

# Measurement Error in Covariates: Session 2, May 27, 2011

Raymond J. Carroll

Department of Statistics

Faculty of Nutrition

**Institute for Applied Mathematics and Computational  
Science**

Texas A&M University

# My Goal Today

- Introduce the types of data for classical measurement error modeling
- The functional approach
- Regression calibration
- SIMEX
- Instrumental variables

# What Will not be Here Today

- Fancy new methods
- Technical details

# Later Lectures

- Measurement error analysis
  - Maximum likelihood
  - Bayes

# Big Theme

- We always have a risk model relating disease to true exposure
- We always have a measurement error model relating observed exposure and true exposure
- The literature is split into methods that have an exposure model (**structural**) versus those that have no exposure model (**functional**)

# The Triple Whammy of Measurement Error

- **Bias** in parameter estimation, which with multiple covariates can lead to incorrect null hypothesis testing
- **Loss of Power** for detecting signals
- **Masking** the features of the data: measurement error causes the model of the observed data to change

# Notation

- Response:  $Y$
- True Covariates Subject to Measurement Error:  $X$
- Covariates Measured Exactly:  $Z$
- Observed Proxies for  $X$ :  $W$

# Emphasis

- I will focus on continuous exposures
- Measurement error with a **categorical exposure  $X$  subject to misclassification** is an even harder topic
- In those problems, the real issue is to get data to estimate the sensitivity and specificity of the surrogate exposure  $W$ , e.g..  
 $\Pr(W = 1|X = 1)$ ,  $\Pr(W = 0|X = 0)$
- The impact of misclassification is often profound

# Nondifferential Measurement Error

- **Definition:** If you can observe  $X$ , you would not bother collecting  $W$
- Technically, the proxies  $W$  are statistically independent of  $Y$  given  $(X, Z)$
- **All my lectures assume nondifferential measurement error**
- The analysis of data subject to differential measurement error is a very different subject

# Programs

- Various Stata, SAS and R programs are at the web site [http://www.stat.tamu.edu/~carroll/matlab\\_programs/MECourse/programs.php](http://www.stat.tamu.edu/~carroll/matlab_programs/MECourse/programs.php)
- Donna Spiegelman also has some SAS macros available at <http://www.hsph.harvard.edu/faculty/donna-spiegelman/software/>

# Some Key Big Picture Ideas

- Thanks to Sholom Wacholder for bringing these up

# Underlying Assumptions

- Nondifferential measurement error of exposure by disease status
- Known model of error (classical, Berkson, etc.) with estimable parameters, and possibly estimable scale
- The data needed to estimate the measurement error parameters and scale (transformation)

# Underlying Assumptions

- In practice, many methods are robust to violations of some of the assumptions
- For example, with nondifferential error, while methods may assume normally distributed classical error, violations of the normality assumption typically do not matter much (except maybe in theory)
- This is why much of the literature is a bit cavalier with such assumptions

# Underlying Assumptions

- Differential measurement error is a real problem
- Learning the error structure is a statistical problem
- Designing validation/calibration studies to learn the error structure is a statistical problem
- If there is a gold standard, or unbiased replicates for an alloyed gold standard, then much can be and has been done

# Underlying Assumptions

- As we have learned from Sholom Wacholder, differential measurement error can wreak havoc with conclusions from observed data and from analysis based on nondifferential assumptions
- The data requirements needed to deal with differential measurement error can include gold standards
- With continuous exposures and differential measurement error, instrumental variables can help in simple models

# Data Needed for a Measurement Error Analysis

# Conundrum

- The Classical Model says that  $W = X + U$ ,

$$U = \text{Normal}(0, \sigma_u^2).$$

- In general, **The measurement error variance  $\sigma_u^2$  cannot be estimated from just  $(Y, W, Z)$  data**
- Question: **what data are needed** to estimate  $\sigma_u^2$ ?
- $W$  could also be multivariate

# Solution #1: Validation Data

- At least in principle, in some cases, one can effectively observe  $X$  in a sub-study
- This is called a **validation study**
- Validation studies are **beautiful** things
- They are **rare**, especially if  $X$  is a long-term exposure
- Of course, if such data exist,  $\sigma_u^2 = \text{var}(W - X)$ .

