BIOS 736, Fall 2023: Notes for Class # 3

CONTINUED: WHAT IS REQUIRED FOR AN ACTUAL ADJUSTMENT?

Shifting gears: Now, let's return to exposure measurement error:

ex) Let's return to the simple linear regression example

Model of interest: $Y = \alpha + \beta X + \varepsilon_y$ "TDM"

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

$\Rightarrow$ We saw that $\hat{\beta}^* \rightarrow \lambda\beta$, where $\hat{\beta}^*$ is the slope estimator from the "naïve" regression of Y on W, and $\lambda = \dfrac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$. $\Rightarrow$ Sometimes called the "reliability coefficient"

* Ideally, we estimate λ using validation or reproducibility data. Note, though, that a sensitivity analysis could easily be carried out here by varying assumed values of σ_x^2 and σ_u^2, and recording the corresponding values of $\hat{\beta} = \hat{\beta}^* / \hat{\lambda}$

$\hat{\beta} = \hat{\beta}^* / \hat{\lambda}$

if not estimable, can vary the assumed λ for sensitivity analysis.

$\Rightarrow$ Consider using validation data to estimate β:

Types of validation data: EXTERNAL or INTERNAL

a) External: (X, W) pairs from an external source, allowing estimation of λ

b) Internal: You measure X on a subset of your own sample, yielding (X, W, Y) data on those subjects

truth
"gold standard"

Validation Study Characteristics

	Advantages	Disadvantages
External	CHEAP, EASY	~~MAY NOT BE AVAILABLE~~ AND/OR "TRANSPORTABLE" ; NOT STATISTICALLY ~~EFFICIENT ; LIKELY~~ REQUIRES NON-DIFFERENTIAL ERROR ASSUMPTION
Internal	CAN BE COSTLY, TIME CONSUMING	SOLVES "TRANSPORTABILITY" PROBLEM ; STATISTICALLY EFFICIENT ; NON-DIFFERENTIALITY NOT REQUIRED

Suppose we have an external validation sample of (X, W) pairs:

For the MEM (perhaps preferably), let's assume

$$\boxed{X = \tau + \lambda W + \varepsilon_X}$$

E(X | W) is linear in W – means E(Y|W) is linear in W (identity link preserved)

Now, to make things easier, let's make fairly standard assumptions about the errors in the 2 regression models:

$$\left\{ \varepsilon_y \overset{\text{iid}}{\sim} (0,\sigma_y^2) \quad \text{and} \quad \varepsilon_X \overset{\text{iid}}{\sim} (0,\sigma_X^2) \right\}$$

Note: Normality of these errors should not be required to get a valid corrected estimate for β, but we may need to assume it to make convenient inferences about β

Two-step process to get a corrected $\hat{\beta}$:

1) Fit the "naïve" model: $Y = \alpha^* + \beta^* W + \varepsilon^*$ on the "main" study (Y, W) data to get $\hat{\beta}^*$ and $\widehat{\text{Var}}(\hat{\beta}^*)$

2) Fit the MEM: $X = \tau + \lambda W + \varepsilon_X$ on the validation study (X, W) data to get $\hat{\lambda}$ and $\widehat{\text{Var}}(\hat{\lambda})$

could use Z's

2 versions of "RC"

Now: $$E(Y \mid W) = \alpha + \beta \boxed{E(X \mid W)} = \alpha + \beta(\tau + \lambda W)$$

$\Rightarrow$ $$\hat{\alpha}^* \to \alpha + \beta\tau \quad \text{and} \quad \hat{\beta}^* \to \lambda\beta$$

Thurston et al. (year?)

So, our corrected estimator becomes: $$\hat{\beta} = \hat{\beta}^* / \hat{\lambda}$$

To get $\widehat{\text{Var}}(\hat{\beta})$, we use the <u>multivariate delta method</u> (a standard statistical tool to approximate the variance of a function of random variables)

In this case, the delta method yields:

$$\left\{ \widehat{\text{Var}}(\hat{\beta}) = \frac{\left(\widehat{\text{Var}}(\hat{\beta}^*) + \hat{\beta}^2\, \widehat{\text{Var}}(\hat{\lambda}) / \hat{\lambda}^2\right)}{\hat{\lambda}^2} \right\}$$

