Determinación del posible sesgo de selección en pruebas mediante la metodología de Heckman

José Andrey Zamora Araya [1]

Resumen

Existen pruebas en el ámbito educativo que son de carácter voluntario. En dichas pruebas surge la duda si los sujetos que realizan la prueba tienen un comportamiento similar a aquellos que no realizan la prueba. La metodología de Heckman explica que el sesgo generalmente aparece en muestras no aleatorias para estimar las relaciones de comportamiento como un error ordinario de especificación por lo que plantea un estimador de dos pasos que facilita la utilización de un método de regresión simple para estimar las funciones de comportamiento. La idea es usar el método para determinar si hay o no sesgo en la selección de la muestra en el ámbito de pruebas educativas.En el caso de la PDM de la UNA, el análisis no muestra evidencia de un sesgo de selección.

Palabras clave: Metodología de Heckman, sesgo de no respuesta, pruebas educativas, matemática, pruebas diagnósticas.

Abstract

There is test in education that are voluntary. In these tests, the question arises whether the test subjects performing a similar behavior to those who do not do. Heckman methodology explained appears a bias nonrandom samples to estimate behavioral relationships as an ordinary of error specification, therefore poses a two-step estimator which facilitates the use of a simple regression method for estimating functions of behavior. The idea is to use the method to determine whether exist or not selection bias in the sample in the field of educational test.

Keywords:Heckman methodology, nonresponse bias, educational testing, mathematics, diagnostic tests.

Modalidad: Ponencia

[1] Escuela de matemáticas Universidad Nacional y Escuela de Estadística Universidad de Costa Rica. Costa Rica. andreyzamora@gmail.com

- **Introducción**

La escuela de matemática de la UNA desde el 2010 ha venido aplicando la prueba de diagnóstico en matemática (PDM) a todos los estudiantes de primer ingreso que tengan en su plan de estudios al menos un curso de matemática. Esto con el fin de evaluar el nivel de conocimientos matemáticos con que ingresan los estudiantes a la institución.

Dicha prueba surge a raíz de los malos resultados académicos de los estudiantes de nuevo ingreso evidenciados sobre todo en los cursos introductorios, particularmente en el curso de matemática general (MAX 0 84) el cual junto con matemática para informática I constituye uno de los principales cursos que ofrece la escuela de matemática a la comunidad universitaria.

El desempeño en el curso MAX 084 es muy bajo, pues el porcentaje de aprobación en el curso en los últimos años ronda el 35%. Entre los docentes que imparten los cursos existen varias hipótesis relacionadas con la poca preparación recibida en el colegio y que se ve reflejada en las aulas. Al respecto, se ha considerado reformular el curso de Matemática General o bien ofrecer una materia que refuerce los conocimientos básicos de aritmética y álgebra, con el fin de que puedan llevar con éxito un curso introductorio de matemáticas a nivel universitario (Zamora, 2010a y Zamora 2010b).

En este sentido la PDM pretende brindar información para poder propiciar medidas tendientes a mejorar el rendimiento de los estudiantes del curso MAX 084, ya que es el principal curso de servicio de la escuela de matemática y que se le ofrece a varias unidades académicas.

No obstante, la PDM es voluntaria por lo que no todos los estudiantes realizan la prueba. Esto plantea la duda sobre la representatividad de los resultados, pues se asume que los sujetos que efectivamente realizaron la prueba se comportan de una manera similar (en cuanto a conocimiento matemático) a aquellos que no realizaron la prueba.

Esta situación se le conoce como sesgo de autoselección y es frecuente en pruebas que tienen un carácter voluntario las pruebas educativas que no son obligatorias, como lo son las pruebas de diagnóstico.

Por ello se requiere de algún método (en este caso el de Heckman) que pueda determinar la presencia o no del sesgo de autoselección ya que esto podría afectar las medidas que se desean implementar sobre rendimiento académico a toda la población universitaria y no solo a los que realizan la PDM.

