Iturbide, Diego
Pagura, José Alberto
Quaglino, Marta Beatriz

***Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística.*

IDENTIFICACIÓN DE FACTORES QUE INFLUYEN EN EL RENDIMIENTO UTILIZANDO MODELOS PARA DATOS CON CENSURA.

1. INTRODUCCION

El crecimiento de la matrícula de carreras como la de Contador Público, que se cursa en nuestra facultad de Ciencias Económicas y Estadística, no se ve acompañado hasta el momento, por un crecimiento acorde en la cantidad de graduados, ya que el mismo ha permanecido casi constante en los últimos diez años. Este fenómeno podría deberse a altos índices de deserción, o bien a un alargamiento de los tiempos empleados para cursar las currículas. Por la modalidad de exigencias en la universidad, la forma en que los alumnos van cumplimentando las distintas etapas de la carrera es despareja.

Si se plantea un estudio de seguimiento de una cohorte para evaluar por ejemplo el tiempo promedio que tardan en obtener el título, aun cuando este seguimiento se haga por un lapso de tiempo muy superior al establecido por el plan de estudios para obtenerlo, quedará un grupo de alumnos, que en la práctica es muy importante, para los cuales solo se sabrá que hasta el momento de observación no se han recibido, pero eso no implica que no lo harán en un tiempo más prolongado.

En el presenta trabajo se utilizan métodos estadísticos que permiten considerar datos de esta naturaleza para calcular los tiempos medios que emplean los alumnos para cubrir una particular etapa académica. Además se estudia si ese tiempo medio es distinto para algunos grupos definidos por sexo, nivel educacional de los padres, condiciones de ingreso, etc. y se estima el grado de influencia de factores socioeconómicos sobre la probabilidad instantánea de cumplir dicha etapa.

2. MATERIAL Y MÉTODOS

En este trabajo se consideran aquellos alumnos que ingresaron a la carrera de Contador Público Nacional en el año 1995 y que hubieran rendido al menos una materia al 31/03/1998. Se estudia esta cohorte porque es la primera para la cual se puso en vigencia la actual reglamentación sobre el ingreso y por estudios anteriores se sabe que los resultados de los exámenes de ingreso tienen una fuerte relación con el rendimiento posterior. Se eliminaron aquellos inscriptos que:

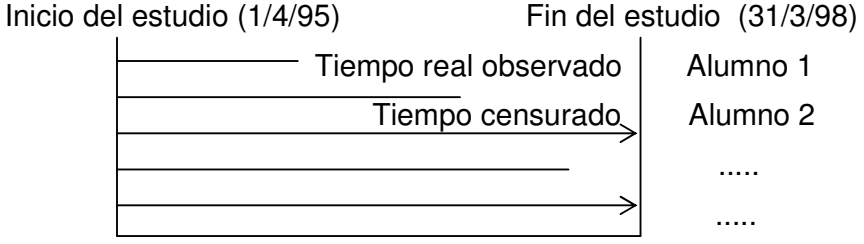
- Realizaron el trámite de ingreso, pero no aprobaron ninguna materia durante tres ciclos lectivos, considerando que su situación era más cercana a la deserción.
- Aprobaron alguna de las materias requeridas por equivalencias puesto que cursaban otra carrera. Por lo general quienes estaban en esta condición eran alumnos avanzados de Licenciatura en Administración de Empresas, que se inscribieron en la carrera de Contador.

Se realizó un seguimiento hasta el 31/03/1998, es decir de 36 meses, estudiando si habían terminado una etapa de la currícula: "aprobar todas las materias de primer año y tres fundamentales de segundo, Sistema de Información Contable II, Matemática II y Administración General".

La información necesaria para realizar el estudio debió recogerse de diferentes fuentes. El departamento alumnado de la facultad proporcionó bases de datos con la situación de cada alumno respecto a las materias rendidas y con los aspectos demográficos, socioeconómicos y de antecedentes educacionales de los alumnos y sus padres, que corresponden al formulario de inscripción SUR I. Por otro lado, se tuvieron en cuenta los resultados de los exámenes de ingreso, que fueron obtenidos de la base de datos elaborada para el desarrollo de la tesina "Un estudio sobre la problemática del ingreso y su vinculación con el rendimiento" realizada por Adriana Solimano.

