Gene-Environment Case-Control Studies

Raymond J. Carroll
Department of Statistics
Faculties of Nutrition and Toxicology

Texas A&M University
http://stat.tamu.edu/~carroll

STATISTICS
TEXAS A&M UNIVERSITY

Lecture 1, Part #1: Outline

- What is a case-control study?
- First example of gene-environment (GE) interaction studies
- Standard analyses of case-control studies
- Case-control studies as a semiparametric model
- Reformulation as a missing data problem
- The case that the probability of disease is known in the population

STATISTICS

A Little Terminology

- **Epidemiologists**: Case control sample
- **Econometricians**: Choice-based sample
- These are exactly the same problems
- Subjects have two choices (or disease states)
- Subjects have their covariates sampled conditional on their choices, i.e.,
  - Random sample from those **with** disease
  - Random sample from those **without** disease

STATISTICS

The Key Point

- You are used to regression: you observe a response D given the values of covariates (G,X)
  - Sometimes this comes from a random sample
  - These are called **Prospective** studies
- Case-control studies are entirely different

STATISTICS

The Key Point

- Case-control studies are entirely different
- You find people with D = 1, and sample at random from them to get (G,X)
- Same for those with D = 0
- These studies are **Retrospective**, i.e., occur after you observe the disease

Israeli Ovarian Cancer Study

- G = the BRCA1/2 mutation
  - Largely confined to Jews of European origin
  - Very deadly: 30%-50% rate of breast and ovariance cancer
- X = use of oral contraceptives
- **Question**: for those with the mutation, does use of an oral contraceptive decrease probability of disease
  - Formalized as an interaction

Israeli Ovarian Cancer Study

- Let $H(x)=\{1+\exp(-x)\}^{-1}$ be the logistic distribution function
- Standard model with no interaction

$$pr(D=1\mid G,X)=H(\beta_0+\beta_1 G+\beta_2 X)$$

- The odds

$$\frac{pr(D=1\mid G,X)}{pr(D=0\mid G,X)}=\exp(\beta_0+\beta_1 G+\beta_2 X)$$

Israeli Ovarian Cancer Study

- The odds ratio of (G,X) versus (G_0,X_0)

$$\frac{pr(D=1\mid G,X)/pr(D=0\mid G,X)}{pr(D=1\mid G_0,X_0)/pr(D=0\mid G_0,X_0)}=\exp\{\beta_1(G-G_0)+\beta_2(X-X_0)\}$$

- For people with the same gene status ($G = G_0$), what is the odds ratio of those with oral contraceptives (X=1) versus not (X_0=0): $\exp(\beta_2)$
- The odds ratio is independent of the gene.

Israeli Ovarian Cancer Study

- Now add an interaction

$$\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\mathbf{G})=\mathbf{H}(\boldsymbol{\beta_0}+\boldsymbol{\beta_1}\mathbf{G}+\boldsymbol{\beta_2}\mathbf{X}+\boldsymbol{\beta_3}\mathbf{XG})$$

- For people with BRCA1/2 (G = G_0=1), the odds ratio of those with oral contraceptives (X=1) versus not (X_0=0): **$\exp(\beta_2+\beta_3)$**
- For those without BRCA1/2: **$\exp(\beta_2)$**
- **Odd ratio depends on gene status**

Basic Problem Formalized

- Case control sample: **D** = disease
- Gene expression: **G**
- Environment: **X**
- We are interested in main effects for G and (X,S) **along with** their interaction

Likelihood Function

- Consider a case-control study without interaction and only a single covariate X

$$\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X})=\mathbf{H}(\boldsymbol{\beta_0}+\boldsymbol{\beta_1}\mathbf{X})$$

- This is a retrospective, case-control study, not a prospective study.

Likelihood Function

- The likelihood is

$$\mathbf{pr}(\mathbf{X}=\mathbf{x}|\mathbf{D}=\mathbf{d})=\frac{\mathbf{pr}(\mathbf{X}=\mathbf{x})\mathbf{H}(\boldsymbol{\beta_0}+\boldsymbol{\beta_1}\mathbf{x})}{\mathbf{pr}(\mathbf{D}=\mathbf{d})}$$

- Note how the likelihood depends on two things:
  - The distribution of X **in the population**
  - The probability of disease in the population
- Unless the latter is known, neither can be estimated, nor can β_0 For this reason, epidemiologists focus on odds ratios

Cornfield's Calculation

- Cornfield (1956) recognized that you can estimate the slope parameter from case control data

$$\frac{\mathbf{pr}(\mathbf{X}=\mathbf{x}|\mathbf{D}=\mathbf{1})/\mathbf{pr}(\mathbf{X}=\mathbf{x}|\mathbf{D}=\mathbf{0})}{\mathbf{pr}(\mathbf{X}=\mathbf{x_0}|\mathbf{D}=\mathbf{1})/\mathbf{pr}(\mathbf{X}=\mathbf{x_0}|\mathbf{D}=\mathbf{0})} = \exp\{\boldsymbol{\beta_1}(\mathbf{x}-\mathbf{x_0})\}$$

- **This is the odds ratio!**

$$\frac{\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X}=\mathbf{x})/\mathbf{pr}(\mathbf{D}=\mathbf{0}|\mathbf{X}=\mathbf{x})}{\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X}=\mathbf{x_0})/\mathbf{pr}(\mathbf{D}=\mathbf{0}|\mathbf{X}=\mathbf{x_0})} = \exp\{\boldsymbol{\beta_1}(\mathbf{x}-\mathbf{x_0})\}$$

Cornfield's Calculation

- Cornfield (1956) recognized from this calculation that you can get the slope β_1 **without even having to think about whether it is a case-control study or not**!
- At least when X is binary.

Andersen and Prentice & Pyke

- Andersen (1970) and more generally Prentice & Pyke (1979) showed that Cornfield was right on all fronts.
- They first showed that if you run an ordinary logistic regression, **ignoring that you have a case-control study**, then you get consistent estimates.

Consistency

- I will only do the details once, so you can get a flavor of the tricks.
- Let X have density function f in the population
- Let the number with D=d in the case control study be denoted as n_d.
- Define

$$\pi_{\mathbf{d}} = \mathbf{pr}(\mathbf{D}=\mathbf{d})$$

$$\boldsymbol{\beta}_{\mathbf{0}}^{*} = \boldsymbol{\beta}_{\mathbf{0}} + \mathbf{log}(\mathbf{n_1}/\mathbf{n_0}) - \mathbf{log}(\pi_1/\pi_0)$$

Consistency

- We will use:

$$\mathbf{H}(\beta_0 + \beta_1 \mathbf{x}) = \{1 - \mathbf{H}(\beta_0 + \beta_1 \mathbf{x})\} \exp(\beta_0 + \beta_1 \mathbf{x})$$

- The intercept is not identified.
- The score of the logistic loglikelihood function that ignores the case control sampling scheme is

$$\sum_{i=1}^{n} \mathbf{X_i}\{\mathbf{D_i} - \mathbf{H}(\beta_0^* + \beta_1 \mathbf{X_i})\}$$

- We have to show that this has mean zero **in the case-control sampling scheme**, for some version of an intercept

Consistency

- Since there are n_d people with D=d, and the likelihood in the case-control data is the distribution of X given D, the expectation of the score is

$$\mathbf{E}\left[\mathbf{n_1 X}\{1 - \mathbf{H}(\beta_0^* + \beta_1 \mathbf{X})\} | \mathbf{D} = 1\right]$$
$$-\mathbf{E}\left[\mathbf{n_0 X H}(\beta_0^* + \beta_1 \mathbf{X}) | \mathbf{D} = 0\right]$$

- Now use the earlier formula for the density of X given D

Consistency

- The expectation of the score is

$$\int \left[\frac{\mathbf{n_1}}{\pi_1}\mathbf{x}\{1 - \mathbf{H}(\beta_0^* + \beta_1 \mathbf{x})\}\mathbf{H}(\beta_0 + \beta_1 \mathbf{x})\right.$$
$$\left. -\frac{\mathbf{n_0}}{\pi_0}\mathbf{x}\mathbf{H}(\beta_0^* + \beta_1 \mathbf{x})\{1 - \mathbf{H}(\beta_0 + \beta_1 \mathbf{x})\}\right]\mathbf{f}(\mathbf{x})\mathbf{dx}$$

- Now use the facts

$$\beta_0^* = \beta_0 + \mathbf{log}(\mathbf{n_1}/\mathbf{n_0}) - \mathbf{log}(\pi_1/\pi_0)$$

$$\mathbf{H}(\beta_0 + \beta_1 \mathbf{x}) = \{1 - \mathbf{H}(\beta_0 + \beta_1 \mathbf{x})\} \exp(\beta_0 + \beta_1 \mathbf{x})$$

$$\mathbf{H}(\beta_0^* + \beta_1 \mathbf{x}) = \{1 - \mathbf{H}(\beta_0^* + \beta_1 \mathbf{x})\} \exp(\beta_0^* + \beta_1 \mathbf{x})$$

- This shows that the term in brackets = 0 identically, as needed!

Inference

- This shows that the logistic regression loglikelihood score that ignores the case-control sampling scheme is an unbiased estimating equation.
- It solves something of the form

$$\mathbf{0} = \sum_{i=1}^{n} \Psi(\mathbf{D_i}, \mathbf{X_i}, \beta_0^*, \beta_1).$$

Inference

$$0 = \sum_{i=1}^{n} \Psi(D_i, X_i, \beta_0^*, \beta_1).$$

- Using the same sort of calculations that I used to show consistency (the same tricks!), it is easy to show that the asymptotic variance of the estimate of β_1 is estimated consistently by using the inverse of the observed prospective Fisher information

Semiparametric Formulation

- The probability of disease is

$$\begin{aligned} pr(D=1) &= \int pr(D=1|X=x) f(x) dx \\ &= \int H(\beta_0 + \beta_1 x) f(x) dx \end{aligned}$$

Semiparametric Formulation

- The likelihood function is

$$\begin{aligned} f_{X|D}(x|d) &= \frac{f(x) pr(D=d|X=x)}{pr(D=d)} \\ &= \frac{f(x) H^d(\beta_0+\beta_1 x)\{1-H(\beta_0+\beta_1 x)\}^{1-d}}{\int H^d(\beta_0+\beta_1 z)\{1-H(\beta_0+\beta_1 z)\}^{1-d} f(z) dz} \end{aligned}$$

- Thus, the model has a parametric part, (β_0, β_1) and a nonparametric part $f(\cdot)$

Semiparametric Formulation

- The likelihood function is

$$\frac{f(x) H^d(\beta_0+\beta_1 x)\{1-H(\beta_0+\beta_1 x)\}^{1-d}}{\int H^d(\beta_0+\beta_1 z)\{1-H(\beta_0+\beta_1 z)\}^{1-d} f(z) dz}$$

- If we make no assumptions about $f(\cdot)$, then we cannot identify $f(\cdot)$ but can identify

$$\{\beta_0^* = \beta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0), \beta_1\}$$

- Prentice and Pyke show that the best estimate possible of β_1 is that from ordinary logistic regression ignoring the case-control sampling scheme

Parametric Formulation?

