BIOS 736, Fall 2023: Class # 1 Notes

INTRODUCTION TO MEASUREMENT ERROR TERMINOLOGY AND CONCEPTS; PROPOSED CORRECTION METHODS

The basic goal (usually):

outcome exposure of interest

Relate Y to X, C → confounders / covariates

where instead of X we observe W (a variable that represents an error-prone substitute for X)

* Often we're interested in a generalized linear model, e.g.,

$$g[E(Y \mid \mathbf{X}=\mathbf{x}, \mathbf{C}=\mathbf{c})] = \alpha + \mathbf{x}'\boldsymbol{\beta} + \mathbf{c}'\gamma$$

error-prone version of $\underset{\sim}{X}$

?? What happens if we use **W** in place of **X** ??

This means instead we fit:

"naive" regression model

$$g[E(Y \mid \mathbf{W}=\mathbf{w}, \mathbf{C}=\mathbf{c})] = \alpha^* + \mathbf{w}'\boldsymbol{\beta}^* + \mathbf{c}'\gamma^*$$

⇒ Usually produces biased estimates of parameters of interest (sometime severe!)

*Note: While often the focus is on measurement error in X (covariate measurement error), you can also have measurement error in the outcome (Y)

ex 1) One-sample t test situation: Suppose instead of Y_i you actually measure

$$Y_i^* = Y_i + U_i \ , \quad \text{where } U_i\text{'s are } \overset{iid}{\sim} \ (0, \sigma_u^2)$$

assume $u \perp Y$

?? What happens to estimate of μ_y ??

$\bar{Y}^*$ is still unbiased

(still valid!)

?? What happens to usual test of H_0: $\mu_y = \mu_{y0}$??

$\bar{Y}^*$ is less precise than $\bar{Y}$ would have been.

(loss of power)

ex 2) If model of interest is linear and any error in Y is random and roughly normal, then it can just be absorbed into the linear model error term:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \ , \quad \text{but we observe } Y_i^* = Y_i + u_i$$

same assumptions as above

$$\Rightarrow \ Y_i^* = Y_i + u_i = \alpha + \beta x_i + (\varepsilon_i + u_i)$$

(estimates of α and β still valid; residual variance is larger)

loss power / precision

* But outcome measurement error not so easily dismissed in other situations… What about, e.g., logistic regression where Y is misclassified?

<u>Some terminology to be familiar with</u>:

* Note: This terminology can ~~be found discussed in the Carroll et al. (2006) text,~~ and in review papers such as Thomas, Stram, and Dwyer (*Ann Rev Pub Hlth*, ~~1993~~)

will post ~~to~~ canvas

a) Systematic vs. random measurement errors

We'll primarily focus on random errors rather than purely systematic ones, as the latter are generally less troublesome.

ex) Systematic errors in simple linear regression of Y on X:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

i) Suppose $W_i = X_i + c$ for all subjects—what happens to the model?

only affects intercept

$$Y_i = \alpha + \beta(w_i - c) + \varepsilon_i = (\alpha - \beta c) + \beta w_i + \varepsilon_i$$

ii) Suppose $W_i = X_i \times c$ for all subjects—what happens to the model?

$$Y_i = \alpha + (\beta/c) w_i + \varepsilon_i$$

*<u>Note</u>: However, sometimes measurement errors can have both a systematic and a random component, in which case ignoring the systematic component may lead to problems.

b) Non-differential vs. differential measurement errors   Important distinction

*There are at least 2 common definitions given for non-differential error: in X

i) $f(Y \mid X, W) = f(Y \mid X)$ the usual definition

once know X, w tells nothing of f(y)

$\Rightarrow$ W is a "surrogate" for X   usually implies non-diff error

error-prone version of X

ii) $f(W \mid X, Y) = f(W \mid X)$ dist^n of w|x does not vary w/ Y

Although i) is more commonly cited as the non-differential assumption, ii) is sometimes motivated naturally and these two expressions are equivalent.

HW 1

LET'S VERIFY THAT i) implies ii):

$$f(Y \mid X, W) = f(Y \mid X) \;\Rightarrow\; \frac{f(Y, X, W)}{f(X, W)} = f(Y \mid X)$$

$$\Rightarrow \frac{f(W \mid X, Y)\, f(X, Y)}{f(X, W)} = f(Y \mid X)$$

$$\Rightarrow \frac{f(W \mid X, Y)\, f(Y \mid X)\, f(X)}{f(X, W)} = f(Y \mid X)$$

$$\Rightarrow f(W \mid X, Y) = \frac{f(X, W)}{f(X)} = f(W \mid X)$$

[similarly, can show that ii) implies i)]