$\Rightarrow$ approximate 95% CI for β is $\hat{\beta} \pm 1.96\sqrt{\widehat{\text{Var}}(\hat{\beta})}$

SAS EXAMPLE 5: Large sample to illustrate measurement error result in simple linear regression

```
data one;
n=200000;
alpha=0;
beta=1; → (2)
sigsqy=1;

sigsqx=.5;
sigsqu=1;

do i=1 to n;
  x=5 + sqrt(sigsqx)*rannor(0);
  u=sqrt(sigsqu)*rannor(0);          → MEM
 z=x+u;
  y=alpha + beta*x + sqrt(sigsqy)*rannor(0);  → TDM
   output;
end;

proc reg;  → W
 model y=z;

proc reg;  → W
 model x=z;
run;
```

$Y = \alpha + \beta X + \epsilon_y$

(W) ← z

The REG Procedure
Model: MODEL1
Dependent Variable: y

"Naive" regression

Number of Observations Used 200000

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00890	0.00341	-2.61	0.0089
z	1	$\hat{\beta}^*$ 0.66388	0.00278 → $\hat{SE}(\hat{\beta}^*)$	239.17	<.0001

$$\frac{\hat{\beta}^*}{\hat{\lambda}} \doteq 2$$

The REG Procedure
Model: MODEL1
Dependent Variable: x

Number of Observations Used 200000

X vs. W

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00282	0.00129	-2.19	0.0287
z	1	$\hat{\lambda}$ 0.33247	0.00105 → $\hat{SE}(\hat{\lambda})$	316.40	<.0001

$\hat{\sigma}_x^2$ get $\hat{\sigma}_u^2$ separately? (X, W) (Y, W)

(X, W)

Note that the data were generated under the classical MEM, i.e., $W = X + U$, U's $\overset{\text{iid}}{\sim}$ $N(0, \sigma_u^2)$. But $\hat{\lambda}$ from fitting the MEM $X = \tau + \lambda W + \varepsilon_x$ faithfully estimates $\lambda = \dfrac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$ -- we could also have estimated σ_x^2 and σ_u^2 directly from the (X, W) data.

could experiment w/ (τ, λ)

* Interestingly, (X, W) data generated under the MEM $X = \tau + \lambda W + \varepsilon_x$ can produce inflation as well as attenuation of the "naïve" $\hat{\beta}^*$, because this MEM does not necessarily imply the "classical" MEM…

can augment w/ other covariates

* It is the MEM of the form $\boxed{X = \tau + \lambda W + \varepsilon_x}$ that most readily leads to extensions that make it possible to handle <u>other covariates</u> in the regression model of interest

ex 2) Suppose the linear model of interest is:

$$1) \qquad Y = \beta_0 + \beta_1 X + \beta_2 C + \varepsilon_y \,, \qquad \text{TDM}$$

where X is measured with error and C is a covariate measured without error.

i.e., we observe W in place of X.

Let's assume the following MEM:

could be other covariates not in TDM.

$$2) \qquad X = \tau + \lambda_1 W + \lambda_2 C + \varepsilon_x$$

As before, assume $\varepsilon_y \overset{\text{iid}}{\sim} (0, \sigma_y^2)$ and $\varepsilon_x \overset{\text{iid}}{\sim} (0, \sigma_x^2)$

Now, we can readily see that

$$
\begin{aligned}
E(Y \mid W, C) &= \beta_0 + \beta_1 E(X \mid W, C) + \beta_2 C \\
&= \beta_0 + \beta_1(\tau + \lambda_1 W + \lambda_2 C) + \beta_2 C \qquad \text{"identity" link is preserved.} \\
&= \underbrace{(\beta_0 + \beta_1 \tau)}_{\beta_0^*} + \underbrace{\beta_1\lambda_1}_{\beta_1^*} W + \underbrace{(\beta_1\lambda_2 + \beta_2)}_{\beta_2^*} C
\end{aligned}
$$

$\Rightarrow$ The "naïve" estimates $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ can be thrown off in all possible directions when we regress Y on (W, C)!

