INTRODUCTION AND OUTLINE

This short course is based upon the book

Measurement Error in Nonlinear Models

R. J. Carroll, D. Ruppert and L. A. Stefanski

Chapman & Hall/CRC Press, 1995

ISBN: 0 412 04721 7

http://www.crcpress.com

OUTLINE OF SEGMENT 1

- What is measurement error?
- Some examples
- Effects of measurement error in simple linear regression
- Effects of measurement error in multiple regression
- Analysis of Covariance: effects of measurement error in a covariate on the comparisons of populations
- The correction for attenuation: the classic way of correcting for biases caused by measurement error

OUTLINE OF SEGMENT 2

- Broad classes of measurement error
  - **Nondifferential**: you only measure an error–prone predictor because the error–free predictor is unavailable
  - **Differential**: the measurement error is itself predictive of outcome
- Surrogates
  - Proxies for a difficult to measure predictor
- Assumptions about the form of the measurement error: additive and homoscedastic
- Replication to estimate measurement error variance
- Methods to disagnose whether measurement error is additive and homoscedastic

OUTLINE OF SEGMENT 3

- Transportability: using other data sets to estimate properties of measurement error
- Conceptual definition of an exact predictor
- The **classical** error model
  - You observe the real predictor **plus** error
- The **Berkson** error model
  - The real predictor is what you observe **plus** error
- Functional and structural models defined and discussed

The project described was supported by Grant Number R44 RR12435 from the National Institutes of Health, National Center for Research Resources. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Research Resources.

OUTLINE OF SEGMENT 4

- The regression calibration method: replace X by an estimate of it given the observed data
- Regression calibration is correction for attenuation (**Segment 1**)in linear regression
- Use of validation, replication and external data
- Logistic and Poisson regression
- Use of an unbiased surrogate to estimate the calibration function

OUTLINE OF SEGMENT 5

- The SIMEX method
- Motivation from design of experiments
- The algorithm
  - The **sim**ulation step
  - The **ex**trapolation step
- Application to logistic regression
- Application to a generalized linear mixed model

OUTLINE OF SEGMENT 6

- Instrumental variables:
  - Indirect way to understand measurement error
  - Often the least informativew
- The IV method/algorithm
  - Why the results are variable
  - IV estimation as a type of regression calibration
- Examples to logistic regression

OUTLINE OF SEGMENT 7

- Likelihood methods
- The Berkson model and the Utah fallout study
  - The essential parts of a Berkson likelihood analysis
- The classical model and the Framingham study
  - The essential parts of a classical likelihood analysis
- Model robustness and computational issues

SEGMENT 1: INTRODUCTION AND LINEAR MEASUREMENT ERROR MODELS REVIEW OUTLINE

- **About This Course**
- **Measurement Error Model Examples**
- **Structure of a Measurement Error Problem**
- **A Classical Error Model**
- **Classical Error Model in Linear Regression**
- **Summary**

ABOUT THIS COURSE

- This course is about analysis strategies for **regression problems in which predictors are measured with error.**
- Remember your introductory regression text ...
  * Snedecor and Cochran (1967), " Thus far we have assumed that X-variable in regression is measured without error. Since no measuring instrument is perfect this assumption is often unrealistic."
  * Steele and Torrie (1980), "... if the X's are also measured with error, ... an alternative computing procedure should be used ..."
  * Neter and Wasserman (1974), "Unfortunately, a different situation holds if the independent variable X is known only with measurement error."
- This course focuses on **nonlinear** measurement error models (MEMs), with some essential review of **linear** MEMs (see Fuller, 1987)

EXAMPLES OF MEASUREMENT ERROR MODELS

- Measures of nutrient intake
  * A **classical** error model
- Coronary Heart Disease vs Systolic Blood Pressure
  * A **classical** error model
- Radiation Dosimetry
  * A **Berkson** error model

MEASURES OF NUTRIENT INTAKE

- Y = average daily percentage of calories from fat as measured by a food frequency questionnaire (FFQ).
- X = true long–term average daily percentage of calories from fat
- The problem: fit a **linear** regression of Y on X
- In symbols, $Y = \beta_0 + \beta_x X + \epsilon$
- **X is never observable**. It is **measured** with error:

MEASURES OF NUTRIENT INTAKE

- Along with the FFQ, on 6 days over the course of a year women are interviewed by phone and asked to recall their food intake over the past year (24–hour recalls).
- Their average % Calories from Fat is recorded and denoted by W.
  - * The analysis of 24–hour recall introduces some error $\Longrightarrow$ **analysis error**
  - * **Measurement error** = **sampling error** + **analysis error**
  - * Measurement error model
    $$W_i = X_i + U_i, \quad U_i \text{ are measurement errors}$$

HEART DISEASE VS SYSTOLIC BLOOD PRESSURE

- Y = indicator of Coronary Heart Disease (CHD)
- X = true long-term average systolic blood pressure (SBP) (maybe transformed)
- Goal: Fot a **logistic** regression of Y on X
- In symbols, $\mathrm{pr}(Y = 1) = H(\beta_0 + \beta_x X)$
- Data are CHD indicators and determinations of systolic blood pressure for $n = 1,600$ in Framingham Heart Study
- X **measured** with error:

HEART DISEASE VS SYSTOLIC BLOOD PRESSURE

- SBP measured at two exams (and averaged) $\Longrightarrow$ **sampling error**
- The determination of SBP is subject to machine and reader variability $\Longrightarrow$ **analysis error**
  - * **Measurement error** = **sampling error** + **analysis error**
  - * Measurement error model
    $$W_i = X_i + U_i, \quad U_i \text{ are measurement errors}$$

THE KEY FACTOID OF MEASUREMENT ERROR PROBLEMS

- Y = response, Z = error-free predictor, X = error-prone predictor, W = proxy for X
- **Observed** are (Y, Z, W)
- **Unobserved** is X
- Want to fit a regression model (linear, logistic, etc.)
- In symbols, $E(Y|Z, X) = f(Z, X, \beta)$
- **Key point**: The regression model in the observed data is not the same as the regression model when X is observed
- In symbols, $E(Y|Z, W) \neq f(Z, W, \beta)$

A CLASSICAL ERROR MODEL

- What you see is the true/real predictor **plus** measurement error
- In symbols, $W_i = X_i + U_i$
- This is called **additive**) measurement error
- The measurement errors U_i are:
  - ∗ independent of all Y_i, Z_i and X_i (independent)
  - ∗ IID$(0,\ \sigma_u^2)$ (IID, unbiased, homoscedastic)

SIMPLE LINEAR REGRESSION WITH A CLASSICAL ERROR MODEL

- Y = response, X = error-prone predictor
- $Y = \beta_0 + \beta_x X + \epsilon$
- Observed data: (Y_i, W_i), $i = 1, \ldots, n$
- $W_i = X_i + U_i$ (**additive**)
- U_i are:
  - ∗ independent of all Y_i, Z_i and X_i (independent)
  - ∗ IID$(0,\ \sigma_u^2)$ (IID, unbiased, homoscedastic)

What are the effects of measurement error on the usual analysis?

SIMULATION STUDY

- Generate $X_1, \ldots, X_{50}$, IID N$(0, 1)$
- Generate $Y_i = \beta_0 + \beta_x X_i + \epsilon_i$
  - ∗ ϵ_i IID N$(0, 1/9)$
  - ∗ $\beta_0 = 0$
  - ∗ $\beta_x = 1$
- Generate $U_1, \ldots, U_{50}$, IID N$(0, 1)$
- Set $W_i = X_i + U_i$
- **Regress Y on X and Y on W and compare**

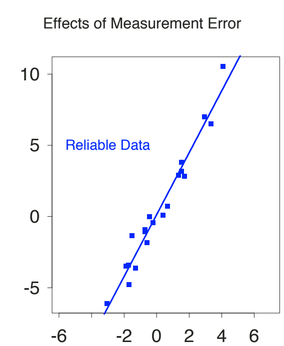

Figure 1: **True Data Without Measurement Error.**

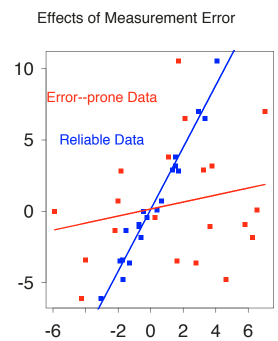

Figure 2: **Observed Data With Measurement Error.**

THEORY BEHIND THE PICTURES: THE NAIVE ANALYSIS

- Least Squares Estimate of Slope:

$$\widehat{\beta}_x = \frac{S_{y,w}}{S_w^2}$$

where

$$\begin{aligned} S_{y,w} \longrightarrow \mathrm{Cov}(Y,\ W) &= \mathrm{Cov}(Y,\ X+U) \\ &= \mathrm{Cov}(Y,\ X) \\ &= \sigma_{y,x} \end{aligned}$$

$$\begin{aligned} S_w^2 \longrightarrow \mathrm{Var}(W) &= \mathrm{Var}(X+U) \\ &= \sigma_x^2 + \sigma_u^2 \end{aligned}$$

THEORY BEHIND THE PICTURES: THE NAIVE ANALYSIS

So

$$\widehat{\beta}_x \longrightarrow \frac{\sigma_{y,x}}{\sigma_x^2+\sigma_u^2} = \left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\right)\beta_x$$

- Note how classical measurement error causes a **bias** in the least squares regression coefficient

THEORY BEHIND THE PICTURES: THE NAIVE ANALYSIS

- The **attenuation factor** or **reliability ratio** describes the bias in linear regression caused by classical measurement error

$$\text{You estimate } \lambda\beta_x; \quad \lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$$

- **Important Factoids**:
  * As the measurement error increases, **more bias**
  * As the variability in the true predictor increases, **less bias**

THEORY BEHIND THE PICTURES: THE NAIVE ANALYSIS

- Least Squares Estimate of Intercept:

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_x \overline{W} \\ &\longrightarrow \mu_y - \lambda \beta_x \mu_x \\ &= \beta_0 + (1-\lambda)\beta_x \mu_x\end{aligned}$$

