Chilean Journal of Statistics
Volume 14, Number 1 June 2023, Pages 26–36
DOI: 10.32372/chjs.14-01-02

Statistical Modeling
Research Paper

On the impact of missing outcomes in linear regression

Eduardo Alarcón-Bustamante[1234,*], Inés M. Varas[12] and Ernesto San Martín[1235]

[1]Department of Statistics, Faculty of Mathematics, Pontificia Universidad Católica de Chile, Santiago, Chile.
[2]Interdisciplinary Laboratory of Social Statistics, Santiago, Chile.
[3]Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI).
[4]Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile.
[5]LIDAM/CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

(Received: 19 May 2023 · Accepted in final form: 27 June 2023)

Abstract

The linear regression model is commonly used for measuring the impact of covariates over an outcome of interest, which is typically measured through the regression coefficients of the model. However, the presence of missing outcomes can seriously affect this interpretation because we have no idea about the potential impact of the covariates on those units with missing outcomes. Here, we illustrate the consequences of the missing outcomes as the interpretation of the regression coefficients in the impact of the selection factors on the performance in the university.

Keywords: Bounded coefficients · Identification bounds · Ignorability · Missing at random · Partial identification

Mathematics Subject Classification: Primary 62J05 · Secondary 62D10.

1. Introduction

Linear regression is one of the main tools to analyze the impact of k covariates or predictors $\mathbf{x}_i$ (here $\mathbf{x}_i$ includes the intercept) on a dependent variable or output y_i. The basic model is typically written as follows:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \qquad i = 1, \ldots, n;$$

or equivalently

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I), \tag{1.1}$$

where $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $n \times 1$ vectors and $\mathbf{X}$ a $n \times k$ matrix whose rows are $\mathbf{x}_i^\top$. The purpose of linear regression is to report the regression coefficients of each covariate on the output.

*Corresponding author. Email: esalarcon@mat.uc.cl

Thus, the regression coefficients $\boldsymbol{\beta}$ are accordingly named the parameters of interest (along with the variance σ^2). These parameters are understood as the effect of the covariates over the outcome and its variability such that the mean of the distribution of $(Y \mid X)$ is explained by X. In order to make inferences about these parameters in Equation (1.1) a distributional assumption is considered for the error term. As shown in the equation, the most common one is a normal distribution.

In empirical research a typical issue is given when some of the outputs are missing, but not the covariates. Some examples where this issue is present are the following:

(1) For selecting applicants for enlistment in the American military, the Armed Services Vocational Aptitud Test (ASVAB) is used (Department of Defense, 1984). The test looks for classify those applicants that will do well in Air Force technical training in the school. Thus, although the test scores are known for all the applicants, the performance in Air Force technical training in the school is known for the selected applicants only. Other examples in personal selection context can be found in Campbell and Knapp (2001). Figure 1 illustrates this issue in the context of university selection where the outcome at the university, the Graded Point Average - GPA for instance, is observed only in selected applicants.

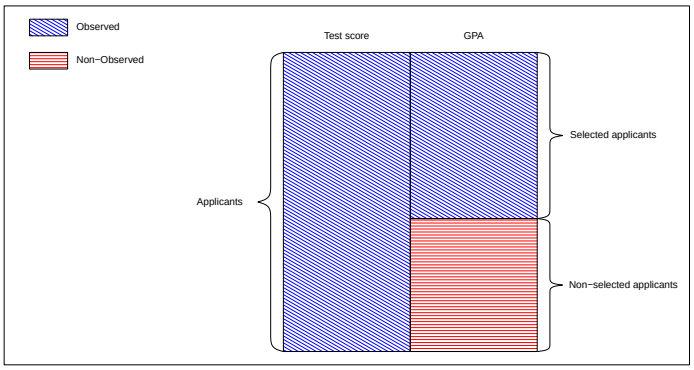

Figure 1. Selection process scheme.

This problem is well-known as the selection problem (also known as the problem of range restriction) and it arises when the data generation process does not fully reveal the behavior of the outcome on the support of the predictors (Manski, 1993).