- The likelihood function is

$$\frac{f(x)H^d(\beta_0+\beta_1 x)\{1-H(\beta_0+\beta_1 x)\}^{1-d}}{\int H^d(\beta_0+\beta_1 z)\{1-H(\beta_0+\beta_1 z)\}^{1-d} f(z)dz}$$

- If we make parametric assumptions about $f(\cdot)$, then we can identify **everything**! For example, X = Normal
- No one does this for two reasons:
  - Model Robustness
  - Computational convenience

Missing Data Formulation

- Suppose you have a large but finite population of size N
- Then, there are $N\pi_1$ with the disease
- There are $N\pi_0$ without the disease
- In a case-control sample, we randomly select n_1 with the disease, and n_0 without.
- The fraction of people with disease status D=d that we observe is

$$\frac{n_d}{N\pi_d}$$

Missing Data Formulation

- Then let's make up a "pretend" study, that has random sampling with missing data
- I take a random sample of size N
- I get to observe (D,X) when D=d with probability $\frac{n_d}{N\pi_d}$
- I will say that $\delta = 1$ if I observe (D,X). Then

$$pr(\delta = 1|D = d, X) = pr(\delta = 1|D = d) = \frac{n_d}{N\pi_d}$$

Missing Data Formulation

- Now let us compute the probability of disease given X and given that (D,X) was observed, namely

$$\begin{aligned} &pr(D = 1|X, \delta = 1) \\ &= \frac{pr(D = 1, \delta = 1|X)}{pr(\delta = 1|X)} \\ &= \frac{pr(D = 1, \delta = 1|X)}{\sum_{d=0}^{1} pr(D = d, \delta = 1|X)} \end{aligned}$$

Missing Data Formulation

- Continuing

$$\begin{aligned}&\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\delta=\mathbf{1})\\&=\frac{\mathbf{pr}(\delta=\mathbf{1}|\mathbf{D}=\mathbf{1},\mathbf{X})\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X})}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\mathbf{pr}(\delta=\mathbf{1}|\mathbf{D}=\mathbf{d},\mathbf{X})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X})}\\&=\frac{(\mathbf{n_1}/\pi_\mathbf{1})\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X})}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}(\mathbf{n_d}/\pi_\mathbf{d})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X})}\end{aligned}$$

- Now use the fact that

$$\frac{\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X})}{\mathbf{pr}(\mathbf{D}=\mathbf{0}|\mathbf{X})}=\mathbf{exp}(\boldsymbol{\beta_0}+\boldsymbol{\beta_1}\mathbf{X})$$

Missing Data Formulation

- Doing the simple algebra, we get

$$\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\delta=\mathbf{1})=\mathbf{H}(\boldsymbol{\beta}_\mathbf{0}^*+\boldsymbol{\beta_1}\mathbf{X})$$

- Implication: By working in a closely related "pretend" or fictitious study, we would
  - Automatically use logistic regression, getting consistency and asymptotic normality
  - Get (asymptotically correct) standard errors from logistic regression
  - Know that we cannot identify the intercept

pr(D=1) Known

- If we know pr(D=1), then we can identify the intercept, since we can estimate

$$\beta_0^*=\beta_0+\mathbf{log}(\mathbf{n_1}/\mathbf{n_0})-\mathbf{log}(\pi_1/\pi_0)$$

- We can also estimate the density of X, since

$$\mathbf{f_X}(\mathbf{x})=\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\pi_\mathbf{d}\mathbf{f_{X|D}}(\mathbf{x}|\mathbf{d})$$

- This does not improve our estimate of β_1

A Semiparametric Derivation

- Here is how to see that ordinary logistic regression is optimal semiparametric if no assumptions are made about the distribution of X in the source population
- Suppose that X is discrete in the population (not the case control sampling scheme).
  - Support point are $\mathbf{x_1},\ldots,\mathbf{x_M}$
  - Also, $\mathbf{pr}(\mathbf{X}=\mathbf{x_m})=\gamma_\mathbf{m}$

A Semiparametric Derivation

- Support points are $\mathbf{x_1, \ldots, x_M}$
- Also, $\mathbf{pr(X = x_m) = \gamma_m}$
- Key identity

$$\mathbf{pr(D = d) = \sum_{m=1}^{M} pr(D = d|X = x_m)\gamma_m} \qquad (1)$$

A Semiparametric Derivation

- The retrospective loglikelihood of the data is

$$\begin{aligned}&\sum_{i=1}^{n} \log\left\{\frac{f_X(X_i)pr(D = D_i|X_i)}{pr(D = D_i)}\right\}\\ &= \sum_{i=1}^{n} \log\left\{\sum_{m=1}^{M} \gamma_m I(X_i = x_m)\right\} + \sum_{i=1}^{n} \log\{pr(D = D_i|X_i)\}\\ &\quad - \sum_{i=1}^{n} \log\left\{\sum_{m=1}^{M} \gamma_m pr(D = D_i|x_m)\right\}\end{aligned}$$

A Semiparametric Derivation

- Differentiate

$$\frac{\sum_{i=1}^{n} I(X_i = x_m)}{\gamma_m} - \sum_{i=1}^{n} \frac{pr(D = D_i|x_m)}{\sum_{j=1}^{M} pr(D = D_i|x_j)\gamma_j}$$

- Now use (1):

$$\begin{aligned}&\frac{\sum_{i=1}^{n} I(X_i = x_m)}{\gamma_m} - \sum_{i=1}^{n} \frac{pr(D = D_i|x_m)}{\pi_{D_i}}\\ &= \frac{\sum_{i=1}^{n} I(X_i = x_m)}{\gamma_m} - \sum_{d=0}^{1} pr(D = d|x_m)\frac{n_d}{\pi_d}\end{aligned}$$

A Semiparametric Derivation

- This means that

$$\gamma_m = \frac{\sum_{i=1}^{n} I(X_i = x_m)}{\sum_{d=0}^{1} pr(D = d|x_m)\frac{n_d}{\pi_d}}$$

- Now go back to the loglikelihood but plug in (1), since the probabilities solve it. The new loglikelihood is

$$\sum_{i=1}^{n} \log\left\{\sum_{m=1}^{M} \gamma_m I(X_i = x_m)\right\} + \sum_{i=1}^{n} \log\{pr(D = D_i|X_i)/\pi_{D_i}\}$$

A Semiparametric Derivation

- Now make the support points equal to the data points (assumed distinct), so that the loglikelihood becomes

$$-\sum_{i=1}^{n}\log\left\{\sum_{d=0}^{1}\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X}_i)\frac{\mathbf{n_d}}{\pi_d}\right\}+\sum_{i=1}^{n}\log\left\{\mathbf{pr}(\mathbf{D}=\mathbf{D}_i|\mathbf{X}_i)/\pi_{\mathbf{D}_i}\right\}$$

- Add an irrelevant constant

$$-\sum_{i=1}^{n}\log\left\{\sum_{d=0}^{1}\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X}_i)\frac{\mathbf{n_d}}{\pi_d}\right\}+\sum_{i=1}^{n}\log\left\{\mathbf{pr}(\mathbf{D}=\mathbf{D}_i|\mathbf{X}_i)\frac{\mathbf{n}_{\mathbf{D}_i}}{\pi_{\mathbf{D}_i}}\right\}$$

A Semiparametric Derivation

$$-\sum_{i=1}^{n}\log\left\{\sum_{d=0}^{1}\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X}_i)\frac{\mathbf{n_d}}{\pi_d}\right\}+\sum_{i=1}^{n}\log\left\{\mathbf{pr}(\mathbf{D}=\mathbf{D}_i|\mathbf{X}_i)\frac{\mathbf{n}_{\mathbf{D}_i}}{\pi_{\mathbf{D}_i}}\right\}$$

- Now use the fact that

$$\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X})=\mathbf{pr}(\mathbf{D}=\mathbf{0}|\mathbf{X})\mathbf{exp}\left\{\mathbf{d}(\beta_0+\beta_1\mathbf{X})\right\}$$

- With some algebra, this becomes

$$-\sum_{i=1}^{n}\log\left[\frac{\mathbf{n_0}}{\pi_0}\left\{1+\exp(\beta_0^*+\beta_1\mathbf{X}_i)\right\}\right]+\sum_{i=1}^{n}\log\left[\exp\left\{\mathbf{D}_i(\beta_0+\beta_1\mathbf{X}_i)\right\}\frac{\mathbf{n}_{\mathbf{D}_i}}{\pi_{\mathbf{D}_i}}\right]$$

A Semiparametric Derivation

$$-\sum_{i=1}^{n}\log\left[\frac{\mathbf{n_0}}{\pi_0}\left\{1+\exp(\beta_0^*+\beta_1\mathbf{X}_i)\right\}\right]+\sum_{i=1}^{n}\log\left[\exp\left\{\mathbf{D}_i(\beta_0+\beta_1\mathbf{X}_i)\right\}\frac{\mathbf{n}_{\mathbf{D}_i}}{\pi_{\mathbf{D}_i}}\right]$$

- With a little more algebra, this is

$$-\sum_{i=1}^{n}\log\left[\left\{1+\exp(\beta_0^*+\beta_1\mathbf{X}_i)\right\}\right] \qquad (2)$$
$$+\sum_{i=1}^{n}\log\left[\exp\left\{\mathbf{D}_i(\beta_0^*+\beta_1\mathbf{X}_i)\right\}\right]$$

- Equation (2) is the loglikelihood for logistic regression with the designated intercept and slope!

