Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 6:
Matching / Stratification

October 15, 2013

1

Lecture Outline

- Quiz
- Matching / Stratification vs Regression
- Directly Standardized Rates
  - Probability Models for Incidence of Disease
  - Example: Colorectal Cancer Incidence in US Whites

2

Quiz
(Pre-test and Survey)

U.S. Colorectal Cancer by Country of Birth

3

Matching / Stratification vs Regression

4

Recall: Adjustment for Covariates

- We "adjust" for other covariates
  - Model effect modification
  - Address confounding
  - Gain precision
- Define groups according to
  - Predictor of interest, and
  - Other covariates
- Compare the distribution of response across groups which
  - differ with respect to the Predictor of Interest, but
  - are the same with respect to the other covariates
    - "holding other variables constant"

5

Recall: Comparing models

Unadjusted $g\left[\theta \middle| X_i\right] = \beta_0 + \beta_1 \times X_i$

Adjusted $g\left[\theta \middle| X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

Science: When is $\gamma_1 = \beta_1$?

When is $\hat{\gamma}_1 = \hat{\beta}_1$?

Statistics: When is $se\left(\hat{\gamma}_1\right) = se\left(\hat{\beta}_1\right)$?

When is $s\hat{e}\left(\hat{\gamma}_1\right) = s\hat{e}\left(\hat{\beta}_1\right)$?

6

Recall: General Results

- These questions can not be answered precisely in the general case
- However, in linear regression we can derive exact results
- These will serve as a basis for later examination of
  - Logistic regression
  - Poisson regression
  - Proportional hazards regression

7

Recall: Linear Regression

- Difference in interpretation of slopes

Unadjusted Model: $E\left[Y_i \middle| X_i\right] = \beta_0 + \beta_1 \times X_i$

- β_1 = Diff in mean Y for groups differing by 1 unit in X
  - (The distribution of W might differ across groups being compared)

Adjusted Model: $E\left[Y_i \middle| X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

- γ_1 = Diff in mean Y for groups differing by 1 unit in X, but agreeing in their values of W

8

Recall: Relationships: True Slopes

- The slope of the unadjusted model will tend to be

$$\beta_1 = \gamma_1 + \rho_{XW}\frac{\sigma_W}{\sigma_X}\gamma_2$$

- Hence, true adjusted and unadjusted slopes for X are estimating the same quantity only if
  - $\rho_{XW} = 0$ (X and W are truly uncorrelated), OR
  - $\gamma_2 = 0$ (no association between W and Y after adjusting for X)

9

Recall: Relationships: True SE

Unadjusted Model $\left[se\left(\hat{\beta}_1\right)\right]^2 = \frac{Var\left(Y \mid X\right)}{nVar\left(X\right)}$

Adjusted Model $\left[se\left(\hat{\gamma}_1\right)\right]^2 = \frac{Var\left(Y \mid X, W\right)}{nVar\left(X\right)\left(1-r_{XW}^2\right)}$

$$Var\left(Y \mid X\right) = \gamma_2^2 Var\left(W \mid X\right) + Var\left(Y \mid X, W\right)$$

$$\sigma_{Y|X}^2 = \gamma_2^2\sigma_{W|X}^2 + \sigma_{Y|X,W}^2$$

10

Binary *W* : Notation

- We can use this notation to explore the benefits of matched analyses
- Suppose Y_{1i} measures "cases" having $Z_{1i} = 1$ and $W_{1i} = w_{1i} = 0,1$
  - Suppose n_{11} and n_{10} count the number with $W_{1i} = 1$ and $W_{1i} = 0$, respectively
- Suppose Y_{0i} measures "cases" having $Z_{1i} = 0$ and $W_{1i} = w_{1i} = 0,1$
  - Suppose n_{01} and n_{00} count the number with $W_{0i} = 1$ and $W_{0i} = 0$, respectively
- (Note: In the following I presume homoscedasticity
  - This will not generally be the case with binary data)

