UFG UNIVERSIDAD FRANCISCO GAVIDIA Tecnología, Innovación y Calidad

Realidad y Reflexión

ISSN 1962-6510

N.° 34 SEGUNDA ÉPOCA

Revista Cuatrimestral • Año 12, No. 34 Enero-Abril 2012 • San Salvador, Centroamerica

Predicción del desempeño académico: número esperado de materias reprobadas como indicador de rendimiento

Raúl E. Avelar, M.Sc.
School of Civil and Construction Engineering.
Oregon State University. Owen Hall 220,
Corvallis, OR, 97331, USA.
raul.avelar@lifetime.oregonstate.edu

Palabras clave:
Desempeño Académico, Rendimiento Académico, Modelos Estadísticos, Métodos Predictivos, Proceso Poisson Homogéneo, Proceso Poisson Heterogéneo

Abstract
Este artículo muestra el desarrollo de un modelo estadístico predictivo de materias reprobadas como indicador de desempeño académico. El modelo se desarrolla usando los datos de una muestra transversal de estudiantes universitarios en El Salvador, Centroamérica. Adicionalmente, este artículo demuestra la importancia de la verificación de las premisas subyacentes a modelos estadísticos, con tal de advertir que dichos métodos dejan de ser apropiados cuando se encuentra evidencia de desviaciones severas de dichas premisas. Como tal fue el caso de la primera propuesta en el presente análisis de modelar las reprobaciones como un proceso Poisson Homogéneo, el presente artículo propone como mejor alternativa y un segundo modelo basado en un proceso Poisson Heterogéneo. Y además propone dos usos potenciales para la herramienta propuesta: La identificación de indicadores tempranos de riesgo. y sus respectivas ponderaciones, y la identificación temprana de alumnos con alto riesgo de reprobación.

Introducción y motivación

Para un investigador la disponibilidad de datos recientes y válidos acerca del fenómeno a investigar es una necesidad apremiante. En términos prácticos, la calidad de la investigación equivale a la calidad y la disponibilidad de los datos disponibles.

Artículo recibido el 23/04/2012 aceptado y arbitrado 21/05/2012

No ha de extrañar entonces el ahínco con que los datos son procurados y el recelo con que son resguardados en cualquier grupo o comunidad de investigadores.

Sin embargo, la investigación en la docencia es un caso excepcional en cuanto a que la generación de datos tiende a ser abundante, así como pequeño el número de investigadores que van más allá de los análisis básicos con tal de maximizar la calidad de la información que se puede obtener. La investigación en la docencia es entonces un frente que ofrece amplias oportunidades para avances importantes, particularmente en El Salvador. Con la intención de animar a cada vez más colegas docentes a ir más allá de comparaciones marginales en sus análisis en investigación de desempeño académico, este artículo propone la utilización de una herramienta estadística relativamente avanzada en la evaluación del desempeño académico. Si este artículo logra su objetivo, entonces el autor esperaría que dichas técnicas avanzadas, o particularmente la herramienta aquí propuesta, sean consideradas regularmente en futuros análisis de desempeño en las aulas universitarias y en la asesoría académica que se brinda a los estudiantes en El Salvador.

Primera parte: Una revisión de las herramientas básicas

En el estudio del desempeño académico, por lo general el investigador busca comparar resultados, típicamente en términos de notas o promedios generales. En tales casos es relativamente sencillo hacer comparaciones marginales usando un indicador de desempeño (e.g. nota alcanzada en una materia X) utilizando los datos de dos grupos de estudiantes. El tipo de preguntas que dichas comparaciones persiguen son de la forma"¿Existe alguna diferencia entre el grupo A de estudiantes y el grupo B, dado que los grupos han sido separados según el criterio C?". El criterio C puede ser cualquier variable cuyo efecto en el desempeño es de interés al investigador (e.g. género, nivel socioeconómico de los estudiantes, método pedagógico empleado para cada grupo, nota obtenida en la clase que es prerrequisito, etc.). Una comparación tal puede ser útil para cuantificar la diferencia entre los dos grupos específicos de la forma "Las notas del grupo A son en promedio 0.5 puntos más altas que las del grupo B". Pero sin más ayuda de la estadística dicho resultado solamente es verdadero para los dos grupos específicos de estudiantes.

Importancia de las premisas en el uso de modelos estadísticos

Si en el ejemplo anterior, además de cuantificar la diferencia, se intentara contestar la pregunta "¿Qué tan probable es encontrar una diferencia de al menos 0.5 puntos entre dos nuevos grupos?", haría falta entonces que los estudiantes en los grupos A y B sean de alguna manera representativos de una población más grande de estudiantes. Si tal es el caso (o si es razonable asumirlo así), y la selección de los estudiantes ha sido tal que su elección se puede considerar independiente, la teoría de probabilidad establece que el resultado de la muestra disponible contiene información sobre cuán típica o atípica es la diferencia de 0.5 puntos para dos cualesquiera muestras de la población. Hace falta echar mano entonces de modelos estadísticos básicos para estimar dicha probabilidad asociada con la diferencia encontrada.

En general, entre más detallada es la pregunta y mayor el alcance de su respuesta, herramientas más complejas serán necesarias y el número de premisas subyacentes al análisis comenzará a crecer. Es importante, entonces, para evitar un análisis errado, revisar dichas premisas y determinar si los modelos estadísticos utilizados son de hecho válidos para el caso particular de los datos a mano.

