boldsymbol{\Delta}\}^{-1} \mathbf{X}$$

where we have

$$\mathbf{V} = \text{diag}\{v(\mu_1), \dots, v(\mu_n)\}$$

$$\boldsymbol{\Delta} = \text{diag}\{g'(\mu_1), \dots, g'(\mu_n)\}$$

IRWLS: Issues

- Q: Why did we switch from MLE to weighted least squares?

Example: GLM estimation: Seizure data

A clinical trial was conducted in order to evaluate the impact of Progabide on the frequency of epileptic seizures. Patients were randomized to either receive or not receive Progabide. The data set contains information on:

- age at start of study (AGE; measured in years)
- baseline seizure count; defined as the number of seizures in the 8 weeks prior to the study's commencement (BASE)
- treatment indicator (Z; 1=treated, 0=placebo)
- seizure counts in each of 4 two-week periods (Y1, Y2, Y3, Y4)

The investigators define the outcome as total post treatment seizure count: $Y_i \equiv \sum_{j=1}^4 Y_{ij}$.

(a) Fit the following model

$$E[Y_i] = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

using ordinary least squares and save the fitted values, \hat{Y}_i , and studentized residuals, \hat{r}_i .

- Systematic component:

$$\mu_i = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

- Random component:

$$Y_i \sim N(\mu_i, \sigma^2)$$

(b) Interpret β_3 coefficient. Test whether the treatment has an effect on the mean number of seizures.

- β_3 : difference in mean seizure count between Progabide and Placebo treatment groups, adjusted for age and baseline seizure count.

- p-value= 0.5887, highly insignificant.
- (c) Considering the data structure, is the above model appropriate?
- NO. Y is the seizure count. 1) Normality and 2) constant variance assumptions are not satisfied
- (d) Plot the \hat{r}_i against \hat{Y}_i . What evidence does this plot provide with respect to the model assumptions?
- Variance increases as the predicted Y increases. Equal variance assumption is not satisfied.
- (e) Plot a histogram of the residuals, and do a q/q plot. Does the normality assumption appear reasonable?
- QQ plot does not follow the 45 degree line. Normality assumption is not satisfied.
- (f) Are the \hat{Y}_i values all reasonable?

- Since Y_i s are seizure counts, \hat{Y}_i should be > 0 . In this data, 10% of $\hat{Y}_i < 0$.

(g) Fit the transformed model,

$$E[T_i] = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i.$$

where $T_i = \log(Y_i + 0.1)$. What is the sense in shifting Y_i in this case?

- to avoid $\log(0)$.

(h) In transforming the response, what issues are we attempting to address?

- Non-normality and non-constant variance (heteroscedasticity)

(i) Based on log-transformed model, test

$H_0 : \beta_3 = 0$. Compare your result to that from the original model.

- P-value=0.0657. Still insignificant at the level $\alpha = 0.05$.

(j) Assess whether the transformation was successful in remedying the lack of adherence to the model assumptions.

- Residual and QQ plots indicate that the transformation addressed the violation of normality and constant variance assumptions to some extent.

(k) Interpret β_3 based on the transformed model.

- β_3 : Difference in mean log seizure count between Progabide and Placebo treatment groups, adjusted for age and baseline seizure count.
- Limitation: β_3 does not represent log mean seizure count difference.

(ℓ) Is $\exp\{\hat{T}_i\}$ the mean seizure count estimate, i.e. estimate of $E(Y_i)$?

- Since

$$\exp\{\hat{T}_i\} \approx \exp\{E(\log(Y_i))\} \neq E\{\exp(\log(Y_i))\} = E(Y_i)$$

So the answer is NO.

(m) Fit the following generalized linear model,

$$\log\{E[Y_i]\} = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i.$$

using PROC GENMOD. Interpret β coefficients.

- Systematic component:

$$\log(\lambda_i) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

- Random component:

$$Y_i \sim \text{Poisson}(\lambda)$$

- β_3 : Difference in log mean seizure count between Progabide and Placebo treatment groups, adjusted for age and baseline seizure count.

(n) Set up an iteratively re-weighted least squares algorithm to fit the previously specified GLM.

- Set β_0 and ζ

- In the j_{th} iteration, calculate

$$\begin{aligned}\eta_{i,j} &= X_i^T \beta_j; \quad \mu_{i,j} = \exp(\eta_{i,j}); \quad v(\mu_{i,j}) = \mu_{i,j} \\ \eta_j &= (\eta_{1,j}, \dots, \eta_{n,j})'; \quad \mu_j = (\mu_{1,j}, \dots, \mu_{n,j})' \\ V_j &= \text{diag}\{v(\mu_{1,j}), \dots, v(\mu_{n,j})\} \\ Z_j &= \eta_j + V_j^{-1}(Y - \mu_j) \\ \beta_{j+1} &= (X^T V_j X)^{-1} X^T V_j Z_j\end{aligned}$$

- Stop when $\|\beta_{j+1} - \beta_j\| < \zeta$

- (o) Using PROC IML, Fit the GLM using iteratively re-weighted least squares.

See the SAS code.

- (p) Recall that patients were followed for 8 weeks before randomization and 8 weeks after. An alternative to modeling number of seizures is to model whether or not the patient's number of seizures was reduced. This implies the binary response variate $I(Y_i < B_i)$.
- Write down an appropriate GLM for this response.

- Let $Y_i^* = I(Y_i < B_i)$

- Systematic component:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 Z_i$$

- Random component:

$$Y_i^* \sim \text{Bernoulli}(\pi_i)$$

- Set up an IRWLS algorithm.

- Use the same procedure in (n) with

$$\pi_{i,j} = \mu_{i,j} = \frac{\exp(\eta_{i,j})}{1 + \exp(\eta_{i,j})}$$

$$v(\mu_{i,j}) = \pi_{i,j}(1 - \pi_{i,j})$$

- Code the IRWLS scheme using IML.

- See the SAS code.

BIOSTAT 651
Notes #6: GLM: Inference

- Lecture Topics:
 - Nested models
 - Hypothesis testing
- Text (Dobson & Barnett, 3rd Ed.): Chapter 5

GLM: Basic Set-Up

- Canonical exponential family,

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

- Key quantities:

$$\begin{aligned} E[Y_i] &= \mu_i \\ V(Y_i) &= v(\mu_i) a(\phi) \\ g(\mu_i) &= \eta_i \\ \eta_i &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

- Connecting parameters:

$$\begin{aligned} \mu_i &= b'(\theta_i) \\ V(Y_i) &= b''(\theta_i) a(\phi) \\ v(\mu_i) &= b''(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i} \end{aligned}$$

- Under canonical link ($\eta_i = \theta_i$):

$$v(\mu_i) = \frac{1}{g'(\mu_i)}$$

Comparison of Nested Models

- Suppose we wish to compare the fit of two models
 - *Full model:*

$$\eta_i = \mathbf{x}_{i1}^T \boldsymbol{\beta}_1 + \mathbf{x}_{i2}^T \boldsymbol{\beta}_2$$

- *Reduced model:*

$$\eta_i = \mathbf{x}_{i2}^T \boldsymbol{\beta}_2$$

where $\mathbf{x}_i^T = (\mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ are $q \times 1$ vectors, with $q = q_1 + q_2$

- The above-listed models are *nested*
 - reduced model contains proper subset of full model's covariates
- Testing the null hypothesis

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \text{ (vs } H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0})$$

Hypothesis Testing: Example 1

- Example: Suppose that the impact of a treatment (Z_i) on Y_i is of interest, with age (A_i) and bodymass index (B_i) serving as adjustment covariates. The n patients in the study are randomized to receive treatment ($Z_i = 1$), or not ($Z_i = 0$).

- Since Z_i was randomly generated, the investigators may be interested in whether A_i and B_i are actually worth including in the model.

- Full model:

$$\eta_i = \beta_0 + \beta_1 Z_i + \beta_2 A_i + \beta_3 B_i$$

- Reduced model:

$$\eta_i = \beta_0 + \beta_1 Z_i$$

- Implies the hypothesis test,

$$H_0 : \beta_2 = \beta_3 = 0 \text{ versus}$$

$$H_1 : \beta_2 \neq 0 \cup \beta_3 \neq 0$$

Hypothesis Testing: Example 2

- Example: As a continuation of the previously-described study, a follow-up study is carried out in which all patients received treatment, although the dose (D_i) differed substantially across patients. The dose/response relationship is of chief interest and, although parsimony is a goal, the investigators are willing to work with more complicated models.

- Two models that the investigator might compare are

$$\eta_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 D_i^3$$

$$\eta_i = \beta_0 + \beta_1 D_i$$

- Hypothesis test,

$$H_0 : \beta_2 = \beta_3 = 0 \text{ versus}$$

$$H_1 : \beta_2 \neq 0 \cup \beta_3 \neq 0$$

Hypothesis Testing: Nested Models

- Note that the hypothesis testing methods we will study in this lecture related to nested models
 - full model contains each of the reduced model's covariates, plus some additional factors
- e.g., Referring back to the previous examples, the models

$$\eta_i = \beta_0 + \beta_1 D_i + \beta_2 A_i$$

$$\eta_i = \beta_0 + \beta_1 Z_i + \beta_2 A_i + \beta_3 B_i$$

are *not nested*

Hypothesis Testing: General Set-Up

- The model (GLM) is given by,

$$\begin{aligned}\eta_i &= \mathbf{x}_{i1}^T \boldsymbol{\beta}_1 + \mathbf{x}_{i2}^T \boldsymbol{\beta}_2 \\ \boldsymbol{\beta} &= \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \\ \mathbf{x}_i^T &= [\mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T]\end{aligned}$$

where $\boldsymbol{\beta}_j$ is a $q_j \times 1$ vector ($j = 1, 2$), with $q = q_1 + q_2$

- Hypothesis of interest:
 - $H_0 : \boldsymbol{\beta}_1 = \mathbf{d}$ versus $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{d}$
 - often, we will set $\mathbf{d} = \mathbf{0}$
- We use the above general framework for the three tests we will now describe:
 - Wald test
 - Score test
 - Likelihood ratio test

Wald Test

- For the Wald test, fit the full model
i.e., *unrestricted* MLE of $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$
- Reference distribution: under H_0 ,

$$X_W^2 \sim \chi_{q_1}^2$$

- Special case (β_1 : scalar):

$$X_W^2 = \left\{ \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} \right\}^2 \sim \chi_1^2$$

Wald Test (continued)

- More generally, can use the Wald test for any null hypothesis of the form $H_0 : \mathbf{C}\boldsymbol{\beta} - \mathbf{d} = \mathbf{0}$
 - then, test statistic given by:

$$\begin{aligned} X_W^2 &= (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T \left\{ \hat{V}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \right\}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \\ &\sim \chi_r^2 \end{aligned}$$

where C is of rank r .

- The variance estimator of $\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}$ is

$$\begin{aligned} V(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) &= V(\mathbf{C}\hat{\boldsymbol{\beta}}) = \mathbf{C}V(\hat{\boldsymbol{\beta}})\mathbf{C}^T \\ &= \mathbf{C}I(\boldsymbol{\beta})^{-1}\mathbf{C}^T \end{aligned}$$

- Combining the results

$$\begin{aligned} X_W^2 &= (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T \left\{ \mathbf{C}I(\hat{\boldsymbol{\beta}})^{-1}\mathbf{C}^T \right\}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \\ &\sim \chi_r^2 \end{aligned}$$

- Recall the linear regression analog:
 - $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
 - General Linear Hypothesis (GLH) test,

$$\begin{aligned}
 F &= \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T \left\{ \hat{V}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \right\}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})}{r} \\
 &= \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T \left\{ \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T \right\}^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})}{\hat{\sigma}^2 r} \\
 &\sim F_{r, n-q}
 \end{aligned}$$

Score Test

- To carry out the score test, compute the *restricted* MLE
i.e., maximize the likelihood under the constraints imposed by H_0
- The test is then based on the original (*unrestricted*) score function
- Score test statistic for $H_0 : \boldsymbol{\beta}_1 = \mathbf{d}$,

$$X_S^2 = U(\hat{\boldsymbol{\beta}}_H)^T I(\hat{\boldsymbol{\beta}}_H)^{-1} U(\hat{\boldsymbol{\beta}}_H)$$
$$\hat{\boldsymbol{\beta}}_H = \begin{bmatrix} \mathbf{d} \\ \hat{\boldsymbol{\beta}}_{2H} \end{bmatrix}$$

where $\hat{\boldsymbol{\beta}}_{2H}$ is computed under H_0

- Null distribution:

$$X_S^2 \sim \chi_{q_1}^2$$

Likelihood Ratio Test

- Likelihood Ratio Test (LRT),
 - fit the full model, hence maximizing the unrestricted likelihood
 - fit the reduced model, computing the H_0 -restricted MLE
 - the LRT statistic is then given by:

$$\begin{aligned}X_L^2 &= 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_H)\} \\ &\sim \chi_{q_1}^2\end{aligned}$$

Comparison of Tests

- The Wald, Score and Likelihood Ratio tests are *asymptotically* equivalent
 - not (numerically) equivalent
 - need not be equal in finite samples
- The LRT tends to perform better than Wald or Score in smaller samples
- Note: the Wald test is not invariant under reparametrization
 - e.g., test of $H_0 : \beta_1 = 0$ could yield a different result from a test of $H_0 : e^{\beta_1} = 1$
- The Score and LR tests are invariant

Example: GLM, Inference; Cell differentiation

Cell differentiation: Researchers are interested in the effect of two agents of immuno-activating ability that may introduce cell differentiation.

Outcome: cell count

Covariates: TNF (tumor necrosis factor), IFN (interferon)

- (a) Write down a model which would allow the TNF effect to depend on IFN.

- Systematic component:

$$\log(\lambda_i) = \beta_0 + \beta_1 TNF_i + \beta_2 IFN_i + \beta_3 TNF_i \times IFN_i$$

- Random component:

$$Y_i \sim \text{Poisson}(\lambda_i)$$

(b) Estimate β and their (Wald) confidence intervals using IRWLS.

- Estimation of $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ can be carried out using IRWLS (See the SAS code).
- 95% Wald confidence interval for $\beta_k (k = 0, \dots, 3)$ is

$$\hat{\beta}_k \pm 1.96 \widehat{SE}(\hat{\beta}_k).$$

Since $\widehat{Var}(\hat{\beta}) = I(\hat{\beta})^{-1}$, $\widehat{SE}(\hat{\beta}_k)$ is the square root of the k th diagonal element of $I(\hat{\beta})^{-1}$. For the Poisson regression,

$$I(\hat{\beta}) = X'VX,$$

where $V = \text{diag}\{v(\mu_1), \dots, v(\mu_n)\}$ with $v(\mu) = \mu$.

(c) Test for IFN and interactions between TNF and IFN. Specifically, write out the Wald statistic and compute it using IML.

- $H_0: \beta_2 = \beta_3 = 0$ vs $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$

- Contrast matrix

$$C = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Under H_0 , $C\hat{\beta}$ asymptotically follows the multivariate normal distribution with mean zero and variance $CI(\hat{\beta})^{-1}C^T$.

- Wald test statistic is

$$X_w^2 = \{C\hat{\beta}\}^T \{CI(\hat{\beta})^{-1}C^T\}^{-1} \{C\hat{\beta}\},$$

which follows χ_2^2 under the null hypothesis.

- In this data, $X_w^2 = 4.4615902$ and the corresponding p-value=0.107443. We cannot reject the null hypothesis.

(d) This time, carry out the Score test. Write out the formula, then do the computations using IML.

- $H_0: \beta_2 = \beta_3 = 0$ vs $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$
- Let $\hat{\beta}_H = (\hat{\beta}_{0H}, \hat{\beta}_{1H}, 0, 0)$ be the MLE of β under

the null hypothesis. Score test statistic is

$$X_s^2 = U(\hat{\beta}_H)^T I(\hat{\beta}_H)^{-1} U(\hat{\beta}_H),$$

which follows χ_2^2 under the null hypothesis.

- In this data, $X_s^2 = 4.4782374$ and the corresponding p-value=0.1065524. We cannot reject the null hypothesis.
- [Here I used the same notation in the note $(\hat{\beta}_H)$. In the today's lecture I used $\hat{\beta}^0$ instead of $\hat{\beta}_H$. But they are the same.]

(e) Carry out the likelihood ratio test (LRT). Write out the formula, then do the computations using IML.

- $H_0: \beta_2 = \beta_3 = 0$ vs $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$
- Maximum likelihood under the null and alternative:

$$l(\hat{\beta}_H) = \sum_{i=1}^n \{Y_i \log(\hat{\lambda}_{H,i}) - \hat{\lambda}_{H,i}\}$$

$$l(\hat{\beta}) = \sum_{i=1}^n \{Y_i \log(\hat{\lambda}_i) - \hat{\lambda}_i\}$$

where $\hat{\lambda}_{H,i} = \exp(X_i^T \hat{\beta}_H)$ and $\hat{\lambda}_i = \exp(X_i^T \hat{\beta})$.

- Likelihood ratio test statistic is

$$X_L^2 = 2\{l(\hat{\beta}) - l(\hat{\beta}_H)\},$$

which follows χ_2^2 under the null hypothesis.

- In this data, $X_L^2 = 4.3365661$ and the corresponding p-value=0.1143738. We cannot reject the null hypothesis.
- [In the today's lecture I used $\hat{\lambda}^0$ instead of $\hat{\lambda}_H$. But they are the same.]

(g) Carry out the LRT using proc genmod by fitting full and reduced models.

- From the proc genmod, $l(\hat{\beta}_H) = 1208.7574$ and $l(\hat{\beta}) = 1210.9257$. LRT test statistic is

$$X_L^2 = 2 * \{l(\hat{\beta}) - l(\hat{\beta}_H)\} = 4.3366$$

- P-value is 0.1143.

(h) Recompute the Wald and LRT, this time using the Contrast statement.

- See the SAS code.

BIOSTAT 651
Notes: GLM Diagnostics

- Topics:
 - Goodness of fit
 - Residuals
 - Influence Measure

Goodness of Fit: General Considerations

- Measure goodness of fit by how well $\hat{\mu}_i$ replace Y_i
 - \mathbf{Y} : n -dimensional
 - $\hat{\boldsymbol{\beta}}$: q -dimensional
- *Saturated model*: n parameters (one per unique data point)
 - fits data perfectly
 - no data reduction
- *Null model*:
 - $\hat{\mu}_i = \hat{\mu}$ for all i
 - e.g., intercept-only model
 - maximum degree of data summarization
 - fit may be very poor
- The above models are use typically useful only for *judging the fit* of the current model

Deviance: Derivation

- *Deviance*: generalization of the sum of squares of residuals in linear regression.
- Derived by first comparing the likelihoods of the *fitted* and *saturated* models,

$$\left\{ \frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})} \right\}^2$$

where $\tilde{\boldsymbol{\theta}}$ is based on the saturated model

- Then, work on the log scale,

$$2 \times \{\ell(\tilde{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}})\}$$

- Now, consider a single data point:

$$\begin{aligned}\ell(\hat{\theta}_i) &= \frac{Y_i \hat{\theta}_i - b(\hat{\theta}_i)}{a(\phi)} \\ \ell(\tilde{\theta}_i) &= \frac{Y_i \tilde{\theta}_i - b(\tilde{\theta}_i)}{a(\phi)}\end{aligned}$$

Deviance: Derivation (continued)

- Take difference, sum over all subjects, remove scaling:

$$D = 2 \sum_{i=1}^n \left[Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \{b(\tilde{\theta}_i) - b(\hat{\theta}_i)\} \right]$$

which is known as the *Deviance*

- $D^* = D/a(\phi)$ is referred to as the *Scaled Deviance*
 - Note: In the book,
 $\frac{1}{a(\phi)} 2 \sum_{i=1}^n \left[Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \{b(\tilde{\theta}_i) - b(\hat{\theta}_i)\} \right]$ is
the Deviance, and $a(\phi)D$ is the Scaled
Deviance.
- When the model fits well, $D^* \sim \chi_{n-q}^2$
asymptotically.

- Examples:

Normal $D = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2$

$$D^* = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2$$

Poisson $D = D^* = 2 \left[\sum_{i=1}^n Y_i \log \frac{Y_i}{\hat{\mu}_i} - \sum_{i=1}^n (Y_i - \hat{\mu}_i) \right]$

Binomial $D = D^* = 2 \sum_{j=1}^m \left[Y_j \log \left(\frac{Y_j}{\hat{\mu}_j} \right) + (n_j - Y_j) \log \left(\frac{n_j - Y_j}{n_j - \hat{\mu}_j} \right) \right]$

Pearson Chi-Square Statistic

- Another measure of a model's fit, the *Pearson Chi-Square Statistic*,

$$X_P^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\hat{V}(Y_i)}$$

- When the model fits well, $X_P^2 \sim \chi_{n-q}^2$ asymptotically.

Goodness of fit tests

- In principle, both (scaled) Deviance and Pearson statistics asymptotically follows χ^2_{n-q} distribution, so we can test GOF.
- However, it does not always work. Especially in logistic regression.
- There exists several modifications, including a test proposed by Hosmer and Lemeshow. (We will cover later)

Comparing GOF Statistics

- Deviance decreases when covariates are added to a model
 - note: applies to *nested* models
- Pearson X^2 has intuitive appeal
- Can carry out hypothesis testing using Deviance
 - applies to nested models
 - equivalent to Likelihood Ratio Test

Difference in Deviances: LRT

- Scaled deviance

- for a given model, with MLE $\hat{\beta}$,

$$D^* = 2 \times \{\ell(\tilde{\beta}) - \ell(\hat{\beta})\}$$

where $\tilde{\beta}$ corresponds to a *saturated* model

i.e., one parameter for each unique covariate pattern

- If we let D_0^* and D_1^* denote the scaled deviances under H_0 and H_1 , respectively, then the LRT can be computed as

$$X_L^2 = D_0^* - D_1^*$$

Residuals

- Deviance and Pearson X^2 are global measures of goodness-of-fit
 - summary of model's fit
- Also useful to evaluate the model's performance for individual subjects or groups of subjects
- Pearson residuals:

$$\hat{r}_i^P = \frac{Y_i - \hat{\mu}_i}{\hat{V}(Y_i)^{1/2}}$$

- Combining the Pearson residuals $\Rightarrow X_p^2$

$$X_P^2 = \sum_{i=1}^n \{\hat{r}_i^P\}^2$$

- Deviance residuals:

- $D = \sum_{i=1}^n D_i$

$$D_i = 2 \left[Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \{b(\tilde{\theta}_i) - b(\hat{\theta}_i)\} \right]$$

- then, define

$$\hat{r}_i^D = \text{sign}(Y_i - \hat{\mu}_i) \sqrt{|D_i|}$$

i.e, such that $D = \sum_{i=1}^n \{\hat{r}_i^D\}^2$

Examples

- Generate data from the following model

$$\log(\lambda_i) = 1 + 0.5x + 0.5x^2, \quad 2 < x < 3$$

$$Y_i \sim \text{Poisson}(\lambda_i)$$

- Use the following model to fit the data
 - True model

$$\log(\lambda_i) = \beta_0 + \beta_1 x + \beta_2 x^2$$

- Missing x^2 term

$$\log(\lambda_i) = \beta_0 + \beta_1 x$$

- Identity link

$$\lambda_i = \beta_0 + \beta_1 x + \beta_2 x^2$$

True Model

Sunday, Jan 13, 2019

The GENMOD Procedure

| Model Information | |
|--------------------|---------|
| Data Set | WORK.A |
| Distribution | Poisson |
| Link Function | Log |
| Dependent Variable | Y |

| | |
|-----------------------------|------|
| Number of Observations Read | 1000 |
| Number of Observations Used | 1000 |

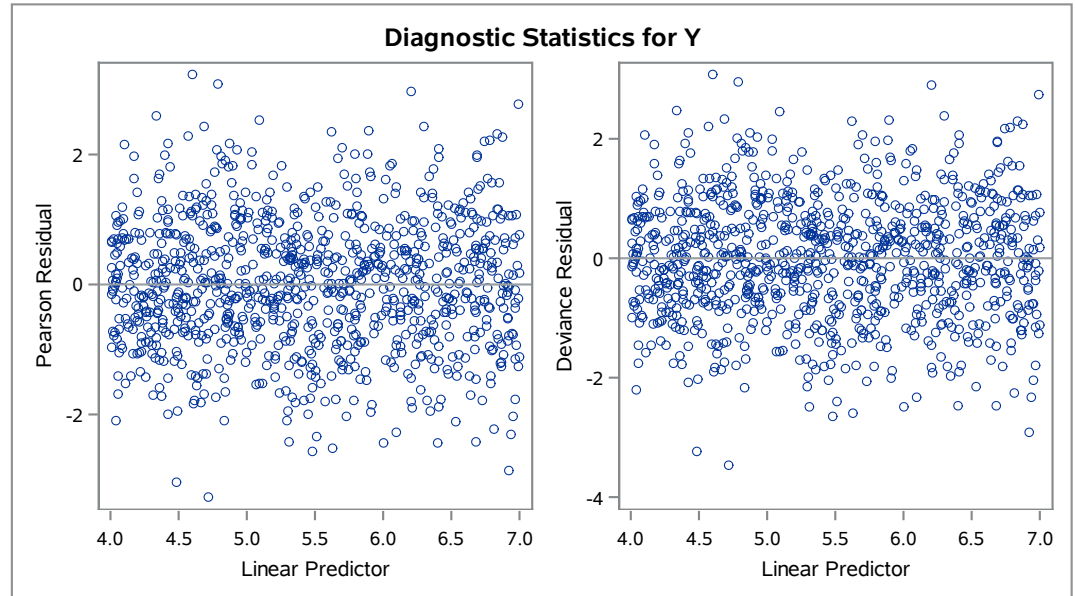
| Criteria For Assessing Goodness Of Fit | | | |
|--|-----|--------------|----------|
| Criterion | DF | Value | Value/DF |
| Deviance | 997 | 974.5321 | 0.9775 |
| Scaled Deviance | 997 | 974.5321 | 0.9775 |
| Pearson Chi-Square | 997 | 972.9738 | 0.9759 |
| Scaled Pearson X2 | 997 | 972.9738 | 0.9759 |
| Log Likelihood | | 1708332.3290 | |
| Full Log Likelihood | | -4127.4402 | |
| AIC (smaller is better) | | 8260.8803 | |
| AICC (smaller is better) | | 8260.9044 | |
| BIC (smaller is better) | | 8275.6036 | |

Algorithm converged.

