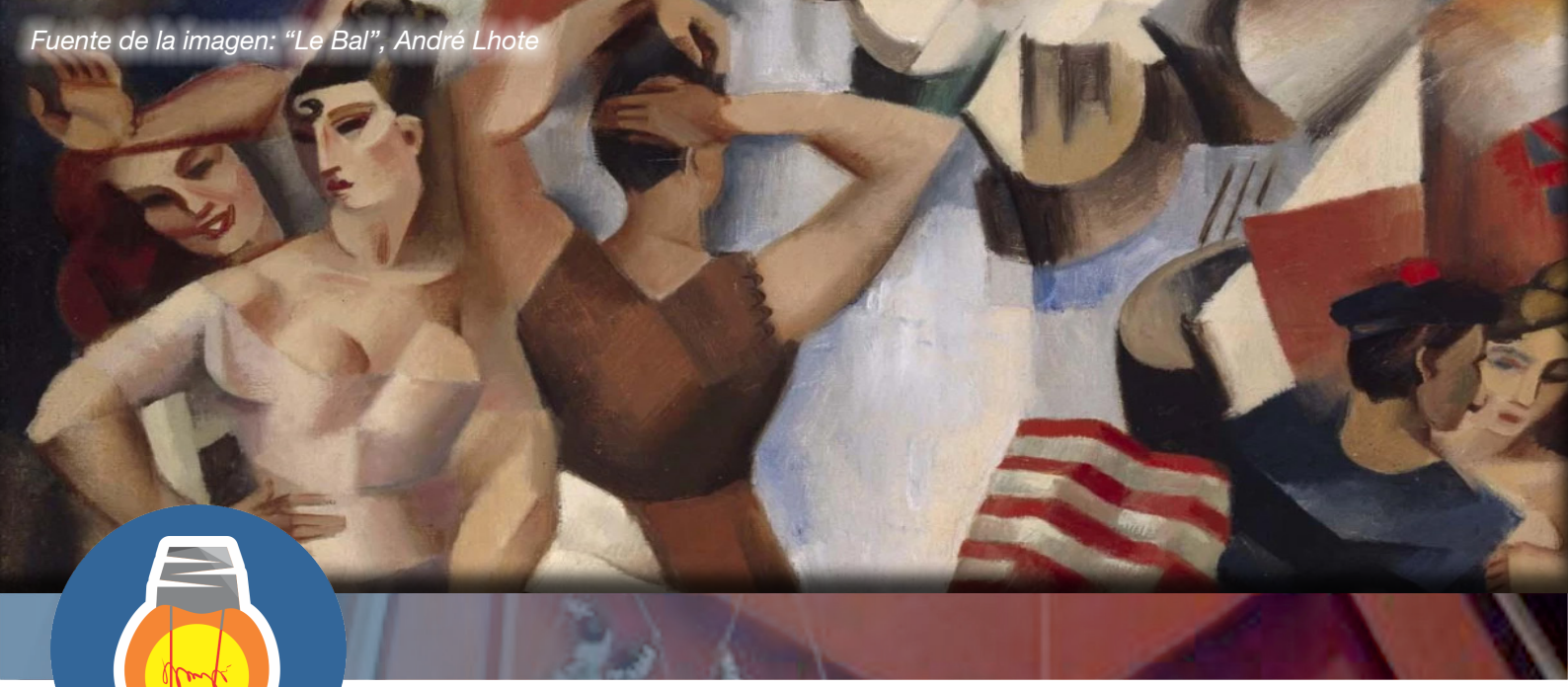
Fuente de la imagen: “Le Bal”, André Lhote

Modelo para el análisis de datos del operativo nacional de evaluación Aprender: ejemplo de uso de la Regresión Logística Ordinal

Jorge Lorenzo

Resumen

Los resultados de los Operativos Nacionales de Evaluación, son fuente de controversia y generan numerosas discusiones, mayormente en medios de comunicación. Consideramos que esta situación invisibiliza el trabajo que se realiza en la comunidad educativa, especialmente investigadoras/es, que se preocupan por dar sentido a los resultados, analizando los datos desde modelos cuantitativos complejos. No obstante, puede notarse una vacancia en el uso de algunos modelos estadísticos multivariados que resultan apropiados para el tipo de datos que contienen las bases usuarias provistas por el Ministerio de Educación en cada operativo realizado. El presente artículo pretende ofrecer un modelo de análisis basado en la Regresión Logística Ordinal, la cual resulta adecuada para trabajar con variables independientes nominales y una dependiente ordinal. En tal sentido, se propone un ejemplo como modo de acercamiento a este tipo de técnica analítica, con la finalidad de explorar las propiedades del modelo estadístico y discutir las diferencias con otros modelos alternativos.

Jorge Lorenzo

Facultad de Filosofía y Humanidades.
Universidad Nacional de Córdoba

Palabras clave: análisis cuantitativo, regresión ordinal, modelo logístico, ejemplo.

Introducción

Desde hace varios años se implementan en Argentina los operativos nacionales de evaluación de estudiantes, con la finalidad de monitorear el rendimiento académico del estudiantado en áreas de conocimiento básicas, generalmente, Lengua, Matemáticas y Ciencias (Mosteiro, 2018). Los resultados de estos operativos despiertan encendidas controversias ni bien se publican dado que, hasta el presente, no se ha registrado una mayoría de estudiantes con rendimiento en un nivel avanzado. Transcurridos los debates, particularmente los que se dan

en los medios de comunicación, la situación queda estancada hasta la aparición de los resultados de nuevas pruebas de evaluación masiva de estudiantes. Esta misma dinámica se observa en evaluaciones internacionales como PISA o ERCE (Rodríguez, et. al. 2018; Rivas, 2015). Los educadores no expertos en análisis cuantitativos, deben conformarse con los informes ejecutivos que se brindan junto con las pruebas, a pesar que las bases usuarias son de dominio público. Solo en el ámbito de la investigación especializada se reportan análisis pormenorizados de aspectos puntuales, muchas veces restringidos a una comunidad particular.