(2) The Current Population Survey (CPS) is used as a source for learning about the labour force for the population of the United States (full information about this survey can be found in https://www.census.gov/programs-surveys/cps.html). It collects, among other variables, the household income in the United States. However, one issue is that some interviewed do not provide information about the income response (Manski, 2007). One type of analysis is related to measure the impact of some covariates over the income (for instance, annual veteran status, school enrollment, contingent employment, worker displacement, and job tenure, among other topics). Nevertheless, although many respondents provide information about the covariates, the income is not fully observed for them. Other important issue in this type of survey are people that do not provide information about the other variables and people that it considered in the sample, do not answer the questionnaire. These types of issues are not treated in this note.

From a statistical perspective, it is possible to use some techniques to handle the missing outcome data problem. A common approach to deal with this situation is eliminating from the sample the unit samples that have missing values. This means that the data for those without the outcome variable are deleted from the database. For instance, Geiser and Studley (2002) conducted a study to assess the impact of SAT scores, X, on the University of California freshman grades, Y. The authors state that only the full observed data is used for the study. When the selection problem arises, also is usually used a technique for correcting the effect of the missingness of the outcome by making some distributional assumptions about them (see for instance Muthén and Hsu, 1993; Lee and Mendoza, 2022; Hsu, 1995; Zimmermann et al., 2017).

Another practice for tackling the missing outcome data problem is assuming that when the covariates are known, the outcome is mean independent of observing the outcome or not (Manski, 1989). More precisely, suppose that the covariates, $\mathbf{x}$, are used to define classes (for instance, sex, race, among others); thus mean independent assumption asserts that

$$E(Y \mid \mathbf{X} = \mathbf{x}, \text{the outcome is observed}) = E(Y \mid \mathbf{X} = \mathbf{x}, \text{the outcome is not observed}) \quad (1.2)$$

This fact allows for making imputations of the missing outcomes by using the conditional mean of the respondents in the same class. This technique is known as conditional imputation (Little and Rubin, 2002). In other literature, this assumption is also known as ignorability or mean missing at random (for technical details see Manski, 2007; Imbens, 2000; Hirano and Imbens, 2004; Rosenmbaum and Rubin, 1983).

A full discussion about imputation techniques of missing values, as well as Maximum Likelihood Expectation-Maximization Imputation, Multiple Imputation, Hot Deck Imputation, among others, can be found in Carpenter and Kenward (2012); Chema (2014); Rubin (2009); Van Buuren (2012). Imputation methods are procedures that allows for dealing with the missing data problem by making some distributional assumptions. However, we never know how the missing data mechanism is. For a discussion about the impact of using these types of assumption see for instance Alarcón-Bustamante (2022, 2023); Davern et al. (2004); Kambourov and Manovskii (2013); Manski (2007, 2016); San Martín and Alarcón-Bustamante (2022).

As it can be appreciated, the missing outcome problem can be reduced to make assumptions about its distribution. The different ways to deal with the missing outcomes will conduct different estimations of the regression coefficients. Thus, conclusions about them can be seriously affected. In this context, different authors have tackled this problem with different procedures by considering some distributional assumptions or missing data mechanisms (see for instance Crambes and Henchiri, 2019; Veras et al., 2020). Our interest is to know the impact of missing outcomes on estimating the regression coefficients, but the problem is to find correct and useful restrictions for the missing outcomes (Manski, 1989). In this note, we illustrate the impact of the presence of missing outcomes in the estimation of the regression coefficients by using weaker assumptions as the above-mentioned ones; thus, instead of reporting a point estimation of the regression coefficients, we provide an interval containing information about the parameter of interest (Tamer, 2010). Thus, the interval contains all the plausible solutions consistent with the assumptions the researcher considers credible (Manski, 1989). In particular, based on the interval defined for the conditional expectation of the outcome for each possible value $\mathbf{x}$ (see details in Section 2.1), we propose to use the results of Stoye (2007) to obtain a region for the regression coefficients when a multiple linear regression model is considered.

