# SIMULATION-BASED EXPLORATION OF SURVEYS WITH NON-RESPONSE

YAP Von Bing
Department of Statistics and Applied Probability
National University of Singapore
stayapvb@nus.edu.sg

*A simple model of random selection and systematic non-response from a linear bivariate population is proposed to illustrate the bias in estimating regression parameters based only on the responses. The problem is explored graphically and numerically via a simulation study implemented in the statistical computing environment R. It highlights the propensity to bias in regression analysis of observational multivariate data. The model can be substantially modified, to make it possible to imagine any dataset as coming from a randomised survey with possible non-response.*

## INTRODUCTION

In using the sample mean to estimate a population mean, bias can stem from both the selection of sample and non-response. The first source can be eliminated by a probability sample, such as a simple random sample. The second source is much harder to avoid. Many surveys are contented with a "high" response rate of 70% or 80%.

Even for a survey using probability sampling with a high response rate, the bias in the estimate can be substantial. For example, consider a population of size 1,000, of which 700 are responders, and 300 are non-responders. Among the responders, 140 have property A: they use smart devices more than eight hours day; among the non-responders, 270 have property A. Suppose a simple random sample of size 100 is taken. Around 70 responses will be obtained, and based of these, the sample rate of A will be around 0.2. However, the population rate of A is (140+270)/1000 = 0.41. So the sample rate has a negative bias, understating the parameter by about half. The conclusion can be demonstrated by a simulation, where the sampling process is repeated many times, and a histogram of the sample rates is presented to be clearly separated from the parameter.

The issue of bias from non-response seems quite obvious to a lay person and is explained and illustrated amply in many textbooks. See, for example, Chocran (1977). Still, it may not inform the analysis of real surveys. It is all too common to see statistical inferences made based on the responses, without any discussion or even awareness that the bias may render the confidence interval or P-value quite uninterpretable.

If the oblivious producer of an estimated population rate may be sensitised to the real possibility of bias in the estimate via a vague recall of theoretical material, there seems to be no analogous recourse for multivariate data. It is as though the bias in any univariate analysis miraculously goes away when one is interested in studying associations among multiple variables. A regression analysis on observational data that discusses bias arising from sampling issues is far from commonly reported.

The aim of this paper is to outline a simulation approach to study biases in estimates and predictions from a linear regression model, where about half of the randomly selected subjects do not respond. First a bivariate population of size 1,000 is generated, which has a straight regression function. The generated population is then fixed for subsequent use. The survey consists of a simple random sample of size 100, and respondents are assumed to reside within a subset specified by another straight line. The procedure is repeated 10,000 times, so that the bias in the estimates and predictions can be quantified. The discussion touches on utility of viewing any dataset as originating from a randomised survey with possible non-response, and on its relevance in statistics and data science.

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
*Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),*
July 2020, Online. Voorborg, The Netherlands. 

## A BIVARIATE MODEL OF NON-RESPONSE

A bivariate population of size 1,000 is set up by repeating the following procedure. A number $x$ is generated from the uniform distribution on the interval (0,4). Then $y = 1 + 2x + e$, where $e$ is generated from the standard normal distribution. The scatterplot of a realised population is shown in Figure 1. This population is kept fixed for the subsequent description.

Imagine a statistician takes a simple random sample of size 100, and then interviews the selected individuals to record their $x$ and $y$ values, in order to estimate the population regression equation, $y = 1 + 2x$. We assume that the values can be obtained without error. However, there is non-response: an individual agrees to an interview exactly if his values satisfy the constraint $-1 < y - x - 2 < 1$. Figure 1 shows one particular sample, where respondents are green circles and non-respondents are red circles. The green line is the regression line based on the respondents, which is too high for low $x$ values, and too low for high $x$ values. Is this just due to random fluctuation, or is there also bias? It might be possible to derive any bias analytically, but a simulation study is a feasible method to find out.

The simulation study repeats the follow steps, many times. In this case, 10,000 times.

(1) Draw a simple random sample of 100 from the fixed population.
(2) Keep points with $-1 < y - x - 2 < 1$, the responses. Record the number of such points.
(3) Fit a regression line to the responses. Record the $y$-intercept and the gradient.