Consideramos que existe una vacancia en los expertos en educación para aplicar modelos de análisis más complejos a estas pruebas, que permitan modelar y comprender patrones de comportamiento en la multiplicidad de datos que aportan los distintos operativos de evaluación de estudiantes (en particular las pruebas Aprender), dada la relevancia de la información disponible para adoptar políticas educativas fundamentadas y para mejorar la gestión de los sistemas educativos, a partir de la evidencia (para una revisión ver Ravela, et. al. 2008). Aplicaciones de análisis multivariados de gran valor para las políticas públicas pueden apreciarse en trabajos que exploran modelos multinivel para encontrar relaciones entre factores que contribuyen a explicar el fracaso escolar (ver Cervini, et. al. 2017). Por otro lado, los distritos y las jurisdicciones escolares pueden servirse de la información que aportan las bases de las evaluaciones Aprender, dado que consta de una matriz de datos de más de 100 variables y 600.000 mil casos. Algunas de esas variables segmentan la base en desagregaciones que pueden ser útiles a las gestiones educativas. No obstante, resulta comprensible que los pedagogos no estén informados de los modelos analíticos que requieren base de datos de estas dimensiones. Sin embargo, la potencia de los ordenadores actuales es suficiente para correr modelos multivariados sobre estas bases con software específico de uso comercial y no comercial. Los modelos analíticos presentan otro tipo de complejidad, más relacionadas a las técnicas estadísticas posibles según el tipo de datos que proveen las bases usuarias (Castro y Lizasoain, 2012).

El trabajo interdisciplinario entre los encargados del gobierno de las instituciones y los técnicos en análisis de datos educativos, cobra una especial relevancia por la disponibilidad de datos abiertos a partir de la sanción de la ley 26.899 (Nakano y Azrilevich, 2017). En este marco se encuadra el presente artículo, que intenta mostrar las potencialidades de una técnica poco utilizada en el análisis de datos educativos: la regresión logística ordinal. Para ello, se desarrolla un ejemplo que intenta mostrar las posibles aplicaciones de dicho modelo, a partir de algunas variables del operativo Aprender 2023. Este artículo se encuadra en lo que Montero y León (2005) describen como estudio de tipo instrumental, en tanto propone la adaptación de un modelo estadístico a datos educativos abiertos.

Modelos de clúster y regresión para variables nominales y ordinales

La base de datos del Aprender 2023 nos ofrece muchas variables nominales que, por definición, es la escala de medición más elemental dado que se trata solo de un sistema clasificatorio. Son pocas las técnicas multivariadas que pueden aplicarse a este tipo de datos. En este sentido, una de las técnicas más usadas en ciencias sociales para mapear el comportamiento de múltiples variables en espacios dimensionales, es el Análisis de Componentes Principales (ACP), que es una técnica estadística que se utiliza para reducir la dimensionalidad de un conjunto de datos y para identificar las direcciones (componentes principales) en las que los datos muestran mayor variabilidad. Sin embargo, el ACP está diseñado específicamente para variables continuas y no es adecuado para variables nominales de manera directa. Una alternativa para trabajar con modelos de esta complejidad es la recodificación de las variables nominales. Es decir, aplicar transformaciones para convertirlas en variables numéricas. El siguiente problema a resolver es asegurar que todas las variables estén en la misma escala, lo cual se resuelve mediante la normalización de los datos. Todo este procedimiento supone una dificultad importante en el momento de interpretar los resultados, v.g. dar sentido a las dimensiones alcanzadas en las soluciones finales.

Ante el inconveniente de aplicar modelos de componente principales, se opta por otros basados en tablas de contingencia múltiples, el más frecuentes es el siguiente: Análisis de Correspondencias Múltiples (ACM), que es una técnica de correlación. Se trata de la extensión del Análisis de Correspondencias Simple (ACS) a múltiples variables categóricas. Permite representar gráficamente asociaciones entre categorías de diferentes variables en un espacio de menor dimensión, similar al Análisis de Componentes Principales (ACP) pero para datos categóricos. La solución final del modelo permite visualizar relaciones entre múltiples variables categóricas. Otros modelos son conocidos como Log-lineales, que son técnicas de

regresión. En muchos casos, se trata de extensiones del modelo de regresión de Poisson que permite analizar relaciones entre varias variables categóricas mediante una representación logarítmica de las frecuencias esperadas en una tabla de contingencia. Descompone los efectos principales e interacciones entre las variables y se usa en estudios de independencia y asociación en tablas de contingencia de varias dimensiones. Ambas técnicas permiten analizar datos categóricos de manera multivariada, pero mientras el ACM se enfoca en encontrar asociaciones y patrones de correlación, los modelos log-lineales buscan explicar relaciones de dependencia entre las variables en un marco de regresión.

Un caso particular de los métodos Log-lineales, es la regresión logística ordinal, que es una técnica estadística utilizada para modelar relaciones entre una variable dependiente categórica ordenada y una o más variables independientes. Se aplica en contextos donde la variable de respuesta presenta un orden inherente, pero las distancias entre categorías no pueden asumirse iguales (Agresti, 2010). La regresión logística ordinal ha sido ampliamente utilizada en estudios de ciencias sociales (Long, 1997; Menard, 2002, Hosmer, Lemeshow y Sturdivant, 2013).

En los operativos Aprender, las métricas de rendimiento se traducen a una escala ordinal, cuyas categorías son: por debajo del básico, básico, satisfactorio y avanzado. Los puntos de corte para estas categorías se explican en el apartado metodológico del operativo. Aquí nos interesa señalar la posibilidad de tratar el rendimiento académico como una variable dependiente ordinal. Como mencionamos anteriormente, otras variables se miden en escalas nominales dicotómicas. En este contexto, la regresión logística ordinal se presenta como un modelo de análisis adecuado para los datos disponibles.