In this study, we illustrate the results by analyzing the effect of selection factors over the performance of enrolled applicants in the Chilean university admission process. The GPA in the first year of the university is considered as the outcome. This quantity is not observed in the non-enrolled applicants but selection factors scores are observed for all of them and

the enrolled ones. In this case, it is clear that the selection factors are important in the selection process because they are used in the following way: better scores in the selection test, it is supposed a better performance (GPA in our case) at the university. Note that, the missing values of the GPA are related to the reason of why they are missing: scores lower than a cut-off is obtained. In consequence, the primary assumption of missing at random is not achieved - the value of the missing variable is not related to the reason why it is missing (Little and Rubin, 2002). In this context, thinking about an ignorability assumption for learning about the conditional mean of the GPA is implausible. Thus, we use assumptions based on the context of the problem to find the mentioned interval.

2. Material and Methods

2.1 Methodology

We know that the regression coefficients must be estimated from a regression model, namely $E(Y \mid \mathbf{X} = \mathbf{x})$. In this note we consider a linear regression. However, due to the presence of missing outcomes, this regression is affected and consequently the estimation of the coefficients too. In order to better understand the anatomy of the problem, let us denote the random variable $Z = 1$ if the outcome is observed and $Z = 0$ if the outcome is not observed. By using the law of total probability (Kolmogorov, 1950) the regression function evaluated at $\mathbf{X} = \mathbf{x}$ decomposes as

$$\begin{aligned} E(Y \mid \mathbf{X} = \mathbf{x}) = {} & E(Y \mid \mathbf{X} = \mathbf{x}, Z = 0)P(Z = 0 \mid \mathbf{X} = \mathbf{x}) \\ & + E(Y \mid \mathbf{X} = \mathbf{x}, Z = 1)P(Z = 1 \mid \mathbf{X} = \mathbf{x}), \end{aligned} \tag{2.3}$$

where $E(Y \mid \mathbf{X} = \mathbf{x}, Z = z)$ is the expectation of Y conditional on $\mathbf{X} = \mathbf{x}$ and $Z = z$; $P(Z = z \mid \mathbf{X} = \mathbf{x})$ for $z \in \{0, 1\}$ is the distribution of Z conditional on $\mathbf{X} = \mathbf{x}$. In what follows, we will denote $P(Z = 0 \mid \mathbf{x}) = m(\mathbf{x})$. Note that, $E(Y \mid \mathbf{X} = \mathbf{x}, Z = 0)$ is impossible to be estimated from the data, because it depends on non-observed outcomes. In the words of Koopmans (1949), the consequence of this non-observability is that the conditional mean, $E(Y \mid \mathbf{X} = \mathbf{x})$ is not identified.

As was mentioned before, currently used procedures for dealing with the problem of missing outcomes are related to finding restrictions to identify this conditional mean. In fact, deleting those unit samples with missing outcomes is analogous to imposing that there are no missing outcomes in the sample, i.e., it is assumed that $P(Z = 0 \mid \mathbf{X} = \mathbf{x}) = 0$. Using this fact, Equation (2.3) is reduced to

$$E(Y \mid \mathbf{X} = \mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x}, Z = 1),$$

and then $E(Y \mid \mathbf{X} = \mathbf{x})$ can be identified through the available information. Consequently, the regression coefficients can be estimated through the coefficients of the observed regression. The same conclusion can be obtained when the identification restriction given in Equation (1.2) is used. Under these approaches, the regression can be point identified (Manski, 1993); thus the coefficients of it can be point estimated.

