An Unbiased Estimation Method for Errors-in-Variables Logistic Regression Model

Sen Yang[*] & Qinxiao Shi[†]

16 December 2018

Abstract

It is well known that if independent variables are measured with erorrs, the ordinary least squared estimates are not unbiased. This is also true for logistic regression models. In this project, we are seeking an unbiased estimation method for logistic regression model. By building up an non-linear equation system of $\mathbb{E}\{\widetilde{\boldsymbol{x_i}}[y_i - \mu_i(\widetilde{\boldsymbol{x_i}})] - B\} = \mathbf{0}$, we are able to get an unbiased estimation by solving the equation system of $\boldsymbol{\beta}$. A simlation study is followed up to validate the improvement of estimation.

Contents

[*]sen.2.yang@uconn.edu; M.S. student at Department of Statistics, University of Connecticut.

[†]qinxiao.shi@uconn.edu; M.S. student at Department of Statistics, University of Connecticut.

1 Introduction

In statistics, errors-in-variables models or measurement error models are regression models that account for measurement errors in the independent variables. In the case when some regressors have been measured with errors, estimation based on the standard assumption leads to inconsistent estimates, meaning that the parameter estimates do not tend to the true values even in very large samples.

For simple linear regression the effect is an underestimate of the coefficient, known as the attenuation bias. It is well known that if independent variables are measured with erorrs, the ordinary least squared estimates are not unbiased. This is also true for logistic regression models.

With a canonical link function, a logistic regression model will be fomulated as:

$$Y_i = \frac{1}{1+\exp(-\boldsymbol{x'\beta})} + \epsilon_i$$

where Y_i denotes the binomial proportion (for a binomial regression model), ϵ_i denotes the error of response, $\boldsymbol{x}$ denotes independent variables and $\boldsymbol{\beta}$ denotes the parameters for the model.

If there is measurement errors for indenpendent variables in a logistic regression model, the estimates are not unbiased in terms of response and predicted response. This project will introduce an unbiased estimation method to remedy attenuation bias so that we could get a better estimates of parameters for logistic regression models with observation errors.

2 Model

2.1 Errors-in-Variables Logistic Regression Model

Suppose that for $i = 1, ..., n$, $Z_i \sim Binomial(m_i, p_i)$, and let $Y_i = \frac{Z_i}{m_i}$ denote the binomial proportion for the i^{th} case, where m_i is known and p_i depends on the vector of covariates $\boldsymbol{x_i} = (1, X_{i1}, X_{i2}, ..., X_{ik})'$. Let z_i be the observed binomial response taking values $0, 1, ..., m_i$ with $y_i = z_i/m_i$.

The regression model for a binomial proportion is

$$\begin{gathered} Z_i \mid p_i \sim Bin(m_i, p_i) \\ logit(p_i) = \eta_i = \boldsymbol{x_i}'\boldsymbol{\beta} \end{gathered}$$

Suppose $\widetilde{\boldsymbol{x_i}}$ is the vector of observed variables and $\boldsymbol{x_i}$ is the vector of latent or true variables. Let $\boldsymbol{u_i}$ be the vector of errors of observation such that,

$$\widetilde{\boldsymbol{x_i}} = \boldsymbol{x_i} + \boldsymbol{u_i}$$

For each $i = 1, ..., n$, suppose we make $\boldsymbol{m}$ measurements for covariates $\boldsymbol{x_i}$, then corresponding matrices for observations, true values and errors are

$$\widetilde{\mathbf{X_i}} = \begin{bmatrix} \widetilde{\boldsymbol{x_{i1}}}' \\ \widetilde{\boldsymbol{x_{i2}}}' \\ \vdots \\ \widetilde{\boldsymbol{x_{im}}}' \end{bmatrix} \qquad \mathbf{X_i} = \begin{bmatrix} \boldsymbol{x_i}' \\ \boldsymbol{x_i}' \\ \vdots \\ \boldsymbol{x_i}' \end{bmatrix} \qquad \boldsymbol{U_i} = \begin{bmatrix} \boldsymbol{u_{i1}}' \\ \boldsymbol{u_{i2}}' \\ \vdots \\ \boldsymbol{u_{im}}' \end{bmatrix}$$