Los métodos utilizados corresponden al denominado "análisis de supervivencia" desarrollado para describir el comportamiento de datos que corresponden al tiempo o duración desde un origen bien definido, hasta la ocurrencia de algún "evento" o un "punto final". Frecuentemente este evento es la muerte o la recidiva de una enfermedad o la falla de alguna componente o equipo, de allí el nombre de "supervivencia". En este trabajo se estudia el "tiempo en meses que un individuo tarda hasta aprobar nueve materias del primer tramo del plan de la carrera de Contador", por lo tanto el evento de interés es favorable: terminar en tiempos curriculares una etapa académica. La metodología que se utiliza podría extenderse en forma similar para el estudio del cumplimiento de otro evento, como por ejemplo la obtención del título universitario.

Los datos de supervivencia no son flexibles a los procedimientos estadísticos estándares, esto sucede por dos razones: una es que en general los datos de supervivencia no están distribuidos simétricamente y la otra es que éstos están frecuentemente censurados. Un dato se dice que es *censurado* cuando el punto final o el evento de interés aún no ha sido observado para el individuo hasta el momento de finalizar el estudio. Esta situación podrían representarse gráficamente como sigue, identificando cada línea los tiempos que diferentes alumnos emplearon hasta aprobar estas nueve materias:



En el análisis de los datos de tiempo de supervivencia hay dos funciones que juegan un rol central, estas son la *función de supervivencia y la función de riesgo.*

Sea T una variable aleatoria, no negativa, que representa el tiempo que un alumno tarda en aprobar las nueve materias definidas. Este evento es favorable, cuanto más corto sea indicará mejor rendimiento. La *función de supervivencia*, se define como la probabilidad de que este tiempo sea mayor que t, es decir:

$$S(t) = P(T>t) = 1-F(t)$$

donde F(t) representa a la función de distribución o función de probabilidad acumulada de T. En términos del problema académico, S(t) representa la probabilidad de que un alumno tarde más de un tiempo t en aprobar las nueve materias definidas, o bien la fracción de alumnos en la población que al tiempo t, aún no han terminado la etapa. S(t) es una función no decreciente y a partir de ella pueden determinarse los percentiles de la distribución.

La *función de riesgo* se define como la probabilidad de que un individuo cumpla el evento al tiempo t, condicionado a que el mismo no lo haya cumplido hasta ese tiempo, es decir, que la función de riesgo representa el riesgo instantáneo de cumplir el evento en un determinado momento, para un individuo que todavía no lo había cumplido. Nuevamente se aclara que en este trabajo el "riesgo" que corre un individuo es favorable. Una definición formal de la función de riesgo, es considerar la probabilidad de que la variable aleatoria T, se encuentre entre t y t+δt, condicionado a que T es mayor o igual a t, es decir:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{P(t \le T < \delta t) / T \ge t]}{\delta t} \right] = \frac{f(t)}{S(t)}$$

siendo f(t) la función de densidad de T.

Las funciones de supervivencia y de riesgo son estimadas por los tiempos de supervivencia observados. Estas estimaciones pueden hacerse a través de métodos no paramétricos, semi paramétricos y/o paramétricos, es decir, suponiendo o no modelos para la distribución de probabilidad de la variable aleatoria T.

Para estudiar la influencia de factores sociales sobre el cumplimiento de este evento, se utiliza el modelo de regresión de Cox que supone que la relación entre las variables explicativas X_1, X_2.... X_p y la función de riesgo de un individuo i es:

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ...+\beta_p x_{pi}) \, h_o(t),$$

siendo $h_o(t)$ la función de riesgo de un individuo para el cual las variables explicativas valen cero. Este modelo también puede ser expresado de la forma:

$$\log \left\{ \frac{h_i(t)}{h_o(t)} \right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + ...+\beta_p x_{pi}$$

por lo tanto el modelo de Cox o de riesgo proporcional, puede considerarse como un modelo lineal para el logaritmo de la razón de riesgo. De este modo cada parámetro β_i puede interpretarse en términos del cambio que produce cada variable explicativa sobre la razón de riesgo.