- Estimate of Residual Variance:

$$\text{MSE} \longrightarrow \sigma_\epsilon^2 + (1-\lambda)\beta_x^2 \sigma_x^2$$

- Note how the residual variance is **inflated**
  * Classical measurement error in X causes the regression to have more noise

MORE THEORY: JOINT NORMALITY

- Y, X, W jointly normal $\Longrightarrow$
  * $Y \mid W \sim$ Normal
  * $E(Y \mid W) = \beta_0 + (1-\lambda)\beta_x \mu_x + \lambda \beta_x W$
  * $\text{Var}(Y \mid W) = \sigma_\epsilon^2 + (1-\lambda)\beta_x^2 \sigma_x^2$

- Intercept is **shifted** by $(1-\lambda)\beta_x \mu_x$
- Slope is **attenuated** by the factor λ
- Residual variance is **inflated** by $(1-\lambda)\beta_x^2 \sigma_x^2$
- **And simple linear regression is an easy problem!**

MORE THEORY: IMPLICATIONS FOR TESTING HYPOTHESES

- Because

$$\beta_x = 0 \quad \text{iff} \quad \lambda \beta_x = 0$$

it follows that

$$[H_0 : \beta_x = 0] \quad \equiv \quad [H_0 : \lambda \beta_x = 0]$$

which in turn implies that **the naive test of $\beta_x = 0$ is valid (correct Type I error rate).**

- The discussion of naive tests when there are **multiple predictor** measured with error, or **error-free** predictors, is **more complicated**
- In the following graph, we show that as the measurement error increases:
  * Statistical power decreases
  * Sample size to obtain a fixed power increases

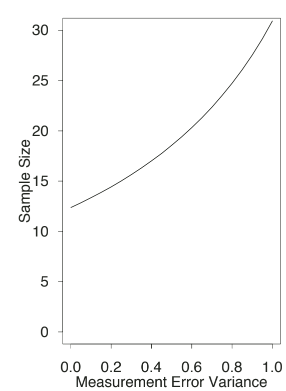

Figure 3: **Sample Size for 80% Power. True slope $\beta_x = 0.75$. Variances $\sigma_x^2 = \sigma_\epsilon^2 = 1$.**

MULTIPLE LINEAR REGRESSION WITH ERROR

- Model

$$\begin{aligned} Y &= \beta_0 + \beta_z^t Z + \beta_x^t X + \epsilon \\ W &= X + U \text{ is observed instead of } X \end{aligned}$$

- Regressing Y on Z and W estimates

$$\begin{pmatrix} \beta_{z*} \\ \beta_{x*} \end{pmatrix} = \Lambda \begin{pmatrix} \beta_z \\ \beta_x \end{pmatrix} \qquad \left[\neq \begin{pmatrix} \beta_z \\ \beta_x \end{pmatrix} \right]$$

- Λ is the **attenuation matrix** or reliability matrix

$$\Lambda = \begin{pmatrix} \sigma_{zz} & \sigma_{zx} \\ \sigma_{xz} & \sigma_{xx} + \sigma_{uu} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{zz} & \sigma_{zx} \\ \sigma_{xz} & \sigma_{xx} \end{pmatrix}$$

- Biases in components of β_{x*} and β_{z*} can be multiplicative or **additive** $\Longrightarrow$
  - ∗ Naive test of $H_0 : \beta_x = 0,\ \beta_z = 0$ is valid
  - ∗ Naive test of $H_0 : \beta_x = 0$ is valid
  - ∗ Naive test of $H_0 : \beta_{x,1} = 0$ is typically not valid ($\beta_{x,1}$ denotes a subvector of β_x)

MULTIPLE LINEAR REGRESSION WITH ERROR

- For X scalar, attenuation factor changes:

$$\lambda_1 = \frac{\sigma^2_{x|z}}{\sigma^2_{x|z} + \sigma^2_u}$$

  - ∗ $\sigma^2_{x|z}$ = residual variance in regression of X on Z
  - ∗ $\sigma^2_{x|z} \leq \sigma^2_x \Longrightarrow$

$$\lambda_1 = \frac{\sigma^2_{x|z}}{\sigma^2_{x|z} + \sigma^2_u} \leq \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_u} = \lambda$$

  - ∗ $\Longrightarrow$ **Collinearity accentuates attenuation**

MULTIPLE LINEAR REGRESSION WITH ERROR

- Amazingly, classical measurement error in X causes iased estimates of β_z:
- Suppose that the regressio of X on Z is $\gamma_0 + \gamma_z Z$
- Then what you estimate is

$$\beta_{z*} = \beta_z + (1 - \lambda_1)\beta_x \gamma_z,$$

- So, there is bias in the coefficient for Z if:
  - ∗ X is correlated with Z
  - ∗ Z is a significant predictor were X to be observed

ANALYSIS OF COVARIANCE

- These results have implications for the **two group ANCOVA.**

∗ **X = true covariate**

∗ **Z = dummy indicator of group**

- We are interested in estimating β_z, the group effect. Biased estimates of β_z:

$$\beta_{z*} = \beta_z + (1 - \lambda_1)\beta_x \gamma_z,$$

  - ∗ γ_z is from $\mathrm{E}(X \mid Z) = \gamma_0 + \gamma_z^t Z$
  - ∗ γ_z is the difference in the mean of X among the two groups.
  - ∗ Thus, **biased unless X and Z are unrelated**.
  - ∗ A randomized Study!!!

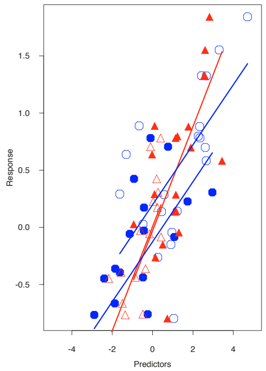

Figure 4: **UNBALANCED ANCOVA. RED = TRUE DATA, BLUE = OBSERVED. SOLID = FIRST GROUP, OPEN = SECOND GROUP. NO DIFFERENCE IN GROUPS.**

CORRENTIONS FOR ATTENUATION

$$\begin{aligned} Y &= \beta_0 + \beta_z^t Z + \beta_x^t X + \epsilon \\ W &= X + U \text{ is observed instead of } X \end{aligned}$$

- Let Σ_{uu} be the measurement error covariance matrix
- Let Σ_{zz} be the covariance matrix of the Z's
- Let Σ_{ww} be the covariance matrix of the W's
- Let Σ_{zw} be the covariance matrix of the Z's and W's
- Ordinary least squares actually estimates

$$\begin{pmatrix} \Sigma_{zz} & \Sigma_{zw} \\ \Sigma_{wz} & \Sigma_{ww} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{zz} & \Sigma_{zw} \\ \Sigma_{wz} & \Sigma_{ww} - \Sigma_{uu} \end{pmatrix} \begin{pmatrix} \beta_z \\ \beta_x \end{pmatrix}.$$

- The correction for attenuation simply fixes this up:

$$\begin{pmatrix} \widehat{\beta}_{z,eiv} \\ \widehat{\beta}_{x,eiv} \end{pmatrix} = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zw} \\ \Sigma_{wz} & \Sigma_{ww} - \Sigma_{uu} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{zz} & \Sigma_{zw} \\ \Sigma_{wz} & \Sigma_{ww} \end{pmatrix} \begin{pmatrix} \widehat{\beta}_{z,ols} \\ \widehat{\beta}_{x,ols} \end{pmatrix}.$$

- In simple linear regression, this means that the ordinary least squares slope is divided by the attenuation to get the correction for attenuation.

SEGMENT 2 NONLINEAR MODELS AND DATA TYPES OUTLINE

- **Differential** and **Nondifferential** measurement error.
- Estimating error variances:
  - ∗ **Validation**
  - ∗ **Replication**
- Using **Replication** data to check error models
  - ∗ **Additivity**
  - ∗ **Homoscedasticity**
  - ∗ **Normality**

THE BASIC DATA

- A **response** Y
- **Predictors** X measured **with error**.
- **Predictors** Z measured **without error**.
- A **major proxy** W for X.
- Sometimes, a **second proxy** T for X.

NONDIFFERENTIAL ERROR

- Error is said to be **nondifferential** if W and T would not be measured if one could have measured X.
  - ∗ It is not clear how this term arose, but it is in commopn use.
- More formally, (W, T) are **conditionally independent** of Y given (X, Z).
  - ∗ The idea: (W, T) provide **no additional information** about Y if X were observed
- This often makes sense, but it may be **fairly subtle** in each application.

NONDIFFERENTIAL ERROR

- Many crucial theoretical calculations revolve around nondifferential error.
- Consider simple linear regression: $Y = \beta_0 + \beta_x X + \epsilon$, where ϵ is independent of X.

$$\begin{aligned} E(Y|W) &= E\left[\{E(Y|X,W)\}|W\right] \\ &= E\left[\{E(Y|X)\}|W\right] \text{ Note} \\ &= \beta_0 + \beta_x E(X|W). \end{aligned}$$

  - ∗ This reduces the problem in general to estimating $E(X|W)$.
- If the error is **differential**, then the second line fails, and no simplification is possible.
- For example,

$$\text{cov}(Y, W) = \beta_x \text{cov}(Y, X) + \text{cov}(\epsilon, W).$$

HEART DISEASE VS SYSTOLIC BLOOD PRESSURE

- Y = indicator of Coronary Heart Disease (CHD)
- X = true long-term average systolic blood pressure (SBP) (maybe transformed)
- Assume $P(Y = 1) = H(\beta_0 + \beta_x X)$
- Data are CHD indicators and determinations of systolic blood pressure for $n = 1600$ in Framingham Heart Study
- X **measured** with error:
  - ∗ SBP measured at two exams (and averaged) $\Longrightarrow$ sampling error
  - ∗ The determination of SBP is subject to machine and reader variability
- ∗ It is hard to believe that the short term average of two days carries any additional information about the subject's chance of CHD over and above true SBP.
- ∗ Hence, **Nondifferential**

IS THIS NONDIFFERENTIAL?

- From Tosteson et al. (1989).
- $Y = I\{\text{wheeze}\}$.
- X is personal exposure to NO_2.
- W = (NO_2 in kitchen, NO_2 in bedroom) is observed in the primary study.

