Measurement Error and Misclassification in statistical models: Basics and applications bcam Bilbao Part 3

Helmut Küchenhoff
Statistical Consulting Unit
Ludwig-Maximilians-Universität München

Bilbao
29-05-2019

Methods

- Functional and structural
- Correction and method of moments and orthogonal regression
- Regression calibration
- Likelihood
- Quasi likelihood
- Bayes

Functional and structural

- **Functional:**
  X fixed unknown constants
  No assumptions about the distribution of X
- **Structural:** X latent random variable
  Use assumptions about the distribution of X

Method of moments

Moments of observed data can be estimated
Solve moments equations
Simple linear regression:

$$
\begin{aligned}
\mu_{X^*} &= \mu_X \\
\mu_Y &= \beta_0 + \beta_1 \mu_x \\
\sigma^2_{x^*} &= \sigma^2_x + \sigma^2_u \\
\sigma^2_y &= \beta_1^2 \sigma^2_x + \sigma^2_\epsilon \\
\sigma_{yx^*} &= \beta_1 * \sigma^2_x
\end{aligned}
$$

Orthogonal regression, total least squares

In linear Regression with classical additive measurement error:

- Assume $\frac{\sigma_\epsilon^2}{\sigma_u^2} = \eta$ is known,
  e. g. no equation error and σ_ϵ is measurement error in Y. Minimize

$$\sum_{i=1}^{n} \left\{ (Y_i - \beta_0 - \beta_1 X_i)^2 + \eta (X_i^* - X_i)^2 \right\}$$

  in $(\beta_0, \beta_1, X_1, X_2, \ldots, X_n)$.
- Total least squares, Van Huffel (1997)
- Technical symmetric applications
- In other applications a problem, assumption of no equation error not realistic

Regression calibration

This simple method has been widely applied. It was suggested by different authors: Rosner et al. (1989) Carroll and Stefanski(1990)

1. Find a model for $E(X|X^*, Z)$ by validation data or replication
2. Replace the unobserved X by estimate $E(X|X^*, Z)$ in the main model
3. Adjust variance estimates by bootstrap or asymptotic methods

- Good method in many practical situations
- Calibration data can be incorporated
- Problems in highly nonlinear models

Regression calibration

- Berkson case: $E(X|X^*) = X^* \longrightarrow$
  Naive estimation = Regression calibration
- Classical : Linear regression X on X^*

$$E(X|X^*) \quad = \quad \frac{\sigma_x^2}{\sigma_{x^*}^2} * X^* + \mu_X * (1 - \frac{\sigma_x^2}{\sigma_{x^*}^2})$$

Correction for attenuation in linear model

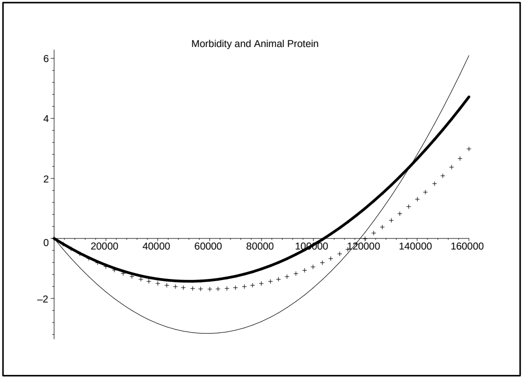
Survival

For Cox Model and rare disease assumption appropriate
Example: MONICA study, Augustin(2002)

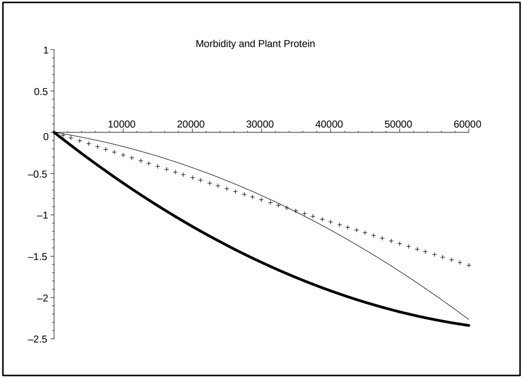
- CHD and fat intake
- Cox-Regression
- Quadratic model
- Classical additive measurement error
- Heteroscedastic measurement error
- Replication (7 days) for estimating measurement error variance