Lyles Huang

c) Structural vs. functional measurement error

measured w/ error

* In the structural approach, the true unknown predictor X is viewed as a random variable following some distribution f(X) in the population of interest

* In the functional approach, traditionally X was treated as a fixed unknown for each experimental unit and the X's might be estimated as nuisance parameters

→ fewer assumptions about measurement error model (MEM)

-- Modern view of "functional" is that X may be fixed or random, but if random then make few or no assumptions about the distn. of X (Carroll et al., 2006)

Robustness vs. efficiency

-- Structural sometimes more 'natural' and/or more powerful

-- Functional more non-parametric ; less risk of misspecifying meas. error model

d) Classical vs. Berkson error

*Relates to how measurement error model is structured. Is it more akin to

$W = X + U$ (classical) → mostly !

OR

$X = W + U$ (Berkson) ?

- Classical seems more 'natural'

think philosophically

- Typical Berkson example: X = doses delivered by a machine, W = number the doctor sets on the dial

- Note: In linear regression, Berkson error does not bias regression parameter estimates

$W = X + u \Rightarrow X = W - u$ → "classical" $E(X|W)$ → linear in w

$w \not\perp u$ $\qquad$ iid$(0, \sigma_u^2)$ $\qquad$ $u \perp W$

Illustration: Berkson error in SLR :

True model: $\quad Y = \alpha + \beta X + \varepsilon$

Suppose we fit: $\quad Y = \alpha^* + \beta^* W + \varepsilon^*$ $\quad$ where $\boxed{X = W + U}$ Berkson MEM

Note: $\quad E(Y|W) = \alpha + \beta E(X\,|\,W)$

$$= \alpha + \beta W \quad \Rightarrow \beta^* = \beta$$

* * Key is that E(X | W) = W (relates to "regression calibration" – discussed later)

one way to address "classical" error

- **Now that we have some terminology down… What sorts of things might the statistician seek to do when measurement error is present?!**

1) Understand potential effects of the measurement error (e.g., Attenuation? → $\hat{\beta}$ biased toward null
Inflation? Loss of power?)
$\hat{\beta}$ biased away from null.

2) May seek sensitivity analysis to get feel for how detrimental the effects could be in a given application

* 3) Determine if study design / available data are sufficient to facilitate a correction for measurement error

* validity first
efficiency second

- Restore validity of parameter or effect estimates
- Allow valid inference and/or confidence intervals — inference
- Efficiency vs. robustness considerations

Two standard examples

I) Simple linear regression with measurement error in X:

$$\text{iid}(0, \sigma_\epsilon^2)$$

TDM

Model of interest: $$Y = \alpha + \beta X + \varepsilon \qquad (1)$$

MEM

Classical error model: $$W = X + U, \quad U\text{'s} \overset{\text{iid}}{\sim} (0, \sigma_u^2), \ U \perp X$$

Note from (1) that $$\beta = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}$$

naive regression

Now, what happens when we fit $Y = \alpha^* + \beta^* W + \varepsilon^*$?

NOTE: $$E(Y \mid W) = \alpha + \beta E(X \mid W)$$

Thus, if E(X|W) is linear in W then "identity link is preserved" !

$$\Rightarrow \beta^* = \frac{\operatorname{Cov}(W,Y)}{\operatorname{Var}(W)} = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(W)} = \frac{\beta \operatorname{Var}(X)}{\operatorname{Var}(W)}$$

$$= \lambda\beta, \text{ where } \lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

a key would be to estimate λ, or σ_u^2, or σ_x^2

- Note that $0 < \lambda < 1$ ("attenuation" of the "naïve" regression coefficient)

β^* is biased toward null

SAS EXAMPLE 1: Non-differential measurement error

true $\beta = 1.5$

```
data one;
  alpha=-1; beta=1.5; sigsq=.25;
  sigsqx=.5; sigsqu=.5;

   do i=1 to 25;
    x=sqrt(sigsqx)*RANNOR(0);
    u=sqrt(sigsqu)*RANNOR(0);
    w=x+u;

    y=alpha + beta*x + sqrt(sigsq)*RANNOR(0);
     output;
   end;
run;

proc reg;
  model y=x;

proc reg;
  model y=w;
run;
```

$n=25$

$x \overset{iid}{\sim} N(0, \sigma_x^2)$

$u \overset{iid}{\sim} N(0, \sigma_u^2)$

$w = x+u$ MEM

$x \perp u$

$u \perp (x, y)$

non-differential error

true x

true regression

naive regression

The REG Procedure
Model: MODEL1
Dependent Variable: y

y vs. x

Number of Observations Used 25

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.20173	0.13883	-8.66	<.0001
x	1	1.58582	0.18058	8.78	<.0001

