{Bounding} a package for identification bounds

Palabras clave: ignorability, missing outcomes, partial identification, regression

Abstract The *Bounding* package has been developed to obtain bound estimations for both the conditional expectation E(Y| X) and the parameters of the regression model. The bounds are obtained under a partial identification approach when there are missing outcomes. The package integrates other R packages such as *tidyverse* to data manipulation and *ggplot2* to export a plot of the estimated bounds. Illustrations are shown in the context of analyzing predictive validity of selection factors in the Chilean university admission system.

Introduction

Regression models are defined to analyze the effect of a set of covariates over the mean of a response variable. From a statistical perspective this represents a model for the expectation of Y, the response, conditional on a (multidimensional) variable X. When there are partial observability of the outcome, the partial identification approach is an alternative for making inferences without distributional assumptions on the non-observables values.

If X represents a random variable and Z = 1 if Y is observed and Z= 0 if not. Then, by the Law of total probability (Kolmogorov, 1950)

$$E(Y|X) = E(Y|X, Z = 1)P(Z = 1|X) + E(Y|X, Z = 0)P(Z= 0|X) \quad (1)$$

where P(Z = 1|X) corresponds to the proportion of observed values of Y given X. The empirical evidence reveals information for estimating E(Y |X, S = 1) but not for estimating E(Y |X, Z = 0). As a consequence, E(Y |X) can not be identified by the data generation process (for details see Manski, 1989). Nevertheless, a restriction can be imposed on the non-observed conditional expectation in order to point identify E(Y |X), namely:

$$E(Y|X, Z = 1) = E(Y|X, Z = 0) \quad (2)$$

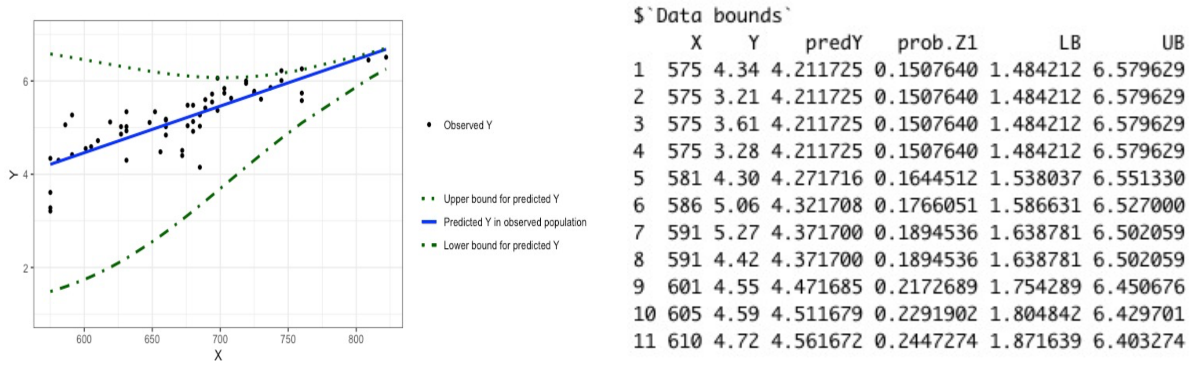
The condition (2) is the formalization of the assumption considered when the regression model is estimated from complete data only. this assumption is also known as ignorability or mean missing at random. In the absence of prior information about the non-observed conditional expectation, E(Y |X, Z = 0), if $Y \in [y_0; y_1]$ then E(Y |X, S = 0) $\in [y_0; y_1]$. By replacing these values in (1) an interval is obtained for all the possible values of E(Y|X). In fact,

$$E(Y|X, Z = 1)P(Z = 1|X) + y_0P(Z = 0|X) \leq E(Y|X) \leq E(Y|X, Z = 1)P(Z = 1|X) + y_1P(Z = 0|X) \quad (3)$$

The interval (3) contains all the plausible regression models that are consistent with the empirical evidence. The extremes of the intervals are known as the identification bounds for E(Y|X). However, in several situations the interest is to analyze the impact of a particular variable of X on the mean values of Y when it is partially observed. In this case, the interest is not the regression model itself but the regression parameter associated to this variable. Under a partial identification approach similarly described here, Stoye (2007) showed bound identifications for the regression parameters of E(Y|X) as is shown in Alarcón-Bustamante et al. (2023).

