Predictive Validity Under Partial Observability

Joint work with Ernesto San Martín and Jorge González

Eduardo Alarcón-Bustamante

[1] Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Chile
[2] LIES Laboratorio Interdisciplinario de Estadística Social, Pontificia Universidad Católica de Chile, Chile

February 5, 2021

Laboratorio Interdisciplinario de Estadística Social

Outline

Motivation

- The analysis of the relationship between **Test Scores** and **Graded Point Average (GPA)** provide an important source of predictive validity evidence of a University Selection Test.

Motivation

Laboratorio Interdisciplinario de Estadística Social

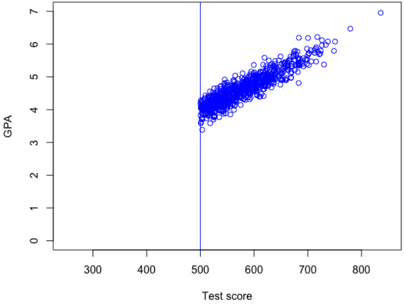

Figura: Real scenario

Warning!

The GPA is observed only in the selected group, whereas the scores of the selection test are observed for the whole population of applicants.

Motivation

Laboratorio Interdisciplinario de Estadística Social

Statistical procedures used for the evaluation of the predictive validity:

- Regression models [1] with truncated distributions (Nawata, 1994; Heckman, 1976, 1979; Marchenko and Genton, 2012), and
- Corrected Pearson correlation coefficient (Thorndike, 1949; Pearson, 1903; Mendoza and Mumford, 1987; Lawley, 1943; Guilliksen, 1950).

Assumption: a prior knowledge for the performance of the whole population, which can be incompatible with the reality (Manski, 2003).

[1] It is formally known as Conditional Expectation of the outcome, Y, given the test scores, X. The conditional expectation is denoted by $\mathbb{E}(Y|X)$

Motivation

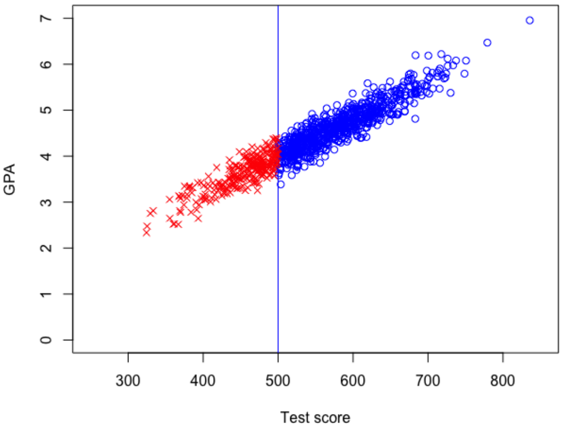

Figura: Assumption for current solutions.

Motivation

Laboratorio
Interdisciplinario de
Estadística Social

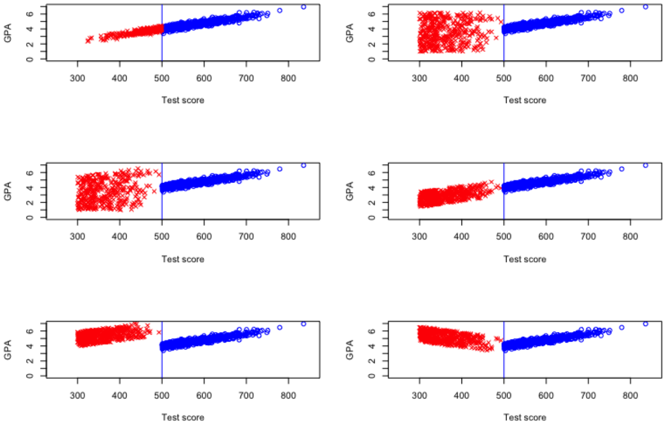

Figura: Some possible scenarios.

Motivation

Laboratorio Interdisciplinario de Estadística Social

In the educational measurement literature, the predictive validity is typically analyzed through the *marginal effect*, that is,

$$M.E^{X} = \frac{d\mathbb{E}(Y|X)}{dX},$$

where, by the Law of Total Probability [2]

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

[2] $Z = 1$ if the outcome is observed and $Z = 0$ otherwise.