3. RESULTADOS

El total de alumnos inscriptos a la carrera de Contador Público Nacional en 1995 fue de 2530, sólo para 2065 de este grupo se cuenta con los resultados de los exámenes de ingreso, por lo tanto en principio se eliminaron 465 alumnos del estudio. A éstos se agregaron 830 (más del 40%) que no habían aprobado ninguna materia durante tres ciclos lectivos, quedando un total de 1235.

Sobre estos 1235 alumnos se realizó un seguimiento de 36 meses, identificando para cada uno de ellos si en dicho lapso habían aprobado o no las nueve materias definidas, encontrándose que el 61,7% (762 alumnos) no habían cumplido ese requisito durante ese tiempo. Para el 38,3% (473 alumnos) restante se calculó el tiempo exacto que habían utilizado hasta aprobar las nueve materias definidas

Utilizando toda esta información, con los tiempos exactos o censurados para cada alumno, se estimaron mediante métodos no paramétricos las funciones de supervivencia considerando subgrupos de alumnos definidos por ciertas características socio-culturales. La tabla siguiente presenta un resumen de los resultados, que incluyen un test de hipótesis para probar si las funciones son estadísticamente diferentes entre los subgrupos y el cálculo del percentil 75 de la distribución.

Tabla 1: Comparación de las funciones de supervivencia

Variable	Frecuencias	% de eventos	Log rank test	Funciones de Superv.	Percentil del 75% (en meses)
Sexo			χ^2=16,01 p=0,0001		
Masculino	513	32,2	las distribuciones difieren	No se cruzan	Q_M = 31
Femenino	722	42,7			Q_F = 24
Domic.proceden.			χ^2=3,19 p=0,0742	Se cruzan	
Rosario	708	40,4	las distribuciones difieren	levemente al	Q_{Ro} = 24
Resto	527	35,5		principio	Q_{Re} = 28
Estado Civil			χ^2=4,56 p=0,0328		
Soltero	1209	38,2	las distribuciones difieren	Se cruzan	Q_S = 25
Casado	26	15,4			Q_C = no $\exists$
Residencia			χ^2=0,46 p=0,4961		
Con familia	1016	37,9	las distribuciones **no** difieren	No se cruzan	Q_F = 25
Independ.	219	40,2			Q_I = 24
Hs. trabajo sem.			χ^2=15,48 p=0,0014		
Hasta 20	89	49,4		Totalmente	Q = 23
De 21 a 35	148	40,5	las distribuciones difieren	cruzadas	Q = 24
36 o mas	142	24,6			Q = no $\exists$
no trabaja	856	39,0			Q = 24
Trabaja			χ^2=0,84 p=0,3602	Se cruzan	
Si	379	36,7	las distribuciones **no** difieren	levemente al	Q_{Si} = 26
No	856	39,0		principio	Q_{No} = 24