El modelo matemático de la regresión logística ordinal puede resumirse como sigue: Sea Y una variable categórica ordinal con J categorías ordenadas ($Y \in \{1,2,\ldots,J\}$, La regresión logística ordinal modela la probabilidad acumulada de que la respuesta Y sea menor o igual a una categoría dada j: La regresión logística ordinal modela la probabilidad acumulada de que la respuesta Y sea menor o igual a una categoría dada j:

$$\text{LogPY} \leq j\text{PY} > j = j + 1x1 + 2x2 + \ldots kxk$$

Donde
$P(Y \leq j)$: es la probabilidad acumulada de que la respuesta esté en la categoría j o inferior,

α_j: representa un umbral específico de cada categoría,

β_k: son los coeficientes asociados a las variables independientes X_k,

X_k: cada una de las variables categóricas.

Este es el modelo logit acumulativo con el que trabaja el software SPSS, y asume que los efectos de los predictores son constantes en todas las categorías de la variable dependiente. Es decir, la relación entre cada par adyacente de categorías es similar en términos de los logaritmos de la razón de probabilidades, lo que implica la restricción de proporcionalidad de la razón de probabilidades o paralelismo (para una revisión de estos conceptos ver McCullagh, 1980).

Aplicar la regresión logística ordinal requiere considerar algunos supuestos que pueden resumirse como sigue: a) las categorías de la variable dependiente deben poseer un orden lógico, aunque las distancias entre ellas sean desconocidas; b) la razón de probabilidades debe ser proporcional, esto significa que los coeficientes (β_k) deben ser los mismos en todas las categorías de la variable dependiente, (Brant, 1990); c) no debe existir dependencia entre los casos analizados; y d) los predictores no deben estar altamente correlacionados. Los supuestos c y d son los más difíciles de cumplir, aunque su violación no es restrictiva para el análisis. De sospechar dependencia y multicolinealidad, se debe proceder con un cuidadoso análisis descriptivo previo de los datos (un tratamiento detallado de estos supuestos puede encontrarse en Hosmer, Lemeshow y Sturdivant, 2013).

Propuesta del modelo

Para ejemplificar la aplicación de un modelo de regresión logística basado en algunas variables del operativo Aprender, en esta sección se describirán las variables seleccionadas junto con el esquema conceptual de dicho modelo. Entre las variables independientes escogidas para el modelo se tomaron las diferentes preguntas agrupadas en la categoría uso del tiempo libre. Aquí, el alumnado responde a una serie de preguntas, que son: a) Usar redes sociales; b) Ver series, películas o videos; d) Leer libros que no te dieron tus maestras o maestros de la escuela; e) Realizar actividades artísticas, deportivas, físicas o juegos al aire libre; f) Juntarte con amigas y amigos; g) Jugar de manera digital (online u offline); h) Aprender idiomas fuera de la escuela; i) Crear contenidos para redes sociales (YouTube, Twitch, Instagram, Tik-Tok, etc.). Este conjunto de variables fue medida en

una escala nominal cuyas principales categorías de respuestas fueron: 1= SI, 2= NO; para los casos en donde no hubo respuesta, aparecen con multimarca, o no disponible, se usaron valores negativos en su codificación por defecto en la base usuaria. En este sentido el filtrado de las respuestas adecuadas para el análisis se realizó mediante la condición de respuesta ≥1. La otra variable independiente fue Sector de Gestión de la escuela, cuya codificación fue: 1=Estatal y 2=Privado; aunque la base no tiene otra codificación, el filtrado de casos se realizó excluyendo casillas vacías.

La variable dependiente utilizada para este ejemplo fue el desempeño en lengua, medido según las siguientes categorías: 1=Por debajo del básico; 2=Básico; 3=Satisfactorio; y 4=Avanzado. A partir del filtrado de las variables independientes, la base no tuvo casos en los que no se registraran respuestas en la variable dependiente.

Conceptualmente, el modelo pretende estimar probabilísticamente, la categoría de pertenencia en los niveles de lectura alcanzado, a partir de la combinación de actividades realizadas (o no) durante el tiempo libre por parte del estudiantado. Por lo tanto, Y es el nivel de desempeño en lectura medido en una variable ordinal; X_j son las diferentes variables independientes correspondiente al uso del tiempo libre; α_j y β_k son los parámetros a estimar en el modelo.

Desarrollo del ejemplo

En la tabla 1 se muestra el total de casos disponibles del total de la base usuaria, luego que se realizara el filtrado de casos en las variables dependientes. Para la variable independiente no se registraron casos nulos o vacíos.

Tabla 1
Total de casos analizados por sector de gestión

	Frecuencia	Porcentaje
Estatal	467308	72,8
Privado	174698	27,2
Total	642006	100,0

Nota: la tabla muestra la cantidad de casos analizados en el modelo, según distribución proporcional por nivel de gestión, luego de filtrar los datos.

Puesto que, para los análisis realizados con SPSS sobre la base usuaria, se usa la codificación por defecto, la tabla 2 reproduce las etiquetas y las referencias a las variables utilizadas en el modelo de regresión.

Tabla 2
Referencias de las etiquetas de las variables independiente y dependientes

Variable Dependiente		
ldesemp	Desempeño en Lengua	Ordinal
Variables Independientes		
sector	Sector de Gestión	Nominal
ap05a	Usar redes sociales	Nominal
ap05b	Ver series, películas o videos	Nominal
ap05c	Leer libros que no te dieron tus maestras o maestros de la escuela	Nominal
ap05d	Realizar actividades artísticas, deportivas, físicas o juegos al aire libre	Nominal
ap05e	Juntarte con amigas y amigos	Nominal
ap05f	Jugar de manera digital (online u offline)	Nominal
ap05g	Aprender idiomas fuera de la escuela	Nominal
ap05h	Crear contenidos para redes sociales (YouTube, Twitch, Instagram, Tik-Tok, etc.)	Nominal

Nota: en esta tabla, la primera columna muestra la etiqueta abreviada de las variables tal como se muestra en las bases usuarias. La segunda columna corresponde a la denominación completa de la variable, y la tercera columna la escala de medición correspondiente.