11

Binary *W* : Marginal Distribution

$$\left.\begin{array}{lll} Y_{0i} \mid W_{0i} = 0 & \sim & \left(\gamma_0,\ \sigma^2\right) \\ Y_{0i} \mid W_{0i} = 1 & \sim & \left(\gamma_0 + \gamma_2,\ \sigma^2\right) \end{array}\right\} \Rightarrow$$

$$Y_{0i} \sim \left(\gamma_0 + \frac{n_{01}}{n_{01}+n_{00}}\gamma_2,\ \sigma^2 + \gamma_2^2\frac{n_{01}n_{00}}{\left(n_{01}+n_{00}\right)^2}\right)$$

$$\left.\begin{array}{lll} Y_{1i} \mid W_{1i} = 0 & \sim & \left(\gamma_0 + \gamma_1,\ \sigma^2\right) \\ Y_{1i} \mid W_{1i} = 1 & \sim & \left(\gamma_0 + \gamma_1 + \gamma_2,\ \sigma^2\right) \end{array}\right\} \Rightarrow$$

$$Y_{1i} \sim \left(\gamma_0 + \gamma_1 + \frac{n_{11}}{n_{11}+n_{10}}\gamma_2,\ \sigma^2 + \gamma_2^2\frac{n_{11}n_{10}}{\left(n_{11}+n_{10}\right)^2}\right)$$

12

Binary *W* : Unadjusted Analysis

$$Y_{0i} \sim \left(\gamma_0 + \frac{n_{01}}{n_{01}+n_{00}}\gamma_2,\ \sigma^2 + \gamma_2^2 \frac{n_{01}n_{00}}{(n_{01}+n_{00})^2}\right)$$

$$Y_{1i} \sim \left(\gamma_0 + \gamma_1 + \frac{n_{11}}{n_{11}+n_{10}}\gamma_2,\ \sigma^2 + \gamma_2^2 \frac{n_{11}n_{10}}{(n_{11}+n_{10})^2}\right)$$

$$\bar{Y}_{1\bullet} - \bar{Y}_{0\bullet} \sim \left(\gamma_1 + \frac{n_{11}n_{00} - n_{10}n_{01}}{(n_{11}+n_{10})(n_{01}+n_{00})}\gamma_2,\ \sigma^2 + \gamma_2^2\left(\frac{n_{01}n_{00}}{(n_{01}+n_{00})^2} + \frac{n_{11}n_{10}}{(n_{11}+n_{10})^2}\right)\right)$$

13

Binary *W* : Analyses Within Subgroups

$$\left.\begin{array}{ll} Y_{1i} \mid W_{1i}=0 & \sim \left(\gamma_0+\gamma_1,\ \sigma^2\right) \\ Y_{0i} \mid W_{0i}=0 & \sim \left(\gamma_0,\ \sigma^2\right) \end{array}\right\} \Rightarrow$$

$$\hat{\Delta}_0 = \bar{Y}_{1\bullet} \mid W_{1i}=0 - \bar{Y}_{0\bullet} \mid W_{0i}=0 \sim \left(\gamma_1,\ \sigma^2 \frac{n_{10}+n_{00}}{n_{10}n_{00}}\right)$$

$$\left.\begin{array}{ll} Y_{1i} \mid W_{1i}=1 & \sim \left(\gamma_0+\gamma_1+\gamma_2,\ \sigma^2\right) \\ Y_{0i} \mid W_{0i}=1 & \sim \left(\gamma_0+\gamma_2,\ \sigma^2\right) \end{array}\right\} \Rightarrow$$

$$\hat{\Delta}_1 = \bar{Y}_{1\bullet} \mid W_{1i}=1 - \bar{Y}_{0\bullet} \mid W_{0i}=1 \sim \left(\gamma_1,\ \sigma^2 \frac{n_{11}+n_{01}}{n_{11}n_{01}}\right)$$