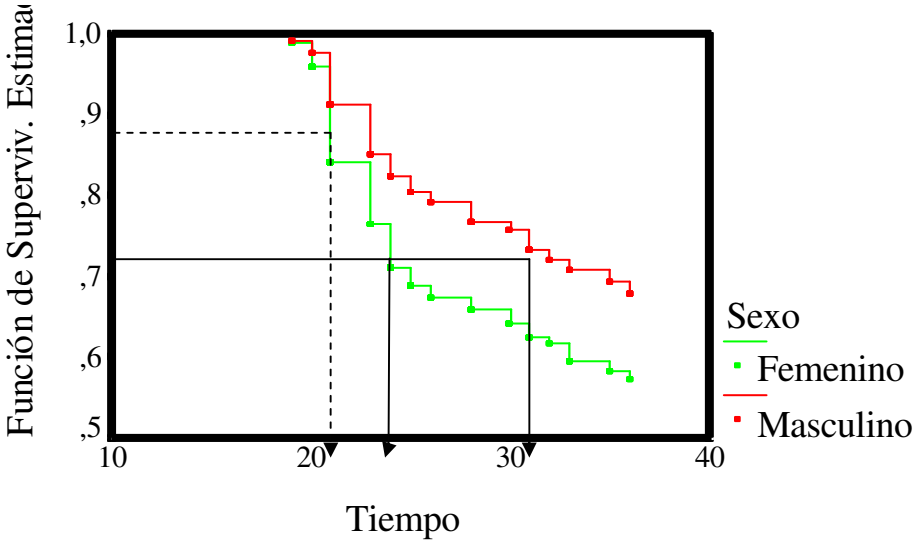
Clase colegio 2º			χ^2=39,97 p=0,0000		
Nacional	301	30,2		Se cruzan	Q_N = 28
Prov y munic	298	34,2	las distribuciones difieren		Q_{PM} = 31
Privado	518	40,5			Q_P = 24
Otro	118	59,3			Q_O = 21
Clase colegio 2º			χ^2=1,89 p=0,1689	Se cruzan	
Oficial	717	36,7	las distribuciones **no** difie-	levemente al	Q_O = 25
Privado	518	40,5	ren	principio	Q_P = 24
Título 2º			χ^2=6,87 p=0,0088	Se cruzan	
Comercial	923	40,5	las distribuciones difieren	levemente al	Q_C = 24
Otro	312	31,7		principio	Q_O = 28
Orient.vocacion.			χ^2=4,49 p=0,0341		
Si recibió	285	43,9	las distribuciones difieren	No se cruzan	Q_S = 24
No recibió	950	36,6			Q_N = 25
Educación padre			χ^2=46,61 p=0,0000		
Bajo	55	25,5	las distribuciones difieren	Se cruzan	Q_B = 36
Medio bajo	550	31,6		levemente al	Q_{MB} = 30
Medio alto	422	40,3		principio	Q_{MA} = 24
Alto	208	55,3			Q_A = 23
Educaciónmadre			χ^2=34,66 p=0,0000		
Bajo	62	25,8	las distribuciones difieren	Se cruzan	Q_B = 35
Medio bajo	481	30,8		levemente al	Q_{MB} = 31
Medio alto	458	41,3		principio	Q_{MA} = 24
Alto	234	51,3			Q_A = 23
Cat.ocup. padre			χ^2=12,26 p=0,0065		
Patrón	361	42,4			Q_P = 24
Cuenta prop.	344	40,1	las distribuciones difieren	Se cruzan	Q_{CP} = 24
Obrero, emp.	467	32,5		levemente al	Q_{OE} = 30
Otro	63	47,6		principio	Q_O = 23
Ingr.contabilidad			χ^2=66,59 p=0,0000		
Aprobaron	765	46,8		Se cruzan	Q_{Ap} = 23
No aprob.	307	21,2	las distribuciones difieren	levemente al	Q_{NA}= no $\exists$
Ausentes	163	30,7		principio	Q_{Au} = 32
Ingr. matemática			χ^2=200,42 p=0,0000		
Aprobaron	480	60,6		Se cruzan	Q_{Ap} = 23
No aprob.	538	23,6	las distribuciones difieren	levemente al	Q_{NA}= no $\exists$
Ausentes	217	25,3		principio	Q_{Au} = 36
Egreso secund.			χ^2=34,61 p=0,0000	Se cruzan	
En 1994	1063	41,8		levemente al	Q = 24
Antes del ´94	172	16,9	las distribuciones difieren	principio	Q = no $\exists$

Estas funciones de supervivencia o funciones de probabilidad antiacumuladas, pueden representarse gráficamente. En el gráfico hay que observar que la curva que se encuentra por debajo de todas, pertenece a la categoría de la variable cuyos individuos cumplen el evento en menor tiempo. Se presentan dos casos para ilustrar los resultados.

La Figura 1 muestra las funciones de supervivencia estimadas por sexo. Para los alumnos de sexo masculino la curva es siempre mayor que la perteneciente al sexo

femenino, es decir, la probabilidad estimada de no cumplir con el evento desde el origen hasta después de algún tiempo t, es mayor para los varones que para las mujeres. Es más probable que los varones aprueben las “primeras” nueve materias en tiempos posteriores a las mujeres.
Otra forma de ver esta característica es a través del percentill 75, señalado en el gráfico con un trazo lleno. El 25% de las mujeres aprueban las materias en 24 meses o menos, mientras que el 25% de los varones lo hacen en 31 meses o menos.
El trazo punteado identifica el percentil 10: solo el 10% tanto de varones como de mujeres tardan 21 meses en aprobar las 9 materias, es decir que la probabilidad de que el tiempo sea superior a 21 meses es del 90%.