El primer paso luego de ejecutar la regresión ordinal es verificar la información de ajuste de los modelos. Los valores tabulares en este caso son fundamentales para evaluar si el modelo propuesto es útil. La tabla 3 muestra el ajuste del modelo según lo presenta SPSS.

Tabla 3
Parámetros de ajuste del modelo

Modelo	Logaritmo de la verosimilitud -2LL	Chi-cuadrado	gl	Sig.
Sólo intersección	61382,011			
Final	13591,258	47790,753	9	0,000

Nota: la siguiente tabla muestra la utilidad del modelo incluyendo predictores, junto con su nivel de significación.

Aquí puede verse el Modelo nulo (solo intersección), y el modelo final (con predictores). El modelo final muestra la diferencia entre ambos. En las columnas encontramos desglosado el logaritmo de la verosimilitud (-2LL), el valor de Chi-cuadrado, los grados de libertad y el nivel de significación. El logaritmo de la verosimilitud es una medida de ajuste del modelo, en este sentido es similar a la suma de cuadrados de los residuos en la regresión lineal. Si el modelo nulo es la predicción más elemental y el modelo final es la predicción incluyendo las variables independientes, se espera un cambio en el valor de este parámetro: que sea menor a medida que se incluyen los predictores. En otras palabras, si se toma como medida de error, se espera que el modelo final muestre un valor menor. La distancia entre los modelos se pondera con el estadístico Chi-cuadrado. Este valor representa la diferencia entre los dos modelos (nulo vs. final); un valor grande de chi-cuadrado sugiere una mejora sustancial. El nivel de significación es el valor p asociado a chi-cuadrado; si $p<0.05$, podemos concluir que nuestro modelo con predictores es significativamente mejor que el modelo nulo. En este ejemplo, tenemos que la inclusión de los predictores mejora la capacidad predictiva del modelo. Esto es, un resultado significativo nos dice que al menos una de las variables independientes está relacionada con la variable dependiente ordinal. La magnitud de chi-cuadrado da una idea de cuán fuerte es esta relación en conjunto, la reducción en el logaritmo de la verosimilitud indica cuánta incertidumbre se ha eliminado al incluir los predictores.

De la tabla anterior se desprende que al introducir las variables independientes se ha mejorado la capacidad predictiva del modelo. Corresponde ahora verificar la bondad de ajuste. SPSS proporciona dos pruebas importantes: Pearson y Desvianza, la tabla 4 muestra estos estadísticos tal como lo presenta el software.

Tabla 4
Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	6764,517	1515	0,000
Desvianza	6690,914	1515	0,000

Nota: la bondad de ajuste del modelo evalúa la capacidad predictiva del modelo propuesto

Aquí es importante entender un aspecto poco intuitivo: a diferencia de la prueba chi-cuadrado presentada en el ajuste del modelo, en estas pruebas de bondad de ajuste se espera no obtener resultados significativos ($p>0.05$). Esto tiene una explicación lógica. La prueba de Pearson compara las frecuencias observadas con las frecuencias esperadas según nuestro modelo, y esto es una forma de medir la discrepancia entre lo que predice nuestro modelo y lo que realmente ocurre en los datos. Un nivel de significación $p>0.05$, indica que no hay diferencias significativas entre las predicciones del modelo y los datos observados. Del mismo modo, la prueba de Desvianza, que es una alternativa a la prueba de Pearson, utiliza el criterio de razón de verosimilitud. Siguiendo este razonamiento, un nivel de significación $p>0.05$, resulta favorable para el modelo. En nuestro análisis nos encontramos con valores de bondad de ajuste significativos, por lo cual entenderíamos que el modelo no se ajusta bien a los datos. Pero esto tiene una explicación; las pruebas de bondad de ajuste son sensibles al tamaño de la muestra y al número de patrones de covariables. Si se tienen muchas combinaciones posibles de variables independientes los resultados son menos confiables. En el ejemplo que estamos desarrollando es necesario considerar que la base de datos donde se corrió el modelo tiene más de seis mil casos; con muestras grandes, las pruebas de bondad de ajuste tienden a volverse significativas con mucha facilidad. Esto ocurre porque estas pruebas son muy sensibles al tamaño de la muestra, lo que puede llevarnos a rechazar modelos que en realidad son útiles en la práctica. En otras palabras, con muestras grandes, estas pruebas pueden detectar diferencias tan pequeñas entre el modelo y los datos que, aunque sean estadísticamente significativas, podrían no ser prácticamente relevantes. En casos, de muestras grandes y pruebas de bondad de ajuste significativas, se debe considerar medidas alternativas de ajuste, tales como los pseudo R-cuadrado (Cox y Snell, Nagelkerke, McFadden). Estos indicadores pueden mostrar un ajuste aceptable, incluso si las pruebas de bondad de ajuste son significativas. En este ejemplo los valores de pseudo R-cuadrado se muestran en la tabla 5.

Tabla 5
Pseudo R-cuadrado

Cox y Snell	0,108
Nagelkerke	0,117
McFadden	0,045

Nota: la tabla muestra los tres estadísticos que ofrece el software SPSS para expresar el porcentaje de varianza explicada por el modelo.

