Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Estimación de la probabilidad de egreso de estudiantes de licenciatura en ciencias de la BUAP usando Regresión Logística

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
LICENCIADA EN MATEMÁTICAS APLICADAS

PRESENTA:
ANA LUISA NIETO MÉNDEZ

DIRECTOR DE TESIS:
DRA. HORTENSIA REYES CERVANTES
DR. FLAVIANO GODÍNEZ JAIMES

Febrero 2015

Agradecimientos

A mi mamá, la Sra. Mónica Méndez pues sin su apoyo no hubiera sido posible realizar esta licenciatura.
A todos aquellos que han sido mis profesores en la licenciatura, de cada uno he aprendido algo interesante.
A la Dra. Hortensia Reyes Cervantes por invitarme al proyecto y su ayuda para realizar ésta tesis, y al Dr. Flaviano Godínez Jaimes por su paciencia para ayudarme.
A la Dra. Gladys Linares Fleites, por se parte del jurado y por su apoyo en el servicio social, al Dr. Hugo Adán Cruz Suárez y al Dr. Fernando Velasco Luna por revisar ésta tesis.
A mis amigos y compañeros de la facultad, en especial a Oscar Rojas, Silvia Blanca Pérez, Verónica Ramírez, José Luis Prado, Marisol Mares, Leonardo Remedios, David Morante, Miguel Ángel Macías, Rafael Fuerte, etc.

A la Vicerrectoría de Investigación y Estudios de Postgrado de la Benemérita Universidad Autónoma de Puebla, por el apoyo económico que permitió la conclusión e impresión de este trabajo mediante el proyecto **Modelación Estadística sobre la eficiencia terminal en dos licenciaturas de la FCFM de la buap-2.**
Y a la Secretaría de Educación Pública por el apoyo económico otorgado.

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias Físico Matemáticas

Estimación de la probabilidad de egreso de estudiantes de licenciatura en ciencias de la BUAP usando Regresión Logística

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
LICENCIADA EN MATEMÁTICAS APLICADAS

PRESENTA:
ANA LUISA NIETO MÉNDEZ

DIRECTOR DE TESIS:
DRA. HORTENSIA REYES CERVANTES
DR. FLAVIANO GODÍNEZ JAIMES

Febrero 2015

Índice

Capítulo 1

Introducción

El inicio de este siglo se ha caracterizado por la necesidad de mejoramiento, de transparencia, de incidencia y trascendencia. Los distintos países, los diversos grupos de liderazgo mundial, en diferentes ámbitos refieren con argumentos peculiares la necesidad de mejorar la calidad de vida, que miden con indicadores económicos relacionados con la educación, la alimentación, la salud, el ambiente, los energéticos; a los que se puede agregar, por su carácter generalizado, la información, los recursos naturales, y el desarrollo tecnológico, entre otros [11].

La enseñanza en general y, en partícular, la relacionada con las ciencias abstractas son de mayor importancia para la sociedad contemporánea. Con el paso del tiempo, las sociedades han conformado instituciones, con la finalidad de articular el saber científico y matemático con la cultura de la sociedad, buscando propiciar en la población una visión científica del mundo. A lo largo de la historia las sociedades han buscado aumentar la cantidad de productos obtenidos por unidad de trabajo invertido. Algunos investigadores dicen que si el incremento de la producción no reduce la calidad, entonces tenemos la noción de **eficiencia**. Su aplicación al campo de la educación superior es directa, ya que la principal función de una Institución de Educación Superior (IES) es la docencia y, por tanto, su eficiencia depende principalmente de la proporción de alumnos que logran egresar o titularse, respecto a aquellos que ingresaron. A este indicador se le ha llamado **eficiencia terminal** y constituye el concepto central de este trabajo [6].

Durante los últimos años, en las IES aumentó el interés por mejorar la calidad del aprendizaje, coadyuvar a disminuir el rezago estudiantil y elevar los índices de eficiencia terminal. Respecto a este último índice en México, de acuerdo a los datos

proporcionados por la Secretaría de Educación Pública durante un encuentro de México-España, en Julio de 2008, sólo 6 de cada 100 niños que ingresan a educación básica pueden llegar a la universidad y, finalmente se mencionó que la eficiencia terminal en México era de aproximadamente del 60%. Durante su Quinto Informe de Gobierno en el año 2011, el ExPresidente de la Nación, Felipe Calderón, mencionó en relación al área educativa, que en los últimos años se elevó la cobertura total de la educación superior (escolarizada y no escolarizada) del 25.2% en la población de 19 a 23 años durante el ciclo 2005-2006, al 30.9% en el ciclo escolar 2010-2011, alcanzando ya la meta propuesta para 2012 de 30%, establecida por el Plan Nacional de Desarrollo [27].

Los métodos de estadística matemática brindan la posibilidad de presentar el conjunto de resultados de la observación en una forma compacta y adecuada para su examen. Ellos permiten separar del conjunto de observaciones la información importante presentándola en forma de un pequeño número de índices de resumen. Si resulta que los datos existentes son insuficientes para comprender la esencia del fenómeno y se requiere llevar a cabo un experimento adicional, los métodos de estadística matemática permiten responder a la pregunta de cómo efectuar dicho experimento para simplificar en grado máximo el trabajo del investigador tanto en la realización del experimento como en la elaboración de los datos experimentales [7].