- **Marco Teórico**

La metodología de Heckman no es comúnmente utilizada en pruebas educativas, su uso es más popular en el ámbito económico para detectar posibles sesgos de autoselección en muestras, no obstante esto no significa que no se pueda utilizar en dichos contextos.

En general, el sesgo de selección puede surgir por dos razones. La primera, porque pueden ser resultado de la autoselección hecha por los individuos o por las unidades de datos que son objeto de la investigación. Y la segunda, por las decisiones referidas a la selección de la muestra realizada por los analistas.

Un ejemplode sesgo de selección lo constituye el mercado de salarios donde las tasas de retorno a la inversión en capital humano sugiere la distinción entre ingresos de personas proviene de diferencias en el nivel de educación alcanzado (cantidad de años de educación) o en relación con los efectos que tienen años adicionales de escolaridad sobre el ingreso salarial (años de experiencia potencial en el mercado laboral); todo lo anterior basado en la teoría del capital humano. (Forero y Gamboa, 2006)

Pero si se considera que un mayor nivel de educación o entrenamiento provoca una mayor probabilidad de participación en el mercado laboral, puesto que tiene mayores costos de oportunidad, puede existir lo que se conoce como sesgo de selección, Heckman (1979).

La técnica de Heckmanusaun modelo Probit que tiene por objeto medir la decisión de participar de acuerdo con las características individuales y del stock de capital humano de las personas. Posteriormente, se incorpora dicha estimación en la ecuación para ingresos del tipo Mincer, donde el salario depende únicamente de las dotaciones de capital humano, (Perlbach y Calderón, S.F).

Por ejemplo, los salarios de las personas no migrantesa menudoofrecen una estimación confiable del salarioque los no migrantes habrían obtenido si hubieran decidido emigrar. En este y otros ejemplos, las funciones de los salarios o los ingresos estimados en muestras seleccionadas, no hacen una estimación poblacionalrigurosa (es decir, una muestra aleatoria) de las funciones salariales.

Las comparaciones de los salarios de los migrantes con los salarios de los no migrantes dan lugar a una estimación sesgada del efecto de un "tratamiento" al azar de la migración.Heckman (1979).

Ahora bien, esta metodología puede aplicarse al caso de pruebas educativas donde los que participan son voluntarios. El posible sesgo surge al comparar los puntajes

de aquellos que asistieron al examen con los puntajes de los que decidieron no realizar la prueba.

En este documento se analizará este fenómeno para la prueba diagnóstica en matemática del 2010 (PDM 2010) con los estudiantes cuya malla curricular incluye el curso de matemática general y de esta manera evaluar el posible sesgo de autoselección.

Cómo se mencionó anteriormente, en el método de Heckman de dos etapas se especifican dos ecuaciones, una ecuación de interés que corresponde a la ecuación que se busca estimar, y una ecuación de selección o participación (ecuación auxiliar) que corresponde a un modelo de selección discreta (Probit o Logit), que mide la probabilidad de estar en la muestra. Es aconsejable en la ecuación de selección elegir variables independientes a la ecuación de interés, (González, 2010).

El modelo de Heckman (1979) de dos etapas recurre a dos ecuaciones que se describen a continuación. (Zamora, 2012)

Considere una muestra aleatoria de I observaciones. Las ecuaciones para el individuo i están dadas por:

$$Y_{1i} = X_{1i}\beta_1 + U_{1i} \quad (1)$$

$$Y_{2i} = X_{2i}\beta_2 + U_{2i} \quad (2)$$

Donde X_{ij} es un vector de 1 X K_j de regresores exógenos, β_j es un vector de K_j X 1 de parámetros y además se supone que:

$$E(U_{ji}) = 0, \text{con } E(U_{ji}U_{j'i''}) = \sigma_{ij'} \quad \text{si} \quad i = i'' \\ = 0 \quad \text{si} \quad i \neq i'' \quad (3)$$