14

Combining Across Subgroups

- Based on the properties of independent, normally distributed estimates

For independent $\hat{\theta}_1 \sim N\left(\theta_1, se_1^2\right),\quad \hat{\theta}_2 \sim N\left(\theta_2, se_2^2\right)$

$$a\hat{\theta}_1 + b\hat{\theta}_2 \sim N\left(a\theta_1 + b\theta_2,\ a^2 se_1^2 + b^2 se_2^2\right)$$

$$\hat{\theta}_1 - \hat{\theta}_2 \sim N\left(\theta_1 - \theta_2,\ se_1^2 + se_2^2\right)$$

$$\hat{\theta}_1 / \hat{\theta}_2 \sim N\left(\frac{\theta_1}{\theta_2}, \frac{1}{\theta_2^2}\left(se_1^2 + \frac{\theta_1^2}{\theta_2^2} se_2^2\right)\right)$$

15

Binary *W* : Average Across Subgroups

- We can use any weighted average

$$\hat{\Delta}_0 = \bar{Y}_{1\bullet} \mid W_{1i}=0 - \bar{Y}_{0\bullet} \mid W_{0i}=0 \sim \left(\gamma_1,\ \sigma^2 \frac{n_{10}+n_{00}}{n_{10}n_{00}}\right)$$

$$\hat{\Delta}_1 = \bar{Y}_{1\bullet} \mid W_{1i}=1 - \bar{Y}_{0\bullet} \mid W_{0i}=1 \sim \left(\gamma_1,\ \sigma^2 \frac{n_{11}+n_{01}}{n_{11}n_{01}}\right)$$

$$\hat{\Delta} = w\hat{\Delta}_0 + (1-w)\hat{\Delta}_1 \sim \left(\gamma_1,\ \sigma^2\left(w^2 \frac{n_{10}+n_{00}}{n_{10}n_{00}} + (1-w)^2 \frac{n_{11}+n_{01}}{n_{11}n_{01}}\right)\right)$$

16

Binary *W* : Average Across Subgroups

- Optimal choice minimizes variance
  - (Solution would differ when we have heteroscedasticity)

$$\hat{\Delta} = w\hat{\Delta}_0 + (1-w)\hat{\Delta}_1 \quad \sim \quad \left(\gamma_1,\ \sigma^2 \left(w^2 \frac{n_{10}+n_{00}}{n_{10}n_{00}} + (1-w)^2 \frac{n_{11}+n_{01}}{n_{11}n_{01}} \right) \right)$$

$$w = \frac{n_{10}n_{00}(n_{11}+n_{01})}{n_{10}n_{00}(n_{11}+n_{01}) + n_{11}n_{01}(n_{10}+n_{00})}$$

17

Example: Mortality by Previous CVD, Sex

- Using inflammatory markers data set
  - Mortality within 4 years is known for everyon
- Descriptive statistics
  - Males: 20.7% if previous CHD; 11.7% if not – RD 0.090
  - Females: 14.7% if previous CHD; 4.9% if not – RD 0.098
- Adjusted analysis could average the subgroup specific RD
  - Weight 50-50? According to M:F ratio in the age range?
  - (Assume no effect modification?)
  - (Average over effect modification?)

18

(Frequency) Matching

- We can use this notation to explore the benefits of matched analyses
- Suppose Y_{1i} measures "cases" having $Z_{1i} = 1$ and $W_{1i} = w_{1i} = 0,1$
  - Suppose n_{11} and n_{10} count the number with $W_{1i} = 1$ and $W_{1i} = 0$, respectively
- We then choose "controls" having $Z_{0i} = 0$ and $W_{0i} = w_{1i} = 0,1$ and measure Y_{0i}
  - We will thus have $n_{01} = n_{11}$ and $n_{00} = n_{10}$

19

Binary *W* : Frequency Matching

- Optimal choice minimizes variance

$$\hat{\Delta} = w\hat{\Delta}_0 + (1-w)\hat{\Delta}_1 \quad \sim \quad \left(\gamma_1,\ \sigma^2 \left(\frac{n_{10}+n_{00}}{n_{10}n_{00}} \right) \left(w^2 + (1-w)^2 \right) \right)$$

$$w = \frac{n_{10}n_{00}}{n_{10}n_{00} + n_{11}n_{01}}$$

20

Generalizations: Stratifications

- Stratified analyses:
  - Analyze within each subgroup
  - Average the results across subgroups
- When averaging across subgroups
  - If there is no effect modification, then we are free to choose weights to minimize variance (maximize precision)
  - If there is effect modification, we will get different answers according to which weights we use
    - Often we use population based weights so our answer will be relevant to some population of interest