$\hat{\beta}$ close to 1.5

$\beta^* \to \lambda\beta$

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Used 25

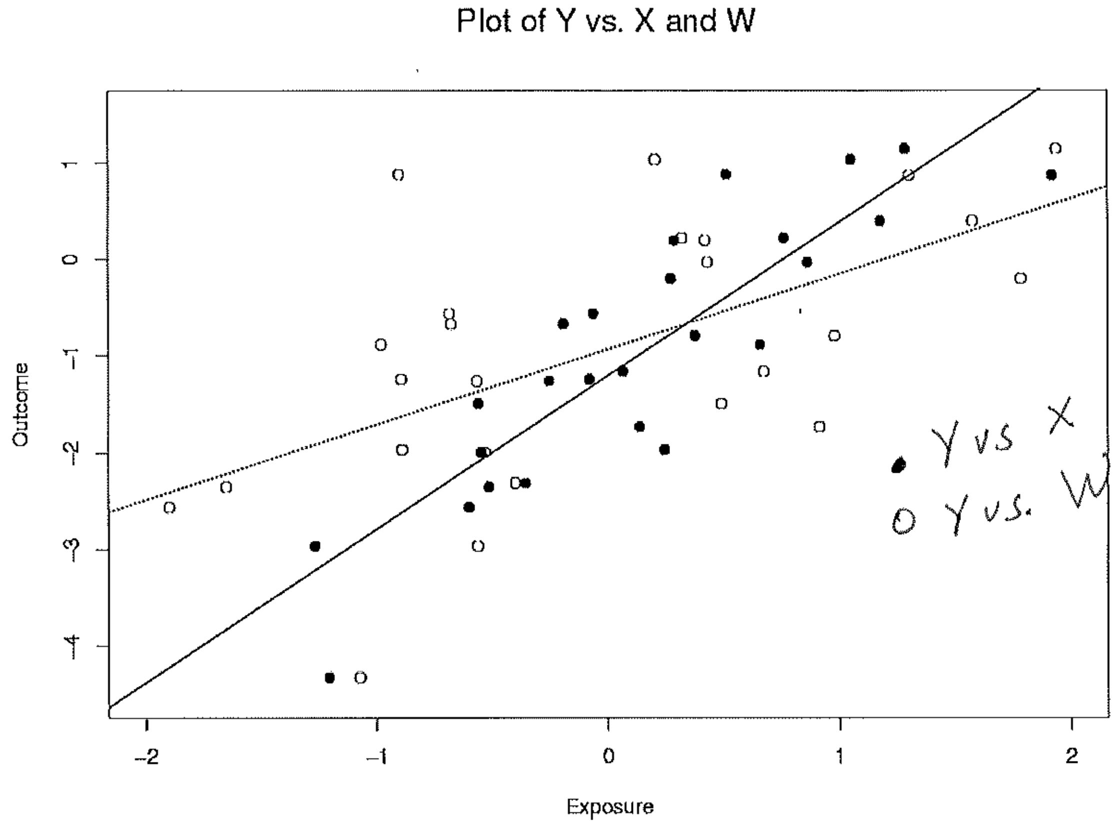
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.92948	0.22798	-4.08	0.0005
w	1	0.77857	0.21949	3.55	0.0017

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} = \frac{.5}{.5+.5} = .5$$

$\hat{\beta}^*$ attenuated (bias to null)

What would overlaid scatter plots look like?