Suponga que se busca estimar la ecuación (1), pero no existen datos para Y_1 para ciertas observaciones. Entonces surge la interrogante de porqué hay datos faltantes y se podría pensar en un posible sesgo muestral. La función de regresión poblacional para la ecuación (1) se puede escribir como:

$$E(Y_{1i}|\, X_{1i}) = X_{1i}\beta_1 \text{ con } i = 1, \dots, I \quad (4)$$

Por su parte, la función de regresión para la sub-muestra de observaciones disponibles está dada por:

$$\begin{aligned} &E(Y_{1i}|\, X_{1i}, \text{regla de selección muestral}) \\ &= X_{1i}\beta_1 + E(U_{1i}|\, \text{regla de selección muestral}) \end{aligned} \quad (5)$$

Con i = 1,…, I por convención se adopta que las primeras $I_1 < I$ son las observaciones disponibles para Y_{1i}. Para el caso general, suponga que los datos están disponibles para Y_{1i} si $Y_{2i} \geq 0$ mientras que si $Y_{2i} < 0$ las observaciones no están disponibles para Y_{1i}. La variable Y_2 representa la probabilidad de pertenecer a la muestra: esta variable tomará el valor de uno si y solo si la variable latente Y_{2i} es mayor a cero lo representa que el individuo pertenece a la muestra. Además, sólo se observará Y_{1i} cuando $Y_2 = 1$, es decir, siempre y cuando el individuo pertenezca a la muestra, en cuyo caso se podrá observar la variable Y_{1i} de la ecuación de interés (González, 2010 y Heckman, 1979).

Teniendo en cuenta tanto la ecuación de interés como la de selección, la ecuación observada es:

$$Y_{1i} = X_{1i}\beta_1 + U_{1i} \text{ si } Y_{2i} > 0 \quad (6)$$

De lo anterior se deriva que el valor esperado de la ecuación observada viene determinado de la siguiente manera:

$$E(Y_{1i}|Y_{2i}) = X_{1i}\beta_1 - \frac{\sigma_{12}}{\sigma_1}\lambda\left(Z'_i\frac{\gamma}{\sigma^2}\right) \quad (7)$$

Donde $\lambda_i = \lambda\left(Z'_i\frac{\gamma}{\sigma^2}\right)$ representa la inversa del ratio de Mills, es decir la probabilidad dadas unas características de que un individuo participe o no en la muestra.

Por otra parte, es aconsejable en la ecuación de selección elegir variables independientes de la ecuación de interés (González, 2010).

$$s = \gamma_0 + \sum_{i=1}^{n} z_i\gamma_i + v\,, v > 0 \; \textit{Ecuación de selección} \quad (8)$$

Donde la variable "s" representa la probabilidad de pertenecer a la muestra, por lo que toma los valores 1 y 0 dependiendo si el individuo pertenece o no a la muestra. Teniendo en cuenta tanto la ecuación de interés como la de selección, la ecuación observada es:

$$y = \beta_0 + \sum_{i=1}^{n} x_i\beta_i + u, \text{si } s > 0 \quad (9)$$

De lo anterior, se concluye que el valor esperado de la ecuación esperada es:

$$E(Y|s) = x_i\beta_i - \frac{\sigma_{ys}}{\sigma_s}\lambda\left(z'_i\frac{\gamma}{\sigma^2}\right) \quad (10)$$

$$y = \beta_0 + \sum_{i=1}^{n} x_i\beta_i + u \text{ con } E(u|x_1, x_2, \dots, x_k) = 0 \text{ Ecuación de interés} \quad (11)$$

Donde γ_i representa la inversa del ratio de Mills, que se puede interpretar como la probabilidad de que una persona, dada ciertas características sea o no parte de la muestra (González, 2010). Además se supone que los términos de error se comportan de la siguiente manera:

$$u_i \sim N(0, \sigma^2_u) v_i \sim N(0,1) \qquad Corr(u_i, v_i) \neq 0 \qquad (12)$$

Heckman (1979) explica que el sesgo generalmente aparece en muestras no aleatorias para estimar las relaciones de comportamiento como un error ordinario de especificación, por lo que plantea un estimador de dos pasos que facilita la utilización de un método de regresión simple para estimar las funciones de comportamiento.