Results differ for assumption of homoscedastic and heteroscedastic measurement error

Results:Animal protein



Plant protein:



Likelihood methods

- Standard inference can be done with standard errors and likelihood ratio tests
- Efficiency
- Combination of different data types are possible
- Sometimes more accurate than approximations

- Difficult to calculate
- Software not available
- Parametric model for the unobserved predictor necessary
- Robustness to strong parametric assumptions

The classical error likelihood

Main model	$[Y \mid X, Z, \beta]$
Error model	$[X^* \mid X, \eta]$
Exposure model	$[X \mid Z, \lambda]$

$$[\mathbf{Y}, \mathbf{X}^* | \mathbf{Z}, \theta] = \prod_{i=1}^{n} \int [y_i | x_i, z_i, \beta][x_i^* | x_i, \eta][x_i | z_i, \lambda] d\mu(x),$$

where $\theta = (\beta, \eta, \lambda)$

- Evaluation by numerical integration

Berkson likelihood

3 components, but the third component contains no information

Main model	$[Y \mid X, Z, \beta]$
Error model	$[X \mid X^*, \eta]$
Exposure model	$[X^* \mid Z, \lambda]$,

$$[\mathbf{Y}, \mathbf{X}^* | \mathbf{Z}, \theta] = \prod_{i=1}^{n} \int [y_i | x_i, z_i, \beta][x_i | x_i^*, \eta][x_i | z_i, \lambda] d\mu(x)$$

$$[\mathbf{Y}, \mathbf{X}^* | \mathbf{Z}, \theta] = \prod_{i=1}^{n} \int [y_i | x_i, z_i, \beta][x_i | x_i^*, \eta] d\mu(x) * const$$

The Berkson likelihood does not depend on the exposure model

Quasi likelihood

If calculation of the likelihood is too complicated use

$$\begin{aligned} E(Y|X^*, Z) &= \int g(X, Z) f_{x|x^*} dx \\ V(Y|X^*, Z) &= \int v(X, Z) f_{x|x^*} dx + Var[g(X, Z)|X^*] \end{aligned}$$

This can be done e.g. for exponential g:

- Poisson regression model
- Parametric survival

Bayes

Richardson and Green (2002)

- Evaluation by MCMC techniques
- Conditional independence assumptions on the three models parts as seen in the likelihood approach
- The latent variable X is treated an unknown parameter
- Different data types can be combined
- Priori distributions for the error model
- Flexible handling of the exposure model

Case study : Occupational Dust and chronic bronchitis

HK/Carroll (1997) and Goessl /HK(2001)
Research question: Relationship between occupational dust and chronic bronchitis

Data form N=1246 workers:

X:	log(1+average occupational dust exposure)
Y:	Chronic bronchitis (CBR)
X^*:	Measurements and expert ratings
Z_1:	Smoking
Z_2:	Duration of exposure

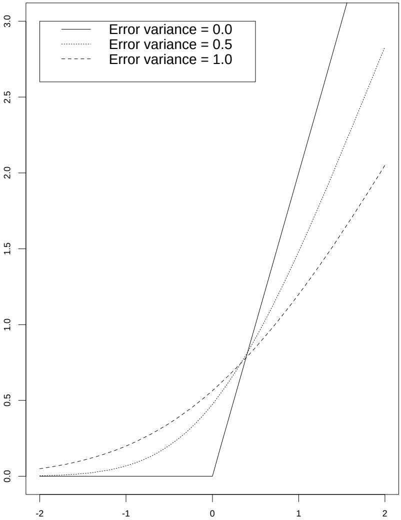
No validation or replication data available!