# Solution #1: Validation Data

- Validation data, which include  $X$ , also allow us to estimate the distribution of true exposure
- They also allow us to understand **whether the classical error model actually holds!**
- Validation study data are really data with **missing data**, in  $X$ , although they are not typical missing data problems because most of the  $X$ 's are missing.

## Solution #2: Replication Data

- In many cases, it is possible to observed **replicated  $W$  data**
- Thus, for the  $i^{\text{th}}$  person, we observed  $(W_{i1}, \dots, W_{im})$  with  $W_{ij} = X_i + U_{ij}$ .
- Replication data allow **easy estimation** of  $\sigma_u^2$  through **ANOVA calculations**
- They also allow **data checking** to see if the additive model with homoscedastic error holds (**details not given, this is an overview**).

## Solution #2: Replication Data

- Replicated biomarkers or 24hr recalls
- Replicated blood pressure measurements
- Replicated monitoring equipment

## Solution #3: Instrumental Variables

- Often forgotten, but widely used in econometrics
- These are variables  $T$  which have the following properties (hopefully)
  - There are **correlated with true exposure**  $X$
  - They are **nondifferential**
  - They are **independent of the measurement error**  $U$
- Convincing oneself (**or referees**) that  $T$  is a proper instrument is hard, because **it cannot be verified numerically**.

# Structural/Functional, Regression Calibration, Non-additive Errors

## Outline

- Functional versus Structural modeling
- Regression calibration
- Multiplicative error

# Functional And Structural Modeling: Classical Error Models

- There are three broad types of approaches
- **Functional modeling**: No assumptions made about the  $X$ 's (could be random or fixed)
- **Classical structural modeling**: Strong parametric assumptions made about the distribution of  $X$ . Generally normal, lognormal or gamma.
- **Flexible structural modeling**: Structural, but flexible parametric family. Tries to get the best of both worlds.

# Advantages of Functional Models

- **FUNCTIONAL**: No need to perform extensive sensitivity analyses.
- Many functional methods are simple to implement (and some are computed using little more than standard software).
- Functionality focuses emphasis on the error model.
- Because of “latent-model” robustness, a functional analysis serves as a useful check on a parametric structural model.

# Disadvantages of Functional Models

- **Efficiency**: Significant loss of efficiency for missing data, thresholds, nonparametric regression
- **Likelihood Paradigm**: Functional methods are not maximum likelihood as generally known.

# Some Functional Methods

- **Regression Calibration**/Substitution
  - Replaces true exposure  $X$  by an estimate of it **based only on covariates** but not on the response.
  - In linear model with additive errors, this is the classical **correction for attenuation**.
  - In Berkson model, one simply ignores the measurement error.
- **SIMEX** is a fairly general functional method.
  - It assumes only that you have an error model and that you can “add on” measurement error to make the problem worse.

## Regression Calibration—Basic Ideas

- **Key idea:** replace the unknown  $X$  by  $E(X|Z, W)$  which depends only on the known  $(Z, W)$ .
- This provides an approximate model for  $Y$  in terms of  $(Z, W)$ .
- Called the “conditional expectation approach” by Lyles and Kupper (1997)

# Regression Calibration—Basic Ideas

- Often works well for logistic, Poisson and Cox Models
- Depends on the measurement error being “**not too large**” in order for the approximation to be sufficiently accurate.
- For some models, if  $\text{var}(X|Z, W)$  is constant then the only bias due to the regression calibration approximation is in the intercept parameter.

# Regression Calibration—Basic Ideas

## Why does regression calibration work?

- Suppose

$$Y = X^T \beta_x + Z^T \beta_z + \epsilon$$

- Then

$$E(Y|Z, W) = E(X|Z, W)^T \beta_x + Z^T \beta_z$$

- Therefore, regression of  $Y$  on  $E(X|Z, W)$  and  $Z$  gives unbiased estimates

# The Regression Calibration Algorithm

- Using various brands of data, develop a model  $E(X|Z, W) = m(Z, W, \gamma)$  and estimate  $\gamma$ .
- Replace  $X$  by  $m(Z, W, \hat{\gamma})$  and run your favorite analysis.
- Obtain standard errors by the bootstrap or the “sandwich method.”
- In linear regression, regression calibration is equivalent to the “correction for attenuation.”