IS THIS NONDIFFERENTIAL?

- From Küchenhoff & Carroll
- $Y = I\{\text{lung irritation}\}$.
- X is **actual personal long–term dust exposure**
- $W =$ is dust exposure as **measured by occupational epidemiology techniques**.
  - ∗ They sampled the plant for dust.
  - ∗ Then they tried to match the person to work area

IS THIS NONDIFFERENTIAL?

- Y = average daily percentage of calories from fat as measured by a food frequency questionnaire (FFQ).
- FFQ's are in wide use because they are inexpensive
- The non–objectivity (self–report) suggests a generally complex error structure
- X = true long–term average daily percentage of calories from fat
- Assume $Y = \beta_0 + \beta_x X + \epsilon$
- X is never observable. It is **measured** with error:
  - ∗ Along with the FFQ, on 6 days over the course of a year women are interviewed by phone and asked to recall their food intake over the past year (24–hour recalls). Their average is recorded and denoted by W.

WHAT IS NECESSARY TO DO AN ANALYSIS?

- In linear regression with classical additive error $W = X + U$, we have seen that what we need is:
  - ∗ **Nondifferential error**
  - ∗ An **estimate of the error variance** $\text{var}(U)$
- How do we get the latter information?
- The best way is to get a subsample of the study in which X is observed. This is called **validation**.
  - ∗ In our applications, generally not possible.
- Another method is to do **replications** of the process, often called **calibration**.
- A third way is to get the value from another similar study.

REPLICATION

- In a **replication** study, for some of the study participants you measure **more than one** W.
- The standard **additive** model with m_i replicates is

$$W_{ij} = X_i + U_{ij}, \quad j = 1, ..., m_i.$$

- This is an unbalanced 1–factor ANOVA with mean squared error $\text{var}(U)$ estimated by

$$\widehat{\sigma}_u^2 = \frac{\Sigma_{i=1}^n \Sigma_{j=1}^{m_i} (W_{ij} - \overline{W}_{i\bullet})^2}{\Sigma_{i=1}^n (m_i - 1)}.$$

- Of course, as the proxy or surrogate for X_i one would use the sample mean $\overline{W}_{i\bullet}$.

$$\begin{aligned} \overline{W}_{i\bullet} &= X_i + \overline{U}_{i\bullet} \\ \text{var}(\overline{U}_{i\bullet}) &= \sigma_u^2 / m_i. \end{aligned}$$

REPLICATION

- **Replication** allows you to test whether your model is basically **additive** with **constant error variance**.
- If $W_{ij} = X_i + U_{ij}$ with U_{ij} symmetrically distributed about zero and independent of X_i, we have a major fact:
  - ∗ **The sample mean and sample standard deviation are uncorrelated.**
- Also, if U_{ij} are normally distributed, then so too are differences $W_{i1} - W_{i2} = U_{i1} - U_{i2}$.
  - ∗ q-q plots of these differences can be used to assess normality of the measurement errors
- Both procedures can be implemented easily in any package.

REPLICATION: WISH

- The WISH study measured caloric intake using a 24–hour recall.
  - ∗ There were 6 replicates per woman in the study.
- A plot of the caloric intake data showed that W was no where close to being normally distributed in the population.
  - ∗ If additive, then either X or U is not normal.
- When plotting standard deviation versus the mean, typical to use the rule that the method "passes" the test if the essential max–to–min is less than 2.0.
  - ∗ **A little bit of non–constant variance never hurt anyone.** See Carroll & Ruppert (1988)

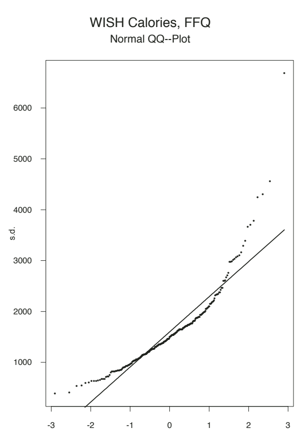

Figure 5: **WISH, CALORIC INTAKE, Q–Q plot of Observed data.** Caloric intake is **clearly** not normally distributed.

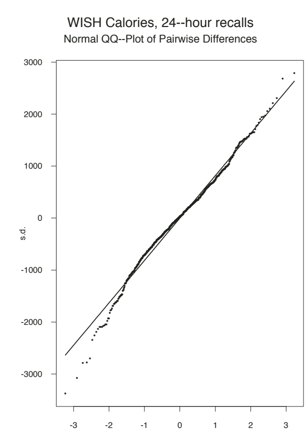

Figure 6: **WISH, CALORIC INTAKE, Q–Q plot of Differenced data.** This suggests that the measurement errors are reasonably normally distributed.

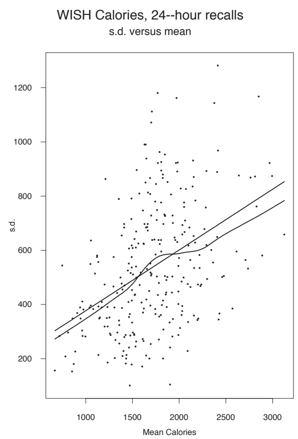

Figure 7: **WISH, CALORIC INTAKE, plot for additivity, loess and OLS.** The standard deviation versus the mean plot suggests lots of non–constant variance. Note how the range of the fits violates the 2:1 rule.

REPLICATION: WISH

- Taking logarithms improves all the plots.

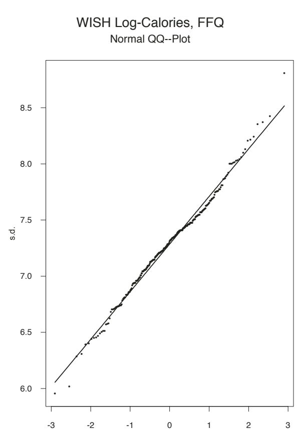

Figure 8: **WISH, LOG CALORIC INTAKE, Q–Q plot of Observed data.** The actual logged data appears nearly normally distributed.

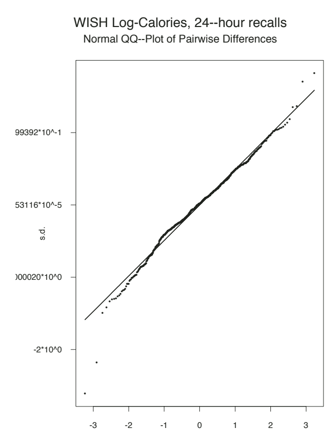

Figure 9: **WISH, LOG CALORIC INTAKE, Q–Q plot of Differenced data.** The measurement errors appear normally distributed.

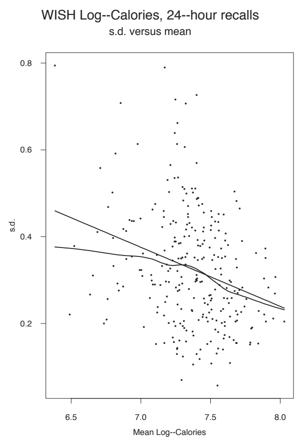

Figure 10: **WISH, LOG CALORIC INTAKE, plot for additivity, loess and OLS.** The 2:1 rule is not badly violated, suggested constant variance of the errors. This transformation seems to work fine.

SUMMARY

- **Nondifferential** error is an important assumption.
  - ∗ In the absence of **validation** data, it is **not a testable assumption**.
- **Additivity, Normality, Homoscedasticity** of errors can be assessed graphically via **replication**
  - ∗ Sample standard deviation versus sample mean.
  - ∗ q–q plots of differences of within–person replicates.

SEGMENT 3: BASIC CONCEPTUAL ISSUES

- **Transportability**: what parts of a measurement error model can be assessed by external data sets
- **What is Berkson? What is classical?**
- **Functional versus structural modeling**

TRANSPORTABILITY AND THE LIKELIHOOD

- In linear regression, we have seen that **we only require knowing the measurement error variance** (after checking for semi–constant variance, additivity, normality).
- Remember that the reliability ratio or attenuation coefficient is
$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \frac{\mathrm{var}(X)}{\mathrm{var}(W)}$$
- **In general though, more is needed**. Let's remember that if we observe W instead of X, then the observed data have a regression of Y on W that effectively acts as if
$$\begin{aligned} E(Y|W) &= \beta_0 + \beta_x E(X|W) \\ &\approx \beta_0 + \beta_x \lambda W. \end{aligned}$$
- If we knew λ, it would be easy to correct for the bias

TRANSPORTABILITY

- It is tempting to try to use outside data and transport this distribution to your problem.
  - * **Bad idea**!!!!!!!!!!!!

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

  - * Note how λ depends on the **distribution of X.**
  - * It is **rarely** the case that two populations have the **same X distribution**, even when the same instrument is used.

EXTERNAL DATA AND TRANSPORTABILITY

- A model is **transportable** across studies if it holds **with the same parameters** in the two studies.
  - * **Internal data**, i.e., data from the current study, is ideal since there is no question about transportability.
- **With external data, transportability back to the primary study cannot be taken for granted.**
  - * Sometimes transportability clearly will not hold. Then the value of the external data is, at best, questionable.
  - * Even is transportability seems to be a reasonable assumption, it is still just that, an assumption.

EXTERNAL DATA AND TRANSPORTABILITY

- As an illustration, consider two nutrition data sets which use exactly the same FFQ
- **Nurses Health Study**
  - * Nurses in the Boston Area
- **American Cancer Society**
  - * National sample
- Since the **same instrument is used**, error properties should be about the same.
  - * But maybe **not the distribution** of X!!!
  - * var(differences, NHS = 47)
  - * var(differences, ACS = 45)

  - * var(sum, ACS = 296)

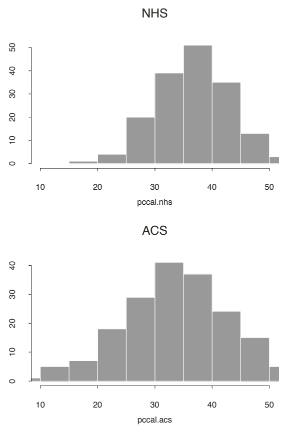

Figure 11: **FFQ Histograms of % Calories from Fat in NHS and ACS**

THE BERKSON MODEL

- The **Berkson model** says that
  True Exposure = Observed Exposure + Error

$$X = W + U_b$$

- Note the difference:
  - * **Classical:** We observe true X plus error
  - * **Berkson:** True X is what we observe (W) plus error
  - * Further slides will describe the difference in detail
- In the linear regression model,
  - * Ignoring error still leads to **unbiased intercept and slope estimates**,
  - * but the **error about the line is increased**.