Motivation

Laboratorio Interdisciplinario de Estadística Social

In the educational measurement literature, the predictive validity is typically analyzed through the *marginal effect*, that is,

$$M.E^X = \frac{d\mathbb{E}(Y|X)}{dX},$$

where, by the Law of Total Probability [2]

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

Warning!

As a consequence of the partial observability of the GPA, the conditional expectation is **not identified**. Hence, the marginal effect is not identified either.

[2] $Z = 1$ if the outcome is observed and $Z = 0$ otherwise.

Motivation

- **Goal**: To learn about the predictive validity of selection tests without considering a prior structure for the performance of the whole population.
- **Strategy**: To make assumptions weaker than current solutions and to find an **Identification region** of values for the marginal effects. (Manski, 1993, 2005, 2007, 2013).

Partial Identification solution

As it was mentioned before,

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|X, Z=0)\mathbb{P}(Z=0|X) + \mathbb{E}(Y|X, Z=1)\mathbb{P}(Z=1|X)$$

Then,

$$\begin{aligned} M.E^X \quad = \quad & \frac{d\mathbb{E}(Y|X, Z=0)}{dX}\mathbb{P}(Z=0|X) + \frac{d\mathbb{E}(Y|X, Z=1)}{dX}\mathbb{P}(Z=1|X) + \\ & [\mathbb{E}(Y|X, Z=1) - \mathbb{E}(Y|X, Z=0)]\frac{d\mathbb{P}(Z=1|X)}{dX} \end{aligned} \tag{2}$$

Partial Identification solution

Assumptions

- If $Y \in [y_0, y_1]$, then,

$$y_0 \leq \mathbb{E}(Y|X, Z = 0) \leq y_1$$

- The marginal effect for the non-selected population exist[3]

$$D_{0x} < \frac{d\mathbb{E}(Y|X, Z = 0)}{dX} \leq D_{1x}$$

[3] if this population had been selected, the score of the selection test would have predicted the outcome with an associated marginal effect

By considering that **the selection process is correct**[4], we can assume that:

- the marginal effect in the non-observed group is positive, i.e.,

$$0 < \left. \frac{d\mathbb{E}(Y|X, Z=0)}{dX} \right|_{X=x}.$$

- The marginal effect in the non-observed group can not be higher that the maximum observed marginal effect, i.e.,

$$\left. \frac{d\mathbb{E}(Y|X, Z=0)}{dX} \right|_{X=x} \leq \max_{x \in X} \left\{ \left. \frac{d\mathbb{E}(Y|X, Z=1)}{dX} \right|_{X=x} \right\}$$

[4] The selection test is such that higher scores would translate to higher values of the outcome

Partial Identification solution

Identification Bonds for the Marginal Effect

Remember that

$$M.E^X = \frac{d\mathbb{E}(Y|X,Z=0)}{dX}\mathbb{P}(Z=0|X) + \frac{d\mathbb{E}(Y|X,Z=1)}{dX}\mathbb{P}(Z=1|X) + [\mathbb{E}(Y|X,Z=1) - \mathbb{E}(Y|X,Z=0)]\frac{d\mathbb{P}(Z=1|X)}{dX} \tag{3}$$

Then, According to the ideas of Manski (1989)

$$M.E^{X=x} \in \left(\left.\frac{d\mathbb{E}(Y|X,Z=1)}{dX}\right|_{X=x} \mathbb{P}(Z=1|X=x) + [\mathbb{E}(Y|X=x,Z=1) - y_0] \left.\frac{d\mathbb{P}(Z=1|X)}{dX}\right|_{X=x};\right.$$

Partial Identification solution

Identification Bonds for the Marginal Effect

Laboratorio Interdisciplinario de Estadística Social

Remember that

$$M.E^X = \frac{d\mathbb{E}(Y|X,Z=0)}{dX}\mathbb{P}(Z=0|X) + \frac{d\mathbb{E}(Y|X,Z=1)}{dX}\mathbb{P}(Z=1|X) + [\mathbb{E}(Y|X,Z=1) - \mathbb{E}(Y|X,Z=0)]\frac{d\mathbb{P}(Z=1|X)}{dX} \tag{3}$$

Then, According to the ideas of Manski (1989)

$$M.E^{X=x} \in \left(\left.\frac{d\mathbb{E}(Y|X,Z=1)}{dX}\right|_{X=x}\mathbb{P}(Z=1|X=x) + [\mathbb{E}(Y|X=x,Z=1) - y_0]\left.\frac{d\mathbb{P}(Z=1|X)}{dX}\right|_{X=x}; \right.$$
$$\max_{x\in X}\left\{\left.\frac{d\mathbb{E}(Y|X,Z=1)}{dX}\right|_{X=x}\right\}\mathbb{P}(Z=0|X=x) + \mathbb{P}(Z=1|X=x)\left.\frac{d\mathbb{E}(Y|X,Z=1)}{dX}\right|_{X=x}$$
$$\left. + [\mathbb{E}(Y|X=x,Z=1) - y_1]\left.\frac{d\mathbb{P}(Z=1|X)}{dX}\right|_{X=x}\right] \tag{4}$$

(Manski, 1989).