Implementing the procedure in R, we obtain 10,000 realisations of three random variables: the number of responses, as well as the y-intercept and the gradient of the fitted regression line. All the associated regression lines are plotted in Figure 2, forming a green cloud, which shows that the particular regression line in Figure 1 is quite typical. The averages of the three sets of realisations are 46.0, 1.73 and 1.26 (to 3 significant figures), and the standard deviations are around 0.00, 0.13 and 0.08. Hence the response rate is about 51%, the population $y$-intercept 1 is overestimated by 0.73, and the population gradient 2 is underestimated by 0.74. The average regression line, $y = 1.73 + 1.26x$, is off the population regression line in a similar manner as the one in Figure 1. We can also examine the bias in predictions. For $x$=0.5, the average regression line predicts $y$ to be $1.73 + 1.26 \times 0.5 = 2.36$, which is higher than the average $y$-value $1 + 2 \times 0.5 = 2$. For $x$=3, the prediction is 5.50, which is lower than 7, the average $y$-value.

In conclusion, the simulation study tells us that if non-response is ignored, there is a bias in estimating the population regression line, which leads to biased predictions of $y$ from $x$. The R code for generating the population and running the simulation study is available upon request.

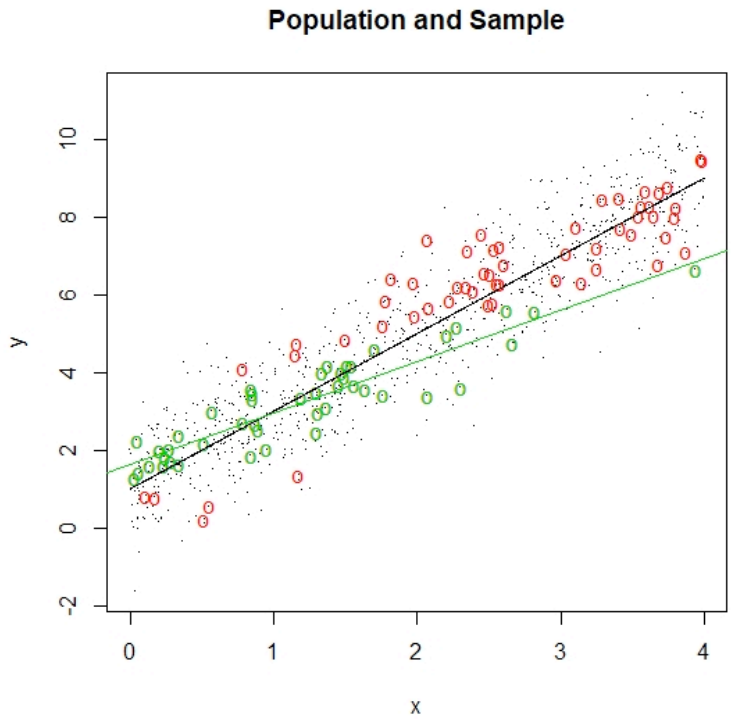


Figure 1: Scatterplot of 1,000 points with $x$ values drawn from the uniform distribution on (0,4) and $y = 1 + 2x + e$, where $e$ is drawn from the standard normal distribution. The regression line $y = 1 + 2x$ is in black. The coloured circles are a simple random sample of 100 people. The green circles are the respondents, satisfying $-1 < y - x - 2 < 1$, and the green line is the corresponding regression line. The red circles are the non-respondents.

## DISCUSSION

Freedman (2009) questions the reliability of statistical analyses on observational data. Our model of non-response may be seen as a concrete example of how bias can arise in using regression analysis of non-random samples to infer parameters. The model is simplistic: the population regression function is a straight line, and the respondents lie within a neighbourhood of another straight line. Yet it illustrates a general point: Whenever the respondents' scatterplot does not represent the population, the regression analyst can be misled. Suppose the population regression function is a piecewise linear function or a nonlinear function, and the responses are still governed by a linear constraint, then ignoring the non-responses obviously leads to a more serious error. Furthermore, it is obvious that responses can be corralled by all sorts of non-linear constraints. Perhaps any "non-random" sample can be imagined as having come from a randomised survey with some specific pattern of non-response. This viewpoint may deserve further consideration. From a pedagogical point of view, if the commonly understood red flag of bias associated with non-response can be thus transferred to regression analyses, then the impact can be immense. It seems plausible that misleading conclusions from a regression analysis of non-random samples are more likely in multiple regression than simple regression. Given the proliferation of large multivariate data sets from observational studies on important current concerns from all empirical sciences, there is thus an urgent need to engage in more theoretical study of this issue, and to cultivate a more critical attitude among the statistics students.