De hecho, la tentación más común no sólo en la investigación docente, sino en muchos otros campos donde se realiza investigación, es el uso indiscriminado de estas herramientas estadísticas, y el blandir de "valores p" (o "p-value", en inglés) de un estadístico como evidencia contundente de los resultados, aún en los casos en que las premisas podrían no estarse cumpliendo. Aunque resulta problemático e impráctico verificar las premisas de cada comparación de desempeño que un docente realice, la verificación es necesaria cada vez que la comparación vaya acompañada de una inferencia sobre la población general de estudiantes.

Volviendo al ejemplo sencillo de la diferencia de notas entre los dos grupos, la estimación de la probabilidad de una diferencia al menos tan extrema como 0.5 puntos se basa en ciertas premisas acerca de la distribución de probabilidad que sigue la variable que se compara (i.e. la nota obtenida). Si dichas premisas no se cumplen, entonces la probabilidad obtenida, o "valor-p", no tiene significado, así como no tendría sentido hablar de la población general de estudiantes basándose en el resultado obtenido.

Si se persiguen preguntas investigativas de mayor complejidad, naturalmente hará falta el uso de herramientas estadísticas más sofisticadas. Por ejemplo, para la pregunta. "¿Qué otros factores, además del criterio C, pueden estar influyendo en la diferencia de desempeño entre los grupos A y B?". El investigador debe controlar directamente la influencia marginal de dichos factores en la diferencia de notas observada entre los grupos. Las herramientas más básicas para este tipo de comparaciones son el análisis de varianza (ANOVA) de dos o más niveles y bloques y la regresión lineal. De hecho el análisis de regresión lineal puede considerarse un caso particular de análisis de varianza. Las premisas importantes a verificar en este caso son: independencia entre los puntos en la muestra, normalidad de los residuos de la regresión, linealidad entre cada variable independiente y los residuos parciales de la variable dependiente, y homocedasticidad. Todas estas premisas son verificables vía gráficos o pruebas estadísticas (y es recomendable utilizarlos), menos la primera. Para hablar de independencia, con propiedad, hace falta examinar el proceso de selección de los datos en la muestra y verificar que las posibilidades de sesgo de selección fueron minimizadas. Típicamente, la estrategia para tal efecto es la selección de la muestra utilizando números aleatorios. En tal caso, se tiene una muestra probabilística, y la premisa de independencia entre los elementos es entonces razonable.

En resumen, es importante resaltar que, a pesar de la utilidad de herramientas estadísticas, el investigador siempre debe estar alerta a las premisas que subyacen a dichos modelos y hacer las correcciones que sean pertinentes con tal de que las conclusiones del análisis no pierdan validez. En la segunda parte de este artículo se describe el desarrollo de una herramienta estadística para el desempeño académico.

Segunda parte: Desarrollo de un modelo predictivo del desempeño académico

Esta sección presenta la aplicación de una técnica estadística, relativamente avanzada, a una muestra transversal de estudiantes de Ingeniería Industrial con el objetivo de pronosticar el desempeño académico esperado a lo largo de la carrera, utilizando solamente información disponible del primer semestre de estudio. Si dicha proyección es confiable, puede entonces utilizarse el modelo propuesto para identificar las materias más críticas, así como a los alumnos con alto riesgo de fracaso académico y brindarles asesoría académica adecuada y oportuna para prevenir tales fracasos.

Datos disponibles

Durante el período de febrero a marzo de 2007, se obtuvieron aleatoriamente 250 expedientes de estudiantes de Ingeniería Industrial en la Universidad Centroamericana de los que derivaron comparaciones marginales de variables en un trabajo de graduación de Ingeniería Industrial(López González, Núñez Reyes y Avalos Lara 2007). Varias tendencias observadas por los autores fueron sometidas a pruebas estadísticas, y los resultados como conjunto ofrecen una interesante caracterización numérica de varios aspectos del desempeño académico. Sin entrar en mayores detalles, resalta entre otras cosas, una hipótesis para la cual los investigadores declararon su análisis inconcluso:

"[En] *el caso del porcentaje de aprobación en segunda matrícula versus el CUM de carrera [Coeficiente de Unidades de Mérito, o promedio general], según los análisis del capítulo No.4 no existe correlación lineal entre estos dos factores, pero según la prueba de hipótesis realizada en el modelo de regresión múltiple, la contribución del porcentaje de aprobación en segunda matricula al CUM de carrera no es insignificante.*" (Lopez Gonzalez, Nuñez Reyes y Avalos Lara 2007, p. 168)

Más allá del cuestionamiento de la premisa de linealidad en el modelo de regresión que los autores utilizan, es de esperarse que tanto las notas obtenidas en las materias básicas así como la matrícula de aprobación tengan una influencia significativa en el futuro desempeño de un estudiante. Como sugiere el citado estudio, el hecho de que una materia ha sido aprobada en primera matrícula es en sí una pieza de información relevante para emitir un pronóstico de éxito académico. Si, por ejemplo, dos estudiantes han aprobado una materia con la misma nota, pero sólo uno de ellos lo ha hecho en primera matrícula, es de esperar que éste último tenga mayores probabilidades de seguir aprobando materias en primera matrícula si se le compara con el otro. En otras palabras, a pesar de que la probabilidad de reprobar materias se espera sea inversamente proporcional a la nota obtenida en las materias básicas, también se espera que sea directamente proporcional a la matrícula de aprobación de dichas materias. El problema con el análisis descrito, que correctamente identifican los autores, es precisamente una violación importante a la premisa de linealidad subyacente a un modelo de regresión múltiple.