Again, we can get "corrected" estimates as follows:

$$
\left[\quad \hat{\beta}_1 = \hat{\beta}_1^* / \hat{\lambda}_1 \quad \text{and} \quad \hat{\beta}_2 = \hat{\beta}_2^* - \hat{\beta}_1 \hat{\lambda}_2 \quad \right]
$$

The variances of these estimators can again be approximated via the multivariate delta method

GLMM

outcome exposure

covariate

Note: These results generalize to the case of multivariate Y, X, and C

(e.g., Spiegelman et al., *Am J of Clinical Nutrition*, 1997)

SAS EXAMPLE 6: Large sample to illustrate measurement error result in multiple linear regression

```
data one;
n=200000;
bet0=0;
bet1=1;
bet2=-2;
sigsqy=1;

tau=.5;
lambda1=.75;
lambda2=.5;
sigsqx=.5;

do i=1 to n;
  z=rannor(0);
  c=ranbin(0,1,.45);
  x=tau + lambda1*z + lambda2*c + sqrt(sigsqx)*rannor(0);
  y=bet0 + bet1*x + bet2*c + sqrt(sigsqy)*rannor(0);
   output;
end;

proc reg;
 model y=z c;
run;
```

true β's

MEM parameters

W

→ MEM

→ TDM

```
                     The REG Procedure
                       Model: MODEL1
                  Dependent Variable: y

          Number of Observations Used       200000

                   Parameter Estimates
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.50633	0.00369	137.13	<.0001
z	1	0.74449	0.00275	270.97	<.0001
c	1	-1.50801	0.00551	-273.88	<.0001

λ^* β_1

λ^* β_2

Note that $\hat{\beta}_1{}^* \cong \beta_1\lambda_1$ and $\hat{\beta}_2{}^* \cong \beta_1\lambda_2 + \beta_2$, where β_1, β_2, λ_1, and λ_2 were set in the simulation!

NOTE: This sort of "corrected estimator" approach also extends to other important types of regression models, such as:

LOGISTIC REGRESSION for binary outcome data

(e.g., Rosner et al., 1990 *Am J of Epidemiology*)

COX PROPORTIONAL HAZARDS REGRESSION for survival data

(e.g., Prentice, 1982, *Biometrika*; Armstrong, 1990 *Am J of Epidemiology*)

The basic idea is the same, i.e.,

1) Fit the "naïve" regression model with (W, C) as predictors to your "main study" sample

2) Fit an MEM, e.g., regress X on (W, C), to your validation study sample

The estimated parameters and the variance-covariance matrices from these two models are used to obtain the "corrected" estimates and their standard errors (via delta method)

* For the approach to apply in logistic and Cox regression, certain assumptions are needed:

a) "rare" disease
b) "low" relative risk
c) "small" to "moderate" measurement error

⇒ For an accessible review, see Spiegelman et al., *Am J of Clinical Nutrition*, 1997

⇒ For more technical insights, see Liang and Liu, 1991 (book chapter)

Kuha, 1994 Sim

QL

↳ Estimating Functions

Now, how could we formulate a likelihood-based approach that would accomplish this same sort of adjustment?

Assume the same TDM and MEM:

TDM $$Y = \beta_0 + \beta_1 X + \beta_2 C + \varepsilon_y, \qquad \varepsilon_y \overset{\text{iid}}{\sim} N(0, \sigma_y^2)$$

(note normality assumed)

MEM $$X = \tau + \lambda_1 W + \lambda_2 C + \varepsilon_x, \qquad \varepsilon_x \overset{\text{iid}}{\sim} N(0, \sigma_x^2)$$

Observed data consist of:

Study Design

1) "Main" study sample of size n_{mn} : (Y, W, C) no X

2) External validation sample of size n_{val} : (X, W, C) → no Y

Let's formulate the likelihood for the data, conditional on (W, C):

⇒ Each main study observation contributes:

$$f(Y_i \mid W_i, C_i) = \int_{-\infty}^{\infty} f(Y_i, X_i \mid W_i, C_i)\, dX_i$$

integrating out the true X

$$= \int_{-\infty}^{\infty} f(Y_i \mid X_i, W_i, C_i)\, f(X_i \mid W_i, C_i)\, dX_i$$

non-diff

$$= \int_{-\infty}^{\infty} \underbrace{f(Y_i \mid X_i, C_i)}_{\text{TDM}}\, \underbrace{f(X_i \mid W_i, C_i)}_{\text{MEM}}\, dX_i \qquad \text{(assuming non-differential)}$$

$$= \int_{-\infty}^{\infty} \underbrace{\left\{ \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-(y_i-\mu_{yi})^2/2\sigma_y^2} \right\}}_{f(y|x,c)} \times \underbrace{\left\{ \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-(x_i-\mu_{xi})^2/2\sigma_x^2} \right\}}_{f(x|w,c)} dX_i$$

where $\underbrace{\mu_{yi} = \beta_0 + \beta_1 x_i + \beta_2 c_i}_{\text{TDM}}$ and $\underbrace{\mu_{xi} = \tau + \lambda_1 w_i + \lambda_2 c_i}_{\text{MEM}}$

$\boxed{\epsilon_{yi} \perp \epsilon_{xi}}$

$\Rightarrow$ Each external validation study observation contributes:

$$f(X_i \mid W_i, C_i) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-(x_i-\mu_{xi})^2/2\sigma_x^2}$$

So the complete likelihood (conditional on W, C) becomes:

\# in main study

$$\prod_{i=1}^{n_{mn}} \int_{-\infty}^{\infty} \left\{ \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-(y_i-\mu_{yi})^2/2\sigma_y^2} \right\} \times \left\{ \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-(x_i-\mu_{xi})^2/2\sigma_x^2} \right\} dx_i$$

\# in validation study

$$\times \prod_{i=1}^{n_{val}} \left\{ \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-(x_i-\mu_{xi})^2/2\sigma_x^2} \right\}$$

might be able to get rid of the integral.

In this linear TDM / linear MEM case (and sometimes for other forms of the TDM), it is possible for us to maximize the likelihood numerically.

SAS EXAMPLE 7: Illustration of the "correction" and ML methods on a simulated dataset containing main / external validation data