Los modelos de regresión logística son una herramienta que permite explicar el comportamiento de una variable con respuesta binaria mediante una o varias variables independientes explicativas de naturaleza cuantitativa y/o cualitativa. Los modelos de respuesta discreta son un caso particular de los modelos lineales generalizados formulados por Nelder y Wedderburn en 1972, al igual que los modelos de regresión lineal o el análisis de varianza [10].

Este modelo, se utiliza cada vez más en ciencias sociales en tanto que posibilita el análisis de relaciones entre variables no métricas que predominan en nuestras disciplinas (esto es, las que normalmente se identifican como variables nominales u ordinales) [4].

1.1 Marco Conceptual y Teórico

El comienzo del siglo XXI se ha caracterizado por significativos avances científicos y tecnológicos, pero a la vez por la agudización de conflictos sociales a escala global,

que han traído por consecuencia la aparición de eventos tales como la globalización y la polarización. Éstos fenómenos han provocado cambios mundiales en dónde los países de mayor adelanto científico, tecnológico y cultural, se han organizado en bloques que favorecen su propio desarrollo. Sin embargo, los países latinoamericanos, entre ellos México, se encuentran en una posición desventajosa, dada su debilidad científica y tecnológica lo que limita su ingreso a la economía de este mundo globalizado. Por lo tanto la Universidad tiene un papel muy importante en el cambio de esta situación al promover la investigación y la formación de recursos humanos competentes. Pero para lograr este objetivo, un paso decisivo será elevar la retención estudiantil en las universidades, ya que a nivel nacional y local la deserción continúa siendo elevada [8].

Considerar la complejidad de la problemática implica abordar una serie de factores que se articulan en curriculares, académicos, psicológicos, sociales, ambientales, económicos, administrativos, institucionales, referente a la normatividad, geográficos; en este sentido las investigaciones en esta línea adquieren importancia ya que muestran a la luz su incidencia en la obtención del grado académico además de buscar el desarrollo de lineas de acción fundamentadas que ofrezcan alternativas de solución a esta problemática [24].

1.1.1 Funciones de la Universidad

Son tres los principales procesos que se dan en la Universidad dada la función que desempeña: el docente - educativo, el investigativo y el de extensión. Mediante la docencia, se prepara al hombre para su labor profesional. Por medio de la investigación se descubren nuevos conocimientos científicos, para resolver problemas sociales usando la ciencia como instrumento, así se desarrolla la cultura. Y finalmente el proceso mediante el cual la Universidad promociona a la sociedad la cultura que ha ido acumulando y también la que puede recibir de la sociedad, se llama extensión [8].

La misión clave de la educación superior además de las tradicionales de educar, es fomentar la investigación, así como contribuir al desarrollo sostenible y al mejoramiento de toda la sociedad a través de:

1. La formación de egresados altamente calificados que sean ciudadanos críticos, participativos y responsables.

2. La formación de espacios en la educación superior que propicien la educación continua.

3. La generación y difusión de conocimientos logrados a través de la investigación científica, promoviéndola, para que los alumnos puedan aplicar su conocimiento en las ciencias sociales, de humanidades y arte, si llegara el caso.

4. La contribución de la educación terciaria para fomentar y difundir las culturas regionales, internacionales e históricas.

5. La protección de los valores sociales.

6. El mejoramiento de la educación, mediante la capacitación del personal docente (UNESCO, 1998) [8].

1.1.2 La Universidad en Puebla

La historia universitaria en Puebla se remonta a más de cuatro siglos, cuya imagen es de transformación permanente, y se ha colocado al ritmo de los tiempos, siempre vinculada a la ciencia, la cultura, y se ha ligado a los intereses del pueblo mexicano. La Benemérita Universidad Autónoma de Puebla (BUAP) es una institución académica que ha experimentado procesos de transformación profunda en las tres últimas décadas del siglo pasado, lo que le ha permitido arribar al siglo XXI como una de las más importantes instituciones del país. En la actualidad, se ha logrado consolidar un proyecto de desarrollo y mejoramiento permanente que le ha dado el reconocimiento de diversos sectores de la sociedad poblana. Hoy en día, las necesidades de la universidad se transforman en una mejor calidad académica del personal docente, una interacción más cercana con el estudiante, y mayor calidad y eficacia del personal administrativo, Gaceta Universitaria [9].

1.1.3 Facultad de Ciencias Físico Matemáticas

En la Facultad de Ciencias Físico Matemáticas (FCFM) de la Benemérita Universidad Autónoma de Puebla (BUAP) existen algunos estudios acerca de los alumnos, tales como los siguientes:

- Un estudio de uso de la regresión logística para estudiar la aprobación de la materia de Matemáticas Básicas de la FCFM en las generaciones 2010 y 2011.

Se encontró que los alumnos tienen malos hábitos de estudio. Las dos generaciones se comportan de manera diferente: la mejor en aprobación es la 2011 y se debe a que estos alumnos van estudiando diariamente los conceptos vistos durante la clase, mientras que los alumnos de la generación 2010 deciden estudiar hasta la fecha del examen [15].

- Un análisis estadístico de algunos factores que afectan el proceso de enseñanza aprendizaje en la FCFM, usando técnicas estadísticas multivariadas, donde se encontró que uno de los principales factores fue el hecho de que la Licenciatura en Matemáticas no fuera su primera opción, ya que esto afecta en el desempeño académico, otro factor que influye es la atención que tiene el maestro en sus cursos, así como los factores económico, social y cultural, entre otros [13].