Aunque existen varias avenidas para ofrecer una solución a este problema, la propuesta en este artículo está basada en una variable distinta al promedio general. Se propone caracterizar el desempeño académico vía el número de reprobaciones en el historial académico del estudiante a lo largo de la carrera. Dicha caracterización tomará en cuenta la información más temprana de desempeño que se espera esté correlacionada con el historial de reprobaciones: las notas y las matrículas de aprobación de las primeras materias básicas del plan de estudio.

Las preguntas investigativas del presente análisis son, entonces:

Pregunta 1. ¿A qué grado es posible predecir el futuro desempeño de los estudiantes basándose solamente en las notas obtenidas en las materias básicas del primer ciclo de la carrera? (Además, ¿Qué tan significativa es la contribución de cada materia a la predicción?).

Pregunta 2: Adicionalmente al posible poder predictivo de las notas de las materias del primer ciclo, ¿influye si dichas materias fueron aprobadas en 2.ª matrícula (o mayor) sobre la probabilidad de futuras reprobaciones?

Especificación del modelo estadístico

La variable de respuesta es, naturalmente, el conteo de reprobaciones de un estudiante en su carrera. Como los predictores más tempranos de desempeño, se escogieron las cuatro materias básicas cursadas en el primer ciclo de la carrera de Ingeniería Industrial en la Universidad Centroamericana: Álgebra Vectorial y Matrices, Matemática I, Química General I, y Principios Gráficos del Diseño. Dado que se intenta predecir el número de reprobaciones en el historial de los estudiantes, para ser parte de este análisis, hace falta que hayan aprobado o reprobado al menos una materia, lo que descalifica aquellos que se encuentran cursando materias por primera vez. Además, hace falta que todos los estudiantes hayan aprobado o reprobado las cuatro materias seleccionadas como predictores al menos una vez cada una. Por estas razones fue necesario reducir significativamente la muestra disponible a 191 de los expedientes en el subsiguiente análisis.

La estrategia de modelado requerida sobrepasa los alcances de la regresión ordinaria de mínimos cuadrados (OLS, por sus siglas en inglés). La regresión ordinaria de mínimos cuadrados asume una varianza transversalmente constante (i.e. homocedasticidad); sin embargo, debido a que el análisis se realiza utilizando una muestra transversal de estudiantes, hará falta controlar variables temporales como medidas de exposición al proceso que el modelo intenta caracterizar. Se espera que la incertidumbre asociada con niveles más avanzados de la carrera, así como con mayores conteos de materias reprobadas, sea mayor. Específicamente, se espera que dicho incremento en la incertidumbre sea proporcional a la medida en el punto de predicción se aleje del momento en que las variables predictores fueron tomadas. En otras palabras, como las variables usadas en el análisis son los resultados del primer ciclo, entonces más incertidumbre debe acompañar una predicción de desempeño al 4.° año de carrera que una predicción al 2.° o 3.°. Esta característica implica dos cosas: el tiempo de la predicción debe ser parte del modelo, y además se espera un incremento en la incerteza monótonamente positivo. Además, se trata en este caso de una variable de respuesta discreta y siempre positiva (i.e. número de materias reprobadas), lo que es contrario del tipo de variable continua e irrestricta que un procedimiento OLS asume en sus premisas.

La literatura existente [(Iwasaki y Tsubaki 2006), (Ramsey y Schafer 2002), (Samaniego 1976), (Wackerly, Mendenhall III y Scheaffer 2008) y (Winkelmann y Zimmermann 1991)] sostiene que el tipo de variable seleccionada para modelar en este caso (el conteo total de materias reprobadas después de un tiempo t en la carrera), puede ser conceptualizada como el resultado de un proceso Poisson. Este tipo de modelos ha sido utilizado con éxito en la investigación de seguridad vial (Avelar y Dixon 2011). El caso más sencillo es el de un proceso Poisson Homogéneo:

$$Prob[Rep = Y] = \frac{e^{-t.\lambda}.(t.\lambda)^{Y}}{Y!}$$

Ecuación 1. Probabilidad de materias reprobadas de un proceso Poisson homogéneo.

Donde Rep es el conteo antes mencionado; 'Y' es una variable discreta con Dominio= N, incluyendo cero; t es el tiempo de la predicción; y lambda es el parámetro de escala de la distribución de probabilidad, con Dominio= R^+. Esta ecuación permite el cálculo de la probabilidad de que el número de materias reprobadas sea igual a un valor. Para el cálculo de dicha probabilidad hace falta conocer los dos parámetros de la función: t y lambda.En el presente caso de modelado, tanto como en otros similares, se utilizó un algoritmo de optimización (e.g. Newton-Raphson) para maximizar la función de verosimilitud de la distribución estadística seleccionada (Maximum Likelihood Estimation, o MLE, por sus siglas en inglés).

El presente problema de modelado entonces se enfoca en caracterizar λ como función de los predictores disponibles. El tiempo debe ser parametrizado de manera simultánea a lambda, como una medida de exposición en el proceso Poisson:

$$MLE:\begin{Bmatrix} t = f(Expos) \\ \lambda = g(Predictores) \end{Bmatrix}$$

Ecuaciones simultáneas 2. Estimación de los parámetros de la función de probabilidad Poisson.

