MEASUREMENT ERROR IN NONLINEAR MODELS

- Segment 1 — Introduction, Linear Models
- Segment 2 — Nonlinear Models, Data Types
- Segment 3 — Functional/Structural, Berkson/Classical
- Segment 4 — Likelihood Methods
- Segment 5 — Bayesian Methods
- Segment 6 — Regression Calibration
- Segment 7 — Simulation Methods
- Segment 8 — Instrumental Variables
- Segment 9 — Nonparametric Methods
-

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

- **Different measures of nutrient intake**
- **Coronary Heart Disease vs Systolic Blood Pressure**
- **Radiation Dosimetry**

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
- 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.
  - ∗ The analysis of 24–hour recall introduces some error $\Longrightarrow$ analysis error
  - ∗ **Measurement error** = **sampling error** + **analysis error**
  - ∗ Measurement error model

    $W_i = X_i + U_i$, U_i 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)
- 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 $\Longrightarrow$ analysis error
  - **Measurement error** = **sampling error** + **analysis error**
  - Measurement error model $W_i = X_i + U_i$, U_i are measurement errors

GENERAL STRUCTURE OF A MEASUREMENT ERROR PROBLEM

- Y = response, Z = error-free predictor, X = error-prone predictor
- $E(Y|Z,X) = f(Z,X,\beta)$
- Observed data **does not generally preserve the regression structure**: $(Y_i,\ Z_i,\ W_i),\ i = 1,\ldots,n$
  - $E(Y|Z,W) \neq f(Z,W,\beta)$
- Error model relating W_i and X_i
  - $W_i = X_i + U_i$ (classical error model)
  - $W_i = \gamma_{0,\text{em}} + \gamma^t_{x,\text{em}} X_i + \gamma^t_{z,\text{em}} Z_i + U_i$ (error calibration model)
  - Other common models

A CLASSICAL ERROR MODEL

- $W_i = X_i + U_i$ (**additive**)
- U_i are:
  * independent of all Y_i, Z_i and X_i (independent)
  * IID$(0,\ \sigma_u^2)$ (IID, unbiased, homoscedastic)

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:
  - $\ast$ independent of all Y_i, Z_i and X_i (independent)
  - $\ast$ 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**

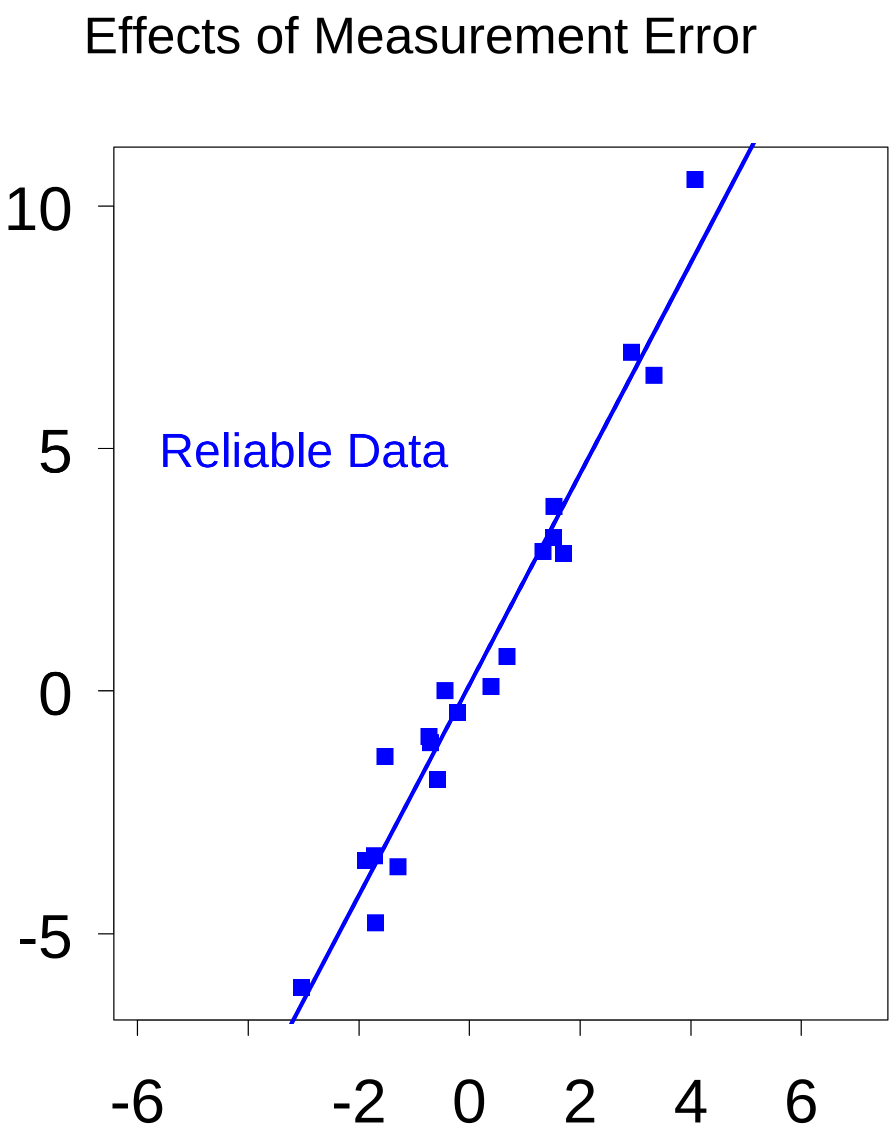

Figure 1: **True Data Without Measurement Error.**

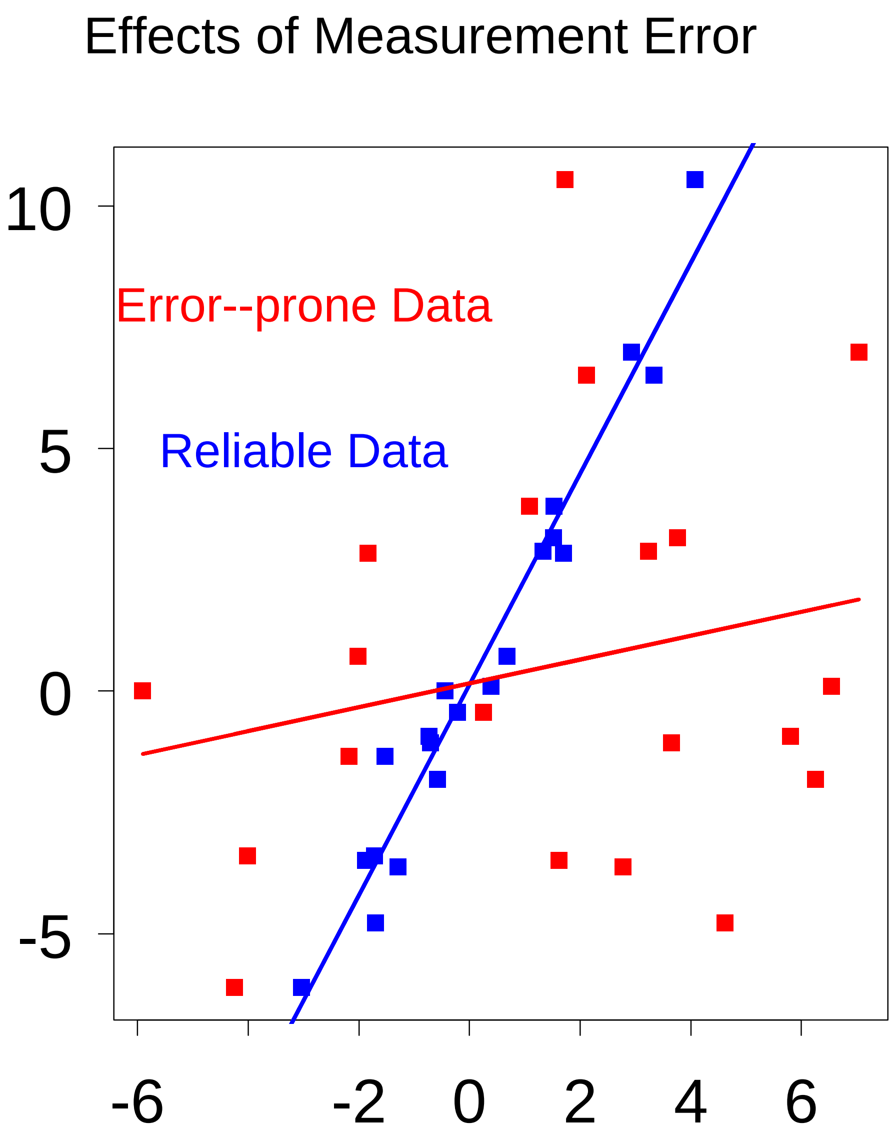

Figure 2: **Observed Data With Measurement Error.**

THEORY BEHIND THE PICTURES: THE NAIVE ANALYSIS

- Least Squares Estimate of Slope:

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

where

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

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

So

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

where

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

is the **attenuation factor** or **reliability ratio**

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:

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

MORE THEORY: JOINT NORMALITY

- $Y,\ X,\ W$ jointly normal $\Longrightarrow$
  - $Y \mid W \sim$ Normal
  - $E(Y \mid W) = \beta_0 + (1-\lambda)\beta_x\mu_x + \lambda\beta_x W$
  - $\mathrm{Var}(Y \mid W) = \sigma^2_\epsilon + (1-\lambda)\beta^2_x\sigma^2_x$

- Intercept is **shifted** by $(1-\lambda)\beta_x\mu_x$
- Slope is **attenuated** by the factor λ
- Residual variance is **inflated** by $(1-\lambda)\beta^2_x\sigma^2_x$
- **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 **naive test of $H_0 : \beta_x = 0$ is asymptotically efficient when $E(X \mid W)$ is linear in W.**

- The discussion of naive tests when there are **multiple predictor** measured with error, or **error-free** predictors, is **more complicated**

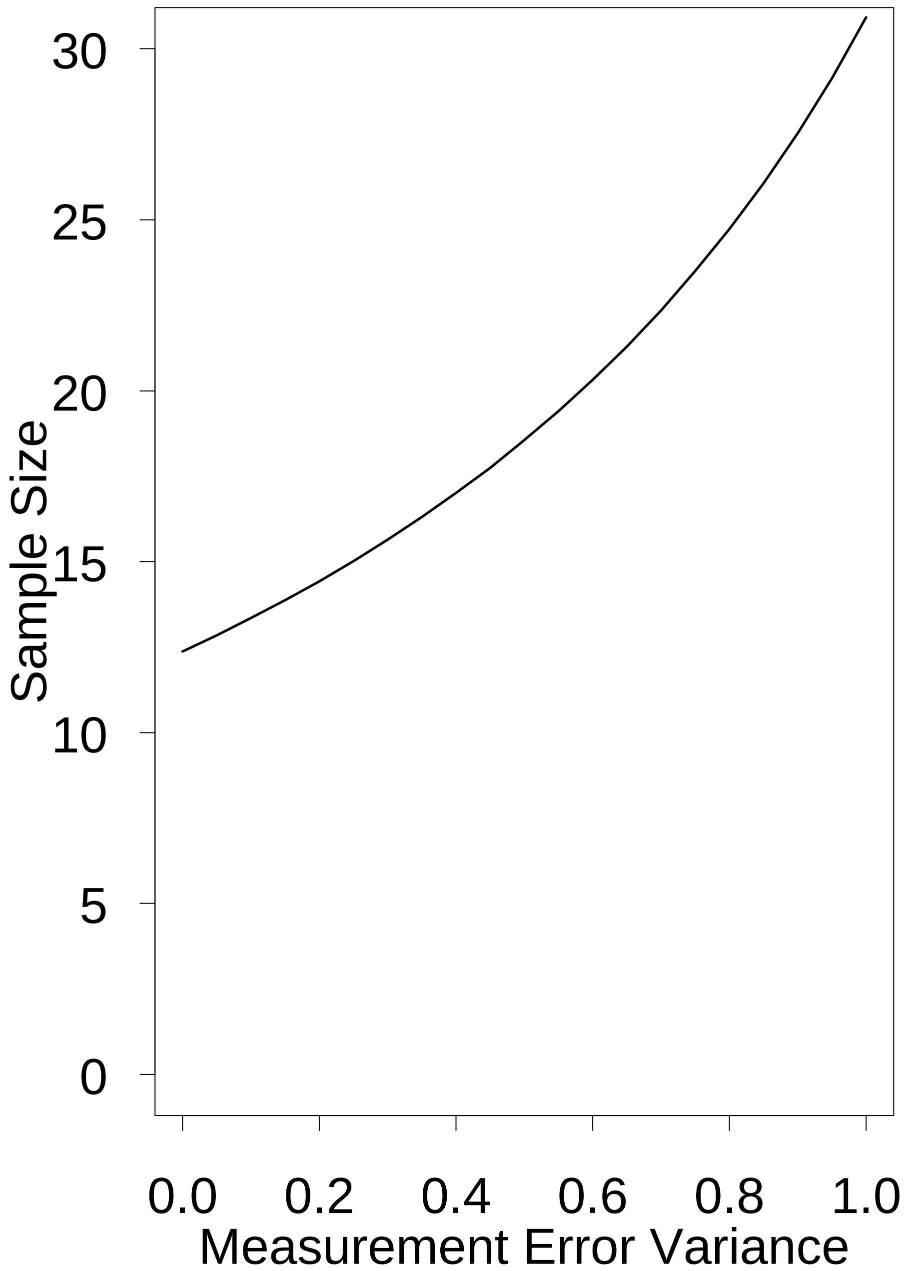

Figure 3: **Sample Size for 80% Power. True slope $\beta_x = 0.75$. Variances $\sigma^2_x = \sigma^2_\epsilon = 1$.**

MULTIPLE LINEAR REGRESSION WITH ERROR

- Model

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

- 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)
  - $*$ Naive test of $H_0 : \beta_z = 0$ is typically not valid (same is true for subvectors)

MULTIPLE LINEAR REGRESSION WITH ERROR

- For X scalar, attenuation factor in β_{x*} is

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

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

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

  * $\Longrightarrow$ **Collinearity accentuates attenuation**

- Biased estimates of β_z:

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

  * Γ_z is from $\mathrm{E}(X \mid Z) = \Gamma_1 + \Gamma_z^{\mathrm{T}} Z$

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 $\mathrm{E}(X \mid Z) = \Gamma_1 + \Gamma_z^t Z$

  ∗ 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!!!

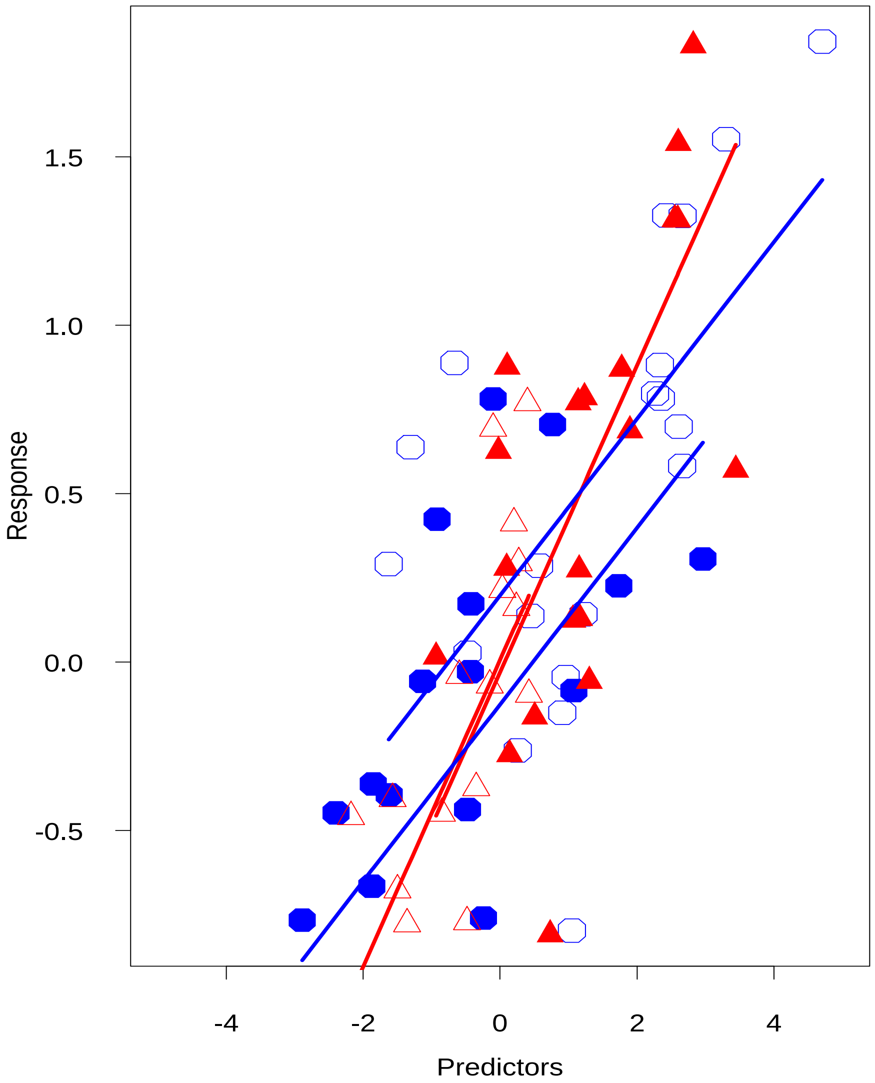

Figure 4: **UNBALANCED ANCOVA. RED = TRUE DATA, BLUE = OBSERVED. SOLID = FIRST GROUP, OPEN = SECOND GROUP. NO DIFFERENCE IN GROUPS.**

Regression	Intercept	Slope	Residual Variance
Y on X	β_0	β_x	σ_ϵ^2
Y on W	$\beta_0 + (1-\lambda)\beta_x\mu_x$	$\lambda\beta_x$	$\sigma_\epsilon^2 + (1-\lambda)\beta_x^2\sigma_x^2$

SUMMARY OF THE EFFECTS OF MEASUREMENT ERROR IN SIMPLE LINEAR REGRESSION

- Regression Model

$$Y = \beta_0 + \beta_x X + \epsilon$$
$$W = X + U$$

- Attenuation Factor (Reliability Ratio)

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

  * $0 < \lambda \leq 1$
  * $\lambda = 1 \iff \sigma_u^2 = 0$

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.
- Sometime, 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.
- More formally, (W, T) are **conditionally independent** of Y given (X, Z).
  - (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,

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

HEARTH 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.
- $W = (NO_2 \text{ in kitchen}, NO_2 \text{ in bedroom})$ 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**.
  - The sampled the plant for dust.
  - Then they tried to match the person to where he/she worked.

IS THIS NONDIFFERENTIAL?

- 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
- 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.
- For illustrative purposes throughout, we will pretend **nondifferential**. **However, recent work of Kipnis and Carroll suggest that this is not the case. It is very subtle!!!**

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** $\mathrm{var}(U)$
- 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 is

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

- This is an unbalanced 1–factor ANOVA with mean squared error 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 $\overline{W}_{i\bullet}$.

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

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 $W_{i1} - W_{i2} = U_{i1} - U_{i2}$.
- 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, I always 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)

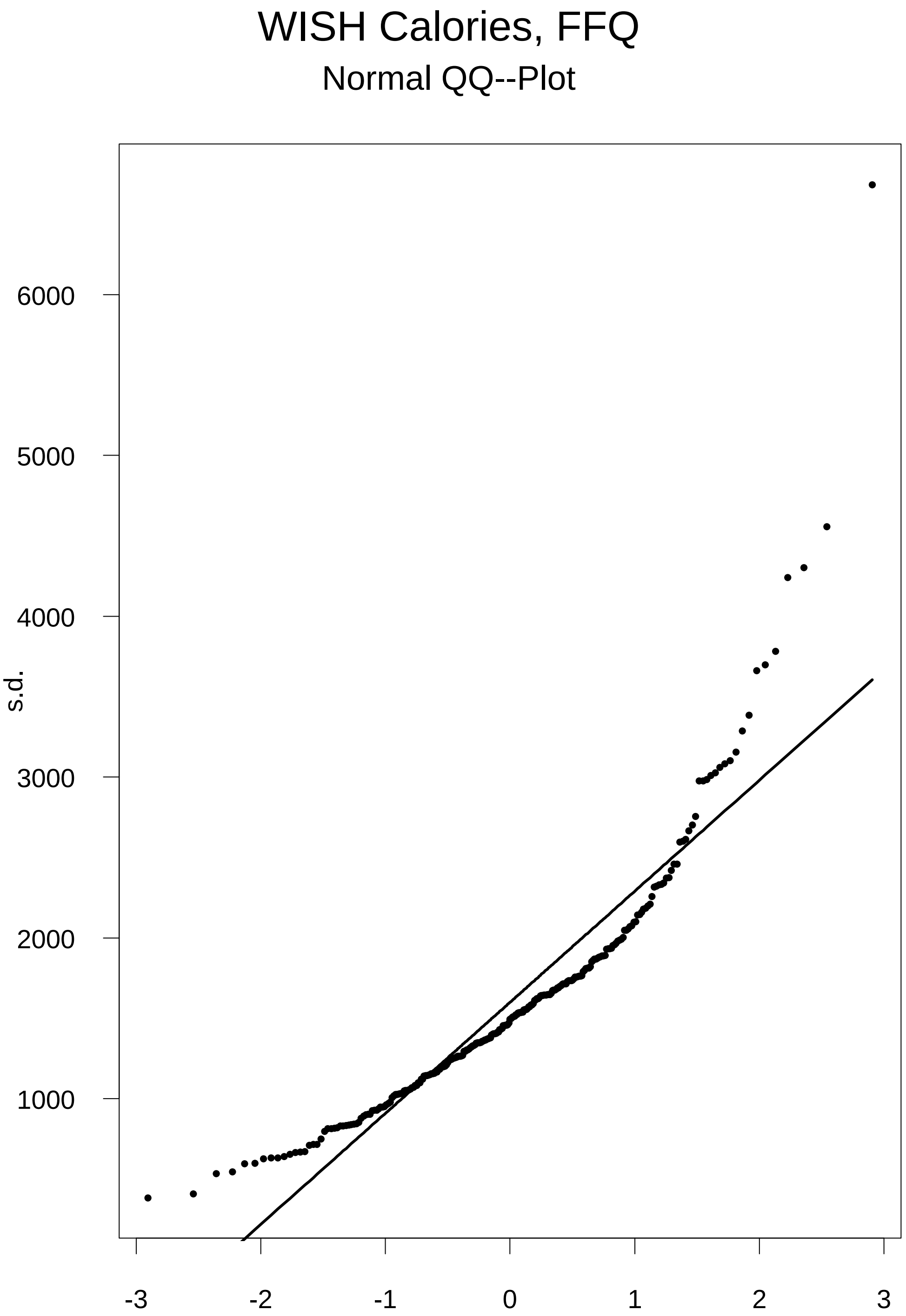

Figure 5: **WISH, CALORIC INTAKE, Q–Q plot of Observed data.**

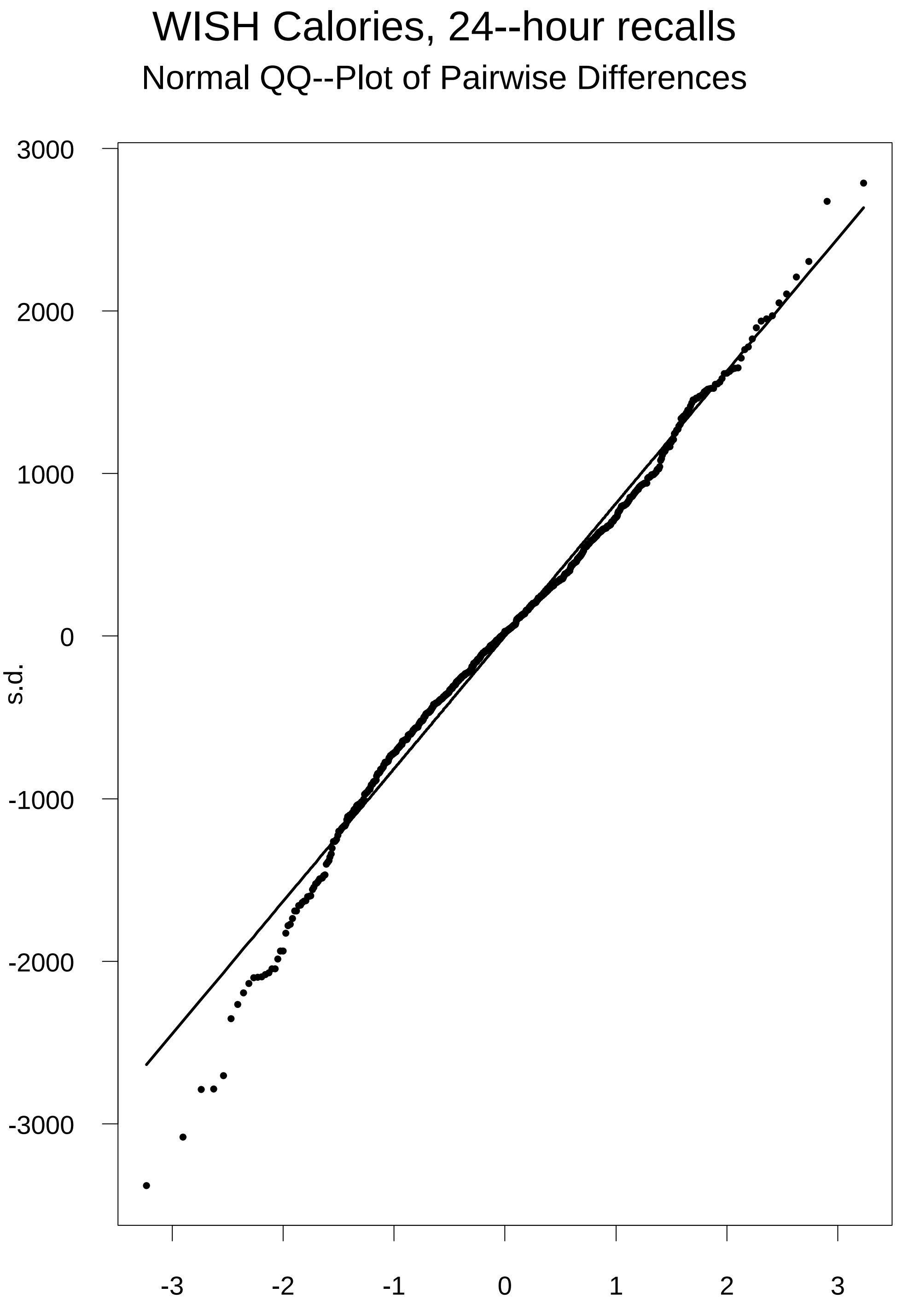

Figure 6: **WISH, CALORIC INTAKE, Q–Q plot of Differenced data.**

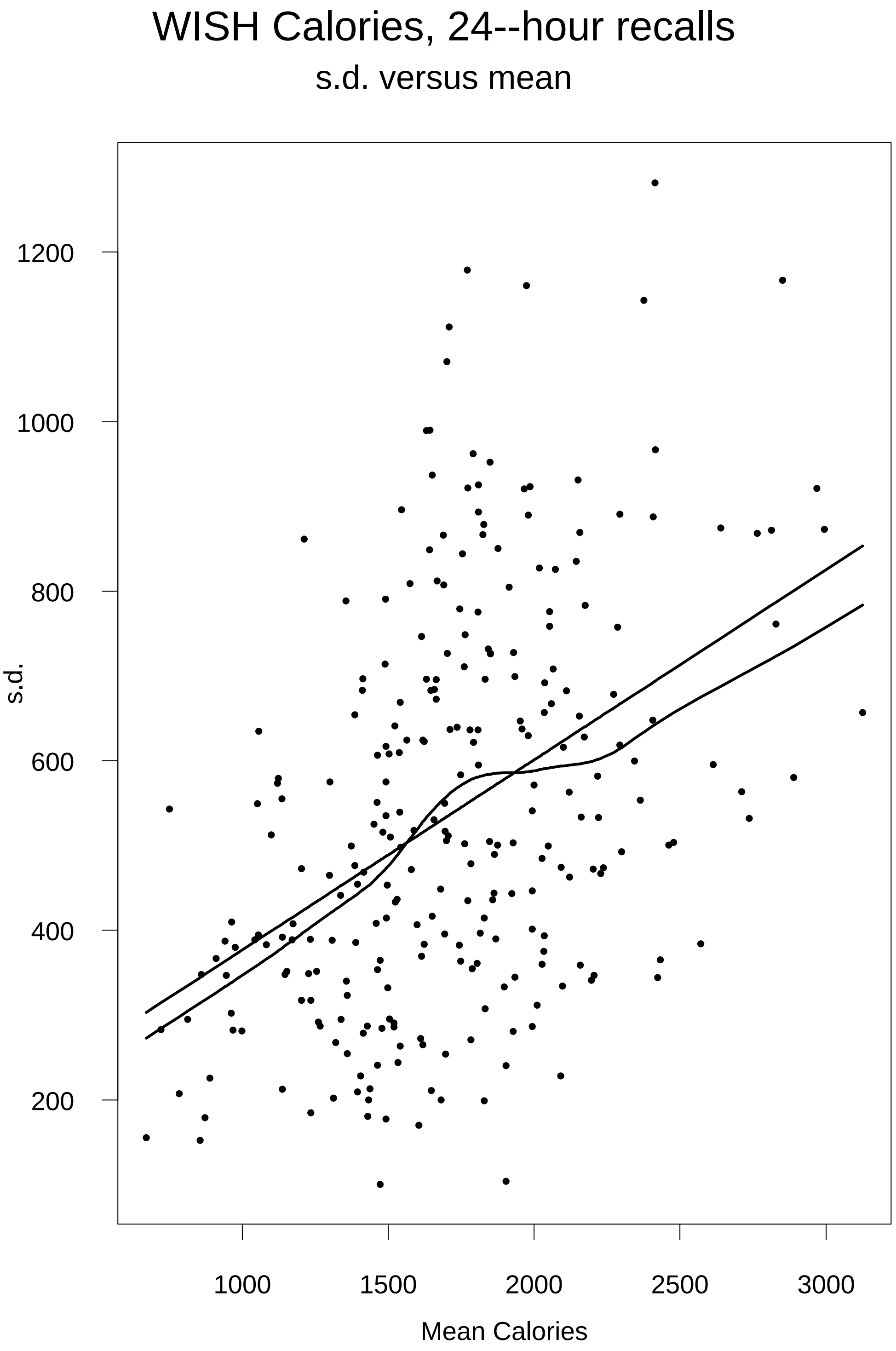

Figure 7: **WISH, CALORIC INTAKE, plot for additivity, loess and OLS.**

REPLICATION: WISH

- Taking logarithms improves all the plots.