b) Differential measurement error

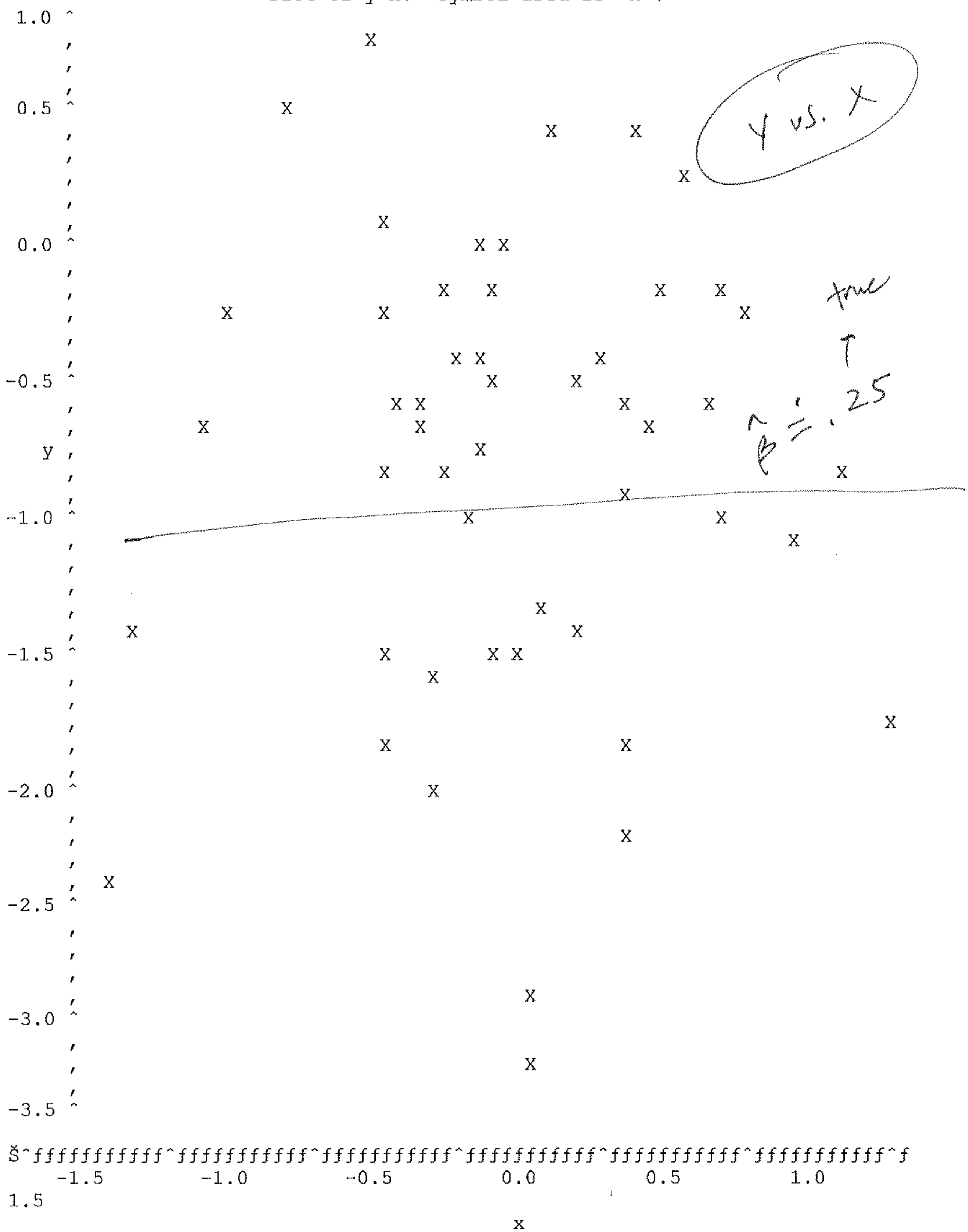
* If $f(Y \mid X, W) \neq f(Y \mid X)$, we can get either attenuation or inflation of the "naïve" estimate ($\hat{\beta}^*$)

ex) Suppose $X \sim N(0, \sigma_x^2)$, $Y = \alpha + \beta X + \varepsilon$ with $\varepsilon \sim N(0, \sigma_\varepsilon^2)$,

and let's assume that $W = X + aY + U$ where $U \sim N(0, \sigma_u^2)$

??? What would $E(Y \mid W)$ look like in this unusual case??

Plot of y*x. Symbol used is 'X'.



NOTE: $$E(Y \mid W) = \alpha + \beta E(X \mid W) + E(\varepsilon \mid W)$$

$\Rightarrow$ Can show from conditional normality that

$$E(Y \mid W) = \alpha^* + (\beta\lambda) W \quad ,$$

where λ is a function of $\left(a,\ \sigma_x^2,\ \sigma_y^2,\ \sigma_u^2,\ \sigma_{xy}\right)$ that could be < 1 OR > 1 !

SAS EXAMPLE 2: Differential measurement error

```
data two;
  alpha=-1; beta=.25; sigsq=1.05;
  sigsqx=.3; sigsqu=1.05;

   do i=1 to 50;
    x=sqrt(sigsqx)*RANNOR(0);
    u=sqrt(sigsqu)*RANNOR(0);
    y=alpha + beta*x +sqrt(sigsq)*RANNOR(0);

    w=x + 2*y + u;                            **NOTE : a = 2 ** ;

     output;
   end;
run;
```