Definimos "Expos" como una o varias variables representativas del tiempo de "exposición" del alumno al proceso Poisson. En este caso, se selecciona Expos simplemente como el número de años que el estudiante lleva en la carrera, y los predictores son estipulados en forma vectorial:

$$MLE:\begin{Bmatrix} t = Expos^{\alpha} \\ \lambda = K.\left[e^{f(\overline{Pred})}\right] \end{Bmatrix}$$

Ecuaciones 3. Parametrización simultánea de los parámetros de la función de Probabilidad Poisson

$\overline{Pred}$ es un vector de dimensión p, tal que cada componente es el valor numérico de los factores predictores a investigar. La composición de este vector se establecerá según la pregunta investigativa que se contemple.

Pregunta 1. ¿A qué grado es posible predecir el futuro desempeño de los estudiantes basándose solamente en las notas obtenidas en las materias básicas del primer ciclo de la carrera? (Además, ¿Qué tan significativa es la contribución de cada materia a la predicción?).

Para el caso de la primera pregunta investigativa solamente, $\overline{Pred}$ es de dimensión p=4, definido como las notas de las cuatro materias de interés:

$$\begin{bmatrix} Nota.Alg \\ Nota.Mat \\ Nota.Qui \\ Nota.Gra \end{bmatrix}$$

La función de exposición temporal se selecciona, en este caso, como una potencia de los años en la carrera. La función del vector de predictores es el producto interno con un vector de parámetros a estimar, multiplicado a su vez por una constante K. Por tanto, la estimación de máxima verosimilitud (MLE) será:

$$MLE:\left\{\begin{array}{l} t = [Años.en.Carrera]^{\alpha} \\ \lambda = K.\left[e^{\left(\vec{\beta}^{T}.\begin{bmatrix} Nota.Alg \\ Nota.Mat \\ Nota.Qui \\ Nota.Gra \end{bmatrix}\right)} \right] \end{array}\right\}$$

Ecuación 4. Parametrización para estimar el proceso Poisson homogéneo.

La solución a este problema puede especificarse en cualquier paquete matemático o estadístico que permita implementar un algoritmo de optimización con función objetivo no-lineal. Para el caso, el autor de este artículo realizó la estimación utilizando el paquete de análisis estadístico R. Resultados de la estimación:

Tabla 1. Resultados de la estimación de un proceso poisson homogéneo.

Parámetro	Estadístico	Límite Inferior	Límite Superior	Significancia estadística
Alfa	1.341	1.169	1.512	< 2e-16
K	37.089	18.766	73.302	< 2e-16
Beta.1	0.031	**-0.078**	**0.141**	0.57428
Beta.2	-0.222	-0.314	-0.131	0.000002
Beta.3	-0.241	-0.325	-0.156	2.27E-08
Beta.4	-0.132	-0.222	-0.042	0.00414

Antes de proceder a la interpretación de los resultados, debe verificarse la credibilidad de las premisas del modelo como se estableció en la primera parte de este artículo. Una prueba de hipótesis preliminar sobre la verosimilitud optimizada del modelo sugiere que se falla en describir adecuadamente la variabilidad en los datos. En esta prueba se obtuvo un p-value de 0.00072 sobre 185 grados de libertad, de una distribución de chi-cuadrada. La noción subyacente a esta prueba es una comparación entre el modelo propuesto y un modelo saturado para la muestra. Se espera que el doble del negativo de la razón de verosimilitudes. (i.e. log-likelihood ratio en inglés) deba converger a la distribución de chi-cuadrada por virtud del teorema de límite central. El valor promedio esperado de dicha medida es entonces 185, el número de grados de libertad del análisis, pero en este caso el estadístico fue de 252.4, resultando en el p-value mencionado. Dicho valor p indica que la plausibilidad de la especificación estadística del modelo ajustándose a los datos es cuestionable.

Hay razones como para creer que el modelo propuesto, entonces, no es el adecuado para los datos, y que las conclusiones que puedan derivarse de él no serían válidas. Se ofrecen dos posibles explicaciones para este impase: (1) el proceso descrito no es un proceso homogéneo de Poisson, como se propuso al principio de este análisis, o (2) hacen falta predictores críticos que aproximen el verdadero proceso homogéneo de Poisson subyacente.

Asumiendo (1), se procede a modelar, como alternativa, un proceso heterogéneo Poisson. Entre la posibilidad de opciones existentes, se selecciona un modelo de proceso heterogéneo tal que:

$$t.\lambda \sim Gamma(t.\lambda_0, \theta)$$

Donde Gamma(.) es la distribución de probabilidad Gamma. La selección de esta distribución no es arbitraria, ya que es fácilmente demostrable, entonces, que:

$$Prob[Rep = Y] = \frac{(t.\lambda_0)^Y}{Y!} \cdot \frac{\Gamma(\theta + Y)}{\Gamma(\theta).(\theta + t.\lambda_0)^Y} \cdot \left(1 + \frac{t.\lambda_0}{\theta}\right)^{-\theta}$$

Ecuación 5. Probabilidad de materias reprobadas de un proceso Poisson heterogéneo (Poisson-Gamma).