Following the ideas of Manski (1989), if Y_0 and Y_1 are the minimum and maximum attainable value of the outcome, although $E(Y \mid \mathbf{X} = \mathbf{x}, Z = 0)$ is impossible to be estimated, we can assure that $Y_0 \leq E(Y \mid \mathbf{X} = \mathbf{x}, Z = 0)$ and $E(Y \mid \mathbf{X} = \mathbf{x}, Z = 0) \leq Y_1$. Thus, from Equation (2.3) it follows that

$$\underline{E}(Y \mid \mathbf{X} = \mathbf{x}) \leq E(Y \mid \mathbf{X} = \mathbf{x}) \leq \overline{E}(Y \mid \mathbf{X} = \mathbf{x}), \tag{2.4}$$

where

$$\underline{E}(Y \mid \mathbf{X}=\mathbf{x}) = E(Y \mid \mathbf{X}=\mathbf{x}, Z=1)(1-m(\mathbf{x})) + Y_0 m(\mathbf{x})$$
$$\overline{E}(Y \mid \mathbf{X}=\mathbf{x}) = E(Y \mid \mathbf{X}=\mathbf{x}, Z=1)(1-m(\mathbf{x})) + Y_1 m(\mathbf{x})$$

This interval is understood as all the possible values of the mean of the outcome - conditional on $\mathbf{X}=\mathbf{x}$ - which are consistent with the observed data and the missing data generation process.

Note that the previous interval gives us information about the regression, but not for the coefficients that govern it. By considering Equation (2.4), Stoye (2007) proposed intervals for the regression coefficients, which we will apply in the linear regression model context: if the conditional mean $E(Y \mid \mathbf{X}=\mathbf{x})$ is bounded by $\Big(\underline{E}(Y \mid \mathbf{X}=\mathbf{x}); \overline{E}(Y \mid \mathbf{X}=\mathbf{x})\Big)$ — as in Equation (2.4), for instance — then the interval that contains any linear combination of the regression parameters, namely $c\boldsymbol{\beta}$, with $c \in \mathrm{R}^k$, is given by

$$c\Big[\int \mathbf{x}^\top \mathbf{x} dF_{\mathbf{x}}\Big]^{-1} \int \mathbf{x}^\top \underline{g}(\mathbf{x}) dF_{\mathbf{x}} \leq c\boldsymbol{\beta} \leq c\Big[\int \mathbf{x}^\top \mathbf{x} dF_{\mathbf{x}}\Big]^{-1} \int \mathbf{x}^\top \overline{g}(\mathbf{x}) dF_{\mathbf{x}}, \tag{2.5}$$

where $F_{\mathbf{x}}$ is the distribution function of $\mathbf{X}=\mathbf{x}$; $\underline{g}(\mathbf{x})$, and $\overline{g}(\mathbf{x})$ are defined such that

- if $c\Big(\int \mathbf{x}^\top \mathbf{x} dF_{\mathbf{x}}\Big)^{-1} \mathbf{x}^\top > 0$, then $\underline{g}(\mathbf{x}) = \underline{E}(Y \mid \mathbf{X}=\boldsymbol{x})$, and $\underline{g}(\mathbf{x}) = \overline{E}(Y \mid \mathbf{X}=\boldsymbol{x})$;
- if $c\Big(\int \mathbf{x}^\top \mathbf{x} dF_{\mathbf{x}}\Big)^{-1} \mathbf{x}^\top \leq 0$, then $\underline{g}(\mathbf{x}) = \overline{E}(Y \mid \boldsymbol{x})$, and $\underline{g}(\mathbf{x}) = \underline{E}(Y \mid \mathbf{X}=\boldsymbol{x})$.

Considering this result, if the interest is computing the interval for β_j, with $j \in \{1, \ldots, k\}$, it is enough to consider c as the j-th canonical vector of R^k. It is important to highlight that in this study we used the empirical distribution of both $\mathbf{X}$ and the conditional distribution of the outcome given the covariates.

2.2 Data description

The university admission system in Chile considers different selection factors. Some of them are related to the performance of applicants in high school namely, the ranking and high school GPA (HS-GPA). Other factors are related to selection tests: two mandatory selection tests (Mathematics and Language and Communication) and two elective ones (Sciences and History, Geography and Social Sciences). All selection factors are in a 150-850 scale and a unique application score is obtained from them to the application process.

The used dataset contains information about the ranking, HS-GPA scores and mandatory admission selection test scores (Mathematics and Language) of all applicants to the School of Biological Sciences at a Chilean university. In addition, for the applicants that were enrolled, the GPA at the first year is also registered.