```
options ps=1000 ls=80;

data main;
nmain=500;
bet0=0;
bet1=1;
bet2=-2;
sigsqy=1;
```

500 main study obs

(β's)

```
alph=.5;
gam1=.75;
gam2=.5;
sigsqx=.5;
```

$\tau, \lambda_1, \lambda_2$

```
do i=1 to nmain;
  wmn=rannor(0);
  cmn=ranbin(0,1,.45);
  xmn=alph + gam1*wmn + gam2*cmn + sqrt(sigsqx)*rannor(0);  -> MEM
  y=bet0 + bet1*xmn + bet2*cmn + sqrt(sigsqy)*rannor(0);    -> TDM
   output;
end;
run;

data valid;
nval=250;
alph=.5;
gam1=.75;
gam2=.5;
sigsqx=.5;
```

250 external validation study obs.

```
do i=1 to nval;
  wval=rannor(0);
  cval=ranbin(0,1,.45);
  xval=alph + gam1*wval + gam2*cval + sqrt(sigsqx)*rannor(0);  -> MEM
   output;
end;
run;

proc reg data=main;
 model y=wmn cmn;
run;
```

1) "Naive regression"

Allows for "corrected" method

```
proc reg data=valid;
 model xval=wval cval;
run;
```

2) Fit MEM to validation data.

SAS IML code for Maximum Likelihood

```
proc iml worksize=70 symsize=250;

nmain=500; nval=250;

  use main;
  read all var{y} into y;
  read all var{wmn} into wmn;
  read all var{cmn} into cmn;
  close main;

  use valid;
  read all var{wval} into wval;
  read all var{cval} into cval;
  read all var{xval} into xval;
  close valid;
```

reading in data

```
** Define function that will be numerically integrated in likelihood **;

START FUNC(xi) global
(bet0,bet1,bet2,sigsqy,tao,lamb1,lamb2,sigsqx,wi,ci,yi,pi);

 muxi=tao+lamb1*wi+lamb2*ci;
 muyi=bet0+bet1*xi+bet2*ci;

 fun1=(1/sqrt(2#pi#max(sigsqy,1E-4)))#exp(-(yi-muyi)##2/(2#max(sigsqy,1E-4)));

 fun2=(1/sqrt(2#pi#max(sigsqx,1E-4)))#exp(-(xi-muxi)##2/(2#max(sigsqx,1E-4)));

 func_i=fun1#fun2;

 return(func_i);