A Semiparametric Derivation

- What this argument shows is that if you make no assumptions about the distribution of X in the population, then
  - You should use logistic regression
  - The intercept is not identified
  - The only way to do better than logistic regression is to make assumptions about the distribution of X in the population

Gene-Environment Case-Control Studies

Raymond J. Carroll
Department of Statistics
Faculties of Nutrition and Toxicology

Texas A&M University
http://stat.tamu.edu/~carroll

STATISTICS
TEXAS A&M UNIVERSITY

Lecture 2: Outline

- GE independence: the case-only method
- Case-control studies with GE independence: a different semiparametric model
- The missing data approach
- The semiparametric MLE approach
- Simulations
- Details of Implementation

STATISTICS

Interactions

- Genetic epidemiology has two major thrusts
- Finding Genes associated with disease
  - QTL mapping
  - Genome-Wide scans
  - Microarray
- Discovering whether the impact of genetics is affected by the environment
  - BRCA1/2 and use of oral contraceptives

STATISTICS

Interactions

- Thus, we need to model interactions
- Most often, this is done via simple multiplicative interactions
- There are some specialized models that have non-multiplicative interactions

STATISTICS

Prospective Models

- Simplest multiplicative interactionlogistic model

$$pr(D = 1 \mid G, X) = H(\beta_0 + \beta_1 G + \beta_2 X + \beta_3 G * X)$$

- General logistic model

$$pr(D = 1 \mid G, X) = H\{\beta_0 + m(G, X, \beta_1)\}$$

- The function m(G,X,β_1) is completely general

Gene-Environment Independence

- In many situations, it may be reasonable to assume G and X are independently distributed **in the underlying population**, possibly after conditioning on strata
- This assumption is often used in gene-environment interaction studies

G-E Independence: Discussion

- Does not always hold!
- **Example**: polymorphisms in the smoking metabolism pathway may affect the degree of addiction

G-E Independence: Discussion

- It is reasonable in many problems
- **Example**: Environment is a treatment in a randomized study under nested case-control sampling
- **Example**: Reasonable when exposure is not directly controlled by individual behavior
  - Radiation exposure for A-bomb survivors
  - Carcinogenic exposure of employees
  - Pesticide exposure in a rural community

G-E Independence: Discussion

- In the Israeli Study, it is hard to believe that one's unknown BRCA1/2 status really influences whether one uses oral contraceptives

Rare Events

- The logistic distribution function is

$$\begin{aligned}\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\mathbf{G}) &= \mathbf{H}\{\boldsymbol{\beta_0}+\mathbf{m}(\mathbf{X},\mathbf{G},\boldsymbol{\beta_1})\} \\ &= \frac{\exp\{\boldsymbol{\beta_0}+\mathbf{m}(\mathbf{X},\mathbf{G},\boldsymbol{\beta_1})\}}{\mathbf{1}+\exp\{\boldsymbol{\beta_0}+\mathbf{m}(\mathbf{X},\mathbf{G},\boldsymbol{\beta_1})\}}\end{aligned}$$

$$\mathbf{pr}(\mathbf{D}=\mathbf{0}|\mathbf{X},\mathbf{G}) = \frac{\mathbf{1}}{\mathbf{1}+\exp\{\boldsymbol{\beta_0}+\mathbf{m}(\mathbf{X},\mathbf{G},\boldsymbol{\beta_1})\}}$$

Rare Events

- For rare events then, The logistic distribution function is

$$\mathbf{pr}(\mathbf{D}=\mathbf{0}|\mathbf{X},\mathbf{G}) \approx \mathbf{1}$$

$$\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\mathbf{G}) \approx \exp\{\boldsymbol{\beta_0}+\mathbf{m}(\mathbf{X},\mathbf{G},\boldsymbol{\beta_1})\}$$

- Now consider that G and X are binary, and a multiplicative interaction

$$\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\mathbf{G}) \approx \exp\{\boldsymbol{\beta_0}+\boldsymbol{\beta_1}\mathbf{X}+\boldsymbol{\beta_2}\mathbf{G}+\boldsymbol{\beta_3}\mathbf{XG}\}$$

Rare Events and Case-Only

- Let $\mathbf{pr}(\mathbf{X}=\mathbf{x},\mathbf{G}=\mathbf{g}) = \mathbf{q}_{\mathbf{X},\mathbf{G}}(\mathbf{x},\mathbf{g})$
- For rare events, the retrospective likelihoods among cases (D=1) satisfies

$$\begin{aligned}&\frac{\mathbf{pr}(\mathbf{X}=\mathbf{1},\mathbf{G}=\mathbf{1}|\mathbf{D}=\mathbf{1})\mathbf{pr}(\mathbf{X}=\mathbf{0},\mathbf{G}=\mathbf{0}|\mathbf{D}=\mathbf{1})}{\mathbf{pr}(\mathbf{X}=\mathbf{1},\mathbf{G}=\mathbf{0}|\mathbf{D}=\mathbf{1})\mathbf{pr}(\mathbf{X}=\mathbf{0},\mathbf{G}=\mathbf{1}|\mathbf{D}=\mathbf{1})} \\ &\approx \frac{\mathbf{q}_{\mathbf{X},\mathbf{G}}(\mathbf{1},\mathbf{1})\mathbf{q}_{\mathbf{X},\mathbf{G}}(\mathbf{0},\mathbf{0})}{\mathbf{q}_{\mathbf{X},\mathbf{G}}(\mathbf{1},\mathbf{0})\mathbf{q}_{\mathbf{X},\mathbf{G}}(\mathbf{0},\mathbf{1})}\exp(\boldsymbol{\beta_3})\end{aligned}$$

- While intriguing, this does no good because we do not know the distribution of (X,G) **in the population**

Gene-Environment Independence

- Now suppose in addition to rare disease, G amd X are independent **in the population**. Then

$$\begin{aligned}
&\frac{\mathbf{pr}(\mathbf{X}=1,\mathbf{G}=1|\mathbf{D}=1)\mathbf{pr}(\mathbf{X}=0,\mathbf{G}=0|\mathbf{D}=1)}{\mathbf{pr}(\mathbf{X}=1,\mathbf{G}=0|\mathbf{D}=1)\mathbf{pr}(\mathbf{X}=0,\mathbf{G}=1|\mathbf{D}=1)} \\
&\quad\approx \frac{\mathbf{q}_{\mathbf{X},\mathbf{G}}(1,1)\mathbf{q}_{\mathbf{X},\mathbf{G}}(0,0)}{\mathbf{q}_{\mathbf{X},\mathbf{G}}(1,0)\mathbf{q}_{\mathbf{X},\mathbf{G}}(0,1)}\mathbf{exp}\,(\beta_3) \\
&\quad= \frac{\mathbf{q}_{\mathbf{X}}(1)\mathbf{q}_{\mathbf{G}}(1)\mathbf{q}_{\mathbf{X}}(0)\mathbf{q}_{\mathbf{G}}(0)}{\mathbf{q}_{\mathbf{X}}(1)\mathbf{q}_{\mathbf{G}}(0)\mathbf{q}_{\mathbf{X}}(0)\mathbf{q}_{\mathbf{G}}(1)}\mathbf{exp}\,(\beta_3) \\
&\quad= \mathbf{exp}\,(\beta_3)
\end{aligned}$$

Gene-Environment Independence

- **Conclusion**: Suppose you are willing to assume
  - Rare disease
  - Gene-Environment independence in the population
  - G, X binary
  - Multiplicative interaction
- Then, **without doing logistic regression**, you can get an approximately correct estimate of the interaction term

Semiparametric Contradiction?

- Simulations show that the case only estimate of the interaction term is much less variable than the logistic regression estimate.
- But, logistic regression is semiparametric efficient.
- Is this a contradiction?

Semiparametric Contradiction?

- It is not a contradiction
- Logistic regression is efficient if **no assumptions** are made about the distribution of (X,G) in the population
- The case only method makes an assumption: X and G are independent in the population
- Clearly, this assumption carries (Fisher) information

General Methods

- The case-only method is very special
- G and X must be binary
- The disease must be rare for all values of (G,X)
  - Not true for BRCA1/2
- It only estimates interaction
- The method is not even formally consistent
- Missing (X,G) data cannot be handled

General Methods

- In part I, we discussed two methods for creating an estimator in the case-control study
- The pretend study trick, where we made up a prospective study with missing data
- The formal semiparametric MLE calculation
- Both resulted in the same, efficient answer
- Let's try again

Pretend Missing Data Formulation

- Assume X and G are independent
- Assume G is binary, and

$$\mathbf{pr}(\mathbf{G}=\mathbf{g})=\mathbf{q_G}(\mathbf{g},\boldsymbol{\theta})$$

- Make no assumptions about the distribution of the often multivariate X
- Recall that $\delta=\mathbf{1}$ if we see (G,X) and

$$\mathbf{pr}(\delta=\mathbf{1}|\mathbf{D}=\mathbf{d},\mathbf{X},\mathbf{G})=\mathbf{pr}(\delta=\mathbf{1}|\mathbf{D}=\mathbf{d})=\frac{\mathbf{n_d}}{\mathbf{N}\pi_\mathbf{d}}$$

Pretend Missing Data Formulation

- Since G is binary, it is easily and robustly modeled in the population, so it makes sense to compute

$$\mathbf{pr}(\mathbf{D}=\mathbf{1},\mathbf{G}=\mathbf{g}|\mathbf{X},\delta=\mathbf{1})$$

- Logistic regression computes

$$\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{G}=\mathbf{g},\mathbf{X},\delta=\mathbf{1})$$

