

# MA540 Mathematical Statistics Project 1

Erik Erhardt, Gang Shen, Li Xiao and Xiao Zhong

December 4, 2002

## 1 Introduction

In a survey experiment it's possible to get less than 100% response rate. In this project we will consider a model to study nonignorable nonresponse. This occurs when there is a substantial nonresponse which is believed to be nonignorable, and there is a relationship between the nonresponse rate and at least one factor of the sample population. The possibility of response and nonresponse is correlated with a linear function of a parameter from the original population. In our example we let this population parameter  $y_i$  be independently and identically distributed normal with some known mean and variance for the whole population. In other words, the respondents and nonrespondants are from the same population, and therefore have the same distribution.

Our goal is to investigate how well the model below can reproduce  $\beta_0$  and  $\beta_1$ . We are estimating  $\beta_0$  and  $\beta_1$  by each  $y_i$ . However, for nonresponse,  $y_i$  is unobservable. Therefore, there is some difficulty in estimating  $\beta_0$  and  $\beta_1$  when the nonresponse rate is high.

## 2 Model discussion

The model we use is

$$\begin{aligned} r_i|p_i &\stackrel{ind}{\sim} \text{Bernoulli}(p_i), i = 1, \dots, n \\ \ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 y_i \\ y_i &\stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2) \end{aligned}$$

Let  $\mu$  and  $\sigma^2$  be known and given from the normal density to generate the population of  $y_i$ . Let  $\beta_0$  and  $\beta_1$  be given but unknown parameters in the linear function to generate  $p_i$ . Given  $p_i$ , the random variable  $r_i$  will conform to a Bernoulli distribution with probability  $p_i$ . We draw  $u_i$  from a Uniform(0,1) density and assign  $r_i = 1$  if  $u < p_i$  indicating “responded”, and  $r_i = 0$  otherwise indicating “did not respond”. By the MLE method our model will generate  $\beta_0$  and  $\beta_1$  so that we can compare these estimated parameters with the “true” ones. Finally, we compare the population  $\beta_0$  and  $\beta_1$  with our estimated parameters to give us some measure of how well our model can estimate  $\beta_0$  and  $\beta_1$ .

Using the model above, we get

$$r_i|y_i \stackrel{ind}{\sim} \text{Bernoulli}\left(\frac{e^{\beta_0 + \beta_1 y_i}}{1 + e^{\beta_0 + \beta_1 y_i}}\right)$$

giving

$$\begin{aligned} f_{R_i, Y_i}(r_i|y_i) &= \left(\frac{e^{\beta_0 + \beta_1 y_i}}{1 + e^{\beta_0 + \beta_1 y_i}}\right)^{r_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 y_i}}\right)^{1-r_i} \\ f_{Y_i}(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \end{aligned}$$

Therefore, our Normal-Logistic model is defined as

$$\begin{aligned} f_{R_i, Y_i}(r_i, y_i) &= f_{R_i, Y_i}(r_i|y_i) f_{Y_i}(y_i) \\ &= \left(\frac{e^{\beta_0 + \beta_1 y_i}}{1 + e^{\beta_0 + \beta_1 y_i}}\right)^{r_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 y_i}}\right)^{1-r_i} \\ &\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \\ & \quad r_i = 0, 1; -\infty < y_i < \infty \end{aligned}$$

We use the MLE on the joint density to estimate  $\beta_0$  and  $\beta_1$ .

$$\prod_{i=1}^n f_{R_i, Y_i}(r_i, y_i) = \prod_{i=1}^n \left\{ \left( \frac{e^{(\beta_0 + \beta_1 y_i) r_i}}{1 + e^{\beta_0 + \beta_1 y_i}} \right) \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \right\}$$

Separating this joint pdf into the response and nonresponse portions gives the following equation. Part (1) is for the responses and (2) is for the non-responses.

$$\prod_{i=1}^n f_{R_i, Y_i}(r_i, y_i) = \prod_{i=1}^r \left\{ \left( \frac{e^{(\beta_0 + \beta_1 y_i)}}{1 + e^{\beta_0 + \beta_1 y_i}} \right) \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \right\} \quad (1)$$

$$\times \prod_{i=r+1}^n \left\{ \left( \frac{1}{1 + e^{\beta_0 + \beta_1 y_i}} \right) \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \right\} \quad (2)$$

In part (1) we know the value of  $y_i$  since these are responses, so this is in terms of  $\beta_0$  and  $\beta_1$  only. In part (2) we do not know the value of  $y_i$  since these are nonrespondants, but we can take the expected value instead. Therefore, we perform a numerical integration of  $y_i$  to get part (2) in terms of  $\beta_0$  and  $\beta_1$  only.