WHAT'S BERKSON? WHAT'S CLASSICAL?

- **In practice, it may be hard to distinguish between the classical and the Berkson error models.**
  - * In some instances, neither holds exactly.
  - * In some complex situations, errors may have both Berkson and classical components, e.g., when the observed predictor is a combination of 2 or more error–prone predictors.
- **Berkson model:** a **nominal value is assigned.**
  - * Direct measures cannot be taken, nor can replicates.
- **Classical error structure:** direct **individual** measurements are taken, and can be replicated but with variability.

WHAT'S BERKSON? WHAT'S CLASSICAL?

- Direct measures possible?
- Replication possible?
- **Classical:** We observe true X plus error
- **Berkson:** True X is what we observe (W) plus error
- **Let's play stump the experts!**
- Framingham Heart Study
  - * Predictor is systolic blood pressure

WHAT'S BERKSON? WHAT'S CLASSICAL?

- All workers with the same job classification and age are assigned the same exposure based on job exposure studies.

- Using a phantom, all persons of a given height and weight with a given recorded dose are assigned the same radiation exposure.

WHAT'S BERKSON? WHAT'S CLASSICAL?

- Long–term nutrient intake as measured by repeated 24–hour recalls.

FUNCTIONAL AND STRUCTURAL MODELING

- Once you have decided on an error model, you have to go about making estimation and inference.
- In classical error models, you have to know the structure of the error.
  - ∗ Additive or multiplicative?
  - ∗ Some experimentation is necessary to give information about the measurement error variance.
- With all this information, you have to decide upon a method of estimation.
- The methods can be broadly categorized as **functional** or **structural**.

FUNCTIONAL AND STRUCTURAL MODELING

- The common linear regression texts make distinction:
  - ∗ **Functional:** X's are **fixed** constants
  - ∗ **Structural:** X's are **random** variables
- If you pretend that the X's are fixed constants, it seems plausible to try to estimate them as well as all the other model parameters.
- This is the functional maximum likelihood estimator.
  - ∗ Every textbook has the linear regression functional maximum likelihood estimator.
- Unfortunately, the functional MLE in nonlinear problems has two defects.
  - ∗ It's really nasty to compute.
  - ∗ It's a **lousy estimator** (badly inconsistent).

FUNCTIONAL AND STRUCTURAL MODELING CLASSICAL ERROR MODELS

- The common linear regression texts make distinction:
  - **Functional:** X's are **fixed** constants
  - **Structural:** X's are **random** variables
- These terms are misnomers.
- All inferential methods assume that the X's behave like a random sample anyway!
- More useful distinction:
  - **Functional:** No assumptions made about the X's (could be random or fixed)
  - **Classical structural:** Strong parametric assumptions made about the distribution of X. Generally normal, lognormal or gamma.

FUNCTIONAL METHODS IN THIS COURSE CLASSICAL ERROR MODELS

- 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, this means to ignore measurement error.
- The SIMEX method (Segment 4) is a fairly generally applicable functional method.
  - It assumes only that you have an error model, and that in some fashion you can "add on" measurement error to make the problem worse.

FUNCTIONAL METHODS CLASSICAL ERROR MODELS

- The strength of the **functional** model is its model **robustness**
  - No assumptions are made about the true predictors.
  - Standard error estimates are available.
- There are **potential** costs.
  - Loss of efficiency of estimation (missing data problems, highly nonlinear parameters)
  - Inference comparable to likelihood ratio tests are possible (SIMEX) but not well–studied.

SEGMENT 4: REGRESSION CALIBRATION OUTLINE

- Basic ideas
- The **regression calibration algorithm**
- **Correction for attenuation**
- Example: NHANES-I
- Estimating the calibration function
  - **validation data**
  - **instrumental data**
  - **replication data**

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).
- Developed as a general approach by Carroll and Stefanski (1990) and Gleser (1990).
  * Special cases appeared earlier in the literature.
- **Generally applicable** (like SIMEX).
  * **Depends on the measurement error being "not too large"** in order for the approximation to be suffciently accurate.

THE REGRESSION CALIBRATION ALGORITHM

- The general algorithm is:
  * Using replication, validation, or instrumental data, develop a model for the regression of X on (W,Z).
  * **Replace** X by the model fits 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."**

AN EXAMPLE: LOGISTIC REGRESSION, NORMAL X

- Consider the logistic regression model

$$\mathrm{pr}(Y=1|X) = \{1+\exp(-\beta_0-\beta_x X)\}^{-1} = H(\beta_0+\beta_x X).$$

- Remarkably, the regression calibration approximation works extremely well in this case

AN EXAMPLE: POISSON REGRESSION, NORMAL X

- Consider the Poisson loglinear regression model with

$$E(Y|X) = \exp(-\beta_0-\beta_x X).$$

- Suppose that X and U are normally distributed.
- Then the regression calibration approximation is approximately correct for the mean
- However, **the observed data are not Poisson**, but are overdispersed
- In other words, **and crucially**, measurement error can destroy the distributional relationship.

NHANES-I

- The NHANES-I example is from Jones et al., (1987).
- $Y = I\{\text{breast cancer}\}$.
- $Z =$ (age, poverty index ratio, body mass index, $I\{\text{use alcohol}\}$, $I\{\text{family history of breast cancer}\}$, $I\{\text{age at menarche} \leq 12\}$, $I\{\text{pre-menopause}\}$, race).
- $X =$ daily intake of saturated fat (grams).
- Untransformed surrogate:
  - ∗ saturated fat measured by 24-hour recall.
  - ∗ considerable error $\Rightarrow$ much controversy about validity.
- **Transformation**: W = log(5 + measured saturated fat).

NHANES-I—CONTINUED

- w/o adjustment for Z, W appears to have a small **protective** effect
- Naive logistic regression of Y on (Z, W):
  - ∗ $\widehat{\beta}_W = -.97$, $\text{se}(\widehat{\beta}_W) = .29$, $p < .001$
  - ∗ again evidence for a protective effect.
- Result is sensitive to the three individuals with the largest values of W.
  - ∗ all were non-cases.
  - ∗ changing them to cases: $p = .06$ and $\widehat{\beta}_W = -.53$, even though only 0.1% of the data are changed.

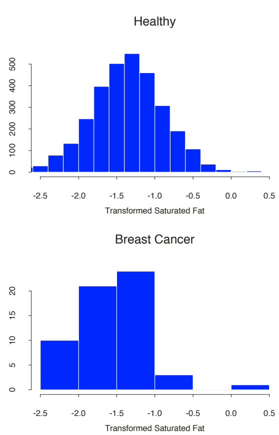

Figure 12: **Histograms of $\log(.05 + \text{Saturated Fat}/100)$ in the NHANES data, for women with and without breast cancer in 10 year follow–up..**

NHANES-I—CONTINUED

- External **replication data**;
  - ∗ CSFII (Continuous Survey of Food Intake by Individuals).
  - ∗ 24-hour recall (W) plus three additional 24-hour recall phone interviews, (T_1, T_2, T_3).
  - ∗ Over 75% of $\sigma^2_{W|Z}$ appears due to measurement error.
- From CSFII:
  - ∗ $\widehat{\sigma}^2_{W|Z} = 0.217$.
  - ∗ $\widehat{\sigma}^2_U = 0.171$ (assuming $W = X + U$)
  - ∗ Correction for attenuation:

$$\begin{aligned}\widehat{\beta}_x &= \frac{\widehat{\sigma}^2_{W|Z}}{\widehat{\sigma}^2_{W|Z} - \widehat{\sigma}^2_u}\widehat{\beta}_w \\ &= \frac{0.217}{0.217 - 0.171}(-.97) = -4.67\end{aligned}$$

  - ∗ 95% bootstrap confidence interval: $(-10.37, -1.38)$.
  - ∗ **Protective effect is now much bigger** but **estimated with much**

ESTIMATING THE CALIBRATION FUNCTION

- **Need to estimate $E(X|Z,W)$.**
  - ∗ How this is done depends, of course, on the type of auxiliary data available.
- **Easy case: validation data**
  - ∗ Suppose one has internal, validation data.
  - ∗ Then one can simply regress X on (Z,W) and transports the model to the non-validation data.
  - ∗ For the validation data one regresses Y on (Z,X), and this estimate must be combined with the one from the non-validation data.
- Same approach can be used for external validation data, but with the usual concern for non-transportability.

ESTIMATING THE CALIBRATION FUNCTION: INSTRUMENTAL DATA: ROSNER'S METHOD

- Internal unbiased instrumental data:
  - ∗ suppose $E(T|X) = E(T|X,W) = X$ so that T is an unbiased instrument.
  - ∗ 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.
  - ∗ then
$$E(T|W) = E\{E(T|X,W)|Z,W\} = E(X|W).$$
- Thus, T regressed on W follows the same model as X regressed on W, although with greater variance.
- One regresses T on (Z,W) to estimate the parameters in the regression of X on (Z,W).