Pretend Missing Data Formulation

- Compute probability of observed (D,G) data in the pretend study

$$\mathbf{pr}(\mathbf{D}=\mathbf{1},\mathbf{G}=\mathbf{g}|\mathbf{X},\delta=\mathbf{1}) = \frac{\mathbf{pr}(\mathbf{D}=\mathbf{1},\mathbf{G}=\mathbf{g},\delta=\mathbf{1}|\mathbf{X})}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\sum_{\mathbf{s}=\mathbf{0}}^{\mathbf{1}}\mathbf{pr}(\mathbf{D}=\mathbf{d},\mathbf{G}=\mathbf{s},\delta=\mathbf{1}|\mathbf{X})}$$

Pretend Missing Data Formulation

- Carrying on the calculations as before we get

$$\begin{aligned}&\mathbf{pr}(\mathbf{D}=\mathbf{1},\mathbf{G}=\mathbf{g}|\mathbf{X},\delta=\mathbf{1})\\ &= \frac{(\mathbf{n_1}/\pi_\mathbf{1})\mathbf{pr}(\mathbf{D}=\mathbf{1},\mathbf{G}=\mathbf{g}|\mathbf{X})}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\sum_{\mathbf{s}=\mathbf{0}}^{\mathbf{1}}(\mathbf{n_d}/\pi_\mathbf{d})\mathbf{pr}(\mathbf{D}=\mathbf{d},\mathbf{G}=\mathbf{s}|\mathbf{X})}\\ &= \\ &= \frac{(\mathbf{n_1}/\pi_\mathbf{1})\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{G}=\mathbf{g},\mathbf{X})\mathbf{pr}(\mathbf{G}=\mathbf{g}|\mathbf{X})}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\sum_{\mathbf{s}=\mathbf{0}}^{\mathbf{1}}(\mathbf{n_d}/\pi_\mathbf{d})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{G}=\mathbf{s},\mathbf{X})\mathbf{pr}(\mathbf{G}=\mathbf{s}|\mathbf{X})}\end{aligned}$$

- But, G and X are independent!

Pretend Missing Data Formulation

- Then

$$\begin{aligned}&\mathbf{pr}(\mathbf{D}=\mathbf{1},\mathbf{G}=\mathbf{g}|\mathbf{X},\delta=\mathbf{1})\\ &= \\ &= \frac{(\mathbf{n_1}/\pi_\mathbf{1})\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{G}=\mathbf{g},\mathbf{X})\mathbf{q_G}(\mathbf{g},\theta)}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\sum_{\mathbf{s}=\mathbf{0}}^{\mathbf{1}}(\mathbf{n_d}/\pi_\mathbf{d})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{G}=\mathbf{s},\mathbf{X})\mathbf{q_G}(\mathbf{s},\theta)}\end{aligned}$$

- Now it is just algebra

Pretend Missing Data Formulation

- Define $\boldsymbol{\Theta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \theta, \boldsymbol{\beta_0^*}) = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \theta, \pi_\mathbf{1})$
- Further define

$$\mathbf{S}(\mathbf{d},\mathbf{x},\mathbf{g},\boldsymbol{\Theta}) = \frac{\mathbf{q}(\mathbf{g},\theta)\mathbf{exp}\left[\mathbf{d}\left\{\boldsymbol{\beta_0^*}+\mathbf{m}(\mathbf{x},\mathbf{g},\boldsymbol{\beta_1})\right\}\right]}{\mathbf{1}+\mathbf{exp}\left\{\boldsymbol{\beta_0}+\mathbf{m}(\mathbf{x},\mathbf{g},\boldsymbol{\beta_1})\right\}}$$

- Then

$$\mathbf{pr}(\mathbf{D}=\mathbf{d},\mathbf{G}=\mathbf{g}|\mathbf{X},\delta=\mathbf{1}) = \frac{\mathbf{S}(\mathbf{D},\mathbf{X},\mathbf{G},\boldsymbol{\Theta})}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\sum_{\mathbf{s}=\mathbf{0}}^{\mathbf{1}}\mathbf{S}(\mathbf{d},\mathbf{X},\mathbf{g},\boldsymbol{\Theta})}$$

Summary

- Define $\boldsymbol{\Theta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \theta, \boldsymbol{\beta_0^*}) = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \theta, \pi_1)$

$$\mathbf{pr}(\mathbf{D}=\mathbf{d}, \mathbf{G}=\mathbf{g}|\mathbf{X}, \delta=\mathbf{1}) = \frac{\mathbf{S}(\mathbf{D}, \mathbf{X}, \mathbf{G}, \boldsymbol{\Theta})}{\sum_{\mathbf{d}=0}^{\mathbf{1}} \sum_{\mathbf{s}=0}^{\mathbf{1}} \mathbf{S}(\mathbf{d}, \mathbf{X}, \mathbf{g}, \boldsymbol{\Theta})}$$

- This pretend likelihood is easily computed.
- Note that it involve both $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_0^*}$ or equivalently both $\boldsymbol{\beta_0}$ and π_1

Summary

- The pretend likelihood involves both $\boldsymbol{\beta_0}$ and π_1
- This holds out the hope that we can estimate the intercept and hence the disease prevalence in the population
- However, it is a pretend likelihood, not the retrospective likelihood.
- Is the method any good?
- (Obviously it is or I would not be talking about it)

Prentice-Pyke Calculation

- Methodology: Start with the retrospective likelihood

$$\mathbf{pr(G=g, X=x \mid D=d)} = \frac{\mathbf{pr(X=x, G=g)}\exp\left[\mathbf{d}\{\beta_0 + \mathbf{m(g, x, \beta_1)}\}\right]\left[\mathbf{1} - \mathbf{H}\{\beta_0 + \mathbf{m(g, x, \beta_1)}\}\right]}{\sum_{\mathbf{x', g'}} \mathbf{pr(X=x', G=g')}\exp\left[\mathbf{d}\{\beta_0 + \mathbf{m(g', x', \beta_1)}\}\right]\left[\mathbf{1} - \mathbf{H}\{\beta_0 + \mathbf{m(g', x', \beta_1)}\}\right]}$$

- The distribution of (X,G) in the population is left unspecified (**the nonparametric part**)
- Semiparametric MLE is usual logistic regression

Environment and Gene Expression

- Methodology: Start with the retrospective likelihood

$$\mathbf{pr(G=g, X=x \mid D=d)} = \frac{\mathbf{pr(X=x)pr(G=g)}\exp\left[\mathbf{d}\{\beta_0 + \mathbf{m(g, x, \beta_1)}\}\right]\left[\mathbf{1} - \mathbf{H}\{\beta_0 + \mathbf{m(g, x, \beta_1)}\}\right]}{\sum_{\mathbf{x', g'}} \mathbf{pr(X=x')pr(G=g')}\exp\left[\mathbf{d}\{\beta_0 + \mathbf{m(g', x', \beta_1)}\}\right]\left[\mathbf{1} - \mathbf{H}\{\beta_0 + \mathbf{m(g', x', \beta_1)}\}\right]}$$

- Note how independence of G and X is used here, see the **red** expressions
- We do not want to model the **often multivariate** distribution of X

Semiparametric MLE

- The retrospective loglikelihood of the data is

$$\sum_{i=1}^{n} \log\left\{\frac{f_X(X_i)q(G_i,\theta)pr(D=D_i|X_i,G_i)}{pr(D=D_i)}\right\}$$

- Again assume that X has a discrete distribution, with support points at the observed data (nonparametric MLE)

Semiparametric MLE

- The retrospective loglikelihood of the data is

$$\sum_{i=1}^{n} \log\left\{\frac{f_X(X_i)q(G_i,\theta)pr(D=D_i|X_i,G_i)}{pr(D=D_i)}\right\}$$
$$= \sum_{i=1}^{n} \log\left\{\sum_{m=1}^{M} \gamma_m I(X_i = x_m)\right\} + \sum_{i=1}^{n} \log\{q(G_i,\theta)\}$$
$$+ \sum_{i=1}^{n} \log\{pr(D=D_i|X_i,G_i)\}$$
$$- \sum_{i=1}^{n} \log\left\{\sum_{m=1}^{M}\sum_{g'=0}^{1} \gamma_m q(g',\theta) pr(D=D_i|X=x_m, G=g')\right\}$$

Semiparametric MLE

- Then, as before, we maximize this in the probabilities $(\gamma_1, \ldots, \gamma_m)$ subject to the key constraint

$$pr(D=d) = \sum_{m=1}^{M}\sum_{g'=0}^{1} \gamma_m q(g',\theta) pr(D=D_i|X=x_m, G=g') \quad (3)$$

Semiparametric MLE

- Then, nearly identical calculations as in the fully nonparametric case are used
- The use of (3) occurs in exactly the same places
- **Result**: The semiparametric profile likelihood is **exactly** the same as that of the pretend missing data likelihood calculation!