We introduce the equations used for numerical integration first with the description following. We consider the two parts of (2) separately as (3) and (4).

$$f_{\beta}(\beta, y_i) = \left( \frac{1}{1 + e^{\beta_0 + \beta_1 y_i}} \right), \quad (\beta = \beta_0, \beta_1) \quad (3)$$

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \quad (4)$$

$$E_{\beta}(f_{\beta}(\beta, y_i)) = \int f_{\beta}(\beta, y_i) f_{Y_i}(y_i) dy_i$$

$$E_{\beta} \left( \frac{1}{1 + e^{\beta_0 + \beta_1 y_i}} \right) = \int \left\{ \left( \frac{1}{1 + e^{\beta_0 + \beta_1 y_i}} \right) \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \right\} dy_i$$

$$E_{\beta}(f_{\beta}(\beta, y_i)) = \frac{1}{M} \sum_{i=1}^M f_{\beta}(\beta, y_i) \quad (5)$$

We perform this integration over  $y_i$  in the following way<sup>1</sup>. We draw  $M = 1000$  samples of  $y_i$  from the density function given by  $f_{Y_i}(y_i)$  in (4) and

---

<sup>1</sup>Refer to Gelman, Carlin, Stern and Rubin in "Bayesian Data Analysis" page 305

evaluate each  $f_{\underline{\beta}}(\underline{\beta}, y_i)$  in (3) at  $y_i$ . Finally, we average over  $f_{\underline{\beta}}(\underline{\beta}, y_i)$  in (5). Now part (2) is in terms of  $\beta_0$  and  $\beta_1$  only.

Define the joint density in terms of  $\beta_0$  and  $\beta_1$  as

$$\begin{aligned} h_{\underline{\beta}}(\beta_0, \beta_1) &= \prod_{i=1}^n f_{R_i, Y_i}(r_i, y_i) \\ &= \prod_{i=1}^r \left\{ \frac{e^{\beta_0 + \beta_1 y_i}}{1 + e^{\beta_0 + \beta_1 y_i}} \right\} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \times (E_{\underline{\beta}}(f_{\underline{\beta}}(\underline{\beta})))^{n-r} \end{aligned}$$

Now we have  $h_{\underline{\beta}}(\beta_0, \beta_1)$  in terms of  $\beta_0$  and  $\beta_1$  only. We employ the Nelder-Mead algorithm to obtain the maximum likelihood estimators for  $\beta_0$  and  $\beta_1$ . A few words about Nelder-Mead. The Nelder-Mead algorithm is used to *minimize* a function. We specify the function  $h_{\underline{\beta}}(\underline{\beta}, y_i)$  as above and maximize the joint density function by minimizing the negative of  $h_{\underline{\beta}}(\underline{\beta}, y_i)$ .

### 3 Results

The plot in Figure 1 on page 7 gives the difference of the population  $\beta_1$  and estimated  $\beta_1$  versus the response rate. From this it seems the estimation of  $\beta_1$  is good for a response rate of at least 0.92. However, a response rate of less than .92 has great variation. The same is true for  $\beta_0$ , shown in the plot in Figure 2 on page 7. In this preliminary experiment, since together  $\beta_0$  and  $\beta_1$  ultimately determine the response rate  $r_i$ , we varied only  $\beta_1$ , fixing  $\beta_0$  at 0 for simplicity.

Similar information is captured in the plots in Figure 3 on page 8 and Figure 4 on page 8. Here it is the population  $\beta$  plotted versus the estimated  $\beta$ .

