

Robustifying Logistic Regression for Nonresponse: An Application to Body Mass Index

William Marjerison
2005-2006 Major Qualifying Project
Department of Mathematical Sciences

The 2006 CIMS MQP Honorable Mention

Abstract -

We predict finite population mean (fpm) body mass index (BMI) of children and adolescents of age-race-sex domains. There are many nonrespondents, and no distribution is assumed for BMI. We study logistic and student's t (variable degrees of freedom) link functions for response indicators in Bayesian regression analysis using the Metropolis sampler. We impute nonrespondents's BMI by placing nonrespondents into cells based on propensity scores. Predictive inference is done using least squares, and we make comparison with a recent method.

1. Introduction

Robustness: Distributional robustification allows the possibility of obtaining similar results as logistic regression using a model with less assumptions about underlying distributions

Project Goals

- **Estimation:** Obtain robust estimates of regression coefficients in logistic regression.
- **Imputation:** Use robust method to fill in body mass index (BMI) for survey nonrespondents given the covariates age, race, and sex.
- **Prediction:** Use another robust method to predict finite population mean BMI nationally for domains formed by age, race, and sex.

■ BMI data comes from the 3rd National Health and Nutrition Examination Survey (NHANES III) by the National Center for Health Statistics (NCHS). Covariates are age (2 to 19 years), race (black or other), sex (male or female). 5185 respondents and 1606 nonrespondents from 35 counties.

■ Nandram & Choi [1] used linear spline regression of BMI on age adjusting for race and sex to predict finite population mean BMI.

Two regressions: logistic and linear

- Student-t(9) cdf is closer to logistic cdf than to normal cdf [2]
- If A and B are logistic and student-t(9) random variables, respectively, the first four moments of A and B/γ are the same, where $\gamma = \pi\sqrt{7/27}$

Model

Bayesian Logistic Regression:

$$\begin{cases} Y_i | \beta \sim \text{Bernoulli}(L(x_i' \beta)) \\ \pi(\beta) = 1, \beta \in R^p \\ \text{Propensity Score} = L(x_i' \beta) \end{cases}$$

Robustified model (mixture of student-t distributions):

$$\begin{cases} Y_i | \beta, \nu \sim \text{Bernoulli}(\pi(x_i' \beta \div \gamma)) \\ \pi(\beta) = 1, \beta \in R^p \\ \nu \sim \pi(\nu), \nu \geq 0 \\ \text{Propensity Score} = \pi(x_i' \beta \div \gamma) \end{cases}$$

2. Posterior Property

■ The marginal posterior distribution of β is

$$\pi(\beta | y) \propto \int \prod_{i=1}^n [\pi(x_i' \beta \div \gamma)]^{y_i} [1 - \pi(x_i' \beta \div \gamma)]^{1-y_i} \pi(\nu) d\nu$$

■ To make inferences, we need to check that the posterior distribution is proper, i.e., the normalization constant

$$\int \int \prod_{i=1}^n [\pi(x_i' \beta \div \gamma)]^{y_i} [1 - \pi(x_i' \beta \div \gamma)]^{1-y_i} \pi(\nu) d\nu d\beta < \infty$$

Dual Model [3]

$$\begin{cases} Y_i | p_i \sim \text{Bernoulli}(p_i) \\ \pi(\beta) = 1, \beta \in R^p \\ \nu \sim \pi(\nu), \nu \geq 0 \\ Z_i | \lambda_i \sim N(0, \lambda_i^{-1}) \\ \lambda_i | \nu \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}) \\ Y_i = \begin{cases} 1, & \text{if } Z_i \geq -x_i' \beta \\ 0, & \text{else} \end{cases} \end{cases}$$

Theorem: Property of Dual Model [4]

- If we place a proper prior on β then the support of ν must contain 9 so that student-t(9) approximation to logistic distribution is valid.
- The number of regression coefficients (including the intercept) can be at most 8.