The Model

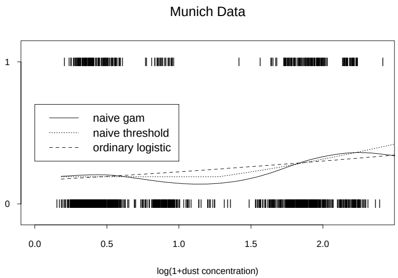
Segmented logistic regression an unknown threshold limiting value (TLV) τ

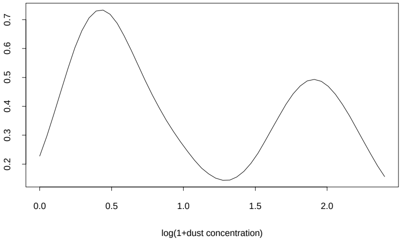
$$\begin{aligned} P(Y=1|X=x,Z=z) &= G(z'\beta_{k-}+\beta_k(x-\tau)_+), \\ (x-\tau)_+ &= \max(0,x-\tau). \end{aligned}$$

Effect of measurement error



Naive analysis





Likelihood

- Probit approximation
- Calculation of the integrals
- Assumption of a mixture of two normals for the exposure model
- Fixed additive measurement error

Regression calibration

- Assumption of a mixture of two normals for the exposure model
- Fixed additive measurement error

Results

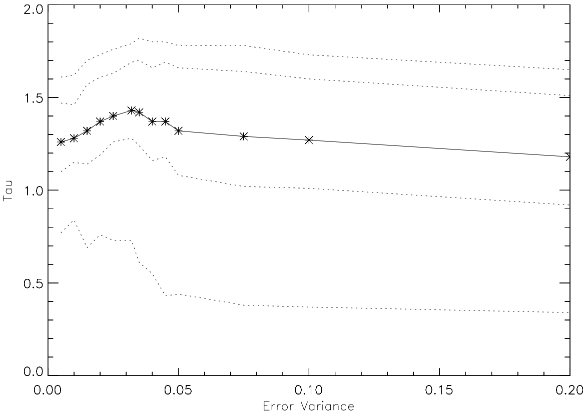
Method	TLV-τ_0	Nom s. e.	boot s.e.
Naive	1.27	.41	.24
Pseudo-MLE	1.76	.17	.21
Regression Calibration	1.75	.12	.19

Tabelle: *Estimated TLV in the Munich data, when* $\sigma_u^2 = 0.035$.

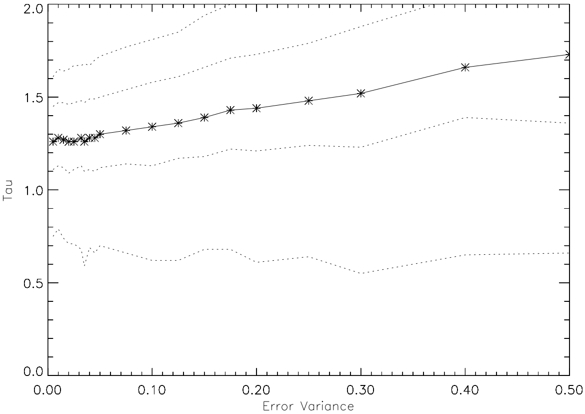
Bayes

- Fixed additive measurement error: Sensitivity analysis
- flat priori for measurement error :No convergence
- Assumption of mixture of normals for exposure model
- Both models Berkson and additive

Results: Estimation of TLV:Classical



Berkson



Conclusions

1. Measurement model essential: High Difference between Effect of Berkson and classical measurement error in most cases!
2. Additive classical non differential measurement error leads to attenuation
3. Many methods available
4. Regression calibration works in many cases
5. ML should be taken into account for Berkson error
6. Bayesian analysis is useful especially if model structure is complex