On quadratic logistic regression models when predictor variables are subject to measurement error

Jakub Stoklosa
School of Mathematics and Statistics,
Evolution & Ecology Research Centre.
The University of New South Wales,
Sydney, Australia.

Joint work with Wen-Han Hwang, Yih-Huei Huang and Elise Furlan

Motivation

- *Logistic regression* with predictor variables (or covariates) is used in a wide variety of applications.

- Such as: biostatistics, ecological, genomics, finance, etc.

- For example, in medical studies:
  - response variables are usually recorded as binary outcomes (*e.g.*, does a patient have diabetes); and
  - predictor variables are often recorded characteristics, attributes or measurements taken on patients (*e.g.*, age or the recorded body mass index values of each patients).

Motivation cont...

- When observed predictor variables are measured with error – *i.e.*, measured imprecisely, then there may be:
    - a loss of statistical power;
    - bias in parameters estimates; and
    - loss of features.
- So the analysis can lead to poor inference.

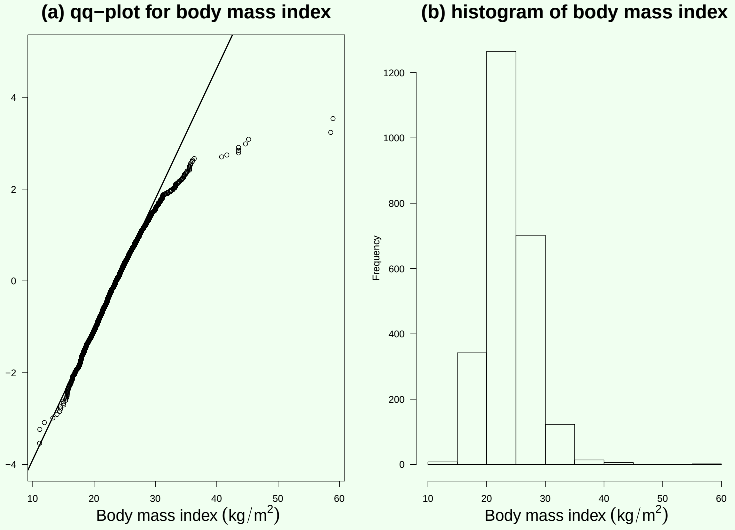
Motivation cont...

- Many *measurement error* models have been developed to account for error-in-predictor variables (Carroll *et al.*, 2006).
- For logistic regression, most of the literature has been primarily developed for parametric linear structures, and less so for quadratic structures.
- Existing methods that can incorporate quadratic models (*e.g.*, regression calibration or SIMEX) usually make the assumption that the distribution of true predictors is normal.

Motivation cont...

- In practice however, this assumption can be quite restrictive.
- Assuming normality on true predictors when in fact they are non-normal can lead to inconsistent parameter estimates.
- For example (see next slide).

Example: Body mass index data of diabetics in Taiwan



Aims

- Our aims are to develop new logistic regression models that:
  - take into account error-in-variables in predictors;
  - allow for quadratic models to be fit;
  - make less restrictive (or no) assumptions on the true predictor, hence leading to consistent estimation; and
  - are more computationally efficient compared to other methods.

Notation

- For $i = 1, \ldots, n$, let Y_i be a random sample of independent binary response variables.

- Let Z_i be categorical and X_i be a continuous covariate, write

$$P(Y_i = 1 \mid Z_i, X_i) = H(\alpha_1 + \alpha_2 Z_i + \beta_1 X_i + \beta_2 X_i^2)$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$ is the logistic function.

- The MLE of $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_1, \beta_2)$ is the root of the following score function:

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} S(\boldsymbol{\theta}, Y_i, Z_i, X_i) = \sum_{i=1}^{n} (Z_i^{\mathsf{T}}, X_i^{\mathsf{T}})^{\mathsf{T}} \{Y_i - H(\boldsymbol{\theta}, X_i, Z_i)\}.$$

Classical measurement error and näive method

- Now, suppose that X_i is measured with *additive* random error and we only have the observed *surrogate* variable W_i.

- We assume that $W_i = X_i + \epsilon_i$ for all i, where $\epsilon_i \sim N(0, \sigma^2)$ is the *measurement error* independent of X_i, Z_i and Y_i.