- En un estudio acerca del proceso de enseñanza aprendizaje en la FCFM-BUAP se encontró que el lugar de Procedencia es un indicador, ya que los alumnos que provienen de otros estados son los que más acreditan la materia de Matemáticas Básicas (este estudio junto con el anterior sólo son para Matemáticas y Matemáticas Aplicadas de las generaciones 2000-2004) [14].

- Un estudio de identificación de factores que intervienen en la reprobación del curso de Matemáticas Básicas de la FCFM de la BUAP [21]. El análisis se realizó del período Primavera 2007 a Otoño 2010, y encontró que los principales factores son:
  1.-El profesor;
  2.-La falta de asistencia a asesorías que compete tanto a estudiantes como a profesores;
  3.- La literatura empleada en el curso.

Se mencionan estos trabajos, ya que sus resultados son importantes y proporcionan los antecedentes para investigar a la población de egresados y no egresados [9].

1.1.4 Eficiencia Terminal

La eficiencia terminal es determinante para conocer la eficacia de un programa de estudio; aunado a esto se encuentra la deserción en los programas lo que ocaciona

un problema serio en el Sistema Educativo Mexicano por su incidencia negativa sobre los procesos políticos, económicos, sociales y culturales del desarrollo nacional [23].

Para comprender cómo se ha venido aplicando este concepto en México, conviene comenzar por una definición normativa. La Dirección General de Planeación, Programación y Presupuesto de la Secretaría de Educación Pública (DGPPP/SEP) la define algebráicamente como la **relación porcentual entre los egresados de un nivel educativo dado y el número de estudiantes que ingresaron al primer grado de este nivel educativo** n **años antes**. Con el fin de controlar el sesgo de estimación por alumnos reprobados, a n se le resta uno. En la evaluación de instituciones educativas se ha dado tal importancia a la eficiencia terminal así definida, que la DGPPP afirma que, es sin lugar a dudas la manifestación de la eficiencia del sistema educativo (SEP, 1977) [23].

En la Tabla 1.1 se observan los porcentajes de eficiencia terminal en Educación Superior Mexicanas en diferentes períodos. En ella podemos observar el porcentaje de eficiencia terminal en las Instituciones de Educación Superior, el cual ha aumentado cada año, al menos en 1%. Sin embargo el porcentaje en cuanto a la cobertura, es decir el porcentaje de alumnos en educación superior, es bajo con un 30% en 2012 (ver Apéndice D) para las definiciones de las instituciones).

1.2 Justificación

El interés por los fenómenos de la deserción y de la eficiencia terminal, constatado por el incremento en el número de estudios al respecto, ha ratificado no tan sólo el consenso sobre la importancia de los mismos, sino que se ha convertido en materia de controversia teórica y empírica. En consecuencia, su naturaleza amerita que se le ubique con el propósito de clarificarlo a través de análisis precisos [34].

La eficiencia terminal es un indicador importante en las metas y objetivos que tienen las universidades hoy en día, aunque se ha visto que con esto no se puede juzgar la calidad de las instituciones y el aprendizaje de los estudiantes. En las instituciones de educación media y superior, se mantienen vivos problemas que a todos preocupa o debiera preocupar, estos son, los asuntos de reprobación, rezago y deserción entre los estudiantes del área de ciencias [3].

Tabla 1.1: Indicadores de Eficiencia Terminal en las IES 2007-2012

Indicador	Unidad de medida	Situación actual	2007	2008	2009	2010	2011	2012
Eficiencia Terminal en Educación Superior.	Porcentaje de egresados.	62.9%	64.1%	65.3%	67%	68%	69%	70%
Matricula de Educación Superior que alcanzan el nivel 1 de las CIEES y son acreditados por la COPAES.	Porcentaje de alumnos en Programas de Educación Superior que alcanzan el nivel 1 y son acreditados.	38.3%	41.9%	45.%	49%	53%	55%	60%
Nivel de Cobertura en Educación Superior.	Matricula escolarizada.	24.3%	25.3%	26.2%	27.3%	28%	29%	30%

Fuente: Programa Sectorial de la Secretaría de Educación Pública, SEP 2007-2012.

Los alumnos que estudian una licenciatura de ciencias se enfrentan a varias problemáticas: dificultad en los temas que se ven y el cambio de ambiente universitario. Ahora, los alumnos deben ser responsables por sí mismos, de sus logros académicos pues se esmeran en sus tareas, estudios y exámenes de las materias que cursan. También los alumnos empiezan a conocer las matemáticas desde sus propiedades elementales, axiomas y se aprende a manejar la lógica matemática para demostrar sus resultados [26].

Además estos estudios nos ayudan como una herramienta apropiada para retroalimentar los programas de formación de profesionales e investigadores en las instituciones de educación superior. Estos también son considerados como mecanismos poderosos de diagnóstico de la realidad, con el potencial de inducir en las instituciones la reflexión a fondo sobre sus fines y sus valores. Los resultados de estos estudios pueden asimismo, aportar elementos para redefinir el proyecto de desarrollo de

aquellas instituciones que se mantienen alerta ante las nuevas necesidades sociales, permitiéndoles reconocer y asumir las nuevas formas de práctica profesional que se requieren para sustentar un proceso social menos inequitativo y dependiente. Los estudios de egresados nos ayudan como una herramienta básica para la definición de políticas en el nivel regional, estatal e incluso nacional y para el diseño de estrategias tendientes a propiciar el desarrollo y el fortalecimiento de todas las instituciones educativas del país [28].