# The Regression Calibration Algorithm: Logistic Regression

- Let  $H$  be the logistic distribution function
- Define the attenuation

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

- Then one can show that

$$\Pr(Y = 1|W) \approx H \left\{ \frac{\beta_0 + \beta_x E(X|W)}{(1 + \beta_x^2 \lambda k^2 \sigma_u^2)^{1/2}} \right\}$$

- The denominator is often very close to 1.0

# Estimating The Calibration Function

- Need to estimate  $E(X|Z, W)$ .
- Easy case: internal validation data
- Then one can simply regress  $X$  on  $(Z, W)$  to estimate  $E(X|Z, W)$
- Use this estimate (transport it) and **replace  $X$  by it** for the non-validation data.
- Run a single regression of  $Y$  on  $(Z, X)$  for validation data and  $Y$  on  $\{Z, E(X|Z, W)\}$  for nonvalidation data.
- Use a different intercept for the two data.

# Estimating The Calibration Function: Instrumental Data

- In the **NCI-AARP Diet and Cancer Study**, fat and energy were measured by an FFQ. This is  $W$ , and is bivariate.
- $Y$  = breast cancer in 10-year followup,  $Z$  are other factors
- The sample size was  $\approx 250,000$  women
- They did a sub-study of 1,000 women, where they used two 24-hour recalls to obtain what we call  $T$ .
- The key feature of this is that  $T = X + U$ , and  $U$  is uncorrelated with  $X$  and  $Z$ .
- Thus,  $T$  is an **unbiased instrument**

# Estimating The Calibration Function: Instrumental Data

- An **internal unbiased instrument** has the property that  $E(T|Z, X) = E(T|Z, X, W) = X$ .
- If  $T$  is expensive to measure, then  $T$  might be available for only a subset of the study.  $W$  will generally be available for all subjects. **This is the NCI-AARP Study.**
- The **rcal algorithm** is to replace  $X$  by the estimate of  $E(X|Z, W)$
- The **calibration function estimate** is found by regressing  $T$  on  $(Z, W)$  in the sub-study.

# Estimating The Calibration Function: Replication Data

- In the **Framingham Heart Study**, we have replicates of (transformed) systolic blood pressure and serum cholesterol
- These will be labelled as  $W_{ij}$ , with  $j = 1, 2$
- Simple ANOVA-type calculations allow estimation of the **calibration function**
- See appendix for details

# Multiplicative Error

- The **multiplicative lognormal error** model is

$$W = X U, \quad \log(U) \sim N(0, \sigma_u^2) \quad (1)$$

- If  $\log(X)$  is normal then we can convert (1) to a **classical additive error** model

$$\log(W) = \log(X) + \log(U)$$

- If we say that  $W_* = \log(W)$ ,  $X_* = \log(X)$ ,  $U_* = \log(U)$ , then we have our old friend:  $W = X + U$ , where  $U_b$  is  $N(0, \sigma_b^2)$  and independent of  $W$
- Thus, if we are willing to work in the log-scale for risk prediction, then regression calibration applies without change.

# Multiplicative Error

- There is considerable controversy about whether to use the transformed scale to estimate risk.
- This is routinely done in nutrition
- Main Advantage: Convenience, because regression calibration can be used.

# Multiplicative Error

- **Disadvantage** of risk modeling in the log scale: Users do not think in terms of logarithms of a predictor.
- The **multiplicative error** model **with outcome linear in exposure** is well-supported by empirical work
- Lyles and Kupper (1997, *Biometrics*):
  - “there is much evidence for this model”
  - “the more biologically relevant predictor is true mean exposure on the **original** scale”

# Multiplicative Error

- Often, **remarkably**, regression calibration with an unbiased instrument **T** works quite well, **ignoring the multiplicative nature of the measurement error**
- We have seen this in nutritional epidemiology
- There is also a substantial literature on direct modeling that attempts to remove most of the bias (see appendix)