Application

Laboratorio Interdisciplinario de Estadística Social

Predictive validity of two mandatory University Selection Tests in Chile, over the GPA of students in the first year in a Chilean university

- $\mathbb{E}(Y|X, Z = 1)$ was estimated by an adaptive local linear regression model using a symmetric Kernel.
- $\mathbb{P}(Z = 1|X)$ was estimated by using a Probit model.

Application

Results

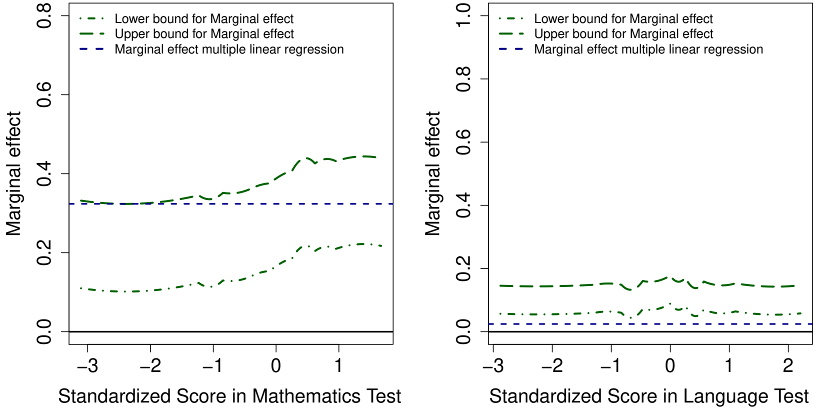

Figura: Identification bounds for the Marginal Effect

- We have presented a method that allows to **learn** about the predictive validity of selection tests through the marginal effect under partial observability.
- Our proposal has the advantage of not assuming any parametric structure for the non-observed group, as we only use desired properties of the selection tests.
- Our proposal has the advantage of interpret the marginal effect as a function of X and not as a number necessarily Alarcón-Bustamante et al. (In press).
- Extending the approach for the scenario where information of more universities is available is a topic in progress.

Thanks for your attention

Eduardo Alarcón-Bustamante
esalarcon@mat.uc.cl
http://alarcon-bustamante.cl

Laboratorio Interdisciplinario de Estadística Social,
Faculty of Mathematics, Pontificia Universidad Católica de Chile.
http://lies.mat.uc.cl

This work was funded by the National Agency for Research and Development (ANID) / Scholarship Program / Doctorado Nacional / 2018-21181007

References I

Alarcón-Bustamante, E., San Martín, E., and González, J. (In press). On the marginal effect under partitioned populations: Definition and interpretation. In Wiberg, M., Molenaar, D., González, J., Böckenholt, U., and Kim, J.-S., editors, *Quantitative Psychology*. Springer International Publishing.

Guilliksen, H. (1950). *Theory of mental tests*. New York, John Willey and Sons.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 46:931–961.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

References II

Lawley, D. (1943). IV.-A note on Karl Pearson s selection formulae. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Science*, 62(1):28–30.

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*. New York: Springer.

Manski, C. (2005). *Social Choice with Partial Knowledge of Treatment Response*. Princeton University Press, New Jersy, 1 edition.

Manski, C. (2007). *Identification for Prediction and Decision*. Harvard University Press.

Manski, C. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

References III

Marchenko, Y. V. and Genton, M. G. (2012). A Heckman Selection-t Model. *Journal of the American Statistical Association*, 107(497):304–317.

Mendoza, J. and Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, 12(3):282–293.

Nawata, K. (1994). Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two-step estimator. *Economics Letters*, 45(1):33–40.

Pearson, K. (1903). Mathematical contribution to the theory of evolution-XI on the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London*, 200(Ser. A):1–66.

Thorndike, R. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.