En este documento se usa el método para determinar si hay o no sesgo en la selección de la muestra de estudiantes que realizaron el diagnóstico, para ello se eligieronalgunas variables asociadas como sede donde se realiza la prueba y carrera elegida.

Las razones por las cuáles se decidió elegir dichas variables son las siguientes, para la variable carrera, se conoce que existen puntajes de ingreso más altos en unas carreras que otras. Se sospecha que los estudiantes cuyas carreras tienen puntajes de ingreso más bajo son más propensos a no presentarse a realizar el PDM.

La zona de residencia indudablemente tiene un impacto a la hora de realizar la PDM, pues tradicionalmente el porcentaje de estudiantes que realiza la prueba en sedes regionales es menor que en la sede central.

En cuanto a las calificaciones de secundaria y la PAA, su selección se debe a que es conocido que las notas previas como las de colegio y la de los exámenes de admisión son buenos predictores del rendimiento académico. (Aitken, 1982; Armenta et al, 2008; Cascón, 2000; Donosso&Schiefelbein, 2007; Guillén & Chinchilla, 2005 y Piñero & Rodríguez, 1998).

Finalmente, en el caso del índice de desarrollo social (IDS) es una aproximación al nivel económico del estudiante que investigaciones como las de Abbate (2008), Creemers (2008), Himmel (2002) y Mujis (2008) han mostrado que son relevantes a la hora de predecir el rendimiento académico.

- **Metodología**

Población y unidad de estudio

La población meta a estudiar serían, todos los estudiantes de nuevo ingreso a la Universidad Nacional, cuya carrera tenga dentro de su plan de estudios el curso de Matemática General (MAX084). La unidad de estudio sería cada estudiante de nuevo ingreso la Universidad Nacional cuya carrera tenga dentro de su plan de estudios el curso de Matemática General (MAX084).

La muestra disponible son los estudiantes de nuevo ingreso que realizaron la prueba diagnóstica en matemática 2010, la cual incluye estudiantes de todas las sedes de la UNA y estudiantes de varias carreras que no deben llevar el curso de Matemática General MAX-084, donde la mayor proporción pertenece a la sede central ubicada en cantón central de Heredia, como se muestra en la tabla 1.

Tabla 1.
UNA. Estudiantes de nuevo ingreso 2010 que realizaron la PDM de acuerdo con el tipo de cursos de matemática que deben matricular.

Categoría	sede central	otras sedes	Total
Estudiantes que deben llevar al menos un curso de matemática	867	264	1131
	61%	49%	58%
Estudiantes que deben matricular MAX084	451	90	541
	59%	34%	52%

Fuente: Elaboración propia.

Descripción de la muestra

Los datos que se refieren a variables socio-demográficas, de los estudiantes que realizaron la PDM 2010, fueron facilitados por el departamento de registro de la UNA y lo concerniente a la PDM 2010, fue facilitado por la Escuela de Matemática de la UNA.

La base de datos estaba constituida originalmente por 1195 registros de los estudiantes de nuevo ingreso a la UNA, durante el primer semestre del año 2010 que realizaron el examen de diagnóstico en matemática aplicado el 03 de febrero de ese año. La base

contiene variables como zona de residencia (rural – urbano), tipo de colegio, tipo de financiamiento del colegio (público, privado o subvencionado), carrera de ingreso, sede universitaria, total de preguntas correctas, nota de examen de diagnóstico, opción seleccionada en cada uno de los ítems de la prueba, que son ítems de opción múltiple, nota en la prueba de aptitud académica. No obstante, la base contiene algunos valores faltantes, sobre todo en las variables carrera de ingreso, además muchos estudiantes, dependiendo de la carrera en que estén empadronados no requieren llevar el curso MAX084 y otros que deben llevarlo al momento de realizar el estudio no lo habían matriculado. Luego de depurar la base de datos, ésta se redujo a 417 casos.