# Simex And Instrumental Variables

# About Simulation Extrapolation

- **Functional**: no assumptions about the true  $X$  values
- **Not model dependent**: like bootstrap and jackknife
- **Complex Problems**: Can attack complex problems not available to Regression Calibration
- **Computer intensive**
- **Not a Panacea**: Bias reduction, but rarely bias elimination

# The Key Idea

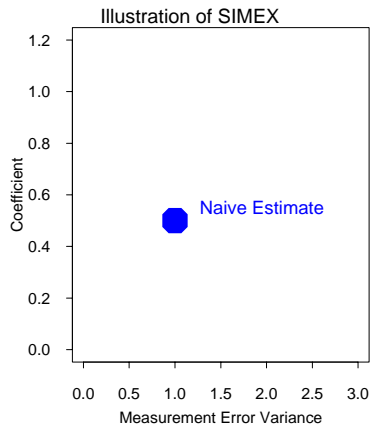
- **A Simulation Experiment**: effects of measurement error on an estimator can be studied with a simulation experiment in which **additional** measurement error is added
- **“Response variable”**: the estimator under study
- **“Independent factor”**: the measurement error variance
- **“Factor levels”**: the variances of the added measurement errors
- **Objective**: study how the estimator depends on the variance of the measurement error

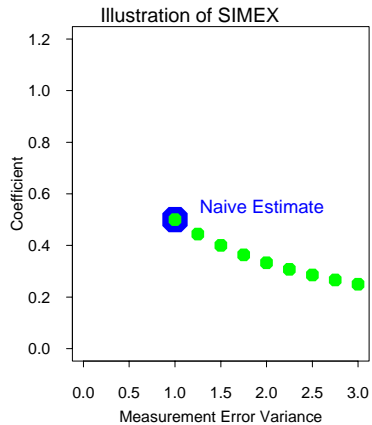
# Outline Of The SIMEX Algorithm

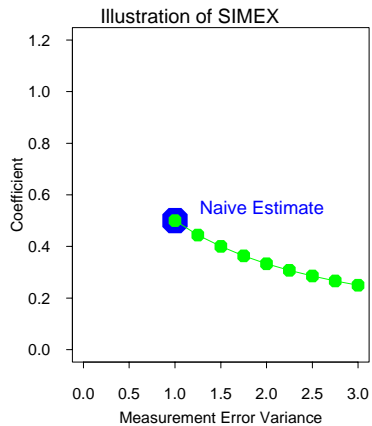
- Add measurement error to variable measured with error
- $\Lambda$  controls amount of added measurement error
- $\sigma_u^2$  increased to  $(1 + \Lambda)\sigma_u^2$

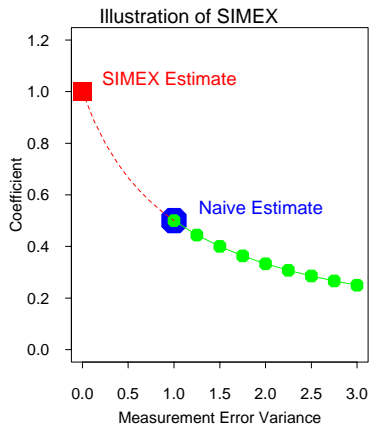
# Outline Of The SIMEX Algorithm

- Recalculate estimates: called pseudo-estimates
- Average over many simulations to remove Monte Carlo variation
- Plot Monte Carlo average pseudo-estimates versus  $\Lambda$
- $\sigma_u^2$  increased to  $(1 + \Lambda)\sigma_u^2$
- Extrapolate to  $\Lambda = -1$ , the case of no measurement error









## SIMEX: Adding Measurement Error (Details)

- **Pick** a few  $\Lambda$  indicating **increased** measurement error
- **Commonly**: For example, choose  $\Lambda = (0.5, 1.0, 1.5, 2.0)$
- **Number of simulations**: do  $B$  simulations, often  $B = 100$ .
- **Pseudo Data**: For  $b = 1, \dots, B$ , create the  $b^{\text{th}}$  pseudo data set:

$$W_{b,i}(\Lambda) = W_i + \sqrt{\Lambda} \text{Normal}(0, \sigma_u^2)_{b,i}$$

- **Refit the pseudo data**: Obtain the  $b^{\text{th}}$  pseudo estimate

$$\hat{\theta}_b(\Lambda) = \hat{\theta}(\{Y_i, W_{b,i}(\Lambda)\}_1^n)$$

# Simulation And Extrapolation Algorithm: Adding Measurement Error (Details)

- **Average** the pseudo estimates over the  $B$  simulations

$$\hat{\theta}(\Lambda) = B^{-1} \sum_{b=1}^B \hat{\theta}_b(\Lambda) \approx E \left( \hat{\theta}_b(\Lambda) \mid \{Y_j, X_j\}_1^n \right)$$