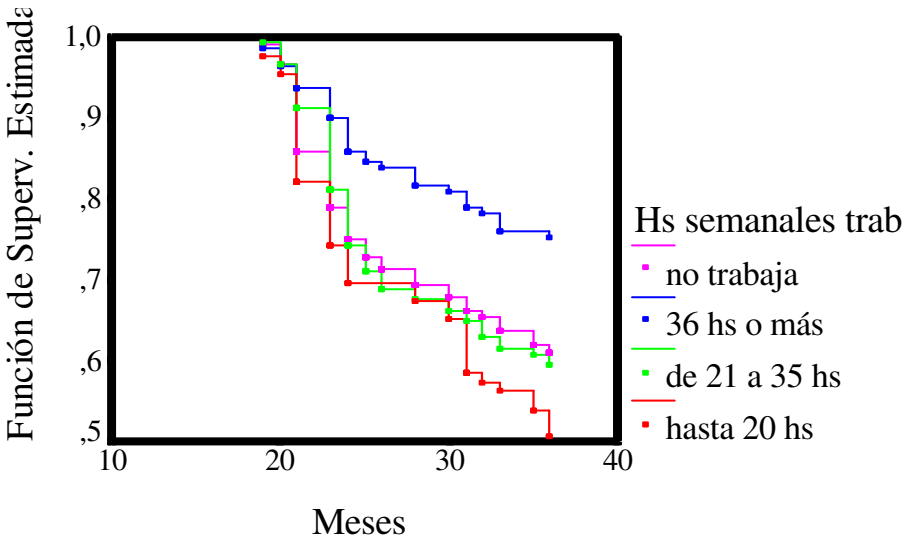
Figura Nº1: Función de supervivencia estimada , según sexo.



Dado que S(t) es mayor que 0,5 para todos los valores de t, la mediana no puede estimarse. Esto significa que más de la mitad del grupo observado no cumplió el evento estudiado en el lapso de observación de 36 meses.

La Figura 2 muestra las funciones de supervivencia para grupos de alumnos con diferente cantidad de *Horas semanales de trabajo.* Los individuos pertenecientes al grupo que trabajan más de 36 horas presentan tiempos mucho más prolongados para cumplir la etapa. Esto también puede apreciarse en los cuartiles puesto que prácticamente, no existen diferencias entre los alumnos que pertenecen a los grupos “no trabajan”, “trabajan 20 horas o menos” y “trabajan entre 21 y 35 horas”.

Figura Nº2: Función de supervivencia estimada según horas semanales de trabajo.



Para completar el trabajo, se ajusta un modelo de regresión de Cox considerando en primer lugar un conjunto de variables explicativas de a una por vez para detectar su posible asociación con la función de riesgo. Las variables que mostraron significación estadística **individualmente** son:

- *Sexo (SEXO)*
- *Título obtenido en el colegio secundario (TIT2º)*
- *Orientación vocacional (OV)*
- *Nivel educacional del padre (EPAD)*
- *Nivel educacional de la madre (EMAD)*
- *Categoría ocupacional del padre (COPAD)*
- *Condición de ingreso en contabilidad (CIC)*
- *Condición de ingreso en matemática (CIM)*
- *Año de egreso del colegio secundario (EG2º)*

En segundo lugar se consideran estas variables en forma conjunta y se aplica un procedimiento de selección paso a paso para seleccionar tanto los efectos principales como las posibles interacciones. Ninguna interacción resulta estadísticamente significativa (p=0,10) por lo que el modelo queda solamente con los efectos principales. La hipótesis de bondad del ajuste de este modelo no se rechaza. (test de hipótesis de nulidad de todos los coeficientes, G=269,031, p=0,0000).

La siguiente tabla presenta la estimación de los parámetros del modelo, su error estándar, la estadística de Wald para probar la hipótesis de que el parámetro es cero en la población y la probabilidad asociada a la estadística bajo la hipótesis nula.