21

Generalizations: Matching

- Frequency matching
  - We ensure that the marginal distribution of each covariate is the same across POI groups
  - Matching on fixed effects
- Individual matching
  - We ensure that the joint distribution (including interactions) of the matching variables are the same across POI groups
  - Matching on fixed effects or random effects
    - Fixed effects: e.g., age, sex, height, weight, smoking behavior
    - Random effects: e.g., hospital, family, community of residence

22

Comparison to Regression

- We use regression to
  - Borrow information across groups
  - Form contrasts (e.g., slope) measuring associations
- As a rule, we can perform stratified analyses within regression
  - Fit dummy variables for each stratum
    - Does not borrow information across strata
  - May have to weight strata appropriately in a weighted regression
  - May have to consider how variances are estimated
    - Only within subgroups, or
    - Borrow information about variance across groups
  - (With binary response variables, issues about variance will also have to consider mean-variance relationships and adequacy of model)

23

Example: Mortality by Previous CVD, Sex

- Descriptive statistics
  - Males: 20.7% if previous CHD; 11.7% if not – RD 0.090
  - Females: 14.7% if previous CHD; 4.9% if not – RD 0.098

```
. regress deadin4 male prevdis m_prevdis
                    |               Robust
           deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
              male |      .0675   .0094     7.16   0.000     .0491    .0860
           prevdis |      .0979   .0158     6.18   0.000     .0669    .1289
         m_prevdis |     -.0080   .0243    -0.33   0.742    -.0557    .0397
             _cons |      .0492   .0045    11.04   0.000     .0404    .0579

. regress deadin4 male prevdis, robust
                    |                    Robust
           deadin4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
              male |      .0656   .0090     7.33   0.000     .0481    .0832
           prevdis |      .0940   .0121     7.75   0.000     .0702    .1177
             _cons |      .0499   .0046    10.84   0.000     .0409    .0589
```

24

Probability Models for Incidence of Disease

25

Risk Sets

- Most often, we recognize that the probability of an event depends in some way upon time
- In many cases, that time dependence is something we merely want to adjust for as we compare different groups
  - It is not as important to contrast the event probability over time
- We thus find it convenient to couch many of our analyses of binary data in terms that also consider "time to event"

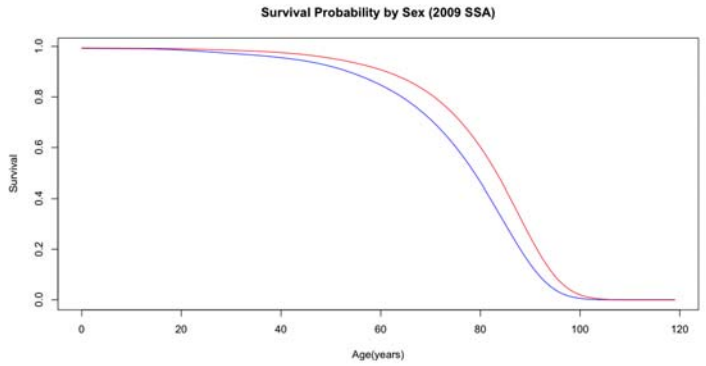
26

Incidence and Mortality Rates (Hazards)

- We are often interested in the rate (over time) at which individuals convert from being "event-free" to having had the event
  - Time can be calendar time, age, study time …
  - (They differ in what we call time zero)
- At each point in time, we essentially compute a proportion
  - Denominator: Individuals who are currently "event-free"
  - Numerator: Among those in the denominator, who converts in the next instant
- Referred to as
  - Epidemiology: incidence and mortality rates, force of mortality
  - Statistics and probability: hazard function