1.3 Objetivos

Determinar los factores académicos que se relacionan con la eficiencia terminal, además caracterizar a los alumnos de la Facultad de Ciencias Físico Matemáticas, separados por grupos: alumnos de Física y Física Aplicada, y alumnos de Matemáticas y Matemáticas Aplicadas. Así probar o descartar afirmaciones acerca de, que tanto la formación académica de un alumno de ciencias influye para que egrese o no de su respectivo plan de estudios.

Capítulo 2

Premilinares

2.1 Modelos Lineales Generalizados

Los Modelos Lineales Generalizados proveen una unificación acerca de los procedimientos estadísticos más comunes, usados en estadística aplicada. Estos tienen aplicaciones en disciplinas tan extensamente variadas como agricultura, demografía, ecología, educación, ingeniería, estudios ambientales y de población, geografía, geología, historia, medicina, ciencias políticas, psicología y sociología [17].

En los años donde los primeros términos fueron introducidos por Nelder y Wedderburn en 1970, los modelos lineales generalizados lentamente llegaron a ser bien conocidos y extensamente usados. La introducción de la idea de modelos lineales generalizados a principios de los 70's, ha tenido mayor importancia sobre la manera de aplicar estadística. En los comienzos, fueron usados ante todo restringidos a estudiantes de estadística bastante avanzados, por que las materias que los explicaban y el software disponible fue dirigido a ellos. Uno de los más importantes logros de los modelos lineales generalizados ha sido promover el papel central de la función verosimilitud en inferencia.

2.1.1 Historia

Los principales adelantos de la visión general de los modelos estadísticos, conocidos como modelos lineales generalizados, prolongados más de un siglo; se pueden resumir como sigue:

- Regresión lineal múltiple- una distribución normal con función de enlace la identidad (Legendre, Gauss: principios del siglo XIX).

- Análisis de la varianza (ANOVA) y diseño de experimentos-una distribución normal con función de enlace la identidad (Fisher: 1920-1935).

- Función de verosimilitud-una aproximación general a la inferencia acerca de cualquier modelo estadístico (Fisher, 1922).

- Familia exponencial-una clase de distribuciones con estadísticas suficientes para los parámetros (Fisher, 1934).

- Análisis probit- una distribución binomial con función de enlace probit (Bliss, 1935).

- Logit para proporciones- una distribución binomial con función de enlace logit (Berkson, 1944; Dyke y Patterson, 1952).

- Modelos log lineales para contar- una distribución Poisson con función de enlace log (Birch, 1963).

- Modelos de regresión para datos de supervivencia- una distribución exponencial con función de enlace log o la recíproca (Feigl y Zelen, 1965; Zippin y Armitage, 1966; Glasser, 1967).

- Polinomios inversos- una distribución Gamma con función de enlace la recíproca (Nelder, 1966) [17].

De este modo, ha sido conocido desde el tiempo de Fisher (1934), que muchas de las distribuciones comúnmente usadas fueron miembros de una familia, la cual es llamada la **Familia Exponencial**. Para el final de los 60's, se hizó una síntesis de varios de estos modelos [17]. En 1972, Nelder y Wedderburn dieron el paso futuro para la unificación de la teoría de modelos estadísticos y, en partícular, los modelos de regresión, publicando su artículo sobre los **Modelos Lineales Generalizados (GLM)**.

2.1.2 La Familia Exponencial

Supóngase que se tiene un conjunto de variables de respuesta, aleatorias e independientes $z_i, (i = 1, \ldots, n)$ y que la función de probabilidad (densidad) puede ser escrita en la forma:

$$f(z_i; \xi_i) = r(z_i)s(\xi_i)exp[t(z_i)u(\xi_i)] = exp[t(z_i)u(\xi_i) + u(\xi_i) + w(\xi_i)], \quad z_i \in \mathbb{R} \quad (2.1)$$

con ξ_i un parámetro de localización indicando la posición donde la distribución está dentro del rango de posibles valores de respuesta. Cualquier distribución que pueda ser escrita de esta manera es un miembro de la familia exponencial (de un parámetro).

La **forma canónica** para la variable aleatoria, el parámetro y la familia se obtiene por $y = f(z)$ y $\theta = u(\xi)$. Las transformaciones uno-uno, son las más simples, pero no es una elección fundamental, el modelo ahora es el siguiente:

$$f(y_i; \theta_i) = exp[y_i\theta_i - b(\theta_i) + c(y_i)] \quad (2.2)$$

donde $b(\theta_i)$ es la constante de normalidad de la distribución. Ahora, $y_i (i = 1, \ldots, n)$ es un conjunto de variables aleatorias independientes con media, digamos μ_i, así tenemos el modelo clásico, que se escribe $y_i = \mu_i + \varepsilon_i$.

2.1.3 Familia de Distribución Exponencial

La familia exponencial puede ser generalizada incluyendo un **parámetro de escala** (constante), digamos ϕ en la distribución tal que:

$$f(y_i; \theta_i; \phi) = exp[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)] \quad y_i \in \mathbb{R} \quad i = 1, 2, \cdots, n.$$

donde θ_i es la forma canónica del parámetro de localización, para alguna función de la media, μ_i.

Ejemplo:

Distribución Normal

$$f(y_i;\mu_i,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}} \quad i = 1,2,\cdots,n \quad -\infty < y_i,\mu_i < \infty \quad 0 < \sigma^2 < \infty.$$

$$= exp\{[y_i\mu_i - \frac{\mu_i^2}{2}]\frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)\}$$

donde $\theta_i = \mu_i$, $b(\theta_i) = \mu_i^2/2$, $a_i(\theta) = \sigma^2$ y $c(y_i,\phi) = -[y_i^2/\sigma^2 + log(2\pi\sigma^2)]/2$.