where $\widetilde{\boldsymbol{x_{ij}}}$ is the observation vector and $\boldsymbol{x_i}$ is the true value vector, and $\boldsymbol{u_{ij}}$ is the corresponding error vector for $j = 1, 2, ..., m$ such that

$$\widetilde{\mathbf{X_i}} = \mathbf{X_i} + \mathbf{U_i}$$

Here, we assume that $\boldsymbol{u_{i1}}, \boldsymbol{u_{i2}}, ..., \boldsymbol{u_{im}} \sim \mathcal{N}_k(\mathbf{0}, \Sigma_i)$ $i.i.d.$ and $\boldsymbol{u_i}$ is independent of $\boldsymbol{x_i}$. The variance-corariance matrix of all errors will be

$$\Sigma_u = \begin{bmatrix} \Sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \Sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \Sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \Sigma_n \end{bmatrix}$$

The erorrs-in-variables logistic regression model is

$$\begin{cases} \widetilde{\boldsymbol{x_i}} = \boldsymbol{x_i} + \boldsymbol{u_i} \\ \boldsymbol{y_i} = \frac{1}{1+\exp(-\boldsymbol{x_i'}\boldsymbol{\beta})} + \boldsymbol{\epsilon_i} \end{cases} \qquad for\ i = 1, 2, ..., n \quad \text{with m observations for each } \boldsymbol{x_i}.$$

2.2 Estimation Method

Let $\mu_i(\boldsymbol{x_i}) = \hat{p}_i = \frac{1}{1+\exp(-\boldsymbol{x_i'}\hat{\boldsymbol{\beta}})}$ be the estimated probability by true data $\boldsymbol{x_i}$.

Similarly, let $\mu_i(\widetilde{\boldsymbol{x_i}}) = \hat{\hat{p}}_i = \frac{1}{1+\exp(-\widetilde{\boldsymbol{x_i'}}\hat{\hat{\boldsymbol{\beta}}})}$ be the estimated probability by observed data $\widetilde{\boldsymbol{x_i}}$.

For general logistic regression model, the estimation is unbiased in terms of binomial proportion and predicted probability. That is $\mathbb{E}[y_i - \mu_i(\boldsymbol{x_i})] = 0$. Therefore, we also have $\mathbb{E}\{\boldsymbol{x_i}[y_i - \mu_i(\boldsymbol{x_i})]\} = \mathbf{0}$.

If there are measurement errors in covariates, similar estimates will lead to a biased estimation, which means

$$B = \mathbb{E}\{\widetilde{\boldsymbol{x_i}}[y_i - \mu_i(\widetilde{\boldsymbol{x_i}})]\} \neq \mathbf{0}.$$

where B is the bias.

However, if we can subtract this bias from the original form, we will get an unbiased estimation. Then, from the unbiased estimation, we can solve the eqution system to get our estimated parameters $\boldsymbol{\beta}$. To be specific,

$$\begin{aligned} B &= \mathbb{E}\{\widetilde{\boldsymbol{x_i}}[y_i - \mu_i(\widetilde{\boldsymbol{x_i}})]\} \\ &= \mathbb{E}\{(\boldsymbol{x_i} + \boldsymbol{u_i})[y_i - \mu_i(\boldsymbol{x_i} + \boldsymbol{u_i})]\} \end{aligned}$$

By first-order Taylor expansion of matrix form,

$$\begin{aligned} \mu_i(\boldsymbol{x_i} + \boldsymbol{u_i}) &= \mu_i(\boldsymbol{x_i}) + \boldsymbol{u_i}' \cdot D\mu_i(\boldsymbol{x_i}) \\ &= \mu_i(\boldsymbol{x_i}) + \boldsymbol{u_i}' \cdot \frac{d\mu_i(\boldsymbol{x_i})}{d\boldsymbol{x_i}} \\ &= \mu_i(\boldsymbol{x_i}) + \boldsymbol{u_i}' \cdot \frac{d\eta_i(\boldsymbol{x_i})}{d\boldsymbol{x_i}} \frac{d\mu_i(\boldsymbol{x_i})}{d\eta_i(\boldsymbol{x_i})} \qquad \text{where } \eta_i(\boldsymbol{x_i}) = \boldsymbol{x_i}'\boldsymbol{\beta} \\ &= \mu_i(\boldsymbol{x_i}) + \boldsymbol{u_i}'\boldsymbol{\beta}\dot{\mu}_i \qquad \text{where } \dot{\mu}_i = \frac{d\mu_i(\boldsymbol{x_i})}{d\eta_i(\boldsymbol{x_i})} \end{aligned}$$