Al analizar la bondad de ajuste y el pseudo R-cuadrado, se aprecia que el modelo predictivo planteado no es el más adecuado, aunque el porcentaje de varianza explicada (~11%) en el contexto de todas las variables que incluye el Operativo, no es menor. Es decir, el modelo intenta predecir el nivel de rendimiento lector solo a partir de ocho variables que describen el uso del tiempo libre, y el sector de gestión de la escuela (siendo estas, variables dicotómicas). En este contexto es que se afirma que la proporción de varianza explicada es estadísticamente importante. No obstante, corresponde profundizar en la lectura de los estadísticos de la tabla 5. Los pseudo R-cuadrado en regresión ordinal no pueden interpretarse exactamente como el R-cuadrado tradicional de la regresión lineal. Sin embargo, nos dan una idea de la magnitud del efecto o la fuerza de la asociación entre las variables predictoras y la variable dependiente. Los valores tabulados sugieren que el modelo está explicando una cantidad relativamente modesta de la variación en la variable dependiente. Concretamente, valores alrededor del 0.1, representa un efecto pequeño. Sin embargo, hay varios factores importantes a considerar: a) con muestras grandes (n > 6000) sugiere que estos efectos, aunque pequeños, son robustos. Con una muestra tan grande, es menos probable que estos resultados se deban al azar. El valor más bajo de McFadden (0.04) no es inusual, ya que este índice tiende a producir valores más conservadores. Valores entre 0.2 y 0.4 se considerarían ya bastante buenos, así que 0.04 indica un efecto pequeño, pero no necesariamente insignificante. Resumiendo, las variables independientes están explicando aproximadamente el 11% de la variación en la variable dependiente. Aunque el efecto es pequeño, con una muestra de más de 6000 casos, podemos confiar en que este efecto es real y no producto del azar. Si se quisiera mejorar el modelo se debería explorar posibles interacciones entre las variables existentes, y verificar si alguna de las variables independientes podría tener una relación no lineal con la variable dependiente (para una revisión de estos conceptos ver Fullerton, 2009; O'Connell, 2006).

Corresponde ahora recorrer los valores de tabla con la estimación de parámetros, para saber qué variables contribuyen a la relación, y cómo se interpreta el ordenamiento de las variables. En la tabla 6 se muestran las estimaciones de los parámetros en la regresión ordinal según SPSS.

Tabla 6
Estimación de parámetros

		Estimación	Error estándar	Wald	gl	Sig.
Umbral	[ldesemp = 1]	-2,858	0,015	36466,189	1	0,000
	[ldesemp = 2]	-1,488	0,014	10719,082	1	0,000
	[ldesemp = 3]	0,259	0,014	333,874	1	0,000
Ubicación	[sector=1]	-1,012	0,006	24774,489	1	,000
	[sector=2]	0	.	.	0	.
	[ap05a=1]	0,410	0,009	1902,485	1	0,000
	[ap05a=2]	0	.	.	0	.
	[ap05b=1]	,040	0,008	22,310	1	0,000
	[ap05b=2]	0	.	.	0	.
	[ap05c=1]	0,135	0,006	522,415	1	0,000
	[ap05c=2]	0	.	.	0	.
	[ap05d=1]	0,003	0,008	,139	1	0,709
	[ap05d=2]	0	.	.	0	.
	[ap05e=1]	-0,373	0,007	2506,915	1	0,000
	[ap05e=2]	0	.	.	0	.
	[ap05f=1]	0,405	0,006	4114,543	1	0,000
	[ap05f=2]	0	.	.	0	.
	[ap05g=1]	0,121	0,006	382,032	1	0,000
	[ap05g=2]	0	.	.	0	.
	[ap05h=1]	-0,566	0,006	8473,989	1	0,000
	[ap05h=2]	0	.	.	0	.

Nota: la tabla muestra la estimación de los parámetros para cada una de las variables incluidas en el modelo. La primera columna contiene las etiquetas de las variables, cuya denominación completa se mostró en la tabla 2.

La estructura de la tabla que muestra SPSS, organiza las filas en dos secciones principales. Umbrales: representan los puntos de corte entre las categorías de nuestra variable dependiente ordinal. Si tenemos 4 niveles en la variable dependiente, veremos 3 umbrales (el número de categorías menos 1). Ubicación: lista las variables independientes dicotómicas. Para cada una, SPSS muestra una categoría de referencia

(generalmente codificada como 0) y compara con la otra categoría (codificada como 1). Para cada fila, la tabla muestra: a) Estimación (β) o coeficiente de regresión, b) Error estándar, c) Estadístico de Wald o prueba de significación, d) Grados de Libertad y e) Significación estadística. En este ejemplo se ha omitido la columna para el Intervalo de confianza (95%).

Las variables que contribuyen significativamente a la relación, son aquellas que en la columna Significación (Sig.), muestran un valor p<0.05. El signo del coeficiente (Estimación) indica la dirección del efecto. Si es positivo aumenta la probabilidad de estar en categorías más altas, de ser negativo, aumenta la probabilidad de estar en categorías más bajas.

En el ejemplo los umbrales nos indican los puntos de corte. De este modo, -2,858 es el punto donde los sujetos tienden a pasar de la categoría 1 a la 2. Para las variables independientes: Sector alcanza la significación estadística (p<0,00), su coeficiente es negativo (-1,012), sugiriendo que los alumnos en esa categoría tienen mayor probabilidad de estar en niveles más bajos. Para la variable Realizar actividades artísticas, deportivas, físicas o juegos al aire libre (ap05d), la variable no alcanza un nivel de significación estadística (p=0,709), por lo cual se considera que no tiene efecto sobre la variable dependiente.