ESTIMATING THE CALIBRATION FUNCTION: REPLICATION DATA

- Suppose that one has unbiased internal replicate data:
  - ∗ n individuals
  - ∗ k_i replicates for the ith individual
  - ∗ $W_{ij} = X_i + U_{ij}$, $i = 1,\ldots,n$ and $j = 1,\ldots,k_i$, where $E(U_{ij}|Z_i,X_i) = 0$.
  - ∗ $\overline{W}_{i\cdot} := \frac{1}{k_i}\Sigma_j W_{ij}$.
  - ∗ Notation: μ_z is $E(Z)$, Σ_{xz} is the covariance (matrix) between X and Z, etc.
- There are formulae to implement a regression calibration method in this case. **Basically, you use standard least squares theory to get the best linear unbiased predictor of X from (W,Z).**
  - ∗ Formulae are ugly, see attached and in the book

ESTIMATING THE CALIBRATION FUNCTION: REPLICATION DATA, CONTINUED

- $E(X|Z,\overline{W})$
$$\approx \mu_x + (\Sigma_{xx}\ \Sigma_{xz}) \begin{Bmatrix} \Sigma_{xx} + \Sigma_{uu}/k & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{Bmatrix}^{-1} \begin{pmatrix} \overline{W} - \mu_w \\ Z - \mu_z \end{pmatrix}. \qquad (1)$$
(best linear approximation = exact conditional expectation under joint normality).
- Need to estimate the unknown μ's and Σ's.
  - ∗ These estimates can then be substituted into (1).
  - ∗ $\widehat{\mu}_z$ and $\widehat{\Sigma}_{zz}$ are the "usual" estimates since the Z's are observed.
  - ∗ $\widehat{\mu}_x = \widehat{\mu}_w = \Sigma_{i=1}^n k_i \overline{W}_i / \Sigma_{i=1}^n k_i$.
  - ∗ $\widehat{\Sigma}_{xz} = \Sigma_{i=1}^n k_i(\overline{W}_{i\cdot} - \widehat{\mu}_w)(Z_i - \widehat{\mu}_z)^t/\nu$
    where $\nu = \Sigma k_i - \Sigma k_i^2 / \Sigma k_i$.

ESTIMATING THE CALIBRATION FUNCTION: REPLICATION DATA, CONTINUED

* $\widehat{\Sigma}_{uu}$

$$= \frac{\Sigma_{i=1}^{n}\Sigma_{j=1}^{k_i}(W_{ij}-\overline{W}_{i\cdot})(W_{ij}-\overline{W}_{i\cdot})^t}{\Sigma_{i=1}^{n}(k_i-1)}.$$

* $\widehat{\Sigma}_{xx}$

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

SEGMENT 5, REMEASUREMENT METHODS: SIMULATION EXTRAPOLATION, OUTLINE

ABOUT SIMULATION EXTRAPOLATION

- Restricted to **classical measurement error**
  * additive, unbiased, independent **in some scale, e.g., log**
  * for this segment:
  * one variable measured with error
  * error variance, σ_u^2, assumed known
- A **functional method**
  * no assumptions about the true X values
- **Not model dependent**
  * like bootstrap and jackknife
- **Handles complicated problems**
- **Computer intensive**
- Approximate, less efficient for certain problems

THE KEY IDEA

- The **effects of measurement error on a statistic can be studied with a simulation experiment in which additional measurement error is added** to the measured data and the statistic recalculated.
- Response variable is the statistic under study
- Independent factor is the measurement error variance
  * Factor levels are the variances of the added measurement errors
- Objective is to study how the statistic depends on the variance of the measurement error

OUTLINE OF THE ALGORITHM

- **Add measurement error !!!** to variable measured with error
  - θ controls amount of added measurement error
  - σ_u^2 increased to $(1+\theta)\sigma_u^2$
- **Recalculate estimates** — called pseudo estimates
- **Plot** pseudo estimates versus θ
- **Extrapolate** to $\theta = -1$
  - $\theta = -1$ corresponds to case of no measurement error

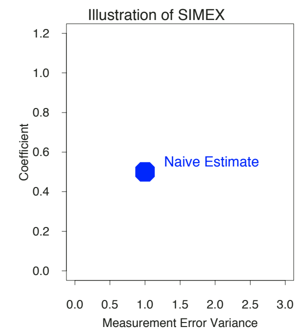

Figure 13: **Your estimate when you ignore measurement error.**

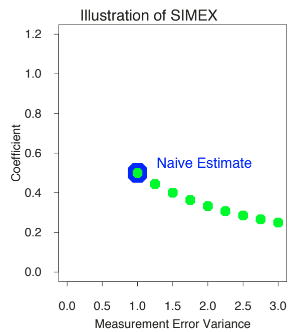

Figure 14: **This shows what happens to your estimate when you have more error, but you still ignore the error.**

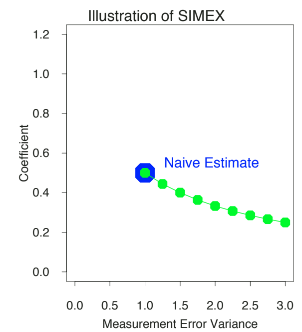

Figure 15: **What statistician can resist fitting a curve?**

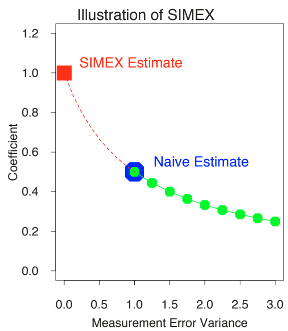

Figure 16: **Now extrapolate to the case of no measurement error.**

OUTLINE OF THE ALGORITHM

- **Add** measurement error to variable measured with error
  - θ controls amount of added measurement error
  - σ_u^2 increased to $(1+\theta)\sigma_u^2$
- **Recalculate** estimates — called pseudo estimates. **Do many times and average for each θ**
- **Plot** pseudo estimates versus θ
- **Extrapolate** to $\theta = -1$
  - $\theta = -1$ corresponds to case of no measurement error

AN EMPIRICAL VERSION OF SIMEX: FRAMINGHAM DATA EXAMPLE

- Data
  - Y = indicator of CHD
  - W_k = SBP at Exam k, $k = 1,\ 2$
  - X = "true" SBP
  - Data, 1660 subjects:

$$(Y_j,\ W_{1,j},\ W_{2,j}),\quad j = 1, \ldots,\ 1660$$

- Model Assumptions
  - $W_1,\ W_2 \mid X$ iid $\mathrm{N}(X,\ \sigma_u^2)$
  - $\Pr(Y = 1 \mid X) = H(\alpha + \beta X)$, H logistic

FRAMINGHAM DATA EXAMPLE: THREE NAIVE ANALYSES:

- Regress Y on $\overline{W}_{\bullet} \longmapsto \widehat{\beta}_{\text{Average}}$
- Regress Y on $W_1 \longmapsto \widehat{\beta}_1$
- Regress Y on $W_2 \longmapsto \widehat{\beta}_2$

θ	Predictor Measurement Error Variance $= (1+\theta)\sigma_u^2/2$	Slope Estimate
-1	0	?
0	$\sigma_u^2/2$	$\widehat{\beta}_{\mathrm{A}}$
1	σ_u^2	$\widehat{\beta}_1,\ \widehat{\beta}_2$

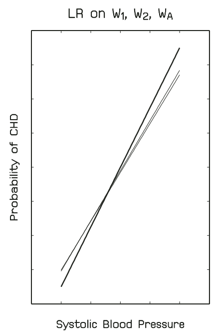

Figure 17: **Logistic regression fits in Framingham using first replicate, second replicate and average of both**

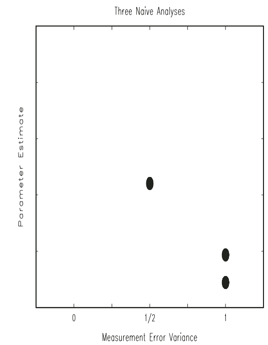

Figure 18: **A SIMEX–type plot for the Framingham data, where the errors are not computer–generated.**

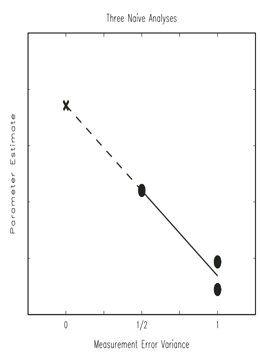

Figure 19: **A SIMEX–type extrapolation for the Framingham data, where the errors are not computer–generated.**

SIMULATION AND EXTRAPOLATION STEPS: EXTRAPOLATION

- Framingham Example: (two points $\theta = 0,\ 1$)
  - ∗ Linear Extrapolation — $a + b\theta$

- In General: (multiple θ points)
  - ∗ Linear: $a + b\theta$
  - ∗ Quadratic: $a + b\theta + c\theta^2$
  - ∗ Rational Linear: $(a + b\theta)/(c + \theta)$

SIMULATION AND EXTRAPOLATION ALGORITHM

- Simulation Step
- For $\theta \in \{\theta_1, \ldots, \theta_M\}$
- For $b = 1, ..., B$, compute:
  * bth pseudo data set
  $$W_{b,i}(\theta) = W_i + \sqrt{\theta}\ \mathrm{Normal}\left(0,\ \sigma_u^2\right)_{b,i}$$
  * bth pseudo estimate
  $$\widehat{\theta}_b(\theta) = \widehat{\theta}\left(\{Y_i, W_{b,i}(\theta)\}_1^n\right)$$
  * the average of the pseudo estimates
  $$\widehat{\theta}(\theta) = B^{-1}\sum_{b=1}^{B}\widehat{\theta}_b(\theta) \approx E\left(\widehat{\theta}_b(\theta) \mid \{Y_j, X_j\}_1^n\right)$$

SIMULATION AND EXTRAPOLATION ALGORITHM

- Extrapolation Step
- Plot $\widehat{\theta}(\theta)$ vs θ $(\theta > 0)$
- Extrapolate to $\theta = -1$ to get $\widehat{\theta}(-1) = \widehat{\theta}_{\mathrm{SIMEX}}$

EXAMPLE: MEASUREMENT ERROR IN SYSTOLIC BLOOD PRESSURE

- Framingham Data:
$$\left(Y_j,\ \mathrm{Age}_j,\ \mathrm{Smoke}_j,\ \mathrm{Chol}_j\ W_{\mathrm{A},j}\right),\quad j = 1, \ldots, 1615$$
  * Y = indicator of CHD
  * Age (at Exam 2)
  * Smoking Status (at Exam 1)
  * Serum Cholesterol (at Exam 3)
  * Transformed SBP
  $$W_{\mathrm{A}} = (W_1 + W_2)/2,$$
  $$W_k = \ln(\mathrm{SBP} - 50)\ \text{ at Exam } k$$
- Consider logistic regression of Y on Age, Smoke, Chol and SBP with transformed SBP measured with error

EXAMPLE: PARAMETER ESTIMATION

- The plots on the following page illustrate the simulation extrapolation method for estimating the parameters in the logistic regression model

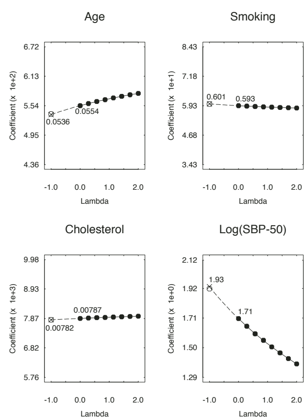

Figure 20:

EXAMPLE: VARIANCE ESTIMATION

- The pseudo estimates can be used for variance estimation.
  - The theory is similar to those for jackknife and bootstrap variance estimation.
  - The calculations, too involved to review here, are similar as well. See Chapter 4 of our book.
- In many cases, **with decent coding**, you can use the **bootstrap** to estimate the variance of SIMEX.