Tabla 2: **Estimaciones Máximo Verosímiles de los parámetros del modelo de Cox.**

Variable	Parámetro estimado	Error estándar	Wald	p = Pr(W>χ^2)
CIM(1)	1,0675 = $\overline{\beta}_1$	0,1102	93,9102	0,0000
CIM(2)	0,0200 = $\overline{\beta}_2$	0,2040	0,0096	0,9219
CIC(1)	0,6957 = $\overline{\beta}_3$	0,1372	25,7005	0,0000
CIC(2)	0,8112 = $\overline{\beta}_4$	0,2267	12,8068	0,0003
EPAD(1)	-0,6207 = $\overline{\beta}_5$	0,3092	4,0294	0,0447
EPAD(2)	-0,5085 = $\overline{\beta}_6$	0,1367	13,8486	0,0002
EPAD(3)	-0,3674 = $\overline{\beta}_7$	0,1269	8,3833	0,0038
EG2º	0,8489 = $\overline{\beta}_8$	0,1970	18,5597	0,0000
SEXO	-0,1944 = $\overline{\beta}_9$	0,0987	3,8829	0,0488
EMAD(1)	-0,2925 = $\overline{\beta}_{10}$	0,2907	1,0124	0,3143
EMAD(2)	-0,3388 = $\overline{\beta}_{11}$	0,1390	5,9450	0,0148
EMAD(3)	-0,2323 = $\overline{\beta}_{12}$	0,1220	3,6255	0,0569

El modelo estimado de regresión de Cox se expresa como sigue:

$$\overline{\mathbf{h}}_\mathbf{i}\mathbf{(t)} = \exp [1{,}076\ \mathbf{CIM}(1) + 0{,}02\ \mathbf{CIM}(2) + 0{,}696\ \mathbf{CIC}(1) + 0{,}811\ \mathbf{CIC}(2) - 0{,}621\ \mathbf{EPAD}(1) - 0{,}509\ \mathbf{EPAD}(2) - 0{,}367\ \mathbf{EPAD}(3) + 0{,}849\ \mathbf{EG2º} - 0{,}194\mathbf{SEXO} - 0{,}292\mathbf{EMAD}(1) - 0{,}339\mathbf{EMAD}(2) - 0{,}232\mathbf{EMAD}(3)]\ \mathbf{h_o(t)}$$

o bien el logaritmo de la razón de riesgo:

$$\mathbf{log}\left\{\frac{\mathbf{h_i(t)}}{\mathbf{h_o(t)}}\right\} = 1{,}076\mathbf{CIM}(1) + 0{,}02\mathbf{CIM}(2) + 0{,}696\mathbf{CIC}(1) + 0{,}811\ \mathbf{CIC}(2) - 0{,}621\mathbf{EPAD}(1) - 0{,}509\mathbf{EPAD}(2) - 0{,}367\mathbf{EPAD}(3) + 0{,}849\mathbf{EG2º} - 0{,}194\mathbf{SEXO} - 0{,}292\mathbf{EMAD}(1) - 0{,}339\mathbf{EMAD}(2) - 0{,}232\mathbf{EMAD}(3)$$

La interpretación de los parámetros, depende de la codificación hecha para las modalidades de cada variable explicativa y son válidas considerando un valor controlado o constante de las demás variables involucradas en el modelo. Por la forma funcional del modelo la interpretación más sencilla es a través de la influencia de exp(β_i) sobre la razón de riesgo.

- Para "*Condición de examen de Ingreso en Matemática aprobado*" (CIM(1)) la razón de riesgo estimada es 2,9081 (2,9081≈ exp 1,076) lo que significa que la posibilidad de que se presente el evento aprobar las primeras nueve materias en un tiempo dado, es casi tres veces mayor para un alumno que "*Aprobó*" el examen de ingreso en matemática que para uno que "*No aprobó*".

- Para "*Condición de examen de Ingreso en Matemática ausente*" (CIM(2)) la razón de riesgo estimada es 1,0202 lo que nos dice que la posibilidad de aprobar las nueve primeras materias en un tiempo dado, para un alumno que estuvo "*Ausente*" en el examen de ingreso en matemática es similar que para aquel que "*No aprobó*" el examen, controlando las demás variables.

- Para un alumno que "*Aprobó*" el examen de ingreso en contabilidad, la chance de rendir bien las primeras nueve materias en un tiempo t, es dos veces mayor que para aquel que "*No aprobó*" este exámen.