- **Plot** the pseudo estimates as a function of  $\Lambda$
- **Recall**:  $\Lambda = -1$  means no measurement error
- **Extrapolate** back to  $\Lambda = -1$

# Simulation And Extrapolation Steps: **Extrapolation**

- In General: (multiple  $\Lambda$  points)
- **Linear** Fit a linear model of the pseudo estimates to  $\Lambda$
- **Quadratic**:  $a + b\Lambda + c\Lambda^2$
- **Rational Linear**:  $(a + b\Lambda)/(c + \Lambda)$  (exact for linear regression)

# Framingham Data: Measurement Error In Systolic Blood Pressure

- **Response**: Indicator of CHD
- **Unobserved Predictor**:  $X$  = long-term average of logarithm of (SBP - 50)
- **Observed Predictor**:  $W$  = observed logarithm of (SBP-50)
- **Other covariates**:  $Z$  = age, smoking status, cholesterol

# Framingham Data in Stata

- **Classical Error**: from additional data,  $\sigma_u^2$  is estimated.
- **Sample Size**: 1,600 +, so many degrees of freedom
- The plots on the following page illustrate the simulation extrapolation method for estimating the parameters in the logistic regression model

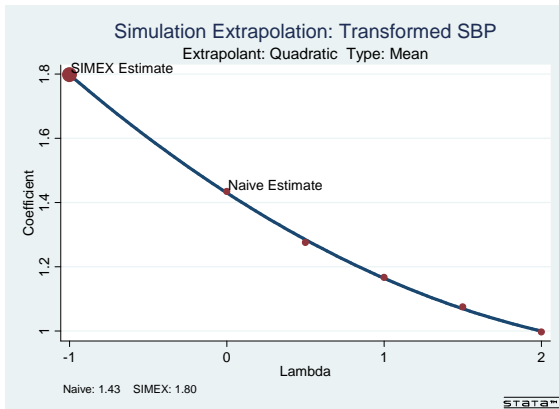
```
. simex (firstchd=age smoke) (w3: lsbp2 lsbp3), bstrap family(binomial) link(logit)
Estimated time to perform bootstrap: 2.18 minutes.
```

```
Simulation extrapolation          No. of obs      =      1615
                                   Bootstraps reps =       199

Residual df =      1611           Wald F(3,1611)  =       27.98
                                   Prob > F         =       0.0000
```

```
Variance Function: V(u) = u(1-u)          [Bernoulli]
Link Function      : g(u) = log(u/(1-u))   [Logit]
```

firstchd	Bootstrap				
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0553085	.0105567	5.24	0.000	.0346022 .0760148
smoke	.6323024	.2709621	2.33	0.020	.100827 1.163778
Cholesterol	.0097426	.0018476	5.27	0.000	.0061186 .0133665
Transformed SBP	1.797655	.494887	3.63	0.000	.826965 2.768346



# Instrumental Variables: Computation

- **Stata**: has many methods for instrumental variable fitting
- **qvf**: This is their program for IV estimation in the nondifferential case
- **Econometricians**: call them exogenous
- **Computation here**: all done in Stata
- **Differential**: The endogenous case for linear regression can be done in Stata

# Instrumental Variables: Rationale

- **Remember**,  $W = X + U$ ,  $U = \text{Normal}(0, \sigma_u^2)$ .
- We need to estimate  $\sigma_u^2$
- **Best way**: observe  $X$  on a subset of the data.
- **Next best**: Replication
- If these are indeed replicates, then we can estimate  $\sigma_u^2$  via a components of variance analysis.
- **Third way**: **Instrumental Variables**.