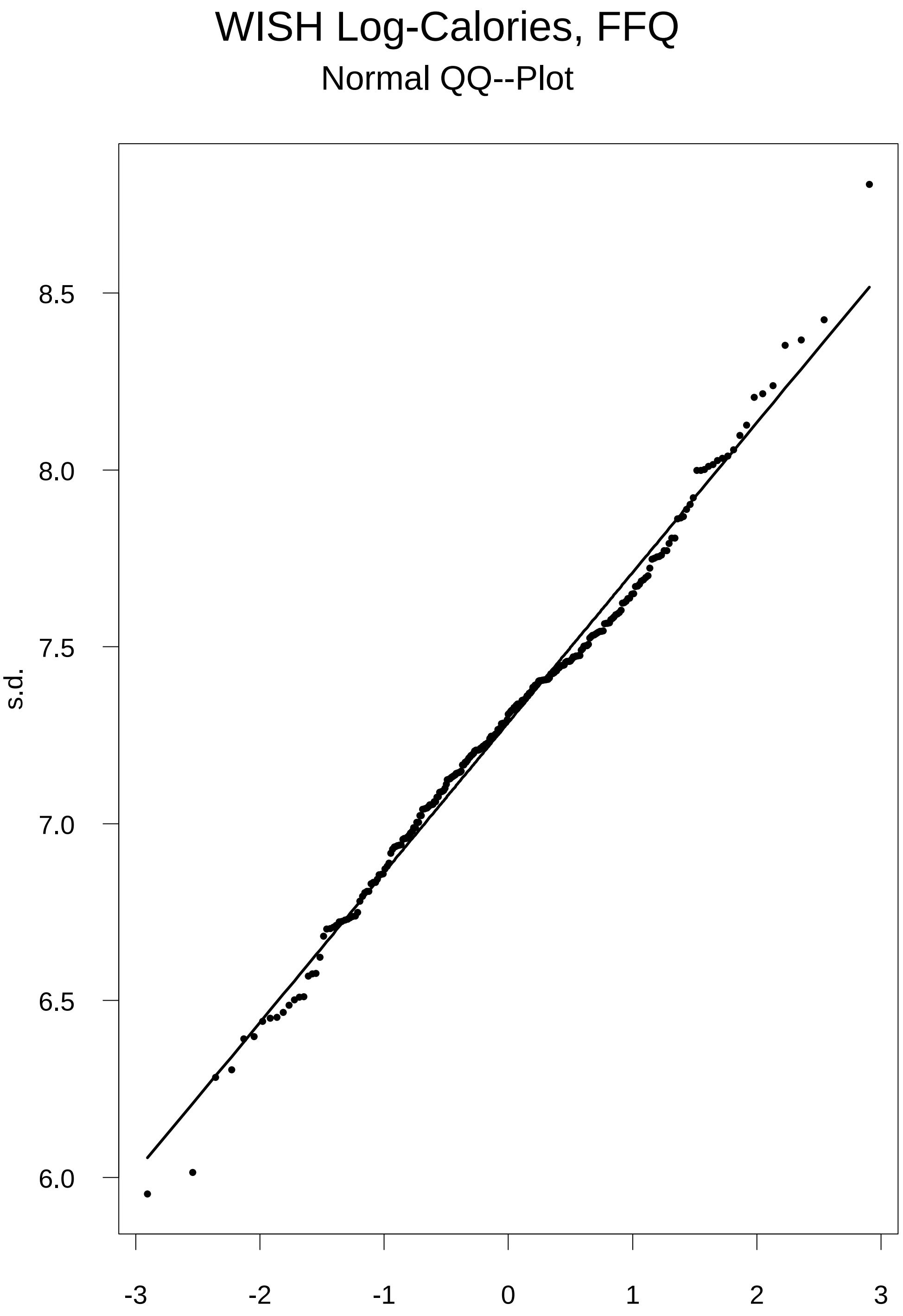

Figure 8: **WISH, LOG CALORIC INTAKE, Q–Q plot of Observed data.**

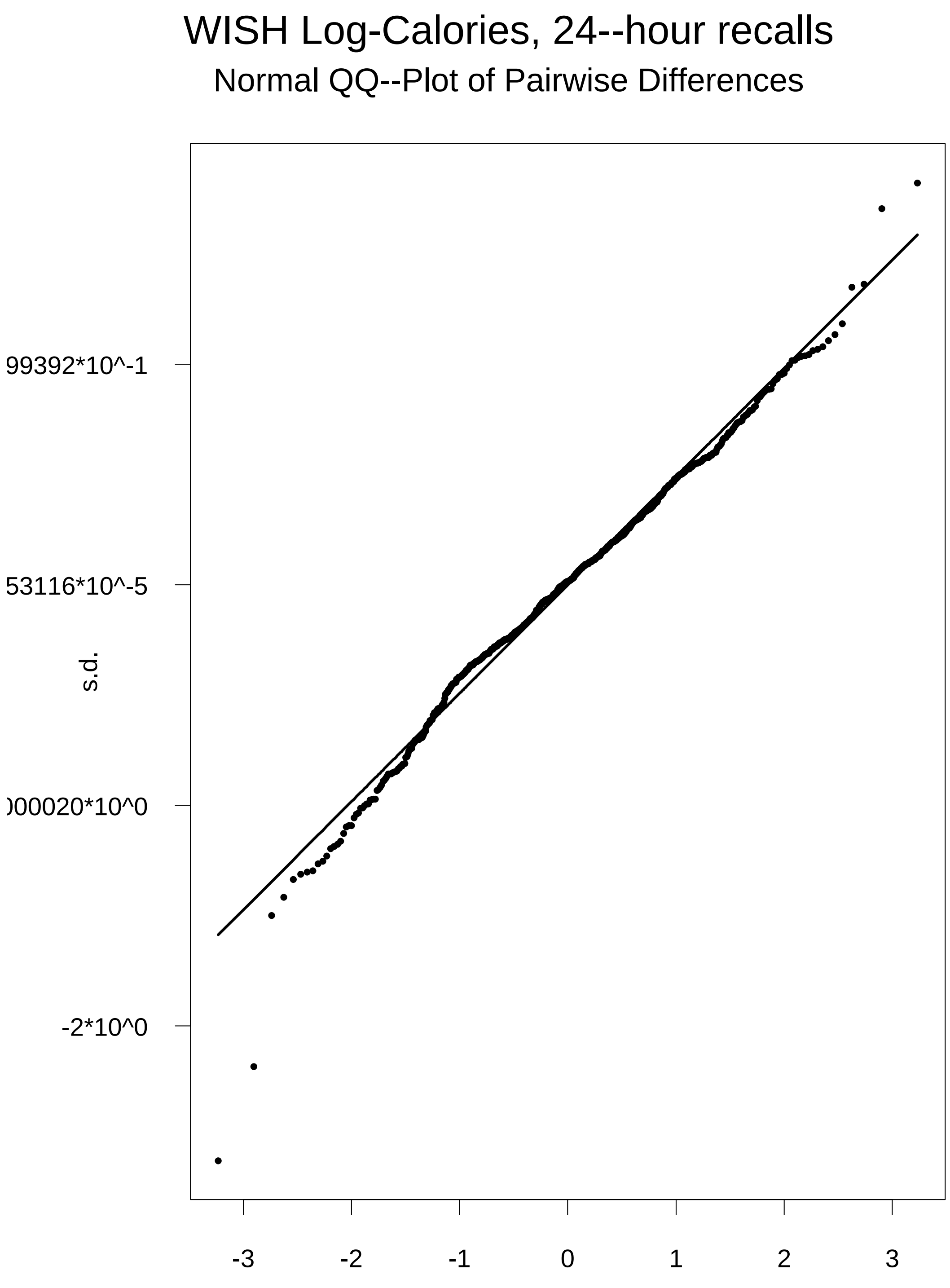

Figure 9: **WISH, LOG CALORIC INTAKE, Q–Q plot of Differenced data.**

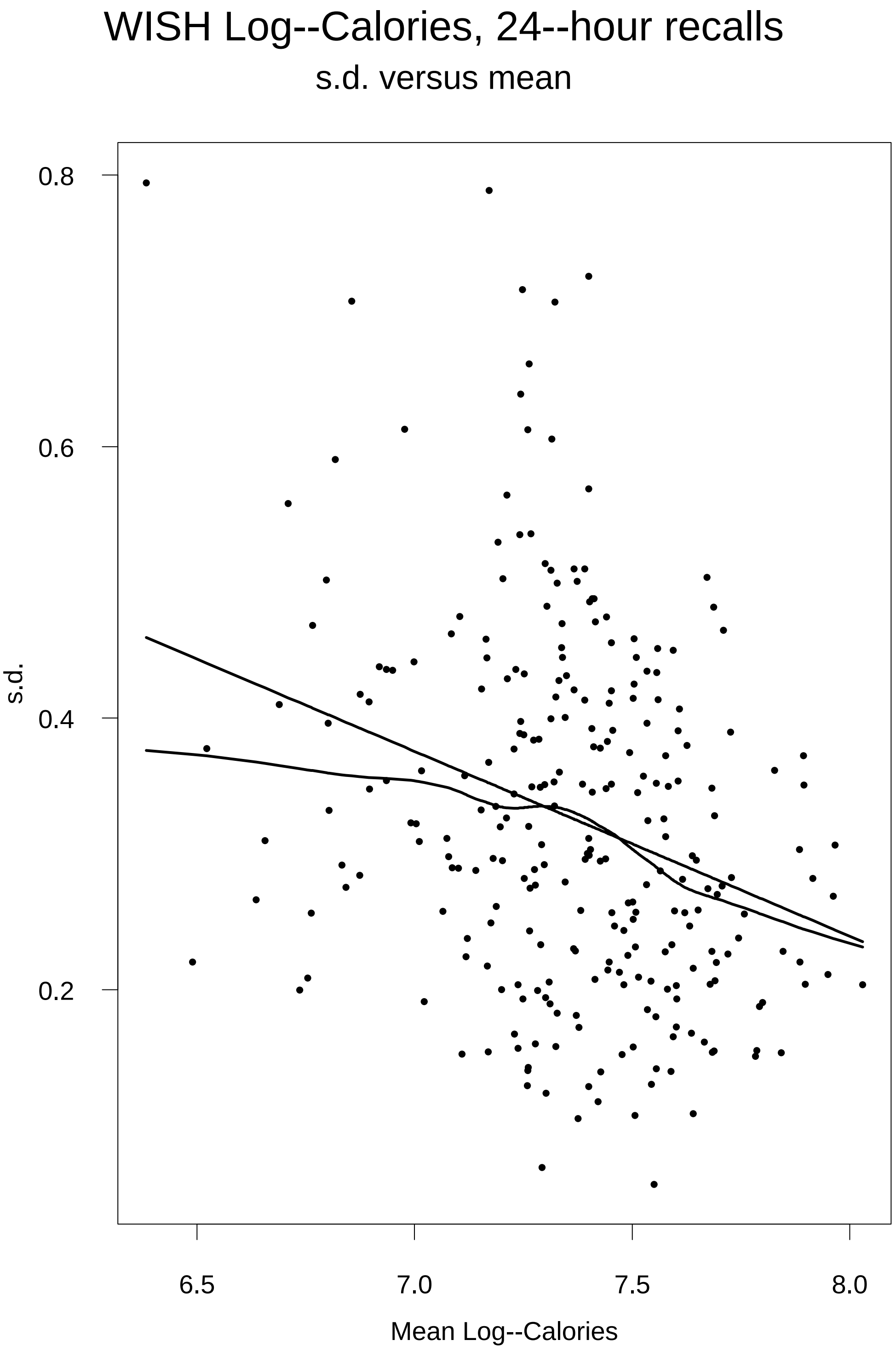

Figure 10: **WISH, LOG CALORIC INTAKE, plot for additivity, loess and OLS.**

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 tested via **replication**
  - $*$ Sample standard deviation versus sample mean.
  - $*$ q–q plots of differences of within–person replicates.

SEGMENT 3: BASIC CONCEPTUAL ISSUES

- **Transportability**
- **Is there an "exact" predictor?**
- **What is Berkson? What is classical?**
- **Functional versus structural modeling**

TRANSPORTABILITY AND THE LIKELIHOOD

- In linear regression, we have seen that **we only require to know the measurement error variance** (after checking for semi–constant variance, additivity, normality).

- **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

$$\begin{aligned} E(Y|W) &= \beta_0 + \beta_x E(X|W) \\ &\approx \beta_0 + \beta_x \lambda W. \end{aligned}$$

- As we will see later, in general problems we can do a likelihood analysis if we know the distribution of X given W.

- 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 this 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

- We say that a model is **transportable** across studies if the model holds **with the same parameters** in the two studies.
  - ∗ **Internal data** 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 entire distribution!!**
  * var(differences, NHS = 47)
  * var(differences, ACS = 45)
  * var(sum, NHS = 152)
  * var(sum, ACS = 296)

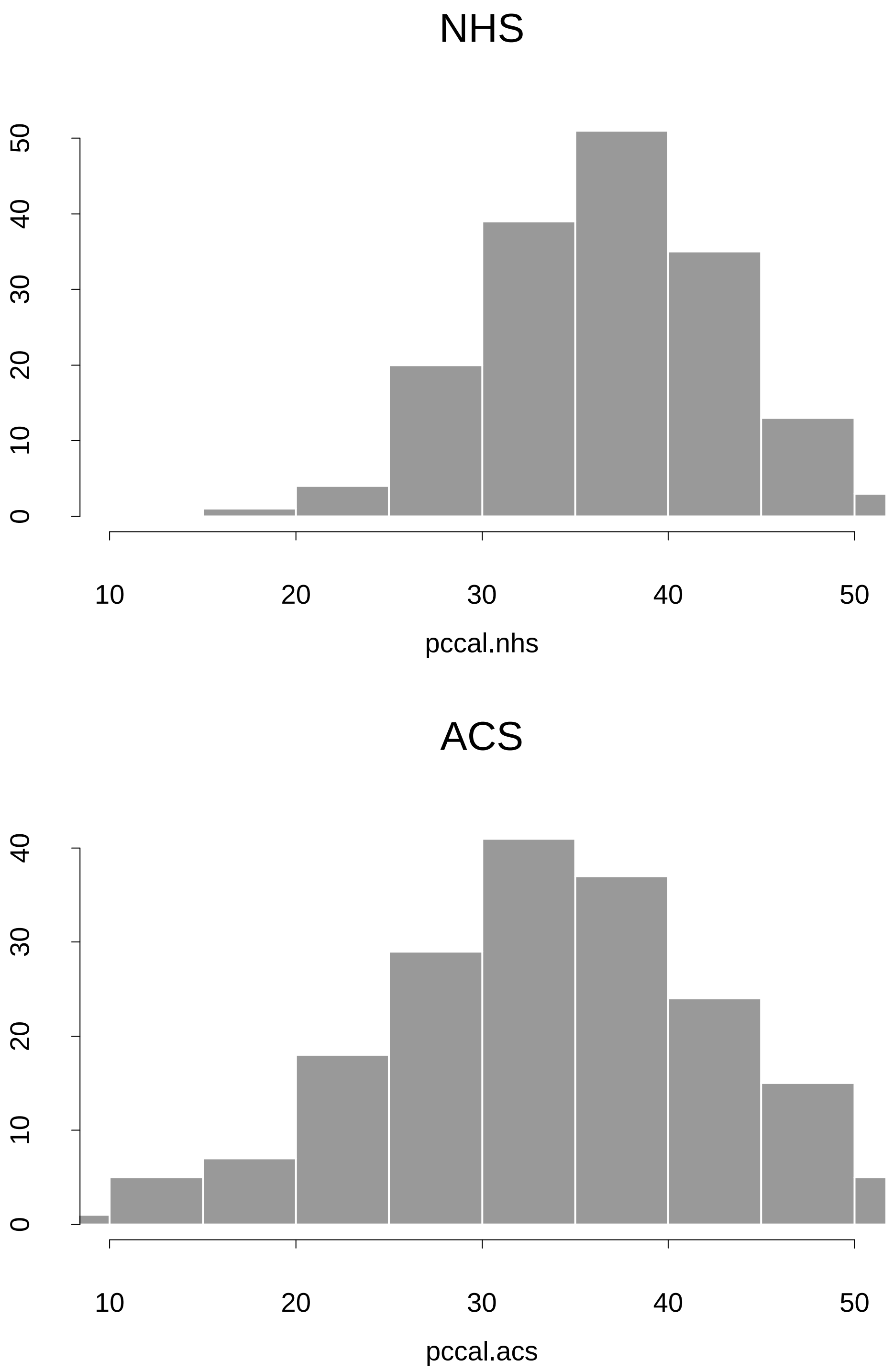

Figure 11: **FFQ Histograms in NHS and ACS**

IS THERE AN "EXACT" PREDICTOR?

- One can distinguish between **two concepts of X**:
  - ∗ the **actual exact predictor**, which may never be observable under any circumstances.
  - ∗ the **operationally-defined exact predictor** which could be observed, albeit at great cost and effort.
- Example: X is long-term caloric intake.
  - ∗ actual X is long-term intake over lifetime.
  - ∗ operationally-defined X is long-term intake since inception of the study, as measured by the best available instrument.

IS THERE AN "EXACT" PREDICTOR? —CONTINUED

- One can construct hypothetical examples where the effects of actual X and of operationally-defined X are quite different, even oppostie in sign.
- Obviously, a policy to change X, say to reduce intake of saturated fat or of calories, affects the actual X, not the operationally defined X.
- Nonetheless, the operationally-defined X is all we can work with.
- **"Gold standard"** generally means the operationally-defined X.
  - **We will take "exact predictor" and "gold standard" to both refer to operationally-defined X.**
  - However, for some authors, **"gold standard"** means the actual exact X even if not operationally-defined.

THE BERKSON MODEL

- The **Berkson model** says that

True Exposure = Observed Exposure + Mean Zero Error

$$X = W + U_b \text{ or } X = W \times U_b$$

- 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?

- The reason why people have trouble distinguishing the two is that in real life, it takes hard thought to do so.

WHAT'S BERKSON? WHAT'S CLASSICAL?

- Direct measures possible?
- Replication possible?
- $X = W + U_b$?
- $W = X + U$?
- **Let's play stump the experts!**
- Framingham Heart Study
  - Predictor is systolic blood pressure

- 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.

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

- Long–term lead exposure in children as measured by repeated personal dosimetry.

- Student lab tech follows protocol for mice in acute inhalation studies to PERC, exposing them to a pre-specified concentration and duration.

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 CLASSICAL ERROR MODELS

- 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 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.

COMMENTS ON FUNCTIONALITY CLASSICAL ERROR MODELS

- Unanswered (unanswerable?) question: how sensitive are parameter inferences to an assumed **STRUCTURAL** model?
  - **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, serves as a useful check on more complex parametric models.
- How much does it "cost" to be functional?
  - There are examples in which the best functional method is inefficient (Missing Data, Thresholds).

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 (Chapter 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.
- It is important not to blindly throw away these methods.
  - ∗ **Model robustness is a real issue.**
  - ∗ Functional methods are often easy to compute.
  - ∗ **Potential for inefficiency $\neq$ inefficiency.**

SEGMENT 4: 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 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 **standard deviation** of U_b is the uncertainty in the data file.

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

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

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

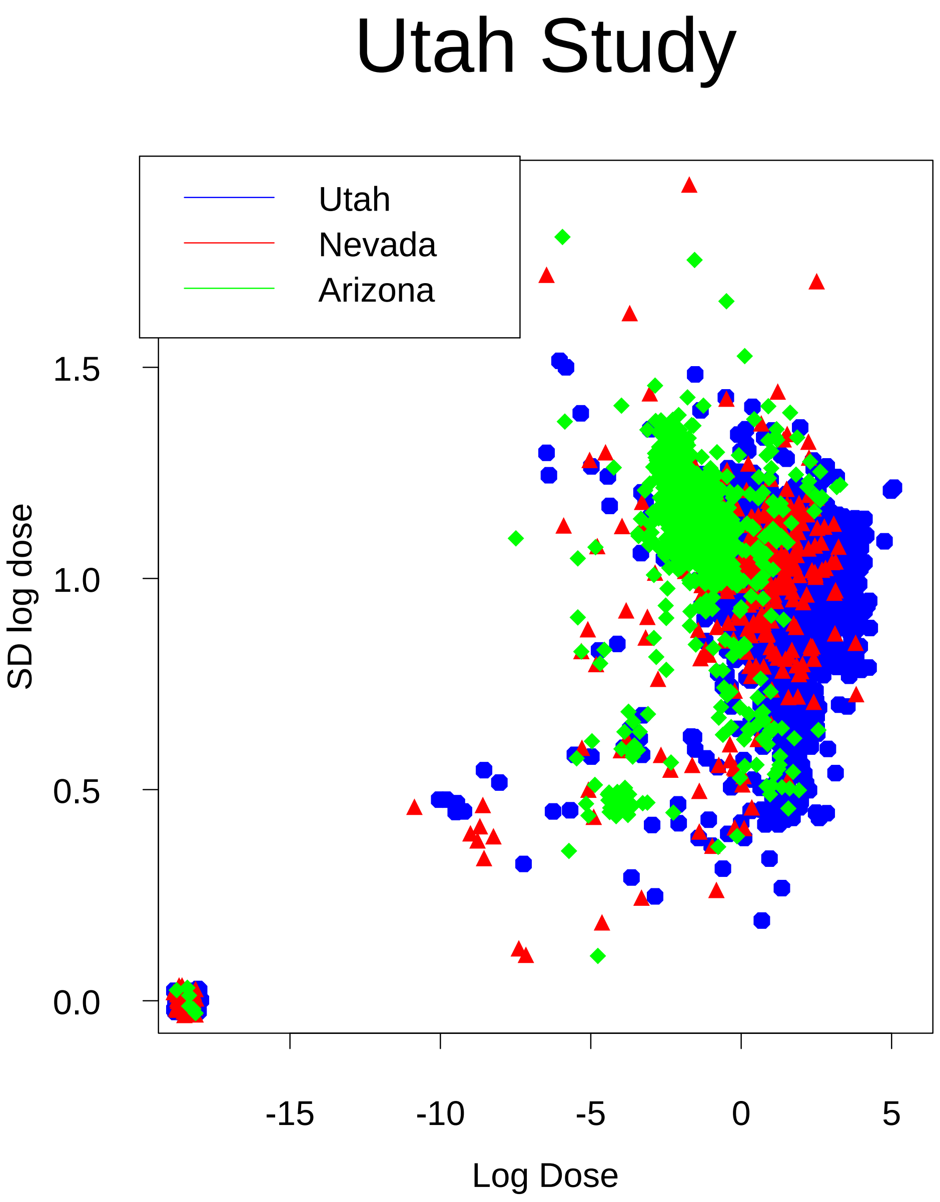

Figure 12: **Log(Dose) and estimated uncertainty in the Utah Data**

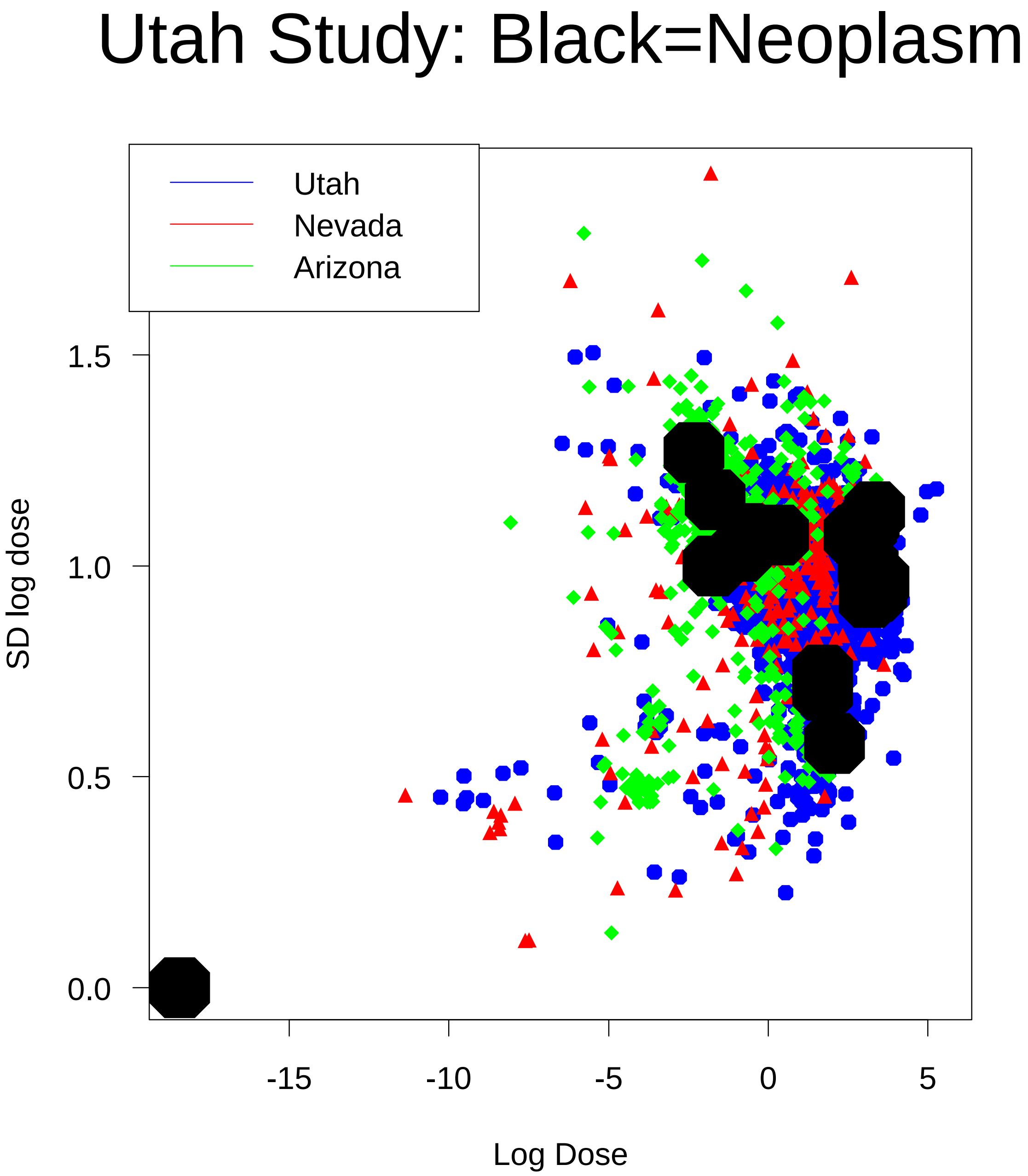

Figure 13: **Log(Dose) and estimated uncertainty in the Utah Data. Large black octogons are the 19 cases of thyroid neoplasm. Note the neoplasm for a person with no dose.**

