Abstraction & Application **13** (2015) 4 − 14 UADY

Análisis de la deserción estudiantil del nivel Introductorio del Instituto Confucio de la UADY

Ernesto Guerrero Lara[a], Henry Pantí Trejo[b], Álvaro Sánchez Marín[c]

Facultad de Matemáticas, Universidad Autónoma de Yucatán, México

[a]ernesto.guerrero@correo.uady.mx, [b]henry.panti@correo.uady.mx, [c]alvaro.sanchez.marin@gmail.com

Abstract

In this paper we compute the probability of a student drops out of Mandarin Chinese courses at the Confucius Institute of the Autonomous University of Yucatan through a logistic regression model with dichotomous variables. In addition, the logistic regression analysis helped to identify the main factors that make students either fail the introductory level of Mandarin Chinese course or desert the program once approved the introductory level, which is the most outstanding of all levels.

Resumen

En este artículo se utilizó un modelo de regresión logística con variables dicotómicas para analizar datos con el fin de estimar la probabilidad de que un alumno deserte de los cursos de Chino Mandarín del Instituto Confucio de la Universidad Autónoma de Yucatán. El análisis de regresión logística nos permitió identificar los principales factores que influyen en que un alumno repruebe o no continúe una vez aprobado el nivel introductorio de los cursos de Chino Mandarín, que ha mostrado ser el de mayor relevancia de todos los niveles.

Keywords and phrases : Logit, Odds, Probability

2010 *Mathematics Subject Classification*: 6207, 62J02, 62P25.

1. Introducción

De principios de siglo a la fecha, la sociedad mundial ha manifestado un interés por conocer y aprender todo lo relacionado con el país China. Siendo China una potencia económica mundial, el idioma Chino Mandarín es una herramienta indispensable para la economía, los negocios, además de ser un puente para conocer y apreciar la cultura de este país.

Desde su creación, los Institutos Confucio en todo el mundo tienen el propósito de promover el conocimiento de la cultura y lengua china a través de conferencias, festivales, simposios y cursos de Chino Mandarín. En el año 2013 había un total de 120 Institutos Confucio a nivel mundial. Un hecho importante, señalado en el reporte anual del 2013 de las oficinas de los Institutos Confucio (HANBAN), es que el número de alumnos inscritos en los cursos de Chino Mandarín impartidos por estas instituciones, aumentó de 655,000 en 2012 a 850,000 en 2013, un aumento que equivale al 30 %, esto muestra el gran interés por estudiar este idioma [1].

En México, particularmente en el Instituto Confucio (**IC**) de la Universidad Autónoma de Yucatán (UADY), los cursos de Chino Mandarín constan de 7 niveles, siendo el nivel Introductorio el primero de

ellos. Los principales objetivos del nivel Introductorio son: familiarizar al alumno con los cuatro tonos del idioma chino mandarín, el *pinyin*, la pronunciación, la escritura, la composición de los caracteres chinos y el dominio del vocabulario de uso cotidiano.

Un problema persistente en cualquier institución docente es la deserción de sus estudiantes. El IC no es la excepción y es un hecho, por demás conocido, la constante preocupación de los directivos del IC por preservar su matrícula. Los factores que pueden conducir a la deserción de un estudiante son diversos. Uno de los objetivos de este trabajo es identificar factores que pudieran favorecer la deserción de un estudiante del IC.

Las opiniones de los docentes y personal administrativo del IC coinciden en que el nivel Introductorio es el que presenta mayor número de deserción de estudiantes. Estas opiniones y experiencias motivaron a enfocar nuestro trabajo en el estudio del comportamiento del estudiantado en este nivel. A su vez, esta información apriori nos permite definir o considerar como deserción de un alumno del IC, al hecho de que el alumno no continúe al nivel posterior al Introductorio (ya sea por reprobación o por cualquier otra circunstancia que se presente).

En nuestra investigación nos enfocamos en estimar la probabilidad de que los alumnos del nivel Introductorio del IC de la UADY no continúen al nivel I en el período inmediato siguiente a la conclusión del nivel Introductorio. Asimismo, se identifican los principales factores que influyen para que un alumno repruebe el curso Introductorio o no continúe al siguiente nivel, aun habiendo aprobado el curso. Identificar estos factores, puede ayudar al IC de la UADY a establecer medidas adecuadas que ayuden a solucionar la problemática de la deserción de alumnos en el nivel introductorio.