FINISH;
```

specifies $\mathcal{L}$

```
** Likelihood equation for FULL ML method **;

START LIKELI1(parms) global
(nmain,nval,y,wmn,cmn,wval,xval,cval,sigsqx,sigsqy,bet0,bet1,bet2,
                            tao,lamb1,lamb2,yi,ci,wi,pi);

bet0=parms[1];
bet1=parms[2];
bet2=parms[3];
sigsqy=parms[4];

tao=parms[5];
lamb1=parms[6];
lamb2=parms[7];
sigsqx=parms[8];
```

TPM

β_1, β_2 of main interest

MEM

```
pi=2*arsin(1);


* External validation study contributions to likelihood ;

func_val=j(nval,1,999);

do t=1 to nval;

 func_val[t,]=(1/sqrt(2#pi#max(sigsqx,1E-4)))#
             exp(-(xval[t,]-tao-lamb1#wval[t,]-
lamb2#cval[t,])##2/(2#max(sigsqx,1E-4)));
end;


* Main study contributions to likelihood ;

        func_mn = j(nmain,1,.);

        do u = 1 to nmain;

                yi = y[u,1];
                ci = cmn[u,1];
                wi = wmn[u,1];

        A   = {.M  .P};

        CALL QUAD(f,'FUNC",A);
                func_mn[u,1]=f;

        end;

    *  print func_mn;

 m2loglik=-2#sum(log(func_mn)) + -2#sum(log(func_val));

   **print m2loglik;

   return(m2loglik);

FINISH LIKELI1;


*****************************************************************************
The following is the main body of the program (which calls the
optimization function, computes the Hessian, etc.)
*****************************************************************************;

START COMP; ** Maximum likelihood method **;

   *create vector of initial parameter estimates for function;

   parms=.2||1.3||-1.5||1.4||.75||.5||.25||.3;

   *options vector for minimization function;
```

$(-\infty, \infty)$

numerical integration

the function we defined

$-2 \ln L$

Calls optimization

initial values

```
   option={0 3};

  **matrix of lower(row 1) and upper(row 2) bound
   constraints on probabilities**;

   con={. . . 0 . . . 0,
        . . . . . . . .};

   *call function minimizer in IML;

   call nlpqn(rc,xres,"likeli1",parms,option,con);

   *create vector of mle's computed using function minimizer;


parms=xres`;

print parms;


   *compute numerical value of Hessian( and covariance matrix)
   using mles calculated above ;

  call NLPFDD(crit,grad,hess,"likeli1",parms);

  cov_mat=2*inv(hess);
  se_vec1=sqrt(vecdiag(cov_mat));

   print se_vec1;


FINISH COMP;

 run COMP;

quit;
```

Quasi-Newton

approximate 2nd derivs ⇒ Hessian, observed info matrix.

SAS OUTPUT:

```
The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Used          500
```

n_{main}

1) "naive" regression

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.51618	0.07645	6.75	<.0001
wmn	1	0.79009	0.05219	15.14	<.0001
cmn	1	-1.44800	0.11294	-12.82	<.0001

$\hat{\beta}_1^*$ $\hat{\beta}_2^*$