Procesamiento: software y procedimientos

Para determinar el posible sesgo en la selección de la muestra se usará el método de Heckman de dos etapas. Para cálculos de medidas descriptivas se usara el software SPSS 17.0 y para el cálculo del método de Heckman se utiliza el programa STATA 10.0.

Para realizar el análisis primero se eligen todos los casos completos, es decir, la población de estudiantes de nuevo ingreso que deben realizar la PDM y deben matricular el curso de Matemática General. Para el caso del estudio la cantidad de estudiantes que cumplen con ambos requisitos es de 977, de los cuales solo 417 realizaron la PDM.

Para la primera etapa del método se construye la ecuación de selección con las variables que en teoría explican la posibilidad de realizar la PDM, que en este caso son sede y carrera. En el archivo de datos se coloca un 1 a los estudiantes que realizaron la prueba y un 0 a los que no y se plantea una regresión lineal cuya variable dependiente es si el estudiante realizó la prueba y cuyas variables independientes son la sede y la carrera.

En la segunda etapa se trata de estimar la nota en la PDM mediante una serie de variables explicativas como lo son la sede, la carrera, la nota del examen de admisión, la nota promedio de 4 y 5 año de colegio en materias básicas y el índice de desarrollo social medido a nivel de distrito de residencia del estudiante.

Análisis de los datos

Para el caso de la muestra de la PDM 2010 se obtuvieron 417 observaciones de un total de 977 estudiantes que representan la población para ese año. La pregunta a contestar es si la muestra presenta un sesgo de autoselección, pues la PDM 2010 fue realizada de manera voluntaria.

Para determinar si está presente un sesgo de selección se realizará una regresión múltiple con variables que pueden estar asociadas a la decisión de realizar la PDM como lo son la sede y la carrera a la cual el estudiante ingresó en el año 2010, pues se tiene la información de dichas variables no solo para los estudiantes que realizaron el PDM sino para toda la población de interés, como se muestra en el Tabla 2.

Tabla 2.

UNA: Número de estudiantes convocados a realizar la PDM 2010 según sede, carrera y participación en la PDM. Febrero 2010.

Sede	Carrera	Código	Realizaron la PDM	% que realizó la PDM	NO realizaron la PDM	Total
Central	Administración	114	126	50,60	123	249
Central	Enseñanza de las ciencias	121	30	37,04	51	81
Central	Biología	122	50	52,63	45	95
Central	Ciencias geográficas	127	20	52,63	18	38
Central	Ingeniería agronómica	128	29	46,77	33	62
Central	Gestión Ambiental	129	10	24,39	31	41
Central	Ingeniería forestal	130	24	60,00	16	40
Central	Cartografía y diseño digital	131	19	50,00	19	38
Central	Comercio y negocios internacionales	161	38	54,29	32	70
Liberia	Administración	201	10	27,78	26	36
Liberia	Comercio y negocios internacionales	211	5	14,29	30	35
Nicoya	Administración	301	15	45,45	18	33
Nicoya	Comercio y negocios internacionales	308	5	17,86	23	28
Pérez Zeledón	Administración	408	12	16,90	59	71
Coto Brus	Administración	501	17	51,52	16	33
Sarapiquí	Administración	701	7	25,93	20	27
Total	-	-	417	42,68	560	977

Fuente: Elaboración propia basado en los datos suministrados por el departamento de registro de la UNA.

La aplicación del método de dos etapas planteado por Heckman para la muestra obtenida, se detalla a continuación (Zamora, 2012).