El nivel Introductorio se ha ofrecido desde la fundación del IC de la UADY en el 2008. La información con la que se cuenta, y que utilizamos para nuestro estudio, es una base de datos de los cinco cursos del nivel Introductorio realizados entre enero del 2012 y diciembre del 2013, período en el cual se realizó el curso Introductorio dos veces de forma trimestral y tres de forma semestral. De los cursos que consideramos, sólo utilizamos los datos de las clases que se imparten de lunes a viernes y cuya edad mínima de los alumnos es 16 años. Las clases tienen una duración de dos horas, una vez por semana. Todo alumno inscrito en el nivel Introductorio debe cumplir con este horario de clases.

Para la realización de este trabajo nos apoyamos en la metodología de los modelos de regresión no lineales denominados modelos de regresión logística (*Logit* para abreviar). Esta metodologá es usada para estimar las probabilidades de deserción de los estudiantes. De igual forma, para el análisis estadístico usamos el *software* libre R.

El presente trabajo está organizado de la siguiente manera. En la Sección 2, de manera resumida se describe el modelo de regresión logística, introduciendo las cantidades que juegan un papel fundamental en el análisis de los datos y las interpretaciones de las mismas. En la Sección 3, se definen las variables consideradas en el modelo y el procedimiento a seguir para el análisis estadístico. Los resultados estadísticos obtenidos a partir del modelo de regresión logística, así como también la interpretación y posibles consecuencias, se encuentran en la Sección 4. Finalmente, en la Sección 5 se establecen las conclusiones generales del trabajo.

2. Modelo de regresión logística

La variable dependiente de nuestra investigación, "No continuar", es una variable con resultados dicotómicos: que el alumno continúe y el alumno no continúe de nivel; para modelar este tipo de datos se utilizan los modelos de regresión no lineales, debido que al utilizar modelos de regresión lineal con resultados dicotómicos se violan supuestos de la regresión lineal, como la homocedasticidad y normalidad de los errores [2]. Los modelos de regresión lineal no son funcionales debido a que los modelos de respuesta binaria tienen una relación gráfica en forma de "S" entre las variables independientes y la variable dependiente, mientras que los modelos de regresión no lineal se ajustan mejor a este tipo de datos.

Entre los modelos no lineales más conocidos se encuentran el modelo Probit y el modelo Logit [3], pero será este último el que utilizaremos en nuestra investigación debido a que como menciona Hosmer-Lemeshow [4], en años recientes se ha popularizado el uso del modelo Logit porque es un modelo cuyos resultados

son de fácil interpretación. Su uso se ha extendido en diversos campos como las finanzas [5], educación [6], investigación biomédica [7], etc.; lo cual nos permite conocer más acerca de la metodología existente para el análisis de información.

A continuación presentamos una breve introducción formal del modelo de regresión logística, considerando los aspectos más importantes que se usan en este trabajo. Nos referimos al libro de Hosmer-Lemeshow [4] para un estudio más detallado respecto a este tema.

Sean $x_1, x_2, \ldots, x_k$ una colección de k variables, las cuales serán las variables independientes del modelo. Denotaremos por x al vector $x = (x_1, x_2, ..., x_k)$ de dimensión k. Sea Y una variable aleatoria tipo Bernoulli que cumple: $Y = 1$ si ocurre un éxito y $Y = 0$ si ocurre un fracaso. Entonces $P(Y = 1|x)$ es la probabilidad de éxito dado el conjunto de variables independientes x. El modelo Logit establece que:

$$P(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}, \tag{2.1}$$

donde β_0 es el intercepto y β_k es el coeficiente de regresión de la k-ésima variable, $k = 1, 2, ..., n$.