A MIXED MODEL

- Data from the Framingham Heart Study
- There were $m = 75$ **clusters** (individuals) with most having $n = 4$ exams, each taken 2 years apart.
- The variables were
  - Y = evidence of LVH (left ventricular hypertrophy) diagnosed by ECG in patients who developed coronary heart disease before or during the study period
  - W = log(SBP-50)
  - Z = age, exam number, smoking status, body mass index.
  - X = average log(SBP-50) over many applications within 6 months (say) of each exam.

A MIXED MODEL

- We fit this as a **logistic mixed model**, with a **random intercept** for each person having mean β_0 and variance θ.
- We assumed that measurement error was independent at each visit.

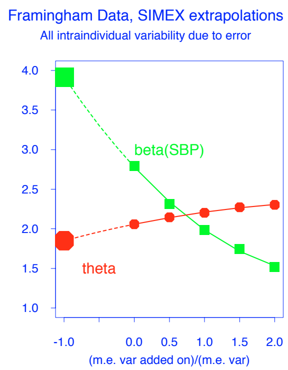

Figure 21: **LVH Framingham data. β(SBP) is the coefficient for transformed systolic blood pressure, while θ is the variance of the person–to–person random intercept.**

SUMMARY

- Bootstrap-like method for estimating bias and variance due to measurement error
- Functional method for classical measurement error
- Not model dependent
- Computer intensive
  - ∗ Generate and analyze several pseudo data sets
- Approximate method like regression calibration

SEGMENT 6 INSTRUMENTAL VARIABLES OUTLINE

- **Linear Regression**
- **Regression Calibration for GLIM's**

LINEAR REGRESSION

- Let's remember what the linear model says.

$$\begin{aligned} Y &= \beta_0 + \beta_x X + \epsilon; \\ W &= X + U; \\ U &\sim \text{Normal}(0, \sigma_u^2). \end{aligned}$$

- We know that if we ignore measurement error, ordinary least squares estimates not β_x, but instead it estimates

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

- λ is the attenuation coefficient or reliability ratio
- **Without information about σ_u^2, we cannot estimate β_x.**

INFORMATION ABOUT MEASUREMENT ERROR

- textbfblueClassical measurement error: $W = X + U$, $U \sim \text{Normal}(0, \sigma_u^2)$.
- **The most direct and efficient** way to get information about σ_u^2 is to observe X on a subset of the data.
- **The next best way is via replication**, namely to take ≥ 2 independent replicates
  - ∗ $W_1 = X + U_1$
  - ∗ $W_2 = X + U_2$.
- If these are indeed replicates, then we can estimate σ_u^2 via a components of variance analysis.
- **The third and least efficient method** is to use **Instrumental Variables**, or IV's
  - ∗ Sometimes replicates cannot be taken.
  - ∗ Sometimes X cannot be observed.
  - ∗ Then IV's can help

WHAT IS AN INSTRUMENTAL VARIABLE?

$$\begin{aligned} Y &= \beta_0 + \beta_x X + \epsilon; \\ W &= X + U; \\ U &\sim \text{Normal}(0, \sigma_u^2). \end{aligned}$$

- In linear regression, an instrumental variable T is a random variable which has three properties:
  - ∗ T is independent of ϵ
  - ∗ T is independent of U
  - ∗ T is related to X.
  - ∗ You only measure T to get information about measurement error: it is not part of the model.
  - ∗ In our parlance, T is a surrogate for X!

WHAT IS AN INSTRUMENTAL VARIABLE?

- **Whether T qualifies as an instrumental variable can be a difficult and subtle question.**
  - ∗ After all, we do not observe U, X or ϵ, so how can we **know** that the assumptions are satisfied?

AN EXAMPLE

$$\begin{aligned} X &= \text{usual (long–term) average intake of Fat (log scale);} \\ Y &= \text{Fat as measured by a questionnaire;} \\ W &= \text{Fat as measured by 6 days of 24–hour recalls} \\ T &= \text{Fat as measured by a diary record} \end{aligned}$$

- In this example, the time ordering was:
  - ∗ Questionnaire
  - ∗ Then one year later, the recalls were done fairly close together in time.
  - ∗ Then 6 months later, the diaries were measured.
- One could think of the recalls as replicates, but some researchers have worried that major correlations exist, i.e., they are not **independent** replicates.
- The 6–month gap with the recalls and the 18–month gap with the questionnaire makes the diary records a good candidate for an instrument.

INSTRUMENTAL VARIABLES ALGORITHM

- The simple IV algorithm in linear regression works as follows:

  STEP 1: Regress W on T (may be a multivariate regression)

  STEP 2: Form the predicted values of this regression

  STEP 3: Regress Y on the predicted values.

  STEP 4: The regression coefficients are the IV estimates.

- Only Step 3 changes if you do not have linear regression but instead have logistic regression or a generalized linear model.
  - $*$ Then the "regression" is logistic or GLIM.
  - $*$ Very simple to compute.
  - $*$ Easily bootstrapped.
- This method is "valid" in GLIM's to the extent that regression calibration is valid.

USING INSTRUMENTAL VARIABLES:MOTIVATION

- In what follows, we will use **underscores** to denote which coefficients go where.
- For example, $\beta_{Y|1\underline{X}}$ is the coefficient for X in the regression of Y on X.
- Let's do a little algebra:

$$\begin{aligned} Y &= \beta_{Y|\underline{1}X} + \beta_{Y|1\underline{X}}X + \epsilon; \\ W &= X + U; \\ (\epsilon, U) &= \text{independent of } T. \end{aligned}$$

- This means

$$\begin{aligned} E(Y \mid T) &= \beta_{Y|\underline{1}T} + \beta_{Y|1\underline{T}}T \\ &= \beta_{Y|\underline{1}X} + \beta_{Y|1\underline{X}}E(X \mid T) \\ &= \beta_{Y|\underline{1}X} + \beta_{Y|1\underline{X}}E(W \mid T) \end{aligned}$$

MOTIVATION

$$\begin{aligned} E(Y \mid T) &= \beta_{Y|\underline{1}T} + \beta_{Y|1\underline{T}}T \\ &= \beta_{Y|\underline{1}X} + \beta_{Y|1\underline{X}}E(X \mid T) \\ &= \beta_{Y|\underline{1}X} + \beta_{Y|1\underline{X}}E(W \mid T) \\ &= \beta_{Y|\underline{1}T} + \beta_{Y|1\underline{X}}\beta_{W|1\underline{T}}T. \end{aligned}$$

- We want to estimate $\beta_{Y|1\underline{X}}$
- Algebraically, this means that the slope Y on T is the product of the slope for Y on X times the slope for W on T:

$$\beta_{Y|1\underline{T}} = \beta_{Y|1\underline{X}}\beta_{W|1\underline{T}}$$

MOTIVATION

- $*$ Equivalently, it means

$$\beta_{Y|1\underline{X}} = \frac{\beta_{Y|1\underline{T}}}{\beta_{W|1\underline{T}}}.$$

- $*$ **Regress Y on T and divide its slope by the slope of the regression of W on T!**

THE DANGERS OF A WEAK INSTRUMENT

- Remember that we get the IV estimate using the relationship

$$\beta_{Y|1\underline{X}} = \frac{\beta_{Y|1\underline{T}}}{\beta_{W|1\underline{T}}}.$$

- This means we divide

$$\frac{\text{Slope of Regression of Y on T}}{\text{Slope of Regression of W on T}}.$$

- **The division causes increased variability.**
  - ∗ If the instrument is very weak, the slope $\beta_{W|1\underline{T}}$ will be near zero.
  - ∗ This will make the IV estimate very unstable.
- **It is generally far more efficient in practice to take replicates** and get a good estimate of the measurement error variance than it is to "hope and pray" with an instrumental variable.

OTHER ALGORITHMS

- The book describes other algorithms which improve upon the simple algorithm, in the sense of having smaller variation.
- The methods are described in the book, but are largely algebraic and difficult to explain here.
- However, for most generalized linear models the two methods are fairly similar in practice.

FIRST EXAMPLE

- WISH Data (Women's Interview Study of Health).

$$\begin{aligned}
X &= \text{usual (long–term) average intake of Fat (log scale);} \\
Y &= \text{Fat as measured by a Food Frequency Questionnaire;} \\
W &= \text{Fat as measured by 6 days of 24–hour recalls} \\
T &= \text{Fat as measured by a diary record}
\end{aligned}$$

- Recall the algorithm:
  - ∗ Regress W on T
  - ∗ Form predicted values
  - ∗ Regress Y on the predicted values.
- Dietary intake data have large error, and signals are difficult to find.