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | |
|--|----|----------|----------------|----------------------------|--------|-----------------|------------|
| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 0.9810 | 0.1904 | 0.6078 | 1.3543 | 26.54 | <.0001 |
| X | 1 | 0.5049 | 0.1469 | 0.2171 | 0.7928 | 11.82 | 0.0006 |
| X2 | 1 | 0.5005 | 0.0281 | 0.4454 | 0.5556 | 316.95 | <.0001 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

Note: The scale parameter was held fixed.

The GENMOD Procedure



Missing x^2

Sunday, Ja

The GENMOD Procedure

| Model Information | |
|--------------------|---------|
| Data Set | WORK.A |
| Distribution | Poisson |
| Link Function | Log |
| Dependent Variable | Y |

| | |
|-----------------------------|------|
| Number of Observations Read | 1000 |
| Number of Observations Used | 1000 |

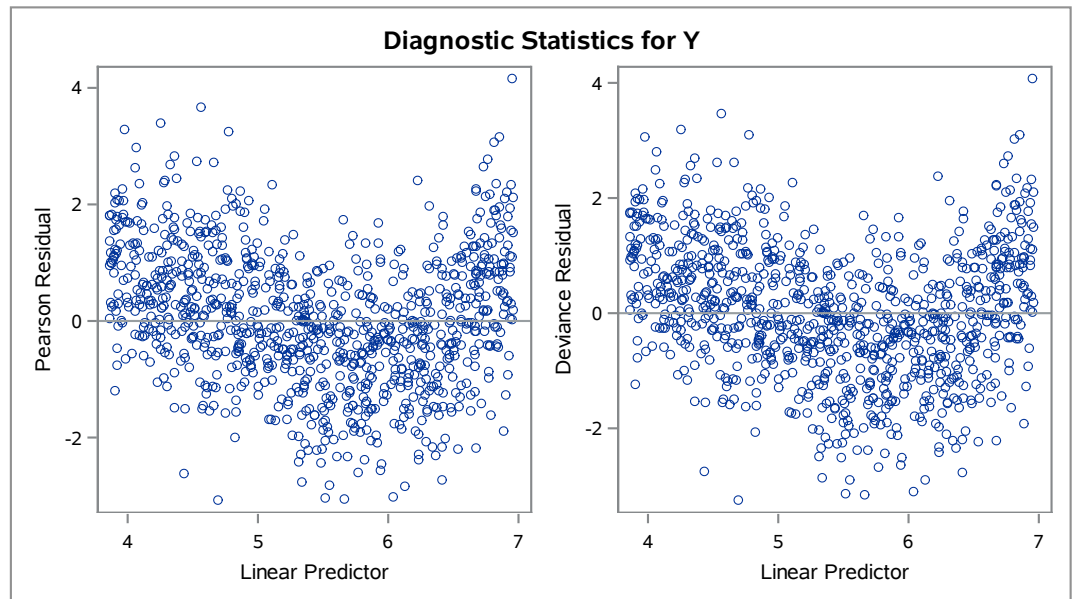
| Criteria For Assessing Goodness Of Fit | | | |
|--|-----|--------------|----------|
| Criterion | DF | Value | Value/DF |
| Deviance | 998 | 1287.4319 | 1.2900 |
| Scaled Deviance | 998 | 1287.4319 | 1.2900 |
| Pearson Chi-Square | 998 | 1299.1010 | 1.3017 |
| Scaled Pearson X2 | 998 | 1299.1010 | 1.3017 |
| Log Likelihood | | 1708175.8791 | |
| Full Log Likelihood | | -4283.8901 | |
| AIC (smaller is better) | | 8571.7802 | |
| AICC (smaller is better) | | 8571.7922 | |
| BIC (smaller is better) | | 8581.5957 | |

Algorithm converged.

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | |
|--|----|----------|----------------|----------------------------|---------|-----------------|------------|
| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | -2.3963 | 0.0205 | -2.4365 | -2.3561 | 13649.5 | <.0001 |
| X | 1 | 3.1187 | 0.0075 | 3.1040 | 3.1333 | 174126 | <.0001 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

Note: The scale parameter was held fixed.

The GENMOD Procedure



Identity link

Sunday, Ja

The GENMOD Procedure

| Model Information | |
|--------------------|----------|
| Data Set | WORK.A |
| Distribution | Poisson |
| Link Function | Identity |
| Dependent Variable | Y |

| | |
|-----------------------------|------|
| Number of Observations Read | 1000 |
| Number of Observations Used | 1000 |

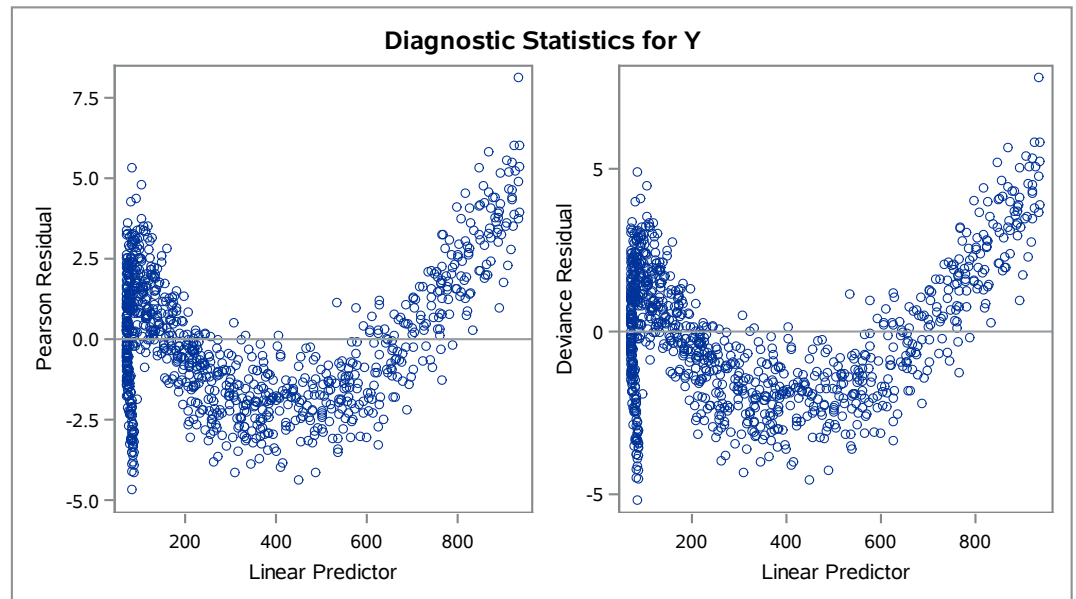
| Criteria For Assessing Goodness Of Fit | | | |
|--|-----|--------------|----------|
| Criterion | DF | Value | Value/DF |
| Deviance | 997 | 4143.7880 | 4.1563 |
| Scaled Deviance | 997 | 4143.7880 | 4.1563 |
| Pearson Chi-Square | 997 | 4142.1266 | 4.1546 |
| Scaled Pearson X2 | 997 | 4142.1266 | 4.1546 |
| Log Likelihood | | 1706747.7011 | |
| Full Log Likelihood | | -5712.0681 | |
| AIC (smaller is better) | | 11430.1363 | |
| AICC (smaller is better) | | 11430.1603 | |
| BIC (smaller is better) | | 11444.8595 | |

Algorithm converged.

| Analysis Of Maximum Likelihood Parameter Estimates | | | | | | | |
|--|----|----------|----------------|----------------------------|----------|-----------------|------------|
| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 5197.430 | 39.1953 | 5120.608 | 5274.251 | 17583.6 | <.0001 |
| X | 1 | -4823.30 | 32.4826 | -4886.96 | -4759.63 | 22049.0 | <.0001 |
| X2 | 1 | 1134.439 | 6.6743 | 1121.357 | 1147.520 | 28890.5 | <.0001 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

Note: The scale parameter was held fixed.

The GENMOD Procedure



Leverage

- In linear regression, the projection matrix (Hat matrix) is

$$H = X(X^T X)^{-1} X^T$$

- h_{ii} , i th diagonal element of H , is called the leverage of the i th observation.
- In GLM, the projection matrix (from IRWLS)

$$H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2}$$

- As the same as the linear regression, h_{ii} is leverage.

- The first order approximation of the variance of raw Pearson residual

$$Var(Y_i - \hat{\mu}_i) \approx (1 - h_{ii})Var(Y_i)$$

- Standardized Pearson residual

$$\hat{r}_i^{PS} = \frac{\hat{r}_i^P}{\sqrt{1 - h_{ii}}}$$

- Similarly, standardized Deviance residual

$$\hat{r}_i^{DS} = \frac{\hat{r}_i^D}{\sqrt{1 - h_{ii}}}$$

Influence measure

- In linear regression, there are a number of diagnostic measures for the influence of one observation based on leave it out, refitting the model, and checking the changes.

- DFBETA

$$DFBETA_i \approx \hat{\beta} - \hat{\beta}_{-i}$$

- Cook's distance

$$\begin{aligned} D_i &= \frac{1}{q\hat{\sigma}^2} (\hat{\beta} - \hat{\beta}_{-i})^T X^T X (\hat{\beta} - \hat{\beta}_{-i}) \\ &= \frac{1}{q} \left(\frac{h_{ii}}{1 - h_{ii}} \right) r_i^2 \end{aligned} \tag{1}$$

- In the linear regression, these statistics can be calculated without refitting the model n times. Explicit shortcut is available based on H .
- In GLM, the exact solution for the explicit shortcut is not available. But the one-step approximation method has been developed to avoid to fitting n times.

- One-step approximation:

- Cook's distance:

$$D_i = \frac{1}{q} \left(\frac{h_{ii}}{1 - h_{ii}} \right) (r_i^{PS})^2$$

- One-step approximation for DFBETA is also available.

Examples

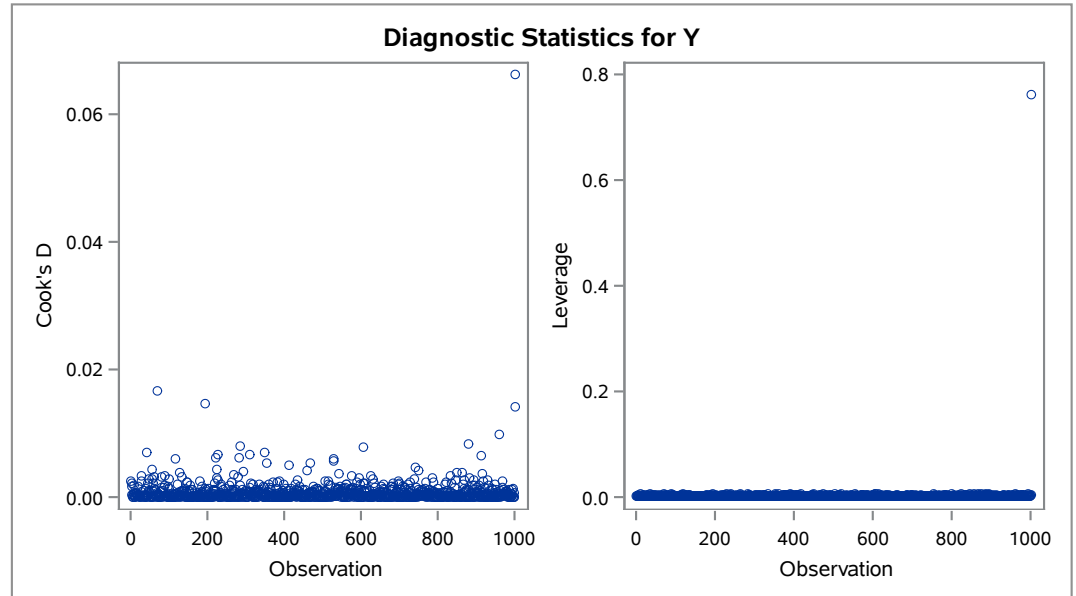
- Previous example:

$$\log(\lambda_i) = 1 + 0.5x + 0.5x^2, \quad 2 < x < 3$$

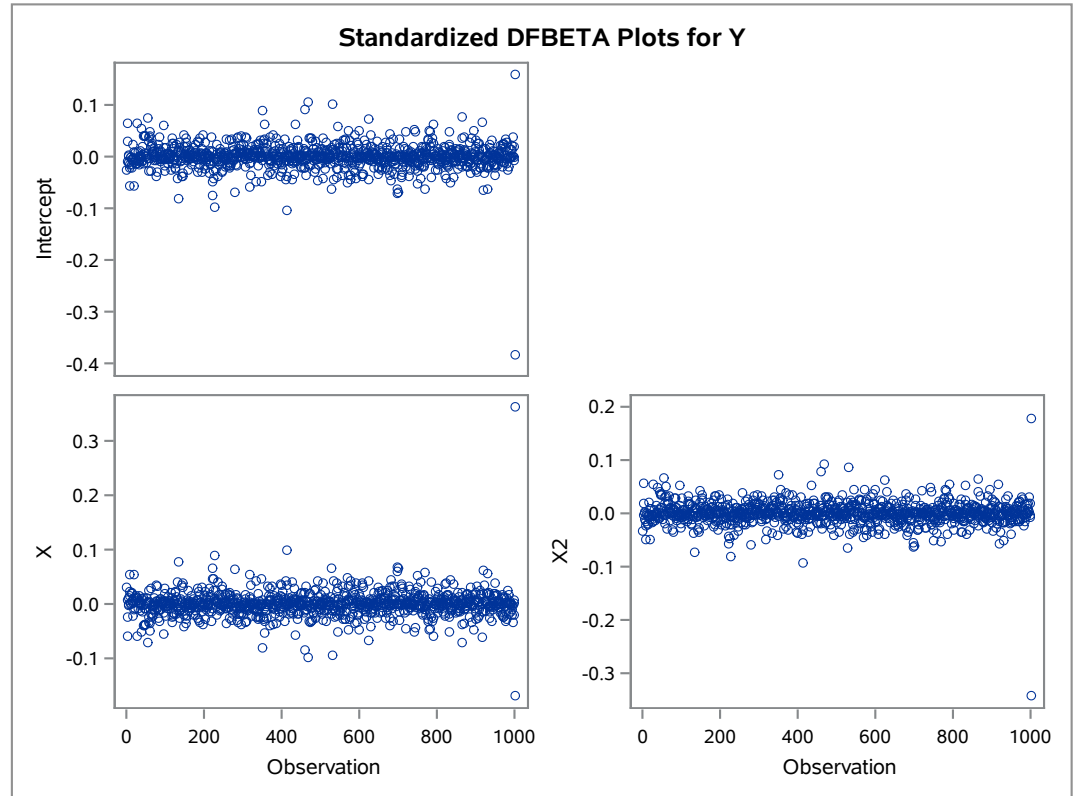
$$Y_i \sim \text{Poisson}(\lambda_i)$$

- Add two outliers (Observation 1001 and 1002)
 - Obs 1001: $X=2$, $Y=0$
 - Obs 1002: $X=3.5$, Y from the true model

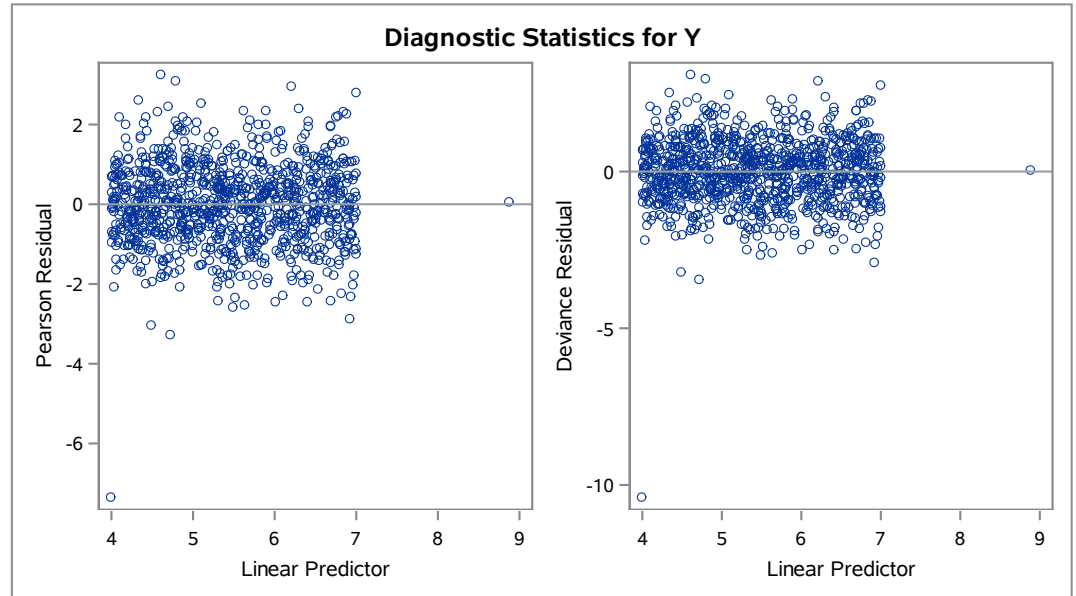
The GENMOD Procedure



The GENMOD Procedure



The GENMOD Procedure



- Obs 1001: $X=2$, $Y=0$
 - Leverage: 0.0037
 - Cook's distance: 0.066
- Obs 1002: $X=3.5$, Y from the true model
 - Leverage: 0.76
 - Cook's distance: 0.014

Multicollinearity

- Explanatory variable (X) are highly correlated with one another.
- Can cause several undesirable consequences.
 - $\hat{\beta}$ will be very unstable.
 - Variances of some $\hat{\beta}$ can be very large.
- Variance inflation factor

$$VIF_j = \frac{1}{1 - R_{(j)}^2}$$

- $R_{(j)}^2$: R^2 obtained from regressing the j th variable against all other variables.
- $VIF = 1$: Not correlated
- $1 < VIF < 5$: moderately correlated
- $VIF > 5$ to 10: highly correlated

- In linear regression, we are concerning about the collinearity in the predictors (X)
- In GLM, we are concerning about the collinearity in the weighted predictor ($V^{1/2}X$)
- SAS proc genmod does not provide VIF, so you have to calculate it using proc reg with the weight statement.

Example: GLM, Inference; Cell differentiation

We keep using the cell differentiation dataset.

Outcome: cell count

Covariates: TNF (tumor necrosis factor), IFN (interferon)

- (a) Consider the following poisson regression model:

$$\log(\lambda_i) = \beta_0 + \beta_1 TNF + \beta_2 IFN + \beta_3 TNF \times IFN$$

Obtain following statistics and save them in a SAS Dataset: (Standardized) Pearson residuals, (Standardized) Deviance residuals, Leverages and Cook's Distances.

- Pearson Residuals:

$$\hat{r}_i^P = \frac{Y_i - \hat{\mu}_i}{\hat{V}(Y_i)^{1/2}} = \frac{Y_i - \hat{\lambda}_i}{\lambda_i^{1/2}}$$

- Deviance Residuals:

$$\begin{aligned} D_i &= 2 \left[Y_i(\tilde{\theta}_i - \hat{\theta}_i) - \{b(\tilde{\theta}_i) - b(\hat{\theta}_i)\} \right] \\ &= 2 \left[Y_i \log \frac{Y_i}{\hat{\lambda}_i} - (Y_i - \hat{\lambda}_i) \right] \end{aligned}$$

$$\hat{r}_i^D = \text{sign}(Y_i - \hat{\lambda}_i) \sqrt{|D_i|}$$

- Leverage :

- Hat matrix: $H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2}$
- Leverage h_{ii} is the i th diagonal element of H

- Standarized Pearson and Deviance residuals:

$$\hat{r}_i^{PS} = \frac{\hat{r}_i^P}{\sqrt{1 - h_{ii}}}; \quad \hat{r}_i^{DS} = \frac{\hat{r}_i^D}{\sqrt{1 - h_{ii}}}$$

- Cook's distance:

$$D_i = \frac{1}{q} \left(\frac{h_{ii}}{1 - h_{ii}} \right) (r_i^{PS})^2$$

(b) In this time, obtain them using proc genmod.

- See the SAS code

(c) Obtain deviance and Pearson's X^2 , and assess the GoF of the model.

- Deviance: 5.09
- Pearson's X^2 : 4.94
- Both Deviance/ DF and Pearson's X^2 / DF are not hugely different from one.
- GOF test
 - H_0 : Model fits the data well
 - Test statistics (use the Pearson's X^2): $D = 4.94$
 - Under the NULL, D follows χ^2_{12}
 - Since $D \leq \chi^2_{12} = 21.026$, we cannot reject H_0 at level 0.05.

(d) Plot the leverages and Cook's distances. Are there any high leverage or high influence observations?

- Observation 16 has $h_{ii} = 0.97 > 2q/n = 0.5$ and Cook's distance = $1.49 > 1$
- High leverage and high influence point.

(e) Obtain VIF for each covariate.

- To get VIF, you first need to obtain the weights (diagonal element of V) using proc genmod (or IML) and use them in proc reg for the weighted least squares. Please see the SAS code.
- VIF
 - β_1 VIF: 1.68
 - β_2 VIF: 2.11
 - β_3 VIF: 2.81

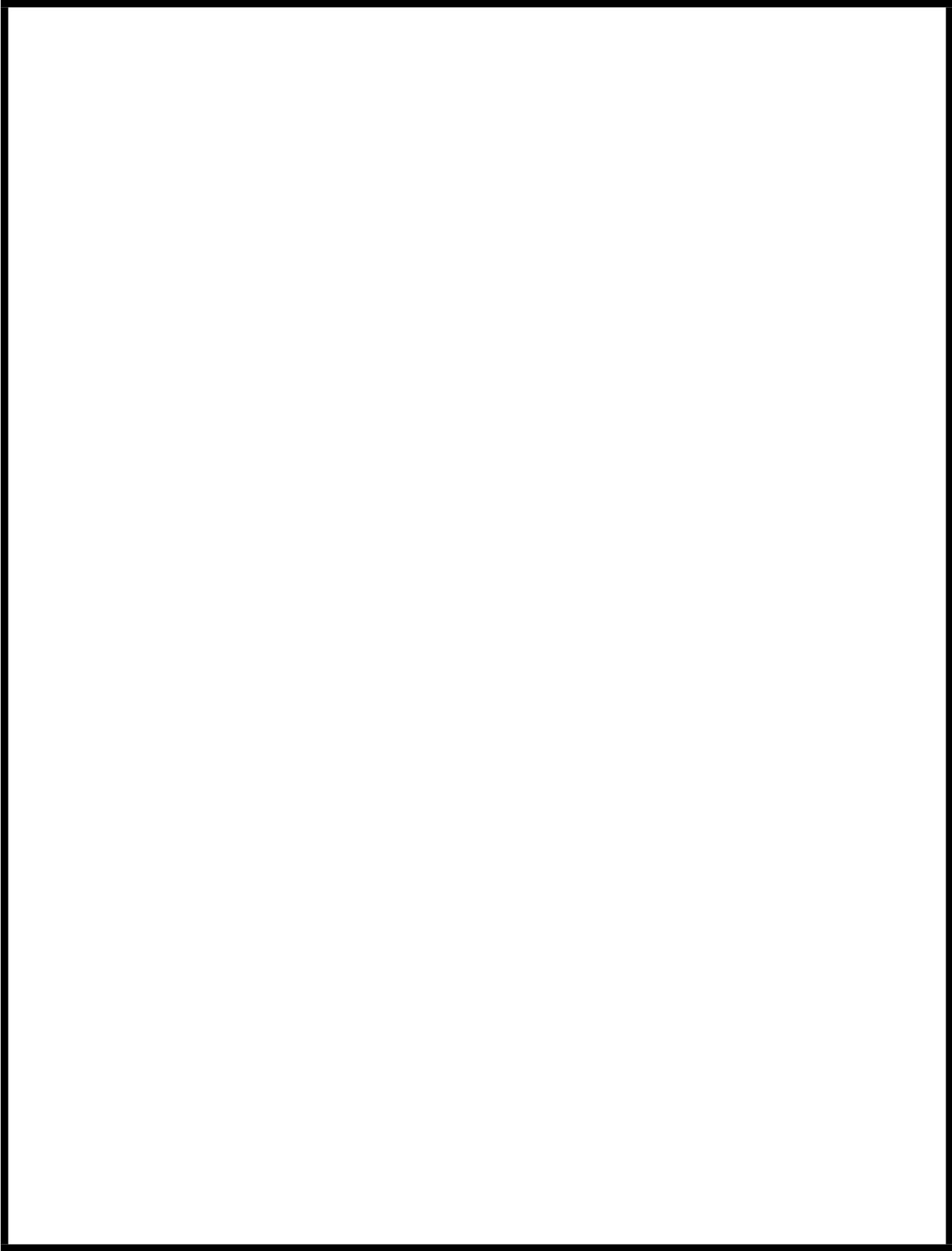
Some covariates are moderately correlated. But no serious multicollinearity problem.

(f) Carry out the LRT test for $H_0 : \beta_2 = \beta_3 = 0$ using Deviance

- $H_0: \beta_2 = \beta_3 = 0$ vs $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$
- From the proc genmod, $D_0^* = 9.4281$ and $D_1^* = 5.0915$. LRT test statistic is

$$X_L^2 = D_0^* - D_1^* = 4.3366$$

- P-value is 0.1143.



BIOSTAT 651

Notes #8: Analysis of Binary Data

- Lecture Topics:
 - Measures of association
 - Sampling mechanisms
 - Potential biases
 - Examples

Data Structure

- Example: Consider a study of liver cancer patients ($n=120$) who have refused conventional therapy. Such patients were randomized to receive either an experimental treatment ($X_i = 1$) or placebo ($X_i = 0$). Patients were then followed for one year, with the response defined as alive ($Y_i = 0$) or dead ($Y_i = 1$). A total of 20 patients refused to be randomized, insisting on receiving the placebo.