Para interpretar los resultados que se obtienen al usar el modelo presentado en (2.1), es importante tener clara la relación de las variables independientes con la variable dependiente y definir apropiadamente la unidad de cambio para las variables independientes; una cantidad que involucra esta dinámica de cambio para las variables independientes, es la Razón de Momios (*Odds Ratio*). Para definir esta cantidad es necesario primero definir los Momios (*Odds*):

$$Momios = \frac{P(Y = 1|x)}{1 - P(Y = 1|x)} \ . \tag{2.2}$$

Los momios se interpretan como la razón que existe entre la probabilidad de éxito de un evento y la probabilidad de fracaso del mismo evento. Considerando el modelo de regresión logística (2.1), los momios se pueden escribir como:

$$Momios = \exp\{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\}. \tag{2.3}$$

Por otro lado, la razón de momios nos ayuda a identificar de que forma afecta el aumento de una unidad de la variable x_i en el modelo de regresión logística, conservando fijos los valores de las variables restantes. De manera precisa, si $Momios_2$ se obtiene usando (2.2) al considerar un aumento de una unidad en la variable x_i, conservando fijos los valores de las variables restantes, y $Momios_1$ es obtenida usando (2.2) con las k variables x_i sin cambio alguno, entonces la razón de momios está dada por:

$$Razón\ de\ Momios = \frac{Momios_2}{Momios_1}.$$

Gracias a la ecuación (2.3) obtenida a partir de la especificación del modelo, es posible escribir la razón de momios en términos del parámetro β_i, esto es:

$$Razón\ de\ Momios = exp\{\beta_i\}. \tag{2.4}$$

La expresión anterior nos permite afirmar que, al aumentar en una unidad la variable x_i, conservando fijos los valores de las variables restantes; si $\beta_i > 0$ entonces ocurre un aumento en la razón de la probabilidades de éxito y fracaso de un evento; mientras que si $\beta_i < 0$ ocurre una disminución en esta razón.

En este trabajo, las variables independientes consideradas son de tipo dicotómico con valores 0 ó 1 y representan variables nominales, por tal razón, un incremento en una unidad de la variable x_i, no tiene sentido de interpretabilidad numérica. Considerando esto último, la razón de momios en este trabajo es utilizada para comparar probabilidades de razones de deserción de alumnos bajo diferentes escenarios.

3. Descripción de variables

La variable dependiente que se utiliza en nuestro análisis es la variable "No continuar", las variables independientes son: sexo, ocupación, descuento, turno, reprobar.

Es importante mencionar que en todo este trabajo se hace la distinción entre "Estudiante" y "Alumno". Un estudiante será una categoría de la variable independiente ocupación (la otra categoría será trabajador), mientras que Alumno será toda persona que estudie en el IC.

Para el análisis, todas las variables antes mencionadas se consideran dicotómicas, los valores y su descripción se proporcionan a continuación:

- No Continuar (NC):

  0: El alumno del IC continúa al nivel 1 del IC en el período inmediato a la conclusión del nivel introductorio.

  1: El alumno del IC no continúa al nivel 1 del IC en el período inmediato a la conclusión del nivel introductorio.

- Sexo (SEX):

  0: Femenino.

  1: Masculino.

- Ocupación (OCU):

  0: Si el alumno pertenece a la categoría denominada "Estudiante", es decir, si el alumno está inscrito en secundaria, preparatoria, licenciatura. Incluimos en esta categoría a los alumnos cuyo rango de edad se encuentre entre 16 a 24 años (estudiantes de licenciatura que realizan prácticas profesionales, servicio social o trabajo de medio tiempo).

  1: Si el alumno pertenece a la categoría denominada "Trabajador", que incluye a todos los alumnos que no cumplen con las características de la categoría "0" de la variable ocupación. Esta categoría incluye a alumnos de Maestría, Doctorado, las personas jubiladas y amas de casa.

- Descuento (DES). Si el alumno del IC pertenece a alguna institución de la UADY o son hijos de algún empleado que pertenece a esta institución, el IC ofrece un descuento en la colegiatura de los alumnos de cualquier nivel.

  0: Si el alumno no tiene el descuento.

  1: Si el alumno tiene el descuento.

- Turno (TUR). Las clases del IC contemplan diversos horarios, algunas inician entre las 8:00 a.m y 12:00 p.m. o entre las 3:00 p.m. y 6:00 p.m. Con base en esto, consideramos la variable Turno como sigue:

  0: Si las clases del alumno del IC comienzan en un horario entre las 8:00 a.m. y las 12:00 p.m. Esta categoría la denominamos "Matutino".