Since we suspected this variation could be due to the unavoidable variation from the sample, we investigated the variation for each response rate. We generate 100 samples for each response rate by drawing a random pair of  $\beta_0$  and  $\beta_1$  ( $-5 < \beta_0 < 3$  and  $0 < \beta_1 < 3$  gave an response range that visited each rate with nearly equal frequency). For the differences between the population and estimated  $\beta$  versus response rate we create scatterplots (similar to Figure 3 and 3). The plot for  $\beta_1$  is in Figure 5 on page 9 and  $\beta_0$  in Figure 6 on page 9. It is clear that the variation is nearly constant for all response rates. Recall that we draw the respondands and the nonrespondants from the same distribution. Therefore, we should expect constant variation since there is no bias when the responses of the non-respondents are the same as those of the respondents<sup>2</sup>. We note that there is more variation in  $\beta_0$ , our intercept, than  $\beta_1$ , but this is left without investigation. This may be worth further investigation.

The strength of the estimation is shown well by the plot of the population versus the estimated  $\beta$  in Figure 7 on page 10 and Figure 8 on page 10.

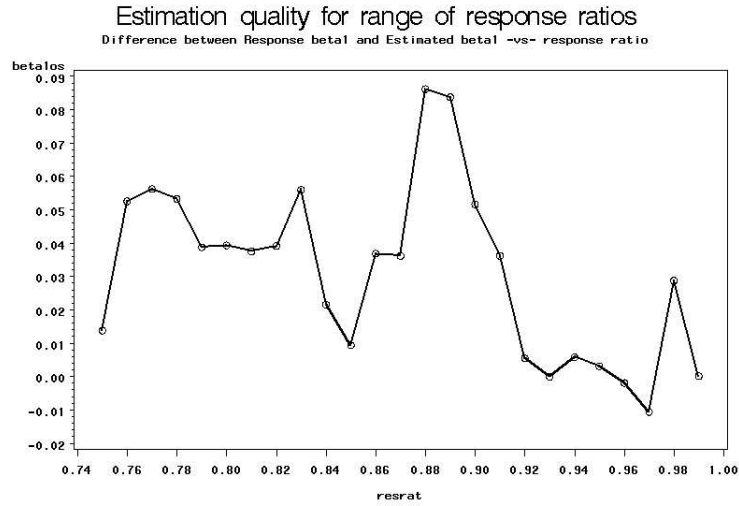
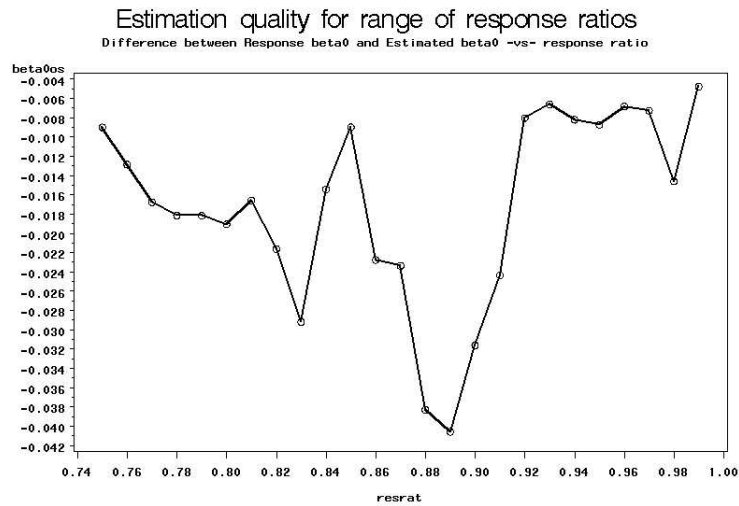
---

<sup>2</sup>Nandram and Choi in “An Assessment of the Impact of Non-response Weighting in Sample Survey Data” (2000)

## 4 Conclusion

We find that regardless of the response rate, the estimation of the parameters  $\beta_0$  and  $\beta_1$  is quite good. We attribute this to the assumption that the respondents and nonrespondents come from the same population. Note in Figure 5 on page 9 and Figure 6 on page 9 that the variation increases only slightly as the response rate decreases from 99% to about 18%. When the response rate is less than 18%, the variation increases more rapidly, but the center for the estimate remains the same as the true parameter.