Summary

- Define $\boldsymbol{\Theta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \theta, \boldsymbol{\beta_0^*}) = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \theta, \pi_1)$
- Further define

$$\mathbf{S}(\mathbf{d}, \mathbf{x}, \mathbf{g}, \boldsymbol{\Theta}) = \frac{\mathbf{q}(\mathbf{g}, \theta)\mathbf{exp}\left[\mathbf{d}\left\{\boldsymbol{\beta_0^*} + \mathbf{m}(\mathbf{x}, \mathbf{g}, \boldsymbol{\beta_1})\right\}\right]}{\mathbf{1} + \mathbf{exp}\left\{\boldsymbol{\beta_0} + \mathbf{m}(\mathbf{x}, \mathbf{g}, \boldsymbol{\beta_1})\right\}}$$

- Then, the semiparametric likelihood function is

$$\mathcal{L}(\mathbf{D}, \mathbf{X}, \mathbf{G}, \boldsymbol{\Theta}) = \frac{\mathbf{S}(\mathbf{D}, \mathbf{X}, \mathbf{G}, \boldsymbol{\Theta})}{\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}} \sum_{\mathbf{s}=\mathbf{0}}^{\mathbf{1}} \mathbf{S}(\mathbf{d}, \mathbf{X}, \mathbf{g}, \boldsymbol{\Theta})}$$

Summary

- When we allow G and X to be correlated, we find that $\mathbf{pr}(\mathbf{D} = \mathbf{1}) = \pi_1$ is confounded with the intercept, i.e., the semiparametric likelihood function combines them into $\boldsymbol{\beta_0^*}$
- When we assume G and X are independent, they are not confounded: π_1 and $\boldsymbol{\beta_0}$ enter the likelihood function separately

Computation

- Sometimes, π_1 is hard to estimate, i.e., the semiparametric likelihood function is flat in this argument
- Sometimes, π_1 is known from disease registries
- Other times, it is easy to place bounds on it

Simulation

- G binary: pr(G=1) = 0.065 and = 0.26
- X = min{10,lognormal(0,1)}
- 500 cases and 500 controls

$$\begin{aligned}&\mathbf{logit}\left\{\mathbf{pr}(\mathbf{D} = \mathbf{1}|\mathbf{X}, \mathbf{G})\right\} \\ &\quad = \boldsymbol{\beta_0} + \mathbf{0.10X} + \mathbf{0.26G} + \mathbf{0.3XG}\end{aligned}$$

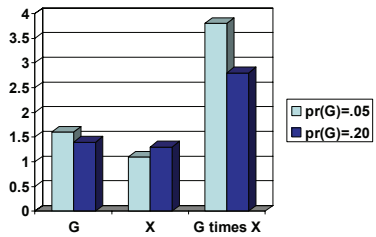
- Here $\boldsymbol{\beta_0}$ was manipulated so that pr(D=1) = 0.05

Simulation

- We assumed that we know a priori that 0.02 < pr(D=1) < 0.07
- This is the kind of information that is routinely available in Western countries
- We computed the mean squared error (MSE) efficiency of our method compared to ordinary logistic regression

Simulation

- MSE Efficiency of Profile method compared to ordinary logistic regression



Computation

- For the case that G is binary, Matlab code is at http://www.stat.tamu.edu
- That code has some options related to model robustness, to be discussed in the subsequent lectures
- It is very simple, just needs an optimizer
- Experts in R could convert the code in about 20 minutes

Rare Disease

- In rare disease cases, some of the formulae simplify
- Remember that we have

$$\mathbf{S}(\mathbf{d},\mathbf{x},\mathbf{g},\boldsymbol{\Theta}) = \frac{\mathbf{q}(\mathbf{g},\theta)\exp\left[\mathbf{d}\left\{\beta_0^* + \mathbf{m}(\mathbf{x},\mathbf{g},\boldsymbol{\beta}_1)\right\}\right]}{1+\exp\left\{\beta_0 + \mathbf{m}(\mathbf{x},\mathbf{g},\boldsymbol{\beta}_1)\right\}}$$

- In rare disease cases, we have treated the denominator as if it were = 1
- In this case β_0 disappears and cannot be estimated

Rare Disease

- In rare disease cases

$$\mathbf{S}_{\text{rare}}(\mathbf{d},\mathbf{x},\mathbf{g},\boldsymbol{\theta},\boldsymbol{\beta_0^*},\boldsymbol{\beta_1}) = \mathbf{q}(\mathbf{g},\boldsymbol{\theta})\mathbf{exp}\left[\mathbf{d}\left\{\boldsymbol{\beta_0^*} + \mathbf{m}(\mathbf{x},\mathbf{g},\boldsymbol{\beta_1})\right\}\right]$$

- Profile likelihood then is

$$\mathcal{L}_{\text{rare}}(\mathbf{D},\mathbf{X},\mathbf{G},\boldsymbol{\theta},\boldsymbol{\beta_0^*},\boldsymbol{\beta_1}) = \frac{\mathbf{q}(\mathbf{G},\boldsymbol{\theta})\mathbf{exp}\left[\mathbf{D}\left\{\boldsymbol{\beta_0^*} + \mathbf{m}(\mathbf{X},\mathbf{G},\boldsymbol{\beta_1})\right\}\right]}{\sum_{\mathbf{d}=0}^{\mathbf{1}}\sum_{\mathbf{s}=0}^{\mathbf{1}}\mathbf{q}(\mathbf{s},\boldsymbol{\theta})\mathbf{exp}\left[\mathbf{d}\left\{\boldsymbol{\beta_0^*} + \mathbf{m}(\mathbf{X},\mathbf{s},\boldsymbol{\beta_1})\right\}\right]}$$

STATISTICS

Gene-Environment Case-Control Studies

Raymond J. Carroll
Department of Statistics
Faculties of Nutrition and Toxicology

Texas A&M University
http://stat.tamu.edu/~carroll

STATISTICS
TEXAS A&M UNIVERSITY

Lecture 3: Outline

- Profile Likelihood as a real likelihood
- First Generalizations
  - G not binary
  - G modeled as a function of X
- Missing genetic data
- Analysis of the Israeli Study Data

STATISTICS

Real Data Complexities

- In the Israeli Study, G is missing in 50% of the controls, and 10% of the cases
- Also, among Jewish citizens, Israel has two dominant ethnic types
  - Ashkenazi (European)
  - Shephardic (North African)

STATISTICS

Real Data Complexities

- The gene mutation BRCA1/2 if frequent among the Ashkenazi, but rare among the Shephardic
- Thus, if one component of X is ethnic status, then pr(G=1 | X) depends on X
- Gene-Environment independence fails here
- What can be done?

STATISTICS

Solution

- What we will do is
  - Build a model for [G | X]
  - Allow missing G data
- Then we will use our two main tools to construct a profile likelihood
  - Pretend Missing Data formulation
  - Formal retrospective likelihood formulation

Model Building and Missing Genes

- Consider a general model:

$$\mathbf{pr}(\mathbf{G}=\mathbf{g}|\mathbf{X})=\mathbf{q}(\mathbf{g}|\mathbf{X},\boldsymbol{\theta})$$

- Allow missing data: you observe $\mathcal{G}$ and you know that $\mathbf{G}\in\mathcal{G}$
- If G is a simple genotype and it is missing, then $\mathcal{G}=(\mathbf{0},\mathbf{1})$
- If G is observed, then $\mathbf{G}=\mathcal{G}$

Model Building and Missing Genes

- We will consider missing genetic status to be missing at random (MAR)

$$\mathbf{pr}(\mathbf{G}\ \mathbf{missing}\ |\mathbf{D},\mathbf{X},\mathbf{G})=\mathbf{pr}(\mathbf{G}\ \mathbf{missing}\ |\mathbf{D},\mathbf{X})$$

- Use known facts that MAR means that you can ignore the missing data mechanism in computing likelihoods

Pretend Data Formulation

- In our previous pretend calculations, we computed

$$\mathbf{pr}(\mathbf{D}=\mathbf{d},\mathbf{G}=\mathbf{g}|\mathbf{X},\delta=\mathbf{1})$$

- Since G can be missing and we observe only $\mathcal{G}$ the pretend likelihood is

$$\mathbf{pr}(\mathbf{D}=\mathbf{d},\mathbf{G}\in\mathcal{G}|\mathbf{X},\delta=\mathbf{1})=\sum_{\mathbf{g}\in\mathcal{G}}\underbrace{\mathbf{pr}(\mathbf{D}=\mathbf{d},\mathbf{G}=\mathbf{g}|\mathbf{X},\delta=\mathbf{1})}_{\text{An old friend}}$$

Pretend Data Formulation

- This means that the pretend profile likelihood is simply

$$\mathcal{L}_{\text{miss}}(\mathbf{D},\mathbf{X},\mathcal{G},\boldsymbol{\Theta}) = \frac{\sum_{\mathbf{g}\in\mathcal{G}} \mathbf{S}(\mathbf{D},\mathbf{X},\mathbf{g},\boldsymbol{\Theta})}{\sum_{\mathbf{d}=0}^{\mathbf{1}}\sum_{\mathbf{s}=0}^{\mathbf{1}} \mathbf{S}(\mathbf{d},\mathbf{X},\mathbf{g},\boldsymbol{\Theta})} \qquad (4)$$

- Nothing could be simpler, if we believe in pretend studies

Semiparametric Calculations

- The observed data are $(\mathbf{D},\mathbf{X},\mathcal{G})$
- The retrospective likelihood function is

$$\mathbf{pr}(\mathbf{X}=\mathbf{x},\mathbf{G}\in\mathcal{G}|\mathbf{D}=\mathbf{d}) = \frac{\mathbf{f}_{\mathbf{X}}(\mathbf{x})\sum_{\mathbf{g}\in\mathcal{G}}\mathbf{q}(\mathbf{g}|\mathbf{x},\boldsymbol{\theta})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X}=\mathbf{x},\mathbf{G}=\mathbf{g})}{\sum_{\mathbf{d}=0}^{\mathbf{1}}\sum_{\mathbf{s}}\int \mathbf{q}(\mathbf{s}|\mathbf{z},\boldsymbol{\theta})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X}=\mathbf{z},\mathbf{G}=\mathbf{s})\mathbf{f}_{\mathbf{X}}(\mathbf{z})\mathbf{d}\mu(\mathbf{z})}$$

Semiparametric Profile Likelihood

- Again, if X is discrete, we would write it as

$$\mathbf{pr}(\mathbf{X}=\mathbf{x}_{\mathbf{m}},\mathbf{G}\in\mathcal{G}|\mathbf{D}=\mathbf{d}) = \frac{\gamma_{\mathbf{m}}\sum_{\mathbf{g}\in\mathcal{G}}\mathbf{q}(\mathbf{g}|\mathbf{x},\boldsymbol{\theta})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X}=\mathbf{x},\mathbf{G}=\mathbf{g})}{\sum_{\mathbf{d}=0}^{\mathbf{1}}\sum_{\mathbf{s}}\sum_{\mathbf{j}=1}^{\mathbf{M}}\gamma_{\mathbf{j}}\mathbf{q}(\mathbf{s}|\mathbf{z},\boldsymbol{\theta})\mathbf{pr}(\mathbf{D}=\mathbf{d}|\mathbf{X}=\mathbf{z},\mathbf{G}=\mathbf{s})}$$

- The calculations are very similar to the ones we have done before.
- The semiparametric profile likelihood is (4)

Inference and Profile Likelihood

- It is easy to do the calculations to show that profile likelihood leads to consistent and asymptotically normal estimators
- I will simply outline the steps when G is observable
- If G is partially missing and we only observe $\mathcal{G}$ then there are only minor changes of notation

Inference and Profile Likelihood

- The biggest technical issue is that it is not true that the individual likelihood scores have mean zero, i.e., it is not true that

$$\mathbf{0} = \mathbf{E}\left\{\frac{\partial}{\partial \boldsymbol{\Theta}}\mathcal{L}(\mathbf{D}, \mathbf{X}, \mathbf{G}, \boldsymbol{\Theta})|\mathbf{D}\right\}$$

- This issue also arises for ordinary logistic regression

Profile Likelihood: General X

- What is true instead is that the likelihood score for the entire case-control study has mean zero, i.e.,

$$\mathbf{0} = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{E}\left\{\frac{\partial}{\partial \boldsymbol{\Theta}}\mathcal{L}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i}, \boldsymbol{\Theta})|\mathbf{D_i}\right\} \quad (5)$$

- We showed this for ordinary logistic regression back in Lecture 1. Much the same argument applies.