  1: Si las clases del alumno del IC comienzan en un horario entre las 3:00 p.m. y las 6:00 p.m. Esta categoría la denominamos "Vespertino".

- Reprobar (REP): En una escala del 0 al 100, se considera al alumno "Aprobado", si su calificación es mayor o igual a 60, por el contrario, si la calificación es inferior a 60 se considera "Reprobado". Es importante señalar que aprobar el nivel Introductorio es un requisito para continuar al nivel siguiente, en este caso el nivel 1.

  0: Si el alumno aprobó el nivel Introductorio.

  1: Si el alumno reprobó el nivel Introductorio.

Lo anterior se puede resumir en la siguiente tabla:

Descripción	**Valor**	**Notación**
No Continuar	0: El estudiante del IC continúa; 1: El estudiante del IC no continúa	NC
Sexo	0: Femenino; 1: Masculino	SEX
Ocupación	0: Estudiante; 1: Trabajador	OCU
Descuento	0: No; 1: Sí	DES
Turno	0: Matutino; 1: Vespertino	TUR
Reprobar	0: Calificación ≥ 60; 1: Calificación < 60	REP

Tabla 1: Descripción de las variables contempladas en el modelo

4. Análisis Estadístico

Como fue mencionado en la introducción, estamos interesados en la probabilidad de no continuar, $P(NC = 1)$. Por la ley de probabilidad total, $P(NC = 1)$ se puede escribir de la manera siguiente:

$$P(NC = 1) = P(NC = 1|REP = 1)P(REP = 1) + P(NC = 1|REP = 0)P(REP = 0). \tag{4.1}$$

Esta ecuación nos permite calcular la probabilidad de no continuar considerando dos escenarios: el primero, que el alumno no continúe por haber reprobado el curso y el segundo, que el alumno no continúe aún habiendo aprobado el curso.

En el reglamento académico del IC se establece que para poder inscribirse al siguiente nivel es necesario aprobar el actual, esto se traduce en $P(NC = 1|REP = 1) = 1$. Usando este hecho, la ecuación (4.1) ahora se escribe como:

$$P(NC = 1) = P(REP = 1) + P(NC = 1|REP = 0)P(REP = 0).$$

Por otro lado, ya que $P(REP = 0) = 1 - P(REP = 1)$, la ecuación anterior se simplifica aún más resultando:

$$P(NC = 1) = P(REP = 1) + P(NC = 1|REP = 0) - P(NC = 1|REP = 0)P(REP = 1). \tag{4.2}$$

De (4.2) se deduce que para estimar la probabilidad de que un alumno no continúe después del nivel Introductorio, $P(NC = 1)$, sólo se requiere estimar $P(REP = 1)$ y $P(NC = 1|REP = 0)$.

Para estimar $P(REP = 1)$ y $P(NC = 1|REP = 0)$ se utilizó la metodología de los modelos de regresión logística. Los resultados obtenidos se detallan a continuación.

4.1. Estimación de $P(REP = 1)$

El análisis estadístico desarrollado en esta sección está basado en una muestra de tamaño 142. El nivel de significancia considerado para todas las pruebas estadísticas es $\alpha = 0.05$. Se divide el procedimiento realizado en 6 fases, las cuales describimos a continuación.

Fase 1: Análisis individual de las variables.

Primero se realizaron pruebas de las variables en conjunto y no presentaron resultados significativos, posteriormente se procedió a analizar las variables individualmente y las distintas combinaciones existentes entre ellas. Se obtuvo que las variables OCU y TUR fueron significativas individualmente. Por lo anterior, se propone un modelo donde se incluyan ambas variables. En la Tabla 2 se exponen los resultados del modelo propuesto y se observa que ambas variables permanecen significativas. Adicionalmente se contempló incluir las variables previamente rechazadas en el modelo propuesto que involucra a las variables OCU y TUR, pero resultaron no significativas en presencia de éstas.