Donde Γ(.) es la Función Gamma, lo que requiere que θ es una variable continua con dominio R^+.
Se tienen ahora tres parámetros a estimar: t,λ_0, y θ. Especificando los primeros dos como antes, y el nuevo parámetro simplemente como θ=θ_o, tenemos entonces:

$$MLE: \begin{Bmatrix} t = [Años.en.Carrera]^{\alpha} \\ \lambda_0 = K.\left[e^{\left(\vec{\beta}^T \times \begin{bmatrix} Nota.Alg \\ Nota.Mat \\ Nota.Qui \\ Nota.Gra \end{bmatrix}\right)} \right] \\ \theta = \theta_0 \end{Bmatrix}$$

Ecuación 6. Especificación de parámetros a estimar (Proceso Poisson heterogéneo Poisson-Gamma).

Los resultados de la nueva estimación se presentan en la Tabla 2. Es interesante ver que los valores de los estadísticos no son muy diferentes a los obtenidos en la primera estimación; pero como era de esperarse, el rango de valores esperados es más amplio en la segunda estimación debido a la presencia de un nuevo parámetro indicativo inverso de dispersión (nótese que en la Ecuación 5 cuando Theta tiende al infinito, el segundo término tiende a 1, y el tercero tiende a $(e^{-t.\lambda})$.

Tabla 2. Resultados de la estimación de un proceso heterogéneo Poisson-Gamma

Parámetros	Estadístico	Límite Inferior	Límite Superior	Significancia
Alfa	1.333	1.144	1.523	< 2e-16
K	48.65	22.236	106.439	< 2e-16
Beta.1	0.016	-0.113	0.146	0.8039
Beta.2	-0.247	-0.36	-0.134	3.00E-05
Beta.3	-0.248	-0.347	-0.148	0.00E+00
Beta.4	-0.128	-0.234	-0.022	1.85E-02
Theta.0	12.667	0.8286	24.5054	0.03598

La misma prueba que se realizó sobre la razón de verosimilitudes para el modelo de proceso Poisson homogéneo falla en rechazar la hipótesis de que la muestra proviene de un proceso Poisson heterogéneo, tal como el modelado (p-value de 0.1278, de una distribución chi-cuadrada sobre 184 grados de libertad, con un estadístico de razón de verosimilitudes de 207.03 proveniente del algoritmo de optimización).

Ahora sí es posible extraer conclusiones válidas, ya que no hay evidencia en contra de que el modelo describa adecuadamente el proceso subyacente a los datos. Es importante mencionar, sin embargo, que esto no significa que el modelo es una representación uno-a-uno del proceso en sí. Tal afirmación resultaría en el erróneo de decir que la cantidad de materias reprobadas por un estudiante en toda su trayectoria académica solo depende de las notas obtenidas en las cuatro materias básicas del primer semestre. Se sabe que la verdadera función en realidad involucra todas y cada una de las materias cursadas alguna vez en la trayectoria académica. El modelo es simplemente una aproximación matemática para describir el resultado de dicho proceso utilizando solamente la información temprana de desempeño, y como tal, es potencialmente más útil para asesorar a estudiantes de alto riesgo.

Interpretación del modelo.

Alfa. Este exponente es significativamente distinto de cero y positivo, lo que indica convincentemente que, como era de esperarse, la probabilidad de dejar materias aumenta con los años que un alumno tiene cursando la carrera independientemente de cualquier indicador de desempeño.

K. Esta constante es positiva como era de esperarse. Sin embargo, su rango de valores es amplio y con grandes magnitudes. No tiene una interpretación física propiamente, pero el amplio rango de valores sugiere que el modelo exhibe grande sensibilidad al estimado de este parámetro, una característica no deseada que podría indicar potencialmente la divergencia del modelo respecto del proceso real subyacente (i.e. no es de esperar que en proceso subyacente el valor de esté predeterminado por una constante fija y totalmente independiente de los predictores.). Sin embargo, el modelo puede tener mérito aún en su carácter empírico si una validación estadística verificara su poder predictivo sobre una muestra de expedientes nueva e independiente.

Theta.0. Esta constante es positiva, con un rango de valores positivos y relativamente pequeños. El rol que esta variable juega dentro de la función de probabilidad de materias reprobadas es esencialmente una medida inversa de dispersión. El hecho de que los valores sean pequeños indica que el modelo, a pesar de poseer poder predictivo, podría tener limitada precisión de los valores predichos, como se discute en la tercera parte de este artículo.

Beta1. El primer componente del vector de constantes (correspondiente a la materia Álgebra Vectorial) es junto al amplio rango de valores K, otro indicativo en contra de la estructura matemática del modelo: se espera que todos los componentes del vector Beta sean negativos (i.e. a mayor nota en cada una de las materias, menor probabilidad de incrementar el número total de materias reprobadas). Sin embargo, el estimador de este componente es positivo. A pesar de ello, el rango de valores para el estimador contiene tanto valores positivos como negativos, lo que es un punto a favor del modelo. Se sigue entonces con el análisis bajo la hipótesis de que este componente es en realidad igual o menor a cero, pero que el tamaño de la muestra es el factor que ha hecho marginal su estimación. Esta es una hipótesis razonable, ya que el modelo no ofrece evidencia al contrario: la variación estimada es tal que el parámetro real puede tener tanto valores positivos como negativos. Con todo, se concluye que la nota de Álgebra Vectorial en este modelo no es un predictor relevante del futuro desempeño de los estudiantes.