MEM differential error

```
proc plot;
  plot y*x='X' y*w='W';
```

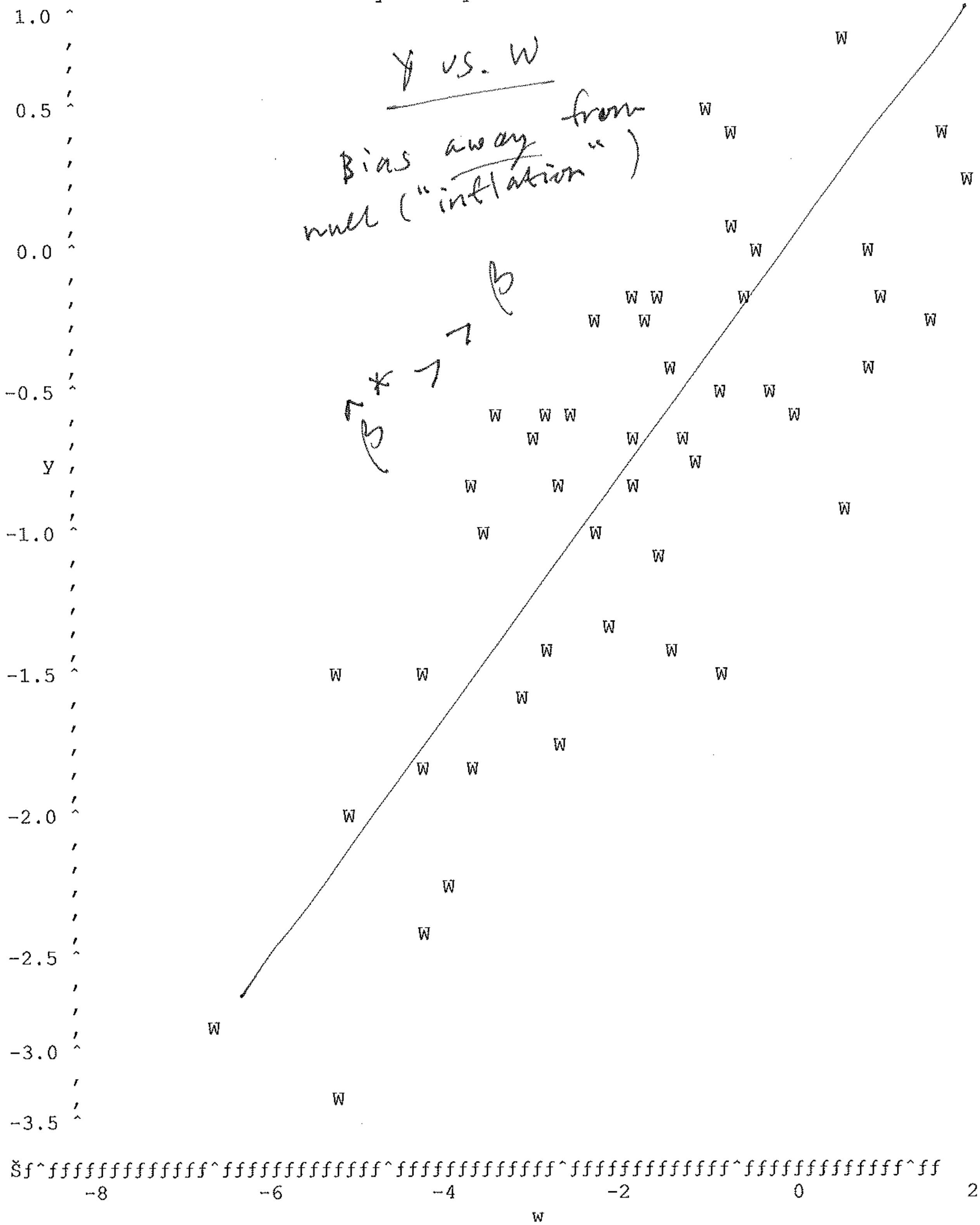
true

```
proc reg;
  model y=x;
```

"naive"

```
proc reg;
  model y=w;
run;
```

Plot of y*w. Symbol used is 'W'.



The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Used 50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.02120	0.02120	0.03	0.8685
Error	48	36.71295	0.76485		
Corrected Total	49	36.73415			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.81233	0.12376	-6.56	<.0001
x	1	0.03598	0.21615	0.17	0.8685

$\hat{\beta}$

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Used 50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	22.23193	22.23193	73.58	<.0001
Error	48	14.50222	0.30213		
Corrected Total	49	36.73415			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.17253	0.10779	-1.60	0.1160
w	1	0.34181	0.03985	8.58	<.0001

misleading

$\hat{\beta}^* \to \beta$

II) Misclassification of X in a 2×2 table

Consider a case-control study that would ideally reflect cell probabilities and yield data based on independent samples as follows:

= "Y" (outcome)