Ocupación + Turno	Estimación	Error Estándar	Valor-Z	P-Valor
Intercepto	-1.6804	0.4449	-3.777	0.00016
OCU	0.9511	0.3789	2.510	0.01208
TUR	0.9830	0.4738	2.075	0.03802

Tabla 2: Análisis de significancia de los parámetros del modelo con OCU y TUR

Fase 2: Interacciones entre variables

Al analizar la interacción entre las variables OCU y TUR, ésta resultó ser significativa. La Tabla 3 contiene los resultados de este análisis. Debido a que la interacción entre las variables OCU y TUR resultó ser significativa, es necesario realizar la prueba de razón de verosimilitudes, para determinar que modelo describe mejor el comportamiento de la variable dependiente; el modelo que incluye la variable OCU*TUR o el modelo que no la incluye.

	Estimación	Error Estándar	Valor-Z	P-Valor
Intercepto	-3.296	1.018	-3.238	0.00120
OCU	4.394	1.305	3.368	0.00076
TUR	2.816	1.048	2.687	0.00720
OCU * TUR	-4.20	1.368	-2.940	0.00329

Tabla 3: Análisis de la significancia de la interacción de las variables OCU y TUR

Fase 3: Prueba de razón de verosimilitudes (*Likelihood Ratio Test*)

Es bien conocido que la prueba de razón de verosimilitudes considera el estadístico:

$$lr = -2ln\frac{L(m_1)}{L(m_2)} = 2[ll(m_2) - ll(m_1)], \tag{4.3}$$

donde $ll(m_2)$ es el valor de la logverosimiltud evaluada en el estimador máximo verosímil considerando el modelo completo con k parámetros y $ll(m_1)$ es el valor de la logverosimiltud evaluada en el estimador máximo verosíimil considerando el modelo reducido con k' parámetros, $k' < k$. El estadístico lr tiene una distribución aproximada ji-cuadrada con $k - k'$ grados de libertad.

En nuestro caso, los modelos completo y reducido se obtienen de (2.1). El modelo completo es considerando las variables $Y = REP$, $x_1 = OCU$, $x_2 = TUR$, $x_3 = OCU * TUR$, en otras palabras, β_i, para $i = 0, 1, 2, 3$ en (2.1) son diferentes de cero. El modelo reducido es obtenido del modelo completo haciendo $\beta_3 = 0$. El modelo completo está compuesto de $k = 4$ parámetros, mientras que el modelo reducido contiene $k = 3$ parámetros.

Desarrollando los cálculos correspondientes se verifica $ll(m_1) = -86.49413$, $ll(m_2) = -80.33364$. Sustituyendo estos valores en la prueba de razón de verosimilitudes, obtenemos lo siguiente:

$$\begin{aligned} lr &= 2[ll(m_2) - ll(m_1)] \\ &= 2[-80.33364 - (-86.49419)] \\ &= 12.32098. \end{aligned}$$

El estadístico lr se distribuye aproximadamente ji-cuadrada con $k - k' = 1$ grado de libertad. De esta forma, el valor p asociado a 12.32098 es $p = 0.0004479 < 0.05$. Por lo tanto, el modelo que incluye las variables OCU, TUR, $OCU * TUR$ se ajusta mejor que el modelo que sólo considera las variables OCU y TUR.

Fase 4: Verificación del ajuste del modelo

A continuación se presenta la tabla para la prueba de bondad de ajuste del modelo [3]:

χ^2	Grados de libertad	P-Valor
27.99679	138	1

Tabla 4: Prueba Ji-Cuadrada

Con base en la Tabla 4 concluimos que no existe evidencia estadística suficiente para rechazar el modelo de regresión logística. Esto significa que el modelo de regresión logística se considera un modelo apropiado para los datos.

Fase 5: Análisis de residuos

Si el modelo de regresión logística ajustado es correcto, entonces $E[Y_i] = \pi(i)$ y asintóticamente se cumpliría que $E[Y_i - \widehat{\pi(i)}] = E[e_i] = 0$. Kutner et al. [3] sugieren que el modelo es correcto si el gráfico del ajuste *lowess* de los residuos ordinarios contra la probabilidad estimada se aproxima a una línea horizontal con intercepto cero, la Figura 1 muestra que en efecto esto ocurre.