If the design matrix $X = (X_{ij}), i = 1, \dots, n, j = 1, \dots, p$ is full column rank and satisfies condition C2 of Chen and Shao (1999), and $\pi(\nu)$ has support $\nu > p$, then

$$I = \int \int \int \int \pi(\nu) \times \prod_{i=1}^n I(Z_i \geq -x_i' \beta) I(Y_i = 1) + I(Z_i < -x_i' \beta) I(Y_i = 0) \\ \times \frac{\nu^{\frac{p}{2}}}{\Gamma(\frac{p}{2})} \lambda_i^{\frac{p}{2}-1} \exp(-\frac{p}{2} \lambda_i) \\ \times \sqrt{\frac{\nu}{2\pi}} \exp(-\frac{1}{2} z_i^2) d\nu d\beta d\lambda_i dZ_i < \infty.$$

The normalized constant is finite, so the posterior distribution is proper.

3. Computation

On computer, our hierarchical model is implemented as

$$\begin{cases} Y_i | \beta, \nu \sim \text{Bernoulli}(\pi(x_i' \beta \div \gamma)) \\ \pi(\beta) = N(\theta_0, \Delta_0) \\ \pi(\nu = a_r) = w_r, r = 1, \dots, 6 \\ \gamma = \pi\sqrt{7/27} \end{cases}$$

where ν takes values of $a_1 = 3, a_2 = 6, a_3 = 9$ (approximately logistic), $a_4 = 12, a_5 = 25, a_6 = 50$ (approximately normal), w_r are specified to consider different degrees of robustness, θ_0, Δ_0 are the posterior mean and the posterior variance (inflated by 1000) of β from logistic regression.

Robustification:

$$\begin{cases} P(Y_i = y | \beta) = \sum_{r=1}^6 w_r [\pi(x_i' \beta \div \gamma)]^y [1 - \pi(x_i' \beta \div \gamma)]^{1-y}, y = 0,1 \\ \text{and} \\ \pi(\beta | y) \propto \left\{ \prod_{i=1}^n P(Y_i = y_i | \beta) \right\} N(\beta | \theta_0, \Delta_0) \end{cases}$$

Monitoring the Metropolis-Hastings (MH) Algorithm

- Look at autocorrelation in the MH algorithm, making sure that autocorrelation is reasonably close to zero.
- Proposal density: mean given by maximizing posterior density using Nelder-Mead algorithm; matrix parameter given by evaluating negative inverse Hessian matrix.
- We adjust the degrees of freedom of our proposal density (multivariate student-t with κ degrees of freedom) so that the jump probability in the MH algorithm is between 0.25 and 0.50.

4. Data Analysis

- In each iteration of the MH algorithm, for each respondent and nonrespondent, we draw β from its marginal posterior distribution. Next, draw ν from its conditional posterior density given β and calculate the propensity score.
- We construct "cells" based on the quintiles of the propensity scores of all respondents and nonrespondents.
- We assign BMIs to nonrespondents based on the (discrete) distribution of BMIs of respondents in the same cell.

95% Credible Intervals for Logistic Regression Coefficients

Model	Intercept	Age	Race	Sex	Race*Sex
0.50 as 9 df	(1.45, 1.88)	(0.97, 1.41)	(-0.31, -0.10)	(-0.05, 0.05)	(-0.06, 0.03)
0.70 as 9 df	(1.44, 1.85)	(0.96, 1.39)	(-0.24, -0.10)	(-0.07, 0.06)	(-0.09, 0.04)
0.90 as 9 df	(1.43, 1.82)	(0.97, 1.40)	(-0.31, -0.18)	(-0.07, 0.06)	(-0.08, 0.04)
0.995 as 9 df	(1.45, 1.86)	(0.97, 1.40)	(-0.31, -0.18)	(-0.07, 0.06)	(-0.08, 0.05)
Student-t(9)	(1.47, 1.82)	(0.98, 1.41)	(-0.31, -0.18)	(-0.07, 0.06)	(-0.08, 0.05)
Logistic	(1.46, 1.61)	(0.99, 1.16)	(-0.30, -0.18)	(-0.07, 0.05)	(-0.08, 0.05)

Note 1: Age and race discriminate between respondents and nonrespondents; sex and the interaction between race and sex are not important.