Cell Probabilities

	X	
D	1	0
1	π_1	$1-\pi_1$
0	π_0	$1-\pi_0$

$\Rightarrow$

Cell Counts

	X	
D	1	0
1	a	b
0	c	d

Here, $\pi_1 = \Pr(X=1 \mid D=1)$ prob of exposure among cases

$\pi_0 = \Pr(X=1 \mid D=0)$ prob of exposure among controls

Parameter of interest: $\boxed{OR = \pi_1(1-\pi_0)/[\,\pi_0(1-\pi_1)]}$

Now, suppose that a surrogate W is used to measure exposure (e.g., response based on questionnaire). W is a misclassified substitute for X, with:

a) The non-differential case

Sensitivity = SE = $\Pr(W=1 \mid X=1) \Rightarrow \Pr(W=0 \mid X=1) = 1 - SE$

Specificity = SP = $\Pr(W=0 \mid X=0) \Rightarrow \Pr(W=1 \mid X=0) = 1 - SP$

Observed table:

	W	
D	1	0
1	A	B
0	C	D

NOTE: Observed cell counts (A, B, C, D) differ from true cell counts (a, b, c, d) that we would have seen without misclassification!

SAS EXAMPLE 3: Non-differential misclassification

cases $(y=1)$

SE SP

```
data one;
 pi1=.4; se=.75; sp=.9; onemsp=1-sp;

  do i=1 to 100;
   y=1;  x=ranbin(0,1,pi1);
   if x=1 then w=ranbin(0,1,se);
   else if x=0 then w=ranbin(0,1,onemsp);
 output;
end;
```

$\pi_1 = .4$

$pr(w=1 \mid x=1) = SE = .75$

$pr(w=1 \mid x=0) = 1-SP$

controls $(y=0)$

```
data two;
 pi0=.2; se=.75; sp=.9; onemsp=1-sp;

  do i=1 to 100;
   y=0;  x=ranbin(0,1,pi0);
   if x=1 then w=ranbin(0,1,se);
   else if x=0 then w=ranbin(0,1,onemsp);
 output;
end;
```

SE and SP same for cases + controls (non-differential misclassification!)

$$\underline{\text{Note}}: \quad \text{True OR} = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)} = \frac{0.4\times0.8}{0.2\times0.6} = 2.67$$

```
data three;
  set one two;
  if y=1 then ychar='dis';   if y=0 then ychar='nondis';
  if x=1 then xchar='expos'; if x=0 then xchar='nonexpos';
  if w=1 then wchar='expos'; if w=0 then wchar='nonexpos';
run;
```

```
proc freq;
  tables xchar*wchar;
  tables ychar*xchar / chisq cmh ;
  tables ychar*wchar / chisq cmh ;
run;
```

Output for Non-differential example:

```
      Table of xchar by wchar

xchar     wchar

Frequency,
Row Pct  ,expos   ,nonex   ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
expos    ,     38 ,     10 ,     48
         ,  79.17 ,  20.83 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
nonex    ,     12 ,    140 ,    152
         ,   7.89 ,  92.11 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total          50      150      200
```

X vs. W

$\hat{SE} = Pr(w=1 | x=1) = \frac{38}{48} = .79$

```
      Table of ychar by xchar

ychar     xchar

Frequency,
Row Pct  ,expos   ,nonex   ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
dis      ,     35 ,     65 ,    100
         ,  35.00 ,  65.00 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
non      ,     13 ,     87 ,    100
         ,  13.00 ,  87.00 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total          48      152      200
```

Y vs. X (valid)

$\hat{OR} = 3.60$ (valid estimate, & truth = 2.67)

```
                Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                     Value     95% Confidence Limits
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control      Mantel-Haenszel           3.6036        1.7662        7.3523
  (Odds Ratio)    Logit                     3.6036        1.7662        7.3523
```

```
                          Table of ychar by wchar

                    ychar     wchar

                    Frequency,
                    Row Pct  ,expos   ,nonex   ,  Total
                    ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
                    dis      ,     32 ,     68 ,    100
                             ,  32.00 ,  68.00 ,
                    ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
                    non      ,     18 ,     82 ,    100
                             ,  18.00 ,  82.00 ,
                    ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
                    Total          50      150      200
```

Y vs. W

$\hat{OR}^* = 2.14 < 2.67$

"naive"

Bias toward null

```
                Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                     Value     95% Confidence Limits
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control      Mantel-Haenszel           2.1438        1.1070        4.1516
  (Odds Ratio)    Logit                     2.1438        1.1070        4.1516
```

b) The differential case: Assume SE and/or SP varies by case-control status

Sensitivity among diseased = $SE_1 = \Pr(W=1 \mid X=1, D=1)$

Sensitivity among controls = $SE_0 = \Pr(W=1 \mid X=1, D=0)$

Specificity among diseased = $SP_1 = \Pr(W=0 \mid X=0, D=1)$

Specificity among controls = $SP_0 = \Pr(W=0 \mid X=0, D=0)$

- Now the effect on the observed OR is much less predictable – could be attenuated, inflated, or "flipped" to other side of null!