La primera etapa consiste en estimar la probabilidad de estar en la muestra (estar dispuesto a realizar la PDM), a partir del modelo Probit aplicado a la ecuación:

$$\text{Select} = \alpha_0 + \alpha_1 \text{ sede} + \alpha_2 \text{ carrera} + \mu \qquad (13)$$

Donde la decisión de realizar la PDM (Select) toma valores 1 para los que deciden realizar la prueba, y 0 para el caso contrario. La sede donde el estudiante realizará sus estudios (sede) se espera que influya sobre la decisión de realizar la PDM ya que si se realizan los estudios en la sede central podría beneficiar la asistencia a la PDM en comparación de asistir a una sede regional; así mismo la carrera a la que ingresa el estudiante puede afectar su decisión de realizar la PDM, pues hay carreras como biología y comercio internacional cuya demanda se ha incrementado en los últimos años y el perfil de entrada de estos estudiantes pueden influir en la decisión de tomar la PDM. Además, hay carreras donde el trabajo administrativo y el compromiso de los jerarcas de las unidades académicas son mayores que en otras. Finalmente se incluye un término de error aleatorio denotado por μ.

Esta ecuación se interpreta como la forma reducida de un modelo en el cual la decisión de participación, que es una variable dicotómica, depende de la sede y la carrera a la que pertenece el estudiante. Esta ecuación se supone lineal en los parámetros y permite estimar la probabilidad predicha de participación de un individuo con ciertas características.

La segunda etapa del método consiste en la estimación de la nota obtenida en la PDM a través de Mínimos Cuadrados Ordinarios aplicados a la ecuación:

$$\text{notadia} = \beta 0 + \beta 1\ \text{sede} + \beta 2\ \text{carrera} + \beta 3\ \text{ids} + \beta 4\ \text{notaexaadm} + \beta 5\ \text{notacolegio} + \hat{\lambda} + \varepsilon \quad (14)$$

donde:

notadia representa la nota en la PDM 2010, sede y carrera son las variables relacionadas con el lugar y la carrera elegida por los estudiantes que espera que afectenla nota obtenida en la PDM. El ids es el índice de desarrollo social del estudiante medido a nivel distrital y que es un estimador del nivel socioeconómico del estudiante que se espera que afecte positivamente el resultado en la PDM al igual que la nota obtenida en el examen de admisión o PAA (notaexaadm) y el promedio de las notas de educación diversificada (notacolegio). Por último, $\hat{\lambda}$ representa la inversa del ratio de Mills y ε es un término de error que sigue una distribución normal.

A partir del modelo así estimado, se obtiene el Inverse Mills ratio o función de supervivencia $\varphi(M\Gamma)$, dado por el cociente entre la función de densidad normal estándar y la función de distribución normal, el cual se inserta en el lado derecho de la ecuación de nota de diagnóstico para corregir el sesgo de selección.

Resultados Método de Heckman

Los resultados del análisis de Heckman, para el caso en cuestion se resumen a continuación en la tabla 3.

Tabla 3.

UNA: Estimación por el método de Heckman en dos etapas.