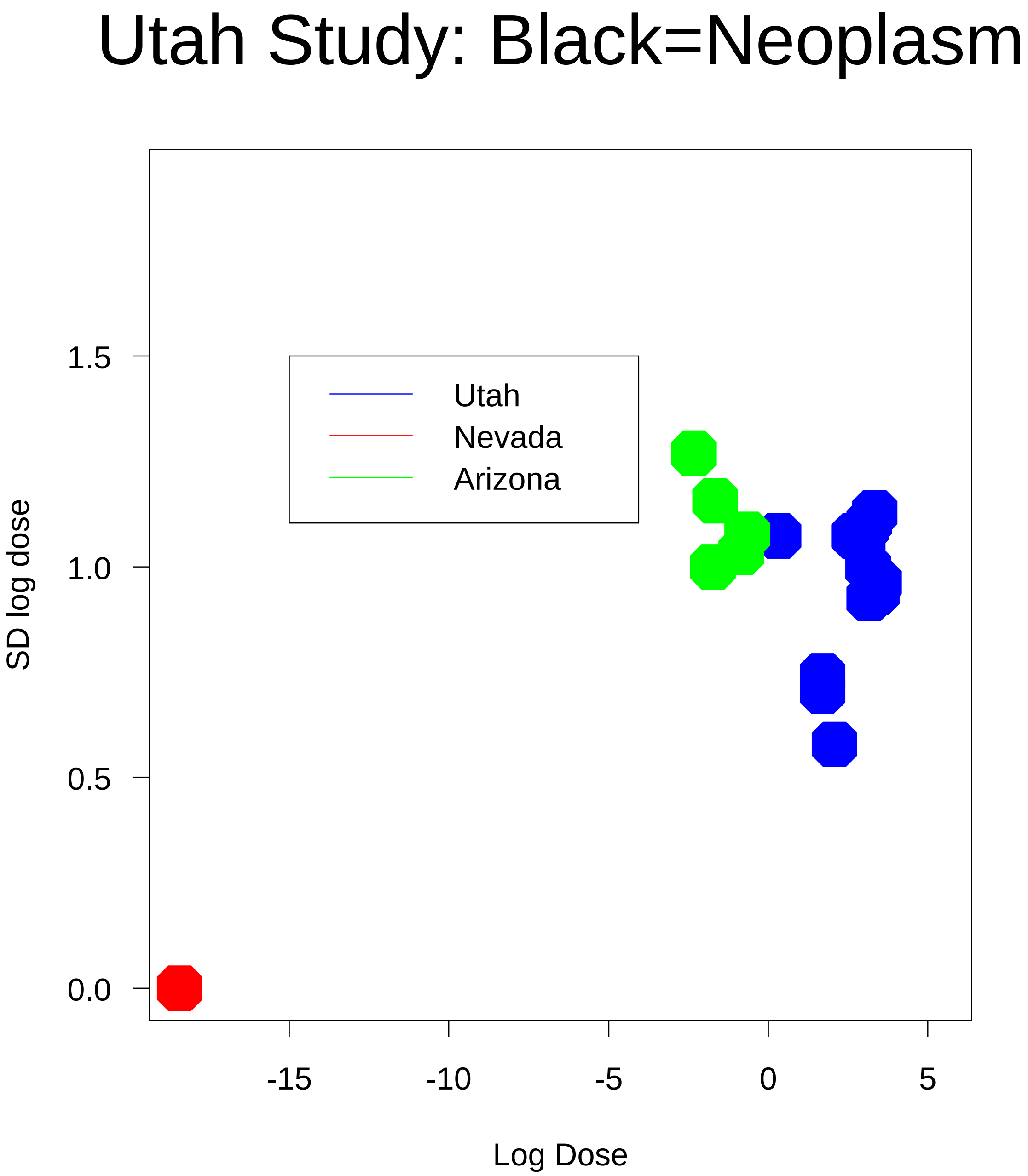

Figure 14: **Log(Dose) and estimated uncertainty in the Utah Data for the thyroid neoplasm cases, by state.**

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.
- 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

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

- 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

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

- 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 log-likelihood 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.
- **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

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

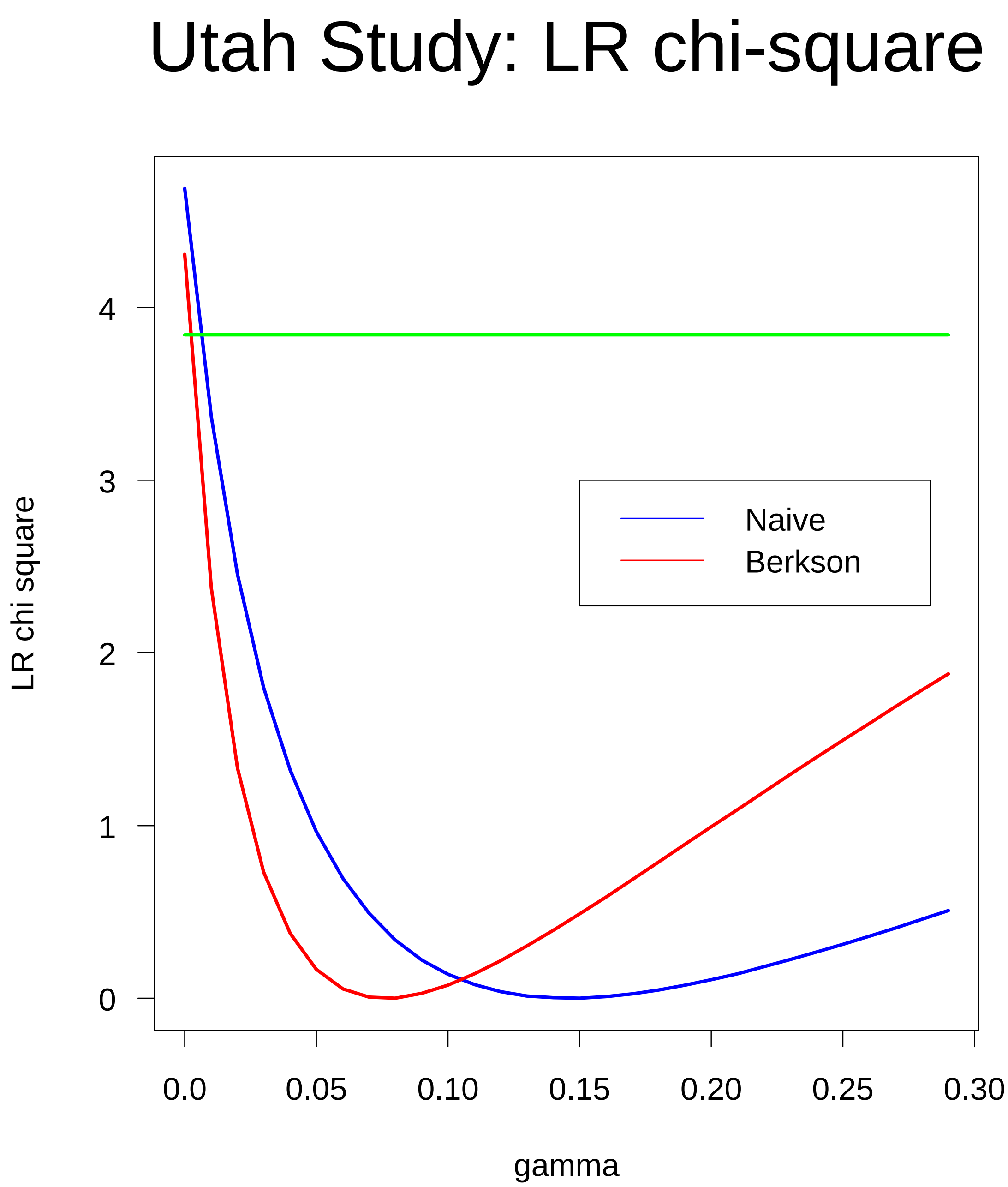

Figure 15: **Likelihood Ratio χ^2 tests for naive and Berkson analyses. Note that the dose effect is statistically significant for both, but that the estimate of γ is larger for the naive than for the Berkson analysis. Very strange.**

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.**
  - $*$ flexible parametric models, e.g., normal mixture models, buy robustness at the cost of complexity.
  - $*$ 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.

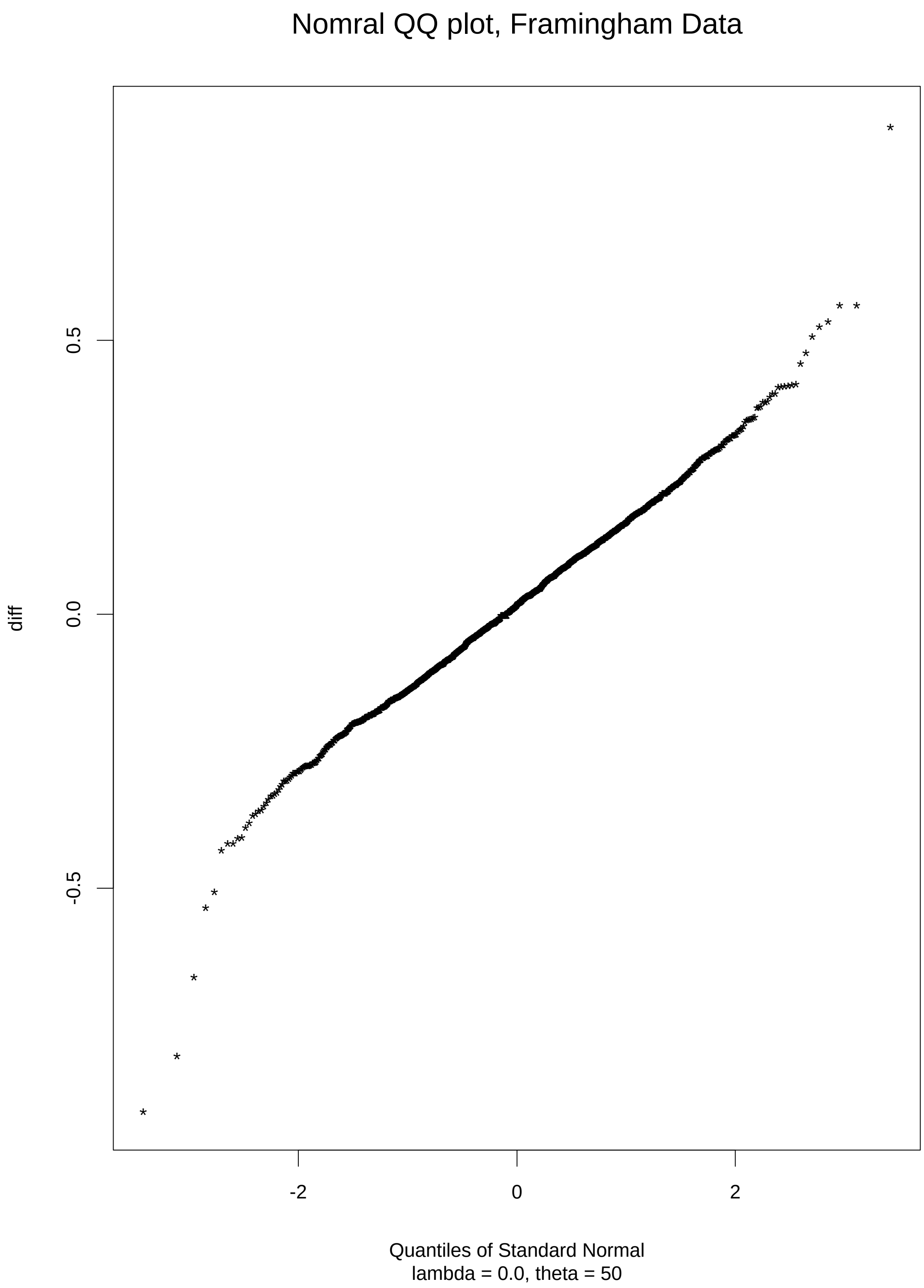

Figure 16: **q–q plot in Framingham for** $\log(\mathbf{SBP} - 50)$

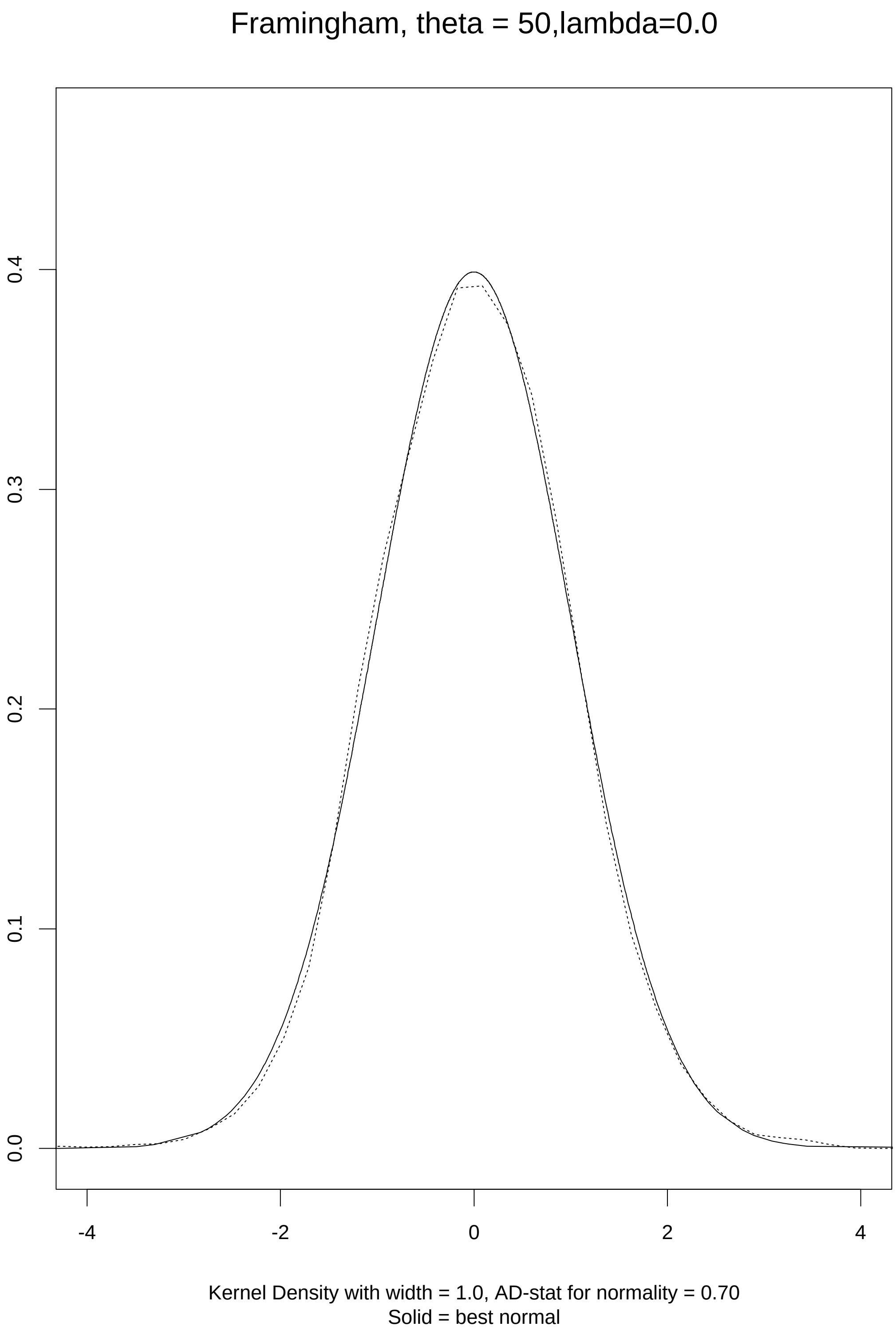

Figure 17: **Kernel density estimate and best fitting normal density plot in Framingham for $\log(\mathbf{SBP} - 50)$**

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

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

  * 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.

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

- We did this in the Framingham data, in **GAUSS**.
  - ∗ 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.

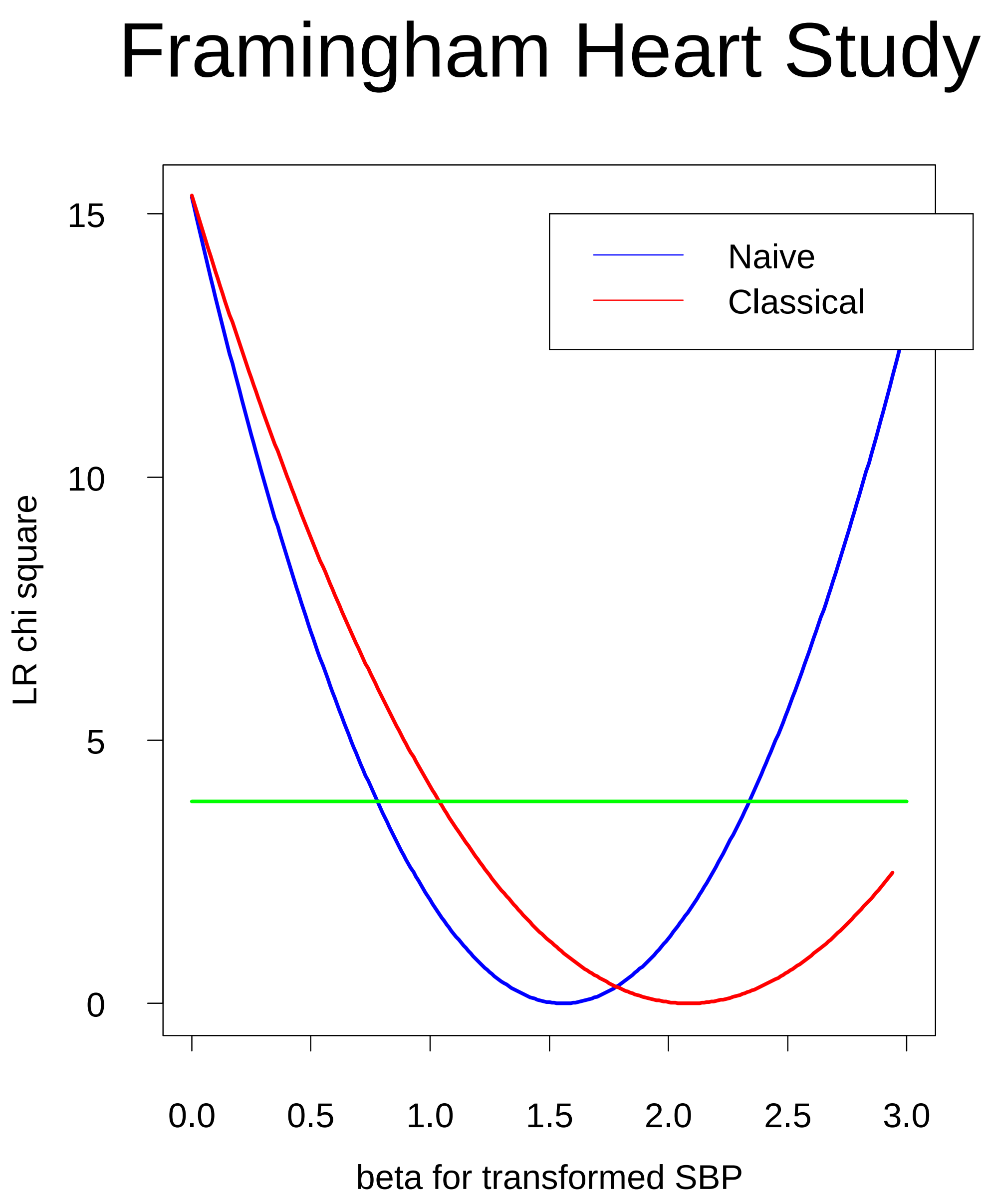

Figure 18: **Profile likelihoods for SBP in Framingham Heart Study.**

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 of less difficult for multiplicative error.

SEGMENT 5: BAYESIAN METHODS: OUTLINE

- **Main ideas**
- **It is not our intention to get into a huge debate about the merits of Bayes versus Frequentist.**
- **Both are useful.**
- **We use both.**

GENERAL BAYES

- In the **Bayesian** framework, there is a parameter, $\mathcal{B}$ and a likelihood function for the data $\underset{\sim}{\mathbf{D}}$ given the parameter, $f(\underset{\sim}{\mathbf{D}}|\mathcal{B})$
  - $*$ **Gibbs sampling** notation calls this $[\underset{\sim}{\mathbf{D}}|\mathcal{B}]$
- There is also a **prior distribution** $f(\mathcal{B})$
  - $*$ Gibbs sampling notation calls this $[\mathcal{B}]$
- The **posterior density** is given by

$$f(\mathcal{B}|\underset{\sim}{\mathbf{D}}) = \frac{f(\underset{\sim}{\mathbf{D}}|\mathcal{B})f(\mathcal{B})}{\int f(\underset{\sim}{\mathbf{D}}|\mathcal{B})f(\mathcal{B})d\mathcal{B}}$$

- Generally, $\mathcal{B}$ is **multidimensional and it is impossible to compute the denominator of the posterior**, and hence to compute posterior mean, posterior confidence intervals, etc.

GENERAL BAYES

- $$f(\mathcal{B}|\widetilde{\underline{\mathbf{D}}}) = \frac{f(\widetilde{\underline{\mathbf{D}}}|\mathcal{B})f(\mathcal{B})}{\int f(\widetilde{\underline{\mathbf{D}}}|\mathcal{B})f(\mathcal{B})d\mathcal{B}}$$

  $$[\mathcal{B}|\widetilde{\underline{\mathbf{D}}}] = \frac{[\widetilde{\underline{\mathbf{D}}}|\mathcal{B}][\mathcal{B}]}{\int [\widetilde{\underline{\mathbf{D}}}|\mathcal{B}][\mathcal{B}]d\mathcal{B}}$$

- There are a variety of means of computing the posterior, or at least approximating it.
- The best known is via **Gibbs sampling**, although **importance sampling** can sometimes be used.

GENERAL BAYES: IMPORTANCE SAMPLING

- In **importance sampling**, one developes an **importance density, $g(\mathcal{B})$**, which is easy to samply from.
- The **importance weights** are
$$w(\mathcal{B}) = \frac{[\widetilde{\underline{\mathbf{D}}}|\mathcal{B}][\mathcal{B}]}{g(\mathcal{B})}.$$
- If you want to estimate $E\left\{h(\mathcal{B})|\widetilde{\underline{\mathbf{D}}}\right\}$, the importance sampler says to take a sample of size B from $g(\cdot)$, and obtain an unbiased estimates as
$$\frac{B^{-1}\Sigma_{j=1}^{B} h(\mathcal{B}_j)w(\mathcal{B}_j)}{B^{-1}\Sigma_{j=1}^{B} w(\mathcal{B}_j)}$$
- This is an **unbiased** estimate.
  - $*$ The **key point** is that **you do not have to integrate anything**, in particular $\int[\widetilde{\underline{\mathbf{D}}}|\mathcal{B}][\mathcal{B}]d\mathcal{B}$.

GENERAL BAYES: IMPORTANCE SAMPLING

- The **major difficulty** with importance sampling is **to find the suitable importance density**.

$$\frac{B^{-1}\,\Sigma_{j=1}^{B}\,h(\mathcal{B}_j)w(\mathcal{B}_j)}{B^{-1}\,\Sigma_{j=1}^{B}\,w(\mathcal{B}_j)}$$

- Because the importance weights are

$$w(\mathcal{B}) = \frac{[D|\mathcal{B}][\mathcal{B}]}{g(\mathcal{B})}.$$

  the **importance density should be chosen so that the weights never explode** (otherwise you have a weighted average with nearly infinite weights sometimes).

- This can be **hard** to accomplish **in multiparameter problems**.

GENERAL BAYES: GIBBS SAMPLING

- **Gibbs sampling is an entire short course in itself!**
- Basically, the idea is to generate values from the parameter $\mathcal{B}$ one at a time.
- So, let's suppose that $\mathcal{B} = (\lambda, \gamma)$.
- The prior is typically an independence prior: $[\lambda][\gamma]$
- The likelihood is $[D|\lambda, \gamma]$
- The posterior is

$$[\lambda, \gamma|\underset{\sim}{\mathbf{D}}] = \frac{[D|\lambda, \gamma][\lambda][\gamma]}{\int [D|\lambda, \gamma][\lambda][\gamma] d\lambda d\gamma}$$

GENERAL BAYES: GIBBS SAMPLING

- **Why use Gibb?s** — To recap a previous slide, it might be hard to find a sensible importance density for the joint posterior $[\lambda, \gamma | \widetilde{\underline{\mathbf{D}}}]]$.
- Because they are **1–dimensional**, it may however be easier to generate observations from the **complete conditionals**:

$$[\lambda | \gamma, \widetilde{\underline{\mathbf{D}}}]; \qquad [\gamma | \lambda, \widetilde{\underline{\mathbf{D}}}].$$

- If you can generate observations from the complete conditionals, you end up with a sequence of random realizations (λ_j, γ_j) for $j = 1, ..., B$
- Geman[2] showed that if B is large enough, under fairly general conditions **you are generating observations from the real joint posterior** $[\lambda, \gamma | \widetilde{\underline{\mathbf{D}}}]$**!**
- We will discuss later how to generate such samples of random variables.

GENERAL BAYES: GIBBS SAMPLING

- The Gibbs sampling algorithm is thus:
  - ∗ Generate observations from the complete conditionals

$$[\lambda|\gamma, \underset{\sim}{\mathbf{D}}]$$
$$[\gamma|\lambda, \underset{\sim}{\mathbf{D}}].$$

  - ∗ Keep the last B of them, and call these (λ_j, γ_j) for $j = 1, ..., B$.
  - ∗ There are a sample from the posterior (approximately).
  - ∗ **They are a highly correlated sample.** Convergence to the stationary distribution may take essentially forever.

GENERAL BAYES: GIBBS SAMPLING

- Armed with the sample, (λ_j, γ_j) for $j = 1, ..., B$, **you can construct summary statistics:**
  - $*$ Sample mean, median, mode, variance, etc.
  - $*$ Sample histogram or kernel density estimate.
- These are all consistent estimates of the posterior quantity: mean, median, mode, variance, density.
- **You can construct credible intervals** (posterior confidence intervals) by the sample percentiles.
- **You can compute posterior probabilities** as well. Suppose you want $\mathrm{pr}(\lambda > 0|\underline{\widetilde{\mathbf{D}}}]$. Then estimate it by the sample percentage of times the generated $\lambda > 0$.

GIBBS SAMPLING: CLASSICAL ERROR

- $[\underline{\widetilde{\mathbf{Y}}}|\underline{\widetilde{\mathbf{X}}}, \underline{\widetilde{\mathbf{Z}}}, \beta]$ is the **likelihood function for the data if the X's were observed.**
- $[\underline{\widetilde{\mathbf{W}}}|\underline{\widetilde{\mathbf{X}}}, \underline{\widetilde{\mathbf{Z}}}, \mathcal{U}]$ is the **likelihood function for the error model if the X's were observed.**
- $[\underline{\widetilde{\mathbf{X}}}|\underline{\widetilde{\mathbf{Z}}}, \mathcal{A}]$ is the **likelihood function for the latent, unobserved X's.**
- **If we observe the X's**, the Gibbs sampling algorithm is
  - $*$ Generate β from complete conditionals for $[\underline{\widetilde{\mathbf{Y}}}|\underline{\widetilde{\mathbf{X}}}, \underline{\widetilde{\mathbf{Z}}}, \beta]$.
  - $*$ Generate $\mathcal{U}$ from complete conditionals for $[\underline{\widetilde{\mathbf{W}}}|\underline{\widetilde{\mathbf{X}}}, \underline{\widetilde{\mathbf{Z}}}, \mathcal{U}]$
  - $*$ Generate $\mathcal{A}$ from complete conditionals for $[\underline{\widetilde{\mathbf{X}}}|\underline{\widetilde{\mathbf{Z}}}, \mathcal{A}]$
- Simple!

GIBBS SAMPLING: CLASSICAL ERROR

- The previous algorithm works if the X's $\widetilde{\underline{\mathbf{X}}}$ were observed.
- The whole point is that they are not.
- **The only change that is necessary is to generate the $\widetilde{\underline{\mathbf{X}}}$ from the appropriate distribution.**
- Since the likelihood is

$$[\widetilde{\underline{\mathbf{Y}}}|\widetilde{\underline{\mathbf{X}}},\widetilde{\underline{\mathbf{Z}}},\beta][\widetilde{\underline{\mathbf{W}}}|\widetilde{\underline{\mathbf{X}}},\widetilde{\underline{\mathbf{Z}}},\mathcal{U}][\widetilde{\underline{\mathbf{X}}}|\widetilde{\underline{\mathbf{Z}}},\mathcal{A}]$$

- You need to generate the $\widetilde{\underline{\mathbf{X}}}$ from the **complete conditional** distribution

$$[\widetilde{\underline{\mathbf{X}}}|\widetilde{\underline{\mathbf{Y}}},\widetilde{\underline{\mathbf{W}}},\widetilde{\underline{\mathbf{Z}}},\beta,\mathcal{U},\mathcal{A}] = \frac{[\widetilde{\underline{\mathbf{Y}}}|\widetilde{\underline{\mathbf{X}}},\widetilde{\underline{\mathbf{Z}}},\beta][\widetilde{\underline{\mathbf{W}}}|\widetilde{\underline{\mathbf{X}}},\widetilde{\underline{\mathbf{Z}}},\mathcal{U}][\widetilde{\underline{\mathbf{X}}}|\widetilde{\underline{\mathbf{Z}}},\mathcal{A}]}{\int[\widetilde{\underline{\mathbf{Y}}}|\widetilde{\underline{\mathbf{X}}},\widetilde{\underline{\mathbf{Z}}},\beta][\widetilde{\underline{\mathbf{W}}}|\widetilde{\underline{\mathbf{X}}},\widetilde{\underline{\mathbf{Z}}},\mathcal{U}][\widetilde{\underline{\mathbf{X}}}|\widetilde{\underline{\mathbf{Z}}},\mathcal{A}]d\widetilde{\underline{\mathbf{X}}}}$$

- Generally, for independent data, you do this one X at a time.