- If $\sigma^2 > 0$, the näive method replaces X_i by W_i and solves

$$G_N(\boldsymbol{\theta}) = \sum_{i=1}^{n} S(\boldsymbol{\theta}, Y_i, Z_i, W_i) = 0. \tag{1}$$

  Generally, $\mathrm{E}\{G_N(\boldsymbol{\theta})\} \neq 0$, which result in biased $\boldsymbol{\theta}$.

Regression Calibration (RC)

- *Regression calibration (RC)* is a convenient approximation method commonly used to adjust for bias.

- Briefly, the RC method replaces W_i and W_i^2 by the following *Best linear Unbiased Estimators*: $\mathrm{E}(X_i \mid W_i)$ and $\mathrm{E}(X_i^2 \mid W_i)$ in the estimating equation (1), respectively.

- If $X_i \sim N(\mu_x, \sigma_x^2)$, the above conditional expectations can be easily calculated

- However, the RC may yield a considerable amount of bias when either σ^2 or β are moderately large.

Refined Regression Calibration (RRC)

- The bias can be reduced by refining the approximation for $\mathrm{E}\{H(\boldsymbol{\theta}, X_i, Z_i) \mid Z_i, W_i\}$.

- Known as *refined regression calibration (RRC)*.

- Specifically, we apply a simple logit-to-normal approximation, and we can show that

$$\begin{aligned}\mathrm{E}\left\{H(\boldsymbol{\theta}, X_i, Z_i) \mid Z_i, W_i\right\} &\approx \mathrm{E}\{\Phi(c\alpha Z_i + c\beta^{\mathsf{T}} X_i) \mid Z_i, W_i\} \\ &\approx H\left\{\frac{\alpha Z_i + \beta^{\mathsf{T}}\mathrm{E}(X_i \mid Z_i, W_i)}{\sqrt{1 + c^2 \mathrm{Var}(\beta^{\mathsf{T}} X_i \mid Z_i, W_i)}}\right\}\end{aligned}$$

  where $c = 1/1.7$ is a constant, see Johnson *et al.* (1995).

Refined Regression Calibration cont...

- Again assuming that $X_i \sim N(\mu_x, \sigma_x^2)$, and with some algebra we can find $\mathrm{E}(X_i \mid Z_i, W_i)$ and $\mathrm{Var}(\beta^{\mathsf{T}} X_i \mid Z_i, W_i)$.

- We let

$$\tilde{p}_i(\boldsymbol{\theta}) = H\left\{\frac{\alpha Z_i + \beta^{\mathsf{T}}\mathrm{E}(X_i \mid Z_i, W_i)}{\sqrt{1 + c^2 \mathrm{Var}(\beta^{\mathsf{T}} X_i \mid Z_i, W_i)}}\right\},$$

and estimate $\boldsymbol{\theta}$ by solving the usual estimating equation.

- But, both RC and RRC need normality assumptions on X_i.

Weighted Corrected Score (WCS)

- So, can we avoid making normality assumption on X_i but still obtain consistent and asymptotically normal estimators?
- An alternative approach is to seek out a "correctable" weighted score function.
- That is, for $i = 1, \ldots, n$, let ω_i be weights so that

$$S_\omega(\boldsymbol{\theta}, Y_i, Z_i, X_i) = \omega_i S(\boldsymbol{\theta}, Y_i, Z_i, X_i)$$

  is an unbiased estimating equation.

Weighted Corrected Score cont...

- Recently, Chen *et al.* (2015) showed that there exists a $S^*_\omega(\boldsymbol{\theta}, Y_i, Z_i, W_i)$, such that

$$\mathrm{E}\left\{S^*_\omega(\boldsymbol{\theta}, Y_i, Z_i, W_i) \mid Z_i, X_i\right\} = S_\omega(\boldsymbol{\theta}, Y_i, Z_i, X_i)$$

  yields consistent and asymptotically normal estimators.

- Chen *et al.* (2015) only considered linear logistic regression.

- We develop similar estimators (or weighted score functions) but specifically for quadratic models.

Weighted Corrected Score cont...

- **Required condition:** Provided that $|\beta_2\sigma^2| < 1$ holds, then we can show the existence of S^*_ω.

- We refer to this as a *weighted corrected score (WCS)* function:

$$G^*_\omega(\boldsymbol{\theta}) = \sum_{i=1}^{n} S^*_\omega(\boldsymbol{\theta}, Y_i, Z_i, W_i)$$

  where $S^*_\omega = ({S^*_{\omega 1}}^T, S^*_{\omega 2}, S^*_{\omega 3})^\mathsf{T}$; the first component is a 2×1 vector (due to Z_i) and the latter two are both scalars.