The observed data are provided in the following table:

| | $Y=0$ | $Y=1$ | total |
|-------|-------|-------|-------|
| $X=0$ | 27 | 43 | 70 |
| $X=1$ | 10 | 40 | 50 |
| total | 37 | 83 | 120 |

Measures of Frequency

- For now, ignore treatment ...
- *Risk* of death: $P(Y_i = 1)$,

$$\hat{P}(Y_i = 1) = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{83}{120} = 0.692$$

- *Odds* of death,

$$\begin{aligned} \text{Odds}_i &= \frac{P(Y_i = 1)}{P(Y_i = 0)} \\ \widehat{\text{Odds}}_i &= \frac{83/120}{37/120} = 2.24 \end{aligned}$$

Measures of Frequency (continued)

- *Odds* is sometimes used to estimate *risk*

| π | $\pi/(1 - \pi)$ |
|-------|-----------------|
| 0.02 | 0.020 |
| 0.04 | 0.042 |
| 0.06 | 0.064 |
| 0.08 | 0.090 |
| 0.1 | 0.111 |
| 0.2 | 0.250 |
| 0.3 | 0.429 |
| 0.4 | 0.667 |
| 0.5 | 1 |

Measures of Association

- Returning to the liver cancer example, we now focus on comparing the treatment and placebo groups

| X | $\hat{\pi}_j$ |
|-----|---------------|
| 0 | 43/70=0.61 |
| 1 | 40/50=0.80 |

where $\pi_j \equiv P(Y_i = 1|X_i = j)$

- Risk ratio

$$RR = \frac{\pi_1}{\pi_0}$$

$$\widehat{RR} = \frac{\widehat{\pi}_1}{\widehat{\pi}_0} = 1.31$$

- excess relative risk: $(RR - 1) \times 100\%$
in our example: 31%

- Risk difference:

$$RD = \pi_1 - \pi_0$$

$$\widehat{RD} = 0.8 - 0.61 = 0.19$$

Difference versus Ratio

- Risk difference and ratio may yield very different interpretations of exactly the same data set
- e.g., Suppose a flu vaccine is being evaluated, with the risk of the UM650 virus as given in the following table

| X | $\hat{\pi}_j$ |
|-----|---------------|
| 0 | 0.01 |
| 1 | 0.003 |

◦ $\widehat{RR} =$

◦ $\widehat{RD} =$

Difference versus Ratio (continued)

- e.g., Suppose a second flu vaccine is being evaluated, this time with the risk of the UM651 virus given by:

| X | $\hat{\pi}_j$ |
|-----|---------------|
| 0 | 0.9 |
| 1 | 0.7 |

- $\widehat{RR} =$

- $\widehat{RD} =$

Measures of Association (continued)

- Odds ratio:

$$\begin{aligned} OR &= \frac{\text{odds}_1}{\text{odds}_0} \\ &= \frac{P(Y_i = 1|X_i = 1)/P(Y_i = 0|X_i = 1)}{P(Y_i = 1|X_i = 0)/P(Y_i = 0|X_i = 0)} \\ &= \frac{\pi_1/(1 - \pi_1)}{\pi_0/(1 - \pi_0)} \end{aligned}$$

- e.g., in the liver cancer example,

$$\widehat{OR} =$$

compare to relative risk: $\widehat{RR} = 1.31$

- The OR is often used to approximate the RR

OR as an Estimator of RR

- How accurately the OR approximates the RR depends on baseline risk
 - consider the table below, where $RR = 1.5$

| π_0 | OR |
|---------|------|
| 0.02 | 1.51 |
| 0.04 | 1.53 |
| 0.06 | 1.55 |
| 0.08 | 1.57 |
| 0.1 | 1.59 |
| 0.2 | 1.71 |
| 0.3 | 1.91 |
| 0.4 | 2.25 |
| 0.5 | 3.00 |

Odds Ratio: Further Considerations

- In addition to its relationship with the RR, the OR is often viewed as an interesting measure in its own right
 - OR can be estimated consistently for biased samples (ex. case-control design)
 - OR is easily computed using logistic regression
- At this point, it is useful to consider the commonly used study designs ...

Observational Study: Study Designs

- Cohort Study:
 - subjects sampled independently of outcome status followed (prospectively or retrospectively) to ascertain outcome
 - RR and OR are both relevant
- Case Control Study:
 - subjects sampled based on outcome status
 - e.g., select 100 *cases* ($Y_i = 1$) and 300 *controls* ($Y_i = 0$) then, obtain treatment/exposure information
 - often used when studying rare diseases
 - Can't use RR
- Cross-sectional Study:
 - both covariate and outcome status are obtained at the same time point often, a common calendar date
 - RR and OR are both relevant

Study Designs: Cohort Studies

- A cohort study may be either *prospective* or *retrospective*
 - Prospective cohort: response variate has *not* been observed at the start of the study
 - Retrospective cohort: response variate has already been observed by the time the study began
- Prospective designs are considered to be less prone to bias
- Retrospective studies are often more cost- and time-efficient
 - e.g, using large pre-collected databases

Observational Study Designs: Case Control vs Cohort

Exposure

Disease

Wu, S.
Brainfacts



Can't use RR, can only use OR because researcher sets the prevalence within the study. Good for rare diseases. In rare diseases, OR approximates RR. In non-rare diseases, the direction of OR and RR are the same, but the actual number obtained for OR and RR are different. You CANNOT obtain a RR for this. It makes no sense to.



Case-Control

Exposure

Disease



RR and OR are both relevant for this. This is sometimes used to test out a new intervention/treatment.

Prospective Cohort

RR and OR are both relevant for retrospective cohorts.

Exposure

Disease



Investigator/Researcher begins their research. When the researcher enters the scene.

KEY



Present



Absent



What we are seeking; the information we are trying to obtain; what we do not know; our question.

Retrospective Cohort

[wikipedia]

Study Designs: Comparisons

- Simulation: comparison between cohort vs case-control designs
 - Smoking is a risk factor for the colorectal cancer. It can increase the risk twice.
 - Assumptions:
 - * Risk for the cancer among non-smoker: 0.05
 - * Prevalence of smoking: 20%
 - Studies
 - * Cohort study with 10,000 samples
 - * Cohort study with 5,000 non-smoker vs 5,000 smokers
 - * Case-control study with 5,000 cases vs 5,000 controls.

Study Designs: Comparisons

- Settings:
 - $Y=1$ (cancer) vs 0 (no-cancer)
 - $X=1$ (smoker) vs 0 (non-smoker)
 - $RR=2$
 - $P(X = 1) = 0.2$
 - $P(Y = 1|X = 0) = \pi_0 = 0.05$
 - $P(Y = 1|X = 1) = 0.1$

Cohort Study

- Sample 10,000 healthy individual without considering smoking status, and follow them several years.
- The observed data (after several years of follow up)

| | Y=0 | Y=1 | total |
|-------|------|-----|-------|
| X=0 | 7613 | 385 | 7998 |
| X=1 | 1808 | 194 | 2002 |
| total | 9421 | 579 | 10000 |

◦ $\hat{\pi}_0 =$ $\hat{\pi}_1 =$

◦ $\widehat{RR} =$

◦ $\widehat{OR} =$

Cohort Study: Use exposures

- Sample 5000 healthy smokers and 5000 healthy non-smokers.
- The observed data (after several years of follow up)

| | Y=0 | Y=1 | total |
|-------|------|-----|-------|
| X=0 | 4748 | 252 | 5000 |
| X=1 | 4465 | 535 | 5000 |
| total | 9213 | 787 | 10000 |

◦ $\hat{\pi}_0 =$ $\hat{\pi}_1 =$

◦ $\widehat{RR} =$

◦ $\widehat{OR} =$

Case-Control

- Sample 5000 cancer patients and 5000 healthy controls.
- Investigate their smoking history.

| | Y=0 | Y=1 | total |
|-------|------|------|-------|
| X=0 | 4022 | 3358 | 7380 |
| X=1 | 978 | 1642 | 2620 |
| total | 5000 | 5000 | 10000 |

○ $\hat{\pi}_0 =$ $\hat{\pi}_1 =$

○ $\widehat{RR} =$

○ $\widehat{OR} =$

Case-Control

- $P(Y = 1|X)$ cannot be estimated, so RR.
- The OR can be accurately estimated
 - use the *Exposure odds ratio*

$$\begin{aligned} EOR &= \frac{\text{odds}(X = 1|Y = 1)}{\text{odds}(X = 1|Y = 0)} \\ &= \frac{P(X = 1|Y = 1)}{P(X = 0|Y = 1)} \cdot \frac{P(X = 0|Y = 0)}{P(X = 1|Y = 0)} \\ &= \dots \\ &= OR \end{aligned}$$

Misclassification Bias

- Misclassification:
 - e.g., some subjects with $Y = 1$ are mistakenly classified as $Y = 0$
 - if random, OR is generally biased towards the null
 - if non-random, bias can be in either direction
- Examples:
 - recall bias (e.g., case-control study)

Recall bias

- Colorectal cancer example (Case-Control)
- 20 % of previous-smokers without cancer misidentify them as non-smokers.

| | Y=0 | Y=1 | total |
|-------|------|------|-------|
| X=0 | 4231 | 3271 | 7502 |
| X=1 | 769 | 1729 | 2498 |
| total | 5000 | 5000 | 10000 |

◦ $\widehat{OR} =$

Selection Bias

- Selection:
 - Sample obtained is not representative of the population intended to be analyzed
- Key: Does the selected sample accurately represent the target population?
 - if not (resulting from the selection mechanism): *selection bias*

Confounding

- Even in the absence of selection or misclassification, bias can still occur
- e.g., Suppose that there is an *unmeasured* covariate, C
 - *confounding* occurs when:
 - (i) C is associated with X
 - (ii) C is associated with Y (i.e., adjusting for X)
- Confounding can lead to substantial bias

Example: Confounding

- Example: A study was carried out to investigate the association between alcohol consumption (X_i) and lung cancer Y_i . A random sample of $n = 220$ Ann Arbor residents was classified based on whether they drank alcohol ($X_i = 1$) or not ($X_i = 0$). The cohort was then followed for 30 years and classified based on whether they had been diagnosed with lung cancer ($Y_i = 1$) or not ($Y_i = 0$).

Observed data are summarized by the following table:

| | $Y_i=0$ | $Y_i=1$ | total |
|---------|---------|---------|-------|
| $X_i=0$ | 91 | 19 | 110 |
| $X_i=1$ | 19 | 91 | 110 |
| total | 110 | 110 | 220 |

- Odds ratio: $\widehat{OR} =$

Example: Confounding (continued)

- However, *if* information on smoking status S_i *had been recorded*, the following data would have been observed

for non-smokers, $S_i = 0$

| | $Y_i=0$ | $Y_i=1$ | total |
|---------|---------|---------|-------|
| $X_i=0$ | 90 | 9 | 99 |
| $X_i=1$ | 10 | 1 | 11 |
| total | 100 | 10 | 110 |

and for smokers, $S_i = 1$

| | $Y_i=0$ | $Y_i=1$ | total |
|---------|---------|---------|-------|
| $X_i=0$ | 1 | 10 | 11 |
| $X_i=1$ | 9 | 90 | 99 |
| total | 10 | 100 | 110 |

- The apparent association between alcohol consumption and lung cancer was completely due to *confounding* by smoking

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 $i = 1, \dots, n$
 - pairs (\mathbf{x}_i, Y_i) are independent
- GLM
 - Systematic component:

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

- Random component:

$$Y_i \sim \text{Binomial}(n_i, \pi_i)$$

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

$$Y_i = 0, \dots, n_i$$

- $n_i \geq 1$
- 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 |
| \vdots | \vdots | \vdots |

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*,

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

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:

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

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

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

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.

$$\hat{\beta}_0 =$$

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

$$\hat{\beta}_1 =$$

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:
 $\exp\{\hat{\beta}_1\} = \exp\{-0.5056\} = 0.603$
 - 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:

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

◦ $\hat{\beta}$:

$$\hat{\beta}_M =$$

$$\hat{\beta}_F =$$

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,

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

- 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 $Y_i \sim \text{Binomial}(n_i, \pi_i)$ with

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

- 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:

$$L_i(\boldsymbol{\beta}) = \pi_i^{Y_i} (1 - \pi_i)^{n_i - Y_i}$$

$$\ell_i(\boldsymbol{\beta}) = Y_i \log \pi_i + (n_i - Y_i) \log(1 - \pi_i)$$

- 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:

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

- 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 $j = 0, \dots, q$

Likelihood Ratio Test

- Likelihood ratio test:

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

- 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.

$$D = 2 \sum_{j=1}^n \left[Y_i \log \left(\frac{Y_i}{n_i \hat{\pi}_i} \right) + (n_i - Y_i) \log \left(\frac{n_i - Y_i}{n_i - n_i \hat{\pi}_i} \right) \right]$$

$$X_p^2 = \sum_{i=1}^n \frac{(Y_i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}$$

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

Goodness of Fit

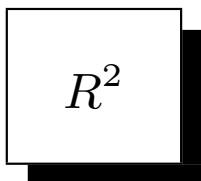
- Deviance and Pearson χ^2 work well when the expected number of events (and non-events) > 5
- When n_i is small, they don't work well.
- SAS does not provide Deviance (and Pearson χ^2) when $n_i = 1$

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 g th 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.



$$R^2$$

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

$$\text{logit}(\pi_i) = \beta_0$$

- Pseudo R^2 (Cox & Snell)

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

- 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 .

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

- 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

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

- Deviance residuals

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

Example: Logistic Regression (Pneumonia Data)

A study carried out on several cohorts of coal miners sought to determine the relationship between duration of coal dust exposure (measured in years) and incidence of severe pneumonia. We use logistic regression to analyze these data.

- (a) What is the overall probability of severe pneumonia?

n_i : number of miners

y_i : number of Pneumonia incidence

x_i : duration (year)

$$\text{overall probability: } \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n n_i} = 0.121$$

- (b) Write down a plausible logistic regression model based on the logit link.

Logistic regression model:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

$$y_i \sim \text{Binomial}(n_i, \pi_i)$$

- (c) Fit the model listed in (b) using PROC LOGISTIC.

$$\hat{\beta}_0 = -4.5383, \quad \exp\{\hat{\beta}_0\} = 0.0107$$

$$\hat{\beta}_1 = 0.0869, \quad \exp\{\hat{\beta}_1\} = 1.091$$

See the SAS code

- (d) Interpret $\hat{\beta}_1$ and $\exp\{\hat{\beta}_1\}$

$\hat{\beta}_1$: expected increase in logit risk (log odds) of Pneumonia for a one-unit increase in exposure.

$\exp\{\hat{\beta}_1\}$: expected odds ratio of Pneumonia for a one-unit increase in exposure.

- (e) Should $\exp\{\beta_1\}$ be an accurate approximation to the relative risk?

$$RR = \frac{\pi_1}{\pi_0} \text{ and } OR = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

It can be shown that

$$OR = RR \frac{1 - \pi_0}{1 - RR\pi_0}. \quad (1)$$

From (1)

$$RR = \frac{OR}{1 - \pi_0 + OR\pi_0} \quad (2)$$

The above equation shows that OR is close to RR when π_0 is small or OR is close to 1. Since the OR estimate ($\exp\{\hat{\beta}_1\}$) is 1.09 and the overall risk is small, we can conclude that the OR estimate is an accurate approximation to RR.

- (f) Interpret $\hat{\beta}_0$ and $\exp\{\hat{\beta}_0\}$

$\hat{\beta}_0$: expected logit risk (log odds) of Pneumonia for a subject with zero exposure.

$\exp\{\hat{\beta}_0\}$: expected odds of Pneumonia for a subject with zero exposure.

(g) Test $H_0 : \beta_1 = 0$ using the Wald statistic.

$$X_w = \frac{\hat{\beta}_1^2}{\widehat{SE}(\hat{\beta}_1)^2} = 34.9410 \gg 3.84 = \chi_{1,0.95}^2$$

Reject the null hypothesis.

(h) Repeat the hypothesis test, but this time using the LRT.

Full model: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$

Reduced model: $\text{logit}(\pi_i) = \beta_0$

$$X_L = -2\{l(\hat{\beta}^0) - l(\hat{\beta})\} = 46.554 \gg 3.84 = \chi_{1,0.95}^2$$

Reject the null hypothesis.

(i) Again, using the score test.

$$X_S = U(\hat{\beta}^0)' I^{-1}(\hat{\beta}^0) U(\hat{\beta}^0) = 44.13 \gg 3.84 = \chi_{1,0.95}^2$$

(j) For the score test just computed, what does $\hat{\beta}^0$ equal?

$$\hat{\beta}^0 = (\hat{\beta}_0^0, 0)'$$

$$\hat{\beta}_0^0 = \text{logit}(\hat{\pi}_0) = \text{logit}(0.121) = -1.983$$

(k) Verify your calculation in (j) using PROC LOGISTIC.

See the SAS code

(l) Calculate pseudo R^2 and generalized R^2 .

Minimal model: $\text{logit}(\pi_i) = \beta_0 \Rightarrow \hat{\pi}^{intercept}$

Fitted model: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i \Rightarrow \hat{\pi}$

- Pseudo R^2 (Cox & Snell)

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

(N is the number of subjects. So in our data, $N = 371$)

- From -2 Log L in SAS output

$$-2l_0 = 274.165; -2l_1 = 227.611$$

$$L_0 = \exp(l_0) = 2.92e-60; L_1 = \exp(l_1) = 3.76e-50$$

$$R^2 = 1 - \left\{ \frac{2.92e-60}{3.76e-50} \right\}^{2/371} = 0.1179$$

You can obtain generalized (Cox & Snell) and max-adjusted (Nagelkerke) R^2 in SAS by using RSQ option in the model statement. See the SAS code.

- Maximum of the Cox & Snell R^2 can be smaller than 1. So Nagelkerke proposed a max adjusted Cox & Snell R^2 . From the SAS output (with RSQ option):

$$\text{max-adjusted } R^2 = 0.6478$$

- (m) Write out the model equation for a saturated logistic model; include an intercept.

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 I(x_i = 15.0) + \cdots + \beta_7 I(x_i = 51.5)$$

- (n) Interpret the parameters from the saturated model.

β_0 : log odds when $x_i = 5.8$

β_1 : log odds ratio between $x_i = 15.0$ vs $x_i = 5.8$

...

- (o) Write out another saturated model, this time *not including* an intercept.

$$\text{logit}(\pi_i) = \beta_1 I(x_i = 5.8) + \cdots + \beta_8 I(x_i = 51.5)$$

- (p) Interpret the parameters from this version of the saturated model.

β_1 : log odds when $x_i = 5.8$

β_2 : log odds when $x_i = 15.0$

...

- (q) Fit the no-intercept saturated model without using PROC LOGISTIC or GENMOD.

$$\hat{\beta}_i = \text{logit}(\hat{\pi}_i), \text{ where } \hat{\pi}_i = y_i/n_i$$

See the SAS code

- (r) Fit the no-intercept saturated model using PROC LOGISTIC. Compare the parameter estimates with your data-step calculations.

See the SAS code

- (s) Fit the saturated model with an intercept using PROC LOGISTIC.

- See the SAS code

- (t) Plot the fitted logits from the saturated and linear logistic model. Does the linear model appear to fit the data based on this plot?

See the SAS code

- To compare these two models, LRT can be used.
- H_0 : linear logistic model fits data well

- Full model ($q = 8$):

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 I(x_i = 15.0) + \cdots + \beta_7 I(x_i = 51.5)$$

- Reduced model ($q = 2$):

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

- LRT test statistic:

$$X_L = 2\{-112.22 + 113.81\} = 3.18 < 12.59 = \chi^2_{6,0.95}$$

- At $\alpha = 0.05$, we cannot reject H_0 . There is not enough evidence that the saturated model fits better than the linear logistic model.

(u) This time, plot the fitted probabilities.

See the SAS code

(v) Change the group level data format to the individual level data format and fit the linear logistic model. Are there any difference?

- Parameter estimates and log likelihood are the same. Full log likelihood are different.

- In the individual level format, deviance and Pearson χ^2 GOF statistics are not presented.

Example: Logistic Regression (Low Birth Weight)

A group of doctors at Baystate Medical Center (in Springfield, MA) sought to determine the factors associated with infant low birth weight (defined as birth weight <2.5 kg). The response variate is a 0/1 indicator for low birth weight (LOW), with the covariates given by the following characteristics of the newborn's mother: age in years (AGE); weight in pounds at last menstrual period (WT); race ("White", "Black", "Other"); smoking status during pregnancy (SMOKE); history of hypertension (HYP); presence of uterine irritability (UI); number of physician visits during the first trimester (FTV); history of premature labor (PTL).

- (a) Compute descriptive statistics on all variables.

See the SAS code.

- (b) Fit a main effects model based on all covariates and using PROC logistic.

$$\begin{aligned} \text{logit}(\pi) &= \beta_0 + \beta_1 AGE + \beta_2 WT + \beta_3 I(\text{white}) \\ &+ \beta_4 I(\text{black}) + \cdots + \beta_9 PTL \end{aligned}$$

- (c) Carry out the Hosmer-Lemeshow goodness of fit test.

- Use lackfit option
- H_0 : the model fits the data well
- HL test statistics: 4.0126, DF= 8
- HL p-value 0.856

- Fail to reject H_0 . This implies that the logistic regression model fits the data well.

(d) What would be the impact on $\hat{\beta}$ of reversing the response variable?

- Model:

$$\text{logit}(\pi_i) = X_i' \beta, \text{ where } \pi_i = \text{Pr}(y_i = 1 | X_i)$$

- New Model:

$$\text{logit}(\pi_i^*) = X_i' \beta^*, \text{ where } \pi_i^* = \text{Pr}(y_i = 0 | X_i)$$

$$\text{Since } \text{logit}(\pi_i^*) = \text{logit}(-\pi_i) = -\text{logit}(\pi_i),$$

$$\beta^* = -\beta$$

(e) Re-fit the main effects model, with $[LOW = 1]$ as the event. Which covariates are predictive of low birth weight?

- Based on Wald test p-values, WT, HYP and PTL are significant.

- You can also use stepwise selection (See the SAS code).
- (f) Re-fit the model, with the AGE, RACE and FTV deleted. Interpret the parameter estimate for SMOKE and WT.
- $\hat{\beta}_{smoke} = 0.5035$; $\hat{\beta}_{WT} = -0.0154$
 - $\hat{\beta}_{smoke}$: estimated log odds ratio of LBW between smoking status, adjusting for the other covariates.
 - $\hat{\beta}_{WT}$: estimated log odds ratio of LBW per pound increase in weight, adjusting for the other covariates.
- (g) Interpret the intercept, $\hat{\beta}_0$.
- $\hat{\beta}_0 = 0.4723$
 - $\hat{\beta}_0$: estimated log odds of LBW when all the covariates = zero.

(h) How would you restructure the design matrix such that the intercept has a more appealing interpretation?

- WT cannot be zero, so use $WT - \overline{WT}$, instead of WT
- \overline{WT} = average value of WT = 130
- Replace WT to $WT - 130$.

(i) Re-fit the model, carrying out your suggestion for improving the interpretation of the intercept. Compare $\hat{\beta}_0$ based on the previous and current models, and reconcile any difference.

- Model (with WT)

$$\text{logit}(\pi) = \beta_0 + \beta_1 WT + \beta_2 \text{Smoke} + \beta_3 \text{Hyp} + \beta_4 \text{Ui} + \beta_5 \text{PTL}$$

- Model (with WT -130)

$$\begin{aligned} \text{logit}(\pi) &= \beta_0^* + \beta_1^* (WT - 130) + \beta_2^* \text{Smoke} + \beta_3^* \text{Hyp} \\ &\quad + \beta_4^* \text{Ui} + \beta_5^* \text{PTL} \\ &= (\beta_0 + \beta_1 130) + \beta_1 (WT - 130) + \beta_2 \text{Smoke} \end{aligned}$$

$$+\beta_3Hyp + \beta_4Ui + \beta_5PTL$$

- $\hat{\beta}_0^* = -1.5239$, which is $\hat{\beta}_0 + 130\hat{\beta}_{WT}$ of the previous model.

- (j) Compare $\hat{\beta}_{WT}$ based on the previous and current models. Comment.

$\hat{\beta}_{WTS}$ are the same.

- (k) Carry out a test of whether the effect of SMOKE depends on either UI or PTL.

- Model: (S: smoke)

$$\begin{aligned} \text{logit}(\pi) &= \beta_0 + \beta_1WT + \cdots + \beta_5PTL \\ &+ \beta_6S \times UI + \beta_7S \times PTL \end{aligned}$$

- H0: $\beta_6 = \beta_7 = 0$

- Wald test statistic: $1.76 < 5.99 = \chi^2_{2,0.95}$

Fail to reject H_0

(l) Based on the interaction model, interpret the parameter estimate for SMOKE.

- Log odds among smokers when UI = PTL=0

$$\text{logit}(\pi_{s=1}) = \beta_0 + \beta_1 WT + \beta_2 S + \beta_3 Hyp$$

- Log odds among non-smokers when UI = PTL=0

$$\text{logit}(\pi_{s=0}) = \beta_0 + \beta_1 WT + \beta_3 Hyp$$

- Now

$$\beta_{smoke} = \beta_2 = \text{logit}(\pi_{s=1}) - \text{logit}(\pi_{s=0})$$

- $\hat{\beta}_{smoke} = 0.6094$

- $\hat{\beta}_{smoke}$: estimated log odds ratio of LBW between smoking status when PTL=UI=0, adjusting for the other covariates.

(m) Based on the interaction model, interpret the

parameter estimate for SMOKE×PTL parameter.

- Log odds ratio between smoker vs non-smoker when PTL=0 (S=Smoke)

$$\begin{aligned} & \text{logit}(\pi_{S=1}) - \text{logit}(\pi_{S=0}) \\ &= \beta_S + \beta_{S \times UI} UI \end{aligned} \quad (1)$$

- When PTL=1

$$\begin{aligned} & \text{logit}(\pi_{S=1}) - \text{logit}(\pi_{S=0}) \\ &= \beta_S + \beta_{S \times UI} UI + \beta_{S \times PTL} \end{aligned} \quad (2)$$

- Now

$$\beta_{S \times PTL} = (2) - (1)$$

or

$$\exp(\beta_{S \times PTL}) = \exp((2)) / \exp((1))$$

- $\hat{\beta}_{S \times PTL} = 0.5245$
- $\hat{\beta}_{S \times PTL}$: log odds ratio of LBW between smoking status is increased by 0.524 for women with history of PTL.