This project was a great opportunity to practice numerical techniques for maximum likelihood estimation using Nelder-Mead. This project was an excellent exercise.

Figure 1: Difference of population and estimated  $\beta_1$  versus response ratesFigure 2: Difference of population and estimated  $\beta_0$  versus response rates

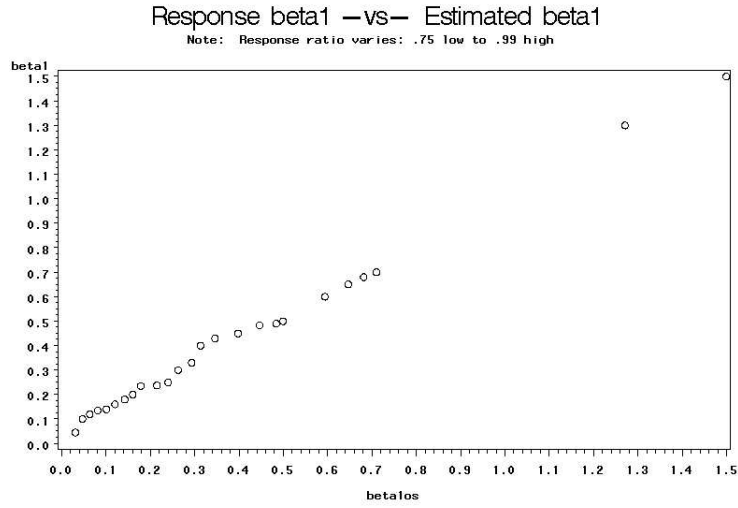


Figure 3: Population  $\beta_1$  versus estimated  $\beta_1$

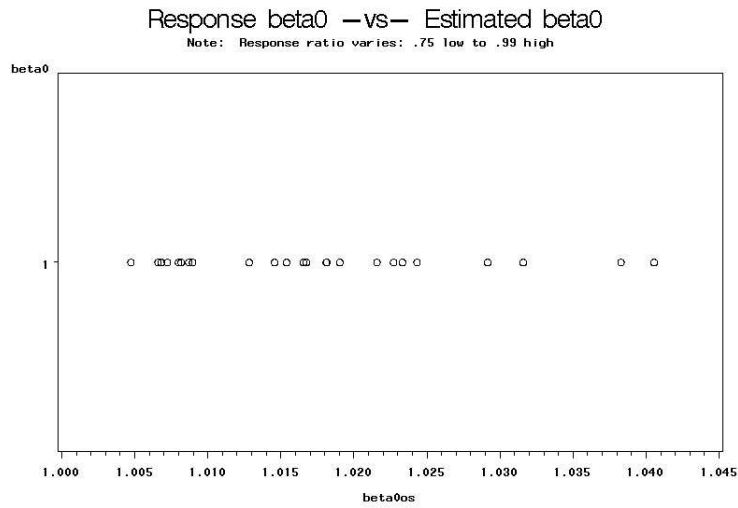


Figure 4: Population  $\beta_0$  versus estimated  $\beta_0$

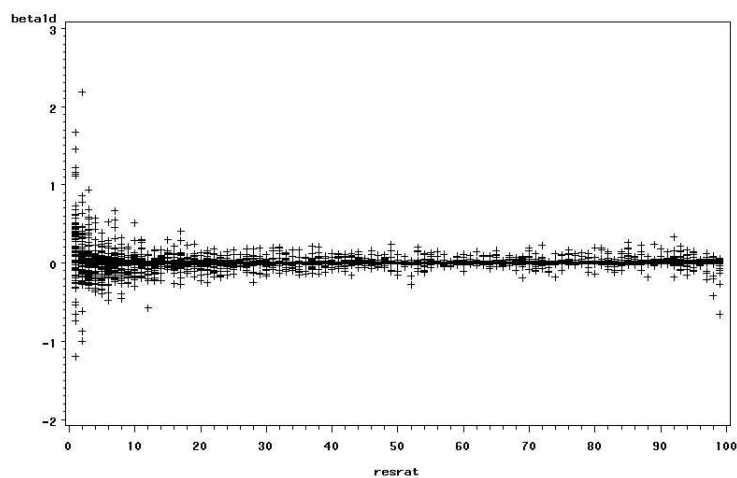


Figure 5: Difference of population and estimated  $\beta_1$  versus 100 samples at each response rate