The dataset contains information about 319 applicants to the mentioned school. The 60.82% of them were not enrolled. Table 1 provides descriptive statistics for scores of each selection factor by applicants status (enrolled - non-enrolled). The dispersion of all selection factor scores are lower for non-enrolled applicants than for the enrolled ones. The mean of enrolled applicants tends to be higher than non-enrolled ones for all selection factors, relation that also happens with medians. This is clearer in Figure 2 where it can be appreciated that in all selection factors the distribution of the scores are fairly symmetric. Regarding to enrolled applicants, the minimum and maximum of the GPA are 2.97 and 6.51, respectively. The distribution of the GPAs is quite symmetric with a median of 5.12 and mean of 5.07. The standard deviation of the GPA is 0.69.

Table 1. Summary statistics of selection factor scores by applicant status.

	Enrolled					Non-enrolled				
Factor	Min	Median	Mean	Max	SD	Min	Median	Mean	Max	SD
Mathematics	543	664	663.86	822	54.19	536	623	627.74	752	42.21
Language	517	663	662.36	807	68.13	489	612	617.00	760	50.23
Ranking	564	758	747.58	850	74.67	523	685	681.89	850	72.32
HS-GPA	564	708	705.10	816	58.72	523	647	647.62	785	53.00

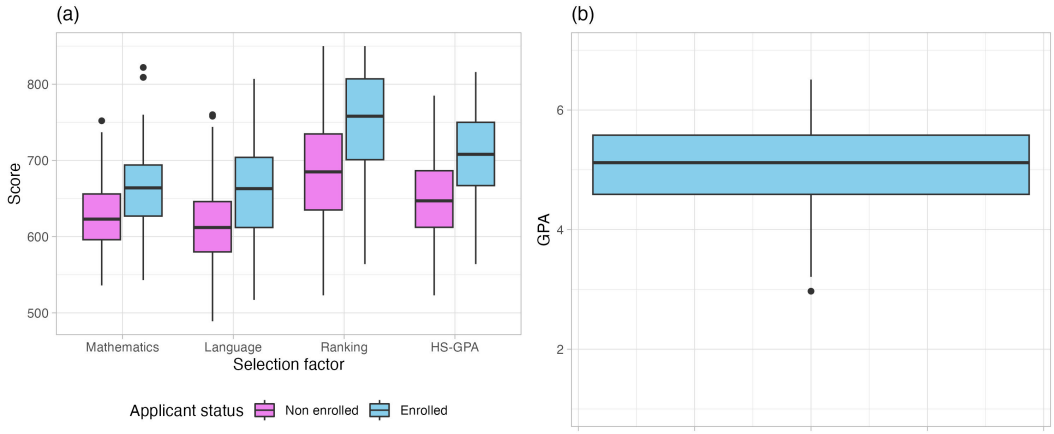

Figure 2. (a) Distributions of selection factor scores by applicant status. (b) Distribution of enrolled applicants GPAs.

3. Results

In order to evaluate the effect of the selection factors over the GPA, in Chile, a multiple linear regression model is typically used. High school performance selection factors and mandatory selection tests are observed for all applicants to the university admission system; thus, these factors are considered in the regression model; for more details, see (Manzi et al., 2008). The analysis is made by using only the complete available information, i.e. considering only those enrolled applicants.

For this purpose, it is considered Y as the GPA, and the covariates as the scores in Mathematics and Language tests in addition to the Ranking and the HS-GPA scores. The used linear regression model is given by

$$E(Y \mid \mathbf{X}) = \beta_0 + \beta_1 \text{Math} + \beta_2 \text{Language} + \beta_3 \text{Ranking} + \beta_4 \text{HS-GPA}.$$

Due to the differences in the scale scores of both GPA and selection factors, to compute point estimation of the regression coefficients and the bounds associated with them, we considered their standardized values.

The computational routine was implemented in the software R, version 4.3.0 (R Core Team, 2023) and it is available online in Alarcón-Bustamante et al. (2023). An advantage of the procedure is that the computational cost is almost zero.