Then, we have

$$
\begin{aligned}
B &= \mathbb{E}\{(\boldsymbol{x_i}+\boldsymbol{u_i})[y_i-\mu_i(\boldsymbol{x_i})-\boldsymbol{u_i}'\boldsymbol{\beta}\dot{\mu}_i] \mid \boldsymbol{x_i}\}\\
&= \mathbb{E}\{\boldsymbol{x_i}[y_i-\mu_i(\boldsymbol{x_i})]-\boldsymbol{x_i}\boldsymbol{u_i}'\boldsymbol{\beta}\dot{\mu}_i+\boldsymbol{u_i}[y_i-\mu_i(\boldsymbol{x_i})]-\boldsymbol{u_i}\boldsymbol{u_i}'\boldsymbol{\beta}\dot{\mu}_i \mid \boldsymbol{x_i}\}\\
&= \mathbb{E}\{-\boldsymbol{u_i}\boldsymbol{u_i}'\boldsymbol{\beta}\dot{\mu}_i \mid \boldsymbol{x_i}\}\\
&= -\Sigma_i\boldsymbol{\beta}\dot{\mu}_i
\end{aligned}
$$

where $\dot{\mu}_i = \frac{d\mu_i(\boldsymbol{x_i})}{d\eta_i(\boldsymbol{x_i})}$. It could be easily calculated in R by ***binomial()$mu.eta*** option in *glm* function.

Now, we can substract this bias at the begining so that we can get an unbiased estimation by solving a non-linear eqution system of $\boldsymbol{\beta}$.

$$\mathbb{E}\{\widetilde{\boldsymbol{x_i}}[y_i-\mu_i(\widetilde{\boldsymbol{x_i}})]-B\} = \mathbf{0}.$$

By Monte Carlo method,

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}\widetilde{\boldsymbol{x}}_{im}[y_i-\mu_i(\widetilde{\boldsymbol{x}}_{im})]+\frac{1}{n}\sum_{i=1}^{n}\hat{\Sigma}_i\boldsymbol{\beta}\dot{\mu}_i = \mathbf{0} \qquad \text{with } N = n\cdot m.$$

Finally, we get a non-linear equation system of $\boldsymbol{\beta}$. Solving it by function ***nleqslv*** in R, we will get an unbiased estimation of parameters $\hat{\boldsymbol{\beta}}$.

3 Simulation Study

3.1 Generate an Errors-in-variables Sample

For simplicity, we assume

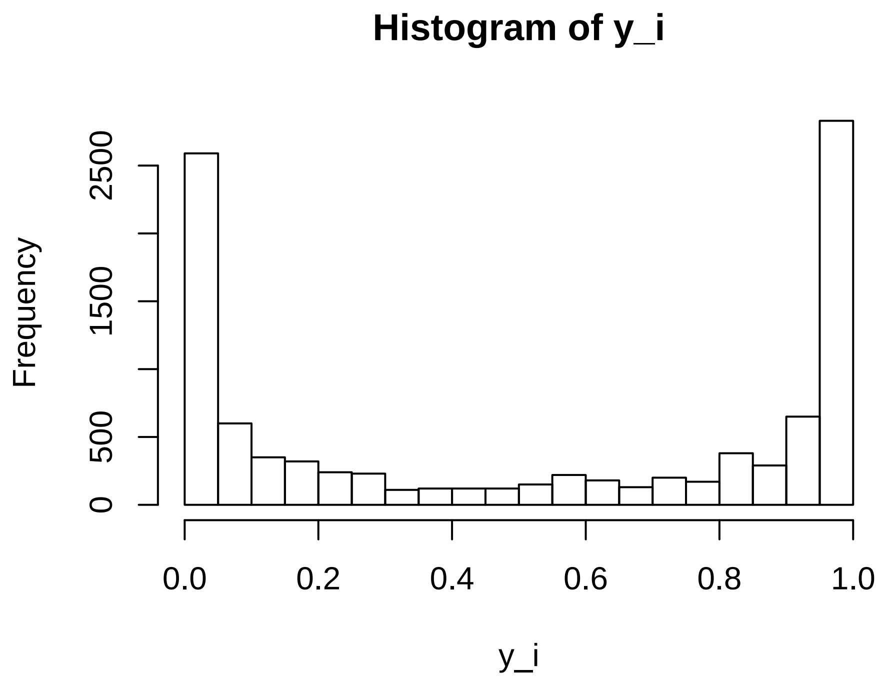
- $\Sigma_1 = \Sigma_2 = ... = \Sigma_n = \Sigma = I$.
- $\boldsymbol{x_i} = (1, X_1, X_2, X_3)$ with $X_1 \sim N(10, 1^2)$, $X_2 \sim N(20, 3^2)$ and $X_3 \sim N(-40, 2^2)$.
- $m = 10, n = 1000$