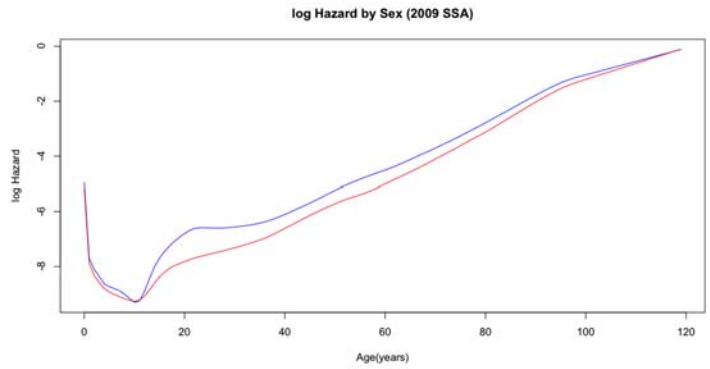
27

Age Effects on Mortality



28

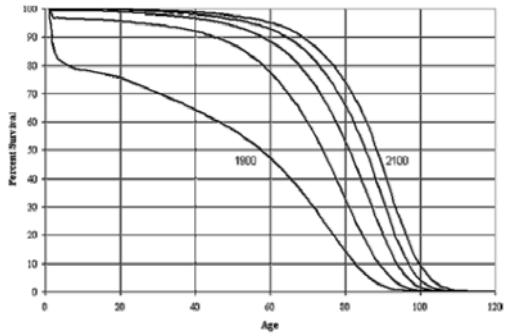
Age Effects on Mortality



29

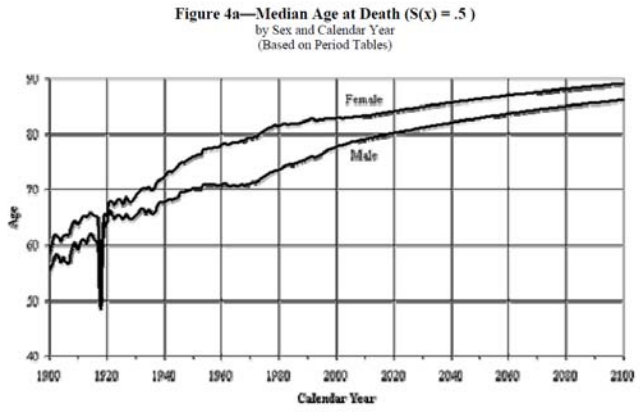
Birth Cohort Effects on Mortality

- Survival curves 1900 to 2100 by 50 year increments



30

Calendar Year Effects on Mortality



31

Hazard Function Notation

- For each individual in some group of interest, T measures the time the event will occur
  - $Y(t)$ is thus an indicator that the event has occurred prior to t
  - T might be infinity

Hazard function (continuous T): for very small Δt

$$\lambda(t) = \Pr(t \le T < t + \Delta t \mid t \le T)$$
$$= \frac{\Pr(t \le T < t + \Delta t)}{\Pr(t \le T)} = \frac{f(t)}{1 - F(t)}$$

$F(t)$ is cumulative distribution function

$f(t)$ is density

32

Hazard Rate Based Inference

- When the changing conversion rate is just a nuisance to our primary question, we still have to worry that time might be
  - An effect modifier and/or
  - A confounder and/or
  - A precision variable.
- Most often we choose some way to adjust for those roles by
  - Using weighted averages of the hazard (e.g., standardized rates)
  - Adjusting in a regression model
    - Poisson models adjusting for person-time at risk
    - Proportional hazards regression models
    - Parametric regression models

33

(Cumulative) Incidence and Mortality

- Sometimes we choose a specific interval of time of greatest interest
  - E.g., incidence of cancer within one year, teenage mortality
- Usually estimated with a simple proportion
  - Denominator: Individuals who are "event-free" at time *a*
  - Numerator: Individuals experiencing event between *a* and *b*
- It does relate to the hazard

$$\text{(Cumulative) incidence between times } a \text{ and } b$$

$$\Pr(a \le T < b \mid a \le T) = 1 - e^{-\int_a^b \lambda(u)\,du}$$

34

(Cumulative) Incidence Based Inference

- Note that if the hazard function is (nearly) constant over some small period of time then

$$\text{(Cumulative) incidence between times } a \text{ and } b$$

$$\Pr(a \le T < b \mid a \le T) = 1 - e^{-\int_a^b \lambda(u)\,du} = 1 - e^{-\int_a^b \lambda\,du} = 1 - e^{-\lambda(b-a)}$$