Table 2 shows the results by using the least square estimator (OLS, Rao, 1973) by using information about the enrolled applicants only. In order to be consistent with the estimation of the bounds for the regression coefficients, we did not assume any distribution in the computation of these regression coefficients. Thus, confidence intervals are not shown for this point estimation.

Note that, almost coefficients are positive, suggesting - preliminary- a positive impact of the selection factors over the GPA on the first year at the university. Thus, mathematics and language test scores, as well as scores associated with the HS-GPA, are such that better scores will be translated as better performance. This is not the case for the coefficient associated with the Ranking of the applicant. However, its impact is relatively small (it is near to be zero); it could be that the impact of this selection factor over the GPA is near null.

Table 2. OLS point estimations for selection factor regression coefficients over GPA.

Factor	Mathematics	Language	Ranking	HS-GPA
Point estimation	0.6340	0.1483	-0.0991	0.2959

In Table 3 the results of the interval computed from Equation (2.5) are shown. The interval was obtained by considering the minimum and maximum attainable GPA, i.e., $Y_0 = 1.0$ and $Y_1 = 7.0$. Thus, the interval contains all the plausible values of the regression coefficients that are compatibles with all the possible regression models. Note that these intervals are quite large and include the zero. Then, the direction of the impact (positive or negative) can not be identified, i.e., due to the large presence of missing outcomes, it is not possible to know if the selection factors will impact in a positive or negative way the GPA. In contrast, when only available data is used, all the coefficients have a clear sign and the direction of the impact can be interpreted in the traditional way. If an interval did not contain the zero value, thus its interpretation would be unambiguous. For instance, if the region is fully positive (negative), we have no doubts that the impact would be positive (negative) over the GPA. However, this is not the case.

A first reaction to the results shown in Table 3 could be that the imposed restriction is no-viable because the probability of obtaining the minimum or maximum GPA is very low. The advantage of the proposed strategy is that we can use other restrictions based on the context of the problem, for instance: instead of using the minimum and maximum attainable GPA, we can use the minimum and maximum observed GPA (2.97 and 6.1, respectively). In Table 4 the resulting interval using this restriction is shown. As it can be appreciated, although under this restriction we reduced the width of the interval for each selection factor coefficient, the interpretation in terms of the impact of the selection factors over the GPA do not change.

Table 3. Bounds on regression coefficients for the regression of the GPA over selection factors assuming that the non-observed conditional expectation is in between the minimum and maximum attainable GPA.

Factor	Mathematics	Language	Ranking	HS-GPA
Lower Bound	-1.8285	-1.9536	-5.1242	-4.7223
Upper bound	2.6538	2.3452	5.1251	5.4032

Table 4. Bounds on regression coefficients for the regression of the GPA over selection factors assuming that the non-observed conditional expectation is in between the minimum and maximum observed GPA.

Factor	Mathematics	Language	Ranking	HS-GPA
Lower Bound	-1.0086	-1.1708	-3.0592	-2.8006
Upper bound	1.6359	1.3655	2.9879	3.1734

4. Discussion and conclusion

Throughout the manuscript, we discussed the different ways to deal with the missing outcomes when the goal is estimating the regression coefficients. Traditional strategies are based on missing at random assumptions; however, as was mentioned, in the selection process this assumption does not make sense because the missing data-generation process depends on the selection factors, causing a not missing at random process (Muthén and Hsu, 1993). In this context, we have proposed to use a partial identification technique to compute bounds for the parameters that contains all the possible values for the regression coefficients of the selection factors. The bounds were estimated by assuming that the conditional mean of the non-enrolled applicants should be in between the minimum and the maximum attainable/observed GPA. Although other assumptions can be used to point identify the regression coefficients, the used assumptions for do it are rarely justified and they not consider the context of the problem. In contrast, the used assumptions in this manuscript are fully justifiable because they are based on the support of the GPA.

From the results of Tables 2, 3, and 4, when a point estimation for the regression coefficients is used under an ignorability assumption, the conclusions about the effect of the selection factors over the GPA are totally different when the bounds are considered. In fact, due to the missing outcomes, it is impossible to know if the effect of the regression coefficients is positive or negative.