Use OR

- $\exp(\hat{\beta}_{S \times PTL}) = 1.690$
- $\exp(\hat{\beta}_{S \times PTL})$: odds ratio of LBW between smoking status is increased by 69% for women with history of PTL.

(n) Test whether the impact of SMOKE on low birth weight is affected by the weight of the mother.

- Model:

$$\begin{aligned} \text{logit}(\pi) &= \beta_0 + \beta_1 WT + \cdots + \beta_5 PTL \\ &+ \beta_6 S \times WT \end{aligned}$$

- Wald test

$$X_w = \frac{0.0174^2}{0.0134^2} = 1.7 < 3.84 = \chi_{1,0.95}^2$$

Fail to reject H_0

(o) Based on this latest interaction model, interpret the SMOKE \times WT parameter estimate.

- $\hat{\beta}_{S \times WT} = 0.0174$
- $\hat{\beta}_{S \times WT}$: log odds ratio between smoking status is

increased by 0.0174 per pound increase in weight, adjusting for the other covariates.

- $\exp(\hat{\beta}_{S \times WT}) = 1.018$
- $\exp(\hat{\beta}_{S \times WT})$: odds ratio between smoking status is increased by 1.8% per pound increase in weight, adjusting for the other covariates.

(p) Based on this latest interaction model, interpret the SMOKE parameter estimate.

- $\hat{\beta}_S = -1.674$
- $\hat{\beta}_S$: estimated log odds ratio between smoking status when WT=0, adjusting for the other covariates.

(q) Re-fit the model, using $(WT - 130)$ in place of WT. Compare results with the preceding interaction model and comment on any differences.

- New model

$$\begin{aligned} \text{logit}(\pi) &= \beta_0^* + \beta_1^*(WT - 130) + \beta_2 S + \cdots \\ &+ \beta_6^* S \times (WT - 130) \end{aligned}$$

- From the original model

$$\begin{aligned} \text{logit}(\pi) &= \beta_0 + \beta_1 WT + \beta_2 S + \cdots + \beta_6 S \times WT \\ &= (\beta_0 + 130\beta_1) + \beta_1(WT - 130) \\ &+ (\beta_2 + 130\beta_6)S + \cdots + \beta_6 S \times (WT - 130) \end{aligned}$$

$\hat{\beta}_{smoke}^*$ is changed to 0.5934.

$\hat{\beta}_{smoke \times WT}$ and $\hat{\beta}_{smoke \times WT}^*$ are the same.

- (r) Re-fit the most recent model, this time using PROC GENMOD.

See the SAS code.

BIOSTAT 651

Notes #10: Case-Control & Link functions

- Lecture Topics:
 - Case-control sampling
 - Link functions

Case-Control sampling

Study Designs

- Study designs:
 - randomized clinical trial
 - observational study
- Randomized trial: gold standard
 - often infeasible (logistics, ethics)
- Observational study: often easier and more cost efficient to study associations
 - cohort
 - case-control

Case-Control Sampling

- Case-Control study:
 - select n_1 diseased subjects
 - select n_0 non-diseased subjects
 - analysis: contrast \mathbf{x}_i between *cases* ($Y_i = 1$) and *controls* ($Y_i = 0$)
- Motivation: study of *rare* diseases
 - more generally, *cost-* and/or *time-efficient* study of disease (more later...)

Analysis of Case-Control Data

- Recall: the exposure odds ratio (EOR) estimated through case-control sampling is equal to the OR of interest
 - i.e., $Y_i = 0, 1$ and $X_i = 0, 1$

$$\begin{aligned} EOR &\equiv \frac{\text{odds}(X_i = 1|Y_i = 1)}{\text{odds}(X_i = 1|Y_i = 0)} \\ &= \vdots \\ &= \vdots \\ &= \frac{\text{odds}(Y_i = 1|X_i = 1)}{\text{odds}(Y_i = 1|X_i = 0)} \equiv OR \end{aligned}$$

- Result extends to the regression setting...

Case-Control Study: General Set-Up

- Consider the following setting:
 - Y_i : binary
 - $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iq})^T$ (assume discrete)
 - population: N subjects; study: n subjects
sample n_1 cases, and n_0 controls
 $S_i =$ sampling indicator
- data:
 - observed (population):
 - assigned (by investigators):
 - observed data (sample):

- Model:

- sampling fractions:

$$\tau_0 \equiv P(S_i = 1|Y_i = 0)$$

$$\tau_1 \equiv P(S_i = 1|Y_i = 1)$$

- model:

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

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

Case-Control Data: MLE

- Estimation proceeds via maximum likelihood
 - Since \mathbf{x}_i is a random variable, the retrospective likelihood is written as

$$L_i(\boldsymbol{\beta}) \propto P(\mathbf{x}_i | Y_i = 1, S_i = 1)^{Y_i} \\ P(\mathbf{x}_i | Y_i = 0, S_i = 1)^{1-Y_i}$$

- Prentice and Pyke (1979) showed that MLE of β from $P(Y_i = 1 | \mathbf{x}_i, S_i = 1)$ is the same as MLE of β from the prospective logistic regression model with retrospective sampling.

Case-Control Data: MLE

- $P(Y_i = 1|\mathbf{x}_i, S_i = 1)$:

$$\begin{aligned} & P(Y_i = 1|\mathbf{x}_i, S_i = 1) \\ = & \frac{P(Y_i = 1, \mathbf{x}_i, S_i = 1)}{P(\mathbf{x}_i, S_i = 1)} \\ = & \frac{P(\mathbf{x}_i|Y_i = 1, S_i = 1)P(Y_i, S_i = 1)}{P(\mathbf{x}_i, S_i = 1)} \quad (1) \end{aligned}$$

- Since

$$P(S_i = 1|\mathbf{x}_i, Y_i = 1) = P(S_i = 1|Y_i = 1),$$

we have

$$\begin{aligned} & P(\mathbf{x}_i|Y_i = 1, S_i = 1) \\ = & \frac{P(S_i = 1|\mathbf{x}_i, Y_i = 1)P(\mathbf{x}_i|Y_i = 1)}{P(S_i = 1|Y_i = 1)} \\ = & P(\mathbf{x}_i|Y_i = 1) \\ = & \frac{P(Y_i = 1|\mathbf{x}_i)P(\mathbf{x}_i)}{P(Y_i = 1)} \quad (2) \end{aligned}$$

- Combine (1) and (2)

$$\begin{aligned}
 P(Y_i = 1 | \mathbf{x}_i, S_i = 1) &= P(Y_i = 1 | \mathbf{x}_i) \frac{P(S_i = 1 | Y_i = 1)}{P(S_i = 1 | \mathbf{x}_i)} \\
 &= \pi(\mathbf{x}_i) \frac{P(S_i = 1 | Y_i = 1)}{P(S_i = 1 | \mathbf{x}_i)}
 \end{aligned}$$

- Odds

$$\begin{aligned}
 \frac{P(Y_i = 1 | \mathbf{x}_i, S_i = 1)}{P(Y_i = 0 | \mathbf{x}_i, S_i = 1)} &= \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \frac{P(S_i = 1 | Y_i = 1)}{P(S_i = 1 | Y_i = 0)} \\
 &= \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \frac{\tau_1}{\tau_0}
 \end{aligned}$$

- Odds ratio between \mathbf{x}_a vs \mathbf{x}_b

$$OR = \frac{\pi(\mathbf{x}_a) / \{1 - \pi(\mathbf{x}_a)\}}{\pi(\mathbf{x}_b) / \{1 - \pi(\mathbf{x}_b)\}}$$

OR is the same as the prospective design OR.

- Logistic regression model with retrospective sampling

$$\begin{aligned}\log\{odds(\mathbf{x}_i)\} &= \mathbf{x}_i' \beta^* \\ &= \mathbf{x}_i' \beta + \log\left(\frac{\tau_1}{\tau_0}\right)\end{aligned}$$

- Intercept:

$$\beta_0^* = \beta_0 + \log\left(\frac{\tau_1}{\tau_0}\right)$$

- Only the intercept is changed.
- If the sampling fractions are known (τ_0 and τ_1), we can estimate true β_0 .

Case-Control: Example

- We return to the lung cancer example in Lecture Note #8 ...
 - Smoking is a risk factor for the colorectal cancer. It can increase the risk twice.
 - Assumptions:
 - * Risk for the cancer among non-smoker: 0.05
 - * Prevalence of smoking: 20%
 - Studies
 - * Case-control study with 5,000 cases vs 5,000 controls.

- Consider a saturated model based on the logit link,

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

- True values

$$\beta_0 = \log \left\{ \frac{0.05}{0.95} \right\} = -2.944$$

$$\beta_0 + \beta_1 = \log \left\{ \frac{0.1}{0.9} \right\} = -2.197$$

$$\beta_1 = 0.747; \quad \exp(\beta_1) = 2.11$$

Example : Case-Control Sampling (logit)

| | Y=0 | Y=1 | total |
|-------|------|------|-------|
| X=0 | 4022 | 3358 | 7380 |
| X=1 | 978 | 1642 | 2620 |
| total | 5000 | 5000 | 10000 |

- we compute the parameter estimates as:

$$\hat{\beta}_0 = \log \left\{ \frac{3358}{4022} \right\} = -0.1804$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log \left\{ \frac{1642}{978} \right\} = 0.518$$

$$\hat{\beta}_1 = 0.698; \quad \exp(\hat{\beta}_1) = 2.01$$

- Calculate sampling fraction:

- * Prevalence:

$$P(Y = 1) = 0.06$$

- * Ratio of the sampling fractions:

$$\frac{\tau_1}{\tau_0} = 0.94/0.06 = 15.6667$$

- Adjust β_0 using the sampling fractions:

$$\hat{\beta}_0 - \log\left(\frac{\tau_1}{\tau_0}\right) = -0.1804 - 2.7515 = -2.9319$$

Outcome-Dependent Sampling

- Case-control study is a special case of what has come to be called *Outcome-Dependent Sampling* (ODS)
 - creative ways to sample cases and controls
 - outcome can be binary, or more complicated structure
 - extension to survival times, clustered data, etc

Link functions

- Dose-response modeling
- Link functions
- Interpretation of parameters
- Issues in case-control sampling

Dose-Response Models: Introduction

- General set-up:
 - evaluate effect of dose on the probability of a specific event
 - covariate: D_i
 - dependent variable: $Y_i = 0, 1$
 - set $\pi_i = \pi(D_i) = P(Y_i = 1|D_i)$
- Based on our study to date, we use the logit link
 - we now explore alternative link functions

Link Functions: Binary Response

- Link functions used for binary data:

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

$$\Phi^{-1}(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$\log \{ -\log(1 - \pi_i) \} = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$-\log \{ -\log(\pi_i) \} = \mathbf{x}_i^T \boldsymbol{\beta}$$

- all are continuous and increasing on $(0, 1)$
- only first two are symmetric

Link Functions: Background

- Often model π_i using a cumulative distribution function

$$\pi(t) = \int_{-\infty}^t f(s)ds$$

where $f(s)$ is a *tolerance* distribution

- characteristics of valid $f(s)$:
 - (i) $f(s) \geq 0$
 - (ii) $\int_{-\infty}^{\infty} f(s)ds = 1$

Connection to Bioassays

- Historically, binomial regression models were motivated by *bioassay* studies
 - response: proportion of events
e.g., percent dead
 - exposure: dose level
e.g., treatment, toxin, contaminant, etc
- Models of the form $g(\pi_i) = \beta_0 + \beta_1 D_i$ were often considered

Example: Uniform Tolerance

- Example: Suppose that the tolerance distribution is $\text{Uniform}(a, b)$:

$$f(s) =$$

$$\pi(t) =$$

- Graphs of $f(s)$ and $\pi(t)$:

Uniform Tolerance

- Connecting dose and tolerance:

$$\pi(x) = \beta_0 + \beta_1 x$$

$$\beta_0 = \frac{-a}{b-a}$$

$$\beta_1 = \frac{1}{b-a}$$

- Need to impose constraints on x , β_0 and β_1
 - standard GLM methods would not apply

Probit Model

- Suppose that a $\text{Normal}(\mu, \sigma^2)$ distribution is used for the tolerance

- $f(s) =$

- $P(T \leq t) =$

- Choosing the probit function as the link:

$$\Phi^{-1}(\pi(x)) = \beta_0 + \beta_1 x$$

$$\beta_0 = -\frac{\mu}{\sigma}$$

$$\beta_1 = \frac{1}{\sigma}$$

Probit Model (continued)

- Probit link is used frequently
 - e.g., $Y_i = I_i(\text{dead})$
 μ referred to as the median lethal dose,
LD(50): dose required to kill 50% of the
members of a tested population in a specific
time.

Logit Link

- Consider the logistic tolerance distribution:

$$f(s) = \frac{\beta_1 \exp\{\beta_0 + \beta_1 s\}}{(1 + \exp\{\beta_0 + \beta_1 s\})^2}$$

which implies that

$$\pi(t) =$$

which implies the *logit* link function

Complementary Log-Log Link

- If tolerance follows the *extreme value* distribution:

$$f(s) = \beta_1 \exp\{(\beta_0 + \beta_1 s) - e^{\beta_0 + \beta_1 s}\}$$

then we obtain

$$\pi(t) = 1 - \exp\{-e^{\beta_0 + \beta_1 t}\}$$

which implies the *complementary log-log* link:

$$\log\{-\log(1 - \pi(x))\} = \beta_0 + \beta_1 x$$

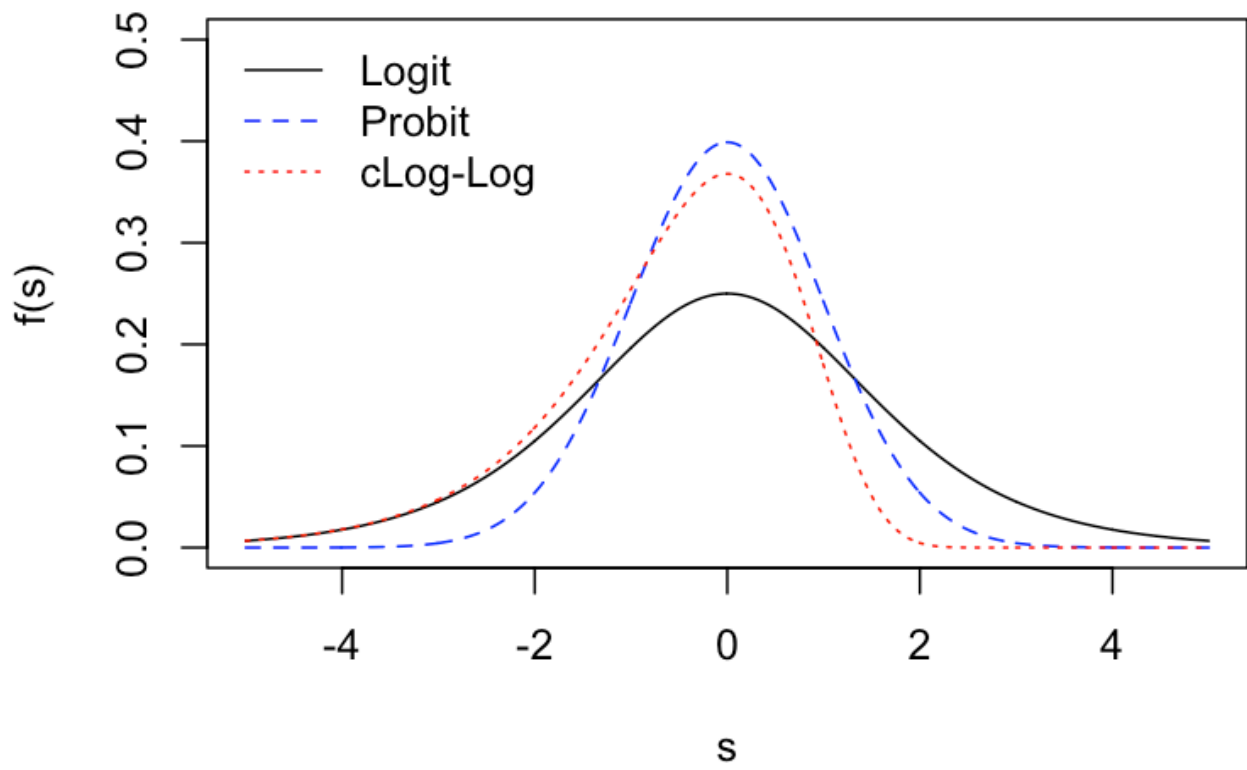
- Related to the hazard ratio
 - Hazard function

$$h(t) = P(T = t | T \geq t) = \frac{f(t)}{1 - \pi(t)}$$

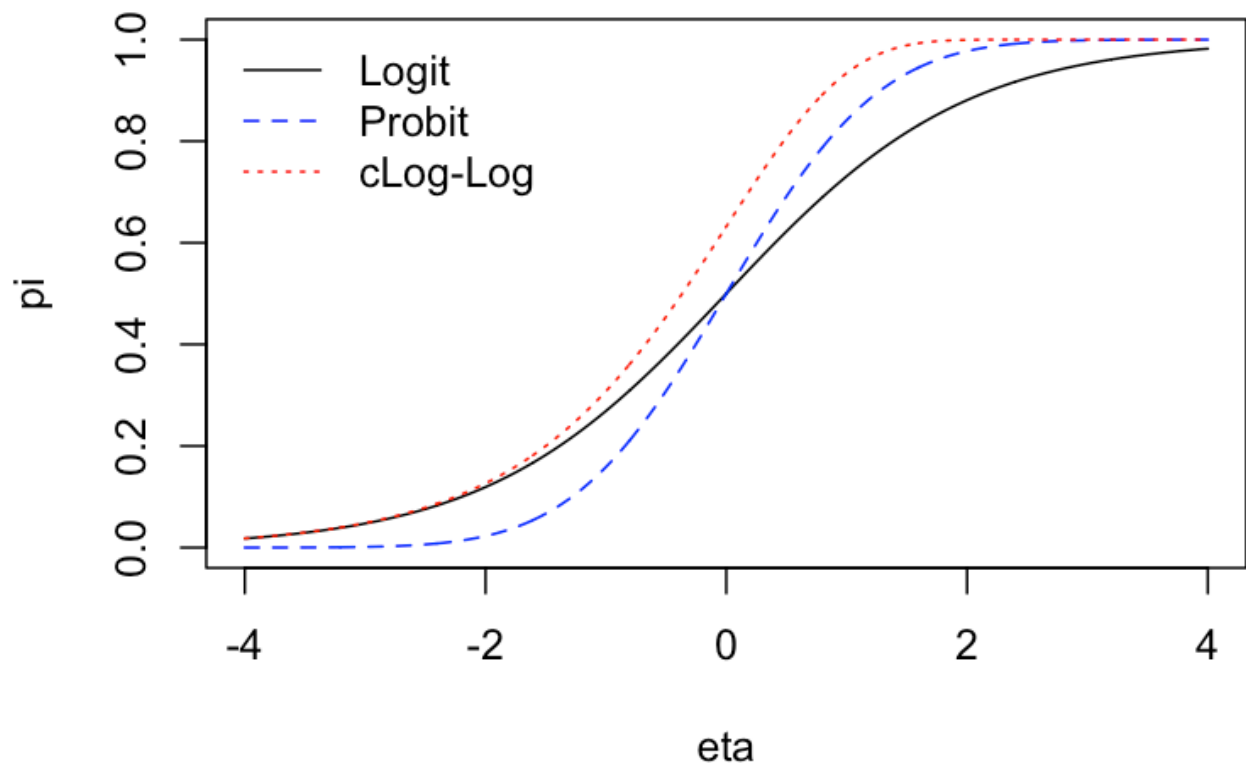
- Hazard ratio:

$$\frac{h(t+1)}{h(t)} = \exp(\beta_1)$$

$f(s)$ functions with $\beta_0 = 0$ and $\beta_1 = 1$



$\pi(x)$ functions



Case-Control Sampling

- Recall that logistic regression provides a consistent OR estimator for case-control sampling
 - e.g, if the model is given by

$$\log \left\{ \frac{\pi_i}{1 - \pi_i} \right\} = \beta_0 + \mathbf{x}_{i1}^T \boldsymbol{\beta}_1$$

then a case-control study will consistently estimate:

- This is a property of the logit link
 - need not hold for alternative link functions

Case-Control Sampling: Example

- Lung cancer example
 - Cohort study with 5,000 non-smoker vs 5,000 smokers
 - Case-control study with 5,000 cases vs 5,000 controls.

Example: Complementary log-log link

- Example: Assume that the true model follows the complementary log-log link,

$$\log \{-\log(1 - \pi_i)\} = \beta_0 + \beta_1 X_i$$

- True values

$$\beta_0 = \log \{-\log(1 - 0.05)\} = -2.97$$

$$\beta_0 + \beta_1 = \log \{-\log(1 - 0.1)\} = -2.25$$

$$\beta_1 = 0.720$$

Example: Cohort Sampling (CLL)

| | Y=0 | Y=1 | total |
|-------|------|-----|-------|
| X=0 | 4748 | 252 | 5000 |
| X=1 | 4465 | 535 | 5000 |
| total | 9213 | 787 | 10000 |

- Parameter estimates:

$$\hat{\beta}_0 = \log \{-\log(4748/5000)\} = -2.962$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log \{-\log(4465/5000)\} = -2.178$$

$$\hat{\beta}_1 = 0.783$$

Example : Case-Control Sampling (CLL)

| | Y=0 | Y=1 | total |
|-------|------|------|-------|
| X=0 | 4022 | 3358 | 7380 |
| X=1 | 978 | 1642 | 2620 |
| total | 5000 | 5000 | 10000 |

- Parameter estimates:

$$\hat{\beta}_0 = \log \{-\log(4022/7380)\} = -0.499$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log \{-\log(978/2620)\} = -0.0147$$

$$\hat{\beta}_1 = 0.484$$

Example: Dose-Response Modeling

In the data set of interest, beetles were exposed to various levels of gaseous carbon disulphide (recorded on the \log_{10} scale) for 5 hours. The observed data are given by:

| j | X_j | n_j | Y_j |
|-----|--------|-------|-------|
| 1 | 1.6907 | 59 | 6 |
| 2 | 1.7242 | 60 | 13 |
| 3 | 1.7552 | 62 | 18 |
| 4 | 1.7842 | 56 | 28 |
| 5 | 1.8113 | 63 | 52 |
| 6 | 1.8369 | 59 | 53 |
| 7 | 1.8610 | 62 | 61 |
| 8 | 1.8839 | 60 | 60 |

We model the dose response relationship between mortality and CS_2 exposure.

(a) Calculate the empirical death probability for each exposure level. Also, calculate the empirical logit, probit and complementary log-log.

- Death Prob.: $\hat{\pi}_j = Y_j/n_j$
- Logit: $\text{logit}_j = \log\{\hat{\pi}_j/(1 - \hat{\pi}_j)\}$
- Probit: $\text{Probit}_j = \Phi^{-1}(\hat{\pi}_j)$
- CLL: $\text{CLL}_j = \log\{-\log(1 - \hat{\pi}_j)\}$

(b) Fit linear regression models between each empirical link and dose. Which link function appears to be most appropriate?

- See SAS code
- CLL link function has the largest R^2 value, so based on R^2 we can select CLL. However other link functions also have very high R^2 values (> 0.95), so you can select link functions based on

the context of study and the interpretation of the results.

(c) Fit a model assuming a probit link.

- Model: $\Phi^{-1}(\pi_j) = \beta_0 + \beta_1 X_j$

(d) Re-fit the model with the probit link, but this time define the end-point as $I(Y_{ij} = 0)$. What do you notice about $\hat{\beta}_1$?

- Model:

$$\Phi^{-1}(\pi_j) = \beta_0 + \beta_1 X_j, \text{ where } \pi_j = Pr(Y_{ij} = 1|X_j)$$

- New Model:

$$\Phi^{-1}(\pi_j^*) = \beta_0^* + \beta_1^* X_j, \text{ where}$$

$$\pi_j^* = Pr(Y_{ij} = 0|X_j)$$

$$\text{Since } \Phi^{-1}(\pi_i^*) = -\Phi^{-1}(\pi_i),$$

$$\beta_1^* = -\beta_1$$

(e) Estimate the LD50 and compare the estimate to a model-free estimator. Compute a 95% confidence interval using IML.

- LD50: x value satisfying $\pi(x) = 0.5$
- Probit model: $\Phi^{-1}\{\pi(x)\} = \beta_0 + \beta_1 x$, where $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$.

$$\hat{\mu} = -\hat{\beta}_0/\hat{\beta}_1 = 1.77 \text{ and } \hat{\sigma} = 1/19.72 = 0.0507.$$

Therefore,

$$\widehat{LD50} = \hat{\mu} = 1.77$$

Model free estimator: 1.78

- By delta method

$$Var\{g(\hat{\beta})\} \approx \frac{\partial g'}{\partial \beta} Var(\hat{\beta}) \frac{\partial g}{\partial \beta},$$

where $g(\beta) = -\beta_0/\beta_1 = LD50$, and

$$\frac{\partial g}{\partial \beta} = (-1/\beta_1, \beta_0/\beta_1^2)'$$

- Using these equations, we can show (use IML)

$$\widehat{Var}\{\widehat{LD50}\} = 0.0000145$$

- 95% CI

$$\widehat{LD50} \pm 1.96\sqrt{\widehat{Var}\{\widehat{LD50}\}} =$$

$$(1.763, 1.778)$$

(f) Fit a model assuming a complementary log-log link.

- Model: $\log\{-\log(1 - \pi_j)\} = \beta_0 + \beta_1 X_j$

(g) Estimate the LD50.