- These weights were trickier to calculate (see next slide), but we now have estimators that are consistent and asymptotically normal, and allow for quadratic structures.

Weighted Corrected Score cont...

- For $j = 1, 2$, we define $D_j = 1 + (-1)^j \beta_2 \sigma^2$ and

$$C_j(\boldsymbol{\theta}, Y_i, Z_i, W_i) = \exp\left\{(-1)^j \frac{1}{2} \boldsymbol{\alpha} Z_i + (-1)^j \frac{1}{2} \frac{\boldsymbol{\beta}^{\mathsf{T}} W_i}{D_j} - \frac{1}{8} \frac{\beta_1^2 \sigma^2}{D_j}\right\}.$$

- The three components of S^*_ω are given as follows:

$$\begin{aligned}
S^*_{\omega 1}(\boldsymbol{\theta}, Y_i, Z_i, W_i) &= Z_i \left\{\frac{Y_i C_1(\boldsymbol{\theta}, Y_i, Z_i, W_i)}{\sqrt{D_1}} + \frac{(Y_i - 1) C_2(\boldsymbol{\theta}, Y_i, Z_i, W_i)}{\sqrt{D_2}}\right\}, \\
S^*_{\omega 2}(\boldsymbol{\theta}, Y_i, Z_i, W_i) &= \left\{\frac{W_i}{\sqrt{D_1^3}} + \frac{\beta_1 \sigma^2}{2\sqrt{D_1}}\right\} Y_i C_1(\boldsymbol{\theta}, Y_i, Z_i, W_i) \\
&\quad + \left\{\frac{W_i}{\sqrt{D_2^3}} - \frac{\beta_1 \sigma^2}{2\sqrt{D_2}}\right\} (Y_i - 1) C_2(\boldsymbol{\theta}, Y_i, Z_i, W_i), \\
S^*_{\omega 3}(\boldsymbol{\theta}, Y_i, Z_i, W_i) &= \left\{\frac{W_i^2 + \beta_1 W_i \sigma^2 + \frac{1}{4}\beta_1^2 \sigma^4}{\sqrt{D_1^5}} - \frac{\sigma^2}{\sqrt{D_1^3}}\right\} Y_i C_1(\boldsymbol{\theta}, Y_i, Z_i, W_i) \\
&\quad + \left\{\frac{W_i^2 - \beta_1 W_i \sigma^2 + \frac{1}{4}\beta_1^2 \sigma^4}{\sqrt{D_2^5}} - \frac{\sigma^2}{\sqrt{D_2^3}}\right\} (Y_i - 1) C_2(\boldsymbol{\theta}, Y_i, Z_i, W_i).
\end{aligned}$$

Simulations: Finite sample performance

- We considered two scenarios where the true distribution for X was set to the following: (1) $X \sim N(0,1)$; and (2) $X \sim (\chi^2_3 - 3)/\sqrt{6}$;

- We simulated measurement error $\epsilon \sim \mathrm{N}(0, \sigma^2)$ to get $W = X + \epsilon$.

- For both scenarios above we set: $\sigma^2 = 0.30$, $n = 200, 1000$ and true parameter values: $\boldsymbol{\theta} = (0.50, 1, -0.30)$.

- We then generated Y and fit the näive model and four logistic regression (measurement error) models for each scenario.

Simulation scenario 1: $X \sim N(0, 1)$

- For further comparison, we also included another consistent method called the extensively corrected score (ECS, Huang *et al.*, 2015).

scenario 1	$\beta_1 = 1$					$\beta_2 = -0.30$				
method	Mean	SD	SE	RMSE	CP	Mean	SD	SE	RMSE	CP
näive	0.73	0.16	0.16	0.65	0.57	-0.15	0.11	0.11	0.78	0.69
RC	0.95	0.21	0.21	0.80	0.92	-0.26	0.20	0.18	0.88	0.92
RRC	1.04	0.26	0.26	0.88	0.95	-0.32	0.26	0.23	0.93	0.96
ECS	1.12	0.40	0.42	0.99	0.96	-0.38	0.48	0.40	1.07	0.96
WCS	1.12	0.33	0.29	0.96	0.93	-0.41	0.36	0.30	1.05	0.91

Table: *Estimates, RMSE and 95% coverage (CP) for* $n = 200$.