FRAMINGHAM HEART STUDY DATA

- We will let age and smoking status be denoted by Z.
- A **reasonable basic 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

- The unknown parameters are β_0, β_z, β_x, μ_x, σ_x^2.
- We first standardized W, Z so that they had sample mean zero and sample variance one. This makes the intercept largely independent of the other values.
- We gave β_0, β_z, β_x, μ_x normal prior distributions with mean zero and variance $1,000$, which is essentially a noninformative prior.
- We gave $\tau_x^2 = 1/\sigma_x^2$ a Gamma prior with mean 1.0 and variance $1,000$. This is an essentially non–informative inverse–Gamma prior for σ_x^2.

FRAMINGHAM HEART STUDY

- We used the program **BUGS** for this problem.
  - ∗ **BUGS** is freely available on the web.
  - ∗ It has unix and Windows versions.
  - ∗ It is based upon adaptive rejection sampling (basically, the candidate density adapts to the problem over time).
  - ∗ **BUGS** can be very slow, because it generates $1,615$ values of the X's each time through.
  - ∗ **BUGS** is not commercial, and tends to be limited in the kinds of problems it can handle.
  - ∗ In MEM's, this means essentially GLIM's, normal X's, normal measurement errors.

FRAMINGHAM HEART STUDY

- We used a "burn–in" period of $4,000$ times through the algorithm, and we then set $B = 10,000$.
- The starting values were from a likelihood analysis.
- We standardized all the values.
- In the naive analysis, we ignored measurement error.
- Here is the output for the run, in terms of the parameter β_x.
- All results were close to the likelihood analyses.

FRAMINGHAM HEART STUDY

NAIVE ANALYSIS

VARIABLE	Mean	SD
========	====	==
beta0	-2.660000	0.108000
betaa	0.470000	0.102000
betas	0.250000	0.104000
betax	0.353000	0.092100
mux	-0.000336	0.024700
taux	1.000000	0.035300

FRAMINGHAM HEART STUDY

MEASUREMENT ERROR ANALYSIS

VARIABLE	Mean	SD
beta0	-2.680000	0.112000
betaa	0.473000	0.098900
betas	0.254000	0.106000
betax	0.475000	0.125000
lambda	0.753000	0.008770
mux	0.000432	0.025100
tauw	1.000000	0.035500
taux	1.330000	0.062700

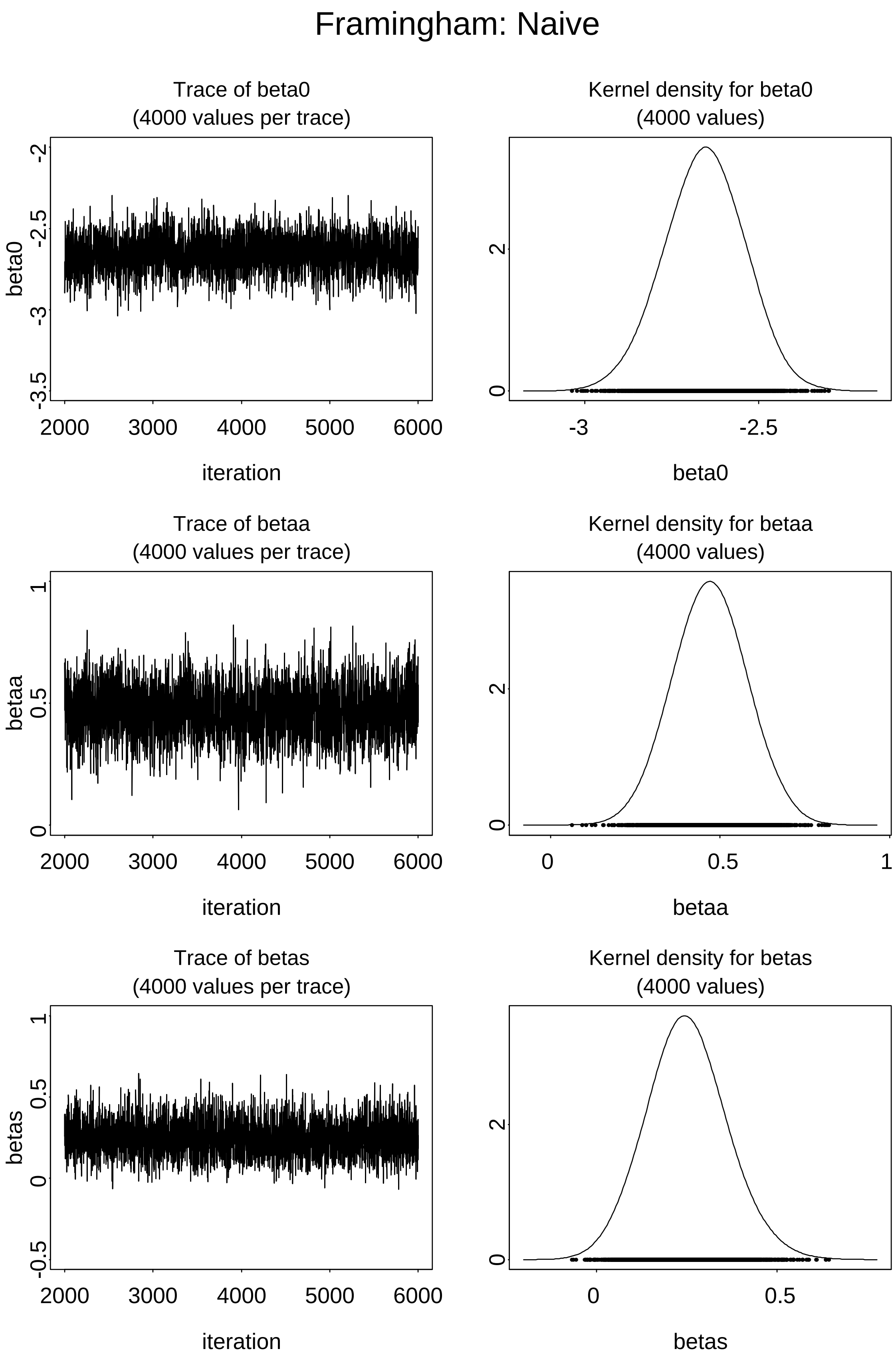

Figure 19: **Bayesian analysis of Framingham data ignoring measurement error.**

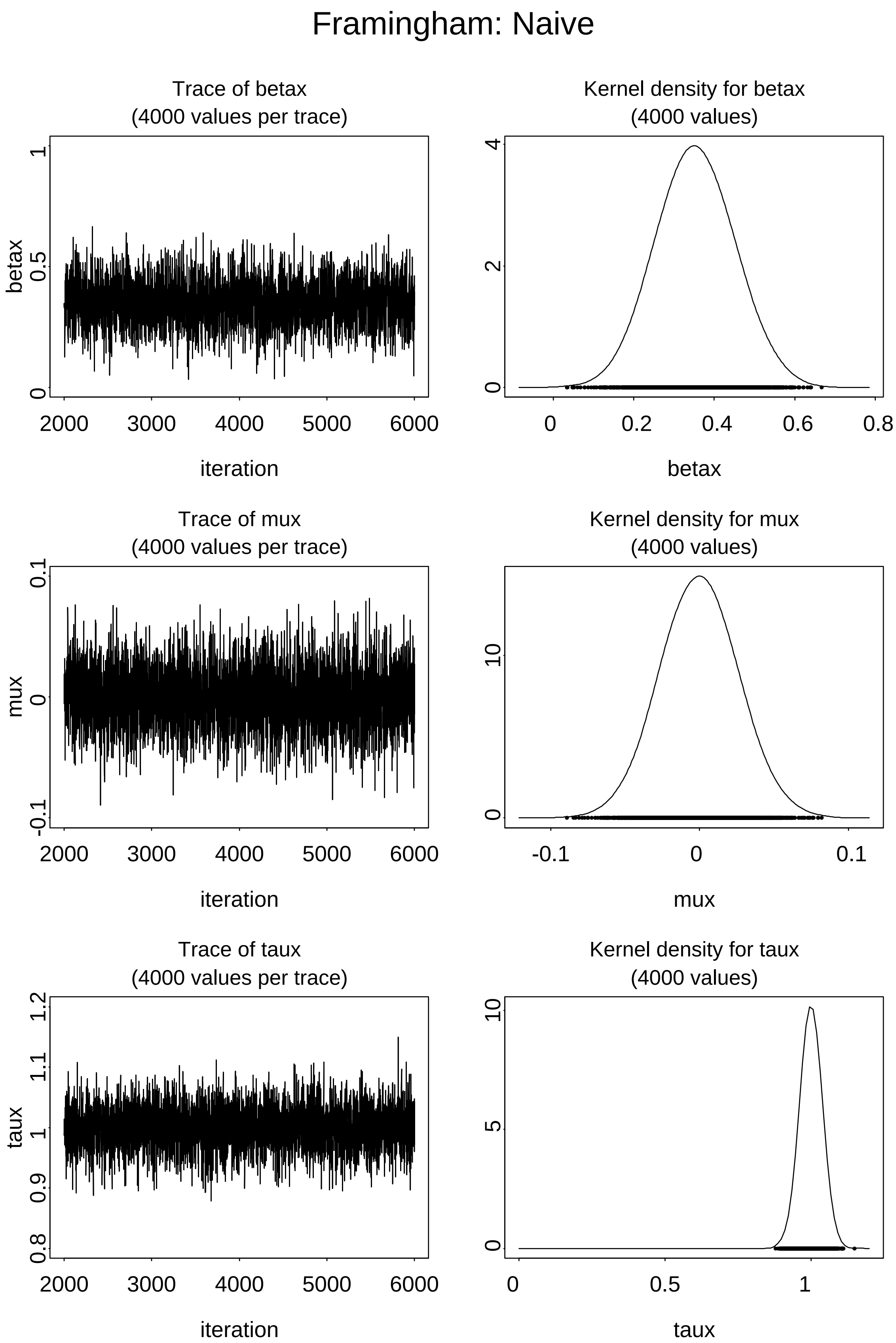

Figure 20: **Bayesian analysis of Framingham data ignoring measurement error.**

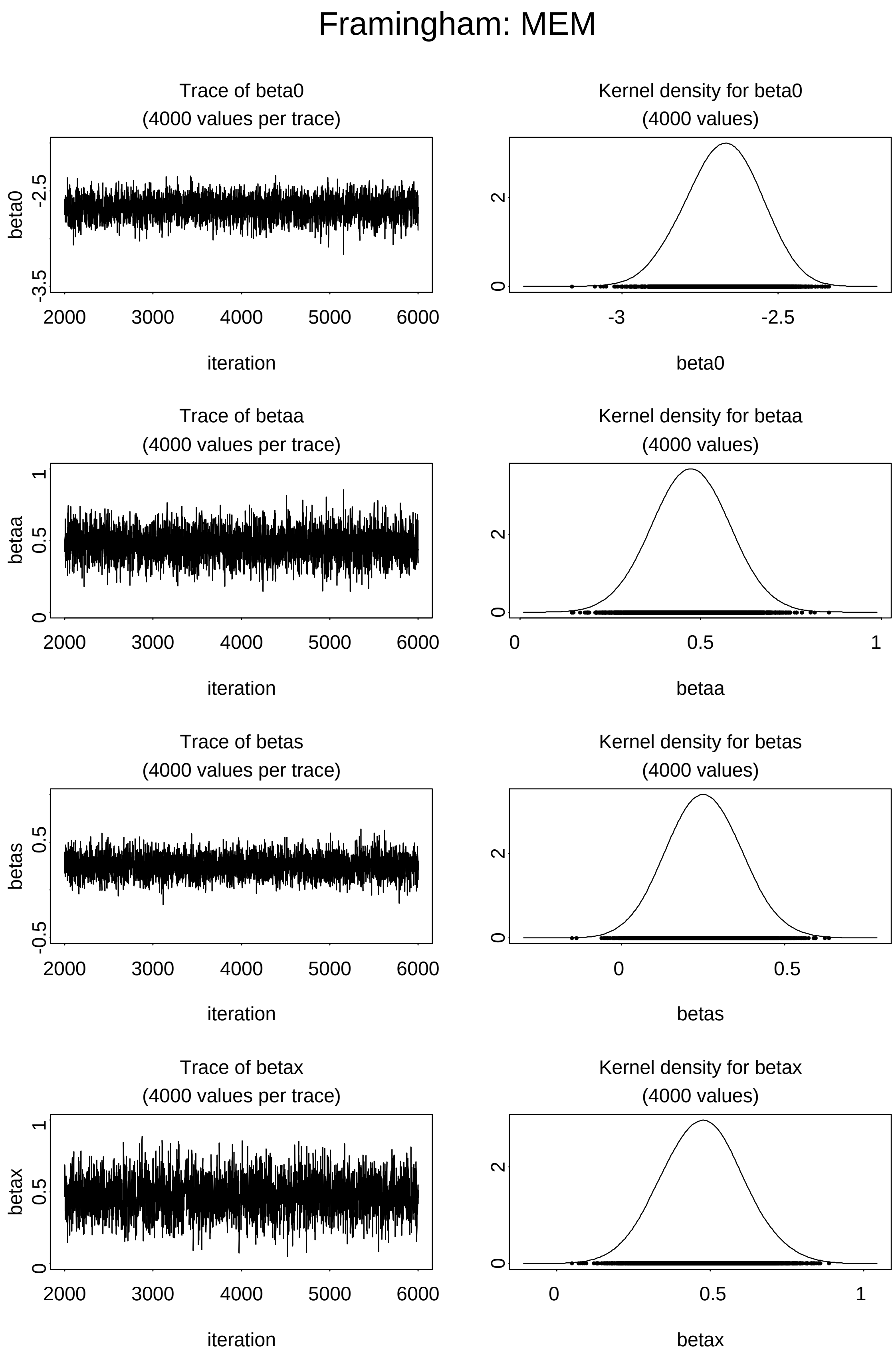

Figure 21: **Bayesian analysis of Framingham data accounting for measurement error.**

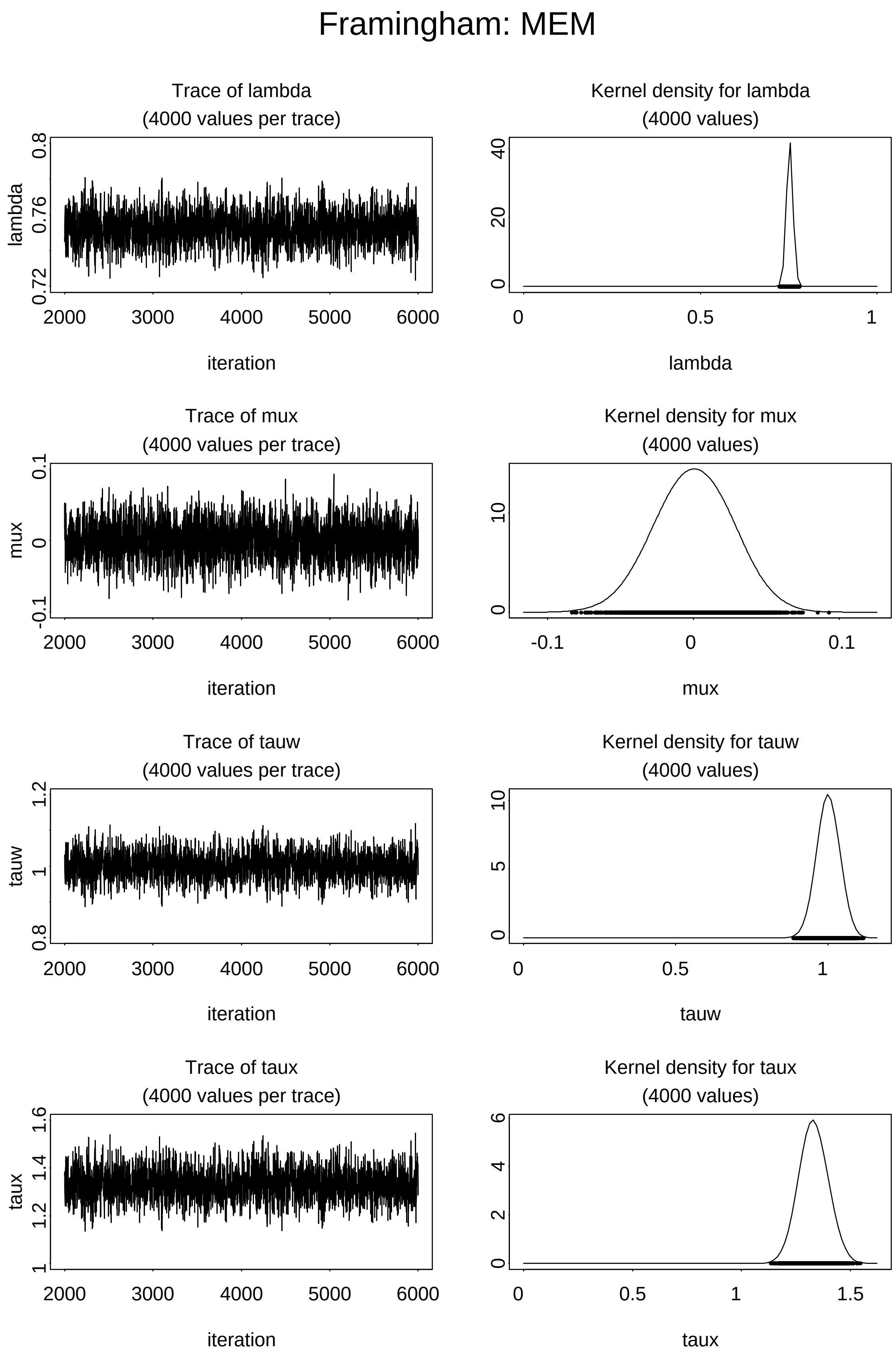

Figure 22: **Bayesian analysis of Framingham data accounting for measurement error.**

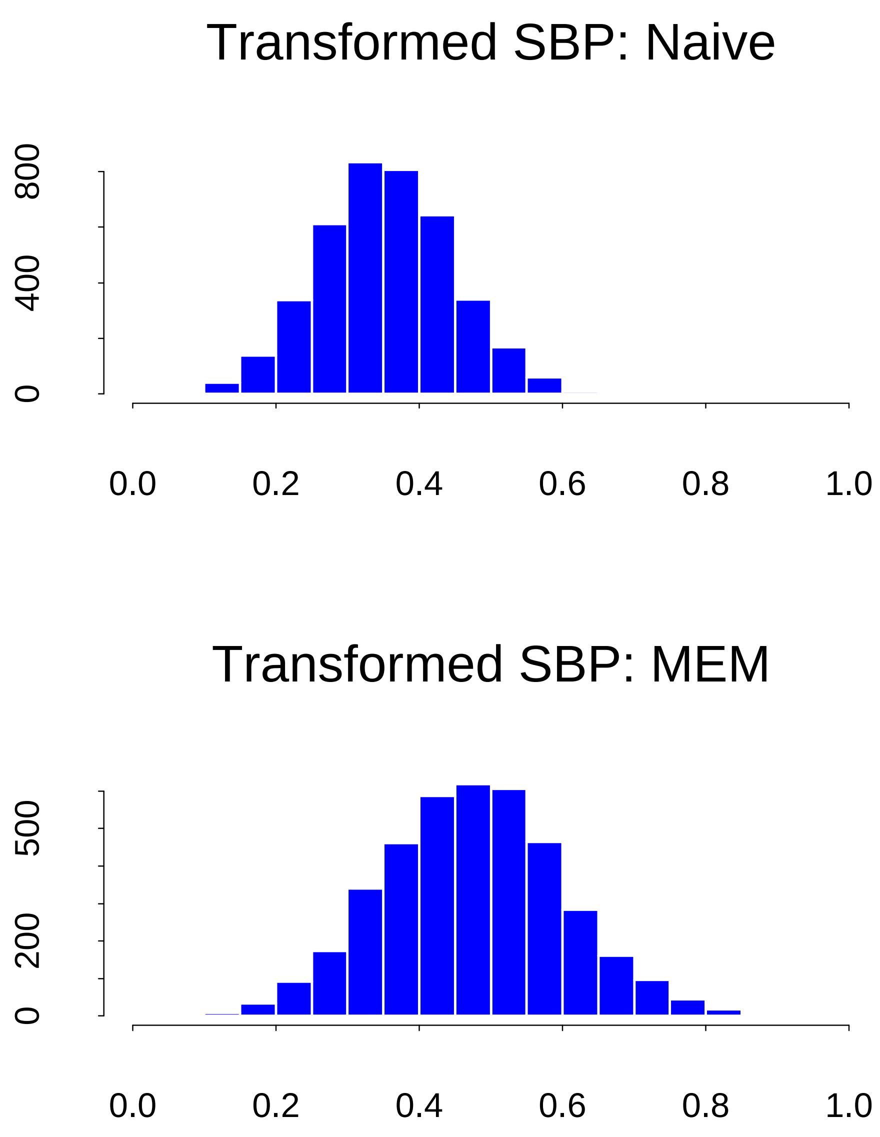

Figure 23: **Bayesian analysis of Framingham data ignoring and accounting for measurement error: transformed Systolic Blood Pressure.**

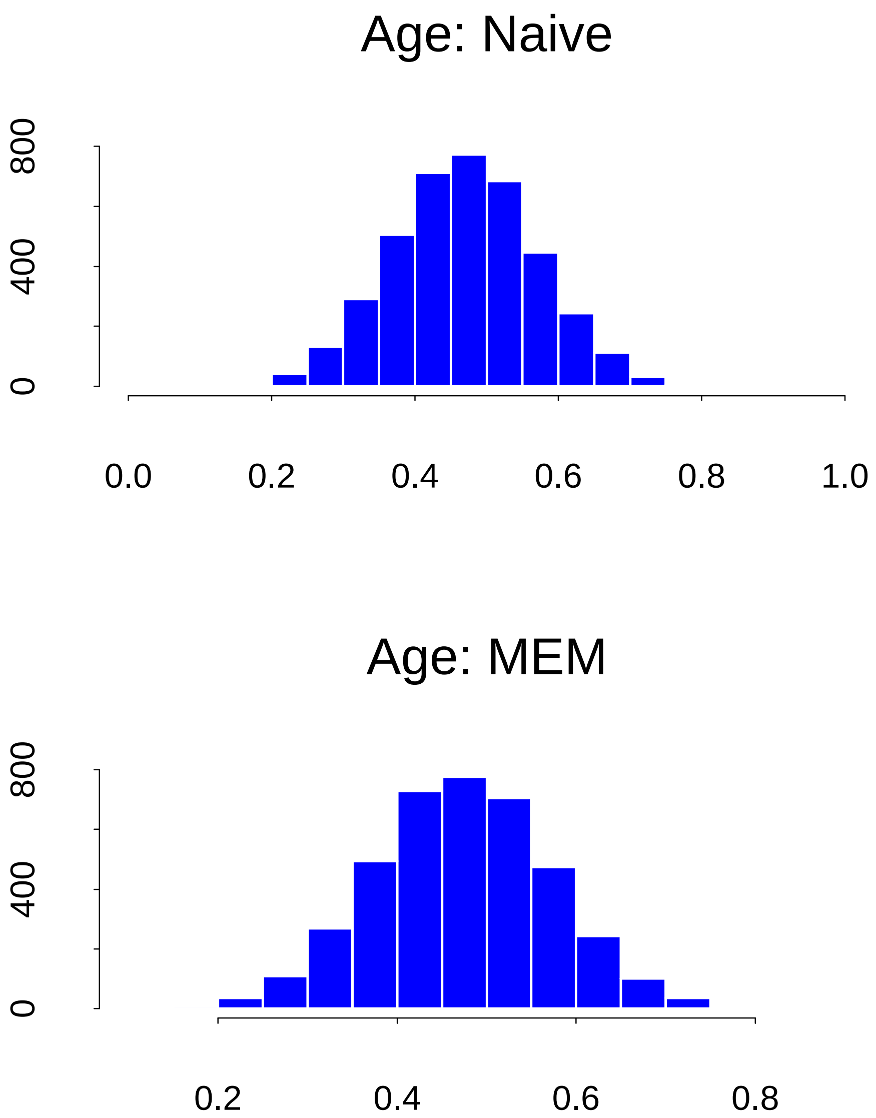

Figure 24: **Bayesian analysis of Framingham data ignoring and accounting for measurement error: Age**

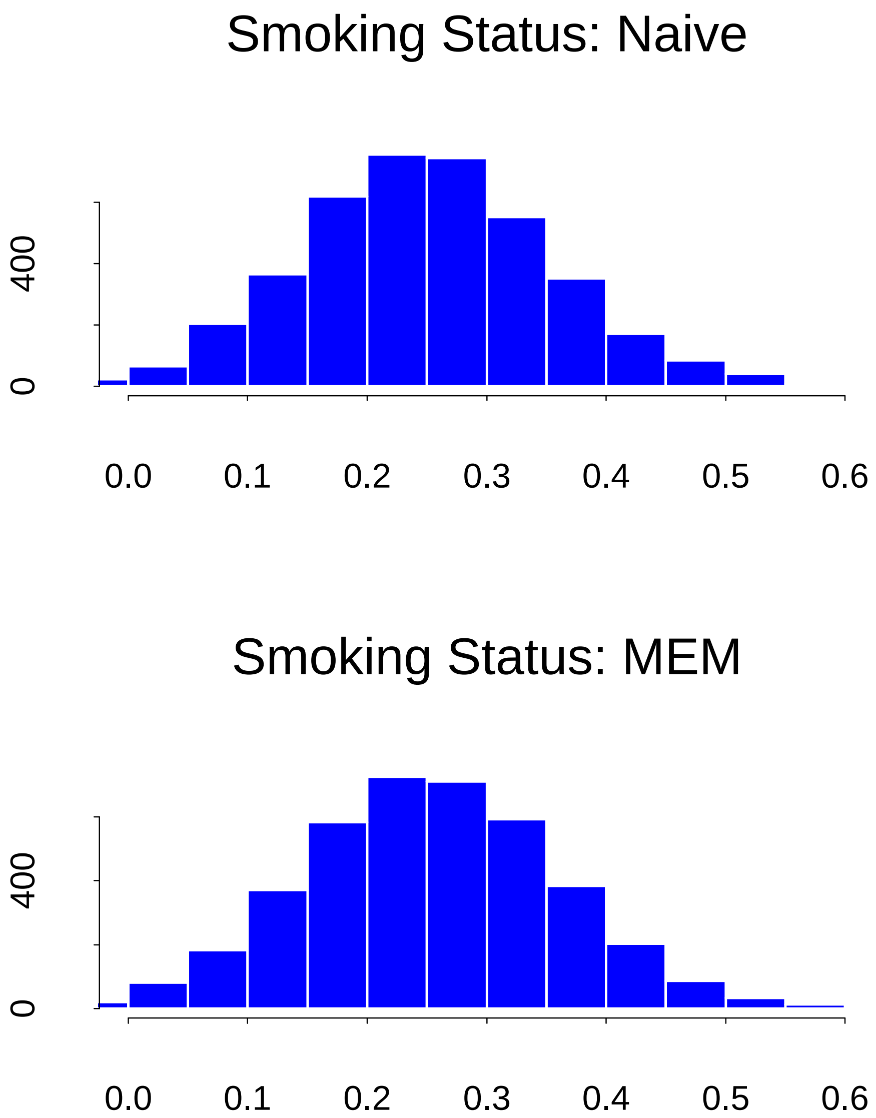

Figure 25: **Bayesian analysis of Framingham data ignoring and accounting for measurement error: Smoking Status.**

SEGMENT 6 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.
- For some models, if var($X|Z,W$) is constant (even a large constant) then the only **bias** due to the regression calibration approximation is in the **intercept** parameter.

THE REGRESSION CALIBRATION ALGORITHM

- The general algorithm is:

  * Using replication, validation, or instrumental data, develop a model $E(X|Z,W) = m(Z,W,\gamma_{cm})$ and estimate γ_{cm}.