By our assumption, $y_i = \frac{1}{1+\exp(-\beta_0-\beta_1X_1-\beta_2X_2-\beta_3X_3)} + \epsilon_i$ with $\epsilon_i \sim N(0,1)$ *i.i.d.*.

Generate Errors-in-variables Sample based on these assumptions.

```
# Generate EIV sample
n<-1000; m<-10; k<-3
set.seed(2019)
U_mtx <- data.frame(u1=rnorm(m*n),u2=rnorm(m*n),u3=rnorm(m*n))
rep_n <- rep(m,n)
X_true <- data.frame(x1=rep(rnorm(n, 10,1),rep_n),x2=rep(rnorm(n,20,3),rep_n),x3=rep(rnorm(n,-4
X_tld <- X_true + U_mtx

eta <- 5 + 0.5*X_true$x1 + 1.5*X_true$x2 + X_true$x3
y_i <- 1/(1+exp(-eta))
hist(y_i)
```




```
z_i <- rbinom(m*n,1,y_i)
```

3.2 Simulation

Given $y, \widetilde{X}_1, \widetilde{X}_2, ...$ and $\Sigma_1 = \Sigma_2 = ... = \Sigma_n = \Sigma$, the variance-corariance matrix of all errors will be

$$\Sigma_u = \begin{bmatrix} \Sigma & 0 & 0 & \dots & 0 \\ 0 & \Sigma & 0 & \dots & 0 \\ 0 & 0 & \Sigma & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \Sigma \end{bmatrix}$$

Therefore, the non-linear equation system will be simplified as

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \widetilde{\boldsymbol{x}}_{im}[y_i - \mu_i(\widetilde{\boldsymbol{x}}_{im})] + \hat{\Sigma}\boldsymbol{\beta}\sum_{i=1}^{n}\sum_{j=1}^{m}\dot{\mu}_i = \mathbf{0}.$$

```
# Estimation
df_true <- data.frame(y=z_i,x1=X_true$x1,x2=X_true$x2,x3=X_true$x3)
df_obs <- data.frame(y=z_i,x1=X_tld$x1,x2=X_tld$x2,x3=X_tld$x3)
true_glm <- glm( y~x1+x2+x3,data=df_true,family="binomial")
obs_glm <- glm( y~x1+x2+x3,data=df_obs,family="binomial")

mean_xobs <- NULL
```

```
for (i in 1:n) {
  rep_mean <- t(matrix(rep(colMeans(X_tld[(m*(i-1)+1):(m*i),]),m), nrow = 3))
  mean_xobs <- rbind(mean_xobs, rep_mean)
}

X_tld_mtx <- data.frame(x0=rep(1,m*n),X_tld)
cov_i <- cov(data.frame(u0=rep(0,m*n),U_mtx))
beta_fn <- function(beta) {
 eqa_v <- colSums(X_tld*(y_i-obs_glm$fitted.values))+
    cov(X_tld-mean_xobs) %*% beta * sum(obs_glm$family$mu.eta(obs_glm$linear.predictors))
 eqa_v
}

nleqslv::nleqslv(c(1,2,3),beta_fn)$x
```

```
## [1]  0.4813892  0.9168533 -1.8547521
```

Reference

[1] Wikipedia. *Errors-in-variables models* https://en.wikipedia.org/wiki/Errors-in-variables_models

[2] Wikipedia. *Taylor series* https://en.wikipedia.org/wiki/Taylor_series

[3] Rhonda Robinson Clark (1982 July). THE ERROR-IN-VARIABLES PROBLEM IN THE LOGISTIC REGRESSION MODEL. *Institute of Statistics Mimeo Series, No. 1407*