Beta 2, 3, y 4. A diferencia de Beta 1, todos estos componentes son negativos, y estadísticamente significativos (i.e. el rango esperado de valores solo contiene números negativos). Se concluye entonces que hay una relación inversa entre la probabilidad de reprobar materias en el futuro y las notas de Matemática, Química I y Principios Gráficos I. Esta relación inversa es esperada: a mayores notas obtenidas en tales materias, menores son las probabilidades de incrementar el número de futuras reprobaciones.

Pregunta 2: Adicionalmente al poder predictivo de las notas de las materias del primer ciclo, ¿influye si dichas materias fueron aprobadas en segunda matrícula o mayor matrícula) en la probabilidad de futuras reprobaciones?

Se estimó otro modelo con un vector de predictores p=8, que contiene además de los parámetros de las notas, las matrículas en que fueron aprobadas las respectivas cuatro materias, como se muestra en la Tabla 3. La respuesta que dicho modelo sugiere a esta pregunta es que sí, hay evidencia estadística de que la matrícula influye: a mayor matrícula de aprobación, mayor probabilidad de incrementar el número de reprobaciones en el futuro (Coeficientes Beta 5, 6, 7, y 8). Sin embargo, solo en dos casos el segundo modelo mostró que el incremento en la probabilidad de reprobaciones es estadísticamente significativa: Álgebra Vectorial y Principios Gráficos (i.e. el rango de valores solamente incluye valores positivos).

Tabla 3. Resultados de la estimación incluyendo las matrículas como predictores.

Parámetros	Estadístico	Límite Inferior	Límite Superior	Significancia
Alfa	1.221	1.014	1.428	0.0001
K	11.344	3.395	37.911	< 2e-16
Beta.1	**0.1**	**-0.039**	**0.239**	**0.1584**
Beta.2	-0.198	-0.318	-0.077	0.0013
Beta.3	-0.234	-0.347	-0.122	0.0001
Beta.4	-0.148	-0.256	-0.04	0.0070
Beta.5	0.221	0.063	0.379	0.0063
Beta.6	**0.032**	**-0.159**	**0.223**	**0.7409**
Beta.7	**0.061**	**-0.084**	**0.205**	**0.4121**
Beta.8	0.296	0.006	0.586	0.0454
Theta.0	15.992	0.508	31.476	0.0081

Cabe resaltar que los componentes de las notas obtenidos en el primer modelo no se ven significativamente afectados por la inclusión de las matrículas en el segundo modelo. Con todo, el valor de K cae a 11.34, con rango de valores [3.395, 37.911], lo que preliminarmente indica que este modelo es más estable y posiblemente mejor aproximado al proceso real subyacente. A vista de lo tal, es probable que la influencia de los componentes de matrícula en el vector de predictores sea más importante en conjunto; aún más allá de simplemente las dos materias para las cuales el efecto resulta estadísticamente significativo (Álgebra Vectorial y Principios Gráficos). Esta observación además refuerza la noción de la hipótesis (2) en el impase que se encontró con el modelo de un proceso Poisson homogéneo. Sin embargo, esa hipótesis no será explorada en el presente artículo por razones de espacio.

Tercera parte: Utilización del modelo predictivo

Habiendo descrito con cierto nivel de detalle el modelo predictivo y habiendo además verificado la plausibilidad de sus premisas, la tercera y última parte de este artículo introduce una discusión sobre dos usos potenciales de tal herramienta.

Análisis de la contribución marginal de cada materia al futuro desempeño académico

Contrario a una regresión ordinaria de mínimos cuadrados, en el caso del modelo utilizado no es posible atribuir puntualmente una contribución de cada variable en el vector de predictores al número de materias reprobadas. En el caso de un procedimiento OLS, la contribución marginal de una variable es simplemente:

$$\frac{\partial}{\partial x_1} E[Y] = \beta_{x1}$$

Sin embargo, en el modelo propuesto, si se aplica la derivada parcial a λ_0 respecto de cualquier predictor, de la ecuación 6 se obtiene que:

$$\frac{\partial}{\partial Pred_1}(\lambda_0) = \beta_{Pred_1} \cdot K. \left[e^{\left(\overline{\beta^T \times Pred}\right)}\right]$$

En términos prácticos esto significa que la contribución marginal del predictor $Pred_1$ en la predicción de materias reprobadas es proporcional a la contribución total de todas las variables multiplicada por la constante β_{Pred_1} Entre mayor es la predicción de materias reprobadas, mayor es la contribución de cada predictor. A pesar de este impase, dos aserciones importantes son posibles aún.

Primero, el signo de la constante β_{Pred_1} indica, como fue discutido en la interpretación del modelo, si la contribución es en directa o en inversa proporción al número esperado de materias reprobadas. Para el caso del modelo propuesto, todos los predictores que han resultado estadísticamente significativos exhiben signos tales y como era de esperarse (i.e. negativos para las notas de las materias y positivos para las matrículas de aprobación respectivas). Aún en los casos de matrículas que no resultaron estadísticamente significativas, los signos de los componentes β fueron como se esperaba. El único caso anómalo en cuanto al signo fue la nota de Álgebra Vectorial, cuya estimación de β resultó positiva. Sin embargo, la incertidumbre que el modelo asocia con este componente permite especular que dicho coeficiente pueda que sea nulo, o aun pueda que sea negativo como se espera en teoría que sea.