  * **Replace** X by $m(Z,W,\widehat{\gamma}_{cm})$ 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."**

LOGISTIC REGRESSION, NORMAL X

- Consider the logistic regression model
  $\text{pr}(Y = 1|X) = \{1 + \exp(-\beta_0 - \beta_x X)\}^{-1} = H(\beta_0 + \beta_x X).$
- Suppose that
  - X and U are normally distributed.
  - Then X given W is normal with mean $E(X|W) = \mu_x(1-\lambda) + \lambda W$ and variance $\lambda\sigma_u^2$.
- If the event is "rare", then $H(\beta_0 + \beta_x X) \approx \exp(\beta_0 + \beta_x X)$.
- Then, by using moment generating functions, the observed data follow

$$
\begin{aligned}
\text{pr}(Y = 1|W) &= E\left[\{\text{pr}(Y = 1|X, W)\}\,|W\right] \\
&= E\left[\{\text{pr}(Y = 1|X)\}\,|W\right] \text{ why??} \\
&\approx E\left\{\exp(\beta_0 + \beta_x X)|W\right\} \\
&= \exp\{\beta_0 + (1/2)\beta_x^2\lambda\sigma_u^2 + \beta_x E(X|W)\} \\
&\approx H\{\beta_0 + (1/2)\beta_x^2\lambda\sigma_u^2 + \beta_x E(X|W)\}.
\end{aligned}
$$

- **The regression calibration approximation works fine in this case.**

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 X given W is normal with mean $E(X|W) = \mu_x(1-\lambda) + \lambda W$ and variance $\lambda\sigma_u^2$.
- Then it is easily shown that **the observed data are not Poisson**, but
  * The mean is loglinear:

  $$E(Y|W) = \exp\{\beta_0 + (1/2)\beta_x^2\lambda\sigma_u^2 + \beta_x E(X|W)\}.$$

  * **The Poisson data 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).
- $Y = I\{$breast cancer$\}$.
- $Z =$ (age, poverty index ratio, body mass index, $I\{$use alcohol$\}$, $I\{$family history of breast cancer$\}$, $I\{$age at menarche $\leq 12\}$, $I\{$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: $W = \log(5+\text{measured saturated fat})$.

NHANES-I—CONTINUED

- w/o adjustment for Z, W appears to have a small **protective** effect, since cases have a smaller value of $\overline{W}$ compared to non-cases.
- Naive logistic regression of Y on (Z, W):
  - $\widehat{\beta}_W = -.97$, 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.

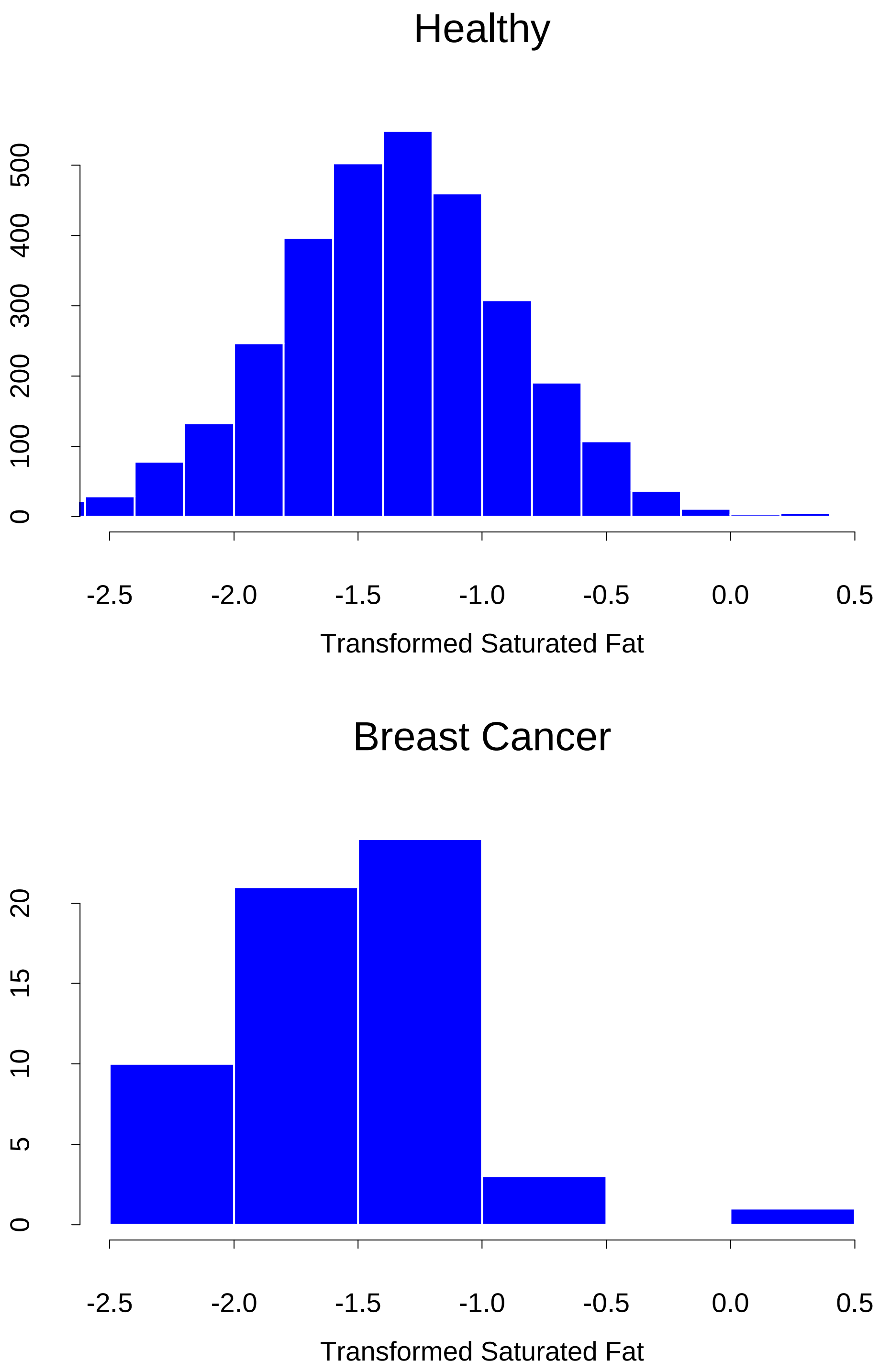

Figure 26: **Histograms of $\log(.05 + \textbf{Saturated Fat}/100)$ in the NHANES data, for women with and without breast cancer in 10 year follow–up..**

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:

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

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

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.
  * of course, 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:
  - $W_{ij} = X_i + U_{ij}$, $i = 1, \ldots, n$ 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

- $E(X|Z,\overline{W})$

$$\approx \mu_x + (\Sigma_{xx}\ \Sigma_{xz}^t) \begin{Bmatrix} \Sigma_{xx} + \Sigma_{uu}/k & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{Bmatrix}^{-1} \begin{pmatrix} \overline{W} - \mu_w \\ Z - \mu_z \end{pmatrix}. \quad (1)$$

(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.
    - $*$ $\widehat{\mu}_x = \widehat{\mu}_w = \Sigma_{i=1}^n k_i \overline{W}_{i\cdot} / \Sigma_{i=1}^n k_i$.
    - $*$ $\widehat{\Sigma}_{xz} = \Sigma_{i=1}^n k_i (\overline{W}_{i\cdot} - \widehat{\mu}_w)(Z_i - \widehat{\mu}_z)^t / \nu$

      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 7, REMEASUREMENT METHODS: SIMULATION EXTRAPOLATION, OUTLINE

- **About Simulation Extrapolation**
- **The Key Idea**
- **An Empirical Version**
- **Simulation Extrapolation Algorithm**
- **Example: Measurement Error in SBP**
- **Summary**

ABOUT SIMULATION EXTRAPOLATION

- For **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
- For bias reduction and variance estimation
  - ∗ like bootstrap and jackknife
- **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
  - λ controls amount of added measurement error
  - σ_u^2 increased to $(1+\lambda)\sigma_u^2$
- **Recalculate estimates** — called pseudo estimates
- **Plot** pseudo estimates versus λ
- **Extrapolate** to $\lambda = -1$
  - $\lambda = -1$ corresponds to case of no measurement error

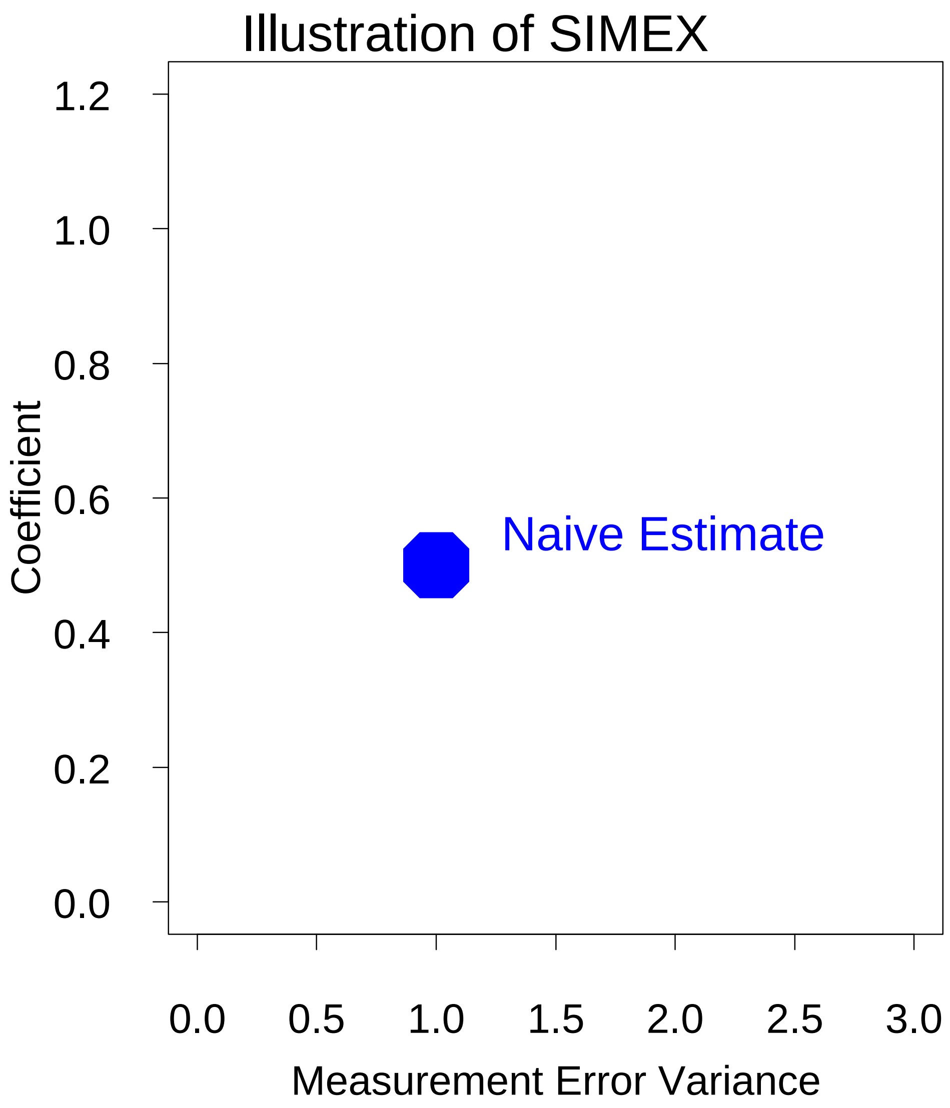

Figure 27: **Your estimate when you ignore measurement error.**

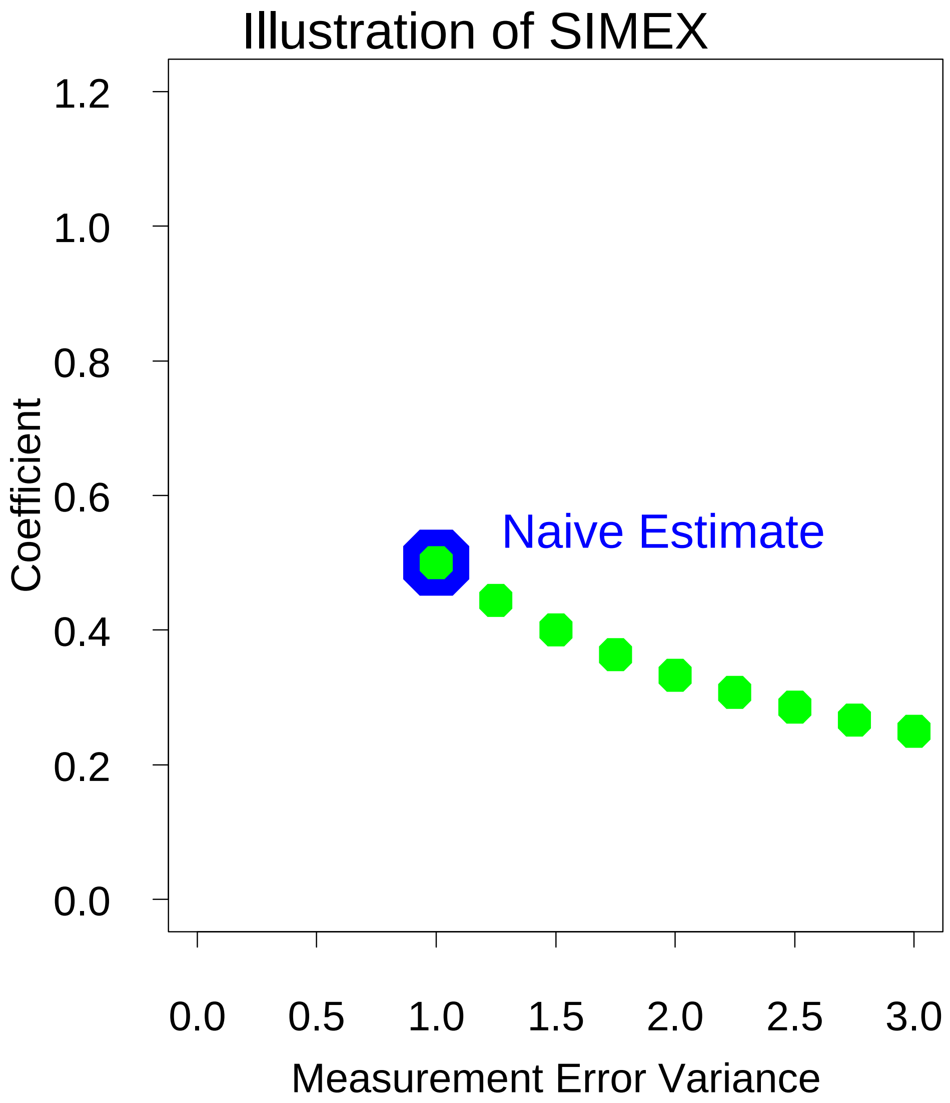

Figure 28: **What heppens to your estimate when you have more error, which you add on by simulation, but you still ignore the error.**

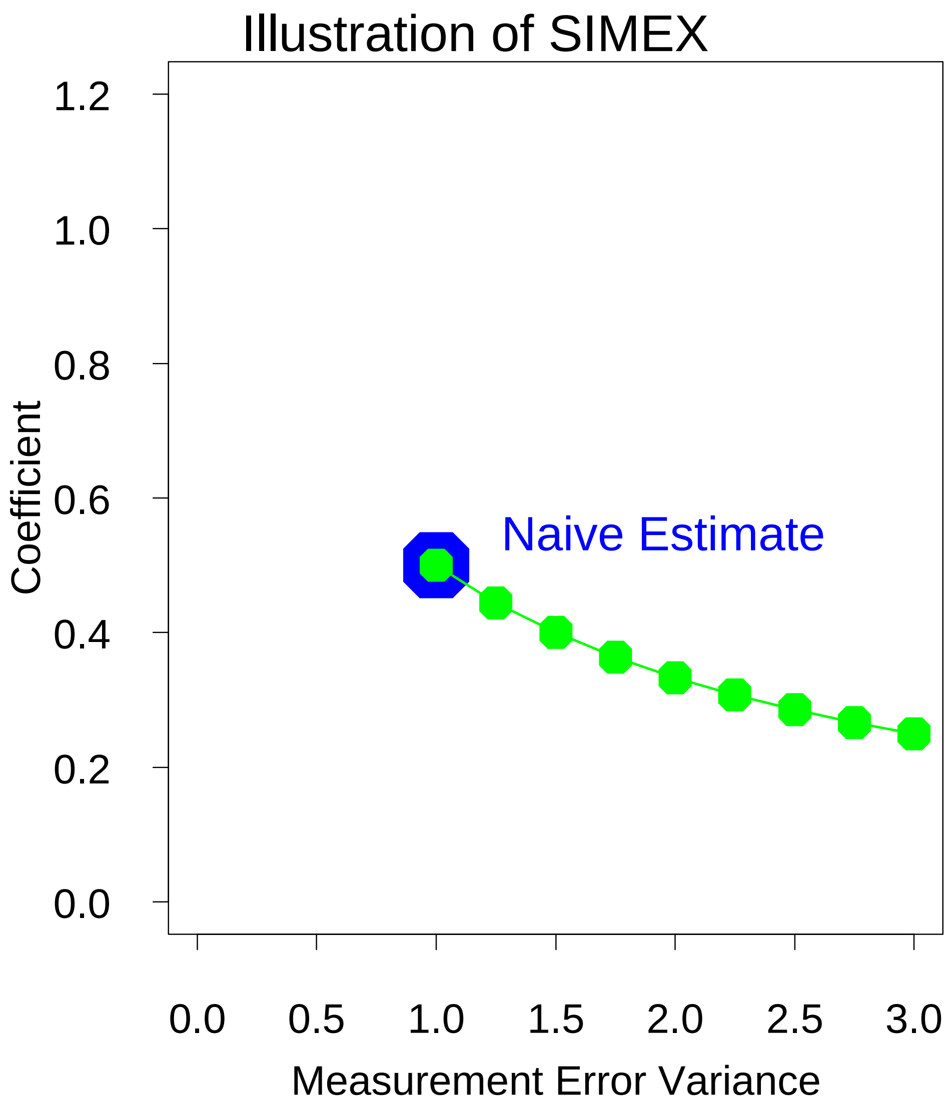

Figure 29: **What statistician can resist fitting a curve?.**

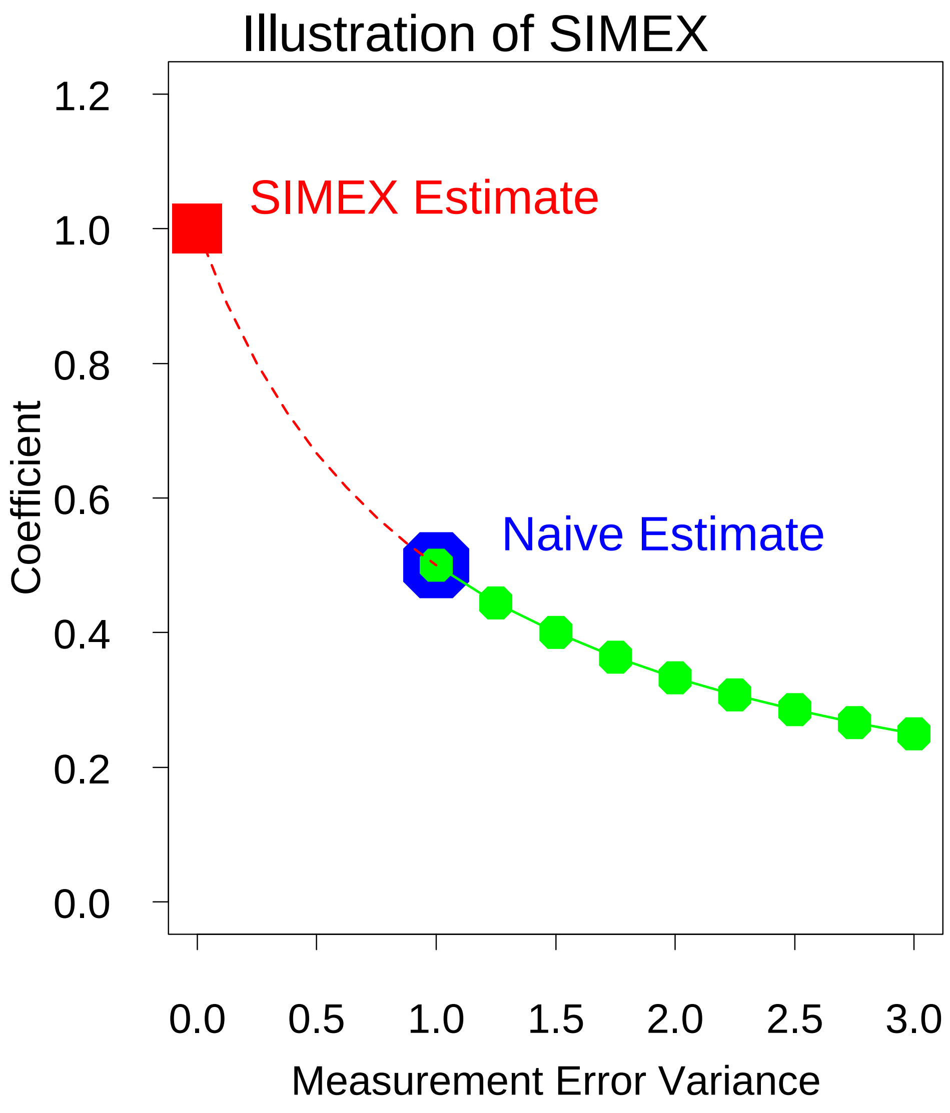

Figure 30: **Now extrapolate to the case of no measurement error.**

OUTLINE OF THE ALGORITHM

- **Add** measurement error to variable measured with error
  - λ controls amount of added measurement error
  - σ_u^2 increased to $(1+\lambda)\sigma_u^2$
- **Recalculate** estimates — called pseudo estimates. **Do many times and average for each λ**
- **Plot** pseudo estimates versus λ
- **Extrapolate** to $\lambda = -1$
  - $\lambda = -1$ 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:

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

- 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

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

FRAMINGHAM DATA EXAMPLE: THREE NAIVE ANALYSES:

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

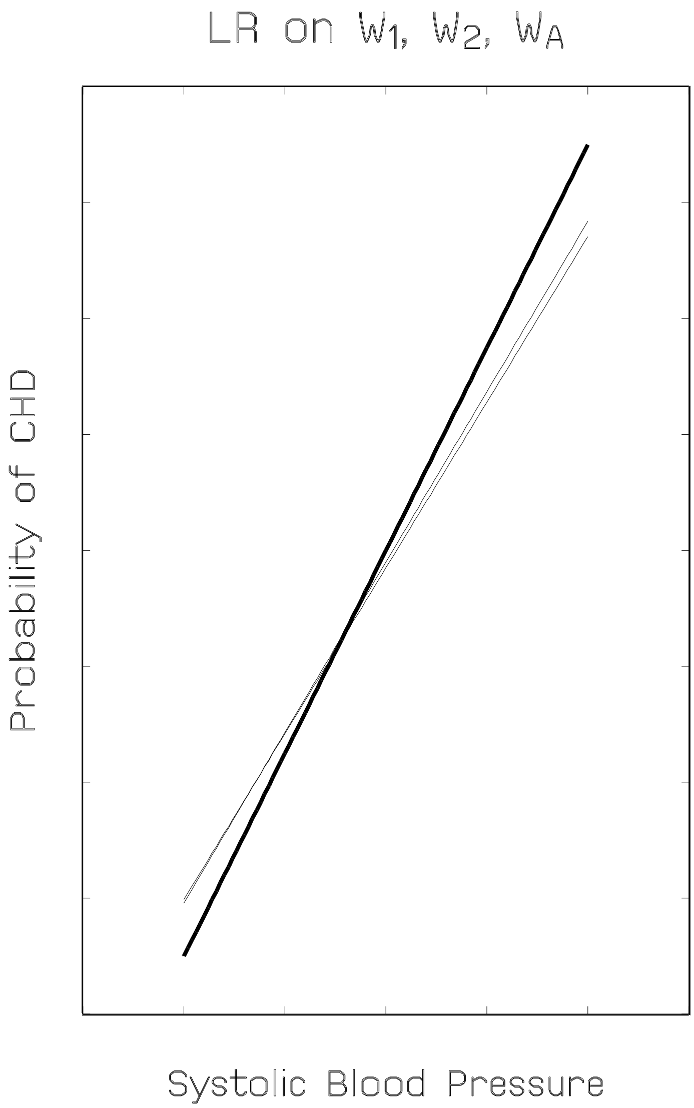

Figure 31: **Logistic regresion fits in Framingham using first replicate, second replicate and average of both**

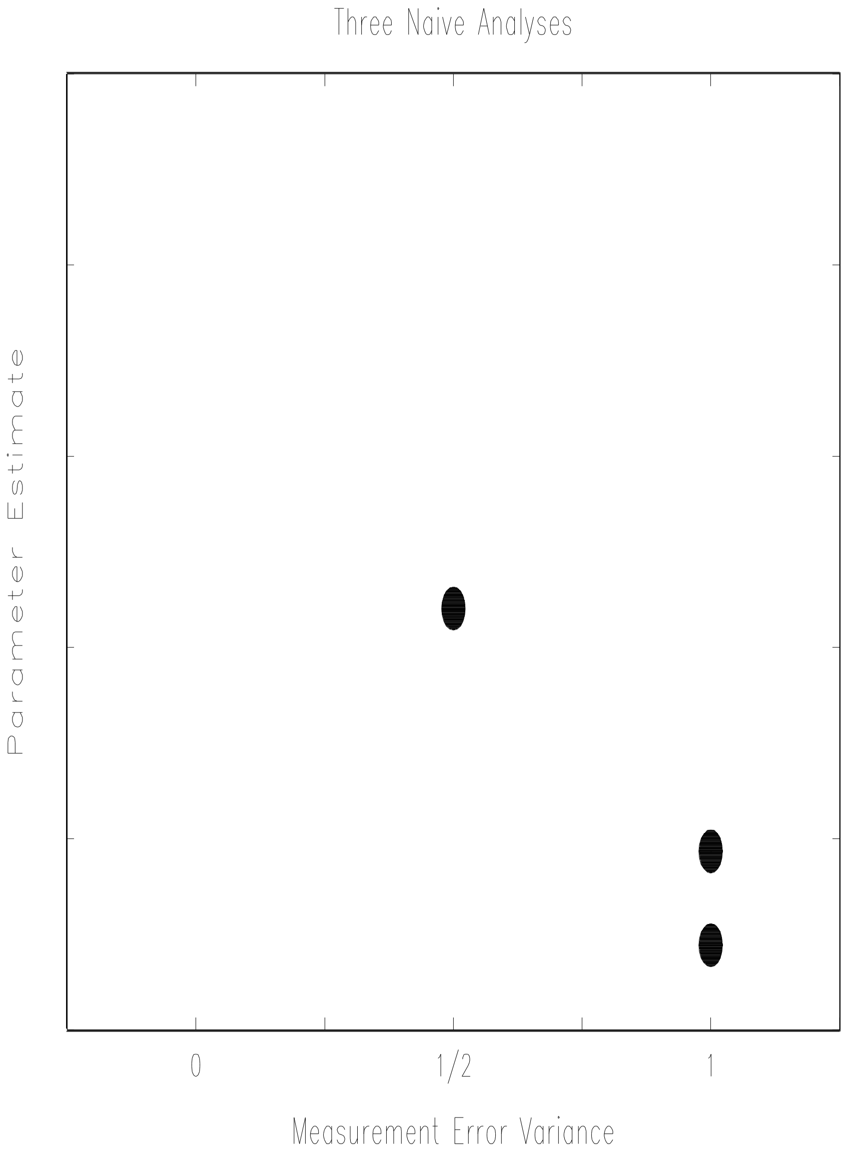

Figure 32: **A SIMEX–type plot for the Framingham data, where the errors are not computer–generated.**