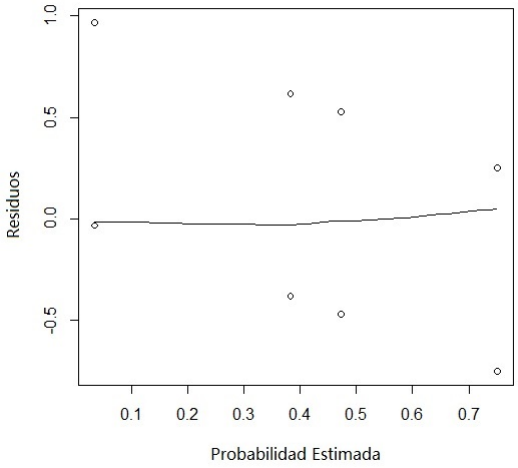

Figura 1: Residuos versus probabilidad estimada

Fase 6: Modelo Final.

Por simplicidad, escribiremos $\widehat{\pi}_1(x)$ para denotar el modelo de regresión logística ajustado que se obtiene considerando las variables $Y = REP$, $x_1 = OCU$, $x_2 = TUR$, $x_3 = OCU * TUR$. De esta forma, con base en los análisis previos, el modelo ajustado de regresión logística es el siguiente:

$$\widehat{\pi}_1(x) = \frac{1}{1 + e^{-(-3.296+4.394*OCU+2.816*TUR-4.020*OCU*TUR)}}. \tag{4.4}$$

Una vez establecido el modelo, podemos realizar estimaciones de $P(REP = 1)$ para los diferentes valores de las variables OCU y TUR. La tabla siguiente es obtenida considerando los diferentes valores de OCU y TUR en (4.4). Escribimos $\widehat{P}(REP = 1)$ para las estimaciones obtenidas.

$\widehat{P}(REP = 1)$	
OCU = 0, TUR = 0	.035708
OCU = 1, TUR = 0	.749885
OCU = 0, TUR = 1	.382252
OCU = 1, TUR = 1	.473524

Tabla 5: Valores $\widehat{P}(REP = 1)$

De los resultados de la Tabla 5 podemos observar que los alumnos más vulnerables a reprobar el curso introductorio son los alumnos que son "Trabajadores" y asisten en el turno "Matutino" (con una probabilidad del 74.98 %), mientras que los alumnos menos propensos a reprobar son los "Estudiantes" del turno "Matutino" (con una probabilidad del 3.57 %).

Con ayuda del modelo ajustado (4.4) y la ecuación (2.4), es posible calcular la razón de momios. Esto se muestra en la tabla siguiente:

Variable	β i	Razón de Momios
Constante	-3.296	0.03700
OCU	4.394	80.96360
TUR	2.816	16.70980
OCU * TUR	-4.020	0.01795

Tabla 6: Razón de Momios de $\widehat{\pi}_1(x)$

Una mejor interpretación de las probabilidades de interés se puede obtener calculando la razón de momios de los submodelos resultantes de (4.4) al considerar $OCU = 0$ (Estudiante), $OCU = 1$ (Trabajador) por separado. Esto se describe a continuación.

4.1.1. Modelo Reprobar en la población de Estudiantes (OCU = 0)

En este submodelo de la ecuación (4.4) consideramos que el alumno es un "Estudiante", es decir, fijamos $OCU = 0$. De esta forma, la ecuación (4.4) se escribe como

$$\widehat{\pi}_2(x) = \frac{1}{1 + e^{-(-3.296+2.816*TUR)}}. \tag{4.5}$$

De (4.5) se obtienen las siguientes razones de momios:

Variable	β i	Razón de Momios
Constante	-3.296	0.03700
TUR	2.816	16.70980

Tabla 7: Razón de Momios de $\widehat{\pi}_2(x)$

Las razones de momios señalan que, en la población "Estudiantes", por cada persona que reprueba en el turno "Matutino", aproximadamente 17 personas reprueban en el turno "Vespertino".

4.1.2. Modelo Reprobar en la población Trabajadores (OCU = 1)

En este caso consideramos que el alumno es un "Trabajador", es decir, el valor de la variable $OCU = 1$, esto se traduce en la siguiente ecuación, obtenida de (4.4),

$$\widehat{\pi}_3(x) = \frac{1}{1 + e^{-(1.098-1.204*TUR)}}. \tag{4.6}$$

De (4.6) se obtienen las siguientes razones de momios:

Variable	β i	Razón de Momios
Constante	1.098	2.9980
TUR	-1.204	0.2999

Tabla 8: Razón de Momios de $\widehat{\pi}_3(x)$

De las razones de momios de la Tabla 8 se puede concluir que, en la población de "Trabajadores", por cada persona que reprueba en el turno "Vespertino", aproximadamente $3(1/0.29)$ personas reprueban en el turno "Matutino".