The approach of combining the data with suitable restrictions for the non-observed data to find an interval containing plausible values for the parameters of interest is known as partial identifiability. This approach has the advantage of imposing restrictions on the non-observed values by using more reasonable (and justifiable) assumptions than those commonly used in the literature. Because the partial identification approach allows for making assumptions based on the context of the problem, it has been used in other fields as intergenerational mobility, test equating, surveys analysis with missing outcomes, decision theory, educational measurement, among others (see for instance San Martín and González, 2022; Pepper, 2000; San Martín and Alarcón-Bustamante, 2022; Manski, 2016; Stoye, 2011; Alarcón-Bustamante et al., 2020). Finally, it is important to highlight that there is a no unique way to model a problem (e.g., by using traditional distributional assumptions). In fact, we used assumptions related to the problem that we intend to model, considering the physiognomy of it. From our viewpoint, there is a necessity to make people know that there is another way to deal with missing values. However, if the researcher is ready to believe in the assumptions rarely justified, it must be willing to assume that stronger assumptions yield conclusions that are more powerful but less credible (the law of decreasing credibility Manski, 2003, p.1).

Author contributions Conceptualization, E.A-B., I.V., E.S.M.; methodology, E.A-B., I.V., E.S.M.; software, E.A-B., I.V., E.S.M.; validation, E.A-B., I.V., E.S.M.; formal analysis, E.A-B., I.V., E.S.M.; investigation, E.A-B., I.V., E.S.M.; data curation, E.A-B., I.V., E.S.M.; writing—original draft preparation, E.A-B., I.V., E.S.M.; writing—review and editing, E.A-B., I.V., E.S.M.; visualization, E.A-B., I.V., E.S.M.; supervision, E.A-B., I.V., E.S.M. All authors have read and agreed to the published version of the paper.

Acknowledgements The authors would like to thank the Editors-in-Chief and the anonymous reviewers for their valuable comments and suggestions which improved substantially the quality of this paper.

Funding E.A-B. was partially funded by the Postdoctoral FONDECYT grant No. 3220422. Also, E.A-B. was partially funded by ANID – Millennium Science Initiative Program - Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI) NCS2021072 and by the FONDEF IDeA I+D ID22I10228 project. E.S.M. was partially funded by ANID – Millennium Science Initiative Program - Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI) NCS2021072.

Conflicts of interest The authors declare no conflict of interest.

References

Alarcón-Bustamante, E., 2022. Ignorar o no ignorar, esa es la cuestión. Cuadernos de Beauchef. Ciencia, Tecnología y Cultura. 6(1), 15-33.

Alarcón-Bustamante, E., 2023. Ignorabilidad: un supuesto clave en la dinámica de la inferencia estadística en ciencias de la salud. Inferencias: Boletín de Bioestadística, 7, 11-13.

Alarcón-Bustamante, E., San Martín, E., and González, J., 2020. Prefictive validity under partial observability. In Wiberg, M., Molenaar, D., González, J., Böckenholt, U., and Kim, JS. (Eds), Quantitative Psycology. Springer, Cham, pp. 135-145.

Alarcón-Bustamante, E., Varas, I.M., and San Martín, E., 2023. ImpactMissingOutcomesRegression. Available at https://github.com/edalarconb/ImpactMissingOutcomesRegression.

Campbell, J.P., and Knapp, D.J., 2001. Exploring the limits of personnel selection and classification. Lawrence Erlbaum Associates, New Jersey, US.

Carpenter, J.R., and Kenward, M.G., 2012. Multiple imputation ant its applications. Wiley, New York.

Chema, J., 2014. A review of missing data handling methods in education research. Review of Educational Research, 84(2), 487-508.

Crambes, C., and Henchiri, Y., 2019. Regression imputation in the functional linear model with missing values in the response. Journal of Statistical Planning and Inference, 201, 103-119.