Profile Likelihood: General X

- We have that for any function R(D,X,G),

$$\mathbf{n}^{-1}\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{E}\left\{\mathbf{R}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i})|\mathbf{D_i}\right\}$$
$$= \sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}} \frac{\mathbf{n_0}}{\mathbf{n}} \mathbf{E}\left\{\mathbf{R}(\mathbf{D}, \mathbf{X}, \mathbf{G})|\mathbf{D} = \mathbf{d}\right\}$$

- Suppose that G has the values $(\mathbf{g_1}, \ldots, \mathbf{g_K})$

Profile Likelihood: General X

- Then it follows by simple, tedious algebra that

$$\sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}} \frac{\mathbf{n_0}}{\mathbf{n}} \mathbf{E}\left\{\mathbf{R}(\mathbf{D}, \mathbf{X}, \mathbf{G})|\mathbf{D} = \mathbf{d}\right\} \quad (6)$$
$$= \frac{\mathbf{n_0}}{\mathbf{n}\pi_0}\int \mathbf{f_X}(\mathbf{x}) \sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}}\sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{K}} \mathbf{R}(\mathbf{d}, \mathbf{x}, \mathbf{g_k})\mathbf{S}(\mathbf{d}, \mathbf{x}, \mathbf{g_k}, \boldsymbol{\Theta})\mathbf{d}\mu(\mathbf{x})$$

- This identity and more tedious algebra shows that the likelihood score for all the data has mean zero.

Consistency and Inference

- The preceding argument and the WLLN proves consistency.
- What about inference?
- We are solving

$$\mathbf{0} = \mathbf{n}^{-1} \sum_{\mathbf{i}=1}^{\mathbf{n}} \frac{\partial}{\partial \boldsymbol{\Theta}} \mathcal{L}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i}, \widehat{\boldsymbol{\Theta}})$$

Consistency and Inference

- Define the usual information matrix and its counterpart

$$\boldsymbol{\mathcal{I}} = \mathbf{n}^{-1} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{E}\left\{ \frac{\partial^2}{\partial \boldsymbol{\Theta} \boldsymbol{\Theta}^\top} \mathcal{L}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i}, \boldsymbol{\Theta}) | \mathbf{D_i} \right\}$$

$$\boldsymbol{\mathcal{I}}_* = \mathbf{cov}\left[\mathbf{n}^{-1/2} \sum_{\mathbf{i}=1}^{\mathbf{n}} \left\{ \frac{\partial}{\partial \boldsymbol{\Theta}} \mathcal{L}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i}, \boldsymbol{\Theta}) | \mathbf{D_i} \right\} \right]$$

Consistency and Inference

- By standard linear expansions

$$\mathbf{n}^{1/2}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \Rightarrow \mathbf{Normal}(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{I}}_* \boldsymbol{\mathcal{I}}^{-1})$$

- Using the fundamental identity (6), it is easy to calculate $\boldsymbol{\mathcal{I}}_*$
- Define

$$\begin{aligned} \mathbf{c}(\mathbf{d}) &= \mathbf{E}\left\{ \frac{\partial}{\partial \boldsymbol{\Theta}} \mathcal{L}(\mathbf{D}, \mathbf{X}, \mathbf{G}, \boldsymbol{\Theta}) | \mathbf{D} = \mathbf{d} \right\} \\ \boldsymbol{\Lambda} &= \sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{1}} \frac{\mathbf{n_d}}{\mathbf{n}} \mathbf{c}(\mathbf{d}) \mathbf{c}^\top(\mathbf{d}) \end{aligned}$$

Consistency and Inference

- Then $\boldsymbol{\mathcal{I}}_* = \boldsymbol{\mathcal{I}} - \boldsymbol{\Lambda}$ and

$$\mathbf{n}^{1/2}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \Rightarrow \mathbf{Normal}(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1} - \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\Lambda} \boldsymbol{\mathcal{I}}^{-1})$$

- The curious part of this is that if you just used the inverse of the observed Fisher information, at least in principle your standard errors would be too large!
- In practice, however, $\boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\Lambda} \boldsymbol{\mathcal{I}}^{-1} \approx \mathbf{0}$

Inference

- In various complex settings, the profile likelihoods $\mathcal{L}_{\text{miss}}(\mathbf{D}, \mathbf{X}, \mathcal{G}, \Theta)$ and $\mathcal{L}(\mathbf{D}, \mathbf{X}, \mathbf{G}, \Theta)$ act numerically like likelihood functions
- Likelihood ratio tests applied to the profile likelihoods have very good numeric properties

Israeli Ovarian Cancer Study

- Population based case-control study
- Study the interplay of BRCA1/2 mutations (G) and two known risk factors (E or X) of ovarian cancer:
  - oral contraceptive (OC) use
  - Parity (number of children)
- **Missing Data**: Approximately 50% of the controls were not genotyped, and 10% of the cases

Israeli Ovarian Cancer Study

- Results reported in Modan et al., NEJM (2001).
- **Their analysis** involves
  - Assumption of parity and OC use are independent of BRCA1/2 mutation status
  - Risk model adjusted for
    - Age
    - Ashkenazi or Shepharic (S)
    - Other risk factors

Israeli Ovarian Cancer Study

- Let S = 1 for Ashkenazi, = 0 otherwise
- S is part of X
- Because the incidence of the BRCA1/2 mutation depends on S, we built a model

$$\mathbf{logit}\left\{\mathbf{pr}(\mathbf{G}=\mathbf{1}|\mathbf{X})\right\} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1\mathbf{S}$$

- This is what we have called $\mathbf{q}(\mathbf{g}|\mathbf{X}, \theta)$

Israeli Ovarian Cancer Study

- Denote X as consisting of
  - Age (A)
  - Ashkenazi status (S)
  - Oral contraceptive use (O)
  - Parity (P)
- The final risk model has

$$\begin{aligned}\mathbf{logit}\left\{\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\mathbf{G})\right\} &= \beta_0+\beta_1\mathbf{A}+\beta_2\mathbf{S}\\ &+\beta_3\mathbf{O}+\beta_4\mathbf{P}+\beta_5\mathbf{G}\\ &+\beta_6\mathbf{O}\cdot\mathbf{G}+\beta_7\mathbf{P}\cdot\mathbf{G}\end{aligned}$$

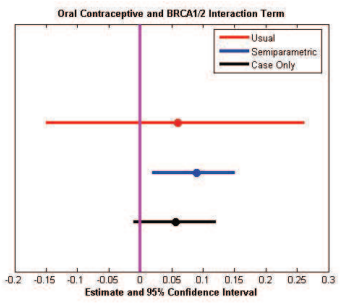
Israeli Ovarian Cancer Study

- The relative risk for BRCA1/2 is $\exp(\beta_5)$
- The interaction terms have relative risks $\exp(\beta_6)$ and $\exp(\beta_7)$
- Among carriers of the mutation (G=1), the relative risk for oral contraceptive use is
$$\exp(\beta_3+\beta_6)$$
- The hypothesis is that this last quantity is < 1, i.e., OC use is protective

$$\begin{aligned}\mathbf{logit}\left\{\mathbf{pr}(\mathbf{D}=\mathbf{1}|\mathbf{X},\mathbf{G})\right\} &= \beta_0+\beta_1\mathbf{A}+\beta_2\mathbf{S}\\ &+\beta_3\mathbf{O}+\beta_4\mathbf{P}+\beta_5\mathbf{G}\\ &+\beta_6\mathbf{O}\cdot\mathbf{G}+\beta_7\mathbf{P}\cdot\mathbf{G}\end{aligned}$$

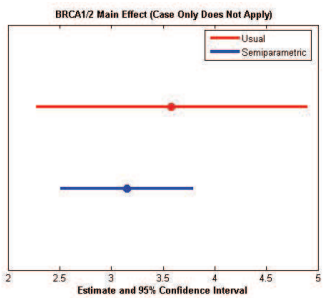
Israeli Ovarian Cancer Study

- **Interaction of OC and BRCA1/2**:



Israeli Ovarian Cancer Study

- **Main Effect of BRCA1/2**:



Israeli Ovarian Cancer Study

- Odds ratio for OC use among carriers **= 1.04 (0.98, 1.09)**
  - No evidence for protective effect
  - Not available from case-only analysis
  - **Length of interval is ½ the length of the usual analysis**

Robustness

- In the simplest case, we assume independence between G and X in the population
- More generally, we assume a model for G given X
- Some model-robustness is obtained by fitting more or less complex models

Robustness

- In the Israeli study, we fit many model for G given X

Gene-Environment Case-Control Studies

Raymond J. Carroll
Department of Statistics
Faculties of Nutrition and Toxicology

Texas A&M University
http://stat.tamu.edu/~carroll

STATISTICS
TEXAS A&M UNIVERSITY

Lecture 4: Outline

- Issue of Robustness
- Straightforward Methods
  - Chi-squared tests for bias
- Model selection strategies
  - Open Problems

STATISTICS

Interactions

- We have shown that simple models can lead to much greater efficiency in understanding interactions
- In general, interactions are hard to understand
- More data for "*nothing*"?