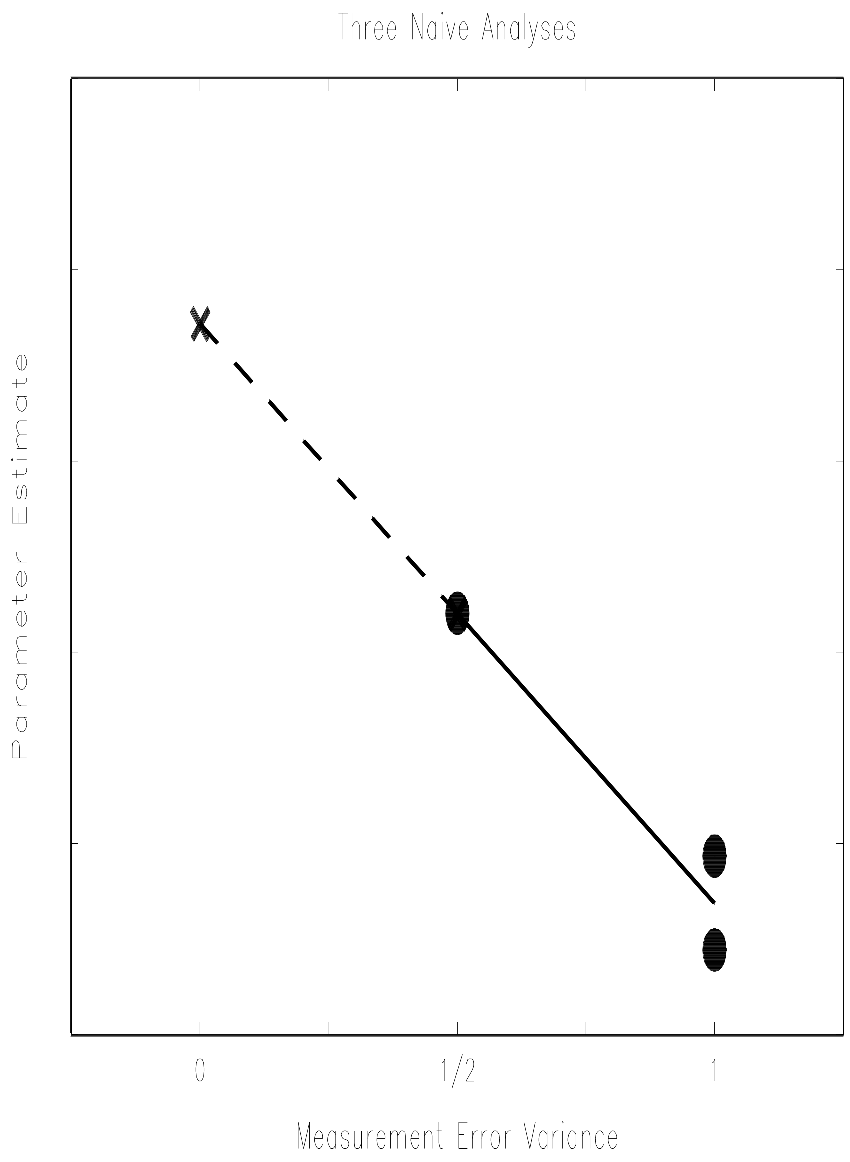

Figure 33: **A SIMEX–type extrapolation for the Framingham data, where the errors are not computer–generated.**

SIMULATION AND EXTRAPOLATION STEPS: SIMULATION

- Adding Measurement Error
- Framingham Example:
  - $\overline{W}_{\bullet} \longmapsto W_1 = \overline{W}_{\bullet} + (W_1 - W_2)/2$
  - $\overline{W}_{\bullet} \longmapsto W_2 = \overline{W}_{\bullet} + (W_2 - W_1)/2$
- In General:
  - Best Estimator $\longmapsto$ Best Estimator +

    $\sqrt{\lambda}$ {Independent N$(0,\ \sigma_u^2)$ Error }
  - λ controls amount (variance) of added measurement error

SIMULATION AND EXTRAPOLATION STEPS: EXTRAPOLATION

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

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

SIMULATION AND EXTRAPOLATION ALGORITHM

- Simulation Step
- For $\lambda \in \{\lambda_1, \ldots, \lambda_M\}$
- For $b = 1\ (1)\ B$, compute:
  - $*$ bth pseudo data set

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

  - $*$ bth pseudo estimate

$$\widehat{\theta}_b(\lambda) = \widehat{\theta}\left(\{Y_i, W_{b,i}(\lambda)\}_1^n\right)$$

  - $*$ the average of the pseudo estimates

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

SIMULATION AND EXTRAPOLATION ALGORITHM

- Extrapolation Step
- Plot $\widehat{\theta}(\lambda)$ vs λ $(\lambda > 0)$
- Extrapolate to $\lambda = -1$ to get $\widehat{\theta}(-1) = \widehat{\theta}_{\text{SIMEX}}$
- For certain models and assuming the extrapolant function is chosen correctly
- $E\left\{\widehat{\theta}(-1) \mid \text{True Data}\right\} = \widehat{\theta}_{\text{TRUE}}$

EXAMPLE: MEASUREMENT ERROR IN SYSTOLIC BLOOD PRESSURE

- Framingham Data:

$$\left(Y_j,\ \text{Age}_j,\ \text{Smoke}_j,\ \text{Chol}_j\ W_{\text{A},j}\right),\quad j = 1, \ldots, 1615$$

  - $*$ Y = indicator of CHD
  - $*$ Age (at Exam 2)
  - $*$ Smoking Status (at Exam 1)
  - $*$ Serum Cholesterol (at Exam 3)
  - $*$ Transformed SBP

$$W_{\text{A}} = \left(W_1 + W_2\right)/2,$$

$$W_k = \ln\left(\textbf{SBP} - 50\right) \ \text{at Exam } k$$

- 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

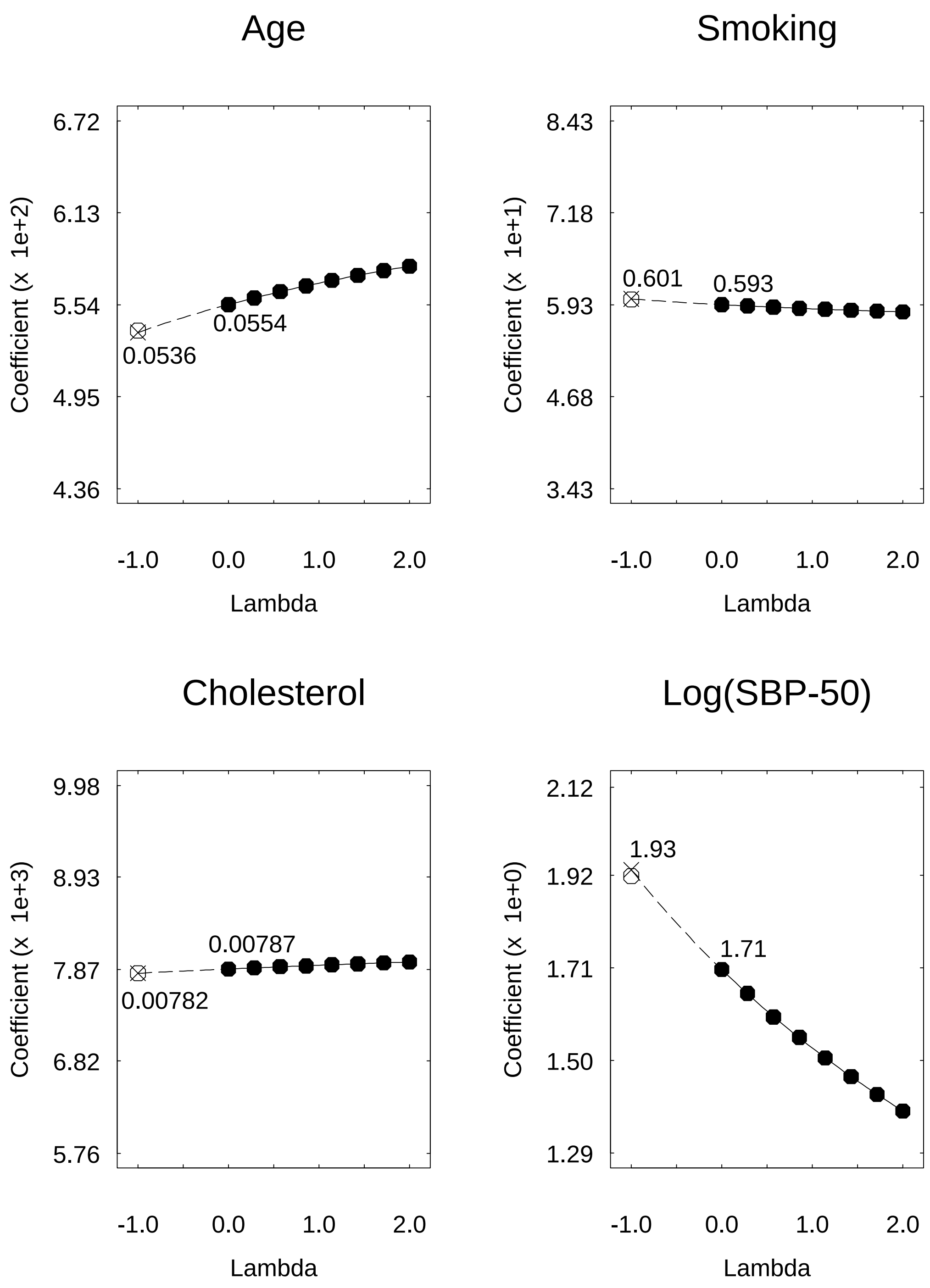
Age
Smoking
Cholesterol
Log(SBP-50)
Coefficient (x 1e+2)
Coefficient (x 1e+1)
Coefficient (x 1e+3)
Coefficient (x 1e+0)
6.72
6.13
5.54
4.95
4.36
8.43
7.18
5.93
4.68
3.43
9.98
8.93
7.87
6.82
5.76
2.12
1.92
1.71
1.50
1.29
0.0536
0.0554
0.601
0.593
0.00787
0.00782
1.93
1.71
-1.0
0.0
1.0
2.0
Lambda

Figure 34:

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.
- The plots on the following page illustrate the simulation extrapolation method for estimating the variances of the bias-adjusted parameter estimates

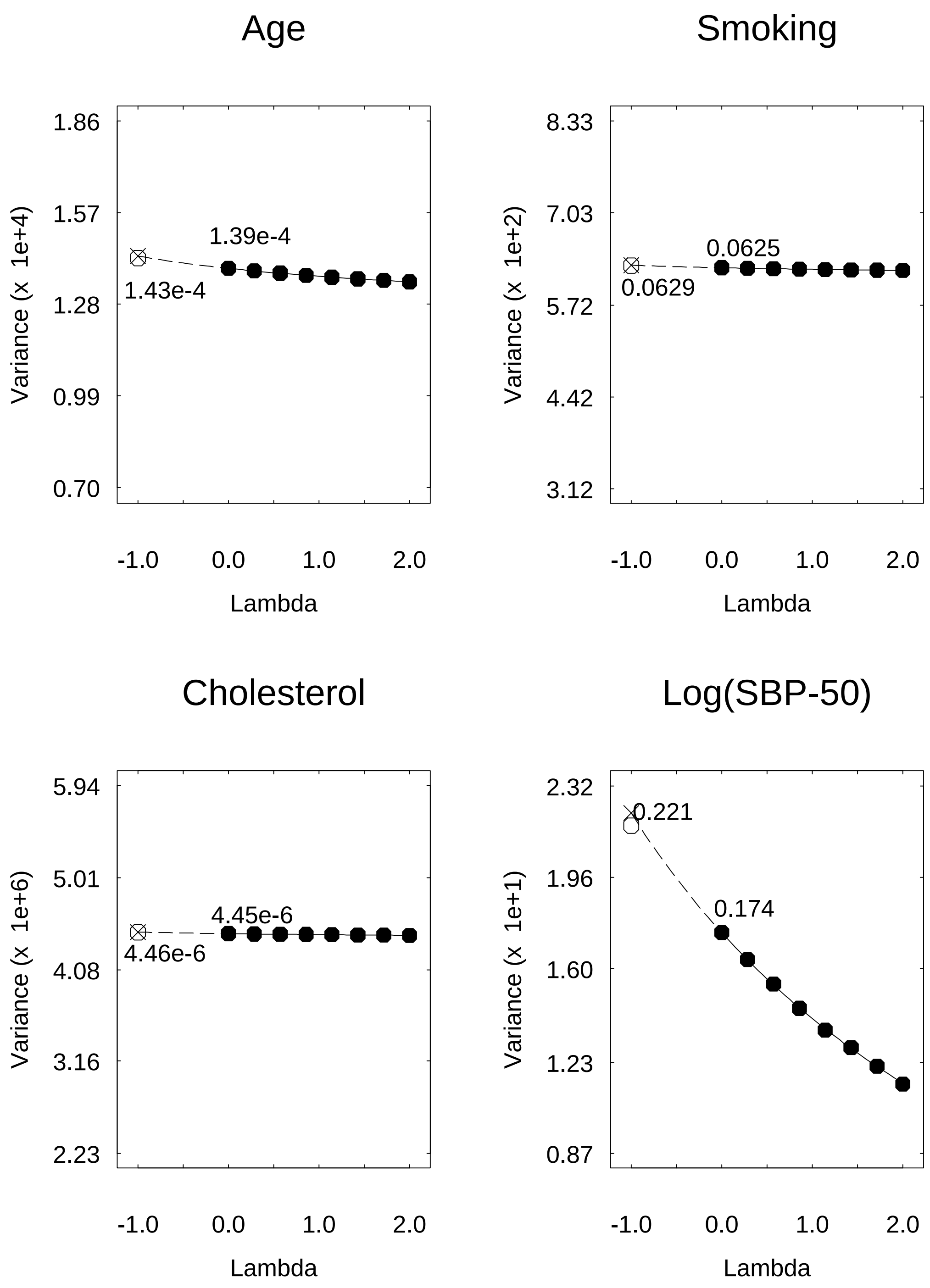
Age
Smoking
Cholesterol
Log(SBP-50)
Variance (x 1e+4)
Variance (x 1e+2)
Variance (x 1e+6)
Variance (x 1e+1)
1.86
1.57
1.28
0.99
0.70
8.33
7.03
5.72
4.42
3.12
5.94
5.01
4.08
3.16
2.23
2.32
1.96
1.60
1.23
0.87
1.39e-4
1.43e-4
0.0625
0.0629
4.45e-6
4.46e-6
0.221
0.174
-1.0
0.0
1.0
2.0
Lambda

Figure 35:

A MIXED MODEL

- We considered data from the Framingham Heart Study

- There were $m = 75$ **clusters** (individuals) with most having $n = 4$ observations, each taken 2 years apart.

- The variables were

  * Y = evidence of LVH 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) reading 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.
- We added error and refit the mixed model using CPQL **a refinement of glimmix**

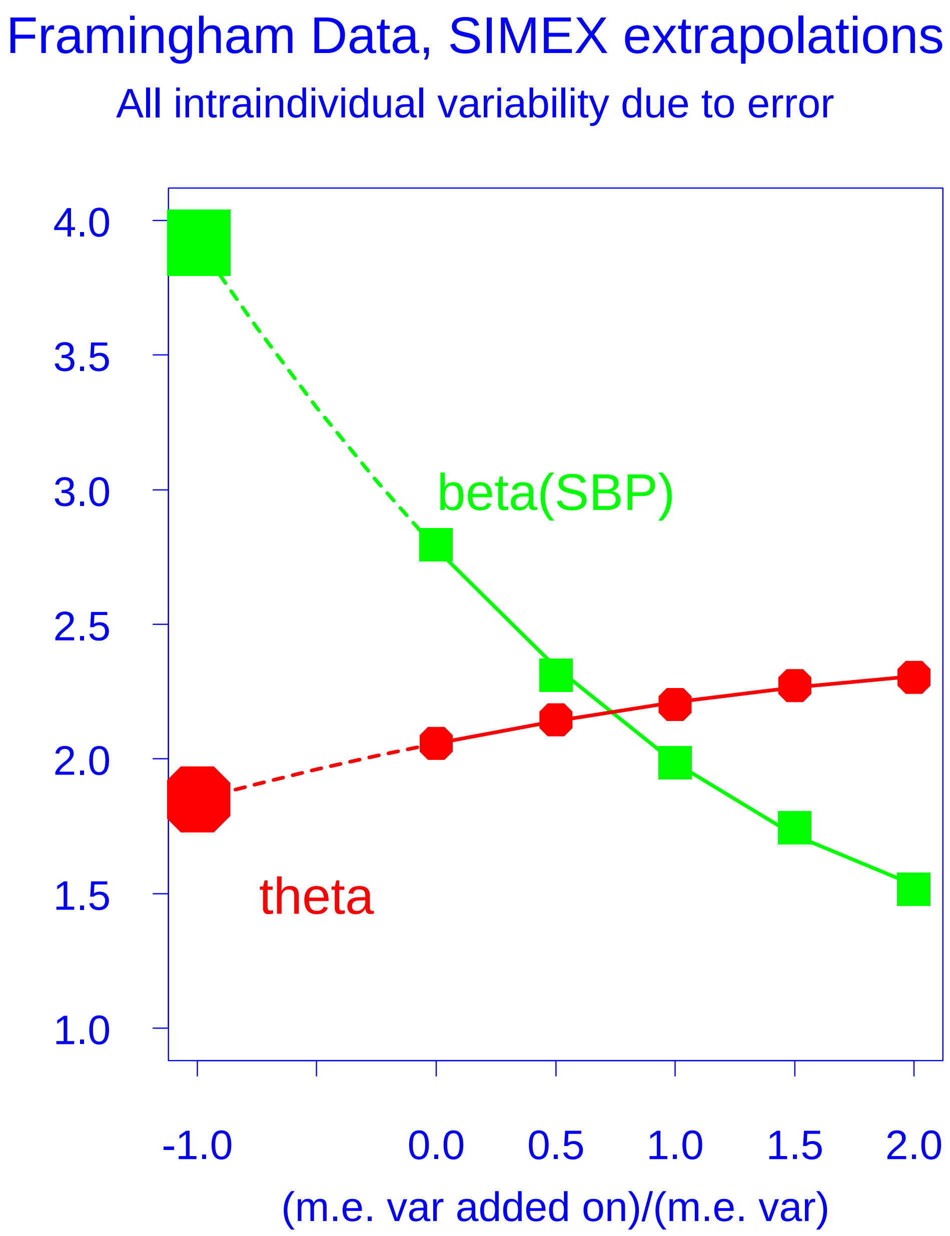

Figure 36: **LVH Framingham data. β(SBP) is the coefficient for transformed systolic blood pressure, while θ is the variance of the person–to–person random intercept.**

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 8 INSTRUMENTAL VARIABLES OUTLINE

- **Linear Regression**
- **Slightly Fancier Approach**

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, OLS estimates not β_x, but instead it estimates

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

- **Without information about σ_u^2, we cannot estimate β_x.**

INFORMATION ABOUT MEASUREMENT ERROR

- Remember, $W = X + U$, $U \sim \text{Normal}(0, \sigma_u^2)$.
- **The most direct and efficient** way to get information about σ_u^2 is to observe X on a subset of the data.
  - Missing data problem, since now X is missing only 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**.
  - 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!
- **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 **red** 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 thus 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 what 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

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

MOTIVATION

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

- 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:

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

  * 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.

IV VIA PREDICTION

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

- We have thus shown that if we have an instrumental variable, we are able to estimate the slope we want, namely
$$\beta_{Y|1\underline{X}}$$

- **Here's another way to do the same thing and get the IV estimator without doing division explicitly:**

  * Regress W on T.
  * Form predicted values: $\beta_{W|\underline{1}T} + \beta_{W|1\underline{T}}T$.
  * Regress Y on these predicted values
  * The slope is the instrumental variables estimate.

- This algorithm works even if X and W are matrices, and T is a matrix possibly of greater rank.

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).

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

- 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.

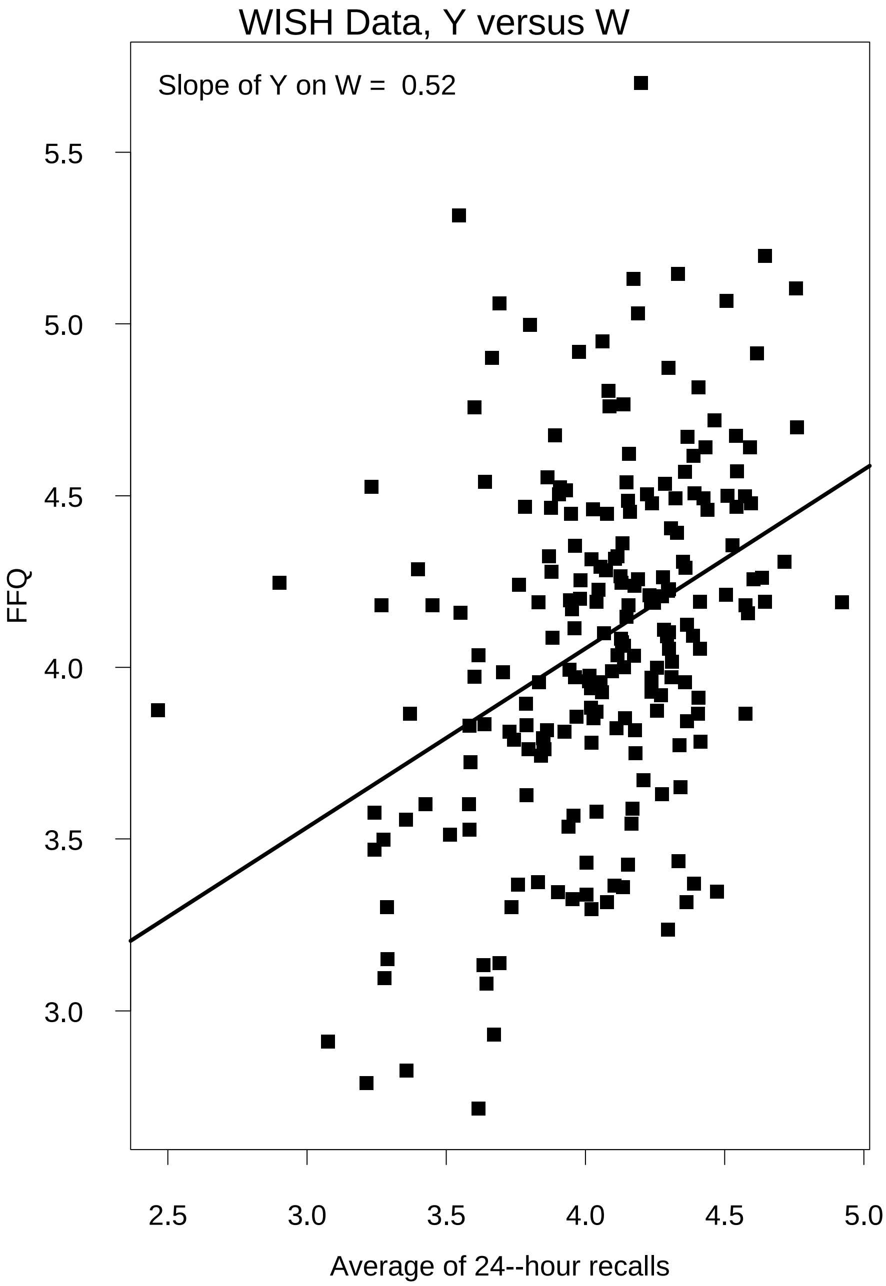

Figure 37: **Wish Data: Regression of FFQ (Y) on Mean of Recalls (W).**

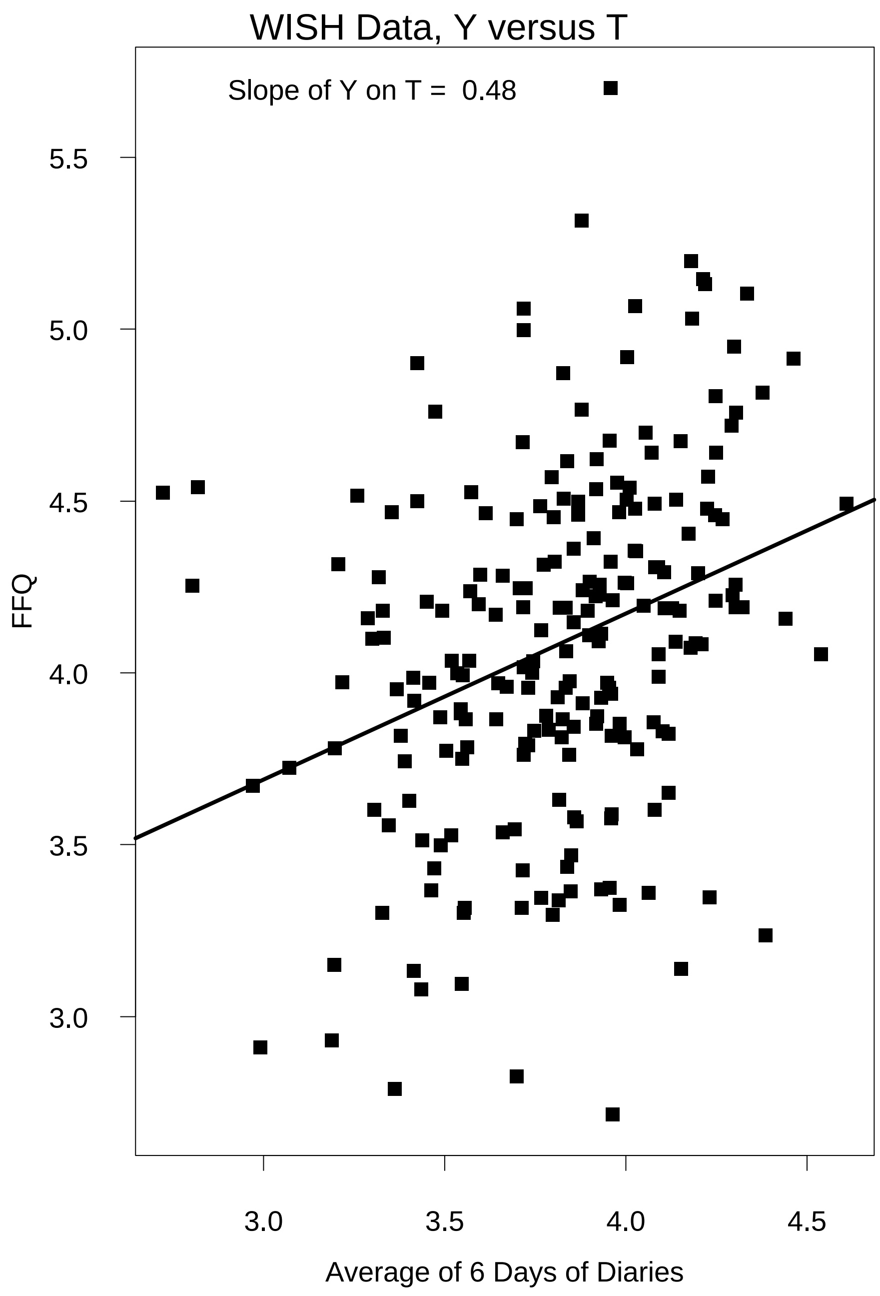

Figure 38: **Wish Data: Regression of FFQ (Y) on Mean of Diaries (T).**

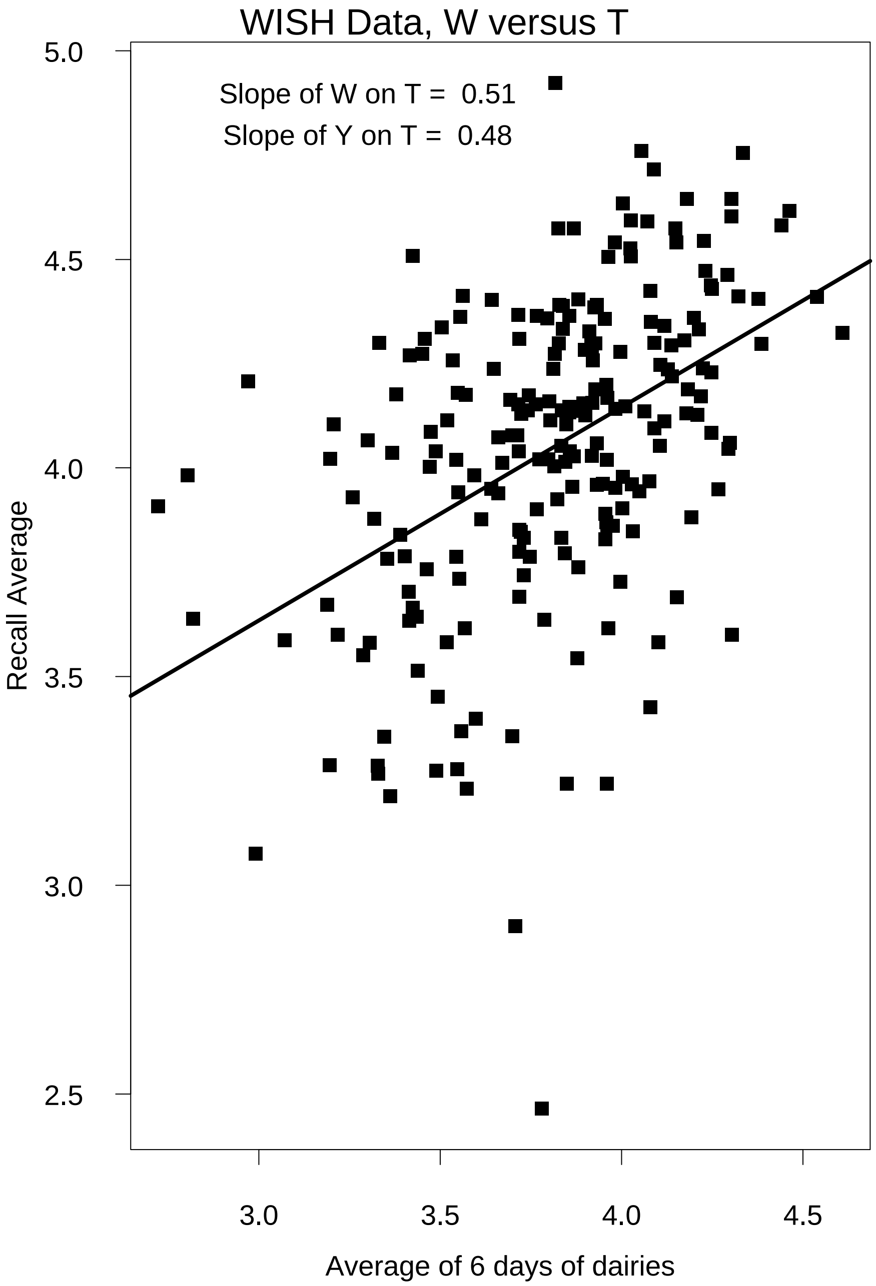

Figure 39: **WISH Data: regression of mean of recalls (W) on mean of diaries (T)**

