37457

Advanced Bayesian Methods

Missing Data and Measurement Error Models

Pima Indians Diabetes Data

body mass index	indicator of diabetes
33.6	1
26.6	0
23.3	1
31.0	1
35.3	0
30.5	1
NA	1
37.6	0
38.0	1
27.1	0
.	.
.	.
.	.
42.0	1
29.7	0
28.0	0
39.1	1
NA	0
19.4	0
24.2	0
.	.
.	.
.	.
30.1	1
30.4	0

Logistic Penalized Spline model

If the data were complete:

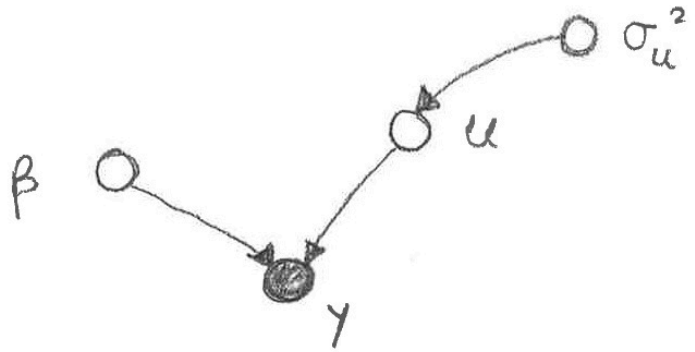
$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\frac{1}{1+\exp\{-f(x_i)\}}\right)$$

$$y_i = \begin{cases} 1 & \text{if diabetes for } i\text{th Pima Indian} \\ 0 & \text{otherwise} \end{cases}$$

$$x_i = \text{body mass index of } i\text{th Pima Indian}$$

$$f(x) = \beta_0 + \beta_1 + \sum_{k=1}^{K} u_k z_k(x), \quad u_k \overset{\text{ind.}}{\sim} N(0, \sigma_u^2).$$

Directed Acyclic Graph for Ordinary (Complete Data) Model



The Simplest Missing Data Model

Define:

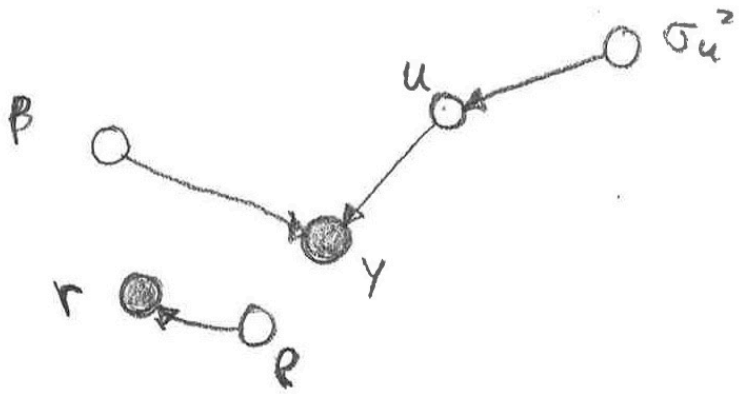
$$r_i = \begin{cases} 1 & \text{if body mass index observed for } i\text{th Pima Indian} \\ 0 & \text{if body mass index missing for } i\text{th Pima Indian} \end{cases}$$

$$r_i|\rho \overset{\text{ind.}}{\sim} \text{Bernoulli}(\rho).$$

This is known as a

missing completely at random model

Directed Acyclic Graph for Missing Completely at Random Model



Caveat of Previous Model

Missingness often depends on the predictor (or response) variables.

e.g. if asked "Are you a cigarette smoker?" in a survey, whether or not you answer may depend on smoking status.

Caveat of Previous Model

Missingness often depends on the predictor (or response) variables.

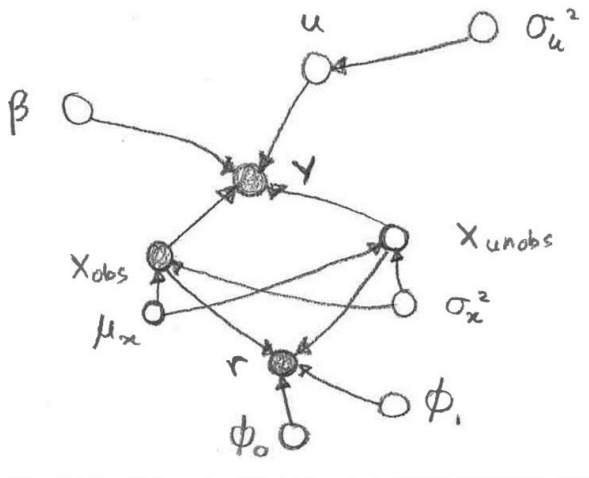
e.g. if asked "Are you a cigarette smoker?" in a survey, whether or not you answer may depend on smoking status.