Variable	Coef.	Std. Err.	z	P>z	Valor del coeficiente en la regresión	Valor P en la regresión
Variable dependiente: Nota del examen de diagnóstico						
IDS	0,00	1,38	0,00	1,00	0,00	1,00
Nota del examen de admisión	0,06	1,21	0,05	0,96	0,06	0,11
Nota promedio del colegio	0,08	2,35	0,03	0,97	0,06	0,36
Carrera de ingreso	**Categoría base administración**					
Enseñanza de las Ciencias	150,98	1978,14	0,08	0,94	-0,91	0,64
Biología	1,92	73,89	0,03	0,98	2,47	0,11
Ciencias Geográficas	-1,63	106,68	-0,02	0,99	-1,07	0,64
Agronomía	46,77	648,12	0,07	0,94	-2,52	0,19
Gestión Ambiental	279,24	3656,82	0,08	0,94	-1,74	0,57
Ingeniería Forestal	-62,29	789,80	0,08	0,94	-1,95	0,35
Cartografía	20,43	301,47	0,07	0,95	-1,11	0,63
Comercio Internacional	53,98	670,11	0,08	0,94	3,04	0,05
Sede	**Categoría base sede central**					
Liberia	297,80	3914,11	-0,08	0,94	-2,98	0,24
Nicoya	160,85	2089,80	0,08	0,94	0,43	0,85
Pérez Zeledón	388,94	5081,48	0,08	0,94	-1,64	0,56
Coto Brus	7,24	161,58	0,04	0,96	-1,44	0,56
Sarapiquí	258,58	3404,93	0,08	0,94	-2,95	0,42
Constante	423,36	5212,56	0,08	0,94	22,15	0,00
mills						
lambda	-531,82	6915,06	-0,08	0,94		
rho	-1,00					
sigma	531,82					
lambda	-531,82	6915,06				

Fuente: Elaboración propia.

Para verificar si el modelo presenta o no el problema de sesgo de selección y para ello es se contrastan las siguientes hipótesis:

$$H_0: \hat{\lambda} = 0$$
$$H_1: \hat{\lambda} \neq 0$$

El test de significancia se realiza mediante la prueba del valor p, que para este caso no recomienda el rechazo de la hipótesis nula (p =0,94). Esto significa que no existe evidencia suficiente para suponer que$\hat{\lambda}$es diferente de cero, es decir, que el posible sesgo de autoselección en la muestra puede ser ignorado o en otras palabras desestimar el sesgo por selección no lleva a un error de especificación en la ecuación.

- **Conclusiones**

Como se ha podido comprobar, la metodología de Heckman ayuda a determinar el posible sesgo de autoselección en muestras que tienen un carácter voluntario, como lo son las pruebas de diagnóstico en el ámbito educativo, (Ordaz, 2008, Psacharopoulos, 2007).

Al trabajar con la muestra de la PDM 2010, con la participación voluntaria de los sujetos, cabía la duda de si la muestra presentaba un sesgo de autoselección, pero de acuerdo con los resultados obtenidos mediante el método de Heckman (p = 0,94), no hay indicios que indiquen la presencia de un sesgo de selección en la muestra considerada.

Cabe resaltar que para aplicar esta técnica se requiere contar con suficiente información relevante tanto de los sujetos que ejecutaron la prueba como aquellos que no. Esto no siempre es fácil de obtener, pues por lo general se conocen bastantes bien las características de los individuos que la realizan la prueba, pero no siempre se conocen las mismas características para aquellos que no efectúan la prueba, (Salas, 2004)

Otro elemento a considerar es que tan relevantes son los datos con que se cuentan para la ecuación de selección, en este caso son las variables sede y carrera, pues se presume que la carrera elegida y la zona de residencia del estudiante, pueden afectar el hecho de que la persona este dispuesta a rendir la prueba.

No obstante, la cantidad de variables que podían asociarse a la ecuación de selección, es decir, tanto estudiantes que efectuaron la prueba como los que no lo hicieron era limitada. Este es un problema muy frecuente cunado no se tiene acceso a variables con un mayor nivel métrico o igualmente importante como lo pueden ser el ingreso económico, la motivación por ingresar a la universidad, la percepción que se tenga de la matemática entre otras.

Lamentablemente es muy difícil contar con este tipo de información, al menos en el caso de la UNA, para toda la población de nuevo ingreso. Sin embargo, los resultados de la prueba de Heckman los cuáles no pudieron determinar la presencia de un sesgo de selección, son razonables, pues los resultados en el curso de Matemática General muestran que las deficiencias detectadas en la PDM las poseen la mayor parte de la población estudiantil de

nuevo ingreso, hay o no hecho el examen de diagnóstico.