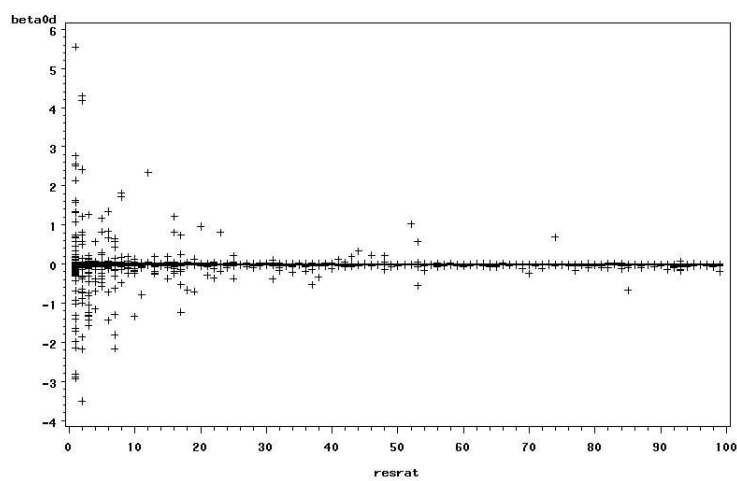


Figure 6: Difference of population and estimated  $\beta_0$  versus 100 samples at each response rate

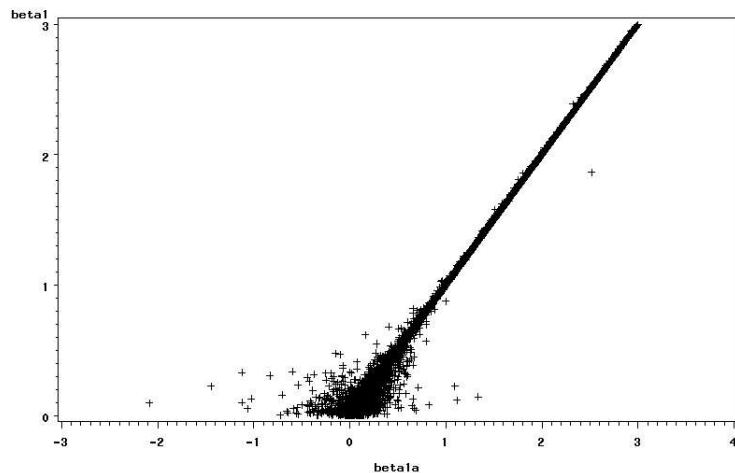


Figure 7: Population  $\beta_1$  versus estimated  $\beta_1$  for 100 samples at each response rate

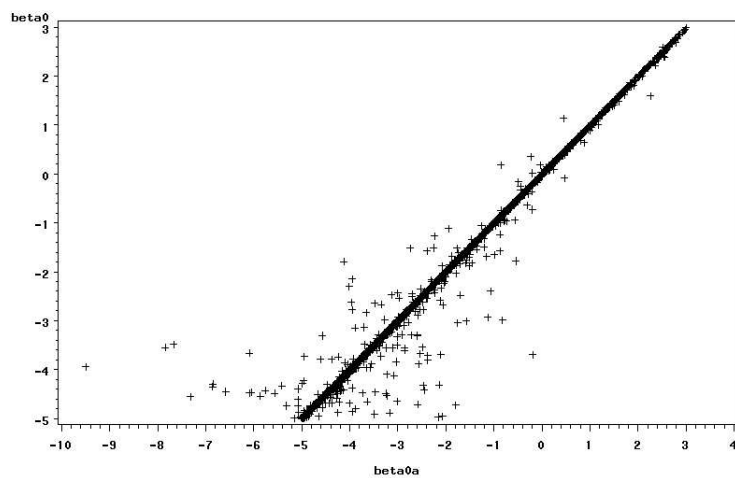


Figure 8: Population  $\beta_0$  versus estimated  $\beta_0$  for 100 samples at each response rate