More sophisticated missing data model for the Pima Indians data:

$$r_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\frac{1}{1+\exp\{-(\phi_0+\phi_1 x_i)\}}\right)$$

where the missingness is not at random.

Directed Acyclic Graph for Missing Not at Random Model



Look at and then run PIDana.R

One Final Topic

MEASUREMENT ERROR MODELS

Coronary Heart Disease Example

$$y_i = \begin{cases} 1 & \text{if coronary heart disease for } i\text{th patient} \\ 0 & \text{otherwise} \end{cases}$$

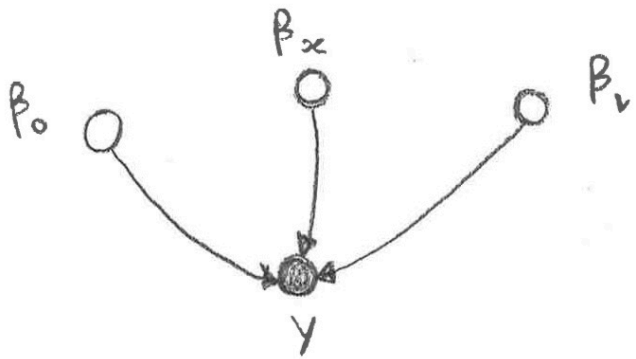
x_i = low density lipoprotein cholesterol level of ith patient

v_i = age in years of ith patient

Ideal (Bayesian logistic regression) model:

$$y_i|\beta_0, \beta_x, \beta_v \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\frac{1}{1+\exp\{-(\beta_0 + \beta_x\, x_i + \beta_v\, v_i)\}}\right)$$

Corresponding Directed Acyclic Graph



Cost of Data Collection

Suppose that

$x \equiv$ low density lipoprotein cholesterol level

costs \$2,500 to measure for each patient.

But

$w \equiv$ total cholesterol level

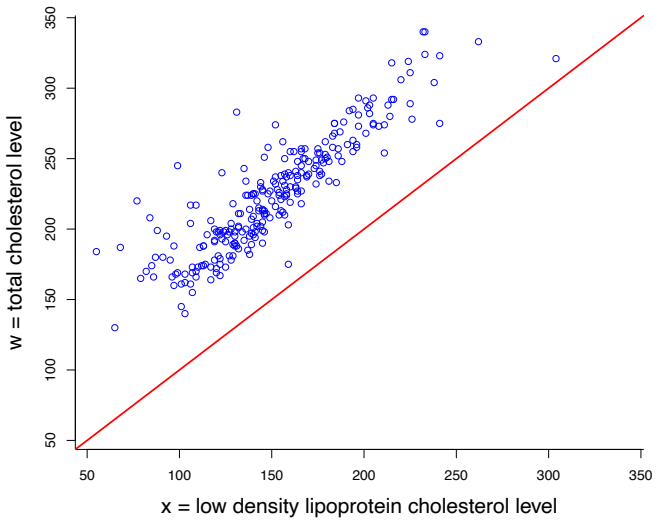
costs \$50 to measure for each patient.

Cost of Data Collection

Suppose that

$x \equiv$ low density lipoprotein cholesterol level

costs \$2,500 to measure for each patient.