Illustration

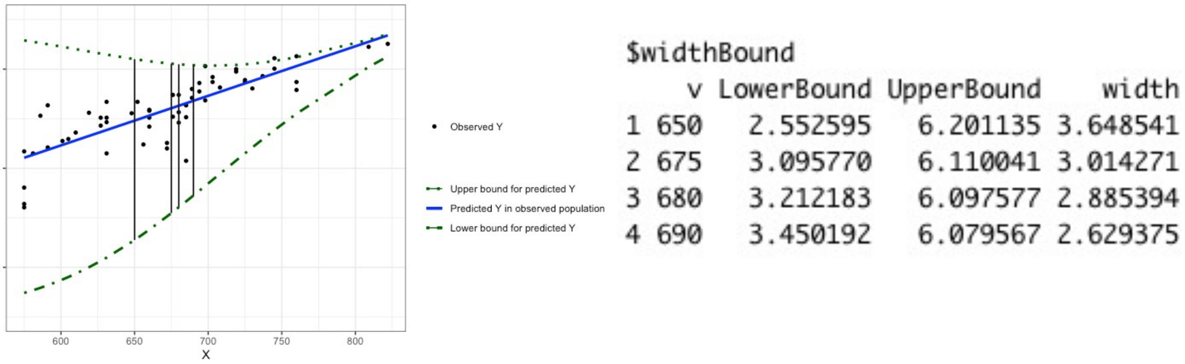
A clear example of partial observability outcome is the selection to the university. We illustrate it in the context of the Chilean university admission process where the interest is to evaluate the effect of selection factors over the performance of students in the university. The scores for selection factors are observed for all applicants to the system. However, the performance in the university (graded point average - GPA, for instance) is observed only on those enrolled ones. From a subset of a real dataset of the Chilean process, using the function *WiderBounds* of the *Bounding* package is obtained the identification bound (3) for the mean of the GPA's (Y) given the mathematic university selection test (X) under a linear model. The function's outputs are the plot of the estimated bounds and the dataset from which the plot is obtained. Illustrations are shown in Figure 1.



```
$`Data bounds`
     X    Y    predY   prob.Z1       LB       UB
1  575 4.34 4.211725 0.1507640 1.484212 6.579629
2  575 3.21 4.211725 0.1507640 1.484212 6.579629
3  575 3.61 4.211725 0.1507640 1.484212 6.579629
4  575 3.28 4.211725 0.1507640 1.484212 6.579629
5  581 4.30 4.271716 0.1644512 1.538037 6.551330
6  586 5.06 4.321708 0.1766051 1.586631 6.527000
7  591 5.27 4.371700 0.1894536 1.638781 6.502059
8  591 4.42 4.371700 0.1894536 1.638781 6.502059
9  601 4.55 4.471685 0.2172689 1.754289 6.450676
10 605 4.59 4.511679 0.2291902 1.804842 6.429701
11 610 4.72 4.561672 0.2447274 1.871639 6.403274
```

Figure 1: (left) Plot of estimated bounds for the mean of GPA given mathematic test scores. (right) Dataset from which the plot is obtained.

Additionally, by using the *widthBound* function the width of the bound intervals are obtained from user specified X value. Both a plot and a table of the results are outputs of the function as shown in Figure 2.



```
$widthBound
    v LowerBound UpperBound    width
1 650   2.552595   6.201135 3.648541
2 675   3.095770   6.110041 3.014271
3 680   3.212183   6.097577 2.885394
4 690   3.450192   6.079567 2.629375
```

Figure 2: (left) Plot of the width of estimated bounds for the used specified mathematic scores. (right) Dataset with summary information of the bounds and the width of the interval.

The identification bounds for the regression coefficients (Stoye, 2007) are also implemented in the Bounding package under the name *Stoye_bounds*

References

Alarcón-Bustamante, E., Varas, I.M., San Martín, E., 2023. On the impact of missing outcomes in linear regression. Chilean Journal of Statistics. (Accepted for publication)

Kolmogorov, A.N., 1950. Foundations of the theory of probability. New York: Chelsea Pub. Co.

Manski, C., 1989. Anatomy of the selection problem. The Journal of Human Resources, 24(3), 343-360.

Stoye, J., 2007. Bounds on generalized linear predictors with incomplete outcome data. Reliable Computing, 13, 293-302.