- $\log\{-\log(0.5)\} = \beta_0 + \beta_1 LD50.$

$$\widehat{LD50} = \frac{\log\{-\log(0.5)\} - \widehat{\beta}_0}{\widehat{\beta}_1}$$

$$= \frac{-0.37 + 39.57}{22.04} = 1.78$$

(h) Re-fit the model with the complementary log-log link, but this time define the end-point as $I(Y_{ij} = 0)$. What do you notice about $\hat{\beta}_1$?

- Model:

$$\log\{-\log(1 - \pi_j)\} = \beta_0 + \beta_1 X_j, \text{ where } \pi_j = Pr(Y_{ij} = 1|X_j)$$

- New Model:

$$\log\{-\log(1 - \pi_j^*)\} = \beta_0^* + \beta_1^* X_j, \text{ where } \pi_j^* = Pr(Y_{ij} = 0|X_j) = 1 - \pi_j$$

Since $\log\{-\log(1 - \pi_j)\} \neq -\log\{-\log(\pi_j)\}$,

$$\beta_1^* \neq -\beta_1$$

(i) Re-fit the all three models using PROC GENMOD and compare them using numeric criteria. Which link function appears to be most appropriate?

| | Deviance | Pearson χ^2 | Likelihood |
|--------|----------|------------------|------------|
| Logit | 11.23 | 10.03 | -186 |
| Probit | 10.12 | 9.51 | -186 |
| CLL | 3.45 | 3.29 | -182 |

- CLL has the smallest deviance (and the smallest Pearson χ^2 and the largest likelihood). So we can conclude that CLL is the most appropriate link function.

BIOSTAT 651
Notes #11:
Conditional Logistic Regression

- Lecture Topics:
 - Matching
 - Analysis of matched pairs
 - Conditional logistic regression
 - Matched case-control studies

Control of confounding

- *Restriction* is a frequently employed method of eliminating confounding in biomedical studies
 - Study on exercise and heart disease
 - Age and gender are confounding factors
 - Restricted the study to men aged 40-65
- Sample restriction has its pros and cons
 - Adv: successfully eliminates confounding
 - Disadv:
 - Can't evaluate the effects of factors that have been restricted for
 - Reduces generality of study's findings

Matching

- Alternative to restriction: *Matching*
 - i.e., match subjects who are very similar with respect to the confounder of concern
- For example, common to match on age, sex, diagnosis
- Matching can be used in each of the studies we've previously described
 - prospective cohort study:
e.g., match treatment A and treatment B subjects by gender and race
 - case-control studies:
e.g., match each case to a control of the same age

Matching (continued)

- Matching eliminates the need to adjust for confounders upon which matching is based
 - not able to estimate the effect of matched covariates
- Unlike restriction, matching need not reduce generality of study
- Analyses of matched data often require methods distinct from those of unmatched study

Matching Schemes

- Various methods are available for matching subjects:
 - 1 : 1 matching
 - 1 : m matching
 - m : n matching
- Depending on the nature of the analysis, matched sets of unequal size may be permitted
- The unit of analysis is typically the *matched set*, as opposed to the subject

Analysis of Matched Pairs

- Consider a study consisting of matched pairs
 - each pair consists of two responses:

 Y_{i1} = response (0,1) from subject 1

 Y_{i2} = response (0,1) from subject 2

 $i = 1, \dots, m$
- Unit of analysis: pair
- Ex. Vaccine study
 - 500 matched pairs. Each pair is matched on gender and age.
 - For example, Pair 1 might be two men, both age 23. Pair 2 might be two women, both age 22.

| Pair | Placebo (Y_{i1}) | Treatment (Y_{i2}) |
|------|----------------------|------------------------|
| 1 | 1 | 0 |
| 2 | 0 | 0 |
| ... | ... | ... |
| 500 | 1 | 1 |

Analysis of Matched Pairs

- Summarize the observed data by the following table:

| | $Y_{i2}=0$ | $Y_{i2}=1$ | total |
|------------|------------|------------|----------|
| $Y_{i1}=0$ | m_{00} | m_{01} | m_{0+} |
| $Y_{i1}=1$ | m_{10} | m_{11} | m_{1+} |
| total | m_{+0} | m_{+1} | m |

McNemar's Test

- Set $\pi_1 = P(Y_{i1} = 1)$ and $\pi_2 = P(Y_{i2} = 1)$
- McNemar's Test
 - $H_0 : \pi_1 = \pi_2$
 - uses only off-diagonal elements
 - test statistic:

$$X_M^2 = \frac{(m_{10} - m_{01})^2}{(m_{10} + m_{01})} \sim \chi_1^2$$

Example: McNemar's Test

- Example: Vaccine study (Page 6)

Y_{i1} = Placebo

Y_{i2} = Treatment

- Responses summarized in the following table:

| | $Y_{i2}=0$ | $Y_{i2}=1$ | total |
|------------|------------|------------|-------|
| $Y_{i1}=0$ | 384 | 18 | 402 |
| $Y_{i1}=1$ | 91 | 7 | 98 |
| total | 475 | 25 | 500 |

- McNemar's Test:

McNemar's Test: SAS Code

- We can carry out McNemar's Test using PROC FREQ:

```
data vaccine;  
  input placebo treatment count;  
  datalines;  
    0    0   384  
    0    1    18  
    1    0    91  
    1    1     7  
  ;  
run;
```

```
proc freq data=vaccine;  
  tables placebo*treatment / agree cmh;  
  weight count;  
run;
```

Regression Analysis of Matched Data

Regression Analysis: Matched Data

- Matched sets are often viewed as *strata*
 - e.g., matching on state (MI, OH, WI, NC) produces $K = 4$ strata
 - e.g., matching by age group (0-14, 15-29, 30-39, 40-49) and diabetes type (I, II, none) produces $K = 12$ strata
- If there are few strata and many subjects in each, then stratum could be incorporated into the \mathbf{x}_i vector
- However, technical issues arise when K is large

Matched Pairs Cohort Study

- The matched-pairs cohort study provides an interesting application of conditional likelihood
- Set-up is as follows:
 - cohort study
 - data consist of matched pairs: $k = 1, \dots, K$
 - observed data for each subject: $(Y_{ik}, \mathbf{x}_{ik})$
covariate: $\mathbf{x}_{ik} = (x_{ik1}, x_{ik2}, \dots, x_{ikq})^T$
parameter of interest: $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$
 - each pair consists of one *treated* and one *untreated* subject
 $x_{1k1} = 1, x_{2k1} = 0$

- Model:

$$\log \left\{ \frac{\pi_{ik}}{1 - \pi_{ik}} \right\} = \alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}$$

Estimation Issues: Stratified Logistic Regression

- Issues in estimating $(\alpha_1, \alpha_2, \dots, \alpha_K, \boldsymbol{\beta}^T)$:

Conditional Logistic Regression

- Alternative to standard logistic regression (applicable to matched data):
 - *conditional* logistic regression
 - uses conditional likelihood
- Set $L_k(\boldsymbol{\beta})$ = conditional likelihood, stratum k
 \propto probability of observed data in stratum k *given some characteristic of stratum*
- Conditional Likelihood:

$$L(\boldsymbol{\beta}) = \prod_{k=1}^K L_k(\boldsymbol{\beta})$$

Matched Pairs: Conditional Likelihood

- Recall (from McNemar's Test): principle that concordant matched pairs provide little information on β
 - therefore, we form the likelihood by conditioning on discordance
- In particular,

$$\begin{aligned} L(\beta) &= \prod_{k=1}^K L_k(\beta) \\ L_k(\beta) &\propto P(\text{observed data, stratum } k) \\ &\propto P(\text{observed data, stratum } k | \text{discordance}) \end{aligned}$$

Matched Pairs: Conditional Likelihood (continued)

- Probability of discordant pair:

$$\begin{aligned} & P(Y_{1k} = 1|\mathbf{x}_{1k})P(Y_{2k} = 0|\mathbf{x}_{2k}) \\ & + P(Y_{1k} = 0|\mathbf{x}_{1k})P(Y_{2k} = 1|\mathbf{x}_{2k}) \\ = & \pi(\mathbf{x}_{1k})\{1 - \pi(\mathbf{x}_{2k})\} \end{aligned} \tag{1}$$

$$+ \pi(\mathbf{x}_{2k})\{1 - \pi(\mathbf{x}_{1k})\} \tag{2}$$

- Conditional probabilities:

$$P(Y_{1k} = 1|\text{discordance}) =$$

$$P(Y_{2k} = 1|\text{discordance}) =$$

Forming Conditional Likelihood: MPC

- Recall that under the assumed model,

$$\pi(\mathbf{x}_{ik}) = \frac{e^{\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}}}$$

- Therefore, we have

$$\begin{aligned} (1) &= \frac{e^{\alpha_k + \mathbf{x}_{1k}^T \boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{x}_{1k}^T \boldsymbol{\beta}}} \frac{1}{1 + e^{\alpha_k + \mathbf{x}_{2k}^T \boldsymbol{\beta}}} \\ (2) &= \frac{1}{1 + e^{\alpha_k + \mathbf{x}_{1k}^T \boldsymbol{\beta}}} \frac{e^{\alpha_k + \mathbf{x}_{2k}^T \boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{x}_{2k}^T \boldsymbol{\beta}}} \end{aligned}$$

- We then obtain

$$\begin{aligned} \frac{(1)}{(1) + (2)} &= \frac{e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \\ \frac{(2)}{(1) + (2)} &= \frac{1}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \end{aligned}$$

Matched Pair: Conditional Likelihood

- Finally, the conditional likelihood is then given by:

$$L_k(\boldsymbol{\beta}) = \left\{ \frac{e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \right\}^{Y_{1k}(1 - Y_{2k})} \times \left\{ \frac{1}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \right\}^{Y_{2k}(1 - Y_{1k})}$$

- Equal to the typical logistic regression likelihood, except:
 - using only discordant pairs
 - one record per matched pair
 - response: $Y_k^* = Y_{1k}$
 - covariate: $\mathbf{x}_k^* = \mathbf{x}_{1k} - \mathbf{x}_{2k}$
 - no intercept term

Matched Case-Control Studies

Matched Case-Control Study

- Matched-data set-up:
 - case-control study
 - total of K strata: $k = 1, \dots, K$
 - n_{1k} cases and n_{0k} controls in stratum k
 - set $n_k = n_{0k} + n_{1k}$
 - K can be quite large, with n_k generally small
 - set $\pi_{ik} = P(Y_{ik} = 1 | \mathbf{x}_{ik})$

- Model:

$$\log \left\{ \frac{\pi_{ik}}{1 - \pi_{ik}} \right\} = \alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}$$

$$\boldsymbol{\beta} =$$

$$\mathbf{x}_{ik}^T =$$

Conditional Likelihood: Case-Control Study

- Set $L_k(\boldsymbol{\beta}) =$ conditional likelihood, stratum k
 \propto probability of observed data in stratum k *given the total number of cases*

- Note: among n_k subjects, number of case assignments:

$$\begin{pmatrix} n_k \\ n_{1k} \end{pmatrix} \equiv c_k$$

- Matched pair: $c_k = 2$

Conditional Likelihood: Matched pair

- Probability of observed data (stratum k):

$$\prod_{i=1}^{n_k} P(\mathbf{x}_{ik} | Y_{ik} = 1)^{Y_{ik}} P(\mathbf{x}_{ik} | Y_{ik} = 0)^{1-Y_{ik}}$$

- *Conditional* probability of observed data (stratum k), *given* the total number of cases (stratum k):

$$\frac{\prod_{i=1}^{n_k} P(\mathbf{x}_{ik} | Y_{ik} = 1)^{Y_{ik}} P(\mathbf{x}_{ik} | Y_{ik} = 0)^{1-Y_{ik}}}{\sum_{j=1}^{c_k} \prod_{i=1}^{n_k} P(\mathbf{x}_{ik} | Y_{i(j)k} = 1)^{Y_{i(j)k}} P(\mathbf{x}_{ik} | Y_{i(j)k} = 0)^{1-Y_{i(j)k}}}$$

- Re-write key probability:

$$\begin{aligned} P(\mathbf{x}_{ik} | Y_{ik} = 1) &= \frac{P(Y_{ik} = 1 | \mathbf{x}_{ik}) P(\mathbf{x}_{ik})}{P(Y_{ik} = 1)} \\ &= \frac{\pi(\mathbf{x}_{ik}) P(\mathbf{x}_{ik})}{P(Y_{ik} = 1)} \end{aligned}$$

Constructing Conditional Likelihood (continued)

- We can then write:

$$L_k(\boldsymbol{\beta}) \propto \frac{\prod_{i=1}^{n_k} \pi(\mathbf{x}_{ik})^{Y_{ik}} \{1 - \pi(\mathbf{x}_{ik})\}^{1-Y_{ik}}}{\sum_{j=1}^{c_k} \prod_{i=1}^{n_k} \pi(\mathbf{x}_{ik})^{Y_{i(j)k}} \{1 - \pi(\mathbf{x}_{ik})\}^{1-Y_{i(j)k}}}$$

- Now, we recall that

$$\pi(\mathbf{x}_{ik}) = \frac{e^{\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}}}$$

such that the denominators in the above RHS cancel out

Constructing Conditional Likelihood (cont'd)

- We are then left with

$$L_k(\boldsymbol{\beta}) \propto \frac{\prod_{i=1}^{n_k} e^{(\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}) Y_{ik}}}{\sum_{j=1}^{c_k} \prod_{i=1}^{n_k} e^{(\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}) Y_{i(j)k}}}$$

- Finally, we arrive at our conditional likelihood:

$$L_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_k} e^{\mathbf{x}_{ik}^T \boldsymbol{\beta} Y_{ik}}}{\sum_{j=1}^{c_k} \prod_{i=1}^{n_k} e^{\mathbf{x}_{ik}^T \boldsymbol{\beta} Y_{i(j)k}}}$$
$$L(\boldsymbol{\beta}) = \prod_{k=1}^K L_k(\boldsymbol{\beta})$$

- It has been shown that $L(\boldsymbol{\beta})$ possesses the key properties of a typical likelihood function ...

Conditional Likelihood: Case-Control v. Cohort

- Q: How does this version of $L_k(\boldsymbol{\beta})$ relate to that used for matched cohort studies?
 - recall: for MPC data:

$$L_k(\boldsymbol{\beta}) = \left\{ \frac{e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \right\}^{Y_{1k}(1-Y_{2k})} \times \left\{ \frac{1}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \right\}^{Y_{2k}(1-Y_{1k})}$$

- and, for matched case-control data (1:1 matching with $Y_{1k} = 1$ and $Y_{2k} = 0$)

$$\begin{aligned} L_k(\boldsymbol{\beta}) &= \frac{e^{\mathbf{x}_{1k}^T \boldsymbol{\beta}}}{e^{\mathbf{x}_{1k}^T \boldsymbol{\beta}} + e^{\mathbf{x}_{2k}^T \boldsymbol{\beta}}} \\ &= \left\{ \frac{e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \right\}^{Y_{1k}(1-Y_{2k})} \end{aligned}$$

Example: Matched Cohort Data

Researchers studied the effect of a new treatment on a fairly frequently occurring skin condition. A total of 79 clinics participated in the study. Each clinic recruited two patients, with one randomized to receive the new treatment, the other receiving placebo. The two patients from a given clinic were matched on various demographic variates, reflecting income, socioeconomic status and baseline general health. Patients evaluated after 30 days post-randomization, then classified based on whether or not their condition improved. Adjustment covariates include age (AGE; recorded in years), SEX, and initial grade of skin condition (1, 2, 3, 4).

(a) The input records are as given below,

1 t f 27 0 1 1 p f 32 0 2

with the matched pair from each center contained in the same record. Read in the data set, then print off the resulting SAS file.

- See the SAS code

(b) Fit a logistic model which ignores the matching by center. Estimate the treatment effect based on this model.

- Model (patient i in clinic k)

$$\begin{aligned} \text{logit}(\pi_{ik}) &= \beta_0 + \beta_1 TRT_{ik} + \beta_2 Female_{ik} \\ &+ \beta_3 AGE_{ik} + \beta_4 Grade_{ik} \end{aligned}$$

$$TRT_{ik} = I(NewTreatment)$$

$$Female_{ik} = I(Female)$$

- Estimates: $\hat{\beta}_1 = 0.8803$, $P - value : 0.0152$

- (c) Why might the unmatched analysis from (b) yield biased parameter estimates?

If the true model is

$$\begin{aligned} \text{logit}(\pi_{ik}) &= \alpha_k + \beta_1 TRT_{ik} + \beta_2 Female_{ik} \\ &+ \beta_3 AGE_{ik} + \beta_4 Grade_{ik} \end{aligned}$$

and at least one α_k is different from the others, we can have biased results since the model in (b) does not adjust for center (k).

- (d) Under what conditions would the parameter estimates from the unmatched analysis be valid?
- If $\alpha_1 = \dots = \alpha_K$
- (e) Fit a model which adjusts for center. Examine the log file, then comment.
- Warning messages. Parameter estimating

procedure is not stable, so the validity of the results is questionable.

- $\hat{\beta}_1 = 1.4049$, $P - value : 0.0058$

(f) Fit the a logistic model using the conditional likelihood from matched pairs described for cohort studies in the lecture notes.

- $\hat{\beta}_1 = 0.7024$, $P - value : 0.0511$

(g) Re-fit the conditional logistic regression model using the strata statement. Compare your results with those obtained previously.

- Results are the same.

(h) Why can't we estimate the center effects in this set-up?

- We used the conditional logistic regression, which factors out the center effect (α_k).
- Even if we try other methods, the small sample size in each center (sample size=2) does not allow to stably estimate the center effect.

Example: Matched Case-Control Study

A group of oncologists studied women living in a retirement community, to determine if there was an association between the use of estrogen and the incidence of endometrial cancer. Cases were matched to controls who were approximately the same age, had the same marital status, and were living in the same community as the case at the time of the case's diagnosis. In addition to estrogen use, information was also collected from each subject on hypertension, gallbladder disease history, and prescription drug use.

(a) Fit a logistic regression model which accounts for the matching and includes estrogen and all available adjustment covariates.

- Among variables used for the matching, age is available.
- Model (patient i in pair k)

$$\begin{aligned} \text{logit}(\pi_{ik}) &= \alpha_k + \beta_1 \text{Estrogen}_{ik} + \beta_2 \text{HYP}_{ik} \\ &+ \beta_3 \text{Gall}_{ik} + \beta_4 \text{Drug}_{ik} + \beta_5 \text{Age}_{ik} \end{aligned}$$

- Conditional likelihood ($Y_{1k} = 1, Y_{2k} = 0$)

$$L = \prod_{k=1}^K L_k(\beta) = \prod_{k=1}^K \frac{\exp\{(x_{1k} - x_{2k})'\beta\}}{1 + \exp\{(x_{1k} - x_{2k})'\beta\}}$$

- Fit the logistic regression with the transformed x and y .
 - Unit is the pair
 - $x_k^* = x_{1k} - x_{2k}$
 - $y_k^* = 1$
 - No intercept

(b) Examine the results, paying particular attention to AGE. Given that age is known to be a risk factor for endometrial cancer, do the results of the fitted model make sense?

- Age (β_5): p value = 0.7135
- Age is not significant because it is used for the matching.

(c) Re-fit the model, this time not adjusting for age. Compare the estrogen effects based on this and the previous model.

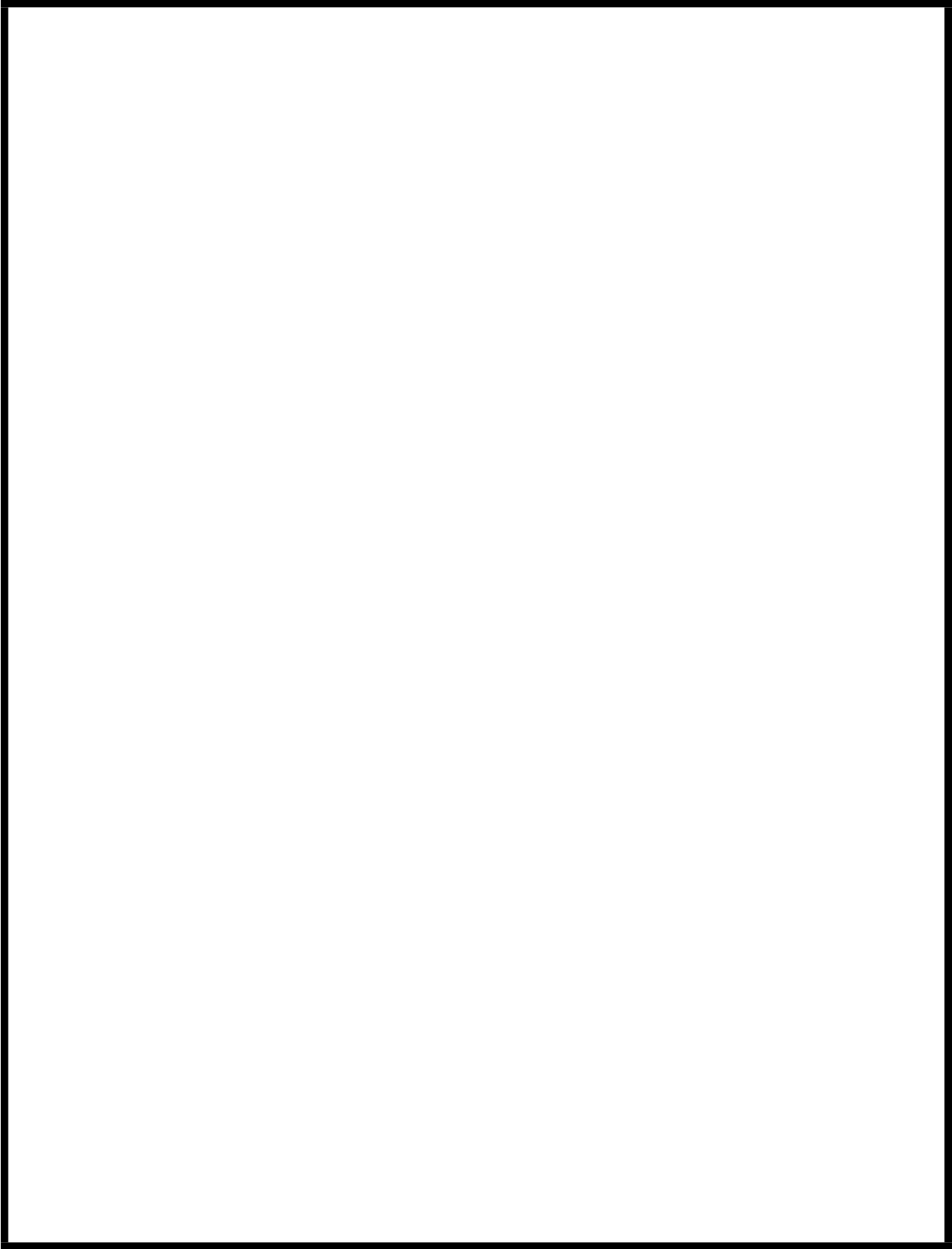
- With age:

$$\hat{\beta}_1 = 2.879, \quad \text{p-value} = 0.001$$

- Without age:

$$\hat{\beta}_1 = 2.814, \quad \text{p-value} = 0.0008$$

- Nearly identical.