El ordenamiento de las filas tiene un propósito lógico. Primero los umbrales, ordenados de menor a mayor categoría; luego las variables independientes, en el orden en que fueron introducidas en el modelo. Esto da coherencia para entender los resultados del modelo. Los umbrales en una regresión ordinal representan los puntos de corte en la escala logística donde la probabilidad de estar en una categoría cambia a la siguiente. Con cuatro niveles de desempeño, tenemos tres umbrales que marcan estas transiciones: ldesemp 1 = -2.858; este primer umbral marca el punto de transición entre “por debajo del básico” y “básico”. El valor negativo nos indica que se necesita una puntuación logística relativamente baja para superar el nivel más bajo. En otras palabras, cuando la combinación de las variables independientes produce un valor menor a -2.858 en la escala logística, es más probable que el estudiante esté en la categoría “por debajo del básico”. Luego, ldesemp 2 = -1.488, el segundo umbral representa el punto de transición entre “básico” y “satisfactorio”. El valor sigue siendo negativo, pero es mayor que el anterior, lo que indica que se necesita una puntuación logística más alta para alcanzar el nivel satisfactorio. Cuando la puntuación logística está entre -2.858 y -1.488, es más probable que el estudiante esté en el nivel “básico”. Finalmente, ldesemp 3 = 0.259, es el tercer umbral, que marca la transición entre “satisfactorio” y “avanzado”. Es interesante notar que este es el único valor positivo, lo cual indica que se necesita una puntuación logística más alta para alcanzar el nivel avanzado. Cuando la puntuación logística está entre -1.488 y 0.259, es más probable que el estudiante esté en el nivel “satisfactorio”, y cuando supera 0.259, es más probable que esté en el nivel “avanzado”. La distribución de estos umbrales marca un punto importante sobre la escala de desempeño: la distancia entre el primer y segundo umbral es 1.37 unidades logísticas (-1.488 - (-2.858)); la distancia entre el segundo y tercer umbral es 1.747 unidades logísticas (0.259 - (-1.488)), lo cual indican que la transición entre niveles no es equidistante en la escala logística. En otras palabras, se requiere un mayor “salto” para pasar de satisfactorio hacia avanzado, que para pasar de por debajo del básico a básico. La escala discrimina bien entre los diferentes niveles de desempeño, ya que los umbrales están claramente separados entre sí.

Continuando con las siguientes filas de la tabla, incorporamos las variables sector de gestión y uso de redes sociales. La estimación para el sector estatal (categoría 1) es -1.012, mientras que para el sector privado (categoría 2) es 0 por ser la categoría de referencia. Este coeficiente negativo nos dice algo importante: los estudiantes de escuelas estatales (comparados con los de privadas) tienen una mayor probabilidad de estar en categorías más bajas de desempeño. Esto es; manteniendo todas las demás variables constantes, estar en una escuela estatal reduce en 1.012 unidades el logaritmo de las probabilidades acumuladas de alcanzar un nivel más alto de desempeño. Para la variable “Usar redes sociales”, la estimación para quienes SÍ usan redes sociales (categoría 1) es 0.410, mientras que para quienes NO usan (categoría 2) es 0 por ser la categoría de referencia. Este coeficiente positivo nos indica que los estudiantes que usan redes sociales tienen mayor probabilidad de estar en categorías más altas de desempeño. Específicamente, usar redes sociales aumenta en 0.41 unidades el logaritmo de las probabilidades acumuladas de alcanzar un nivel más alto de desempeño.

Relacionando estas estimaciones con los umbrales anteriores (-2.858, -1.488, 0.259): para un estudiante de escuela estatal que usa redes sociales, su puntuación logística base sería: (-1.012) + (0.410) = -0.602. Esta puntuación nos ayuda a predecir en qué nivel de desempeño es más probable que se

encuentre el estudiante. Puesto que -0.602 está por encima de -1.488 pero por debajo de 0.259, este estudiante tendría mayor probabilidad de estar en el nivel "satisfactorio". En cambio, para un estudiante de escuela privada que usa redes sociales: (0) + (0.410) = 0.410. Como 0.410 es mayor que 0.259, este estudiante tendría mayor probabilidad de estar en el nivel "avanzado".

La función de enlace Logit nos indica que estamos trabajando con el logaritmo natural de las probabilidades acumuladas, lo que significa que para obtener probabilidades reales necesitaríamos transformar estos valores usando la función exponencial. Para convertir estos valores en probabilidades que podamos interpretar fácilmente, se utiliza la función exponencial (e^x). Para nuestro ejemplo: un estudiante de escuela estatal que usa redes sociales:

Valor logit = -1.012 (estatal) + 0.410 (usa redes) = -0.602
Para calcular las probabilidades acumuladas, se toma cada umbral (U), usando la fórmula:
$P(Y \leq k) = e^{(Uk - X)} / (1 + e^{(Uk - X)})$

Donde:
Uk: es el valor del umbral
X: valor logit (-0.602 para uso de redes sociales)
Al aplicar esta fórmula para cada nivel se obtiene:
Por debajo del básico:

$P(Y \leq 1) = e^{(-2.858 - (-0.602))} / (1 + e^{(-2.858 - (-0.602))}) = e^{(-2.256)} / (1 + e^{(-2.256)}) \approx 0.095$ o 9.5%
Básico o inferior:
$P(Y \leq 2) = e^{(-1.488 - (-0.602))} / (1 + e^{(-1.488 - (-0.602))}) = e^{(-0.886)} / (1 + e^{(-0.886)}) \approx 0.292$ o 29.2%
Satisfactorio o inferior:
$P(Y \leq 3) = e^{(0.259 - (-0.602))} / (1 + e^{(0.259 - (-0.602))}) = e^{(0.861)} / (1 + e^{(0.861)}) \approx 0.703$ o 70.3%

Calculando las probabilidades específicas para cada nivel, se obtiene:
Por debajo del básico = $P(Y \leq 1) = 9.5\%$
Básico = $P(Y \leq 2) - P(Y \leq 1) = 29.2\% - 9.5\% = 19.7\%$
Satisfactorio = $P(Y \leq 3) - P(Y \leq 2) = 70.3\% - 29.2\% = 41.1\%$
Avanzado = $1 - P(Y \leq 3) = 1 - 0.703 = 29.7\%$
Interpretación práctica: para un estudiante de escuela estatal que usa redes sociales, las probabilidades de estar en cada nivel son:
9.5% de probabilidad de estar por debajo del básico
19.7% de probabilidad de estar en nivel básico
41.1% de probabilidad de estar en nivel satisfactorio
29.7% de probabilidad de estar en nivel avanzado
Comparado con un estudiante de escuela privada que usa redes sociales, se aplican los mismos cálculos con valor logit = 0 + 0.410 = 0.410. Obtendríamos en este caso:
3.8% de probabilidad de estar por debajo del básico
10.2% de probabilidad de estar en nivel básico
35.5% de probabilidad de estar en nivel satisfactorio
50.5% de probabilidad de estar en nivel avanzado

Esta comparación nos muestra claramente el efecto del sector escolar: los estudiantes de escuelas privadas tienen una probabilidad mayor de alcanzar el nivel avanzado (50.5% vs 29.7%) cuando se mantiene constante el uso de redes sociales.