STATISTICS

Interactions

- A lot of people worry that the gain in efficiency is simply due to the model
- Well yes, of course!
- But this does not mean that the model is very badly wrong

STATISTICS

Robustness

- Our methods are based on a model for G given X
  - One can build full models
  - One can build nonparametric models given strata
  - One can build independence given strata
- There are a host of possible models
- Nonetheless, assumptions are being made to get gains in efficiency
- How can we protect ourselves against model misspecification?

Robustness Warning

- One can get out of hand worrying about model robustness
- Remember, if we make no assumptions about how G and X are related, the best we can do is to use ordinary logistic regression
- My best analogy: linear regression versus full nonparametric regression

Robustness Warning

- My best analogy: linear regression versus full nonparametric regression
- A better analogy:
  - Linear regression
  - Semiparametric regression
  - Nonparametric regression
- We need realistic models that build **some** robustness, not **total** robustness

Chi-squared Tests

- There is some information about model robustness in the data
- Ordinary logistic regression (OLR) makes no assumptions
- Major (statistically significant) discrepancies between our method and OLR are evidence

Chi-squared Tests

- More specifically, assume G is observable
- Ordinary logistic regression (OLR) differentiates the loglikelihood and solves an estimating equation

$$\mathbf{0} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \psi_{\mathrm{OLR}}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i}, \widehat{\boldsymbol{\beta}}^{*}_{\mathbf{0},\mathrm{OLR}}, \widehat{\boldsymbol{\beta}}_{\mathbf{1},\mathrm{OLR}})$$

- Of course, in the limit, OLR estimates

$$(\boldsymbol{\beta}^{*}_{\mathbf{0}}, \boldsymbol{\beta}_{\mathbf{1}})$$

Chi-squared Tests

- Our Semiparametric logistic regression (SLR) differentiates the semiparametric loglikelihood and solves an estimating equation

$$\mathbf{0} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \psi_{\mathrm{SLR}}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i}, \widehat{\boldsymbol{\beta}}_{\mathbf{0},\mathrm{SLR}}, \widehat{\boldsymbol{\beta}}_{\mathbf{1},\mathrm{SLR}}, \widehat{\theta}_{\mathrm{SLR}}, \widehat{\pi}_{\mathbf{1},\mathrm{SLR}})$$

- Of course, in the limit, if the SLR specifications are correct, then SLR estimates

$$(\boldsymbol{\beta}_{\mathbf{0}}, \boldsymbol{\beta}_{\mathbf{1}}, \theta, \pi_{\mathbf{1}})$$

Chi-squared Tests

- In general, however, allowing for bias the best we can say is that SLR estimates

$$(\boldsymbol{\beta}_{\mathbf{0},\mathrm{SLR}}, \boldsymbol{\beta}_{\mathbf{1},\mathrm{SLR}}, \theta_{\mathrm{SLR}}, \pi_{\mathbf{1},\mathrm{SLR}})$$

- What we want to do is to test whether

$$\mathbf{H_0} : \boldsymbol{\beta}_{\mathbf{1}} = \boldsymbol{\beta}_{\mathbf{1},\mathrm{SLR}}$$

Chi-squared Tests

- This suggests solving both simultaneously. Let

$$\boldsymbol{\Omega} = (\boldsymbol{\beta}^{*}_{\mathbf{0}}, \boldsymbol{\beta}_{\mathbf{1}}, \boldsymbol{\beta}_{\mathbf{0},\mathrm{SLR}}, \boldsymbol{\beta}_{\mathbf{1},\mathrm{SLR}}, \theta_{\mathrm{SLR}}, \pi_{\mathbf{1},\mathrm{SLR}})$$

- Create the estimating function

$$\boldsymbol{\Psi}(\mathbf{D}, \mathbf{X}, \mathbf{G}, \boldsymbol{\Omega}) = \left\{ \begin{array}{c} \psi_{\mathrm{OLR}}(D, X, G, \boldsymbol{\beta}^{*}_{0}, \boldsymbol{\beta}_{1}) \\ \psi_{\mathrm{SLR}}(D, X, G, \boldsymbol{\beta}_{0,\mathrm{SLR}}, \boldsymbol{\beta}_{1,\mathrm{SLR}}, \theta_{\mathrm{SLR}}, \pi_{1,\mathrm{SLR}}) \end{array} \right\}$$

Chi-squared Tests

- Then solving both simultaneously:

$$\mathbf{0} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{\Psi}(\mathbf{D_i}, \mathbf{X_i}, \mathbf{G_i}, \widehat{\mathbf{\Omega}})$$

- This is just an M-estimator, and we know that we can show that, conservatively

$$\mathbf{n^{1/2}} \left(\widehat{\mathbf{\Omega}} - \mathbf{\Omega} \right) \Rightarrow \text{Normal}(\mathbf{0}, \mathbf{\Sigma})$$

Chi-squared Tests

- Thus

$$\mathbf{\Omega} = (\beta_0^*, \boldsymbol{\beta_1}, \boldsymbol{\beta}_{\mathbf{0},\text{SLR}}, \boldsymbol{\beta}_{\mathbf{1},\text{SLR}}, \theta_{\text{SLR}}, \pi_{\mathbf{1},\text{SLR}})$$

$$\mathbf{n^{1/2}} \left(\widehat{\mathbf{\Omega}} - \mathbf{\Omega} \right) \Rightarrow \text{Normal}(\mathbf{0}, \mathbf{\Sigma})$$

- Hence, to test whether the semiparametric MLE is affected by bias, we just want to test whether

$$\mathbf{H_0} : \boldsymbol{\beta_1} = \boldsymbol{\beta}_{\mathbf{1},\text{SLR}}$$

Chi-squared Tests

- Thus

$$\mathbf{\Omega} = (\beta_0^*, \boldsymbol{\beta_1}, \boldsymbol{\beta}_{\mathbf{0},\text{SLR}}, \boldsymbol{\beta}_{\mathbf{1},\text{SLR}}, \theta_{\text{SLR}}, \pi_{\mathbf{1},\text{SLR}})$$

$$\mathbf{H_0} : \boldsymbol{\beta_1} = \boldsymbol{\beta}_{\mathbf{1}.\text{SLR}}$$

- This is a simple contrast, can be tested via a chi-squared test.
- You can also test whether specific components are biased

Chi-squared Tests

- Of course, $\mathbf{\Sigma}$ can be estimated by the bootstrap if one is so inclined
- You have to be careful to maintain the case-control study
- Bootstrap the cases and the controls separately

Model Selection/Averaging

- It is possible to go straight at the problem with model selection and averaging strategies
- Suppose G is binary. Split up the other covariates:
  - Strata S
  - Major possible effects Z
  - All other stuff W

Model Selection/Averaging

- We would write down a model given strata and obvious things:

$$\mathbf{pr(G=g|S,Z,W)=H(\alpha_0+S\alpha_1+Z\alpha_2)}$$

- A more general model is also possible but loses efficiency

$$\mathbf{pr(G=g|S,Z,W)=H(\alpha_0+S\alpha_1+Z\alpha_2+W\alpha_3)}$$

Model Selection/Averaging

- A more general model is also possible but loses efficiency

$$\mathbf{pr(G=g|S,Z,W)=H(\alpha_0+S\alpha_1+Z\alpha_2)}$$

$$\mathbf{pr(G=g|S,Z,W)=H(\alpha_0+S\alpha_1+Z\alpha_2+W\alpha_3)}$$

- Should you include W into the model?
- This is an issue of model selection

Model Selection/Averaging

- It would seem that a reasonable approach would be to build a big logistic model involving main effects and interactions
- Then apply model selection/averaging strategies, e.g., LASSO, etc.
- This has not been tried

Model Selection/Averaging

- Given the indirect nature of the information about G that is involved, I think pretty good model robustness can be achieved by simple models

STATISTICS

Gene-Environment Case-Control Studies

Raymond J. Carroll
Department of Statistics
Faculties of Nutrition and Toxicology

Texas A&M University
http://stat.tamu.edu/~carroll

STATISTICS
TEXAS A&M UNIVERSITY

Lecture 5: Outline

- Haplotypes
  - What are they
  - As a missing data problem
- Analysis
  - Straightforward
  - Via strata
  - More general
  - Model robustness

STATISTICS

Haplotypes

- Most people do not have single gene disorders
- This has motivated disease modeling via haplotypes
- Remember, we have a pair of chromosomes

STATISTICS

Haplotypes

- From Wikipedia:
  - A **haplotype** (Greek *haploos* = simple) is the genetic constitution of an individual chromosome
- Think of this as knowing
  - The genetics you have inherited from your mother
  - The genetics you have inherited from your father

STATISTICS

SNP

- A SNP (single nucleotide polymorphism) is a site in the DNA
- At a site, we might inherit the base pair $\mathbf{A_m}$ from our mother, and a different base pair $\mathbf{a_f}$ from our father
- Our genotypes for that site can be one of ($\mathbf{A_m}$,$\mathbf{A_f}$), ($\mathbf{A_m}$,$\mathbf{a_f}$), ($\mathbf{a_m}$,$\mathbf{A_f}$) or ($\mathbf{a_m}$, $\mathbf{a_f}$)

SNP

- The data cannot tell us what comes from mother and what comes from father, so the unorderd genotypes that we can observe at a site are (**A**,**A**), (**A**,**a**) or (**a**,**a**)
- When we dealt with BRCA1/2, this means that there was an **a** at either the site for BRCA1 or the site for BRCA2, or both

Haplotypes

- Haplotypes consist of what we get from our mother and father at more than one site
- Mother gives us the haplotype h_m = ($\mathbf{A_m}$,$\mathbf{B_m}$)
- Father gives us the haplotype h_f = ($\mathbf{a_f}$,$\mathbf{b_f}$)
- Our diplotype is $\mathbf{H^{dip}}$ = {($\mathbf{A_m}$,$\mathbf{B_m}$), ($\mathbf{a_f}$,$\mathbf{b_f}$)}

Haplotypes

- Our diplotype is $\mathbf{H^{dip}}$ = {($\mathbf{A_m}$,$\mathbf{B_m}$), ($\mathbf{a_f}$,$\mathbf{b_f}$)}
- Various genetic models are phrased in the number of copies of a particular haplotype we inherit.
- In the example above, if we are worried about the (**a**,**b**) haplotype, we inherit one copy