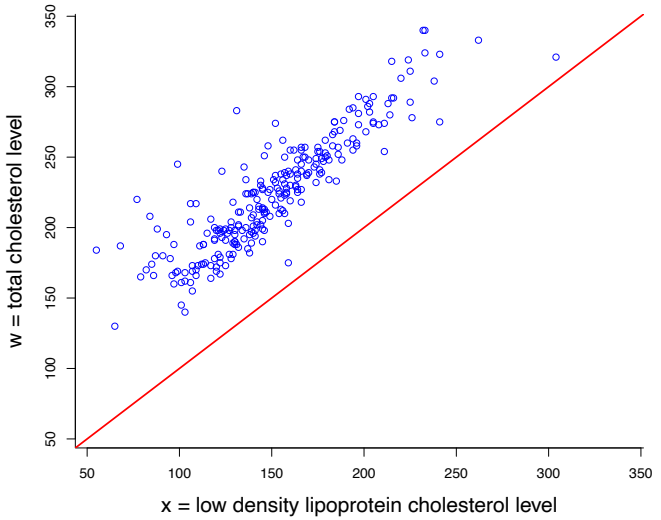


Options (with Limited Budget)

OPTION A

Only use EXPENSIVE x with a smaller sample of, say, about 30 people.

$\Longrightarrow$ lower power to detect effect of x.



Options (with Limited Budget)

OPTION A

Only use EXPENSIVE x with a smaller sample of, say, about 30 people.

$\Longrightarrow$ lower power to detect effect of x.

OPTION B

Use CHEAP w with a large sample of, say, 1500 people.

$\Longrightarrow$ hoping that w is a good surrogate for x.

Model the relationship between w and x using a relatively small validation data set and incorporate this into the model with all w data (large) and x data (small).

$\Longrightarrow$ measurement error model!

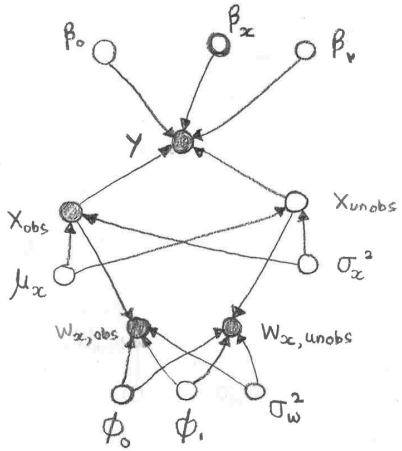
Measurement Model Add-On

$$w_i|x_i, \phi_0, \phi_1, \sigma_w^2 \sim N(\phi_0 + \phi_1\, x_i, \sigma_w^2),$$

$$x_i|\mu_x, \sigma_x^2 \overset{\text{ind.}}{\sim} N(\mu_x, \sigma_x^2),$$

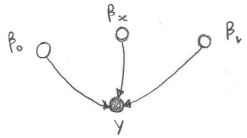
with the x_is only partially observed.

Directed Acyclic Graph for Measurement Error Model

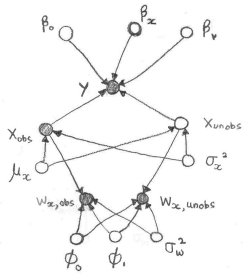


Look at CHDmeaErrAna.R

Both Directed Acyclic Graphs Together



Ordinary Model



Measurement Error Model

Results for Effect on Coronary Heart Disease

Using the fancy measurement error model we get the effect of low density lipoprotein cholesterol level as follows:

$\widehat{\beta}_x = 1.11$ with 95% credible interval $(0.0289, 2.31)$

i.e. a statistically significant effect

If we had just used total cholesterol data then we would have gotten

$\widehat{\beta}_w = 0.556$ with 95% credible interval $(-0.065, 1.22)$

lack of significance; not using the better predictor.

THE END OF

37457 SPRING 2024

MATERIAL

WHAT'S LEFT TO DO IN 37457?

Laboratory 4 (next Wednesday)

Assignment 8 (due next Wednesday)
NOTE EARLIER HELP SESSION!!

Practice final exam
(handed out in Class 12)

Practice final exam solutions
(will be put on web-site)

Final exam on 13th November

Receive your grade (December)

OPTIONAL 20 MINUTE PRESENTION IN SECOND HALF OF TODAY'S CLASS

THEME: What we did not get time to cover.

WHAT'S LEFT TO DO IN 37457?

Laboratory 4 (next Wednesday)

Assignment 8 (due next Wednesday)
NOTE EARLIER HELP SESSION!!

Practice final exam
(handed out in Class 12)

Practice final exam solutions
(will be put on web-site)

Final exam on 13th November

Receive your grade (December)

OPTIONAL 20 MINUTE PRESENTION IN SECOND HALF OF TODAY'S CLASS

THEME: What we did not get time to cover.

OR: What we would cover if we had 12 more weeks?
(and 12 more assessment tasks)