Para continuar con este ejemplo, repasaremos el modelo de regresión en SPSS interpretando la ecuación de regresión para las siguientes variables relevantes; Variable dependiente: nivel de desempeño en lengua (1=Por debajo del básico, 2=Básico, 3=Satisfactorio, 4=Avanzado). Variables independientes: a) sector de gestión (1 = Estatal (-1.012), 2 = Privado (0, categoría de referencia), b) Usar redes sociales (1 = Sí (0.410), 2 = No (0, categoría de referencia), c) Juntarte con amigas y amigos (1 = Sí (-0.373), 2 = No (0, categoría de referencia), d) Jugar de manera digital (1 = Sí (0.405), 2 = No (0, categoría de referencia), d) Aprender idiomas fuera de la escuela (1 = Sí (-0.566), 2 = No (0, categoría de referencia). Umbrales: θ_1 = -2.858 (transición a básico), θ_2 = -1.488 (transición a satisfactorio), θ_3 = 0.259 (transición a avanzado).

Tomando como referencia un caso específico; un estudiante con el siguiente perfil: Escuela estatal (sector=1), Usa redes sociales (ap05a=1), Se junta con amigas y amigos (ap05e=1), Juega de manera digital (ap05f=1); Aprende idiomas fuera de la escuela (ap05h=1). Su valor logit total es igual a
(-1.012) + (0.410) + (-0.373) + (0.405) + (-0.566) = -1.136

Calculando las probabilidades acumuladas usando la función exponencial:

Probabilidad de estar en "por debajo del básico" o inferior:

$P(Y \leq 1) = e^{(-2.858 - (-1.136))} / (1 + e^{(-2.858 - (-1.136))}) = e^{(-1.722)} / (1 + e^{(-1.722)}) \approx 0.151$ o 15.1%
Probabilidad de estar en "básico" o inferior:
$P(Y \leq 2) = e^{(-1.488 - (-1.136))} / (1 + e^{(-1.488 - (-1.136))}) = e^{(-0.352)} / (1 + e^{(-0.352)}) \approx 0.413$ o 41.3%

Probabilidad de estar en “satisfactorio” o inferior:
P(Y ≤ 3) = e^(0.259 - (-1.136)) / (1 + e^(0.259 - (-1.136)))= e^(1.395) / (1 + e^(1.395))≈0.802 o 80.2%

Probabilidades específicas para cada nivel:
Por debajo del básico = 15.1%
Básico = 41.3% - 15.1% = 26.2%
Satisfactorio = 80.2% - 41.3% = 38.9%
Avanzado = 100% - 80.2% = 19.8%

Sintetizando, para un estudiante con este perfil específico, las probabilidades predichas son:

15.1% de probabilidad de estar por debajo del básico
26.2% de probabilidad de estar en nivel básico
38.9% de probabilidad de estar en nivel satisfactorio
19.8% de probabilidad de estar en nivel avanzado

La mayor probabilidad (38.9%) corresponde al nivel satisfactorio, seguido por el nivel básico (26.2%). Esto sugiere que, con esta combinación de características, el estudiante tiene más probabilidades de alcanzar un nivel de desempeño intermedio. Si ahora modificamos el perfil del estudiante, de modo que en la variable sector de gestión corresponde a escuela privada, las probabilidades cambian en el siguiente sentido. Primero, los valores umbrales se mantienen, esto es: θ_1 = -2.858 (transición a básico), θ_2 = -1.488 (transición a satisfactorio), y θ_3 = 0.259 (transición a avanzado). Pero para este nuevo perfil, el valor logit total es: a) Escuela privada (sector=2): 0 (categoría de referencia), b) Usa redes sociales (ap05a=1): 0.410, c) Se junta con amigas y amigos (ap05e=2): 0 (categoría de referencia), d) Juega de manera digital (ap05f=1): 0.405, e) Crear contenidos para redes sociales (ap05h=2): 0 (categoría de referencia). Su valor logit total = 0 + 0.410 + 0 + 0.405 + 0 = 0.815.

Calculando las probabilidades acumuladas usando la función exponencial para cada nivel:

Probabilidad de estar en “por debajo del básico” o inferior:
P(Y ≤ 1) = e^(-2.858 - 0.815) / (1 + e^(-2.858 - 0.815))= e^(-3.673) / (1 + e^(-3.673))≈ 0.025 o 2.5%

Probabilidad de estar en “básico” o inferior:
P(Y ≤ 2) = e^(-1.488 - 0.815) / (1 + e^(-1.488 - 0.815))= e^(-2.303) / (1 + e^(-2.303))≈ 0.091 o 9.1%

Probabilidad de estar en “satisfactorio” o inferior:
P(Y ≤ 3) = e^(0.259 - 0.815) / (1 + e^(0.259 - 0.815))= e^(-0.556) / (1 + e^(-0.556))≈ 0.364 o 36.4%

Ahora podemos calcular las probabilidades específicas para cada nivel:
Por debajo del básico = 2.5%
Básico = 9.1% - 2.5% = 6.6%
Satisfactorio = 36.4% - 9.1% = 27.3%
Avanzado = 100% - 36.4% = 63.6%

Sintetizando, este nuevo perfil muestra una distribución de probabilidades muy diferente al caso anterior. Observemos los cambios más significativos: La probabilidad de estar en nivel avanzado aumentó a 63.6% (comparado con el 19.8% del perfil anterior). Este incremento sustancial se debe principalmente a: Estar en una escuela privada (eliminando el efecto negativo de -1.012 del sector estatal), No tener presente la condición Crear contenidos para redes sociales (eliminando el efecto negativo de (-0.566). Las probabilidades de estar en niveles más bajos disminuyeron considerablemente:

Solo 2.5% de probabilidad de estar por debajo del básico (vs 15.1% anterior)
6.6% de probabilidad de estar en nivel básico (vs 26.2% anterior)
27.3% de probabilidad de estar en nivel satisfactorio (vs 38.9% anterior)
Esta comparación nos permite ver claramente cómo el sector escolar y la ausencia de factores con coeficientes negativos tienen un impacto sustancial en el desempeño esperado. El perfil actual tiene una probabilidad mucho mayor de alcanzar niveles más altos de desempeño, especialmente el nivel avanzado.