# What Is An Instrumental Variable?

$$Y = \beta_0 + \beta_x X + \epsilon;$$

$$W = X + U;$$

$$U \sim \text{Normal}(0, \sigma_u^2).$$

In linear regression, an instrumental variable  $T$  is a random variable which has three properties:

- **Independence of errors**:  $T$  is independent of  $\epsilon$  and  $U$
- **Predictiveness**:  $T$  is related to  $X$ .
- **Surrogacy**: You only measure  $T$  to get information about measurement error: it is not part of the model.
- **The Problem**: Whether  $T$  qualifies as an instrumental variable can be a difficult question.

# Using Instrumental Variables: Motivation

$$Y = \beta_0 + \beta_x X + \epsilon;$$

$$W = X + U;$$

$$T = \text{instrument.}$$

- **Cool Fact:**  $\beta_x = \text{cov}(Y, T) / \text{cov}(W, T)$
- **Conclusion:** Estimate  $\beta_x$  by substituting sample covariances

# Using Instrumental Variables

- All instrumental variable methods use some version of relating  $\text{cov}(Y, T)$  to  $\text{cov}(W, T)$  with some version of division by the latter.
- **Danger**: The division causes increased variability.
- **Weak Instrument**:  $T$  is not well-correlated with  $X$ , then the IV estimate very unstable.
- Regression calibration is almost always more powerful than IV methods

# Logistic Regression Example

$Y$  = Evidence of Coronary Heart Disease (binary)

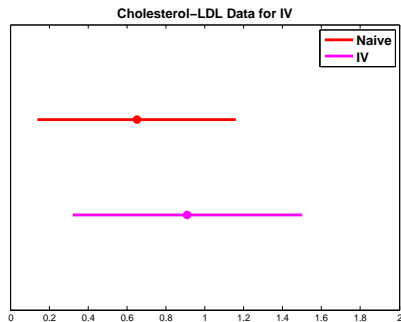
$W$  = Cholesterol Level

$T$  = LDL

$Z$  = Age and Smoking Status

- **Is it an Instrument?** This is not clear
- LDL might not be a surrogate, i.e., it might add information about CHD even if true long-term mean cholesterol were known
- There are no replicates here to compare with, **hence no regression calibration**

# CHD-Cholesterol With LDL as an Instrument



# Logistic Regression Example: Results

- We used LDL/100, Cholesterol/100, Age/100
- The naive logistic regression leads to the following analysis:
  - Slope and bootstrap s.e. for Cholestrol: 0.65 and 0.29
  - Slope and bootstrap s.e. for Smoking: 0.065 and 0.260
  - Slope and bootstrap s.e. for Age: 7.82 and 4.26
- The IV logistic regression leads to the following analysis:
  - Slope and bootstrap s.e. for Cholestrol: 0.91 and 0.33
  - Slope and bootstrap s.e. for Smoking: 0.056 and 0.259
  - Slope and bootstrap s.e. for Age: 7.79 and 4.31
- Note here how only the coefficient for cholesterol is affected by the measurement error.

# STATA Function cme

- cme estimates generalized linear models with covariate measurement error by maximum likelihood.
- **3 submodels**: outcome model, measurement model, true covariate model.
- The outcome model is a generalized linear model
- Classical additive measurement error model
- The true covariate model is a linear regression of the true covariate on the observed covariates.

## STATA Function qvf

- qvf fits generalized linear models using IRLS (maximum quasi-likelihood) and is similar in syntax to glm.
- Results of qvf and glm (using IRLS) should be very similar.
- qvf has support for instrumental variables and implements a very fast built-in bootstrap (different from the regular Stata bootstrap command).

# STAT Function rcal

- rcal fits generalized linear models for measurement error data using regression calibration.
- rcal allows one or more (see comments) covariates measured with errors and uses regression calibration to estimate the missing covariates.
- It will allow replicate data or a user specified measurement error covariance matrix.
- It implements a very fast internal bootstrap (different from the regular Stata bootstrap command).