2.2 Las tres componentes de un GLM

Considere de nuevo el modelo de regresión lineal. Este modelo ha sido escrito clasicamente, como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{donde} \quad \varepsilon_i \sim N(0,\sigma^2) \quad i = 1,2,\cdots,n \tag{2.3}$$

pero es más común verlo como:

$$\mu_i = \beta_0 + \beta_1 x_i \tag{2.4}$$

donde μ_i es la media de una distribución normal con varianza constante σ^2. De este modelo simple, no es necesariamente claro que está formado de tres componentes. Se observan dos de ellos, la distribución de probabilidad y la estructura lineal, en menor medida se observa el tercero, la función de enlace. A continuación, se ve en más detalle las tres componentes.

- **Distribución de la Respuesta o Estructura del Error (Componente aleatorio)**

  Las $y_i (i = 1, \ldots, n)$ son variables aleatorias independientes con medias μ_i. Estas proporcionan alguna distribución de la familia de distribución exponencial, con una constante el parámetro de escala.

- **Componente sistemático**
  El componente sistemático de un GLM referido a un vector $\eta_i, \cdots, \eta_n$ de variables explicatorias mediante un modelo lineal. Sea x_{ij} el valor del predictor $j \quad (j = 1, \cdots, p)$ para el sujeto i. Entonces

$$\eta_i = \sum_j x_{ij}\beta_j \quad i = 1, \cdots, n.$$

  Esta combinación lineal de variables explicatorias es llamado el predictor lineal.

- **Función de Enlace**

  Si $\theta_i = \eta_i$, la definición de la generalización del modelo lineal esta completa. La relación entre la media de la $i-$ésima observación y su predictor lineal está dada por una función enlace, $g_i(\cdot)$:

$$\begin{aligned} \eta_i &= g_i(\mu_i) \\ &= x_i^T \beta. \end{aligned} \tag{2.5}$$

Esta función debe ser monótona y diferenciable. Usualmente la misma función de enlace es usada para todas las observaciones. Entonces, la función de enlace canónica es aquella función la cual transforma la media a el parámetro de localización canónico de la familia de distribución exponencial. En la Tabla 2.1 se observan las funciones de enlace de las distribuciones más comúnes.

Con la función de enlace canónica y todos los parámetros desconocidos de la estructura lineal tenemos sufientes estadísticas (véase Apéndice B), si la distribución de la respuesta es un miembro de la familia de distribución exponencial y el

Tabla 2.1: Funciones de Enlace Canónicas de las principales distribuciones

Distribución	Funciones de Enlace Canónicas
Poisson	Log $\eta_i = log(\mu_i)$
Binomial	Logit $\eta_i = log[\frac{\pi_i}{1-\pi_i}] = log[\frac{\mu_i}{\eta_i-\mu_i}]$
Normal	Identidad $\eta_i = \mu_i$
Gamma	Recíproca $\eta_i = \frac{1}{\mu_i}$

parámetro de escala es conocido. En la Tabla 2.2, se pueden observar las características de las distribuciones más comúnes, se observan la Bernoulli o la Poisson.

Tabla 2.2: Características de las Distribuciones

	Rango de y	Distribución$f(y)$	$\mu(\theta)$	Varianza V(μ)	Término $a(\Phi)$
Bernoulli B(μ)	{0,1}	$\mu^y(1-\mu)^{1-y}$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$	1
Binomial $B(k,\mu)$	{0,...,k}	$\binom{a}{b}\mu^y(1-\mu)^{k-y}$	$\frac{ke^\theta}{1+e^\theta}$	$\mu(1-\frac{\mu}{k})$	1
Poisson $P(\mu)$	$\{0,1,2,\ldots\}$	$\frac{\mu^y}{y!}e^-\mu$	$exp(\theta)$	μ	1
Geométrica $Geo(\mu)$	$\{0,1,2,\ldots\}$	$(\frac{\mu}{1+\mu})^y(\frac{1}{1+\mu})$	$\frac{e^\theta}{1-e^\theta}$	$\mu+\mu^2$	1
Binomial Negativa $NB(\mu,k)$	$\{0,1,2,\ldots\}$	$\binom{k+y-1}{y}(\frac{\mu}{k+\mu})^y(\frac{k}{k+\mu})$	$\frac{ke^\theta}{1-e^\theta}$	$\mu+\frac{\mu^2}{k}$	1
Exponencial $Exp(\mu)$	$(0,\infty)$	$\frac{1}{\mu}exp(\frac{-x}{\mu})$	$\frac{-1}{\theta}$	μ^2	1
Gamma $G(\mu,\psi)$	$(0,\infty)$	$\frac{1}{\Gamma(\psi)}(\frac{\psi}{\mu})^\psi exp(\frac{-\psi y}{\mu})y^{\psi-1}$	$\frac{-1}{\theta}$	μ^2	$\frac{1}{\psi}$
Normal $N(\mu,\psi^2)$	$(-\infty,\infty)$	$\frac{exp\{-(y-\mu)^2/2\psi^2\}}{\sqrt{2\pi\psi}}$	θ	1	ψ^2

2.3 El modelo de Regresión Logística

Los métodos de regresión se han convertido en un componente integral de cualquier análisis de datos, con la descripción de la relación entre una variable de respuesta y una o más variables explicatorias. Muy a menudo la variable de respuesta es discreta, teniendo dos o más valores posibles. El modelo de regresión logística es el modelo de regresión más frecuentemente utilizado para el análisis de estos datos. Antes de iniciar un estudio a fondo del modelo, es importante entender que el objetivo de un análisis utilizando este modelo es, encontrar el modelo mejor ajustado y más parsimonioso, interpretable para describir la relación entre una variable de respuesta y un conjunto de variables independientes (explicatorias). Las variables independientes son a menudo llamadas **covariables**.