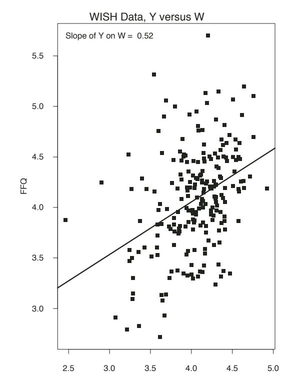

Figure 22: **Wish Data: Regression of FFQ (Y) on Mean of Recalls (W).**

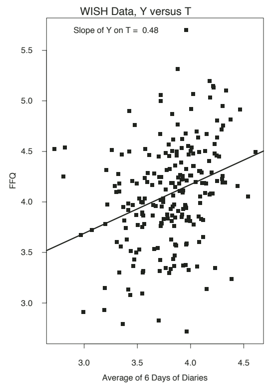

Figure 23: **Wish Data: Regression of FFQ (Y) on Mean of Diaries (T).**

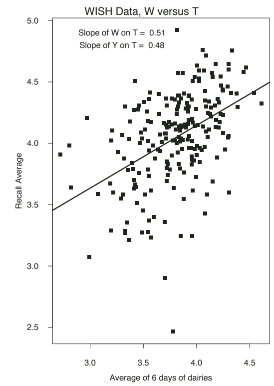

Figure 24: **WISH Data: regression of mean of recalls (W) on mean of diaries (T)**

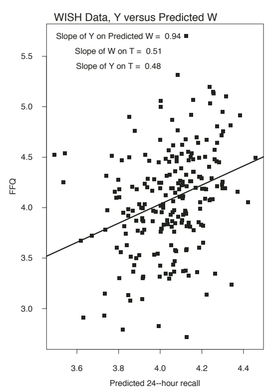

Figure 25: **WISH Data: Regression of FFQ (Y) on the Predictions from the regression of recalls (W) on diaries (T)**

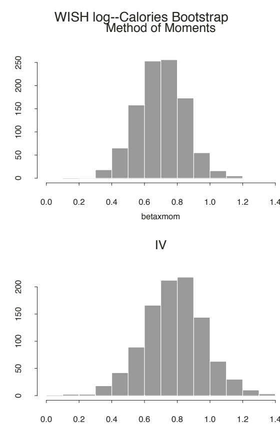

Figure 26: **Bootstrap sampling, comparison with SIMEX and Regression Calibration**

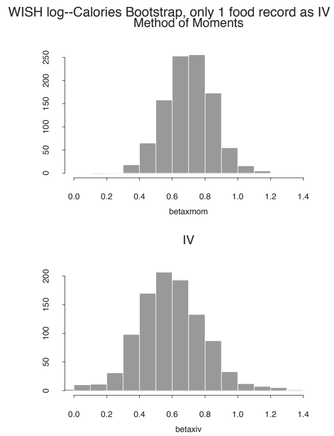

Figure 27: **Bootstrap sampling, comparison with SIMEX and Regression Calibration, when the instrument is of lower quality and one one of the diaries is used.**

FURTHER ANALYSES

- **The naive analysis** has
  - ∗ Slope = 0.4832
  - ∗ OLS standard error = 0.0987
  - ∗ Bootstrap standard error = 0.0946
- **The instrumental variable analysis** has
  - ∗ Slope = 0.8556
  - ∗ Bootstrap standard error = 0.1971
- **For comparison purposes**, the analysis which treats the 6 24–hour recalls as independent replicates has
  - ∗ Slope = 0.765
  - ∗ Bootstrap standard error = 0.1596
- Simulations show that if the 24–hour recalls were really replicates, then the EIV estimate is less variable than the IV estimate.

SEGMENT 7: LIKELIHOOD METHODS OUTLINE

- **Nevada A–bomb test site data**
  - ∗ **Berkson likelihood analysis**
- **Framingham Heart Study**
  - ∗ **Classical likelihood analysis**
- Extensions of the models
- Comments on Computation

NEVADA A–BOMB TEST FALLOUT DATA

- In the early 1990's, Richard Kerber (University of **Utah**) and colleagues investigated the effects of 1950's Nevada A–bomb tests on thyroid neoplasm in exposed children.
- Data were gathered from Utah, Nevada and Arizona.
- Dose to the thyroid was measured by a complex modeling process (more later)
- If true dose in the log–scale is X, and other covariates are Z, fit a **logistic** regression model:

$$\mathrm{pr}(Y = 1|X, Z) = H\left[Z^{\mathrm{T}}\beta_z + \log\{1 + \beta_x \exp(X)\}\right].$$

NEVADA A–BOMB TEST FALLOUT DATA

- **Dosimetry** in radiation cancer epidemiology is a **difficult and time–consuming process**.
- In the fallout study, many factors were taken into account
  - ∗ Age of exposure
  - ∗ Amount of milk drunk
  - ∗ Milk producers
  - ∗ I–131 (a radioisotope) deposition on the ground
  - ∗ Physical transport models from milk and vegetables to the thyroid
- **Essentially all of these steps have uncertainties associated with them.**

NEVADA A–BOMB TEST FALLOUT DATA

- The investigators worked initially in the log scale, and propogated errors and uncertainties through the system.
  - ∗ Much of how they did this is a **mystery to us**.
  - ∗ They took published estimates of measurement errors in food frequency questionnaires in milk.
  - ∗ They also had estimates of the measurement errors in ground deposition of I–131.
  - ∗ And they had **subjective** estimates of the errors in transport from milk to the human to the thyroid.

NEVADA A–BOMB TEST FALLOUT DATA

- Crucially, and as usual in this field, the data file contained not only the **estimated dose** of I–131, but also an **uncertainty** associated with this dose.
- For purposes of today we are going to assume that the error are **Berkson** in the log–scale:

$$X_i = W_i + U_{bi}.$$

  - ∗ The **variance** of U_b is the uncertainty in the data file.

$$\mathrm{var}(U_{bi}) = \sigma_{bi}^2 \text{ known}$$

- And to repeat, the dose–response model of major interest is

$$\mathrm{pr}(Y = 1|X, Z) = H\left[Z^{\mathrm{T}}\beta_z + \log\{1 + \beta_x \exp(X)\}\right].$$

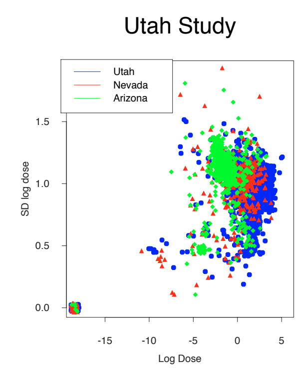

Figure 28: **Log(Dose) and estimated uncertainty in the Utah Data**

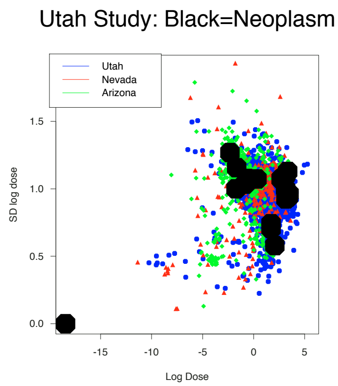

Figure 29: **Log(Dose) and estimated uncertainty in the Utah Data. Large black octogons are the 19 cases of thyroid neoplasm. Note the neoplasm for a person with no dose.**

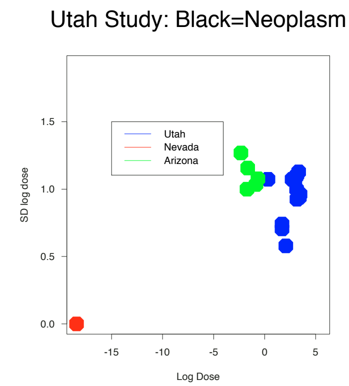

Figure 30: **Log(Dose) and estimated uncertainty in the Utah Data for the thyroid neoplasm cases, by state.**

BERKSON LIKELIHOOD ANALYSIS

- How do we analyze such data?
- We propose that in the **Berkson model**, the only **real** available methods for this complex, heteroscedastic nonlinear logistic model have to be based on **likelihood methods.**
- Let's see if we can understand what the **likelihood** is for this problem.
- **The first step** in any likelihood analysis is to write out the likelihood if there were no measurement error.

BERKSON LIKELIHOOD ANALYSIS

- As a generality, we have a likelihood function for the underlying model in terms of a parameter Θ:

$$\begin{aligned}
&\log\{f_{Y|Z,X}(y|z,x,\Theta)\} \\
&= Y\log\left(H[Z^{\mathrm{T}}\beta_z + \log\{1+\beta_x\exp(X)\}]\right) \\
&+(1-Y)\log\left(1 - H[Z^{\mathrm{T}}\beta_z + \log\{1+\beta_x\exp(X)\}]\right)
\end{aligned}$$

BERKSON LIKELIHOOD ANALYSIS

- **The next step** in the Berkson context is to write out the likelihood function of **true exposure** given the **observed covariates**.
- As a generality, this is

$$f_{X|Z,W}(x|z,w,\mathcal{A}) = \sigma_b^{-1}\phi\left(\frac{x-w}{\sigma_b}\right);$$
$$\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2).$$

- This calculation is obviously dependent upon the problem, and can be more or less difficult.

BERKSON LIKELIHOOD ANALYSIS

- **Likelihood for underlying model**: $f_{Y|Z,X}(y|z,x,\Theta)$
- **Likelihood for error model**: $f_{X|Z,W}(x|z,w,\mathcal{A})$
- We observe only (Y,W,Z).
- **Likelihood for Y given** (W,Z) is

$$\begin{aligned} & f_{Y|W,Z}(y|w,z,\Theta,\mathcal{A}) \\ &= \int f_{Y,X|W,Z}(y,x|w,z,\Theta,\mathcal{A})dx \\ &= \int f_{Y|Z,X}(y|z,x,\Theta)f_{X|Z,W}(x|z,w,\mathcal{A})dx. \end{aligned}$$

BERKSON LIKELIHOOD ANALYSIS

- The likelihood function $f_{Y|W,Z}(y|w,z,\Theta,\mathcal{A})$ can be computed by **numerical integration**.
- The maximum likelihood estimate maximizes the loglikelihood of all the data.

$$L(\Theta,\mathcal{A}) = \sum_{i=1}^{n} \log f_{Y|Z,W}(Y_i|Z_i,W_i,\Theta,\mathcal{A}).$$

- **Maximization program can be used.**

BERKSON LIKELIHOOD ANALYSIS: SUMMARY

- Berkson error modeling is relatively straightforward in general.
- **Likelihood for underlying model**: $f_{Y|Z,X}(y|z,x,\Theta)$
  - ∗ **Logistic nonlinear model**
- **Likelihood for error model**: $f_{X|Z,W}(x|z,w,\mathcal{A})$
  - ∗ In our case, the Utah study data files tells us the Berkson error variance for each individual.