scenario 1	$\beta_1 = 1$					$\beta_2 = -0.30$				
method	Mean	SD	SE	RMSE	CP	Mean	SD	SE	RMSE	CP
näive	0.73	0.07	0.07	0.63	0.03	-0.15	0.05	0.05	0.77	0.11
RC	0.94	0.09	0.09	0.77	0.90	-0.25	0.08	0.08	0.85	0.89
RRC	1.01	0.11	0.11	0.82	0.95	-0.30	0.10	0.10	0.88	0.95
ECS	1.05	0.15	0.15	0.86	0.96	-0.33	0.13	0.13	0.91	0.97
WCS	1.04	0.13	0.12	0.84	0.94	-0.33	0.12	0.11	0.91	0.93

Table: *Estimates, RMSE and 95% coverage (CP) for* $n = 1000$.

Simulation scenario 2: $X \sim (\chi^2_3 - 3)/\sqrt{6}$

scenario 2	$\beta_1 = 1$					$\beta_2 = -0.30$				
method	Mean	SD	SE	RMSE	CP	Mean	SD	SE	RMSE	CP
näive	0.58	0.18	0.17	0.59	0.27	-0.11	0.11	0.09	0.75	0.38
RC	0.76	0.24	0.22	0.69	0.77	-0.19	0.20	0.15	0.82	0.87
RRC	0.80	0.28	0.26	0.73	0.80	-0.21	0.25	0.18	0.84	0.91
ECS	1.11	0.60	0.54	1.05	0.95	-0.35	0.35	0.32	0.96	0.97
WCS	1.15	0.47	0.39	1.02	0.92	-0.39	0.37	0.25	1.01	0.93

Table: *Estimates, RMSE and 95% coverage (CP) for $n = 200$.*

scenario 2	$\beta_1 = 1$					$\beta_2 = -0.30$				
method	Mean	SD	SE	RMSE	CP	Mean	SD	SE	RMSE	CP
näive	0.56	0.07	0.07	0.56	0.00	-0.13	0.03	0.03	0.75	0.00
RC	0.73	0.10	0.10	0.64	0.19	-0.21	0.06	0.06	0.82	0.62
RRC	0.75	0.11	0.11	0.65	0.36	-0.22	0.06	0.06	0.82	0.75
ECS	1.03	0.20	0.19	0.84	0.96	-0.32	0.10	0.09	0.91	0.96
WCS	1.01	0.16	0.16	0.83	0.95	-0.31	0.08	0.07	0.89	0.95

Table: *Estimates, RMSE and 95% coverage (CP) for $n = 1000$.*

Case Study: Diabetes survey data

- First, we obtained an approximate value for σ^2 using validation data.
- We then fitted each model using body mass index as a covariate with quadratic terms.

method	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
näive	-4.26 (1.13)	0.18 (0.08)	-0.00271 (0.00167)
RC	-5.02 (1.33)	0.23 (0.09)	-0.00370 (0.00181)
RRC	-5.06 (1.36)	0.24 (0.10)	-0.00375 (0.00185)
ECS	-4.92 (1.33)	0.22 (0.09)	-0.00341 (0.00173)
WCS	-4.86 (1.40)	0.22 (0.10)	-0.00345 (0.00194)

Table: *Estimates and standard errors (in parentheses) for each method.*

Conclusion and Further Work

- Two new methods (RRC and WCS) for quadratic logistic regression models were comparable (and in some cases better) than known methods.
- However, some additional conditions were still needed.
- We could consider quadratic Berkson (measurement) error models.
- We could also try to extend these methods to other link functions *e.g.*, `probit` or log-linear Poisson models.

Bibliography

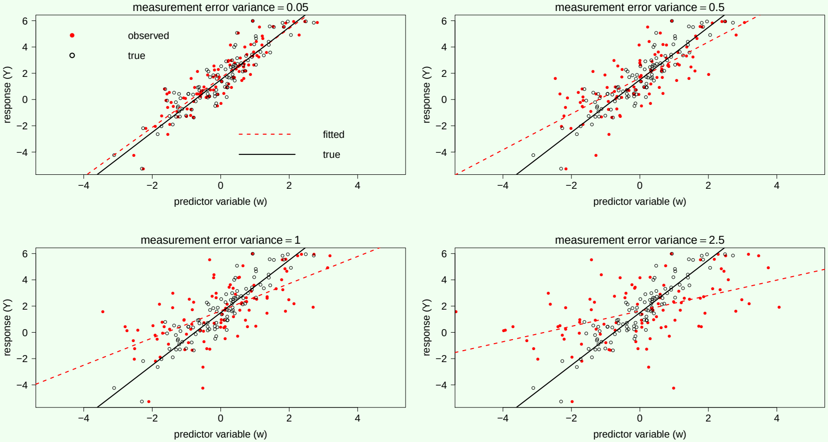
Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* 2nd Ed. London: Chapman & Hall/CRC.