Recent development of data science has prompted statisticians to pay more attention to prediction rather than inference (Sanders 2019). This is a timely, or perhaps delayed, response to the 2001 landmark paper of Breiman, who contrasted "data modeling" and "algorithmic modeling". At the extreme, the goal

is to predict a variable as well as possible from a few other variables via some function or algorithm, and there is no need for a statistical model for the data. Even in this activity, the analyst will do well to pay heed to the possibility that the sample being analysed, the training data set, may not represent the population that well. So, the new framework, of representing such non-representativeness as a randomised survey with non-response, may be of enduring relevance in the rapidly changing field of data analysis.

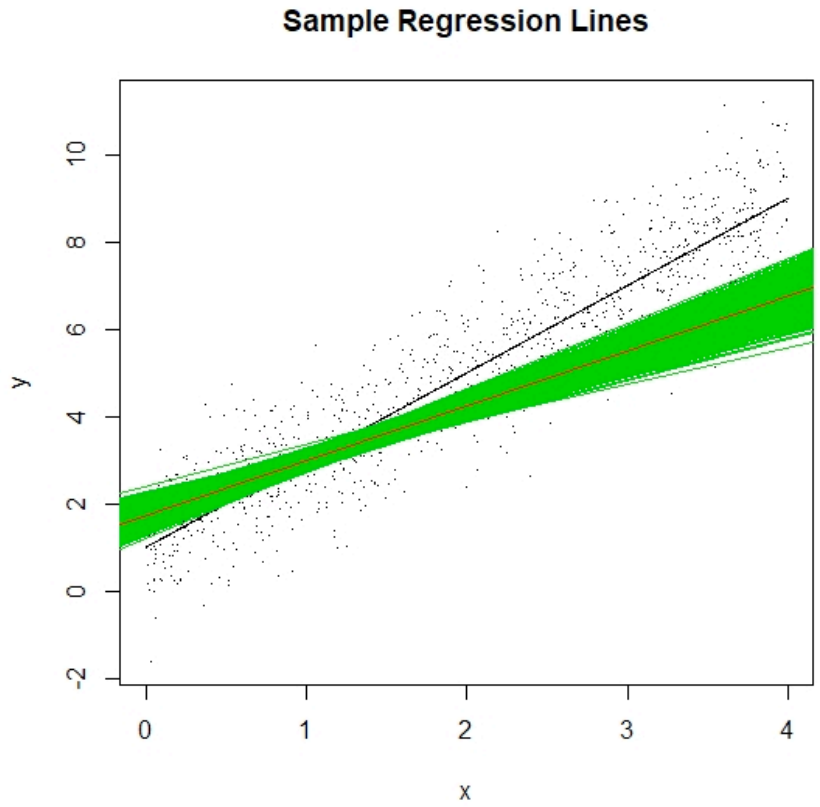


Figure 3: The green shaded area consists of 10,000 realised regression lines from the simulation study. The red line is the average regression line, $y = 1.73 + 1.26x$.

REFERENCES

Breiman, L. Statistical Modeling: The Two Cultures. *Statistical Science*, 16:199-231 (2001).

Chocran, W.G. *Sampling Techniques, 3e*. Wiley (1977).

Freedman, D.A. Statistical Assumptions as Empirical Commitments, in *Statistical Models and Causal Inference*, edited by Collier, D., Sekhon, J.S., and Stark, P.B. (2009)

Sanders, N. A Balanced Perspective on Prediction and Inference for Data Science in Industry. *Harvard Data Science Review* 1.1 (2019).