4.2. Estimación de $P(NC = 1|REP = 0)$

El análisis estadśtico en esta sección está basado en una muestra de tamaño 92. El nivel de significancia considerado para todas las pruebas estadísticas es $\alpha = 0.05$. El procedimiento es similar al de la sección anterior, difiere en algunas fases debido a la simplicidad del modelo ajustado que resulta.

Fase 1: Análisis individual de las variables.

En un análisis previo se realizaron pruebas de las variables en conjunto y no presentaron resultados significativos, posteriormente se procedió a analizar las variables individualmente y las distintas combinaciones existentes entre ellas. La única variable significativa fue la variable TUR. Adicionalmente se analizó incluir a las variables previamente rechazadas en el modelo que incluye a la variable TUR, pero ninguna resultó ser significativa en presencia de ésta.

Fase 2: Verificación del ajuste del modelo

χ^2	Grados de libertad	P-Valor
19.23471	89	1

Tabla 9: Prueba Ji-Cuadrada.

El valor p dado en la Tabla 9 nos permite concluir que no existe evidencia estadística suficiente para rechazar el modelo de regresión logística. Por lo tanto, el modelo de regresión logística se considera un modelo apropiado para los datos.

Fase 3: Análisis de Residuos

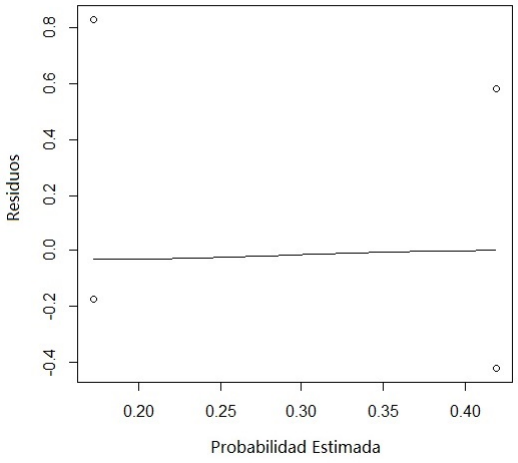

Figura 2: Residuos versus probabilidad estimada

Podemos observar que el gráfico del ajuste *lowess* de los residuos contra las probabilidades estimadas obtenidas a partir del ajuste del modelo, se aproxima a una línea horizontal con intercepto cero, por lo cual se sugiere que el modelo es adecuado.

Fase 4: Modelo Final

El modelo ajustado resultante se presenta a continuación. Por simplicidad escribimos $\widehat{\pi}_4(x)$ para denotar el modelo de regresión logística obtenido para $P(REP = 1|REP = 0)$. Aquí, $x = TUR$ es la única variable significativa.

$$\widehat{\pi}_4(x) = \frac{1}{1 + e^{-(-1.5686+1.2432*TUR)}}. \tag{4.7}$$

A partir de la ecuación (4.7) se pueden obtener las razones de momios:

Variable	β i	Razón de Momios
Constante	-1.5686	0.2083
TUR	1.2432	3.4666

Tabla 10: Razón de Momios de $\widehat{\pi}_4(x)$

De la Tabla 10 se obtiene que para los alumnos aprobados, por cada persona que no continúa en el turno "Matutino", aproximadamente 3 personas no continúan en el turno "Vespertino".

Finalmente, evaluando la ecuación (4.7) con los valores de la variable turno, obtenemos la siguiente tabla de probabilidades:

$\widehat{P}(NC = 1\|REP = 0)$	
TUR = 0	0.172416
TUR = 1	0.419360

Tabla 11: Valores $\widehat{P}(NC = 1|REP = 0)$

Observando los resultados de la Tabla 11 podemos notar que un alumno aprobado en el turno "Vespertino" tiene una mayor probabilidad de no continuar estudiando Chino Mandarín (41.93 %) que un alumno aprobado en el "Matutino" (17.24 %).