SAS EXAMPLE 4: Differential misclassification

$\rightarrow SE_1, SP_1$

```
data four;
  pi1=.4; se=.8; sp=.75; onemsp=1-sp;

  do i=1 to 100;
   y=1;  x=ranbin(0,1,pi1);
   if x=1 then w=ranbin(0,1,se);
    else if x=0 then w=ranbin(0,1,onemsp);
 output;
end;
```

cases

$\rightarrow SE_0, SP_0$

```
data five;
  pi0=.2; se=.65; sp=.95; onemsp=1-sp;

  do i=1 to 100;
   y=0;  x=ranbin(0,1,pi0);
   if x=1 then w=ranbin(0,1,se);
    else if x=0 then w=ranbin(0,1,onemsp);
 output;
end;
```

controls

```
data six;
  set five four;
  if y=1 then ychar='dis';   if y=0 then ychar='nondis';
  if x=1 then xchar='expos'; if x=0 then xchar='nonexpos';
  if w=1 then wchar='expos'; if w=0 then wchar='nonexpos';
run;


proc freq;
  tables xchar*wchar;
  by y;
```

```
proc freq;
  tables ychar*xchar / chisq cmh;
  tables ychar*wchar / chisq cmh;
run;
```

Output for differential example:

y=0 controls

```
                 Table of xchar by wchar

xchar     wchar

Frequency,
Row Pct   ,expos    ,nonex   ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
expos     ,     11 ,      8 ,     19
          ,  57.89 ,  42.11 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
nonex     ,      4 ,     77 ,     81
          ,   4.94 ,  95.06 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total           15       85      100
```

$\hat{SE}_0 = .58$

$\hat{SP}_0 = .95$

y=1 cases

```
                 Table of xchar by wchar

xchar     wchar

Frequency,
Row Pct   ,expos    ,nonex   ,  Total
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
expos     ,     31 ,      6 ,     37
          ,  83.78 ,  16.22 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
nonex     ,     11 ,     52 ,     63
          ,  17.46 ,  82.54 ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total           42       58      100
```

$\hat{SE}_1 = .84$

$\hat{SP}_1 = .82$

Table of ychar by xchar

Y vs. X

ychar	xchar		
Frequency, Row Pct	expos	nonex	Total
dis	37 37.00	63 63.00	100
non	19 19.00	81 81.00	100
Total	56	144	200

truth = 2.67

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control	Mantel-Haenszel	2.5038	1.3153	4.7661
(Odds Ratio)	Logit	2.5038	1.3153	4.7661

Table of ychar by wchar

Y vs. W

ychar	wchar		
Frequency, Row Pct	expos	nonex	Total
dis	42 42.00	58 58.00	100
non	15 15.00	85 85.00	100
Total	57	143	200

Bias away from null

$\hat{OR}^* >> 2.67$ truth

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control	Mantel-Haenszel	4.1034	2.0841	8.0794
(Odds Ratio)	Logit	4.1034	2.0841	8.0794

Correcting for Measurement Error: General Considerations

("structural")

At least in the structural case, it is often useful to consider the following three-model paradigm (e.g., Clayton, 1992):

True Disease Model ("TDM")

1 e.g., $g[E(Y|X,\ C)] = \alpha + \beta X + \gamma C$

Measurement Error Model ("MEM"):

2 e.g., $W = X + U$ classical

or $X = \tau_0 + \tau_1 W + \tau_2 C + e$

can serve as an MEM; facilitates regression calibration

Predictor Distribution Model ("PDM"):

3 e.g., $X\text{'s} \overset{iid}{\sim} N(0, \sigma_x^2)$

not always necessary, but part of Clayton's paradigm

* Note: In practice, a feasible correction for measurement error usually requires information (i.e., assumed values or estimates) about "nuisance" parameters such as those comprising the MEM and the PDM!

* Such estimates are often obtained via "validation" or "reproducibility" data

⇒ Brings in study design considerations!