Haplotypes

- If
  - We could observe the two haplotypes inherited
  - We modeled in terms of the number of specific haplotypes inherited
- Then all the previous lectures apply
- With haplotype-environment independence within strata, one can obtain huge increases in efficiency by semiparametric MLE

Haplotypes

- Unfortunately, we cannot presently observe the two haplotypes
- We can only observe genotypes
- Thus, if we were really $H^{dip} = \{(A_m,B_m), (a_f,b_f)\}$, then the data we would see would simply be the unordered set (**A,a,B,b**)

Missing Haplotypes

- Thus, if we were really $H^{dip} = \{(A_m,B_m), (a_f,b_f)\}$, then the data we would see would simply be the unordered set (**A,a,B,b**)
- However, this is also consistent with a different diplotype, namely $H^{dip} = \{(a_m,B_m), (A_f,b_f)\}$
- Note that the number of copies of the (**a,b**) haplotype differs in these two cases
- The true diploid = haplotype pair is missing

Missing Haplotypes

- There are cases where the number of a particular haplotype is known.
- Suppose we are interested in the number of haplotype pairs that are (**a,b**)
- If the unordered set is (**a,a,B,b**), then there is one such haplotype
- If the unordered set is (**a,a,b,b**), then there are two such haplotypes

Missing Haplotypes

- There are three basic approaches to the problem of unordered (unphased) haplotypes
  - By fancy Bayesian model based on a population genetics model, infer the structure of the diplotypes via their posterior probabilities
    - Then pretend you know the truth
  - Using a population genetics model, treat the missing haplotype data as missing data
  - Use only the unordered data for which the number of copies of a haplotype is known

Haplotypes

- In general, if there are m SNP, then there are

$$2^{m} \text{ haplotypes}$$

$$2^{2m} \text{ diplotypes}$$

- It is typical to model only the most common haplotype effects, collapsing those that have < 1% frequency into a single category

Haplotypes

- There is a nice EM-based program for estimating haplotype frequencies

http://www.bios.unc.edu/~lin/hapstat/

- It accepts data in text format with SAS missing data conventions
- There is also an R package haplo.core

Haplotypes

- Generally, the possible genotypes are labelled as
  - a = most common
  - A = least common
- Then the codes for a single genotype are
  - aa = 0
  - Aa = 1
  - AA = 2
- Missing genotypes are labeled as "."

Haplotype Frequency Estimation

- Suppose there are 5 SNPs
- Then the 32 haplotypes are 00000, 00001, …, 11111
- The data are entered as follows

Haplotype Frequency Estimation

- Status = case-control status, 1 = case

Status	Age	Gender	SNP1	SNP2	SNP3	SNP4	SNP5
1	48	0	2	1	0	2	2
1	49	0	1	2 .		2 .	
1	40	0	0	2	2	0 .	
1	44	1 .		1	1	1	1
1	24	0	1	1	1	1	1
1	48	1	0	2	1	1 .	
1	48	1	2	0	0 .		2
1	36	1	0	2	2	0	0
1	48	1	0	2	2	0	0
1	44	1	0	2	1	1	1
1	22	0 .		2	2	0	0
1	43	0	2	0	2	0	0
1	22	0	2	0	0	2 .	
1	30	0	1	1	1	0	0
1	36	1	0 .		1	1	1
1	44	0	0	2	2	0 .	
1	41	1	2	1	0	2	2
1	48	1	1	2 .		2	2
1	50	0	1	1	1	1 .	
1	26	0	2	0 .		2	2

Haplotype Frequency Estimation

- The program is flexible, but for example it assumes Hardy-Weinberg equilibrium (HWE)
- Suppose there are K haplotypes. Define

$$\mathbf{pr}(\mathbf{haplotype} \ = \mathbf{h_k}) = \boldsymbol{\pi_k}$$

- Then HWE means that

$$\mathbf{pr}\left\{\mathbf{H}^{\mathrm{dip}} = (\mathbf{h_j}, \mathbf{h_k})\right\} = \boldsymbol{\pi}_{\mathbf{jk}}^{\mathrm{dip}} = \boldsymbol{\pi_j \pi_k}$$

Haplotype Frequency Estimation

- The results of an EM algorithm are

Subjects: 2000 Cases: 1000 Controls: 1000

Haplotype	Controls	Cases	Combined
00000	0.0000	0.0000	0.0000
00001	0.0000	0.0000	0.0000
00010	0.0000	0.0000	0.0000
00011	0.0057	0.0019	0.0040
00100	0.0054	0.0039	0.0045
00101	0.0000	0.0000	0.0000
00110	0.0010	0.0000	0.0005
00111	0.0000	0.0000	0.0000
01000	0.0000	0.0000	0.0000
01001	0.0000	0.0000	0.0000
01010	0.0000	0.0000	0.0000
01011	0.1375	0.1129	0.1249
01100	0.2501	0.3490	0.2998
01101	0.0005	0.0006	0.0006
01110	0.0000	0.0000	0.0000
01111	0.0020	0.0025	0.0022
10000	0.0140	0.0079	0.0109
10001	0.0000	0.0000	0.0000
10010	0.0000	0.0000	0.0000
10011	0.3338	0.3231	0.3287
10100	0.0629	0.0514	0.0570
10101	0.0000	0.0000	0.0000
10110	0.0363	0.0195	0.0278
10111	0.0000	0.0000	0.0000
11000	0.0000	0.0000	0.0000
11001	0.0000	0.0000	0.0000
11010	0.0000	0.0000	0.0000
11011	0.1390	0.1178	0.1283
11100	0.0088	0.0089	0.0090
11101	0.0000	0.0000	0.0000
11110	0.0000	0.0000	0.0000
11111	0.0029	0.0006	0.0018

There are only 6 haplotypes that have frequency greater than 1.1%

They would be used for risk modeling, using genetic risk models

Various brands of model selection might be used

Haplotype Frequency Estimation

- The results of an EM algorithm are

Subjects: 2000 Cases: 1000 Controls: 1000

Haplotype	Controls	Cases	Combined
00000	0.0000	0.0000	0.0000
00001	0.0000	0.0000	0.0000
00010	0.0000	0.0000	0.0000
00011	0.0057	0.0019	0.0040
00100	0.0054	0.0039	0.0045
00101	0.0000	0.0000	0.0000
00110	0.0010	0.0000	0.0005
00111	0.0000	0.0000	0.0000
01000	0.0000	0.0000	0.0000
01001	0.0000	0.0000	0.0000
01010	0.0000	0.0000	0.0000
01011	0.1375	0.1129	0.1249
01100	0.2501	0.3490	0.2998
01101	0.0005	0.0006	0.0006
01110	0.0000	0.0000	0.0000
01111	0.0020	0.0025	0.0022
10000	0.0140	0.0079	0.0109
10001	0.0000	0.0000	0.0000
10010	0.0000	0.0000	0.0000
10011	0.3338	0.3231	0.3281
10100	0.0629	0.0514	0.0570
10101	0.0000	0.0000	0.0000
10110	0.0363	0.0195	0.0278
10111	0.0000	0.0000	0.0000
11000	0.0000	0.0000	0.0000
11001	0.0000	0.0000	0.0000
11010	0.0000	0.0000	0.0000
11011	0.1390	0.1178	0.1283
11100	0.0088	0.0089	0.0090
11101	0.0000	0.0000	0.0000
11110	0.0000	0.0000	0.0000
11111	0.0029	0.0006	0.0018

It is standard in data like this to combine the rare haplotypes and the most common one into a reference for risk modeling.

This means that they are absorbed into the intercept in the logistic regression

Haplotype Modeling

- There are many possible models available
- In the example, there are 5 non-referent haplotypes
- Order them as $(\mathbf{h_1}, \ldots, \mathbf{h_5})$
- In an additive model, having two copies of one of these haplotypes doubles the effect, i.e.,

$$\mathbf{pr}\{\mathbf{D} = \mathbf{1} | \mathbf{H}^{\mathrm{dip}} = (\mathbf{h_j}, \mathbf{h_k})\}$$
$$= \mathbf{H}\left[\beta_\mathbf{0} + \sum_{\ell=\mathbf{1}}^{\mathbf{5}} \beta_\ell \{\mathbf{I}(\mathbf{h_j} = \mathbf{h}_\ell) + \mathbf{I}(\mathbf{h_k} = \mathbf{h}_\ell)\}\right]$$

Haplotype Modeling

- These models then also include main effects for X and interaction terms
- There are a **great many** possible models!
- Biological background can help cut down on the number of possible models
- EM algorithms are used to fit the models

Haplotype Fitting

- Models that assume haplotype-environment independence are straightforward to fit
- The remaining issue is how to gain robustness against deviations from this assumed independence
- The hard part is that unlike single genotypes, haplotype modeling needs dimension reduction

Robustness

- We build robustness by specifying models for diplotype given the environmental variables
- We first run a program to get a preliminary estimate of haplotype frequency
- We use the most frequent haplotype as a reference haplotype (similar to a control)

Haplotypes

- **Approach**: Start with a logistic model for the unobserved haplotypes H given covariates X

$$\log\left\{\frac{pr\left\{H^{dip}=(h_j,h_k)\,|\,X\right\}}{pr\left\{H^{dip}=(h_{ref},h_{ref})\,|\,X\right\}}\right\}=\gamma_{0jk}+\gamma_{1jk}X$$

- Also, to get the intercepts above, impose HWE at the population level

$$pr\left\{H^{dip}=(h_j,h_k)\right\}=\pi_{jk}^{dip}=\pi_j\pi_k$$

Haplotypes Analysis

- The resulting method adds robustness
- EM-algorithms enable fast computation
- The method is semiparametric efficient

Haplotypes

- **Technical Model for Diplotypes**:

$$pr\left\{H^{di}=(h_j,h_k)\right\}=\gamma_{0jk}+\gamma_{1jk}X$$

subject to HWE

Thanks!

http://stat.tamu.edu/~carroll

STATISTICS
TEXAS A&M UNIVERSITY