```
The REG Procedure
Model: MODEL1
Dependent Variable: xval

Number of Observations Used          250
```

n_{val}

2) MEM x vs (w, c)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.53223	0.05792	9.19	<.0001
wval	1	0.70166	0.04530	15.49	<.0001
cval	1	0.45419	0.08766	5.18	<.0001

By "correction" method, we have:

$$\hat{\beta}_1 = \hat{\beta}_1^*/\hat{\lambda}_1 = .790/.702 = 1.126$$

and

$$\hat{\beta}_2 = \hat{\beta}_2^* - \hat{\beta}_1\hat{\lambda}_2 = -1.449 - 1.126(.454) = -1.959$$

Also, delta method gives: $$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \frac{\left(\widehat{\mathrm{Var}}(\hat{\beta}_1^*) + \hat{\beta}_1^2\,\widehat{\mathrm{Var}}(\hat{\lambda}_1)/\hat{\lambda}_1^2\right)}{\hat{\lambda}_1^2}$$

$$= [.052^2 + 1.126^2(.045)^2/.702^2]/.702^2 = .016 \quad \Rightarrow \quad \boxed{\widehat{SE}(\hat{\beta}_1) = 0.127}$$

IML Output for Numerical Maximum Likelihood Analysis

Optimization Start
Parameter Estimates

N	Parameter	Estimate	Gradient Objective Function	Lower Bound Constraint	Upper Bound Constraint
1	X1	0.200000	416.334356	.	.
2	X2	1.300000	274.288317	.	.
3	X3	-1.500000	229.541040	.	.
4	X4	1.400000	-46.177305	0	.
5	X5	0.750000	767.095075	.	.
6	X6	0.500000	-475.242554	.	.
7	X7	0.250000	309.382471	.	.
8	X8	0.300000	-726.570035	0	.

Value of Objective Function = 2432.7995114

The SAS System

Dual Quasi-Newton Optimization

Dual Broyden - Fletcher - Goldfarb - Shanno Update (DBFGS)
Gradient Computed by Finite Differences

Parameter Estimates	8
Lower Bounds	2
Upper Bounds	0

Optimization Start

Iter	Rest arts	Func Calls	Act Con	Objective Function	Obj Fun Change	Max Abs Gradient Element	Step Size	Slope Search Direc
1	0	3	0	2247	186.0	239.2	0.0447	-17413
2	0	6	0	2204	42.3299	135.1	0.280	-446.5
3	0	8	0	2199	5.8373	110.3	0.148	-70.174
.								
.								
.								

GCONV convergence criterion satisfied.
~~NOTE~~: At least one element of the (projected) gradient is greater than 1e-3.

Value of Objective Function = 2162.8544808 $-2\ln L$, at the MLE's

```
parms

-0.083118
1.1260542
-1.959428
0.9771172
0.5322234
0.7016479
0.4541908
0.4664709
```

$= \hat{\beta}_1$

$= \hat{\beta}_2$

identical to "corrected" estimates