Lo que distingue a un modelo de regresión logística a partir del modelo de regresión lineal es que la variable de respuesta en la regresión logística es **binaria** o **dicotómica**. Ésta diferencia entre la regresión logística y lineal se refleja tanto en la forma del modelo y sus supuestos. Una vez que esta diferencia se explica, los métodos empleados en el análisis de regresión logística utilizados se siguen, más o menos, de los mismos principios generales utilizados en la regresión lineal.

Sea $Y = (y_1, \cdots, y_n)$ un vector de respuestas independientes cada y_i $i = 1, \cdots, n$ sólo puede tomar uno de los dos valores 0 o 1, y sea $\pi(x) = P(y_i = 1)$ con x una variable explicatoria. Usualmente, los datos binarios resultan de una relación no lineal entre $\pi(x)$ y x. Un cambio fijo en x frecuentemente tiene un menor impacto cuando $\pi(x)$ está cerca de 0 o 1 que cuando $\pi(x)$ está cerca de 0.5. En la práctica, las relaciones no lineales entre $\pi(x)$ y x son frecuentemente monótonas, con $\pi(x)$ creciendo continuamente o $\pi(x)$ decreciendo continuamente si x crece. La curva más importante con esta forma tiene el modelo con la ecuación:

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}. \tag{2.6}$$

Este es el modelo de regresión logística. Si x tiende a ∞, entonces $\pi(x)$ tiende a 0 cuando $\beta < 0$ y si $\pi(x)$ tiende a 1 cuando $\beta > 0$. La función enlace por la cual el modelo de regresión logística es un GLM es en términos de odds es:

$$g(x) = ln(\frac{\pi(x)}{1 - \pi(x)}) = \beta_0 + \beta_1 x. \tag{2.7}$$

El lado izquierdo es el logaritmo de la probabilidad de éxito de y_i [37]. El modelo asume que este log-odds (o logit) es una función lineal del predictor x. La función de probabilidad del modelo se puede escribir en la forma de la familia exponencial como:

$$\pi(x)^y(1-\pi(x))^1 - y = (1-\pi(x))exp\{yln(\frac{\pi(x)}{1-\pi(x)})\}. \tag{2.8}$$

La importancia de esta transformación es que $g(x)$ tiene muchas de las propiedades deseables del modelo de regresión lineal. El logit, $g(x)$, es lineal en sus parámetros, es continua en $\mathbb{R}$, dependiendo del rango de x.
Cuando Y es una variable aleatoria binaria, la esperanza sin condiciones de Y es la probabilidad de que el evento ocurra [20],

$$E(y_i) = [1xP(y_i = 1)] + [0xP(y_i = 0)] = P(y_i = 1). \tag{2.9}$$

Para el modelo de regresión se toma la esperanza condicional:

$$E(y_i|x) = [1xP(y_i = 1|x)] + [0xP(y_i = 0|x)] = P(y_i = 1|x). \tag{2.10}$$

2.4 Ajuste del modelo de Regresión Logística

Supóngase que tenemos una muestra de n observaciones independientes (x_i, y_i), $i = 1, 2, \cdots, n$, donde y_i denota el valor de una respuesta de una variable dicotómica y x_i es el valor de la variable independiente para el i-ésimo sujeto. Además, asumimos que la variable de respuesta está codificada como 0 o 1, representando la ausencia o presencia de la característica, respectivamente. Ajustando el modelo de regresión logística en la ecuación (2.6) a un conjunto de datos, requiere que se estime los valores de β_0 y β_1, parámetros desconocidos.

En regresión lineal, el método usado más frecuente para estimar los parámetros desconocidos es el de los **mínimos cuadrados**. El método en general de la estimación que conduce a la función de mínimos cuadrados en el marco del modelo de regresión lineal (cuando los términos de error se distribuyen normalmente) se llama máxima verosimilitud. En un sentido general, el método de máxima verosimilitud produce valores para los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observado. Para aplicar este método debemos primero construir una función, llamada la **función de verosimilitud**. Esta función de verosimilitud expresa la probabilidad de los datos observados como una función de los parámetros desconocidos. Los **estimadores de máxima verosimilitud** de los parámetros, son los

valores que maximizan esta función. Por lo tanto, los estimadores resultantes son los que están estrechamente relacionados con los datos observados.

Si Y está codificada como 0 o 1 entonces la expresión para $\pi(x)$ dada en la ecuación (2.6) nos provee (para valores arbitrarios de $\beta = (\beta_0, \beta_1)$, el vector de parámetros) la probabilidad condicional de que Y es igual a 1 dado x y $1 - \pi(x)$ da la probabilidad condicional que Y es igual a cero dado x, $P(Y = 0|X = x)$. Por lo tanto, para los pares (x_i, y_i), donde $y_i = 1$, la contribución a la función de verosimilitud es $\pi(x_i)$, y para este par donde $y_i = 0$, la contribución a la función de verosimilitud es $1 - \pi(x_i)$, donde la cantidad $\pi(x_i)$ denota el valor de $\pi(x)$ calculado en x_i. Una manera conveniente de expresar la contribución a la función de verosimilitud para el par (x_i, y_i) es tomar la expresión

$$\pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}. \tag{2.11}$$

Como las observaciones se asumen independientes, la función de verosimilitud se obtiene como el producto de los términos dados en la expresión (2.11):