Chen, J., Hanfelt, J. J., and Huang, Y. (2015). A simple corrected score for logistic regression with errors-in-covariates. *Communications in Statistics, Series A - Theory and Methods* **44**, pp. 2024–2036.

Huang, Y. H., Wen, C. C., and Hsu, Y. H. (2015). The extensively corrected score for measurement error models. *Scandinavian Journal of Statistics* **42**, pp. 911–924.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions, Volume 2, 2nd edition*. New York: Wiley.

Simulation example of bias

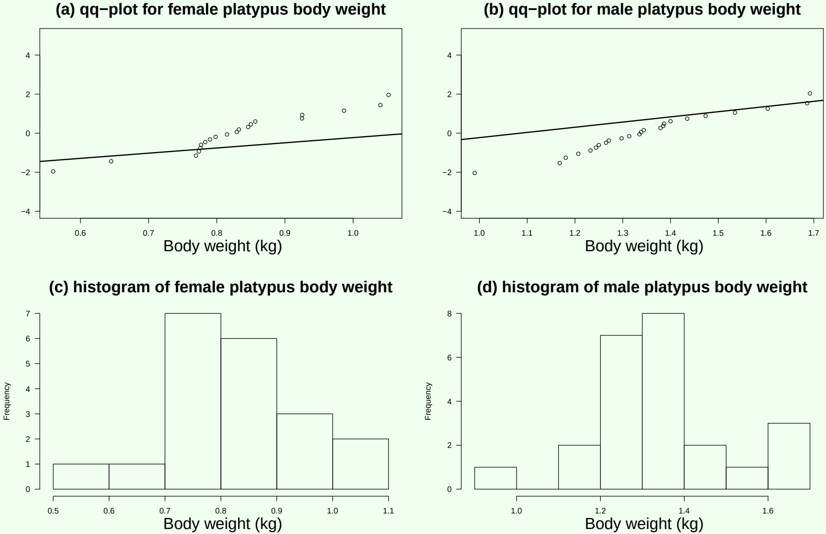


Regression Calibration cont. . .

- It follows that the conditional expectations are: $\mathrm{E}(X_i \mid W_i) = \mu_{x_i|w_i}$ and $\mathrm{E}(X_i^2 \mid W_i) = \sigma^2_{x_i|w_i} + \mu^2_{x_i|w_i}$ where $\mu_{x_i|w_i} = \mu_x + \frac{\sigma_x^2}{\sigma_w^2}(W_i - \mu_w)$, $\sigma^2_{x_i|w_i} = \frac{\sigma^2}{\sigma_w^2}\sigma_x^2$ and $\sigma_w^2 = \sigma_x^2 + \sigma^2$.

- Note that $\mu_x = \mu_w$, such that μ_x can be estimated by $\overline{W}$, and since σ^2 is given then $\sigma_x^2 = \sigma_w^2 - \sigma^2$ can be similarly estimated.

Example 2: Platypus body weight for males and females



Case Study 2: Platypus capture–recapture data

- Here, the main interest was estimating capture probabilities for each gender type using body weight as a covariate.

method	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
naive	-8.58 (4.80)	-0.47 (1.01)	12.32 (8.23)	-5.34 (3.12)
RC	-8.79 (5.60)	-0.44 (1.05)	12.70 (9.60)	-5.53 (3.66)
RRC	-9.17 (6.19)	-0.45 (1.09)	13.38 (10.63)	-5.82 (4.08)
ECS	-10.77 (10.04)	-0.66 (2.29)	16.38 (17.91)	-7.04 (7.43)
WCS	-11.87 (6.71)	-0.91 (1.47)	18.22 (11.54)	-7.66 (4.24)

Table: *Estimates and non-parametric bootstrap standard errors (in parentheses) for each method. Note that* $\widehat{\alpha}_2$ *here is the gender effect.*