- La situación de aquellos individuos que estuvieron "*Ausente*" en el examen de ingreso en contabilidad es que la chance de rendir las primeras nueve materias en un tiempo dado, es 2,25 veces mayor que para los que "*No aprobaron*" el examen. Este resultado, si bien no parece tan claro, estaría indicando que la condición académica dada por la no aprobación del examen de contabilidad es peor aún que la de quienes estuvieron ausentes, evidencia una falta de preparación con la que podrían contar quienes estuvieron ausentes.

- Para la variable "*Nivel de Educación del Padre*" se puede decir que la chance de rendir las primeras nueve materias en un tiempo t, es menor para un individuo cuyo padre tiene un nivel de educación "*Bajo*" respecto a uno cuyo padre tiene un nivel de educación "*Alto*". Interpretando en forma inversa tenemos que para un individuo cuyo padre tiene un "*Nivel de Educación Alto*" la chance de aprobar las primeras nueve materias en un tiempo dado es:

Un 86% mayor (1/0,5375=1,86) que para aquel individuo cuyo padre tiene un "*Nivel de Educación Bajo*"

Un 66% mayor (1/0,6014=1,66) que para aquel individuo cuyo padre tiene un "*Nivel de Educación Medio Bajo*"

Un 44% mayor (1/0,6925=1,44) que para aquel individuo cuyo padre tiene un "*Nivel de Educación Medio Alto*"

- Para "*Año de Egreso del Secundario*" la razón de riesgo es 2,3371 lo que significa que, un alumno que ingresa a la facultad inmediatamente después de terminar el colegio secundario tiene mayores posibilidades de aprobar las primeras nueve materias en un tiempo t, respecto a uno que ingresa a la facultad habiendo terminado el secundario hace más de un año.

- Para "*Sexo*" tenemos que las mujeres tienen un 21% más de chance que los hombres, de aprobar las primeras nueve materias en un tiempo dado.

- En el caso de la variable "*Nivel de Educación de la Madre*" se puede decir que la chance de rendir bien las primeras nueve materias en un tiempo dado es un 40% mayor para un individuo cuya madre tiene un nivel "*Alto*" respecto a uno cuya madre tenga un nivel "*Medio bajo*".

Los otros dos niveles no se interpretan puesto que los coeficientes estimados no son estadísticamente significativos.

4. CONSIDERACIONES FINALES

Los métodos de análisis de tiempos de supervivencia, extensamente utilizados en las áreas médica y de estudios de fiabilidad industrial, han permitido abordar un tema del área educativa, teniendo en cuenta la información brindada por el conjunto total de alumnos de una cohorte, tanto los que habían aprobado el evento de interés estudiado, como aquellos que hasta el tiempo final de observación no habían

aprobado todas la materias especificadas. Haber excluído a este conjunto hubiera significado no sólo una importante pérdida de información, sino dar una subestimación del tiempo real que se tarda en cumplir una etapa de la currícula. Este tiempo real medio estimado, es muy superior al establecido por el plan de estudios.

Por otra parte, un modelo de regresión como el riesgos proporcionales, permitió determinar, sin el supuesto de ninguna distribución de probabilidad especial sobre los tiempos empleados, la forma en que influyen factores socio culturales de interés, sobre la probabilidad instantánea de cumplir esta etapa, encontrando como características que favorecen al rendimiento a la aprobación de los exámenes de ingreso de contabilidad y matemática, el alto nivel educacional de los padres y el comenzar la carrera universitaria inmediatamente después del nivel secundario.

Los resultados obtenidos brindan nuevos aspectos sobre una problemática de interés en nuestro ámbito educativo y el uso de los métodos empleados podría extenderse al estudio de nuevos temas.

5. BIBILIOGRAFÍA

LAWLEES,J.F.(1982), *Statistical Models and Methods for Lifetime Data*, Ed. Wiley.

LEE, E.T. (1980), *Statistical Methods for Survival Data Analysis*, Wadsworth, CA.

MEEKER, W.Q. and ESCOBAR, L.A., (1998) *Statistical Methods for Reliability Data*, Ed. Wiley.