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i)_i^y)^{1-y_i}. \tag{2.12}$$

El principio de máxima verosimilitud establece que utilicemos como nuestra estimación de β el valor que maximiza la expresión en la ecuación (2.12). Sin embargo, es más fácil matemáticamente para trabajar con el logaritmo de la ecuación (2.12). Esta expresión, el log-verosimilitud, se define como

$$L(\beta) = ln(l(\beta)) = \sum_{i=1}^{n} \{y_i ln(\pi(x_i)) + (1 - y_i) ln(1 - pi(x_i))\}. \tag{2.13}$$

Para encontrar el valor de β que maximiza $L(\beta)$ diferenciamos $L(\beta)$ con respecto a β_0 y β_1 y el conjunto de expresiones resultantes se iguala a cero. Éstas ecuaciones, conocidas como las *ecuaciones de verosimilitud*, son:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{2.14}$$

y

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0. \tag{2.15}$$

Para la regresión logística las expresiones en las ecuaciones (2.14) y (2.15) son no lineales en β_0 y β_1, y por lo tanto requiere métodos especiales para su solución (un método para este tipo de modelos es el de mínimos cuadrados ponderados). El valor de β dado para las soluciones (2.14) y (2.15) son llamados los *estimadores de máxima verosimilitud* y se denota como $\hat{\beta}$. Ver [33] para ver una expresión de $\hat{\beta}$. Una interesante consecuencia de la ecuación (2.14) es que :

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i). \tag{2.16}$$

Esto es, la suma de los valores observados de Y es igual a la suma de los valores predichos (esperados).

2.5 El modelo de Regresión Logística Múltiple

Sea una colección de p variables independientes denotadas por el vector $X = (x_1, x_2, \cdots, x_p)$. Sea la probabilidad condicional que la respuesta está presente denotada por $P(Y = 1|X) = \pi(X)$. La función logit del modelo de regresión logística múltiple está dado por la ecuación

$$g(X) = ln(\frac{\pi(X)}{1 - \pi(X)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad x_i \in \mathbb{R}, \quad i = 1, \cdots, p \tag{2.17}$$

donde, para el modelo de regresión logística múltiple,

$$\pi(X) = \frac{e^{g(X)}}{1 + e^{g(X)}}. \tag{2.18}$$

Si alguna de las variables independientes son variables discretas, de escala nominal tal como raza, sexo, grupo de tratamiento y así sucesivamente, es incorrecto incluirlas en el modelo como si fueran variables de escala de intervalo. Los números que se utilizan para representar los distintos niveles de estas variables de escala nominal no son más que identificadores, y no tienen ninguna significación numérica.

En esta situación, el método de elección es usar una colección de **variables de diseño** (o **variables dummy**). Supongamos, por ejemplo, que una de las variables independientes es raza, que se ha codificado como "blanco", "negro" y "otros". En este

caso, dos variables de diseño son necesarias. Una de las estrategias de codificación posible es que cuando la respuesta es blanco, las dos variables de diseño, D1 y D2, se igualarán a cero; cuando la respuesta es negro, D1 se establece igual a 1, mientras que D2 seguiría igual a 0; cuando la raza del demandado es otro, usaríamos D1=0 y D2 = 1.

En general, si una variable nominal de escala tiene k posibles valores, entonces $k-1$ variables de diseño son necesarias. La razón para usar uno menos que el número de valores es que, a menos que se indique lo contrario, los modelos disponen de un término constante. Para ilustrar la notación usada para las variables de diseño en este texto, suponga que la variable independiente j-ésima x_j tiene k_j niveles. Las $k_j - 1$ varibles de diseño se denotan como $D_j l$ y los coeficientes de estas variables se denotan como $\beta_j l$, $l = 1, 2, \cdots, k_j - 1$. Por lo tanto, el logit para un modelo con p variables, con la $j-$ésima variable discreta es:

$$g(X) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p. \tag{2.19}$$

Con algunas pocas excepciones, suprimimos la suma y doble subíndice necesarios para indicar cuando se utilizan variables de diseño, cuando se habla del modelo de regresión logística múltiple.

2.6 Ajuste del modelo de Regresión Logística Múltiple

Asumimos que tenemos una muestra de n observaciones independientes (x_i, y_i), $i = 1, 2, \cdots, n$. Ajustar el modelo requiere que obtengamos las estimaciones del vector $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$. El método de estimación usado en el caso multivariable es el de máxima verosimilitud. La función de verosimilitud es casi idéntica a la dada en la ecuación (2.12) con el único cambio que $\pi(X)$ se define ahora como en la ecuación (2.18). Aquí serán $p+1$ ecuaciones de verosimilitud que son obtenidas diferenciando la función log-verosimilitud con respecto a los $p+1$ coeficientes. Las ecuaciones de verosimilitud que resultan se expresan como sigue:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{2.20}$$

y

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0 \tag{2.21}$$

para $j = 1, \cdots, p$.

Sea $\hat{\beta}$ denota la solución a estas ecuaciones. Por lo tanto, los valores ajustados para el modelo de regresión logística múltiple son $\hat{\beta}$. El método de estimación de varianzas y covarianzas de los coeficientes estimados se sigue de la teoría de estimación de máxima verosimilitud.