Segundo, a pesar de que los efectos marginales de cada variable son dependientes de las otras variables, aún es posible hacer una comparación relativa de las contribuciones de cada predictor, ya que

$$\frac{\frac{\partial}{\partial Pred_1}(\lambda_0)}{\frac{\partial}{\partial Pred_2}(\lambda_0)} = \frac{\beta_{Pred_1}}{\beta_{Pred_2}}$$

La Tabla 4 muestra las comparaciones relativas relevantes de los componentes β del modelo. Dicha tabla no incluye la nota de la materia Álgebra Vectorial, ya que, como se discutió, el estimador asociado con dicha nota resultó anómalo y aún es posible que sea nulo según los mismos resultados.

Tabla 4. Resultados de la estimación incluyendo las matrículas como predictores

			Notas			Matrículas de Aprobación			
			Matemática	Química	Principios Gráficos	Álgebra Vectorial	Matemática	Química	Principios Gráficos
			Beta. 2	Beta.3	Beta.4	Beta.5	Beta.6	Beta.7	Beta.8
			-0.198	-0.234	-0.148	0.221	0.0032	0.061	0.296
Notas	Matemática	Beta. - 2 0.198	1.00	0.85	1.34	-	-	-	-
	Química	Beta. – 3. 0.234	1.18	1.00	1.58	-	-	-	-
	Principios Gráficos	Beta. – 4 0.148	0.75	0.63	1.00	-	-	-	-
Matrículas de Aprobac-ión	Álgebra Vectorial	Beta. – 5 0.221	-	-	-	1.00	6.91	3.62	0.75
	Matemática	Beta. – 6 0.032	-	-	-	0.14	1.00	0.52	0.11
	Química	Beta. – 7 0.061	-	-	-	0.28	1.91	1.00	0.21
	Principios Gráficos	Beta. – 8 0296	-	-	-	1.34	9.25	4.85	1.00

Como puede observarse, Química I es la nota que parece jugar un rol mayor en la predicción de materias reprobadas. Por cada punto obtenido en esta nota se espera una reducción en el número de materias reprobadas 18% mayor que la reducción asociada con cada punto de la nota de Matemática I, y 58% mayor que la reducción asociada con cada punto de la nota de Principios Gráficos.

La comparación entre matrículas de aprobación es ciertamente más extrema. Dos materias sobresalen en este respecto: Álgebra Vectorial y Principios Gráficos. Por cada vez que se reprueba la primera, el incremento en el número de futuras reprobaciones. es 6.91 veces el incremento asociado con cada vez que se reprueba Matemática, y 3.62 veces el incremento asociado con cada vez que se reprueba Química. De manera similar, cada vez que Principios Gráficos es reprobada, se espera un incremento en futuras reprobaciones de alrededor de 34% más que el incremento asociado con Álgebra Vectorial, 9.25 veces el incremento asociado con reprobar Matemática y 4.85 veces el incremento asociado con reprobar Química.

Sin embargo, el nivel de incerteza asociado con la estimación de los componentes β de Matemática y Química, permite aún especular que al igual que la nota de Álgebra vectorial, pueda que el reprobar cada una de estas dos materias sea irrelevante para la predicción de futuras reprobaciones.

Determinación de probabilidades de futuras reprobaciones

El uso más directo del modelo propuesto es como una herramienta de asesoría académica para alumnos de segundo semestre. Una vez se tiene los resultados de las notas y aprobaciones de las materias del primer semestre (que en caso de haber reprobado alguna, debe haber tomado más de un semestre en ser aprobadas), es posible utilizar la ecuación 5 para predecir la probabilidad de reprobar un número particular de materias a distintos puntos en la carrera.

Por ejemplo, puede ser de particular interés acotar la probabilidad de terminar la carrera sin ninguna materia reprobada. En tal caso, se tiene que $Rep = 0$. Después de simplificarla, la ecuación 5 se reduce entonces a:

$$Prob[Rep = 0] = \left(1 + \frac{t.\lambda_0}{\theta}\right)^{-\theta}$$

donde t, λ_0 y θ son según la Ecuación 6, y los estimadores correspondientes son los mostrados en la Tabla 3.

La probabilidad de reprobar al menos una materia viene dada entonces por la ecuación 7.

$$Prob[Rep \geq 1] = 1 - \left(1 + \frac{t.\lambda_0}{\theta}\right)^{-\theta}$$

Ecuación 7. Probabilidad de reprobar al menos una materia según el modelo propuesto

Para brindar un ejemplo concreto del uso de la ecuación 7, considérense los tres casos hipotéticos que se presentan en la Tabla 5.

Tabla 5. Casos hipotéticos para la estimación de probabilidades de reprobación.

Variable	Materia	Estudiante 1	Estudiante 2	Estudiante 3
Nota	Matemática	9.1	6.2	6.2
	Química	8.8	7.1	7.1
	Principios Gráficos	9.5	7.0	7.0
Matrícula de Aprobación	Álgebra Vectorial	1	1	2
	Matemática	1	1	2
	Química	1	1	1
	Principios Gráficos	1	1	1
Probabilidad	**Al menos una reprobada al año 2**	**22.0%**	**60.6%**	**69.6%**
	Al menos una reprobada al año 4	**43.7%**	**87.7%**	**92.9%**

Para el estudiante 1, cuyas notas del primer ciclo son encomiables, quien aprobó todas las materias en primera matrícula, el modelo predice probabilidades relativamente bajas de reprobar al menos una materia al final del segundo y cuarto años (22% y 43.7% respectivamente). Con todo, es interesante observar que aún con notas tan favorables en el primer trimestre de la carrera, todavía no es posible predecir con mayor certeza que en el futuro no reprobará materias basándose en los resultados del primer semestre solamente.