```
se_vec1

0.1132073
0.1035315
0.1555587
 0.156429
0.0575707
0.0450221
0.0871346
0.0417232
```

$= \hat{SE}(\hat{\beta}_1)$

vs. .127 from Δ-method

* Note that ML agrees with the "correction" method for parameter estimation, with a small difference in the estimated standard error of $\hat{\beta}_1$

Now, let's look at measurement error in a predictor (X) in logistic regression:

NOTE: This is a special case of the problem considered by Rosner et al. (1990) and others. Rosner et al. approach the problem by obtaining the convergence result associated with the "naïve" regression, and "correcting" the naïve estimate via internal validation data.

Here, let's first set up the problem from a maximum likelihood (ML) perspective:

Assume a logistic regression model with $Y_i \mid X_i \sim \text{Bernoulli}(p_i)$

Model of interest ("TDM"): $$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i \quad ,$$

where $$p_i = \frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)}$$

Let's assume the classical error model: $W_i = X_i + U_i$, $U\text{'s} \overset{\text{iid}}{\sim} (0, \sigma_u^2)$

Note that we can say W is a "surrogate" for X $\Rightarrow$ non-diff. error

Further, let's go ahead and assume the following:

$$X_i\text{'s} \overset{\text{iid}}{\sim} N(\mu_x, \sigma_x^2) \quad \text{and} \quad U_i\text{'s} \overset{\text{iid}}{\sim} N(0, \sigma_u^2) \quad \text{and} \quad U_i \perp X_i \text{ for all i}$$

i) Let's first assume we have <u>external</u> validation data

$\Rightarrow$ "Main" study data consist of (Y, W) pairs measured on a total of n_m subjects

no X

$\Rightarrow$ Validation study data consist of (X, W) pairs measured on a total of n_v subjects

external

no Y

$\Rightarrow$ The two data sources are completely independent

$\Rightarrow$ By necessity, assume "transportability" and non-differential measurement error

* Can attempt to set up the likelihood in a similar fashion as on pp. 9-10:

$\Rightarrow$ Each main study observation contributes:

$$f(Y_i \mid W_i) = \int_{-\infty}^{\infty} f(Y_i, X_i \mid W_i)\, dX_i$$

non-diff.

$$= \int_{-\infty}^{\infty} \underbrace{f(Y_i \mid X_i)}_{\text{TDM}}\, \underbrace{f(X_i \mid W_i)}_{\text{MEM}}\, dX_i \qquad \text{(assuming non-differential)}$$

* Note: From conditional MVN results, we have $X_i \mid W_i \sim$ Normal, where

$$E(X_i \mid W_i) = \mu_X + \frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2}(w_i - \mu_X) \quad \text{and} \quad Var(X_i \mid W_i) = \sigma_X^2 - \frac{\sigma_X^4}{\sigma_X^2 + \sigma_u^2}$$

$\Rightarrow$ Each external validation study observation contributes:

$$f(X_i \mid W_i)$$

Can attempt to maximize the complete likelihood numerically (e.g., use a program similar to the one shown earlier in these notes)

ii) Let's next assume we have internal validation data

no X

$\Rightarrow$ "Main" study data consist of (Y, W) pairs measured on a total of n_m subjects

$\Rightarrow$ Validation study data consist of (Y, X, W) triples measured on a total of n_v subjects

gold standard

$\Rightarrow$ The two data sources are no longer distinct – but assume validation study subjects are selected at random and we have independence across subjects

$\Rightarrow$ This time, don't have to worry about "transportability"; also, in theory we could assess whether there was a need to model differential measurement error

Overall likelihood should look something like:

$$L = \overbrace{\left\{\prod_{i=1}^{n_m} \int_{-\infty}^{\infty} \underbrace{f(Y_i \mid X_i)}_{\text{TDM is now based on logistic regression}} f(X_i \mid W_i)\, dX_i \right\}}^{\text{main study subjects}} \times \overbrace{\left\{\prod_{i=1}^{n_v} f(Y_i \mid X_i) f(X_i \mid W_i)\right\}}^{\text{validation subjects}}$$

NOTE: A possible issue is that in the logistic regression case, the integral needed to specify the likelihood contributions can be much more difficult to deal with numerically!

What are some ways in which people have tried to deal with this problem?

a) Regression calibration: As in our linear regression example, could replace the unknown exposure X by its conditional expectation E(X | W)

$\Rightarrow$ Appears to work well under certain conditions (see pg. 8), but it is an approximation and can fail badly if the true exposure effect in the TDM (β) and/or the measurement error variance is large

(see, e.g., Thurston, Spiegelman and Ruppert, 2003)

Kuha, 1994 Sim $\rightarrow$ $\beta\lambda$

b) Probit approximation to the logistic function: It has been shown that $H(t) \cong \Phi(t/k)$, where k is a constant, $H(t) = \dfrac{\exp(t)}{1+\exp(t)}$. Using this result, one can derive that

$$\Pr(Y_i = 1 \mid W_i) \cong H\left(\frac{\alpha + \beta E(X_i \mid W_i)}{\sqrt{1 + \beta^2 \mathrm{Var}(X_i \mid W_i)/k^2}} \right)$$

$\Rightarrow$ This gets rid of the integral! Usually the recommended value of k is about 1.7

(see, e.g., Carroll et al., 1984; Carroll et al. text, 2006; Lyles and Kupper, 2012)

class #1

c) Pseudo-likelihood: Numerical problems may be reduced by estimating nuisance parameters in the MEM separately, and then inserting those estimates in place of those parameters in the likelihood. Does not get rid of the integral, but often improves numerical stability

(theory: Gong and Samaniego, 1981; ME application: e.g., Lyles and Kupper, 2012)

→ class #1

d) Quasi-likelihood: See basic form of estimating equations in Class #1 notes; this is like a refined version of regression calibration where Var(X|W) appears in the denominator – also may have some robustness advantages

(see, e.g., Liang and Liu, 1991; Lyles and Kupper, 1997)

↓ book chapter

e) Take advantage of modern software (e.g., SAS NLMIXED procedure) to deal with the integral needed for full ML

→ logistic regression TDM

(see Messer and Natarajan, 2008 → intriguing because may allow for multiple covariates measured with error in a full ML approach)

Further Things to Think About:

1) For internal validation data, what about trying Greenland's idea of weighting two closed-form estimators (see Class # 2 notes)?

2) Consider the case where 2 (or more) covariates are measured with error. How could we proceed?

a) What happens if we make multivariate normality assumptions (see Spiegelman et al., *Am J of Clinical Nutrition*, 1997)?

b) Could we relax the MVN assumption and still make progress?