- This "piecewise exponential" model is often used as a basis for inference
  - The "exponential distribution" has a constant hazard
  - The exponential distribution is "memorylessness"
    - Independent intervals are independent
      - Within or between individuals
  - Also be thought of as Poisson approximation to binomial and/or times between events in Poisson process

35

Person-year Based Analyses

- We divide time into small intervals
  - Small age intervals will have common risk
  - Small follow-up time intervals
- We estimate person-years of observation
  - Each person may contribute to several categories
  - Sum across individuals for each category
- Estimate risk within those intervals
- Compare risk ratio across POI groups

36

Directly Standardized Rates

- Stratum specific weights chosen based on population

$$Y_i \mid X_i = x \quad \sim \quad P(\lambda_x t_i) \doteq N(\lambda_x t_i, \lambda_x t_i)$$

$$\hat{\lambda}_x = \frac{\sum_{i:X_i=x} Y_i}{\sum_{i:X_i=x} t_i} \quad \dot{\sim} \quad N\left(\lambda_x, \frac{\lambda_x}{\sum_{i:X_i=x} t_i}\right)$$

$$\hat{\lambda} = \frac{\sum_x w_x \hat{\lambda}_x}{\sum_x w_x}$$

37

Quiz Answers

38

Example: Incidence of Colorectal Cancer by Birthplace

39

Example

- We are interested in exploring the incidence of colorectal cancer by birthplace among whites in the US
- Cases identified through the SEER registry 1973-1987
- Available data
  - US, 25 non-US, unknown
  - Age in 5 year groups
  - Sex
- Denominator data from US census data

40

Analysis Model

- Effect modification in question?
- Potential confounding?
  - Variables causally associated with cancer incidence
  - Variable associated with birthplace in sample
- Precision?

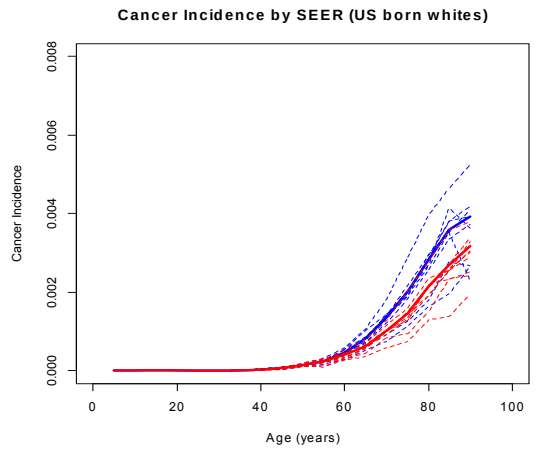
41

Analysis Model

- Effect modification in question?
  - Analysis within sex subgroups?
- Potential confounding?
  - Variables causally associated with cancer incidence
  - Variable associated with birthplace in sample
  - Age?
- Precision?