BIOSTAT 651
Notes #12:
Multinomial Data

- Lecture Topics:
 - Models for multinomial data
 - Interpretation of parameters
 - Examples
- Text (Dobson & Barnett, 3rd Ed.): Chapter 8

Nominal, Ordinal Scales

- *Nominal* data:
 - state of residence after graduation:
MI, OH, WI, IL, NC
 - political affiliation:
Republican, Independent, Democrat
 - T.V. channel preferences:
Fox, CNN, MSNBC, do-not-watch-TV
- *Ordinal* data:
 - pain:
slight, moderate, severe
 - grade:
A+, A, A-, B+, B, B-

Multinomial Data

- Often in practice, the response is multinomial (> 2 categories)
- Simplification: reduce to 2 categories; employ logistic regression
 - may be met with skepticism by investigators and/or reviewers
 - how such grouping affected the results will be an issue
 - may result in a considerably less informative analysis

Multinomial : Exponential Family

- Suppose $\mathbf{Y} \sim \text{Multinomial}(n, \boldsymbol{\pi})$ with $J + 1$ categories.

$$\mathbf{Y} = (Y_0, Y_1, \dots, Y_J)'$$

$$\boldsymbol{\pi} = (\pi_0, \pi_1, \dots, \pi_J)'$$

Y_j : number of subjects with $Y = j$

π_j : probability of $Y = j$

- Moments:

$$E[Y_j] =$$

$$V(Y_j) =$$

$$\text{cov}(Y_j, Y_k) =$$

- Probability function:

$$p(\mathbf{Y}; \boldsymbol{\pi}) = \binom{n}{Y_0 \ Y_1 \ \dots \ Y_J} \prod_{j=0}^J \pi_j^{Y_j}$$

Multinomial : Exponential Family (continued)

- Write as an exponential family:

$$p(\mathbf{Y}; \boldsymbol{\pi}) = \exp \left\{ \sum_{j=0}^J Y_j \log \pi_j + \log C(n; \mathbf{Y}) \right\}$$

- Consider the constraints:

$$\begin{aligned} - \sum_{j=0}^J Y_j &= n \\ - \sum_{j=0}^J \pi_j &= 1 \end{aligned}$$

Multinomial : Exponential Family (cont'd)

- Incorporating these constraints, we re-write

$$p(\mathbf{Y}; \boldsymbol{\pi}) =$$

$$\begin{aligned} & \exp \left\{ Y_0 \log \pi_0 + \sum_{j=1}^J Y_j \log \pi_j + \log C(n; \mathbf{Y}) \right\} \\ = & \exp \left\{ \left(n - \sum_{j=1}^J Y_j \right) \log \pi_0 + \sum_{j=1}^J Y_j \log \pi_j \right. \\ & \quad \left. + \log C(n; \mathbf{Y}) \right\} \\ = & \exp \left\{ \sum_{j=1}^J Y_j \log \left(\frac{\pi_j}{\pi_0} \right) + n \log \pi_0 + \log C(n; \mathbf{Y}) \right\} \end{aligned}$$

- This is an exponential family with:

$$t(\mathbf{Y}) =$$

$$\boldsymbol{\theta} =$$

Deriving Multinomial Parameters

- Examining the natural parameter:

$$\begin{array}{rcl} \frac{\pi_1}{\pi_0} & = & e^{\theta_1} \\ & \vdots & \\ & \vdots & \\ \frac{\pi_J}{\pi_0} & = & e^{\theta_J} \end{array}$$

- Therefore, we obtain

$$\sum_{j=1}^J e^{\theta_j} = \frac{1}{\pi_0} \sum_{j=1}^J \pi_j = \frac{1}{\pi_0} (1 - \pi_0)$$

such that

$$\pi_0 =$$

$$\pi_j =$$

Properties of Multinomial

- We then have

$$\begin{aligned} b(\boldsymbol{\theta}) &= -n \log \pi_0 \\ &= n \log \left(1 + \sum_{j=1}^J e^{\theta_j} \right) \end{aligned}$$

- The canonical link: since $\theta_j = \log(\pi_j/\pi_0)$

$$\log \frac{\pi_{ij}}{\pi_{i0}} = \mathbf{x}_i^T \boldsymbol{\beta}_j$$

for $j = 1, \dots, J$

Multinomial: Regression Framework

- We now consider regression modeling of multinomial outcomes
- General set-up:

Y_i = response, subject i

set $Y_{ij} = I(Y_i = j) \quad j = 0, 1, \dots, J$

such that $Y_i = \sum_{j=0}^J j \times Y_{ij}$

covariate: $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iq})'$

- Outcome probabilities:

$$\pi_{ij} = \pi_j(\mathbf{x}_i) = P(Y_i = j | \mathbf{x}_i)$$

Generalized Logits Model

- Generalized logits model:

$$\log \left(\frac{\pi_{ij}}{\pi_{i0}} \right) = \mathbf{x}_i^T \boldsymbol{\beta}_j$$

- one category defined as the *reference*
 - distinct intercepts and regression coefficients
- Response probabilities:

$$P(Y_i = j | \mathbf{x}_i) = \pi_{ij} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{j=1}^J \exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}$$

$$P(Y_i = 0 | \mathbf{x}_i) = \pi_{i0} = \frac{1}{1 + \sum_{j=1}^J \exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}$$

- Direct generalization of logistic regression to > 2 response categories

MLE: Generalized Logits Model

- Set $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \dots, \boldsymbol{\beta}'_J)$

dimension = $(q + 1) \times J$

- Likelihood is given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \prod_{j=0}^J \pi_{ij}^{Y_{ij}}$$

- Subbing in $\boldsymbol{\beta}$, then taking log:

$$\begin{aligned} \ell(\boldsymbol{\beta}) = & \sum_{i=1}^n \left\{ -Y_{i0} \log \left(1 + \sum_{\ell=1}^J e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell} \right) \right. \\ & \left. + \sum_{j=1}^J Y_{ij} \log \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}}{1 + \sum_{\ell=1}^J e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell}} \right) \right\} \end{aligned}$$

- Recalling that $Y_{i0} = 1 - \sum_{j=1}^J Y_{ij}$

we then obtain:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^n \left\{ - \left(1 - \sum_{j=1}^J Y_{ij} \right) \log \left(1 + \sum_{\ell=1}^J e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell} \right) \right. \\
&\quad \left. + \sum_{j=1}^J Y_{ij} \mathbf{x}_i^T \boldsymbol{\beta}_j - \sum_{j=1}^J Y_{ij} \log \left(1 + \sum_{\ell=1}^J e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell} \right) \right\} \\
&= \sum_{i=1}^n \left\{ - \log \left(1 + \sum_{\ell=1}^J e^{\mathbf{x}_i^T \boldsymbol{\beta}_\ell} \right) + \sum_{j=1}^J Y_{ij} \mathbf{x}_i^T \boldsymbol{\beta}_j \right\}
\end{aligned}$$

- The score function is then given by:

$$U(\boldsymbol{\beta}) = \begin{pmatrix} U_1(\boldsymbol{\beta}) \\ \vdots \\ U_J(\boldsymbol{\beta}) \end{pmatrix}$$

with j th sub-vector:

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_{ij} - \pi_{ij}) \mathbf{x}_i$$

MLE: Generalized Logits Model (continued)

- Information matrix:

dimension of $J(\boldsymbol{\beta})$: $\{J(q+1)\} \times \{J(q+1)\}$

$$J(\boldsymbol{\beta}) = \begin{pmatrix} J_{11}(\boldsymbol{\beta}) & \cdots & J_{1J}(\boldsymbol{\beta}) \\ & J_{22}(\boldsymbol{\beta}) & \vdots \\ & \cdots & J_{JJ}(\boldsymbol{\beta}) \end{pmatrix}$$

(j, j) th block: $\sum_{i=1}^n \pi_{ij}(1 - \pi_{ij}) \mathbf{x}_i^T \mathbf{x}_i$

(j, k) th block: $-\sum_{i=1}^n \pi_{ij}\pi_{ik} \mathbf{x}_i^T \mathbf{x}_i$

- Note that $U_j(\boldsymbol{\beta})$ and $J_{jj}(\boldsymbol{\beta})$ are analogous to a logistic regression which considers categories j and 0 only

Gen Logit Model: Example

- Example: We reconsider the study of childhood asthma. The objective was to determine the role of gender in pre-school asthma incidence. Children enrolled in the study ($n=100$) were followed prospectively in order to determine whether they were hospitalized for asthma. Suppose now that some children were kept overnight after being admitted ($Y_i = 2$), while others were tested then discharged the same day ($Y_i = 1$). Of course, many children did not suffer asthma ($Y_i = 0$).

Gen Logit Model: Example (continued)

The observed data are summarized by the following table:

| | $Y_i=0$ | $Y_i=1$ | $Y_i=2$ | total |
|---------|---------|---------|---------|-------|
| $F_i=0$ | 24 | 14 | 22 | 60 |
| $F_i=1$ | 21 | 10 | 9 | 40 |

- The model is given by:

$$\log \left\{ \frac{\pi_{ij}}{\pi_{i0}} \right\} = \beta_{0j} + \beta_{1j} F_i$$

for $j = 1, 2$

- need to estimate 2 intercepts, and 2 differences
- Note: this model is *saturated*

Gen Logit Example: Interpretation)

- Interpretation of parameters:

- e.g., set $j = 1$; intercept:

$$\beta_{01} = \log \left\{ \frac{P(Y_i = 1|F_i = 0)}{P(Y_i = 0|F_i = 0)} \right\}$$

- Q: Does this lead to an odds?

- set $j = 1$; difference:

$$\beta_{01} + \beta_{11} = \log \left\{ \frac{P(Y_i = 1|F_i = 1)}{P(Y_i = 0|F_i = 1)} \right\}$$

$$\begin{aligned} \beta_{11} &= \log \left\{ \frac{P(Y_i = 1|F_i = 1)}{P(Y_i = 0|F_i = 1)} \right\} \\ &\quad - \log \left\{ \frac{P(Y_i = 1|F_i = 0)}{P(Y_i = 0|F_i = 0)} \right\} \end{aligned}$$

$$\exp\{\beta_{11}\} = \frac{P(Y_i = 1|F_i = 1)}{P(Y_i = 0|F_i = 1)} / \frac{P(Y_i = 1|F_i = 0)}{P(Y_i = 0|F_i = 0)}$$

- Q: Is this an odds ratio?

Gen Logit Model: Example (cont'd)

- Estimated intercepts:

$$\hat{\beta}_{01} = \log \left(\frac{14/60}{24/60} \right) = -0.539$$

$$\hat{\beta}_{02} = \log \left(\frac{22}{24} \right) = -0.087$$

- Estimated differences:

$$\hat{\beta}_{01} + \hat{\beta}_{11} = \log \left(\frac{10/40}{21/40} \right) = -0.742$$

$$\hat{\beta}_{02} + \hat{\beta}_{12} = \log \left(\frac{9/40}{21/40} \right) = -0.847$$

- Estimated gender effects:

$$\exp\{\hat{\beta}_{11}\} = 0.81$$

$$\exp\{\hat{\beta}_{12}\} = 0.48$$

Ordinal data: Cumulative Logit Model

- Generalized logit model is applicable when there are > 2 response categories
 - applicable whether or not the response categories are ordered
 - does not exploit the ordering of categories
- Several other modeling options exist for ordinal responses
- Cumulative logit model:

$$\begin{aligned}\log \left\{ \frac{P(Y_i \leq j)}{P(Y_i > j)} \right\} &= \log \left\{ \frac{\pi_0 + \pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} \right\} \\ &= \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}_j\end{aligned}$$

for $j = 0, 1, \dots, (J - 1)$

- distinct intercepts and covariate parameters

Ordinal data: Proportional Odds Model

- Cumulative logit model: consider the following setting:
 - $(J + 1)$ response levels
 - \mathbf{x}_i : $q \times 1$ covariate
 - number of parameters: $(q + 1) \times J$
- Proportional odds model:
 - Assumes $\gamma = \gamma_0 = \gamma_1 = \cdots = \gamma_{J-1}$
 - fewer parameters to explain to investigator

$$\log \left\{ \frac{P(Y_i \leq j)}{P(Y_i > j)} \right\} = \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}$$

- Trade-off: PO model entails additional assumptions namely, *proportionality*

Ordinal data: Proportional Odds Model

- Proportionality assumption

- Odds:

$$odds(Y_i \leq j | \mathbf{x}_1) = \frac{P(Y_i \leq j | \mathbf{x}_1)}{P(Y_i > j | \mathbf{x}_1)} = \exp(\gamma_{0j}) \exp(\mathbf{x}_1^T \boldsymbol{\gamma})$$

- Odds ratio (independent of j):

$$\frac{odds(Y_i \leq j | \mathbf{x}_1)}{odds(Y_i \leq j | \mathbf{x}_2)} = \exp((\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\gamma})$$

- Number of parameters: $J + q$

Proportional Odds Model: Example

- Example: Consider a study of incomes of students after completing their degree. Information on the student (age, gender, GPA) and program of study (region, level of program) is captured by a covariate, \mathbf{x}_i . The response variate has three levels: low ($Y_i = 0$), medium ($Y_i = 1$) and high ($Y_i = 2$).

Prop Odds Model: Example (continued)

- Suppose the cut-point is $j = 0$:

$$\log \left\{ \frac{P(Y_i \leq 0)}{P(Y_i > 0)} \right\} = \gamma_{00} + \mathbf{x}_i^T \boldsymbol{\gamma}_{10} \quad (1)$$

which separates ‘low’ from ‘medium, high’

- Another choice could be $j = 1$:

$$\log \left\{ \frac{P(Y_i \leq 1)}{P(Y_i > 1)} \right\} = \gamma_{01} + \mathbf{x}_i^T \boldsymbol{\gamma}_{11} \quad (2)$$

which separates ‘high’ from ‘medium, low’

- A proportional odds model would assume that (1) and (2) hold, but with the constraint that $\boldsymbol{\gamma}_{10} = \boldsymbol{\gamma}_{11}$

Prop Odds Model: Example (continued)

- Consider a reduced model, for which $\mathbf{x}_i^T = [1, A_i, G_i]$, where A_i equals age (years) and G_i is an indicator for graduate degree (ref = undergraduate).
 - proportional odds model:

$$\log \left\{ \frac{P(Y_i \leq 0)}{P(Y_i > 0)} \right\} = \gamma_{00} + \gamma_A(A_i - 20) + \gamma_G G_i$$

$$\log \left\{ \frac{P(Y_i \leq 1)}{P(Y_i > 1)} \right\} = \gamma_{01} + \gamma_A(A_i - 20) + \gamma_G G_i$$

- Interpreting the parameters (e.g., $j = 0$):

$$e^{\gamma_{00}}:$$

$$e^{\gamma_G}:$$

Prop Odds Model: Example (cont'd)

- Based on proportional odds model:
 - can estimate odds through

$$\frac{P(Y_i \leq j)}{P(Y_i > j)} = e^{\gamma_{0j} + \mathbf{x}_i^T \gamma_1}$$

- probabilities:

$$P(Y_i \leq j) = \frac{e^{\gamma_{0j} + \mathbf{x}_i^T \gamma_1}}{1 + e^{\gamma_{0j} + \mathbf{x}_i^T \gamma_1}}$$

- cell probabilities:

$$P(Y_i = j) = P(Y_i \leq j) - P(Y_i \leq j - 1)$$

Test for Proportionality

- Compare the models with and without the proportionality assumption
 - H0: $\gamma = \gamma_0 = \gamma_1 = \dots = \gamma_{J-1}$
 - Full model: without the proportionality assumption ($j = 0, \dots, J - 1$)

$$\log \left\{ \frac{P(Y_i \leq j)}{P(Y_i > j)} \right\} = \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}_{1j}$$

- Reduced model: with the assumption

$$\log \left\{ \frac{P(Y_i \leq j)}{P(Y_i > j)} \right\} = \gamma_{0j} + \mathbf{x}_i^T \boldsymbol{\gamma}$$

- SAS carries out score test.
- Reference distribution: ...

Example: Multinomial Regression

A retrospective cohort study conducted at the University of Massachusetts sought to evaluate the attitudes and characteristics associated with mammography. A total of $n = 412$ women were classified into one of the following categories:

- *never* had a mammogram ($Y_i = 0$)
- had a mammogram *within the past year* ($Y_i = 1$)
- had a mammogram, but *more than one year ago* ($Y_i = 2$).

Covariates of of interest were based on responses to a questionnaire and included:

SYMP REQ: 1=SA, 2=A, 3=D, 4=SD

“You do not need a mammogram unless you develop symptoms”

PERC BEN: 5, . . . , 20

“Score the potential benefit of a mammogram”

FAM HIST: 0=No, 1=Yes

“Do you have a mother or sister with breast cancer?”

TAUGHT BSE: 0=No, 1=Yes

“Were you taught how to do a breast self-exam?”

MAMM DET: 1=NL, 2=SL, 3=VL

“How likely would a mammogram detect a new breast tumor?”

- (a) Recode the response variate such that response value could be treated as ordered.
- Recode:
 - *never* had a mammogram ($Y_i = 0$)
 - had a mammogram *within the past year* ($Y_i = 2$)
 - had a mammogram, but *more than one year ago* ($Y_i = 1$).
- (b) To begin, we focus on family history. Print out a cross-classification of mammography experience and family history.
- See the SAS code
- (c) Write down a generalized logit model pertaining to the 2×3 table.

$$\log \left(\frac{\pi_{ij}}{\pi_{i0}} \right) = \beta_{0j} + \beta_{1j} X_i \quad j = 1, 2$$

X_i : Family history (1=YES, 0=NO)

$$\pi_{ij} = P(Y_i = j | X_i)$$

- (d) Fit the generalized logit model to the table, by hand.

$$\hat{\beta}_{01} = \log(63/220) = -1.25$$

$$\hat{\beta}_{11} = \log(11/14) - \log(63/220) = 1.01$$

$$\hat{\beta}_{02} = \log(85/220) = -0.95$$

$$\hat{\beta}_{12} = \log(19/14) - \log(85/220) = 1.26$$

- OR: $\exp(\hat{\beta}_{11}) = 2.74$ $\exp(\hat{\beta}_{12}) = 3.51$

- (e) Re-estimate the family history effect parameters, this time using 2×2 table calculations.

- Table $Y = 0$ vs $Y = 1$

$$OR = (220 * 11) / (63 * 14) = 2.744$$

- Table $Y = 0$ vs $Y = 2$

$$OR = (220 * 19) / (85 * 14) = 3.51$$

- OR estimates are identical.

- (f) Estimate the standard errors of $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$.

NOTE: standard error of logOR in 2×2 table:

$$SE = \sqrt{1/n_{11} + 1/n_{10} + 1/n_{01} + 1/n_{00}}$$

$$\widehat{SE}(\beta_{11}) = \sqrt{1/220 + 1/63 + 1/14 + 1/11} = 0.4275$$

$$\widehat{SE}(\beta_{12}) = \sqrt{1/220 + 1/85 + 1/19 + 1/11} = 0.3747$$

- (g) Re-fit the GL model, this time using PROC LOGISTIC. Compare your parameter estimates and SEs to those obtained by hand.

- See the SAS code (Results are identical)

- (h) Interpret $\exp\{\hat{\beta}_{11}\}$.

Odds ratio of having a mammogram comparing with and without family history, given that the women did not have a mammogram within the past year.

- (i) Carry out a Wald test of $H_0 : \beta_{11} = \beta_{12} = 0$.

Test statistic: $X_w^2 = 12.01$

P-value: 0.0025

Reject H_0 .

- (j) Carry out a likelihood ratio test of the hypothesis listed in (i).

Full model $-2 \log L$: 792.340

Reduced model $-2 \log L$: 805.198

Test statistic: $X_L^2 = 805.198 - 792.340 = 12.858$

$12.858 > 5.99 = \chi_{2,0.95}^2$

Reject H_0 .

- (k) Write out an appropriate proportional odds model, again focusing only on family history.

$$\log\left\{\frac{P(Y_i \leq j|X_i)}{P(Y_i > j|X_i)}\right\} = \gamma_{0j} + \gamma_1 X_i \quad j = 0, 1$$

- (ℓ) Fit the PO model using PROC LOGISTIC.

See the SAS code

- (m) Interpret $\exp\{\hat{\gamma}_1\}$ based on the PO model.

- $\exp\{\hat{\gamma}_1\} = 0.355$

Odds ratio of being in lower mammogram categories comparing with and without family history is 0.355

- (n) Estimate the probability that a woman with a

family history of breast cancer had a mammogram more than one year ago.

$$P(Y_i \leq 0 | X_i = 1) = \frac{\exp(\widehat{\gamma_{00}} + \widehat{\gamma_1})}{1 + \exp(\widehat{\gamma_{00}} + \widehat{\gamma_1})} = 0.344$$

$$P(Y_i \leq 1 | X_i = 1) = \frac{\exp(\widehat{\gamma_{01}} + \widehat{\gamma_1})}{1 + \exp(\widehat{\gamma_{01}} + \widehat{\gamma_1})} = 0.547$$

$$P(Y_i = 1 | X_i = 1) = P(Y_i \leq 1 | X_i = 1) - P(Y_i \leq 0 | X_i = 1) = 0.203$$

- (o) Carry out a score test of the proportionality assumption.

$$X_s^2 = 0.5951 < 3.84 = \chi_{1,0.95}^2$$

Cannot reject H0

There is no strong evidence that the proportionality assumption is not satisfied.

(p) List the two models being compared in the test for proportionality.

- H0 (3 parameters)

$$\log\left\{\frac{P(Y_i \leq j|X_i)}{P(Y_i > j|X_i)}\right\} = \gamma_{0j} + \gamma_1 X_i \quad j = 0, 1$$

- H1 (4 parameters)

$$\log\left\{\frac{P(Y_i \leq j|X_i)}{P(Y_i > j|X_i)}\right\} = \gamma_{0j} + \gamma_{1j} X_i \quad j = 0, 1$$

BIOSTAT 651
Notes #13: Poisson Regression

- Lecture Topics:
 - Motivation: count, rate data
 - Inference procedures
 - Examples
- Text (Dobson & Barnett, 3rd Ed.): Chapter 9

Data Structure: Count Response

- Often in biomedical studies, the response variate is a *count*
 - response = event count
- Examples:
 - number of incident lung cancer cases in Michigan, 2000-2004
 - number of acute liver failure cases diagnosed between 01/01/1990 and 12/31/1999, among males age 40-64 working in either Michigan, Ohio or Illinois.
 - number of physician visits for influenza among University of Michigan students in 2005

Data Structure: Count, Rate

- If follow-up time (or, *exposure*) is constant across all subjects, we can model the mean count directly

- more generally, we model rate, where:

$$\text{rate}_i = \frac{E(\text{count}_i)}{\text{exposure}_i}$$

- Rates are typically reported as events per unit time (or person-time)
 - e.g., cases per 100,000 patient-years
- Units of time dimension are arbitrary
 - must use same units for all observations.

Example: Equal Exposure

- Example: A group of clinicians wishes to study the association between air temperature and physician claims for asthma. The setting is Ann Arbor, with observed data given in the following table:

| Month | Avg. Temp | Claims |
|-------|-----------|--------|
| Sep | 51 | 16 |
| Oct | 44 | 14 |
| Nov | 40 | 9 |
| Dec | 29 | 6 |
| Jan | 23 | 11 |
| Feb | 21 | 8 |
| Mar | 32 | 15 |
| Apr | 41 | 17 |
| May | 62 | 7 |
| Jun | 78 | 5 |

Examples: Unequal Exposure

- Example: A study carried out at the University of Michigan Department of Internal Medicine examined hospital admission patterns among dialysis patients. Each patient was followed from the initiation of dialysis until 12/31/2009. Below is a subset of the records from the analysis file:

| IDNUM | YEARS | ADMITS | DIAB | AGE |
|-------|-------|--------|------|-----|
| 1 | 0.6 | 5 | 1 | 61 |
| 2 | 4.2 | 11 | 0 | 56 |
| 3 | 1.3 | 7 | 1 | 45 |
| 4 | 2.2 | 8 | 0 | 66 |

Poisson Distribution

- General data structure:

Y_i = event count

\mathbf{x}_i = covariate vector

T_i = exposure

observed data: (\mathbf{x}_i, Y_i, T_i) for $i = 1, \dots, n$

- Assume that:

- triplets (\mathbf{x}_i, Y_i, T_i) are *iid*

- Probability function:

- Moments:

- Rate function:

Poisson Regression Model

- Rate model:

$$\log \lambda_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

- written in terms of mean function:

$$\log E \left[\frac{Y_i}{T_i} \right] = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$\log E[Y_i] = \log T_i + \mathbf{x}_i^T \boldsymbol{\beta}$$

- the term $\log(T_i)$ is referred to as an offset

Poisson Regression as a GLM

- Poisson regression model is a special case of a GLM
 - link function:
 - mean function:
 - variance function:
- Standard implementation features the canonical link

Interpretation of Parameters

- Consider a Poisson model with a single covariate where (for now) X_i is continuous

$$\log \lambda_i = \beta_0 + \beta_1(X_i - \overline{X})$$

- Interpretation of β_0 :

- Interpretation of β_1 :

Interpreting Parameters

- Suppose now that the covariate is binary:

$$X_i = I_i(\text{treated})$$

$$\log \lambda_i = \beta_0 + \beta_1 X_i$$

- Interpretation of β_0 :

- Interpretation of β_1 :

Poisson Regression: Estimation

- Applying general GLM results to the case where $Y_i \sim$ follows the Poisson model:

$$\lambda_i = \lambda(\mathbf{x}_i) = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$$

- Score function:

$$\begin{aligned} U(\boldsymbol{\beta}) &= \sum_{i=1}^n \mathbf{x}_i^T (y_i - \mu_i) = \sum_{i=1}^n \mathbf{x}_i^T (y_i - T_i \lambda_i) \\ &= \sum_{i=1}^n \mathbf{x}_i (y_i - T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}) \end{aligned}$$

- Information matrix:

$$\begin{aligned} J(\boldsymbol{\beta}) &= \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T v(\mu_i) \\ &= \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}} \end{aligned}$$

Poisson Model: Maximum Likelihood

- From first principles,

$$\begin{aligned} L_i(\boldsymbol{\beta}) &\propto \frac{\exp\{-T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}\} \left(T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)^{Y_i}}{Y_i!} \\ &= \exp\{-T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}\} \left(T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)^{Y_i} \end{aligned}$$

$$\ell_i(\boldsymbol{\beta}) = -T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}} + Y_i \log T_i + Y_i \mathbf{x}_i^T \boldsymbol{\beta}$$

$$U_i(\boldsymbol{\beta}) = \mathbf{x}_i(Y_i - T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}})$$

$$J_i(\boldsymbol{\beta}) = \mathbf{x}_i \mathbf{x}_i^T T_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}$$

Poisson Regression: SAS

- Like other GLMs, Poisson models can be fitted using PROC GENMOD
- e.g., Reconsider the data set from Slide #5, with input records of the form:

| IDNUM | YEARS | ADMITS | DIAB | AGE | LOG_YRS |
|-------|-------|--------|------|-----|---------|
| 1 | 0.6 | 5 | 1 | 61 | -0.5108 |
| 2 | 4.2 | 11 | 0 | 56 | 1.4351 |
| 3 | 1.3 | 7 | 1 | 45 | 0.2624 |
| 4 | 2.2 | 8 | 0 | 66 | 0.7885 |

```
PROC GENMOD DATA=dialysis;  
  MODEL admits = diab age / DIST=Poisson LINK=log  
                                OFFSET=log_yrs;  
RUN;
```

Example: Poisson Regression

- We analyze the coronary heart disease (CHD) data. The study observed $n = 3,154$ males ages 40-50. The study featured a prospective cohort design, with staggered entry and the observation period concluding on 12/31/70. Men were followed for an average of 8 years, and the number of CHD cases was recorded. Risk factors of interest included smoking, blood pressure and behavior type (A and B).
- (a) Read in and print out the analysis file.
- See the SAS code.
- (b) Fit a main effects model using Poisson regression, but do *not* use an offset. Comment on the validity of such an approach, making reference to the data set.

Y: CHD count

T_i : person year

Model

$$\begin{aligned} \log(\mu_i) &= \log(\lambda_i) = \beta_0 + \beta_1 I(type_A) + \beta_2 Bp \\ &+ \beta_3 I(cig = 2) + \beta_4 I(cig = 3) \\ &+ \beta_5 I(cig = 4) \end{aligned}$$

This model is not valid since T_i s (pearson year) are unequal.

- (c) Examine the parameter estimates based on the *no-offset* Poisson model (significance, direction).

Many regression coefficients are significant. It seems like that the directions of bp and smoke effects are not correct.

- (d) Under what circumstances would the *no-offset*

model be valid?

If $T_i = T$ for all i .

- (e) Fit another main effects model, this time using an offset. Test the significance of each term in the model using the Wald test.