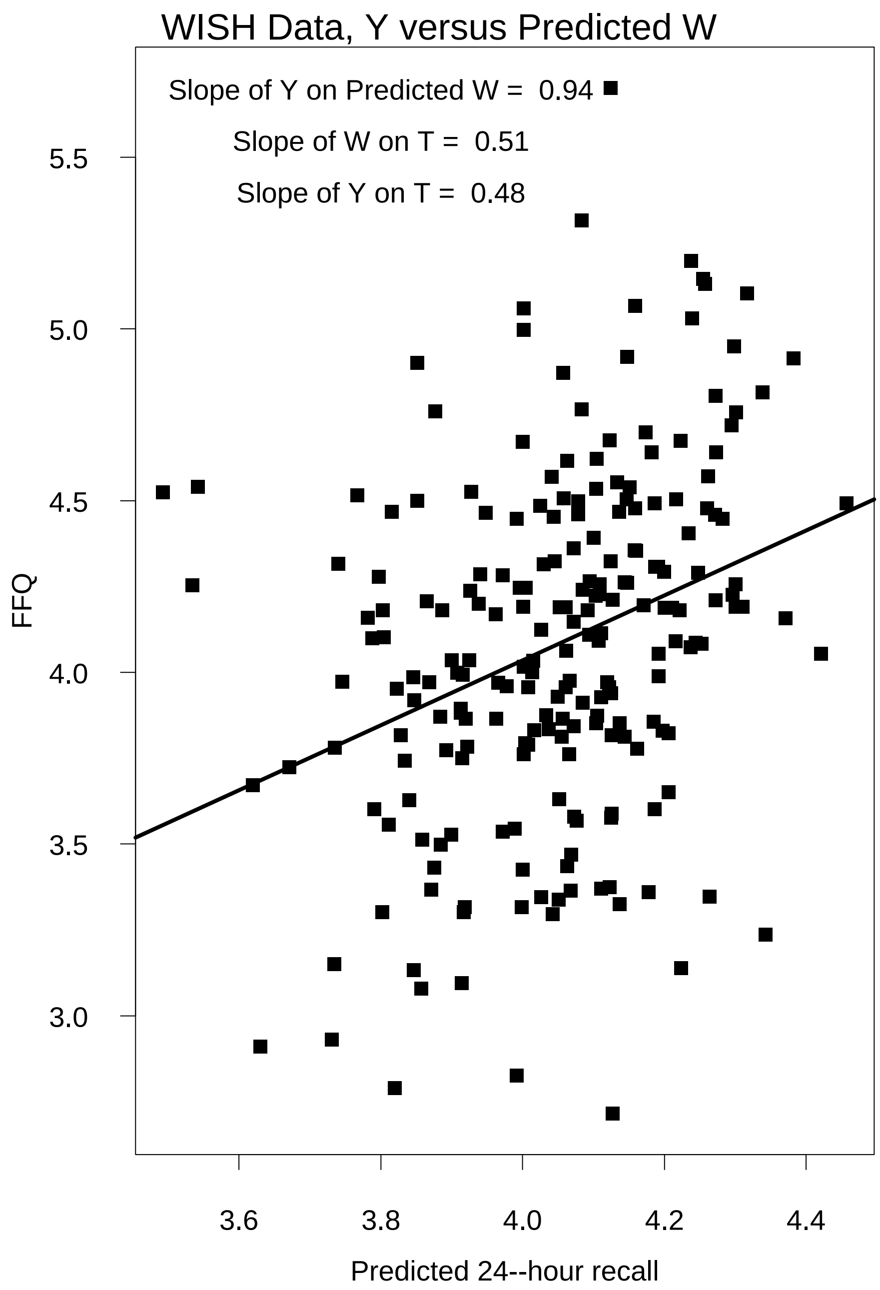

Figure 40: **WISH Data: Regression of FFQ (Y) on the Predictions from the regression of recalls (W) on diaries (T)**

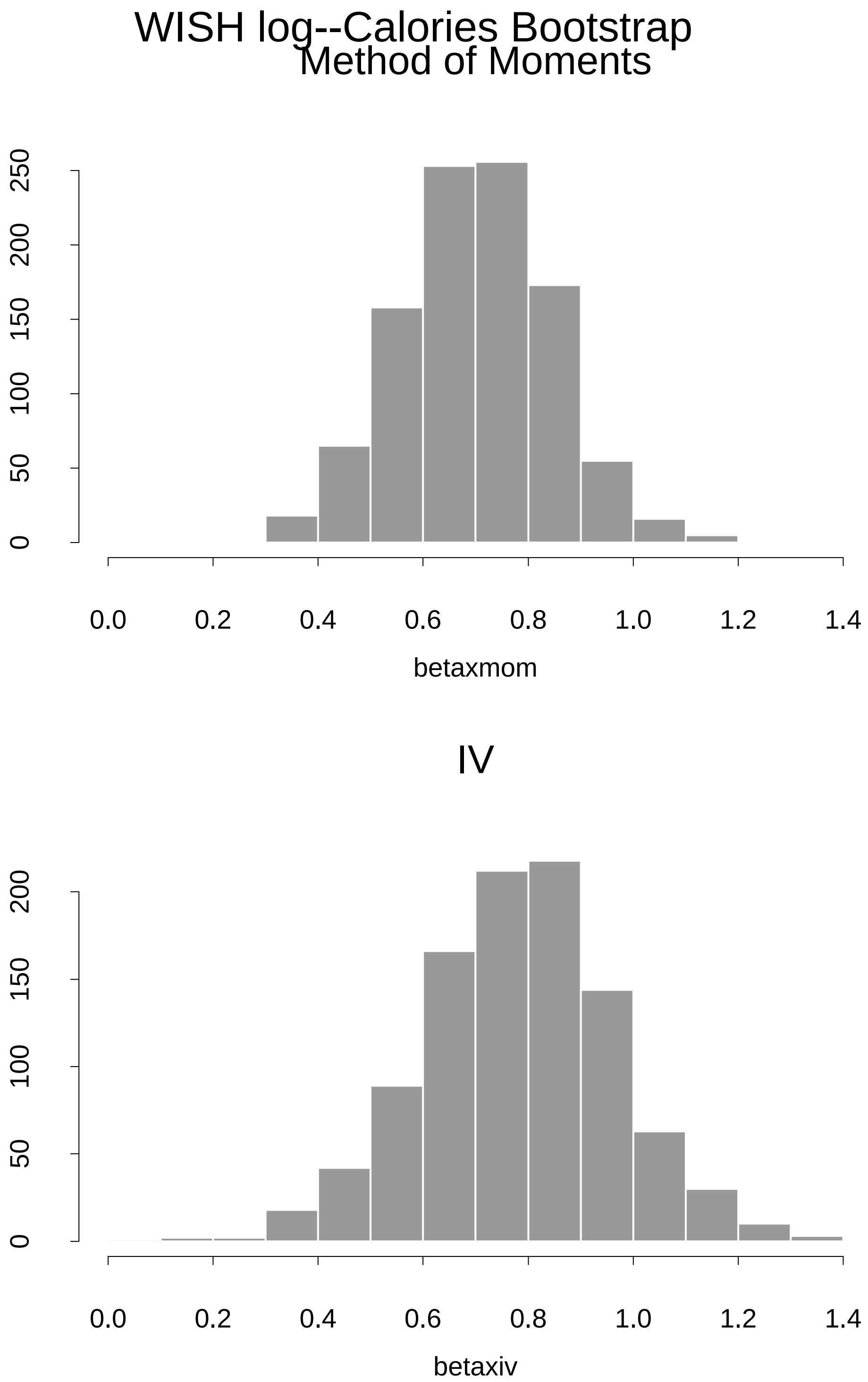

Figure 41: **Bootstrap sampling, comparison with SIMEX and Regression Calibration**

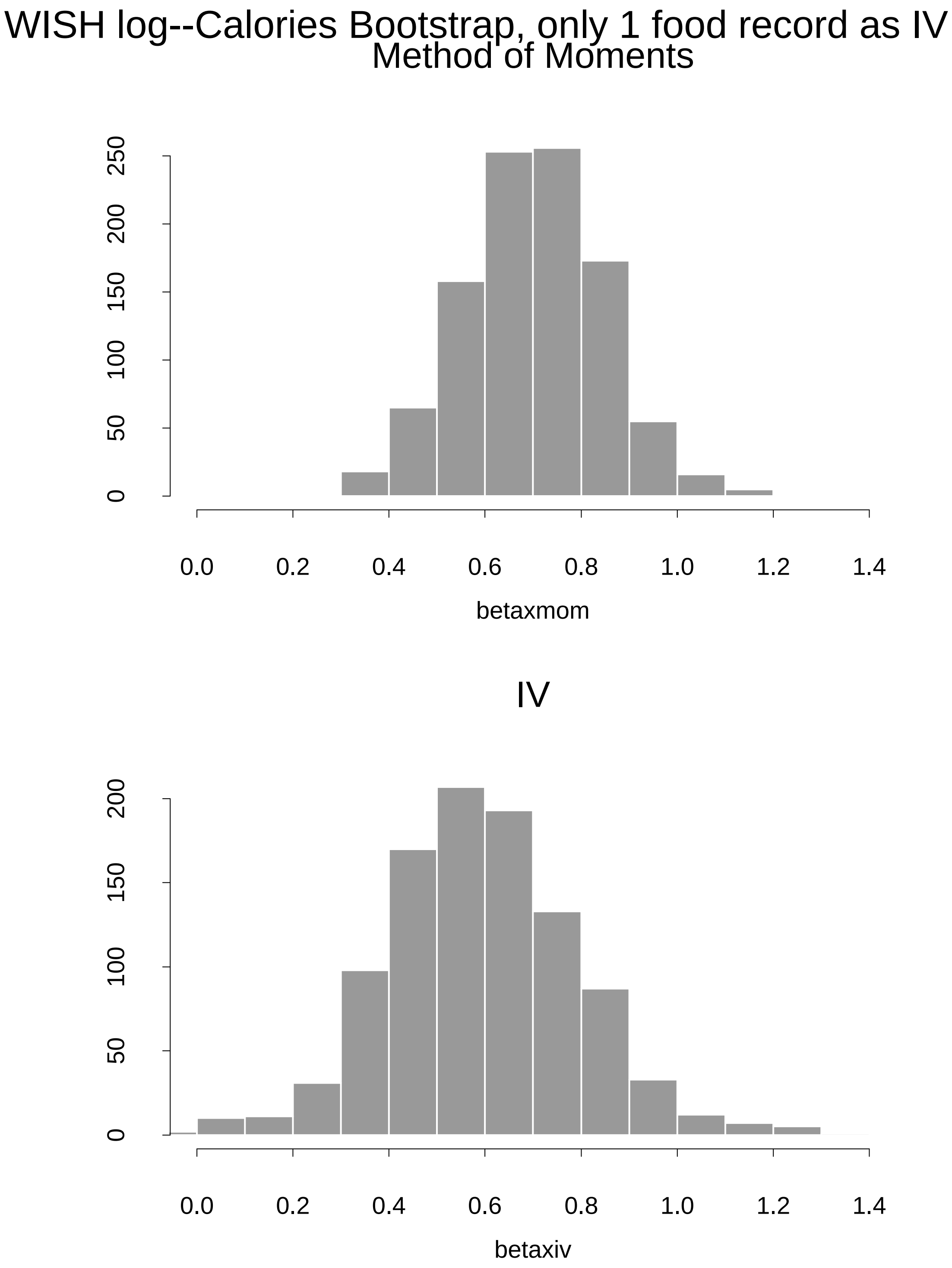

Figure 42: **Bootstrap sampling, comparison with SIMEX and Regression Calibration, when the instrument is of lower quality and one one of the diaries is used.**

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 8: NONPARAMETRIC REGRESSION: OUTLINE

- **We illustrate the two main ideas of the short course (SIMEX and likelihood) on the problem of non-parametric regression.**
- **Only SIMEX is computed**
- **The ideas remain the same. The technical details are more challenging.**

WHAT IS NONPARAMETRIC REGRESSION?

- Ordinary nonparametric regression is a field with an enormous amount of papers.
- Kernels, loess (Splus), regression splines, smoothing splines, orthogonal series, etc.
- The basic problem is to relate a response Y to a predictor X:

$$E(Y|X) = m(X).$$

- The nonparametric part arises when **no assumptions are made about the form of $m(x)$, other than smoothness.**
- Splus has good spline methods standard. Kernel methods on the statlib, and also at

http://www.orie.cornell.edu/∼ruppert

THE PROBLEM OF MEASUREMENT ERROR

- We are interested in nonparametric regression when the predictor X cannot be observed exactly.
- We'll assume the **classical error model** $W = X + U$, where U is a mean zero, variance σ_u^2 normal random variable.

  * The normality is not important.
  * For today, measurement error variance σ_u^2 assumed known.

  – The effects of measurement error are:

  * bias estimates of the regression line
  * increase the perceived variability about the regression line.

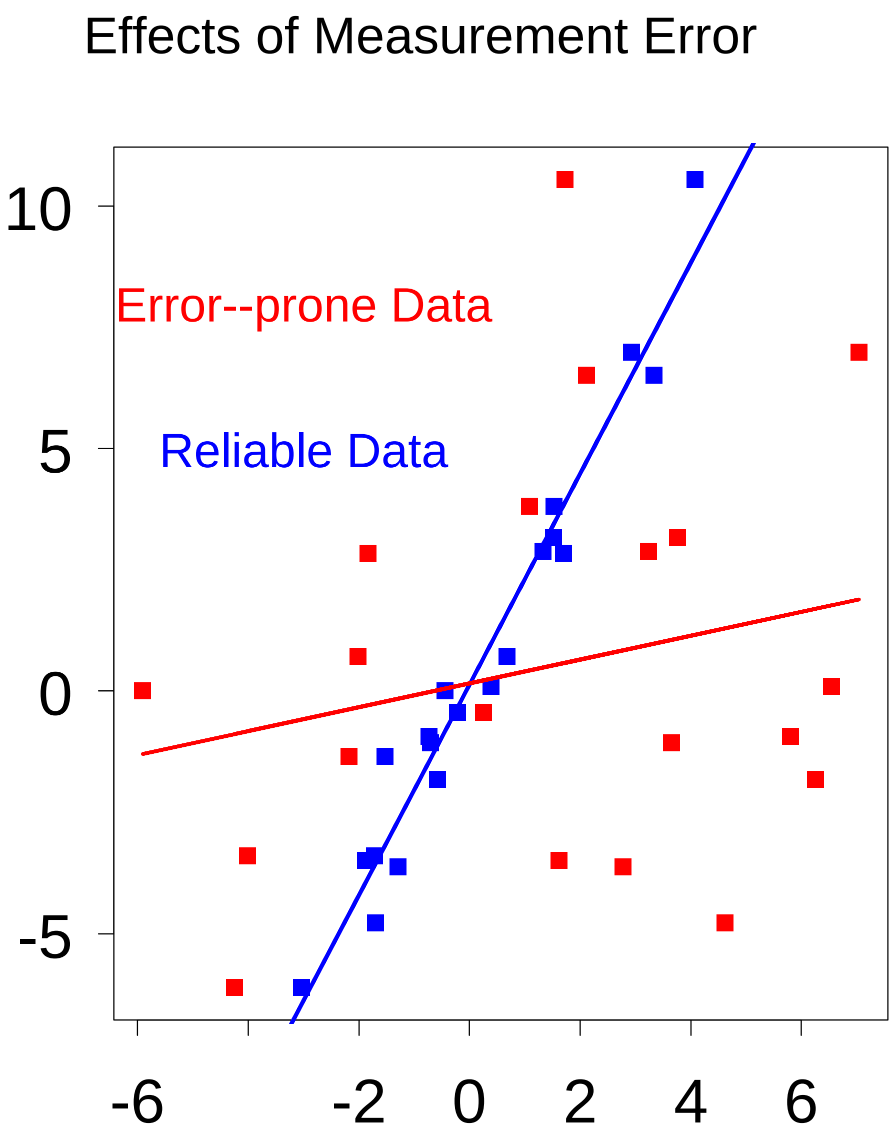

Figure 43: **Observed Data With Measurement Error.**

KERNEL AND LOCAL LINEAR ESTIMATORS

$$E(Y|X) = m(X)$$

- We want to remind you that **kernel and loess estimators are formed by weighted linear regression.**

- If you want to estimate the function $m(\cdot)$ at x_0, the weights are large for $X \approx x_0$ and they decrease as X moves away from x_0.

- The actual estimate at x_0 is the **prediction of the weighted least squares fit** at x_0.

SIMEX

- Let's remember the SIMEX algorithm, in pictures.

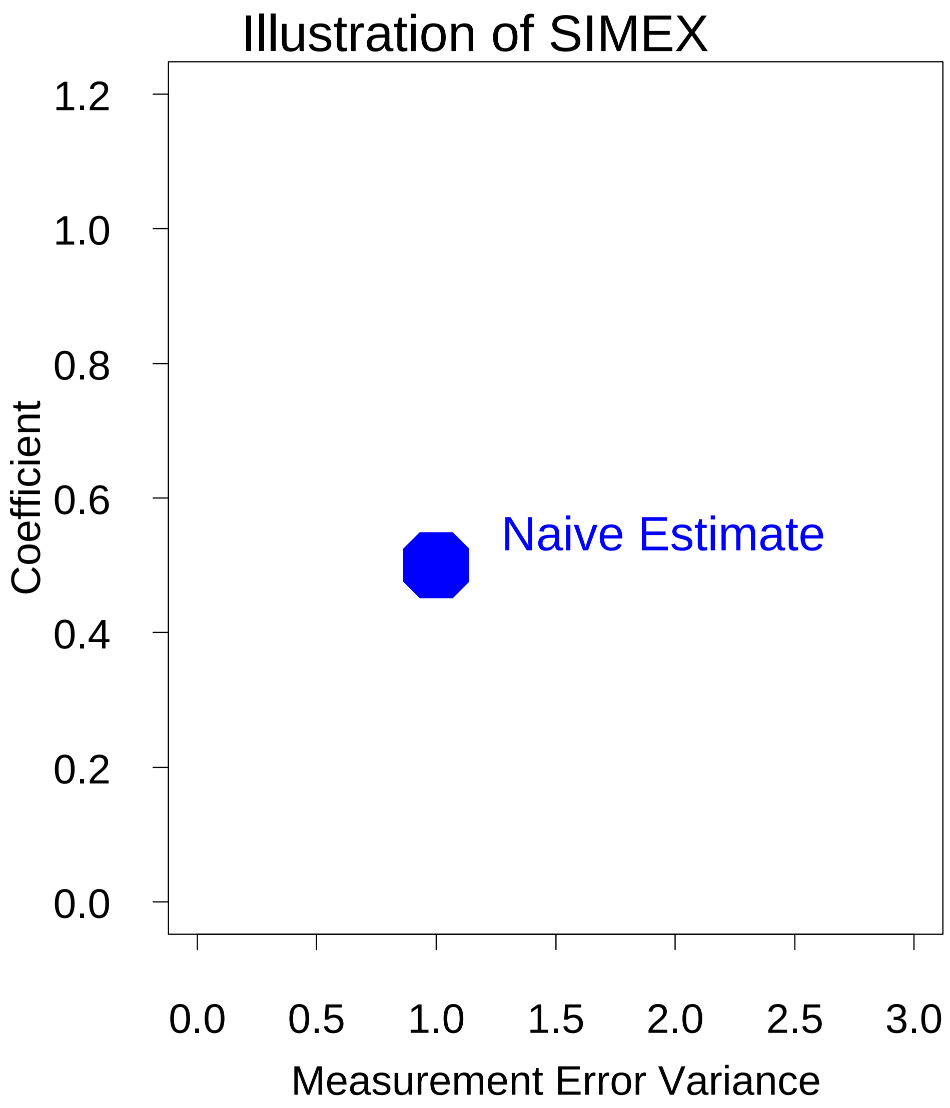

Figure 44: **Your estimate when you ignore measurement error.**

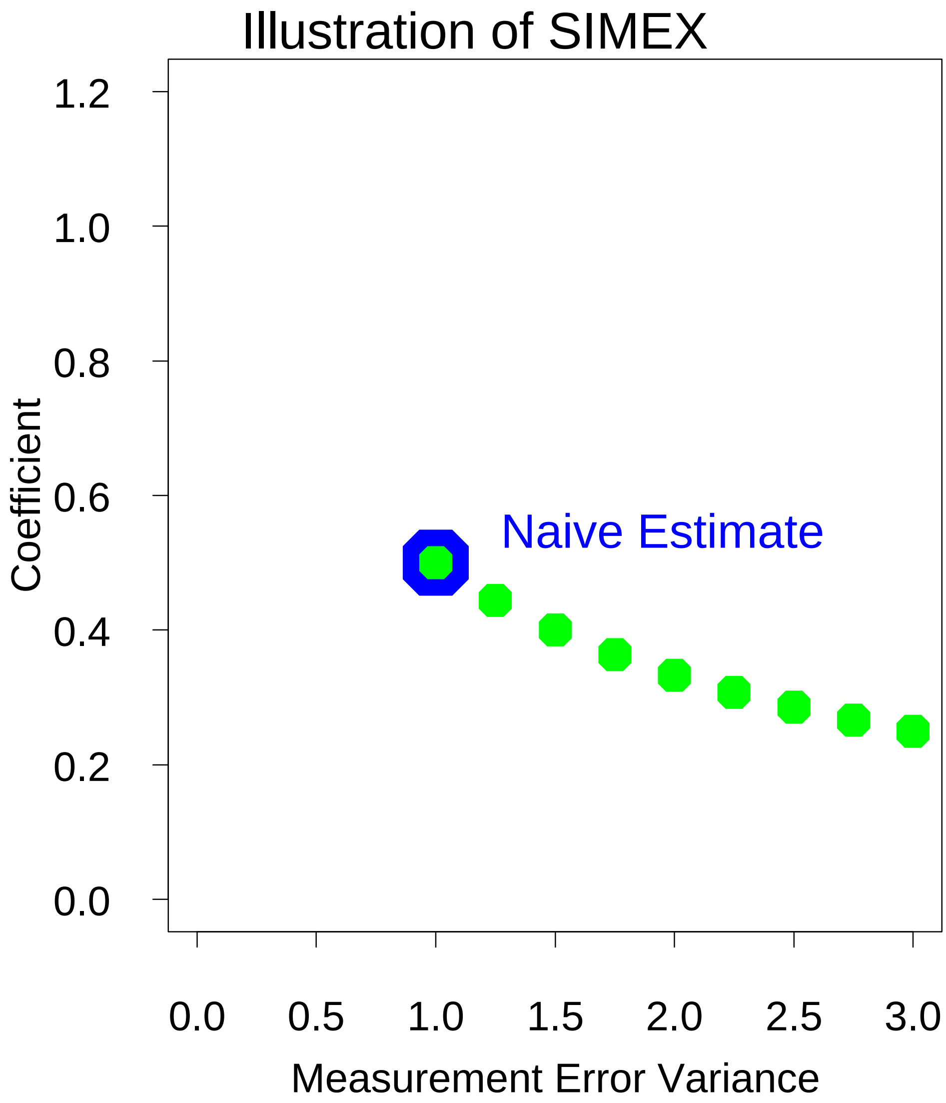

Figure 45: **What heppens to your estimate when you have more error, which you add on by simulation, but you still ignore the error.**

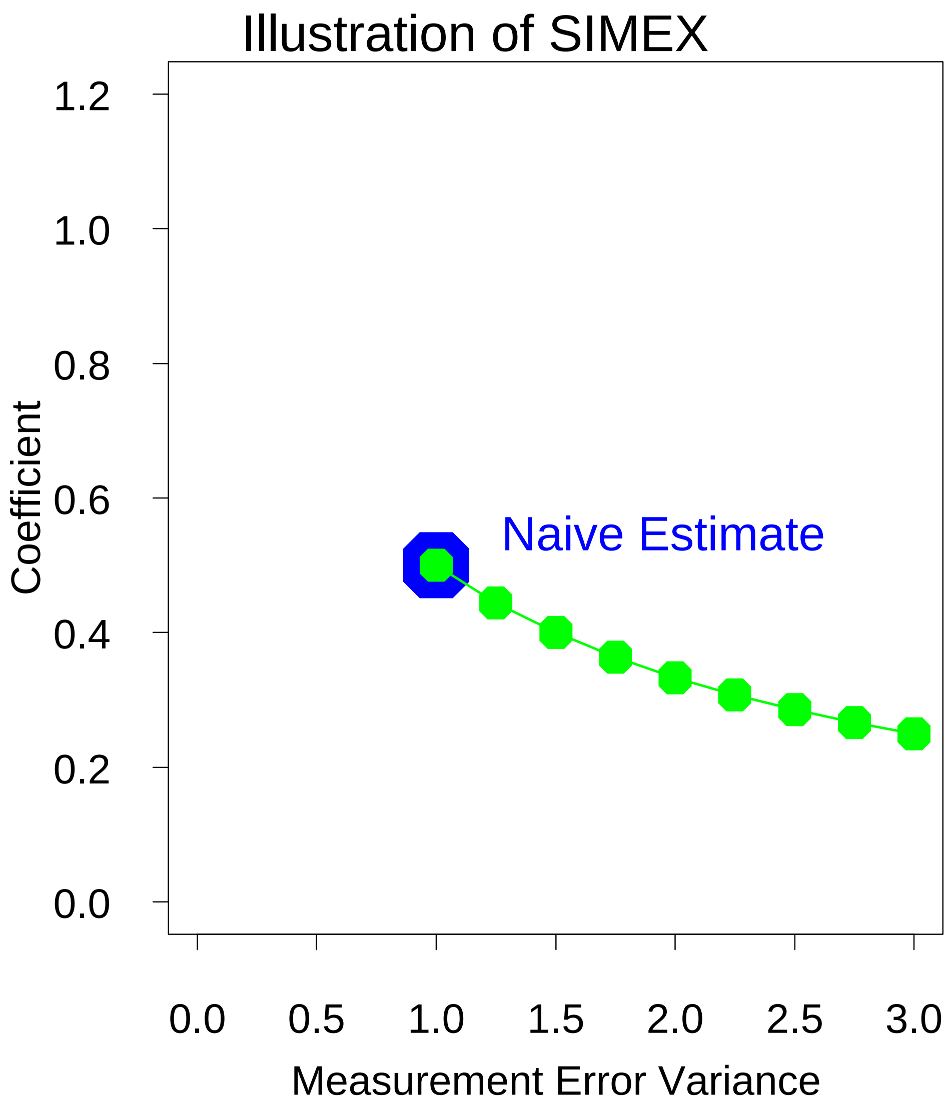

Figure 46: **What statistician can resist fitting a curve?.**

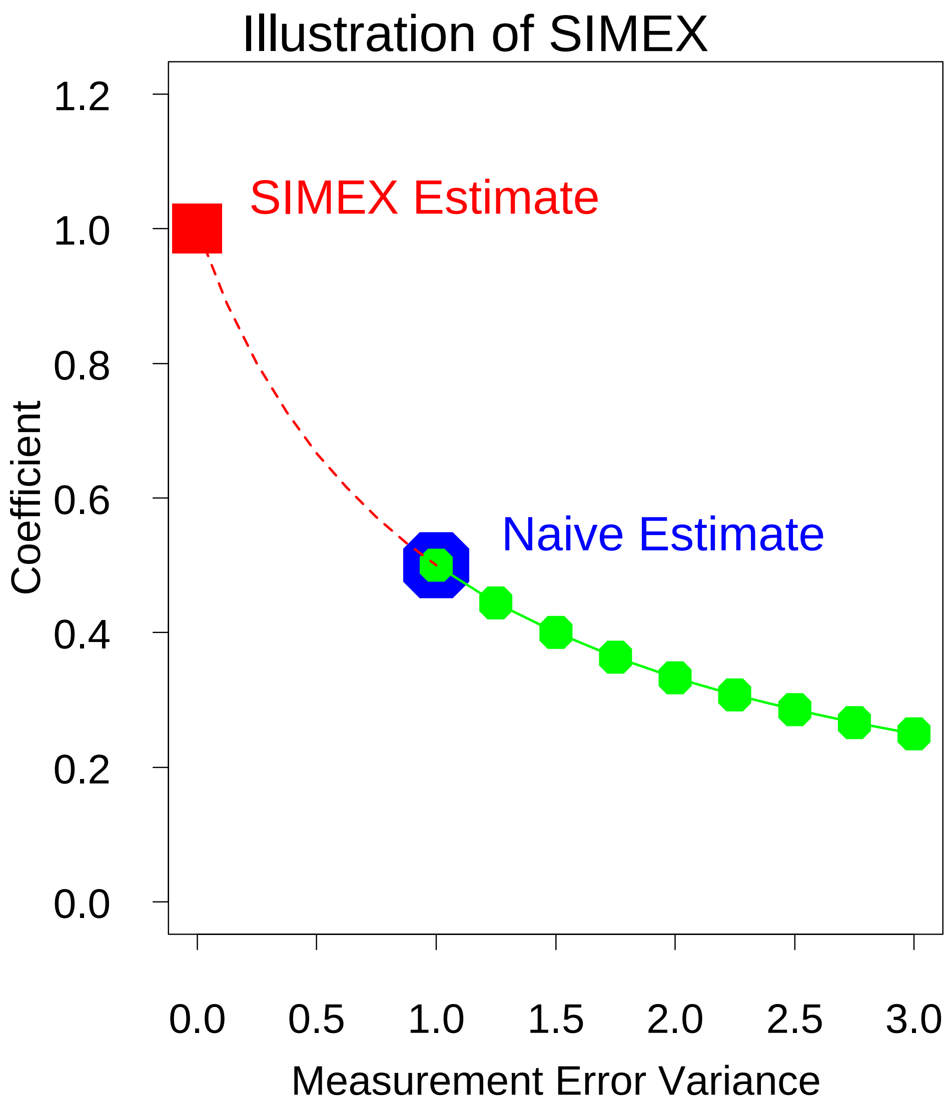

Figure 47: **Now extrapolate to the case of no measurement error.**

SIMEX

- SIMEX works for just about any problem.
- We have a paper to appear in *Biometrika* which shows that **SIMEX works for nonparametric regression just as well as it works for parametric regression**.
- It can be applied to your favorite method of nonparametric regression, as follows.
  - Replace a parametric estimator by a nonparametric one.
    - Extrapolate **on a point–by–point** basis.
    - Thus, if you want the function at x_0, **extrapolate the fits at** x_0.
- We have a submitted paper which shows how the idea works for longitudinal models.