## STATA Function simex

- simex fits generalized linear models for measurement error data using SIMEX.
- `simex` allows one or more (see comments) covariates measured with errors and uses simulation extrapolation to estimate the missing covariates.
- It will allow replicate data or a user specified measurement error covariance matrix.
- It supports a very fast internal bootstrap (different from the regular Stata bootstrap command).

## Framingham Analysis Using [rcal](#)

- Regress CHD on age, smoking status and transformed systolic blood pressure (SBP)
- There are two replicate error-prone measurements of true SBP
- **Special Note**: We use both replicates, see Chapter 5.4
- Logistic regression is used.
- Bootstrap standard errors are computed.

# Framingham Analysis Using rcal

- Response  $Y = \text{firstchd}$
- Replicates of error-prone covariate:  $(W_1, W_2) = \text{lsbp2}, \text{lsbp3}$
- Default name of the mean of these replicates is  $W3$
- Exact covariates Z: age, smoke
- STATA Statement: `rcal (firstchd=age smoke) (w3: lsbp2 lsbp3), bstrap family(binomial) link(logit)`

# rcal Output

```
. rcal (firstchd=age smoke) (w3: lsbp2 lsbp3), bstrap family(binomial) link(logit)
```

```
Regression calibration          No. of obs      =      1615
                               Bootstrap reps   =       199

Residual df =      1611        Wald F(3,1611)   =      26.88
                               Prob > F         =      0.0000
                               (IRLS EIM)        Scale param     =      .941383

Variance Function: V(u) = u(1-u)      [Bernoulli]
Link Function      : g(u) = log(u/(1-u)) [Logit]
```

firstchd	Bootstrap				
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0512144	.0103718	4.94	0.000	.0308709 .071558
smoke	.5804847	.2312271	2.51	0.012	.1269472 1.034022
w3	2.081765	.4308277	4.83	0.000	1.236723 2.926806
_cons	-14.53173	1.804268	-8.05	0.000	-18.07069 -10.99277

# SIMEX in STATA

- **The STATA Statement:**
- **simex** (firstchd=age smoke) (w3: lsbp2 lsbp3), bstrap family(binomial) link(logit)

# SIMEX Output

```
. simex (firstchd=age smoke) (w3: lsbp2 lsbp3), bstrap family(binomial) link(logit)
```

```
Estimated time to perform bootstrap: 2.18 minutes.
```

```
Simulation extrapolation          No. of obs      =      1615
                                   Bootstraps reps  =       199

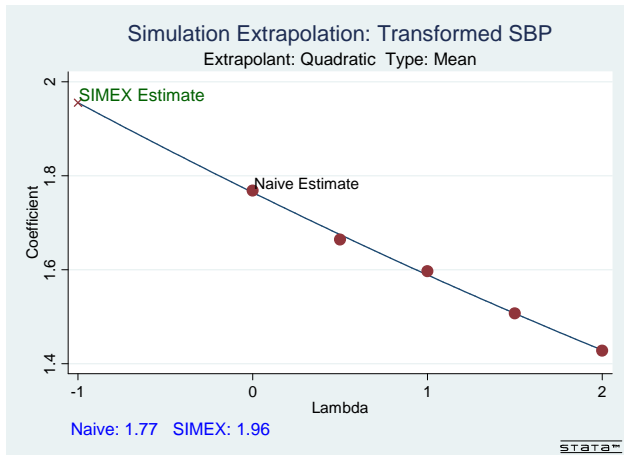
Residual df =      1611           Wald F(3,1611)  =      27.98
                                   Prob > F         =      0.0000
```

```
Variance Function: V(u) = u(1-u)          [Bernoulli]
```

```
Link Function      : g(u) = log(u/(1-u))   [Logit]
```

firstchd	Bootstrap					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0512939	.0093325	5.50	0.000	.0329889	.069599
smoke	.5780199	.2324534	2.49	0.013	.122077	1.033963
w3	1.942437	.3917411	4.96	0.000	1.174061	2.710813
_cons	-13.92484	1.67293	-8.32	0.000	-17.20619	-10.6435

# SIMEX Graph



# Multiplicative Error

- Assume lognormal  $X$  and lognormal measurement error
- Let  $\mu_{\ell w} = E\{\log(W)\}$  and let  $\lambda$  be the attenuation in the log scale.