4.3. Estimación de $P(NC = 1)$

Usando la fórmula (4.2) con las estimaciones de las probabilidades $P(REP = 1)$ y $P(NC = 1|REP = 0)$ contenidas en las Tablas 5 y 11, respectivamente, obtenemos los valores estimados de la probabilidad de que un alumno no continúe, $P(NC = 1)$:

$\widehat{P}(NC = 1)$	
OCU = 0, TUR = 0	.201967
OCU = 1, TUR = 0	.793008
OCU = 0, TUR = 1	.641310
OCU = 1, TUR = 1	.694293

Tabla 12: Valores $\widehat{P}(NC = 1)$

De la tabla anterior se puede concluir que los alumnos más vulnerables a no continuar son los "Trabajadores" en el turno "Matutino" (con una probabilidad estimada del 79.3 %) y los menos vulnerables son los "Estudiantes" en el turno "Matutino" (con una probabilidad del 20.19 %).

Notamos que en el turno "Vespertino", la diferencia de probabilidades estimadas entre los alumnos "Estudiantes" y "Trabajadores" es de apenas 5 %, mientras que para los alumnos del turno "Matutino" existe una diferencia del 59 % entre los "Estudiantes" y los "Trabajadores".

5. Conclusiones

Del análisis previo se obtiene que la probabilidad estimada de que un alumno no continúe en el Instituto Confucio en el período inmediato a la conclusión del nivel introductorio, es muy baja para los "Estudiantes" en el horario "Matutino" y muy alta para los "Trabajadores" en el horario "Matutino". En general, los resultados sugieren que para un "Trabajador" es difícil continuar, esto es de esperarse debido a que un "Trabajador" tiene un mayor número de responsabilidades en comparación con un "Estudiante".

A través de este estudio se obtuvo que las variables "Ocupación" y "Turno" fueron los factores que tienen mayor influencia en los alumnos para reprobar el curso introductorio y el factor "Turno" tiene mayor influencia para que un alumno no continúe el curso introductorio habiéndolo aprobado. En conclusión, el factor predominante para que un alumno no continúe es el "Turno".

Es importante señalar que una encuesta realizada previo a este artículo, a 50 alumnos del IC que no habían continuado en el IC, señaló que los horarios, que determinan el "Turno" del IC, fueron un factor predominante para dejar de asistir a las clases, debido a que la locación donde se imparten las clases es un lugar donde no se tiene un suficiente número de rutas de autobuses o éstos no cumplen con el horario establecido, lo cual dificultaba que los alumnos asistieran a tiempo a sus clases y esta complicación los orientara a desertar del IC.

Como conclusión general, consideramos que la modificación de los horarios de las clases ayudaría a reducir las probabilidades de no continuar al siguiente nivel y por ende no desertar del IC. Tener horarios nocturnos acordes al término de las jornadas laborales que concluyen entre las 5:00 p.m. y 6:00 p.m. podría facilitar que más alumnos asistan a las clases y logren continuar en el IC.

Referencias

[1] Confucius Institute Headquarters (HANBAN) *2013 Confucius Institute Annual Development Report 2013*.

[2] Long, J. Scott. *Regression Models for Categorical and Limited Dependent Variables.* EUA, Sage Publications Inc., 1997.

[3] Kutner, Michael H; Nachtsheim, Christopher J.; Neter, John, Li, William. *Applied Linear Statistical Models.* EUA, McGraw-Hill Irwin, 2004.

[4] Hosmer, David W.; Lemeshow, Stanley. *Applied Logistic Regression.* EUA, John Wiley & Sons Inc., 2000.

[5] Dutta, Avijan.; Bandopadhyay, Gautam; Sengupta, Suchismita. "Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression" *International Journal of Business and Information*, Vol. 7, Núm. 1 (Junio 2012), p. 105-136.

[6] Singh, Mamta; West, Sandr. "Students Enrolled in Biology Majors Pre-Requisite Courses: Effect of High School Academic Performance" *American Journal of Educational Research*, Vol. 2, Núm 4 (2014), p. 183-188.

[7] Dreiseitl, Stephan; Ohno-Manchado, Lucila; "Logistic regression and artificial neural network classification models: a methodology review" *Journal of Biomedical Informatics*, Vol. 35, Núm 5-6 (Octubre 2002), p. 352-359.