Model (offset = $\log(T_i)$)

$$\begin{aligned}\log(\mu_i) - \log(T_i) &= \log(\lambda_i) = \beta_0 + \beta_1 I(\text{type}_A) + \beta_2 Bp \\ &+ \beta_3 I(\text{cig} = 2) + \beta_4 I(\text{cig} = 3) \\ &+ \beta_5 I(\text{cig} = 4)\end{aligned}$$

BP and smoke directions are corrected.

- (f) Interpret $\exp\{\hat{\beta}_0\}$.

$$\exp\{\hat{\beta}_0\} = \exp^{-5.47} = 0.0042$$

Estimated CHD rate per person year for a type B non-smoker with BP < 140.

Per person year is a small time unit, so the rate is very small.

- (g) Describe the estimated Type A effect, referring to the SAS output.

$$\exp\{\hat{\beta}_1\} = \exp^{0.76} = 2.14$$

CHD rate ratio between type A and type B, adjusting for other covariates.

- (h) Test the null hypothesis of no SMOKE effect, testing all categories simultaneously.

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

Test statistic: 24.32

P-value < 0.001

Reject H_0

- (i) Does the model fit the data well? Carry out GoF.

H_0 : the model fits data well

Test statistics: $D=19.28$

Since $D > 18.3 = \chi^2_{10,0.05}$, we can reject H_0 at $\alpha = 0.05$

The model does not fit the data well.

The $D/df > 1$, which indicates that there is an over dispersion problem.

- (g) Re-fit the model, with PY replaced by $PY/1000$. Compare the parameter estimates to those obtained previously.

New model:

$$\begin{aligned} \log(\mu_i) - \log(T_i/1000) &= \log(\lambda_i^*) = \beta_0^* + \beta_1^* I(\text{type}_A) \\ &+ \beta_2^* Bp + \beta_3^* I(\text{cig} = 2) \\ &+ \beta_4^* I(\text{cig} = 3) + \beta_5^* I(\text{cig} = 4) \end{aligned}$$

- Since $\beta_0^* + \log(T_i) - \log(1000) = \beta_0 + \log(T_i)$,
 $\beta_0^* = \beta_0 + \log(1000)$. Therefore,

$$\widehat{\beta}_0^* = \widehat{\beta}_0 + 6.91 = -5.47 + 6.91 = 1.44$$

$$\exp^{1.44} = 4.22$$

- $\beta_i^* = \beta_i$ for all $i = 1, \dots, 5$

BIOSTAT 651
Notes #14: Overdispersion (revised)

- Lecture Topics:
 - Causes of overdispersion
 - Estimating scale parameter
 - Random effect models
 - Generalized estimating equations

Overdispersion

- *Overdispersion*: variance exceeds that under the assumed model

- e.g., Binomial data

$$\text{Var}(Y_i) > v(\mu) = n\mu(1 - \mu)$$

- e.g., Poisson response

$$\text{Var}(Y_i) > v(\mu) = \mu$$

- Under-dispersion can also occur
 - less common

Overdispersion: Causes

- Overdispersion can result for several reasons
 - heterogeneous populations
 - unmeasured covariate
 - events *within cell* are correlated

Overdispersion: Causes

- Example: Population Heterogeneity
 - Suppose there exists a binary covariate, Z_i , and that

$$Y_i|Z_i = 0 \quad \sim \quad \text{Poisson}(\lambda_0)$$

$$Y_i|Z_i = 1 \quad \sim \quad \text{Poisson}(\lambda_1)$$

$$P(Z_i = 1) \quad = \quad \pi$$

$$E(Y_i) \quad = \quad \pi\lambda_1 + (1 - \pi)\lambda_0 = \mu$$

$$\begin{aligned} \text{Var}(Y_i) &= E(\lambda_1 Z_i + \lambda_0(1 - Z_i)) \\ &+ \text{Var}(\lambda_1 Z_i + \lambda_0(1 - Z_i)) \\ &= \mu + (\lambda_1 - \lambda_0)^2 \pi(1 - \pi) \end{aligned}$$

Overdispersion: Impact on Analysis

- As implied, overdispersion generally involves the assumed variance structure being inconsistent with that actually underlying the data
 - in GLMs, $\hat{\beta}$ is generally unbiased
 - however, $\widehat{SE}(\hat{\beta}_j)$ may be substantially biased
- As a result:
 - hypothesis tests tend to be anti-conservative
 - CIs tend to be artificially narrow
- For under-dispersion, effect is in the opposite direction

Overdispersion: Case of Poisson Model

- Any GLM is subject to misspecification
- Poisson model is especially susceptible

Accommodating Overdispersion

- Several methods have been developed for handling overdispersion
 - estimating scale parameter (quasi-likelihood)
 - random effects models
 - generalized estimating equations
- Each method involves modifying the original
 - (i) model assumptions
 - (ii) estimation procedures

Quasi-likelihood

- By our previously derived Exponential family and GLM results:

$$V(Y_i) = a(\phi) v(\mu_i)$$

- Suppose we relax the assumption that $a(\phi) = 1$
 - e.g., common to assume that $a(\phi) = \phi$,
 $E[Y_i] = \mu_i$
 $V(Y_i) = \phi v(\mu_i)$
 - Note: Impact on score and information:

$$U(\boldsymbol{\beta}) = \frac{1}{\phi} X^T (Y - \mu)$$

$$J(\boldsymbol{\beta}) = \frac{1}{\phi} X^T V X$$

Estimating Scale Parameter

- As implied, $\hat{\beta}$ can still be computed by solving

$$\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0}$$

- We now need a method for estimating ϕ
- Pearson Chi-square statistic:

$$X_P^2(\phi) = \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\hat{V}(Y_i)}$$

- under some conditions

$$X_P^2(\phi) \sim \chi_{n-q}^2$$

Estimating Scale Parameter (continued)

- Then, using the fact that

$$V(Y_i) = a(\phi) v(\mu_i)$$

and the assumption that

$$a(\phi) = \phi$$

along with MoM concepts suggests setting

$$X_P^2(\phi) = \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\phi v(\hat{\mu}_i)} \approx (n - q)$$

which implies the Pearson-based scale estimator:

$$\hat{\phi}_P = \frac{X_P^2(1)}{n - q}$$

Estimating Scale Parameter: Examples

- e.g., $Y_i \sim \text{Poisson}(\mu_i)$:

$$\hat{\phi}_P =$$

- e.g., $Y_i \sim \text{Binomial}(n_i, \pi_i)$

$$\hat{\phi}_P =$$

- Note: can also use an estimator based on the Deviance:

$$\hat{\phi}_D = \frac{D}{n - q}$$

- often gives similar results

Impact of Scale Parameter on Inference

- Estimation of β is unaffected

Q: Why?

- Standard errors are modified:

uncorrected: $\widehat{SE}(\widehat{\beta}_j)$

corrected:

Dispersion Parameter: SAS Code

- Easy to estimate ϕ in SAS

e.g., Poisson regression:

```
PROC GENMOD DATA=dialysis;  
  MODEL admits = diab age / DIST=Poisson LINK=log  
                                OFFSET=log_yrs  
                                PSCALE;  
  
RUN;
```

- For $\hat{\phi}_D$ estimator, use DSCALE option

Random effect model

- To use a hierarchical random effect model for over dispersion data
 - Binomial data: Beta-binomial regression
 - Count data: Negative-binomial regression

Negative binomial regression

- Introduce a random effect term to model extra variation.

$$\begin{aligned} Y_i | \theta_i &\sim \text{Poisson}(\theta_i) \\ \theta_i &= \exp(X_i \beta + \epsilon_i) \\ &= \exp(X_i \beta) \exp(\epsilon_i) = \mu_i z_i \end{aligned} \quad (1)$$

- z_i follows gamma distribution
 $\Gamma(\text{shape} = \delta, \text{rate} = \delta)$, so $E(z_i) = 1$

Negative binomial regression

$$P(Y_i = y | \mu_i, \delta) = \frac{\Gamma(\delta + y)}{\Gamma(\delta)\Gamma(y + 1)} \left(\frac{\delta}{\delta + \mu_i} \right)^\delta \left(\frac{\mu_i}{\delta + \mu_i} \right)^y$$

- Mean and variance:

$$\begin{aligned} E(Y_i) &= \mu_i = \exp(X_i\beta) \\ \text{Var}(Y_i) &= \mu_i + \frac{1}{\delta} \mu_i^2 \end{aligned}$$

- Additional parameter δ in the variance.

Negative binomial regression: SAS code

- Use dist=negbin

```
PROC GENMOD DATA=dialysis;  
  MODEL admits = diab age / DIST=negbin  
                                     OFFSET=log_yrs;  
RUN;
```

Generalized Estimating Equations

- Consider the score function for a canonical GLM,

$$U(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu})$$

obtained through standard ML theory

- Suppose now that the model assumptions are in question
 - e.g., quite possible that $E[Y_i] = \mu_i$, but that other properties connected with the GLM may not hold
 - e.g., Poisson case: variance structure

GEE: Poisson Case

- Suppose that Y_i is an event count
 - seems natural to assume $Y_i \sim \text{Poisson}(\mu_i)$
 - and, under a (canonical) GLM:
$$E[Y_i] = T_i \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}$$
- The effect of covariates on $E[Y_i] = \mu_i$ is of chief interest
- However, $V(Y_i)$ is likely of little inherent interest
 - assumptions regarding $V(Y_i)$ are best avoided
 - In GLM we assume $V(Y_i) = \mu_i$
 - need to do a valid inference when $V(Y_i) \neq \mu_i$

GEE: Intro

- Recall: Method-of-Moments estimation
 - equate sample and population moments
 - solve for parameters of interest

- GEE (Liang & Zeger, 1986): Solve

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i) = 0$$

- $D_i = \partial \mu_i / \partial \boldsymbol{\beta}_i^T$
- V_i : assumed variance of Y_i

- GEE can be viewed as a regression analog of MoM
 - i.e., provided that the model for the mean structure is correct,
 $S(\boldsymbol{\beta})$ is a zero-mean estimating equation

- In univariate data, $S(\beta)$ is the same as the score function $U(\beta)$ when V_i is an assumed variance in GLM

- $D_i = 1/g'(\mu_i)\mathbf{x}_i^T$
- $V_i = a(\phi)v(\mu_i)$

$$\begin{aligned} S(\beta) &= \sum_{i=1}^n \frac{Y_i - \mu_i}{a(\phi)v(\mu_i)g'(\mu_i)} \mathbf{x}_i \\ &= U(\beta) \end{aligned} \tag{2}$$

- With canonical link function ($v(\mu_i) = 1/g'(\mu_i)$)

$$\begin{aligned} S(\beta) &= \frac{1}{a(\phi)} \sum_{i=1}^n (Y_i - \mu_i) \mathbf{x}_i \\ &= \frac{1}{a(\phi)} \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu}) \end{aligned} \tag{3}$$

GEE: Applicability

- GEE is usually thought of as a method for correlated responses
 - MLE requires a fully-specified model
 - not so easy in some cases ...
- GEE only requires specification of the *mean* and *covariance* terms
 - consistency of $\hat{\beta}$ requires that the mean be modeled correctly
 - does *not* require that covariance be correctly specified
- More generally, GEE can be used with univariate data, in order to avoid pitfalls of model misspecification

GEE: Properties

- Key GEE result: a zero-mean estimating equation should yield a consistent estimator of β
 - requires that the assumptions that lead to the zero-mean property hold
 - note: *not* required that the estimating equation be based on ML
- GEE $S(\beta)$ is a zero-mean estimating equation

$$S(\beta) = \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i) = 0$$

- $E(S(\beta_0)) = 0$ has mean zero provided that $E[Y_i | \mathbf{x}_i] = \mu_i$
- This mean zero property does not depend on V_i
- If $V_i = a(\phi)v(\mu_i)$ in GLM, $S(\beta)$ is the same as the score function from the log-likelihood

- In GEE, β is estimated by the solution to $S(\beta) = \mathbf{0}$ without framing S as the derivative of the log-likelihood function
 - note: standard ML-based inference procedures do not apply to GEE estimators

GEE: Inference

- Provided that $E[S(\beta_0)] = \mathbf{0}$ and under some (mild) regularity conditions:
 - $\hat{\beta} \rightarrow \beta_0$
 - $V(\hat{\beta}) \approx H(\beta_0)$

$$H(\beta_0) = H_1(\beta_0)^{-1} H_2(\beta_0) H_1(\beta_0)^{-1}$$

$$H_1(\beta_0) = \sum_{i=1}^n D_i^T V_i^{-1} D_i$$

$$H_2(\beta_0) = \sum_{i=1}^n D_i^T V_i^{-1} \text{Var}(Y_i) V_i^{-1} D_i$$

- When V_i is correctly specified: $V_i = \text{Var}(Y_i)$
 - $H_1(\beta_0)^{-1} H_2(\beta_0) H_1(\beta_0)^{-1} = H_1(\beta_0)$
- Canonical link with $V_i = \text{Var}(Y_i)$:
 - $H_1(\beta_0) = X^T V X / a(\phi) = J(\beta_0)$
- If V_i is misspecified, the variance would be larger, so the efficiency decreases.

GEE: Inference (continued)

- When estimated through GEE, $V(\hat{\boldsymbol{\beta}})$ can be estimated by the *robust* variance estimator,

$$\hat{V}(\hat{\boldsymbol{\beta}}) = H_1(\hat{\boldsymbol{\beta}})^{-1} \hat{H}_2(\hat{\boldsymbol{\beta}}) H_1(\hat{\boldsymbol{\beta}})^{-1}$$

$$\hat{H}_2(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n D_i^T V_i^{-1} (y_i - \hat{\mu}_i) (y_i - \hat{\mu}_i)^T V_i^{-1} D_i$$

- also known as the *sandwich* estimator

Generalized Estimating Equations: SAS Code

- *Working independence* assumption:

e.g., Poisson regression GEE:

```
PROC GENMOD DATA=dialysis;  
  CLASS idnum;  
  MODEL admits = diab age / DIST=Poisson LINK=log  
                                OFFSET=log_yrs;  
  REPEATED SUBJECT=idnum / TYPE=ind;  
RUN;
```


GEE: Efficiency

- Increase efficiency by correctly specifying the variance
- For longitudinal data (correlated outcomes), typically

$$\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$$

- several options for specifying \mathbf{R}_i
- correct \mathbf{R}_i increase the efficiency
- Although you misspecified \mathbf{R}_i , you still can do a valid inference.

Example: Overdispersion

We return to the clinical trial analyzed in the first part of the course; to recap ...

A clinical trial was conducted in order to evaluate the impact of Progabide on the frequency of epileptic seizures. Patients were randomized to either receive or not receive Progabide. The data set contains information on:

- age at start of study (AGE; measured in years)
- baseline seizure count; defined as the number of seizures in the 8 weeks prior to the study's commencement (BASE)
- treatment indicator (Z; 1=treated, 0=placebo)
- seizure counts in each of 4 two-week periods (Y1, Y2, Y3, Y4)

The investigators define the outcome as total post treatment seizure count: $Y_i \equiv \sum_{j=1}^4 Y_{ij}$.

- (a) Read in the data file and calculate correlations among $Y1, Y2, Y3, Y4$
- See SAS Code
 - There are very strong positive correlations among $(Y1, Y2, Y3, Y4)$. For example, $corr(Y1, Y2) = 0.87$.
- (b) When $Y_i \equiv \sum_{j=1}^4 Y_{ij}$ is used for the outcome in poisson regression, do you think there will be an overdispersion problem?
- In this data, $(Y1, Y2, Y3, Y4)$ are not independent. Dependency among them can induce overdispersion.
 - Suppose $Y_{ij} \sim Poisson(\lambda_{ij})$, then

$$E(Y_i) = E\left(\sum_{j=1}^4 Y_{ij}\right) = \sum_{j=1}^4 \lambda_{ij}$$

$$Var(Y_i) = Var\left(\sum_{j=1}^4 Y_{ij}\right)$$

$$= \sum_{j=1}^4 \lambda_{ij} + 2 \sum_{j < k} Cov(Y_{ij}, Y_{ik})$$

When $Cov(Y_{ij}, Y_{ik}) > 0$ for all (j, k) ,
 $Var(Y_i) > E(Y_i)$.

- (c) Estimate the covariate-adjusted treatment effect through a Poisson regression model. Do you have an evidence of overdispersion?

- Model

$$\begin{aligned} \log(\lambda_i) &= \beta_0 + \beta_1 Age + \beta_2 Base + \beta_3 Z \\ Y_i &\sim Poisson(\lambda_i) \end{aligned}$$

- Since $Deviance/DF = 10.18 \gg 1$, we can conclude that there exists an overdispersion problem.

- (d) Re-estimate (and re-test) the treatment effect by estimating the scale parameter (quasi-likelihood)

- Number of parameters (including intercept): 4

$$\hat{\phi} = \frac{D}{59-4} = 10.18$$

$$\sqrt{\hat{\phi}} = \sqrt{10.179} = 3.19$$

$\hat{\beta}_3$ is not changed, but the standard error estimate of $\hat{\beta}_3$ is changed.

- SE of β_3 :

$$\widehat{SE}(\hat{\beta}_3) = 0.0465 * 3.19 = 0.148$$

(e) Carry out a Wald test for the age and treatment effects using quasi-likelihood.

- $H_0: \beta_1 = \beta_3 = 0$ vs $H_1: \beta_1 \neq 0$ or $\beta_3 \neq 0$
- Contrast matrix

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Test statistic:

$$X_w^2 = \{C\hat{\beta}\}^T \{C\hat{\phi}I(\hat{\beta})^{-1}C^T\}^{-1} \{C\hat{\beta}\},$$

- In this data, $X_w^2 = 5.34$ and the corresponding

p-value=0.069. We cannot reject the null hypothesis.

- (f) Carry out a likelihood ratio test for the age and treatment effects by fitting full and reduce models. Make a comment on the validity of the test result.

- Full model:

$$\log(\lambda_i) = \beta_0 + \beta_1 Age + \beta_2 Base + \beta_3 Z$$

- Reduced Model:

$$\log(\lambda_i) = \beta_0 + \beta_2 Base$$

From SAS output $l_{full} = 555.85$ and $l_{reduced} = 522.99$.

$$X^2_l = 2 * (555.85 - 522.99) = 65.72$$

- LRT test statistic is too large. Does not seem to valid.

- (g) Carry out the likelihood ratio test, in this time, using the same scale parameter (estimated from the full model) for both full and reduced models.

With fixing $\text{scale}=3.1905$, $|_{\text{reduced}} = 553.19$.

$$X^2_l = 2 * (555.85 - 553.19) = 5.32$$

- Similar to the Wald test statistics

- (h) In this time, carry out both Wald and LRT using the contrast statement.

See the SAS code.

- (e) Re-estimate (and re-test) the treatment effect using the negative binomial regression.

- $\hat{\beta}_3 = -0.2112$
- $p - \text{value} : 0.1681$

- (g) Carry out LRT for the age and treatment effects by fitting the full and reduce models.

- $H_0: \beta_1 = \beta_3 = 0$ vs $H_1: \beta_1 \neq 0$ or $\beta_3 \neq 0$

- From SAS output $l_{full} = 5846.28$ and $l_{reduced} = 5844.59$.

$$X_l^2 = 2*(5846.28 - 5844.59) = 3.38 < 5.99 = \chi_{2,0.05}^2$$

- Cannot reject H_0 at $\alpha = 0.05$

(f) Re-estimate (and re-test) the treatment effect using GEE.

- For GEE fit we use the same variance structure in poisson GLM.

$$Var(Y_i) = V_i = a(\phi)v(\mu_i) = \mu_i$$

Since $D_i = \partial\mu_i/\partial\beta_i^T = 1/g'(\mu_i)X_i^T = v(\mu_i)X_i^T$,
GEE equation is

$$S(\beta) = \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i) = \sum_{i=1}^n X_i (Y_i - \mu_i)$$

- $\hat{\beta}$ from GEE is the same as MLE

- Sandwich estimator of variance:

$$Var(\hat{\beta}) = H_1(\hat{\beta})^{-1} H_2(\hat{\beta}) H_1(\hat{\beta})^{-1}$$

$$\begin{aligned}
H_1(\widehat{\beta}) &= \sum_{i=1}^n D_i^T V_i^{-1} D_i = \sum_{i=1}^n X_i X_i^T \widehat{\mu}_i \\
H_2(\widehat{\beta}) &= \sum_{i=1}^n D_i^T V_i^{-1} \text{Var}(Y_i) V_i^{-1} D_i \\
&= \sum_{i=1}^n X_i X_i^T (Y_i - \widehat{\mu}_i)^2
\end{aligned}$$

BIOSTAT 651

Notes #15: Bootstrap Methods

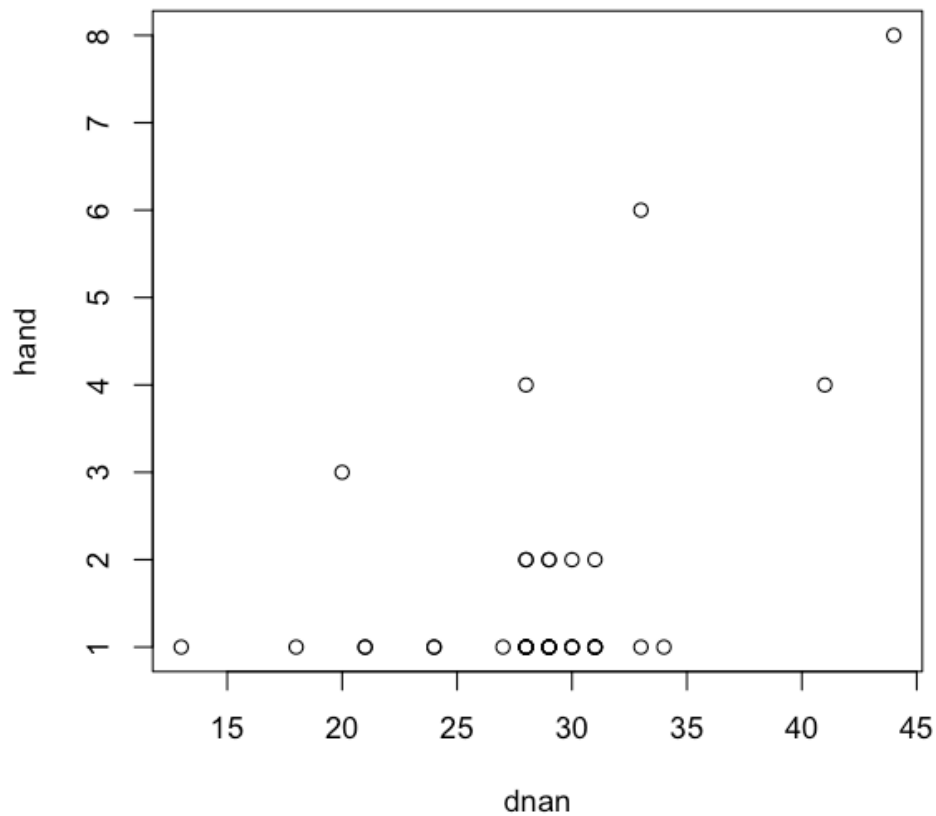
- Lecture Topics:
 - Basic idea
 - Hypothesis test
 - Regression
- Based on following book and slides.
 - Davison and Hinkley (1997) *Bootstrap Methods and their Application*. Cambridge University Press
 - Davison (2006) *Bootstrap Methods and their Application*. Short course slides

Bootstrap

- Bootstrap: simulation methods to estimate sampling distribution of almost any statistics.
- Developed by Bradley Efron in "Bootstrap methods: another look at the jackknife" (1979)
- Useful when
 - Standard assumptions are not valid (small n , invalid regression assumptions, etc)
 - Complex problem with no (reliable) theory ex. theoretical distribution of a statistic of interest is complicated or unknown
 - or (almost) anywhere else.

Example: Handedness data

- Investigate relationship between genetic measurement (dnan) and left-handedness (hand).



Example: Handedness data

- Question: Is there any dependence between dnan and hand for these $n = 37$ individuals?
- Sample coefficient $\hat{\theta} = 0.509$
- Confidence interval (from the bivariate normal model): $CI(0.221, 0.715)$
- Issues?