- **Calibration function:**

$$E(X|W) = W^\lambda \exp\left\{(1 - \lambda)\mu_{\ell w} + \lambda\sigma_u^2/2\right\}$$

- Note that the calibration function is **Not linear in  $W$**

# Multiplicative Error

- Remember one of the **triple whammy**: measurement error can hide features
- **Linear regression**:  $E(Y|X) = \beta_0 + \beta_1 X$

- Then with multiplicative error, the observed data have

$$E(Y|W) = \beta_0 + \left[ \beta_1 \exp \left\{ (1 - \lambda) \mu_{\ell w} + \lambda \sigma_u^2 / 2 \right\} \right] W^\lambda$$

- Note how the observed data have a regression that is not linear in  $W$ .

# Multiplicative Error

- **Solution**: Regress  $Y$  on  $W^\lambda$  and divide the slope estimate by  $\exp\{(1 - \lambda)\mu_{\ell w} + \lambda\sigma_u^2/2\}$
- **Deja vu all over again**: This is regression calibration
- Note that the regression of  $Y$  on  $W$  is not linear even though the regression of  $Y$  on  $X$  is linear

# Estimating The Calibration Function: Replication Data

- Suppose that one has unbiased internal replicate data, as in Framingham
- **Model:**  $W_{ij} = X_i + U_{ij}$ ,  $i = 1, \dots, n$  and  $j = 1, \dots, k_i$ , where  $E(U_{ij}|Z_i, X_i) = 0$ .
- **Within-person mean:**  $\bar{W}_{i\cdot} := k_i^{-1} \sum_j W_{ij}$ .
- **Notation:**  $\mu_z$  is  $E(Z)$ ,  $\Sigma_{xz}$  is the covariance (matrix) between  $X$  and  $Z$ , etc.

# Estimating The Calibration Function: Replication Data

- Use **standard least squares theory** to get the **best linear unbiased predictor** (BLUP) of  $X$  from  $(W, Z)$ :

$$E(X|Z, \overline{W}) \approx \mu_x + (\Sigma_{xx} \Sigma_{xz}^t) \left\{ \begin{array}{cc} \Sigma_{xx} + \Sigma_{uu}/k & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{array} \right\}^{-1}$$

- BLUP = **exact conditional expectation under joint normality**.
- Needed**: estimate the unknown  $\mu$ 's and  $\Sigma$ 's.

# Estimating The Calibration Function: Replication Data, Continued

- **Formulae**: Tedious but standard ANOVA-type
- **Mean and Covariance of  $Z$** :  $\hat{\mu}_z$  and  $\hat{\Sigma}_{zz}$  are the “usual” estimates since the  $Z$ 's are observed.
- **Mean of  $X$** :  $\hat{\mu}_x = \hat{\mu}_w = \sum_{i=1}^n k_i \overline{W}_i / \sum_{i=1}^n k_i$ .
- **Covariance of  $X$  and  $Z$** :  

$$\hat{\Sigma}_{xz} = \sum_{i=1}^n k_i (\overline{W}_i - \hat{\mu}_w)(Z_i - \hat{\mu}_z)^t / \nu$$
 where  $\nu = \sum k_i - \sum k_i^2 / \sum k_i$  is obtained from equating moments.

# Estimating The Calibration Function: Replication Data, Continued

- The **Measurement error covariance matrix** and the **Covariance matrix of  $X$**  are estimated from **ANOVA**
- Remember,  $W_{ij} = X_i + U_{ij}$  is simply a **1-way ANOVA** with factors being the individuals. Then

$$\hat{\Sigma}_{uu} = \frac{\sum_{i=1}^n \sum_{j=1}^{k_i} (W_{ij} - \bar{W}_{i\cdot})(W_{ij} - \bar{W}_{i\cdot})^t}{\sum_{i=1}^n (k_i - 1)}.$$

- 

$$\hat{\Sigma}_{xx} = \left[ \left\{ \sum_{i=1}^n k_i (\bar{W}_{i\cdot} - \hat{\mu}_w)(\bar{W}_{i\cdot} - \hat{\mu}_w)^t \right\} - (n-1)\hat{\Sigma}_{uu} \right] / \nu.$$