Note 2: Posterior means are very similar, but posterior standard deviations are slightly larger in the more robust models.

Predicting Finite Population Mean BMI

- Least squares regression of log BMI onto age, race, sex, race*sex; errors iid with no distributional assumption. BF = black female, BM = black male, OF = other female, OM = other male

R-RS	Model	age				10-14				15-19			
		2-4	5-9	10-14	15-19	2-4	5-9	10-14	15-19	2-4	5-9	10-14	15-19
L	Logistic	(15.05, 16.66)	(17.65, 18.56)	(19.75, 20.81)	(22.03, 23.11)	(15.05, 16.66)	(17.65, 18.56)	(19.75, 20.81)	(22.03, 23.11)	(15.05, 16.66)	(17.65, 18.56)	(19.75, 20.81)	(22.03, 23.11)
	Student-9	(15.05, 16.66)	(17.65, 18.57)	(19.75, 20.82)	(22.03, 23.42)	(15.05, 16.66)	(17.64, 18.55)	(19.75, 20.78)	(22.03, 23.30)	(15.05, 16.65)	(17.64, 18.56)	(19.74, 20.81)	(22.02, 23.40)
	0.70 as 9 df	(15.05, 16.65)	(17.64, 18.55)	(19.75, 20.78)	(22.03, 23.30)	(15.05, 16.65)	(17.64, 18.56)	(19.75, 20.78)	(22.03, 23.30)	(15.05, 16.65)	(17.64, 18.56)	(19.74, 20.81)	(22.02, 23.40)
	0.90 as 9 df	(15.05, 16.65)	(17.64, 18.56)	(19.75, 20.78)	(22.03, 23.42)	(15.05, 16.65)	(17.64, 18.56)	(19.75, 20.78)	(22.03, 23.42)	(15.05, 16.65)	(17.64, 18.56)	(19.74, 20.81)	(22.02, 23.40)
	0.995 as 9 df	(15.05, 16.58)	(17.64, 18.56)	(19.83, 20.82)	(22.30, 23.42)	(15.05, 16.58)	(17.64, 18.56)	(19.83, 20.82)	(22.30, 23.42)	(15.05, 16.58)	(17.64, 18.56)	(19.83, 20.82)	(22.30, 23.42)
M	Logistic	(15.20, 16.19)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.81)	(15.20, 16.19)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.81)	(15.20, 16.19)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.81)
	Student-9	(15.20, 16.19)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.82)	(15.20, 16.19)	(17.23, 18.00)	(19.26, 20.27)	(21.48, 22.82)	(15.20, 16.19)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.82)
	0.70 as 9 df	(15.20, 16.20)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.81)	(15.20, 16.20)	(17.23, 18.00)	(19.26, 20.27)	(21.48, 22.81)	(15.20, 16.20)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.81)
	0.90 as 9 df	(15.20, 16.20)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.81)	(15.20, 16.20)	(17.23, 18.00)	(19.26, 20.27)	(21.48, 22.81)	(15.20, 16.20)	(17.23, 18.00)	(19.26, 20.28)	(21.48, 22.81)
	0.995 as 9 df	(15.25, 16.12)	(17.22, 18.05)	(19.34, 20.27)	(21.64, 22.83)	(15.25, 16.12)	(17.22, 18.05)	(19.34, 20.27)	(21.64, 22.83)	(15.25, 16.12)	(17.22, 18.05)	(19.34, 20.27)	(21.64, 22.83)
O	Logistic	(15.38, 17.05)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.07)	(15.38, 17.05)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.07)	(15.38, 17.05)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.07)
	Student-9	(15.37, 17.05)	(17.29, 18.50)	(19.41, 21.03)	(21.71, 23.25)	(15.37, 17.05)	(17.29, 18.49)	(19.40, 21.03)	(21.71, 23.25)	(15.37, 17.05)	(17.29, 18.50)	(19.41, 21.03)	(21.71, 23.25)