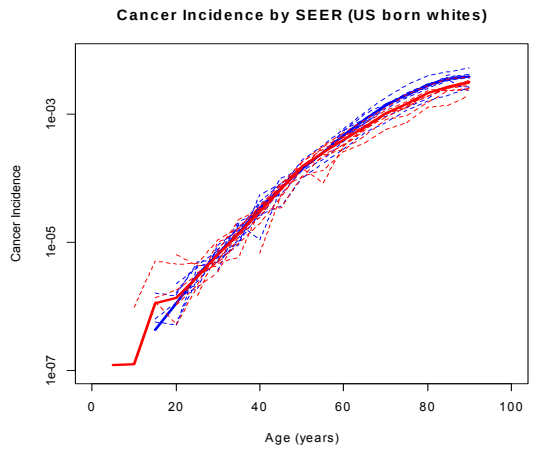
42

Associations with SEER



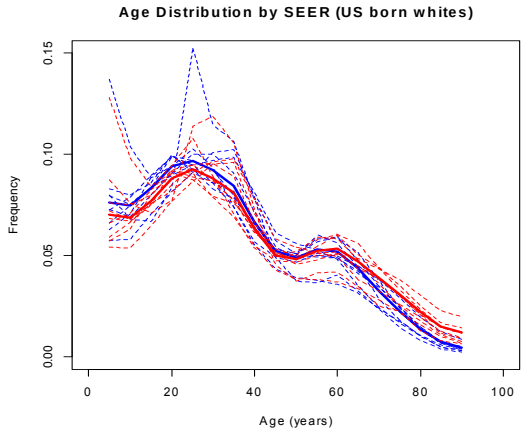
43

Associations with Response (log)



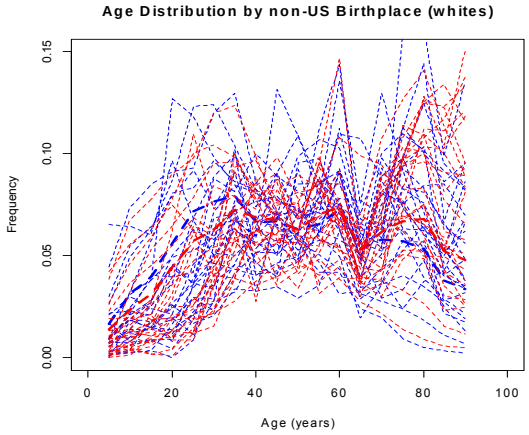
44

Associations with POI



45

Associations with POI



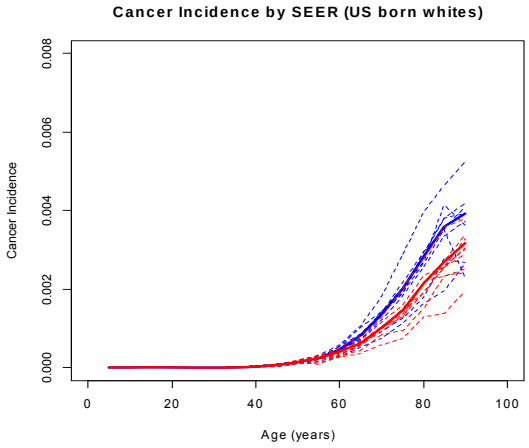
46

Example

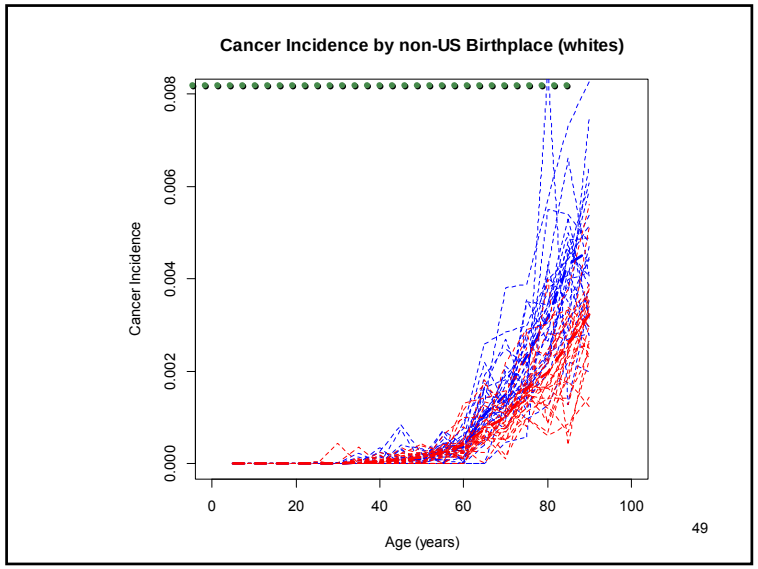
- We need to worry about
  - How we summarize over age
  - Confounding by age across country of birth

47

Associations with POI



48



Example

- We need to worry about missing data
- Missing completely at random
- Missing at random
- Missing not at random

50