Bootstrap

- Estimate distribution of $\hat{\theta}$ using resampling
- Statistical model: data $(y_1, \dots, y_n) \sim F$, unknown
 - Handedness data: $y = (\text{dnan}, \text{hand})$
 - θ : correlation coefficient
- For $r = 1, \dots, R$
 - resample y with replacement:
 $y_r^* = (y_{1r}^*, \dots, y_{nr}^*)$
 - Compute bootstrap $\hat{\theta}_r^*$ using y_r^*
- Repeat R times !

$$\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_R^*$$

Bootstrap

- R code (boot package)

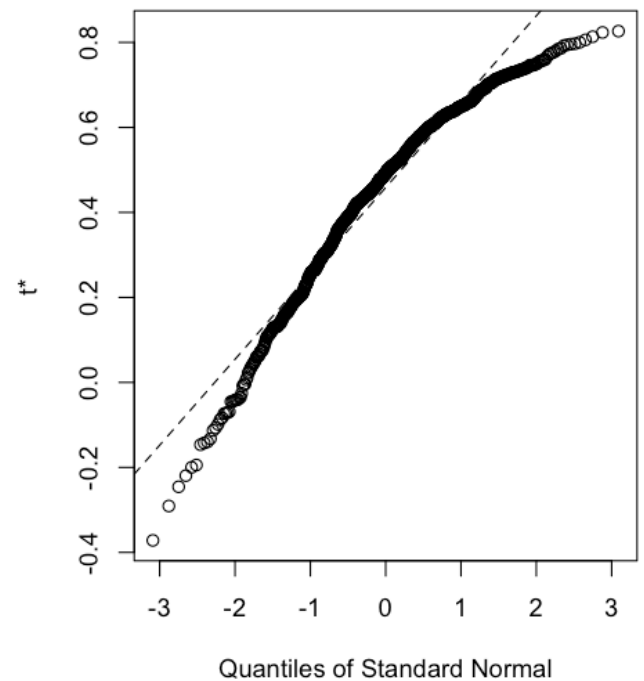
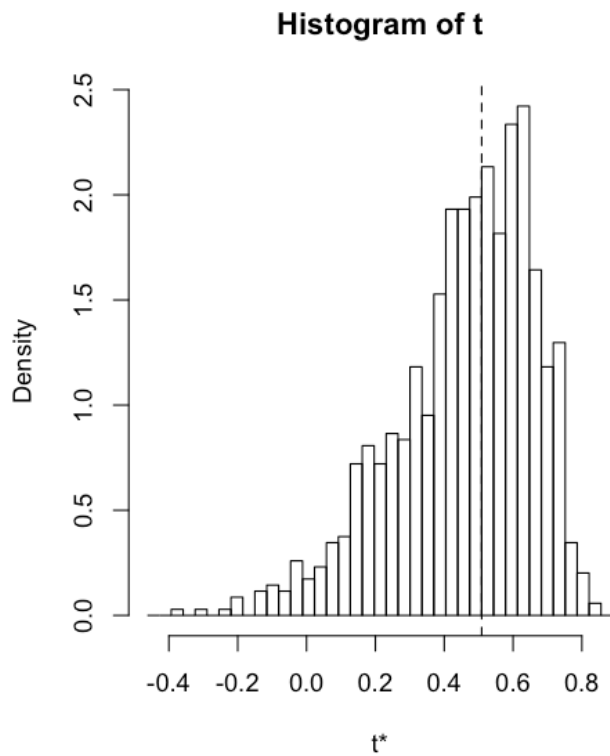
```
library(boot)
data(claridge)
```

```
Corr <- function(d, f){
  cor(d[f,1], d[f,2])
}
boot.out<-boot(claridge, Corr, R=1000)
```

```
plot(boot.out)
boot.ci(boot.out)
```

Bootstrap

- Distribution of $t = \hat{\theta}_r^*$ (R=1000)
- CI: (0.0912, 0.8071) (BCa)



Why does Bootstrap work

- Statistical model: data $(y_1, \dots, y_n) \sim F$
- Estimate distribution of $\hat{\theta}$
 - Key issue: what is the variability of $\hat{\theta}$ when samples are repeatedly taken from F ?
- Suppose F is known - we can answer the previous question by
 - Analytical calculation
 - Simulation

Why does Bootstrap work

- Assume F is known
- For $r = 1, \dots, R$
 - generate random sample
$$y_r^* = (y_{1r}^*, \dots, y_{nr}^*) \sim F$$
 - Compute $\hat{\theta}_r^*$ using y_r^*
- Use $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_R^*$ to estimate sampling distribution of $\hat{\theta}$
- If $R \rightarrow \infty$, Monte Carlo error would disappear.

Why does Bootstrap work

- But we don't know F !
 - Estimate F using the empirical distribution function \hat{F}_n
 - Generate random samples from \hat{F}_n
- For $r = 1, \dots, R$
 - generate random sample
 $y_r^* = (y_{1r}^*, \dots, y_{nr}^*) \sim \hat{F}_n$
 - Compute $\hat{\theta}_r^*$ using y_r^*
- Bootstrap (re)samples are iid samples from \hat{F}_n

Bootstrap estimators

- Variance

$$Var_B(\hat{\theta}) = \frac{1}{R-1} \sum_{i=1}^R (\hat{\theta}_r^* - \hat{\theta}_{(.)}^*)^2$$

$$\hat{\theta}_{(.)}^* = \frac{1}{R} \sum_{i=1}^R \hat{\theta}_r^*$$

- Bias

- Bias: $E(\hat{\theta}) - \theta$
- Bootstrap estimator of Bias:

$$Bias_B = \frac{1}{R} \sum_{i=1}^R \hat{\theta}_r^* - \hat{\theta}$$

- Bias corrected estimator

$$\hat{\theta}_{BC} = \hat{\theta} - Bias_B$$

Bootstrap Confidence Intervals

- There are several versions of CI
- Normal confidence interval
 - If $\hat{\theta}$ approximately normal, then $\hat{\theta} \sim N(\theta + \beta, v)$, where β is a bias
 - With known β and v , $(1 - 2\alpha)$ CI is

$$\theta - \beta \pm Z_{\alpha} v^{1/2}$$

- Replace β and v to their Bootstrap estimates.
- Percentile interval
 - Estimate CI nonparametrically
 - Use α and $1 - \alpha$ quantiles of bootstrap samples to estimate CI

$$\hat{\theta}_{(R+1)\alpha}^*, \quad \hat{\theta}_{(R+1)(1-\alpha)}^*$$

- Studentized-t (Percentile-t) Bootstrap Confidence Interval
 - Generalize Student-t statistic to bootstrap setting
 - Require variance formula V for $\hat{\theta}$ computed from (y_1, \dots, y_n)
 - R bootstrap copies of $(\hat{\theta}_r^*, \hat{V}_r^{*1/2})$

$$T_1^* = \frac{(\hat{\theta}_1^* - \hat{\theta})}{\hat{V}_1^{*1/2}}, \dots, T_r^* = \frac{(\hat{\theta}_r^* - \hat{\theta})}{\hat{V}_r^{*1/2}}$$

- CI

$$\hat{\theta} - \hat{V}^{1/2} T_{(R+1)\alpha}^*, \quad \hat{\theta} - \hat{V}^{1/2} T_{(R+1)(1-\alpha)}^*$$

- Bias corrected, accelerated (BCa) percentile interval
 - Shift and scale the percentile bootstrap confidence interval to compensate for bias
 - Replace percentile interval with

$$\hat{\theta}_{(R+1)\alpha_1}^*, \quad \hat{\theta}_{(R+1)(1-\alpha_2)}^*$$

where α_1 and α_2 were chosen to improve CI.

Handedness data

- Bias: -0.0401
- SE: 0.208
- CI:
 - Normal $(0.1631, 0.9551)$
 - Percentile $(-0.0402, 0.7465)$
 - BCa $(0.0912, 0.8071)$

Hypothesis test

- Testing problem
 - data (y_1, \dots, y_n)
 - Model M_0 to be tested
 - test statistic $T_{obs} = T(y_1, \dots, y_n)$, with large values giving evidence against H_0
- P-value: $Pr(T \geq T_{obs} | M_0)$
 - Small p-value indicates evidence against M_0
- Issue: P-values are often hard to calculate

Hypothesis test

- Estimate P-values by simulating from the fitted null hypothesis model \hat{M}_0
- For $r = 1, \dots, R$
 - generate random sample
 $y_r^* = (y_{1r}^*, \dots, y_{nr}^*) \sim \hat{M}_0$
 - Compute T_r^* from y_r^*
- P-value estimate

$$\hat{p} = \frac{1 + \#[T_r^* \geq T_{obs}]}{1 + R}$$

Hypothesis test

- Handedness data: are *dnan* and *hand* positively associated?
- Observed Correlation: $\hat{\theta} = 0.509$

$$T_{obs} = 0.509^2 = 0.259$$

- Null hypothesis: independence

$$F(dnan, hand) = F_1(dnan)F_2(hand)$$

- For $r = 1, \dots, R$
 - Simulate bootstrap samples independently from $\hat{F}_1(dnan_1, \dots, dnan_n)$ and $\hat{F}_2(hand_1, \dots, hand_n)$, then put them together $(dnan_1^*, hand_1^*), \dots, (dnan_n^*, hand_n^*)$
 - Calculate $T_r^* = \hat{\theta}_r^{*2}$

- P-value estimate (R=10000)

$$\hat{p} = \frac{1 + \#[T_r^* \geq T_{obs}]}{1 + R}$$

Hypothesis test

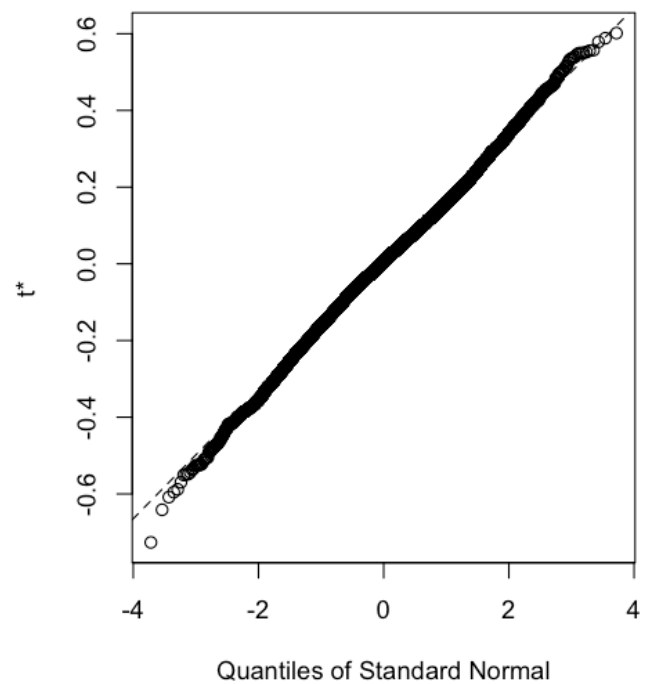
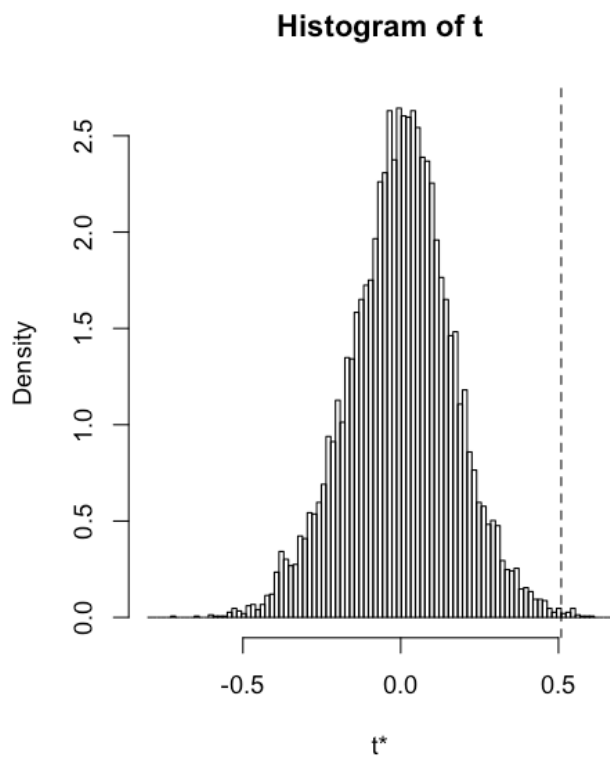
- R code (boot package)

```
set.seed(100)
# Bootstrap p-values
R<-10000
New.D<-c(claridge$dnan, claridge$hand)
Corr1 <- function(d, f, n){
  x<-d[f]
  cor(x[1:n], d[(n+1):(2*n)])
}
boot.out1<-boot(New.D, Corr1, R=R
, strata=rep(c(1,2), c(n,n)), n=n)

n1<-sum(boot.out1$t^2 >= boot.out1$t0^2)
Pval.boot = (n1+1)/(R+1)
```

Hypothesis test

- Bootstrap p-value: 0.0041
- Histogram under the null (bootstrap)



Hypothesis test

- Alternatively confidence interval can be used for hypothesis test
- Handedness data:
 - 95% CI does not include 0
 - Can reject H_0 at $\alpha = 0.05$

Hypothesis test

- Instead of using Bootstrap, Permutation can be used for the hypothesis test

- For $r = 1, \dots, R$
 - Take samples from

$$(dnan_1, hand_{1^*}), \dots, (dnan_n, hand_{n^*})$$

where $(1^*, \dots, n^*)$ is random permutation of $(1, \dots, n)$

- Calculate $T_r^* = \hat{\theta}_r^{*2}$
- Handedness data, permutation p-value=0.0049
 - Nearly identical to bootstrap p-value=0.0041

Linear Regression

- Independent data $(x_1, y_1), \dots, (x_n, y_n)$ with

$$y_i = x_i^T \beta + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2)$$

- Studentized residuals

$$e_i = \frac{y_i - x_i^T \hat{\beta}}{(1 - h_i)^{1/2}} \sim (0, \sigma^2)$$

- Two main resampling schemes
- Model based resampling (or residual resampling)

$$y_i^* = x_i^T \beta + \epsilon_i^*, \quad \epsilon_i^* \sim EDF(e_1 - \bar{e}, \dots, e_n - \bar{e})$$

- Fixed design X , but not robust to model failure

- Case resampling

$$(x_i, y_i)^* \sim EDF[(x_1, y_1), \dots, (x_n, y_n)]$$

- Varying design X , but robust
- Assume (x_i, y_i) sampled from population

GLM

- Case resampling can be used for GLM.
- There exist approximation methods for residual resampling.

GLM: Seizure count

- Seizure count data (Overdispersion!)

```
> dat<-read.table("./seizure1.txt", header=FALSE)
> colnames(dat)<-c("Y1","Y2","Y3","Y4", "Z","base", "age")
> dat$Y<-dat$Y1+dat$Y2+dat$Y3+dat$Y4
>
> out<-glm(Y ~ age+base+Z, data=dat, family=poisson)
> summary(out)
```

Call:

```
glm(formula = Y ~ age + base + Z, family = poisson, data = da
```

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|---------|--------|---------|
| -5.8949 | -2.0883 | -0.9471 | 0.7746 | 11.0049 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) | |
|-------------|-----------|------------|---------|----------|-----|
| (Intercept) | 2.072832 | 0.115817 | 17.897 | < 2e-16 | *** |
| age | 0.018678 | 0.003336 | 5.599 | 2.15e-08 | *** |
| base | 0.022615 | 0.000510 | 44.346 | < 2e-16 | *** |
| Z | -0.184221 | 0.046487 | -3.963 | 7.41e-05 | *** |
| --- | | | | | |

GLM: Seizure count

- Bootstrap $R = 1000$

```
> # case resampling
> Seizure <- function(d, f){
+ d1 = d[f,]
+ out<-glm(Y ~ age+base+Z, data=d1, family=poisson)
+ out1<-summary(out)$coefficients[4,1]
+ return(out1)
+ }
> boot.out<-boot(dat, Seizure, R=1000)
> boot.out
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

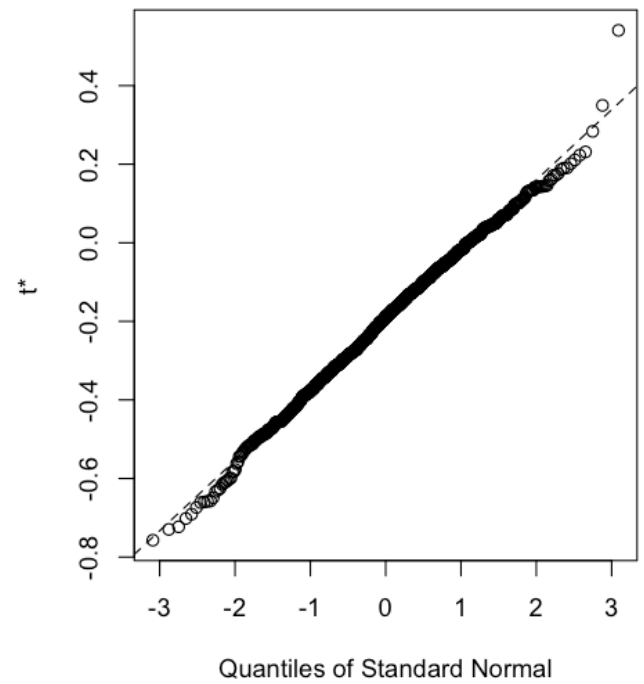
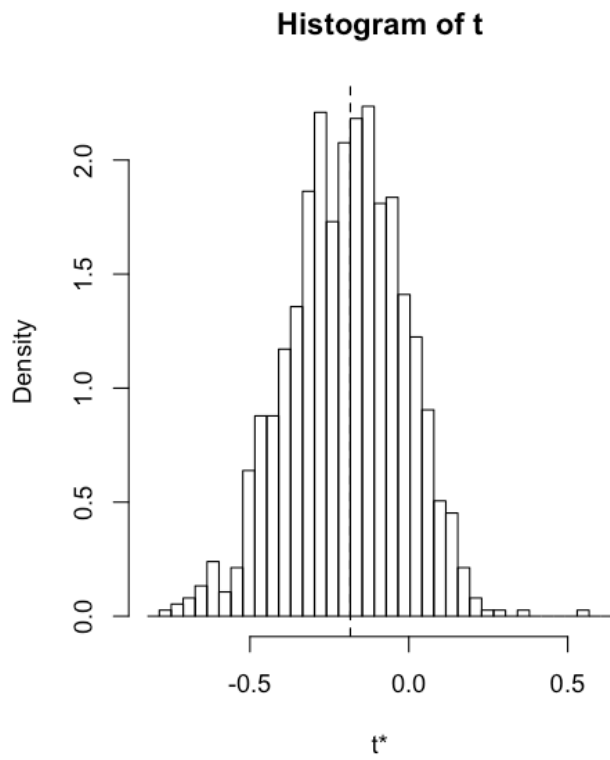
```
boot(data = dat, statistic = Seizure, R = 1000)
```

Bootstrap Statistics :

| | original | bias | std. error |
|-----|------------|-------------|------------|
| t1* | -0.1842214 | -0.01292668 | 0.1786147 |

GLM: Seizure count

- CI: $(-0.5437, 0.1412)$ (BCA)



SAS example (jackboot macro)

```
data seizure1;
    infile "~/BIOSTAT651/seizure1.txt";
input Y1 Y2 Y3 Y4 Z base age;
Y_tot=y1+y2+y3+y4;
idnum=_N_;
run;

%inc "~/BIOSTAT651/jackboot.sas";

%macro analyze(data=,out=);
    options nonotes;
proc genmod data=&data;
    model Y_tot = age base Z / dist=Poisson link=log;
    ods output ParameterEstimates=&out(drop=DF StdErr
LowerWaldCl UpperWaldCL ChiSq ProbChiSq);
    %bystmt;
run;

    options notes;
%mend;
ODS SELECT NONE;
%boot(data=seizure1,samples=1000, id=Parameter, random=123);
%bootci(bca,alpha=.05, id=Parameter)
```

```
ODS SELECT ALL;
```

```
proc print data=BOOTSTAT;  
run;
```

```
proc print data=BOOTCI;  
run;
```

```
/* Get Bootstrap dist for Z */  
data BOOTDIST1;  
set BOOTDIST;  
if Parameter ne "Z" then delete;  
run;
```

```
proc UNIVARIATE data=BOOTDIST1;  
var Estimate;  
histogram;  
run;
```

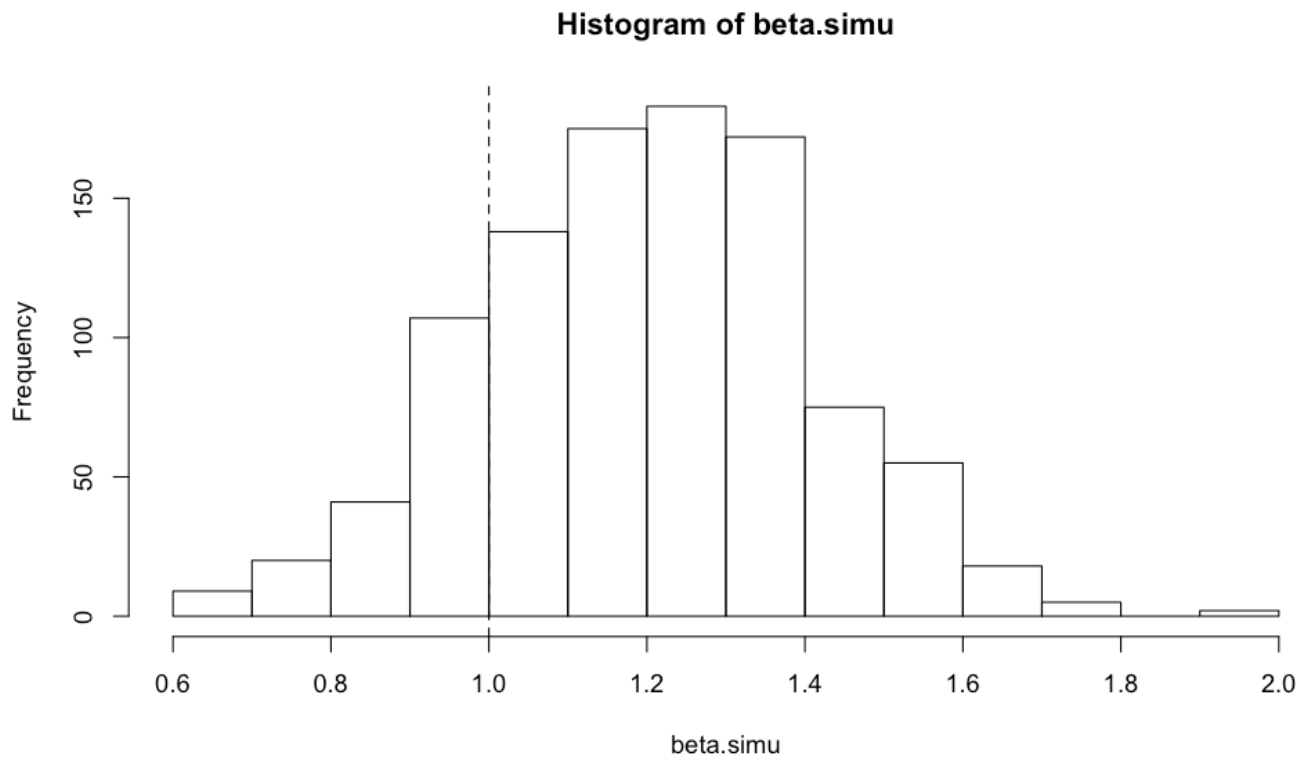
GLM: too many strata

- 100 strata, each has 6 samples
- In each stratum, 3 individuals received treatment ($X_{ki} = 1$) and 3 received placebo ($X_{ki} = 0$).
- Logistic regression model:

$$\text{logit}(\pi_{ki}) = \alpha_k + \beta X_{ki}, \quad (k = 1, \dots, 100; i = 1, \dots, 6)$$

GLM: too many strata

- The true $\beta = 1$
- Generate data 1000 times and get the distribution of $\hat{\beta}$
- Mean $\hat{\beta} = 1.2 \rightarrow \text{Bias} = 0.2$



GLM: too many strata

- Carry out 1000 bootstrap to estimate the bias

```
> # case resampling
> StrataDat <- function(d, f){
+ d1 = d[f,]
+ out<-glm(Y ~ X + factor(Strata) -1, data=d1
+ , family=binomial)
+ out1<-summary(out)$coefficients[1,1]
+ return(out1)
+ }
> boot.out<-boot(dat, StrataDat, R=1000)
> boot.out
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

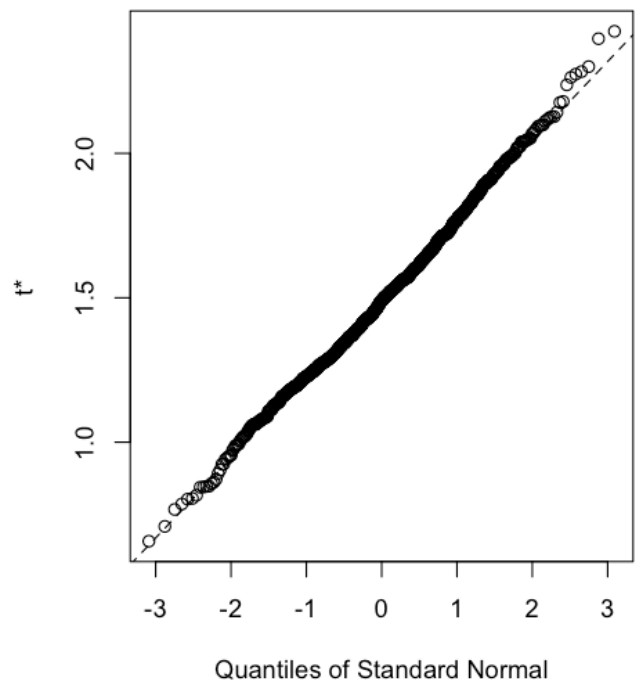
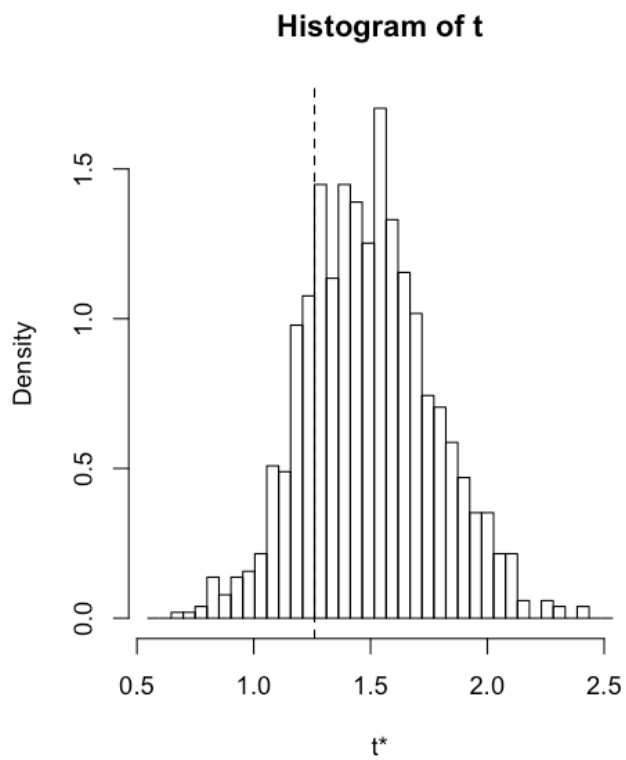
```
boot(data = dat, statistic = StrataDat, R = 1000)
```

Bootstrap Statistics :

| | original | bias | std. error |
|-----|----------|-----------|------------|
| t1* | 1.260071 | 0.2332586 | 0.2748678 |

GLM: too many strata

- Bootstrap bias estimate = 0.233



Bootstrap

- Bootstrap is a very useful tool to estimate sampling distribution of statistics
- In regression model, you can use either case-resampling or residual-resampling
 - In GLM, residual-sampling can be done (using approximation), but case-sampling is more widely used.
- There exist several R packages (ex. boot package)
- In SAS, you can use jackboot macro.