WHAT IF WE HAVE A DISTRIBUTION FOR X?

$$E(Y|X) = m(X), W = X + U$$

- The **SIMEX method makes no assumptions about the distribution of X.**
    - $*$ This is a striking robustness property.
- **Suppose however that we knew the distribution of X, and that U is normally distributed.**
    - $*$ Could we get anywhere?
    - $*$ We in particular know the distribution of X given W!

WHAT IF WE HAVE A DISTRIBUTION FOR X?

$$E(Y|X) = m(X), W = X + U$$

- Suppose that $m(x)$ is a polynomial: $m(x) = \Sigma_{j=0}^{J} \beta_j x^j$. Then, $E(Y|W) = \Sigma_{j=0}^{J} \beta_j E(X^j|W)$.
- Since X given W is known, so too is $E(X^j|W)$, and hence we can estimate the β's by ordinary least squares with "predictors" $E(X^j|W)$.

REGRESSION SPLINES

$$E(Y|X) = m(X), W = X + U$$

- Many functions are well fit by what are called **regression splines.**
- These are functions which start out as polynomials, but flexibility is added by placing "knots" at convenient locations and modifying the function locally.
- Specifically, regression splines of order J with K knots look like:

  $$E(Y|X) = \Sigma_{j=0}^{J} \beta_j X^j + \Sigma_{j=1}^{K} \beta_{j+J}(X - \xi_j)_+^J$$

  * The terms $\xi_1, ..., \xi_K$ are knots, to build flexibility
  * $(X - \xi)_+ = (X - \xi)I(X - \xi > 0)$
- There is a large literature on fitting these things, selecting knots, etc.
  * **Close cousins: smoothing splines.**

REGRESSION SPLINES

$$
\begin{aligned}
E(Y|X) &= m(X), \quad W = X + U; \\
E(Y|X) &= \sum_{j=0}^{J} \beta_j X^j + \sum_{j=1}^{K} \beta_{j+J}(X - \xi_j)_+^J
\end{aligned}
$$

- If we know the joint distribution of X and U, then the observed data have

$$
E(Y|W) = \sum_{j=0}^{J} \beta_j E(X^j|W) + \sum_{j=1}^{K} \beta_{j+J} E\{(X - \xi_j)_+^J|W\}
$$

- **This is just a linear model in the β's!!!!**
- Thus, it would appear fairly natural that if you take a structural approach, then you can estimate the regression functions.
  - $*$ Possibly even more efficiently than via SIMEX.

REGRESSION SPLINES

- Typically, with regression splines, one lays down a fairly large number of knots.
  - **Because there are a lot of knots, there is a potential for overfitting.**
  - Typically, the solution is to penalize the likelihood in a **ridge regression** way.
  - Hence, it is typical to minimize something like
$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{J} \beta_j X^j - \sum_{j=1}^{K} \beta_{j+J} (X - \xi_j)_+^J \right\}^2 + \lambda \sum_{j=1}^{K} \beta_{j+J}^2$$
- λ is the smoothing constant.
  - Choice of λ in a set of data is a technical area undergoing vigorous development, and would be another short course in itself.
- For our examples, we will just write down a very few knots (≤ 3), and we will ignore the smoothing parameter.

AIDS DATA EXAMPLE

- Y = risk of hospitalization in a fixed period for a patient with AIDS.
- X = true CD4 count, actually

$$\log(\text{CD4}/100).$$

- Studies have shown that the measurements errors in X are additive, with a known variance. The reliability is 0.70.
- An initial fit to the data using gam in Splus suggests nonlinearity.

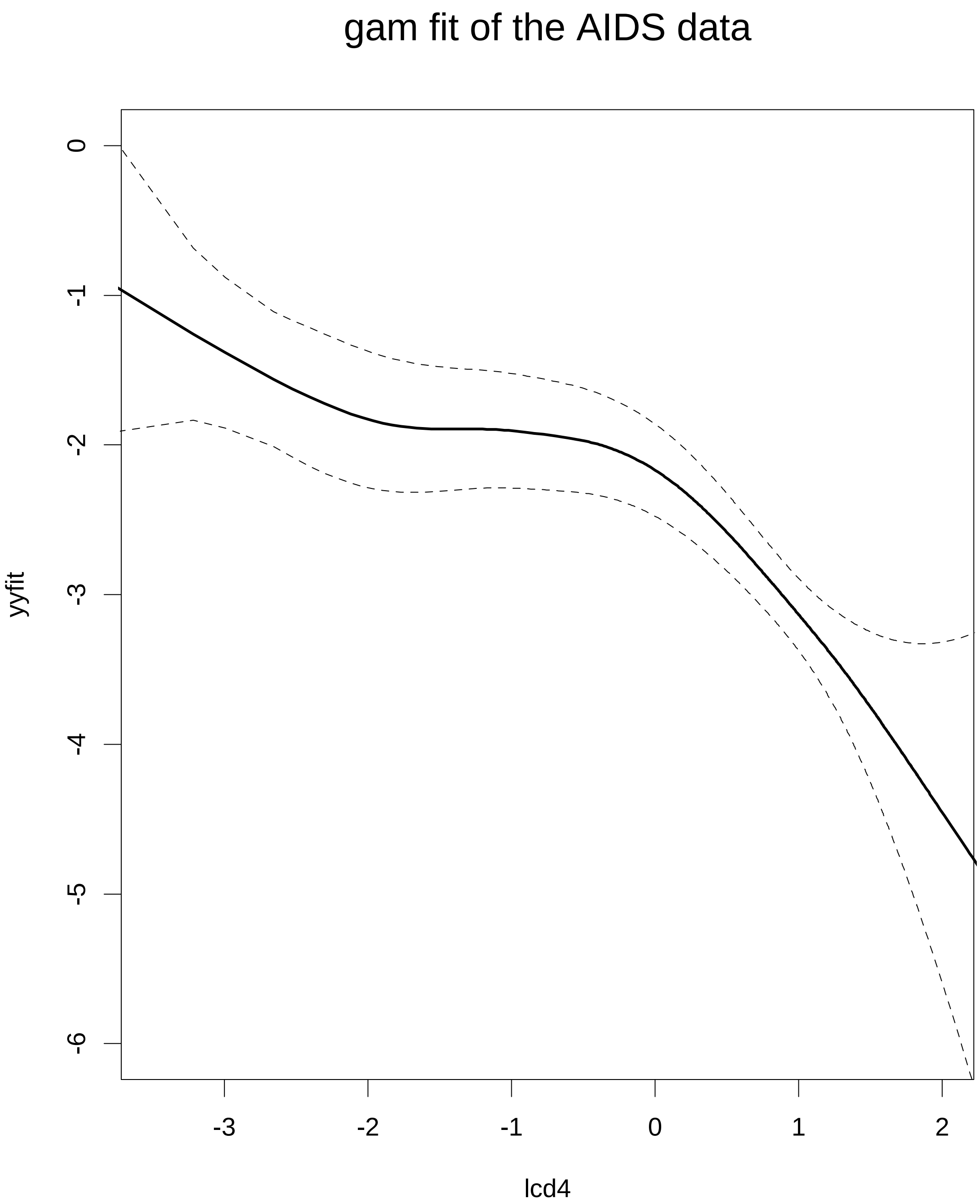

Figure 48: **AIDS data ignoring measurement error. GAM fit in Splus, which seems somewhat badly fitting, see sample logits.**

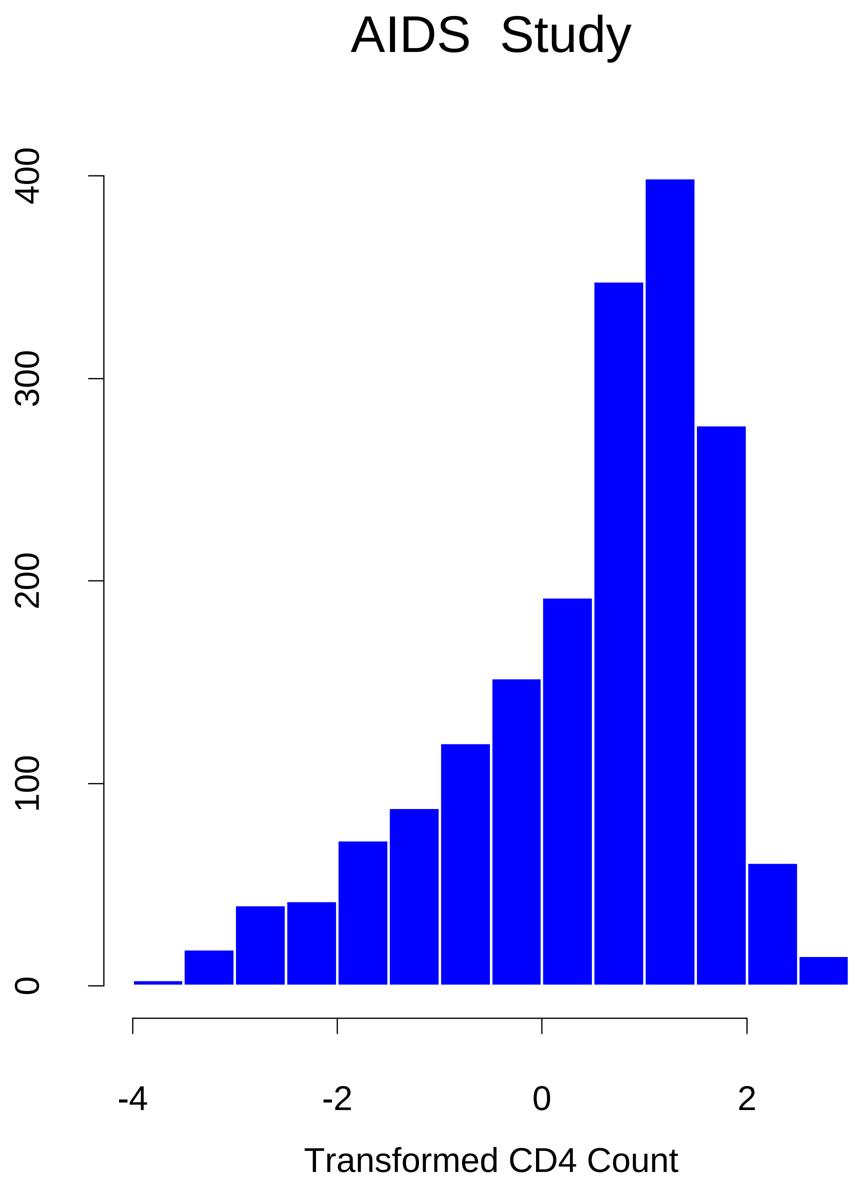

Figure 49: **AIDS data: histogram of transformed CD4 counts.**

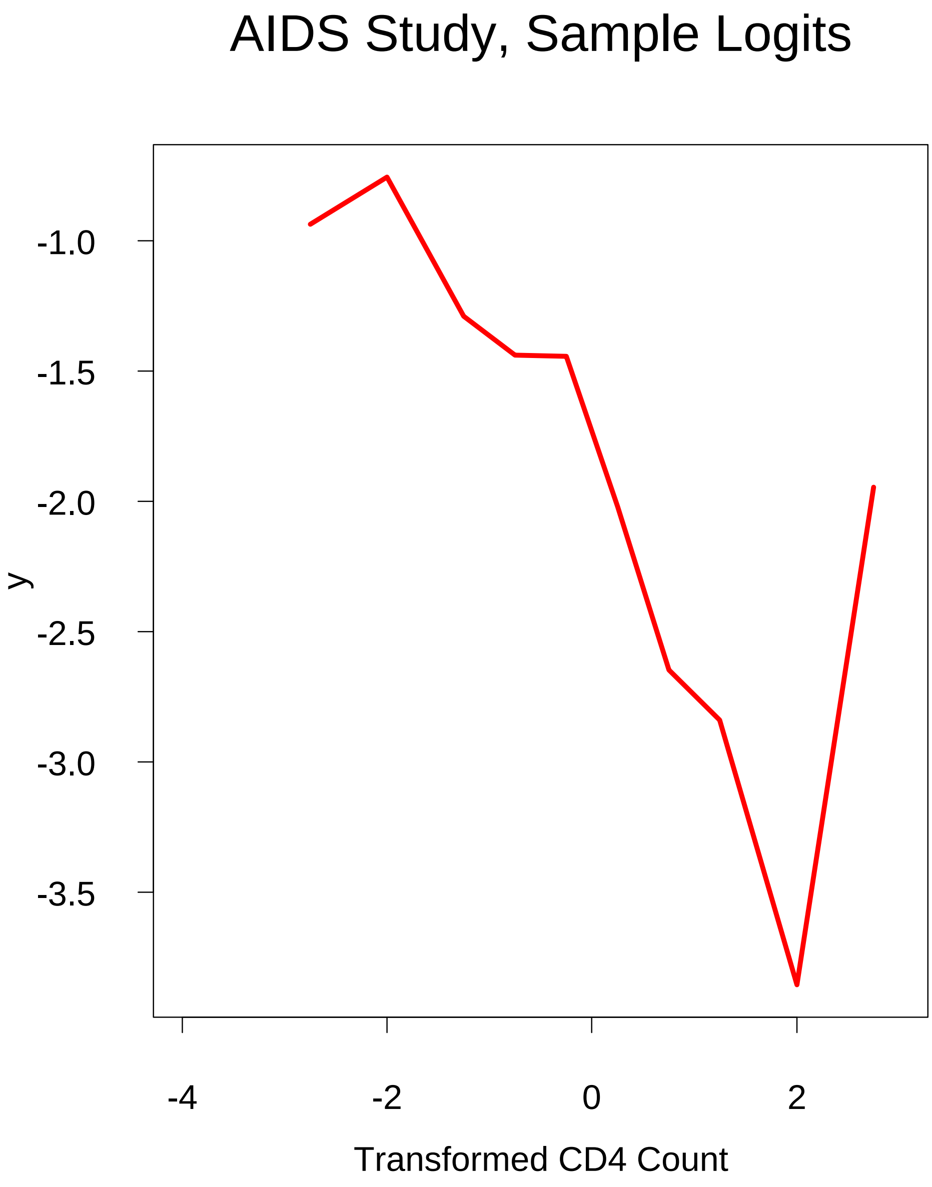

Figure 50: **AIDS data: Sample Logits of Probability of hospitalization**

AIDS DATA EXAMPLE

- The plot suggests placing knots at -2, -1, 0, 1 and fitting a regression spline.
- We will compare four methods
  - Linear logistic
  - Quadratic logistic
  - Spline ignoring measurement error
  - SIMEX spline
- We'd expect of course that the SIMEX fit will exaggerate the peaks and valleys of the naive spline fit, because it wants to deattenuate.

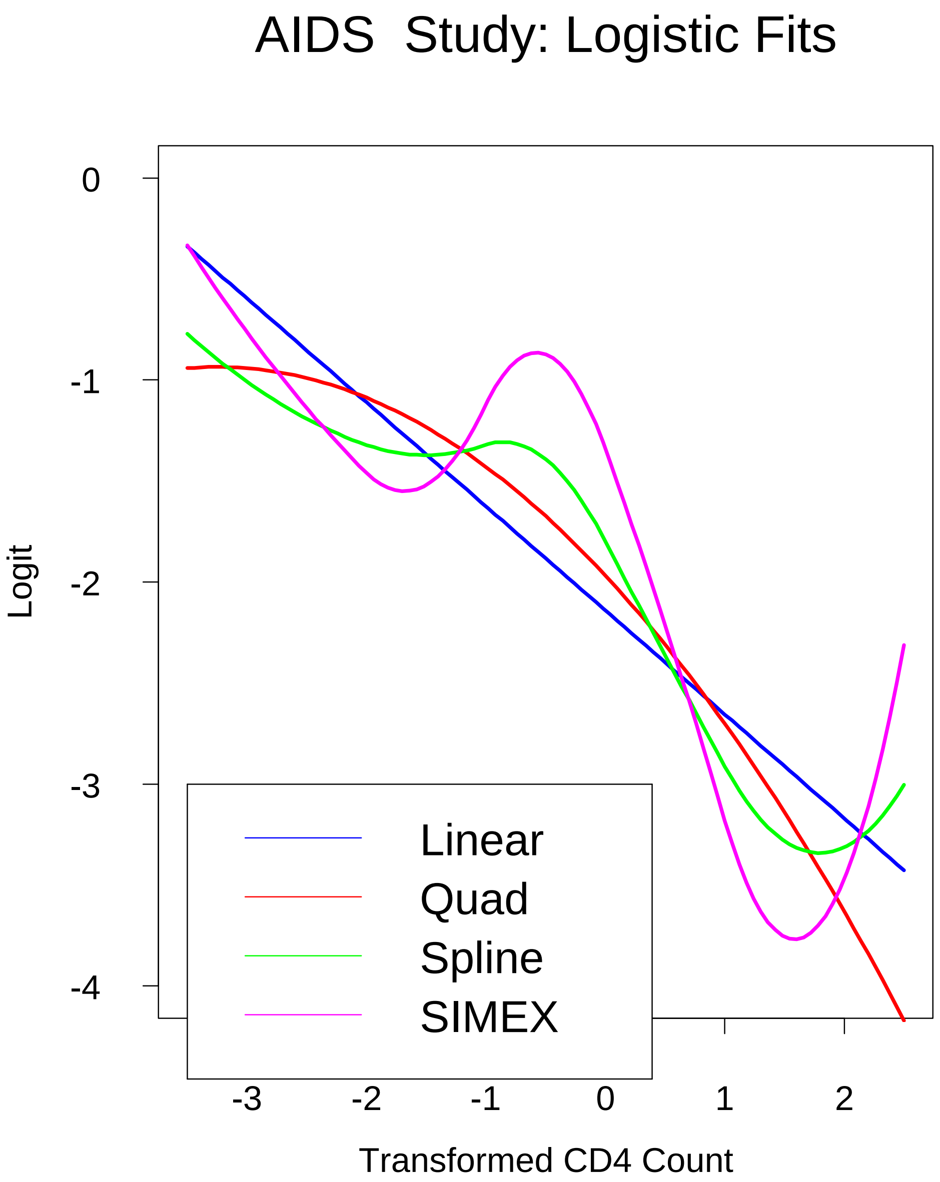

Figure 51: **AIDS data: Naive fit and SIMEX spline fit.**

AIDS DATA EXAMPLE

- I generated data according to the model that has a reliability of 0.75.
- I fit linear and quadratic splines.
- We will compare to methods
  - Spline ignoring measurement error
  - SIMEX spline
- Note how SIMEX tends to get to the peaks and valleys a bit better.
- Note too why it is better to use more knots and penalize the changes at the knots.
- This is our major current research interest.

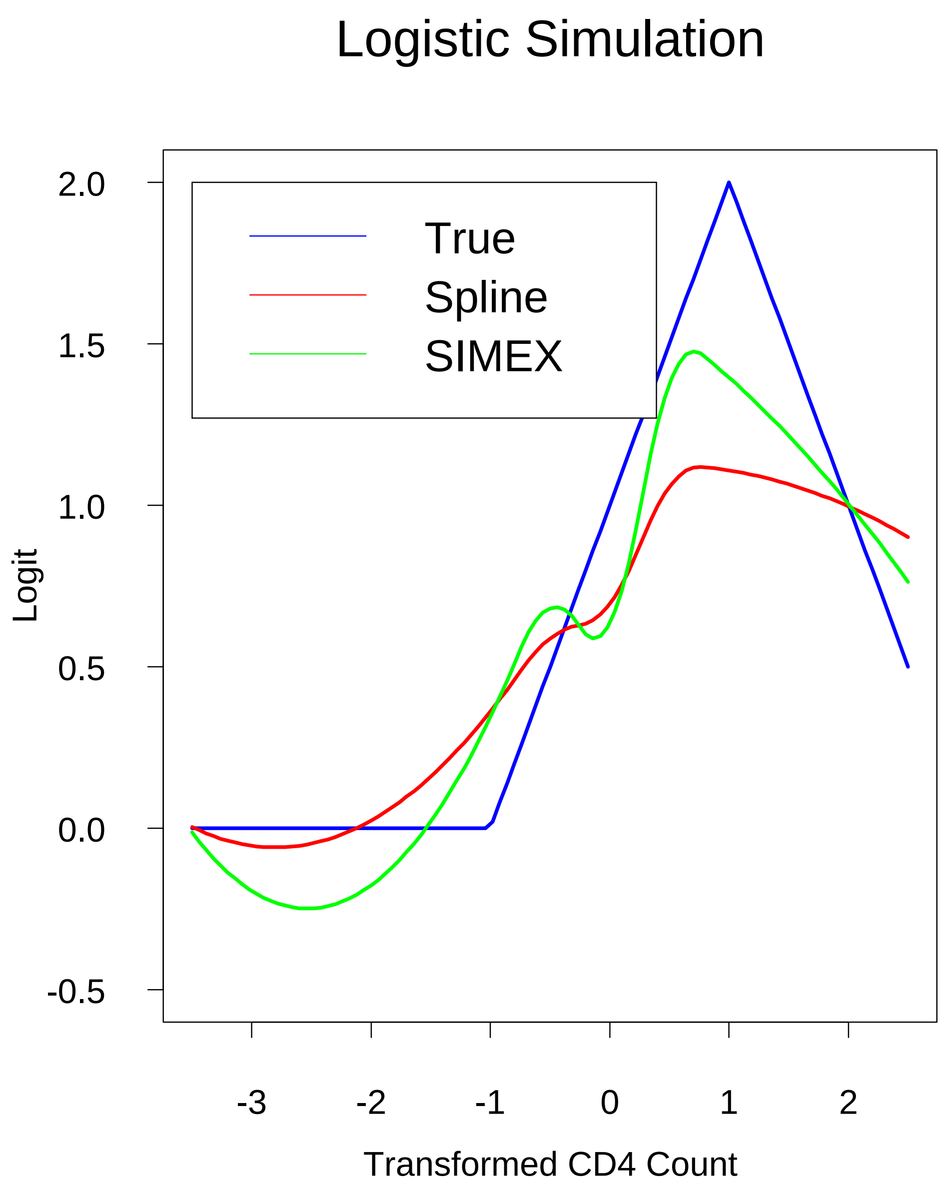

Figure 52: **Simulated data, quadratic splines: Naive fit and SIMEX spline fit.**

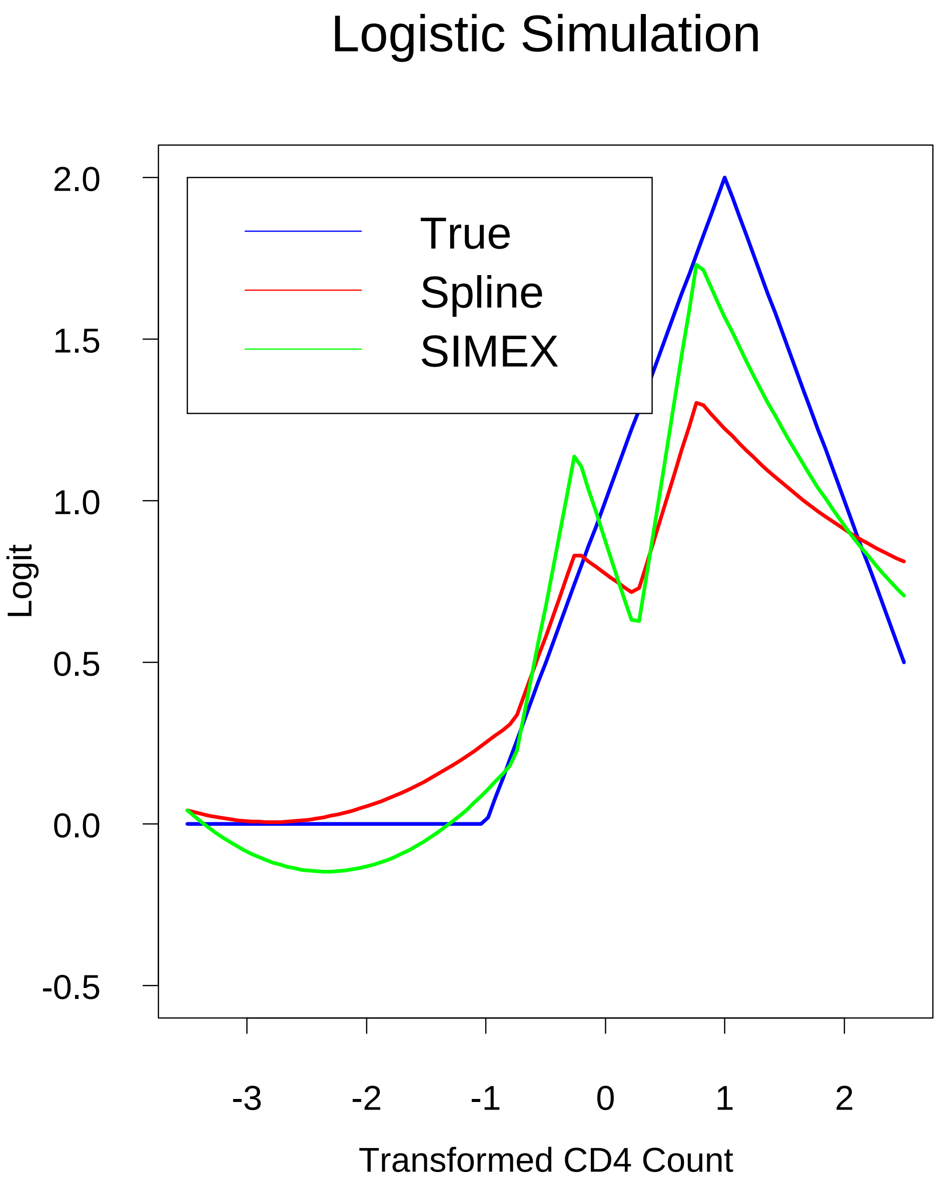

Figure 53: **Simulated data, linear splines: Naive fit and SIMEX spline fit.**