BERKSON LIKELIHOOD ANALYSIS: SUMMARY

- **Overall likelihood** computed by **numerical integration**.

$$f_{Y|W,Z}(y|w,z,\Theta,\mathcal{A}) = \int f_{Y|Z,X}(y|z,x,\Theta) f_{X|Z,W}(x|z,w,\mathcal{A})dx.$$

- The maximum likelihood estimate maximizes

$$L(\Theta,\mathcal{A}) = \sum_{i=1}^{n} \log f_{Y|Z,W}(Y_i|Z_i,W_i,\Theta,\mathcal{A}).$$

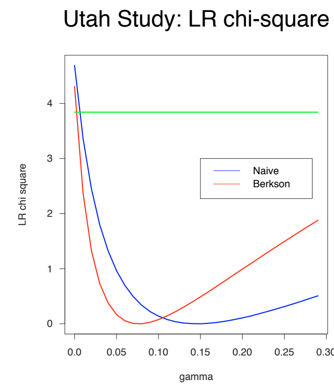

Figure 31: **Likelihood Ratio χ^2 tests for naive and Berkson analyses. Note that the dose effect is statistically significant for both, but that the estimate of γ is larger for the naive than for the Berkson analysis. Very strange.**

CLASSICAL ERROR LIKELIHOOD METHODS—MAIN IDEAS

- There are major differences and complications in the classical error problem with doing a likelihood analysis.
- We will discuss these issues, but once we do we are in business.
- **INFERENCE AS USUAL:**
  * Maximize the density to get point estimates.
  * Invert the Fisher information matrix to get standard errors.
  * Generate likelihood ratio tests and confidence intervals.
  * **These are generally more accurate that those based on normal approximations.**

CLASSICAL ERROR LIKELIHOOD METHODS—STRENGTHS

- **STRENGTHS:** can be applied to a wide class of problems
  * including discrete covariates with misclassification
- **Efficient**
  * makes use of assumptions about the distribution of X.
  * can **efficiently combine different data types**, e.g., validation data with data where X is missing.
  * Linear measurement error with missing data is a case where maximum likelihood seems much more efficient than functional methods.

CLASSICAL ERROR LIKELIHOOD METHODS—WEAKNESSES:

- **Need to parametrically model every component of the data (structural not functional)**
  - ∗ Need a parametric model for the unobserved predictor.
  - ∗ **robustness is a major issue** because of the **strong parametric assumptions.**
  - ∗ Special computer code may need to be written
  - ∗ but can use packaged routines for numerical integraton and optimization.

FRAMINGHAM HEART STUDY DATA

- The aim is to understand the relationship between coronary heart disease (CHD $= Y$) and systolic blood pressure (SBP) in the presence of covariates (age and smoking status).
- SBP is known to be **measured with error.**
  - ∗ If we define $X = \log(\text{SBP} - 50)$, then about 1/3 of the variability in the observed values W is due to error.
  - ∗ **Classical error is reasonable here.**
  - ∗ The measurement error is essentially known to equal $\sigma_u^2 = 0.01259$
- Here is a q–q plot of the observed SBP's (W), along with a density estimate.

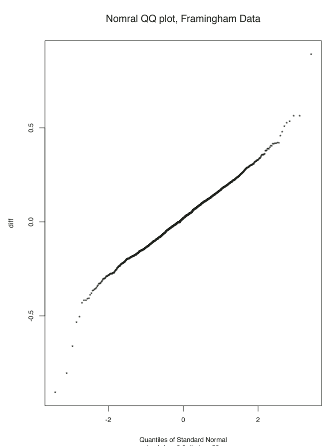

Figure 32: **q–q plot in Framingham for** $\log(\textbf{SBP} - 50)$

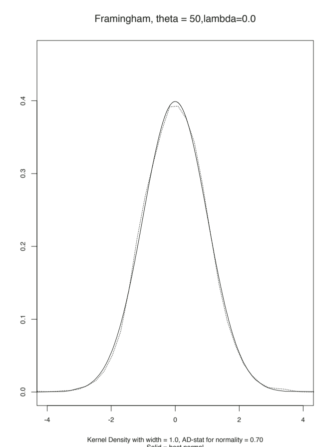

Figure 33: **Kernel density estimate and best fitting normal density plot in Framingham for** $\log(\textbf{SBP} - 50)$

FRAMINGHAM HEART STUDY DATA

- We will let age and smoking status be denoted by Z.
- A reasonable model is **logistic regression.**

$$\begin{aligned}\mathrm{pr}(Y=1|X,Z) &= H(\beta_0+\beta_z^{\mathrm{T}}Z+\beta_x X);\\ &= 1./\left\{1+\exp(\beta_0+\beta_z^{\mathrm{T}}Z+\beta_x X)\right\}.\end{aligned}$$

- **A reasonable error model** is

$$W = X + U, \sigma_u^2 = 0.01259.$$

- W is only *very* weakly correlated with Z. Thus, **a reasonable model for X given Z** is

$$X \sim \mathrm{Normal}(\mu_x, \sigma_x^2).$$

FRAMINGHAM HEART STUDY DATA

- We have now specified everything we need to do a likelihood analysis.
  * **A model for Y given (X,Z)**
  * **A model for W given (X,Z)**
  * **A model for X given Z.**
- The unknown parameters are β_0, β_z, β_x, μ_x, σ_x^2.
- We need a formula for the likelihood function, and for this we need a little theory.

LIKELIHOOD WITH AN ERROR MODEL

- Assume that we observe (Y,W,Z) on every subject.
- $f_{Y|X,Z}(y|x,z,\beta)$ is the density of Y given X and Z.
  * this is the **underlying model of interest.**
  * the density depends on an unknown parameter β.
- $f_{W|X,Z}(w|x,z,\mathcal{U})$ is the conditional density of W given X and Z.
  * **This is the error model.**
  * It depends on another unknown parameter $\mathcal{U}$.
- $f_{X|Z}(x|z,\alpha_2)$ is the density of X given Z depending on the parameter $\mathcal{A}$. This is the **model for the unobserved predictor**. This density may be hard to specifiy but it is needed. This is where **model robustness** becomes a big issue.

LIKELIHOOD WITH AN ERROR MODEL—CONTINUED

- The joint density of (Y,W) given Z is

$$\begin{aligned}&f_{Y,W|Z}(y,w|z,\beta,\mathcal{U},\mathcal{A})\\ &= \int f_{Y,W,X|Z}(y,w,x|z)dx\\ &= \int f_{Y|X,Z,W}(y|x,z,w,\beta)f_{W|X,Z}(w|x,z,\mathcal{U})\\ &\qquad \times f_{X|Z}(x|z,\mathcal{A})dx\\ &= \int f_{Y|X,Z}(y|x,z,\beta)f_{W|X,Z}(w|x,z,\mathcal{U})\\ &\qquad \times f_{X|Z}(x|z,\mathcal{A})dx.\end{aligned}$$

  * The assumption of **nondifferential measurement error** is used here, so that $f_{Y|X,W,Z} = f_{Y|X,Z}$.
  * The integral will ususally be calculated numerically.
  * The integral is replaced by a sum if X is discrete.
  * Note that $f_{Y,W|Z}$ depends of $f_{X|Z}$—again this is why robustness is a worry.

LIKELIHOOD WITH AN ERROR MODEL—CONTINUED

- The log-likelihood for the data is, of course,

$$L(\beta, \alpha) = \sum_{i=1}^{n} \log f_{Y,W}(Y_i, W_i|\beta, \alpha).$$

- The log–likelihood is often computed numerically,
- Function maximizers can be used to compute the likelihood analysis.

LIKELIHOOD WITH AN ERROR MODEL—CONTINUED

- If X is scalar, generally the likelihood function can be computed numerically and then maximized by a function maximizer.

$$f_{Y,W|Z}(y, w|z, \beta, \mathcal{U}, \mathcal{A})$$
$$= \int f_{Y|X,Z}(y|x, z, \beta) f_{W|X,Z}(w|x, z, \mathcal{U}) f_{X|Z}(x|z, \mathcal{A}) dx.$$

- We did this in the Framingham data.
  * We used starting values for β_0, β_z, β_x, μ_x, σ_x^2 from the naive analysis which ignores measurement error.
  * We will show you the profile loglikelihood functions for β_x for both analyses.

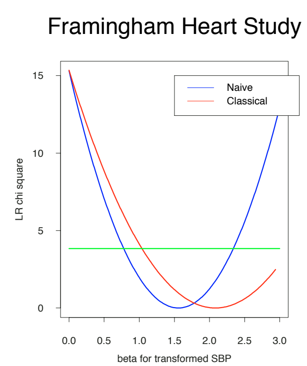

Figure 34: **Profile likelihoods for SBP in Framingham Heart Study.**

A NOTE ON COMPUTATION

- It is almost always better to **standardize the covariates** to have sample mean zero and sample variance one.
- Especially in logistic regression, this improves the accuracy and stability of numerical integration and likelihood maximization.

A NOTE ON COMPUTATION

- **Not all problems are amenable to numerical integration to compute the log–likelihood**
  - * **Mixed GLIM's** is just such a case.
  - * In fact, for mixed GLIM's, the likelihood function *with no measurement error is not computable*
- In these cases, specialized tools are necessary. Monte–Carlo EM (McCulloch, 1997, JASA and Booth & Hobert, 1999, JRSS–B) are two examples of Monte–Carlo EM.

EXTENSIONS OF THE MODELS

- It's relatively easy to write down the likelihood of complex, nonstandard models.
  - * So likelihood analysis is a good option when the data or scientific knowledge suggest a nonstandard model.
- For example, multiplicative measurement error will often make sense. These are additive models in the log scale, e.g., the Utah data.
- Generally, the numerical issues are no more or less difficult for multiplicative error.