Validation data:

cheap/easy

Main study usually consists of (Y, W, C) measurements

often just (X, W)

$\Rightarrow$ Bring in some (X, W, C) data (external or internal)

nice: might (Y, X, W, C) on a subset of your subjects

* In theory this requires a "gold standard" method to assess X

Reproducibility data:

For all or some subjects, collect replicates of the surrogate (e.g., W_{i1}, W_{i2},…, W_{ip})

* For certain MEMs, may yield estimates of the necessary nuisance parameters

Brief List of Some Proposed Measurement Error Correction Strategies

a) Correction based on convergence in probability

$\Rightarrow$ Suppose we fit: $\quad g[E(Y \mid W,\ C)] = \alpha^* + \beta^* W + \gamma^* C$

Can we determine convergence results, e.g.,

$$\hat{\beta}^* \rightarrow \lambda\beta \qquad \text{and} \qquad \hat{\gamma}^* \rightarrow \varphi\gamma$$

-- Clean, intuitive…but not always applicable. Sometimes works "approximately".

b) Replacing W by E(X | W) ("Regression calibration")

⇒ Suppose we fit: $g[E(Y \mid X^*, C)] = \alpha^* + \beta^* X^* + \gamma^* C$

where $X^* = E(X \mid W, C)$

⇒ In some important cases, $\beta^* = \beta$ (at least approximately)

<u>Notes</u>:

Method b) is perhaps most commonly applied, but not always guaranteed to produce consistent estimator

Methods a) and b) are closely linked and usually lead to similar or identical results; often if a) is "valid" then b) is "valid" (and vice-versa)

c) Maximum Likelihood (ML)

Consider a likelihood function (joint distribution of the data, viewed as a function of the parameters to be estimated), as follows:

$$L(\boldsymbol{\theta}, \boldsymbol{\tau};\ \mathbf{Y}, \mathbf{W}) = \prod_{i=1}^{n} f(Y_i, W_i;\ \boldsymbol{\theta}, \boldsymbol{\tau})$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ represent vectors of "primary" (e.g., from the TDM) and "nuisance" (e.g., from the MEM and PDM) parameters, respectively

Now note that:

$$f(Y_i, W_i; \boldsymbol{\theta}, \boldsymbol{\tau}) \quad = \quad \int_{-\infty}^{\infty} f(Y_i, W_i, X_i; \boldsymbol{\theta}, \boldsymbol{\tau})\, dX_i$$

$$= \quad \int_{-\infty}^{\infty} f(Y_i \mid X_i, W_i; \boldsymbol{\theta}, \boldsymbol{\tau})\, f(X_i \mid W_i; \boldsymbol{\tau})\, f(W_i; \boldsymbol{\tau})\, dX_i$$

* If measurement error is <u>non-differential</u>, then:

$$= \quad f(W_i; \boldsymbol{\tau}) \int_{-\infty}^{\infty} f(Y_i \mid X_i; \boldsymbol{\theta})\, f(X_i \mid W_i; \boldsymbol{\tau})\, dX_i$$

$$\Rightarrow \qquad L(\boldsymbol{\theta}, \boldsymbol{\tau};\ \mathbf{Y}, \mathbf{W}) \;=\; \prod_{i=1}^{n} f(W_i; \boldsymbol{\tau}) \left\{ \prod_{i=1}^{n} \int_{-\infty}^{\infty} f(Y_i \mid X_i; \boldsymbol{\theta})\, f(X_i \mid W_i; \boldsymbol{\tau})\, dX_i \right\}$$

This invites estimation of $\boldsymbol{\tau}$ "separately" based only on the W data; $\hat{\boldsymbol{\tau}}$ could then be "plugged in" to the 2nd piece (in braces)

$\Rightarrow$ "Pseudo-likelihood" approach (Gong and Samaniego, 1981, *Annals of Statistics*)

(Still may be a substantial computational problem to maximize the function in braces)

d) Quasi-likelihood (QL)

Quasi-score function:

$$S(\boldsymbol{\theta},\boldsymbol{\tau}) \;=\; \sum_{i=1}^{n}\left(\frac{d\,\mathrm{E}(\mathrm{Y_i}\mid \mathrm{W_i})}{d\boldsymbol{\theta}}\right)' \frac{[\mathrm{Y_i}-\mathrm{E}(\mathrm{Y_i}\mid \mathrm{W_i})]}{\mathrm{Var}(\mathrm{Y_i}\mid \mathrm{W_i})} \;\stackrel{set}{=}\; 0$$

$\Rightarrow$ Generally must incorporate external estimate ($\hat{\boldsymbol{\tau}}$), as in pseudo-likelihood approach

$\Rightarrow$ "Fix-up" needed to get $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ ("sandwich" estimator)

(Liang and Liu, 1991; Liu and Liang, 1991 *Statistics in Medicine*)

e) Semi- or Non-parametric approaches

$\Rightarrow$ Trying to relax MEM distributional assumptions, e.g., by estimating $f(X \mid Z)$ nonparametrically to avoid misspecifying it

(Pepe and Fleming, 1991 *JASA*; Robins et al. 1994, 1995)