En el caso del estudiante 2, que no reprobó ninguna materia durante el primer semestre, pero sus notas fueron bajas en comparación al estudiante 1, el modelo predice un riesgo más alto de reprobar materias en el futuro. Sus probabilidades de haber reprobado una o más materias al final del segundo año son de 60.6% y su riesgo aún es mayor al final del cuarto año de carrera (probabilidad de 87.7%).

Finalmente, en el caso del estudiante 3, cuyas notas son las mismas que las del estudiante 2, las probabilidades de reprobación aumentan, lo que no debe extrañar, ya que aun entre las materias del primer semestre tiene dos reprobadas (i.e. Álgebra Vectorial y Matemática fueron aprobadas en segunda matrícula). En este caso, las probabilidades de futuras reprobaciones al final del segundo año incrementan en un 9%, y al final del cuarto año aumentan en 5.2% respecto del estudiante 2.

Conclusiones

El presente artículo describe el desarrollo de una herramienta estadística para evaluar el desempeño académico universitario. La base de dicha herramienta es vincular los resultados tempranos en materias básicas a la probabilidad de reprobaciones a lo largo de la carrera. Inicialmente, se propuso una herramienta basada en modelar un proceso Poisson homogéneo; pero fue necesario un segundo modelo de proceso Poisson Heterogéneo debido a dudas encontradas durante la verificación de las premisas del primero modelo.

Este artículo presenta además dos ejemplos del uso de la herramienta: (1) identificación de indicadores críticos de riesgo, y sus respectivas contribuciones marginales; y (2) Identificación de alumnos con alto riesgo de reprobación.

El análisis sobre las contribuciones marginales indicó que el desempeño en todas las materias del primer ciclo puede utilizarse en la estimación del futuro desempeño: en el caso de Álgebra vectorial solamente la matrícula de aprobación parece ser el predictor relevante; en contraste, el predictor relevante es la nota obtenida para los casos de Matemática y Química. Solamente en el caso de Principios Gráficos del Diseño fueron buenas predictoras ambas la nota y la matrícula de aprobación.

A pesar de los esfuerzos en pro de la rigurosidad del presente análisis, si el uso primario del modelo propuesto ha de ser predictivo más que informativo, el autor recomendaría realizar una validación estadística usando una segunda muestra independiente de expedientes. La herramienta como es aquí presentada no ha de ser usada para poblaciones estudiantiles distintas a la de Ingeniería Industrial de la Universidad Centroamericana, ya que los resultados y las probabilidades son representativos de esa población únicamente. Si fuera de interés el usar la herramienta para otra población estudiantil, es necesario calibrarla utilizando suficientes datos provenientes de tal población. Aun después de calibrada, la herramienta debe preferentemente ser validada con más datos de la nueva población de estudiantes.

Finalmente, debe resaltarse que después de todo, la propuesta de este artículo sólo es un modelo estadístico. Los resultados y los usos delineados están sujetos a la calidad del ajuste del modelo de probabilidad propuesto; y sus premisas, a los datos disponibles para el análisis. Probablemente es factible la construcción de modelos más sofisticados para contestar preguntas más complejas. En tal caso, más datos serían necesarios, pero dados los grados de libertad que la muestra de estudiantes disponible ofrece, el autor considera que los resultados son robustos y válidos, y recomienda ambos usos sugeridos para dicha herramienta.

Trabajos citados

Avelar, Raúl E., y Karen K. Dixon. Modelling the Safety Effect of Advisory Speed Signs: A Bivariate Multiplicative Factor on Number of Crashes based on the Speed Differential and the Side Friction Demand. CD-ROM. Indianapolis, IN: Transportation Research Board, 2011.

Iwasaki, Masakazu, y Hiroe Tsubaki. «Bivariate Negative Binomial Generalized Linear Models for Environmental Count Data.» Journal of Applied Statistics, 2006: 909-923.

López González, Miguel Ángel, Edwin José Núñez Reyes, y Carlos Alejandro Ávalos Lara. Análisis de los flujos de la población estudiantil de Ingeniería Industrial a través del plan de estudio. Trabajo de Graduación, San Salvador, El Salvador: UCA, 2007.

Ramsey, Fred L., y Daniel W. Schafer. The Statistical Sleuth. A Course in Methods of Data Analysis. Second Editon. Pacific Grove, CA: Duxbury, 2002.

Samaniego, Francisco J. «A Characterization of Convoluted Poisson Distributions with Applications to Estimation.» Journal of the American Statistical Association, 1976: 475-479.

Wackerly, Dennis D., William Mendenhall III, y Richard L. Scheaffer. Mathematical Statistics with Applications. 7th Edition. Toronto, Canadá: Thomson, 2008.

Winkelmann, R., y K. Zimmermann. «A New Approach for Modeling Economic Count Data.» Economics Letters 37, 1991: 139-143.