Davern, M., Blewett, L.A., Bershadsky, B., and Arnold, N., 2004. Missing the mark? imputation bias in the current population survey's state income and health insurance coverage estimates. Journal of Official Statistics, 20(3), 519-549.

Department of Defense, 1984. Test Manual for the Armed Services Vocational Aptitud Battery. North Chicago, IL: United States Military Entrance Processing Command.

Geiser, S., and Studley,R., 2002. UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California. Educational Assessment, 8(1), 1-26.

Hirano, K., and Imbens, G.W., 2004. The propensity score with continuous treatments. In Shewhart, W.A., and Wilks, S.S. (Eds), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family, Chapter 7, 73-84. Wiley, New York.

Hsu, J.W.Y., 1995. Sampling behaviour in estimating predictive validity in the context of selection and latent variable modelling: A monte carlo study. British Journal of Mathematical and Statistical Psychology, 48(1), 75-97.

Imbens, G., 2000. The role of the propensity score in estimating dose-response functions. Biometrika, 87(3), 706-710.

Kambourov, G., and Manovskii, I., 2013. A cautionary note on using (March) current population survey and panel study of income dynamics data to study worker mobility. Macroeconomic Dynamics, 17, 172-194.

Kolmogorov, A.N., 1950. Foundations of the Theory of Probability. Chelsea Publications, New York.

Koopmans, T.C., 1949. Identification problems in economic model construction. Econometrica, 17(2), 125-144.

Lee, S., and Mendoza, J., 2022. The biasing effects of selection and attrition on estimating the mean. British Journal of Mathematical and Statistical Psychology, 76(1), 106-130.

Little, R.J., and Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley, New York.

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

Manski, C., 1993. Identification problems in the social sciences. Sociological Methodology, 23, 1-56.

Manski, C., 2003. Partial Identification of Probability Distributions. Springer, New York.

Manski, C., 2007. Identification for Prediction and Decision. Harvard University Press, Cambridge, MA.

Manski, C., 2016. Credible interval estimates for official statistic with survey nonresponse. Journal of Econometrics, 191, 293-301.

Manzi, J., Bravo, D., del Pino, G., Donoso, G., Martínez, M., and Pizarro, R., 2008. Estudio de la validez predictiva de los factores de selección a las universidades del consejo de rectores, admisiones 2003 al 2006. Technical report, Comité Técnico Asesor, Honorable Consejo de Rectores de las Universidades Chilenas.

Muthén, B.O., and Hsu, J.-W.Y., 1993. Selection and predictive validity with latent variable structures. British Journal of Mathematical and Statistical Psychology, 46(2), 255-271.

Pepper, J., 2000. The intergenerational transmission of Welfare receipt: A nonparametric bounds analysis. The Review of Economics and Statistics, 82, 472-488.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

Rao, C.R., 1973. Linear statistical inference and its applications. Wiley, New York.

Rosenmbaum, P., and Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Rubin, D.B., 2009. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

San Martín, E., and Alarcón-Bustamante, E., 2022. Dissecting Chilean surveys: the case of missing outcomes. Chilean journal of Statistics, 13(1) 17-45.

San Martín, E., and González, J., 2022. A critical view on the NEAT equating design: Statistical modelling and identifiability problems. Journal of Educational and Behavioral Statistics, 47(4), 406-437.

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

Stoye, J. 2011. Minimax regret treatment choice with covariates or with limited validity of experiments. Journal of Econometrics, 166, 138-156.

Tamer, E., 2010. Partial identification in econometrics. Annual Review of Economics, 2(1), 167–195.

Van Buuren, S., 2012. Flexible imputation of missing data. CRC Press, Boca Raton, FL.

Veras, M.B.A., Mesquita, D.P.P., Mattos, C.L.C., and Gomes, J.P.P., 2020. A sparse linear regression model for incomplete datasets. Pattern Analysis and Applications, 23, 1293-1303.

Zimmermann, S., Klusmann, D., and Hampe, W. 2017. Correcting the predictive validity of a selection test for the effect of indirect range restriction. BMC Medical Education, 17(1), 2460.