	0.70 as 9 df	(15.37, 17.05)	(17.29, 18.49)	(19.40, 21.03)	(21.71, 23.25)	(15.37, 17.05)	(17.29, 18.50)	(19.41, 21.03)	(21.71, 23.25)	(15.37, 17.05)	(17.29, 18.50)	(19.41, 21.03)	(21.71, 23.25)
	0.90 as 9 df	(15.39, 17.00)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.51)	(15.39, 17.00)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.51)	(15.39, 17.00)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.51)
	0.995 as 9 df	(15.37, 16.90)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.00)	(15.37, 16.90)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.00)	(15.37, 16.90)	(17.30, 18.50)	(19.41, 21.03)	(21.72, 23.00)
S	Logistic	(16.01, 16.60)	(17.77, 18.54)	(19.62, 20.76)	(21.61, 23.28)	(16.01, 16.60)	(17.77, 18.54)	(19.62, 20.76)	(21.61, 23.28)	(16.01, 16.60)	(17.77, 18.54)	(19.62, 20.76)	(21.61, 23.28)
	Student-9	(15.39, 16.79)	(17.31, 18.65)	(19.42, 20.80)	(21.73, 23.27)	(15.39, 16.79)	(17.31, 18.64)	(19.41, 20.79)	(21.72, 23.27)	(15.39, 16.79)	(17.31, 18.64)	(19.42, 20.79)	(21.72, 23.27)
	0.70 as 9 df	(15.41, 16.81)	(17.33, 18.67)	(19.43, 20.82)	(21.74, 23.29)	(15.41, 16.81)	(17.33, 18.67)	(19.43, 20.82)	(21.74, 23.29)	(15.41, 16.81)	(17.33, 18.67)	(19.43, 20.82)	(21.74, 23.29)
	0.90 as 9 df	(15.39, 16.79)	(17.31, 18.64)	(19.41, 20.80)	(21.72, 23.27)	(15.39, 16.79)	(17.31, 18.64)	(19.41, 20.80)	(21.72, 23.27)	(15.39, 16.79)	(17.31, 18.64)	(19.41, 20.80)	(21.72, 23.27)
	0.995 as 9 df	(16.46, 16.74)	(17.74, 18.53)	(19.58, 20.80)	(21.13, 22.57)	(16.46, 16.74)	(17.74, 18.53)	(19.58, 20.80)	(21.13, 22.57)	(16.46, 16.74)	(17.74, 18.53)	(19.58, 20.80)	(21.13, 22.57)

5. Conclusion

- Our robust mixture models give similar estimated regression coefficients as the Bayesian logistic regression, but with somewhat larger posterior variance.
- Our robust mixture models give estimates for finite population mean BMI that are virtually identical to Bayesian logistic regression without compromising precision and similar to [1].

References

1. B. Nandram and J.W. Choi, Hierarchical Bayesian nonignorable nonresponse regression models for small areas: an application to the NHANES data, *Survey Methodology*, 31, 73-84, 2005.
2. G.S. Mudholkar and E.O. George, A remark on the shape of the logistic distribution, *Biometrika*, 65, 667-668, 1978.
3. J.H. Albert and S. Chib, Bayesian analysis of binary and polychotomous response data, *J. Amer. Stat. Assoc.*, 88, 669-679, 1993.
4. M. Chen and Q. Shao, Propriety of posterior distribution for dichotomous quantal response models, *Proc. AMS*, 129, 293-302, 1999.

Acknowledgments

Special thanks to:

- ✓ Prof. William Martin
- ✓ Dr. Jai Won Choi, National Center for Health Statistics/CDC

Advisor: Prof. Balgobin Nandram