Discusión

La intención de este artículo ha sido mostrar las propiedades de un modelo estadístico particular, la regresión logística ordinal, para el análisis de datos categóricos (nominales y ordinales), tal y como se obtienen de las bases de datos usuaria del Operativo Aprender. Hemos aplicado este modelo como ejemplo, tomando un conjunto de variables independientes relativas al uso del tiempo libre y la gestión escolar, y como variable dependiente el nivel de rendimiento lector. Puesto que la intención fue mostrar las bondades del modelo estadístico, es que no se avanzó con ninguna explicación teórica del impacto que pudieran tener los distintos modos de usar el tiempo libre sobre el desempeño en lengua. Como se mostró en la presentación de los resultados, el conjunto de variables independientes escogidas, explican un porcentaje modesto de la varianza sobre el rendimiento lector. Por lo tanto, con una guía teórica apropiada y con hipótesis a priori derivadas de la misma, el modelo estadístico se

muestra como una herramienta versátil para toda una gama de variables categóricas contenidas en las bases usuarias disponibles. En otras palabras, la regresión logística ordinal, es una base metodológica sólida para indagar en profundidad aspectos relacionados al desempeño académico mediante una combinación de variables teóricamente orientada, que permitan: a) mejorar los ajustes del modelo, b) optimizar los parámetros a estimar, y c) aumentar la proporción de varianza a explicar.

Otra ventaja que conviene señalar es la facilidad de aplicación en comparación con otros modelos multivariados que requieren transformaciones sobre las variables originales, tal el caso del ACP. Por otro lado, las ecuaciones de regresión generadas por el software requieren que los valores de probabilidad sean obtenidos de los valores logísticos originales. En tal caso, la función de transformación de programabilidad que ofrece SPSS permite automatizar todo el cálculo. En este sentido, el software también permite ejecutar programas de Python para el mismo propósito.

Finalmente, este ejemplo intenta mostrar que la selección de un modelo apropiado, puede servir para explotar mucha información de las bases disponibles de los operativos de evaluación de estudiantes. Muchas preguntas, que no necesariamente surgen del campo de la investigación educativa, sino más bien de las gestiones, pueden reunir a educadores sin experticia y expertos en análisis de datos en una sinergia de trabajo, en donde los primeros proponen las preguntas pertinentes, y los segundos modelizan las respuestas en función de la información empírica disponible. Consideramos que este será en un futuro un ámbito de trabajo que construya saberes a partir del ejercicio de la interdisciplina, especialmente a partir de la progresiva difusión de herramientas de inteligencia artificial dedicada a crear códigos para el análisis estadístico, especialmente en software que hasta el presente no han sido amigables a los científicos sociales (v.g. R o Python).

Referencias

Agresti, A. (2010). Analysis of ordinal categorical data (2.ª ed.). John Wiley & Sons.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics, 46(4), 1171–1178.

Castro, M., & Lizasoain, L. (2012). Las técnicas de modelización estadística en la investigación educativa: Minería de datos, modelos de ecuaciones estructurales y modelos jerárquicos lineales. Revista Española de Pedagogía, 70(251), 131–148. http://www.jstor.org/stable/23766443

Cervini, R., Dari, N., & Quiroz, S. (2017). Repitencia y rendimiento escolar en la educación primaria de América Latina: Los datos del TERCE. En R. Cervini (Comp.), El fracaso escolar: Diferentes perspectivas disciplinarias (pp. xx–xx). Universidad Nacional de Quilmes.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. Sociological Methods & Research, 38(2), 306–347. https://doi.org/10.1177/0049124109346162

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3.ª ed.). John Wiley & Sons.

Long, J. S. (1997). Regression models for categorical and limited dependent variables. Sage Publications.

McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2), 109–142.

Menard, S. (2002). Applied logistic regression analysis (2.ª ed.). Sage Publications.

Ministerio de Capital Humano, Secretaría de Educación. (2016–2024). Bases usuarias de los Operativos Aprender. https://www.argentina.gob.ar/educacion/evaluacion-informacion-educativa/aprender

Montero, I., & León, O. G. (2005). Sistema de clasificación del método en los informes de investigación en psicología. International Journal of Clinical and Health Psychology, 5(1), 115–127.

Mosteiro, C. E. (2018). Las evaluaciones Aprender en Argentina: Algunas reflexiones respecto a su definición de calidad. X Jornadas de Sociología de la Universidad Nacional de La Plata. http://jornadassociologia.fahce.unlp.edu.ar/x-jornadas/actas/MosteiroPONMesa20.pdf

Nakano, S., & Azrilevich, P. A. (2017). El acceso abierto y la implementación de la Ley 26.899 en la Argentina. En VII Conferencia Internacional sobre Bibliotecas y Repositorios Digitales de América Latina.

O'Connell, A. A. (2006). Logistic regression models for ordinal response variables (Vol. 146). Sage Publications.

Ravela, P., Arregui, P., Valverde, G., Wolfe, R., Ferrer, G., Martínez, F., Aylwin, M., & Wolff, L. (2008). *Las evaluaciones educativas que América Latina necesita* (Documento de Trabajo N.º 40). PREAL.

Rivas, A. (2015). *América Latina después de PISA: Lecciones aprendidas de la educación en siete países* (2000–2015). CIPPEC.

Rodríguez, L. R., Vior, S. E., & Más Rocha, S. M. (2018). *Las políticas de evaluación de la calidad educativa en Argentina* (2016–2018). Educação & Realidade, 43, 1405–1428.