2.7 Selección de variables

Los criterios para la inclusión de una variable en un modelo pueden variar de un problema a otro y de una disciplina científica a otra. El enfoque tradicional para la construcción de modelos estadísticos implica buscar el modelo más parsimonioso que aún refleja con precisión la verdadera experiencia, resultado de los datos. La justificación para minimizar el número de variables en el modelo es que el modelo resultante es más probable que sea estable numéricamente, y se adopte más fácilmente para su uso. Entre más variables incluidas en un modelo, los errores estándares estimados se hacen más grandes.

La justificación de este enfoque es proporcionar un control tan completo de los factores de confusión como sea posible dentro del conjunto de datos dados. Esto se basa en el hecho de que es posible para las variables individuales no exhibir fuerte confusión, pero cuando se toman en conjunto, una considerable confusión puede estar presente en los datos, (véase [16]). El principal problema con este enfoque es que el modelo puede estar "sobreajustado", produciendo estimaciones numéricamente inestables. El sobreajuste se caracteriza típicamente por coeficientes estimados y/o errores estándares estimados grandes irreales.

Esto puede ser especialmente difícil en problemas en los que el número de variables en el modelo es grande en relación con el número de sujetos y / o cuando la proporción global de responder $(y = 1)$ está cerca de 0 o 1. Los siete pasos si-

guientes describen un método de selección de variables que llamamos **selección intencionada**.

Paso 1: La selección intencionada comienza con un análisis univariable cuidando de cada variable independiente. Para las variables categóricas se sugiere hacer esto a través de un análisis de Tablas de Contingencia estándar de los resultados $(y = 0, 1)$ frente a los k niveles de la variable independiente. La prueba habitual de chi-cuadrado de razón de verosimilitud con $k - 1$ grados de libertad es exactamente igual al valor de la prueba de razón de verosimilitud para la significación de los coeficientes para la $k - 1$ variables de diseño en un modelo de regresión logística univariable que contiene esa sola variable independiente.

Dado que la prueba de chi-cuadrado de Pearson es asintóticamente equivalente a la prueba de chi-cuadrado de razón de verosimilitud, también puede ser utilizado, para aquellas variables que presentan por lo menos un nivel moderado de asociación, para estimar las razones de posibilidades individuales (junto con límites de confianza) con uno de los niveles del grupo de referencia.

Para las variables continuas, el mejor análisis univariable implica ajustar un modelo de regresión logística univariable para obtener el coeficiente estimado, el error estándar estimado, la prueba de razón de verosimilitud para la significación del coeficiente, y el estadístico de Wald.
Un análisis alternativo, que es casi equivalente a nivel univariable y que puede ser preferido en un entorno aplicado se basa en el t-test de dos muestras (ver Apéndice A).

El análisis univariable basado en la prueba t se puede utilizar para determinar si la variable debe incluirse en el modelo ya que el $p-$valor debe ser del mismo orden de magnitud que la de la estadística de Wald, valor de la prueba, o radio de verosimilitud de regresión logística. A través de la utilización de estos análisis identificamos, como candidatos para un primer modelo multivariable, cualquier variable cuya prueba univariable tiene un p valor menor de 0.25, junto con todas las variables importantes.

Paso 2: Ajustar el modelo multivariable que contiene todas las covariables identificadas para su inclusión en el paso 1, enseguida se evalúa la importancia de cada covariable utilizando el $p-$valor de su estadístico de Wald. Variables que no contribuyen, en los niveles tradicionales de significación estadística, deben ser eliminadas y un nuevo ajuste del modelo. El nuevo modelo, más pequeño, debe ser comparado con el anterior modelo, más grande, mediante la prueba de razón de verosimilitud parcial. Esto es especialmente importante si más de un término ha sido retirado del modelo, que es siempre el caso cuando en una variable categórica con más de dos niveles ha sido incluido el uso de dos o más variables dummy que parecen ser no significativas. Además, hay que prestar atención para asegurarse de que las muestras utilizadas para adaptarse a los modelos más grandes y más pequeños son las mismas.

Paso 3: Tras el ajuste del modelo más pequeño reducido comparamos los valores de los coeficientes estimados en el modelo más pequeño de sus respectivos valores desde el modelo más grande. En particular, debemos estar preocupados por cualquier variable cuyo coeficiente ha cambiado marcadamente en magnitud. Esto indica que una o más de las variables excluidas son importantes en el sentido de proporcionar un ajuste necesario del efecto de las variables que permanecieron en el modelo. Tal variable(s) debe añadirse de nuevo en el modelo. Este proceso de eliminación, montaje, y verificación continua, ciclado por los pasos 2 y 3, hasta que parece que todas las variables importantes se incluyen en el modelo y los excluidos son clínicamente y / o estadísticamente insignificantes. En este proceso se recomienda que se debe proceder lentamente eliminando pocas covariables a la vez [16].

Paso 4: Añadir cada variable no seleccionada en el paso 1 para el modelo obtenido en la conclusión del ciclo a través de los pasos 2 y 3, una a la vez, y de verificar su significación, ya sea por el estadístico de Wald, $p-$valor o la prueba parcial de razón de verosimilitud, si es una variable categórica con más de dos niveles. Este paso es vital para las variables de identificación que, por sí mismas, no están significativamente relacionadas con el resultado, pero hacen una contribución importante en la presencia de otras variables.

Paso 5: Una vez que hemos obtenido un modelo que contiene las variables esenciales, se examinan más de cerca las variables del modelo. La cuestión de los niveles apropiados para las variables categóricas se deben haber abordado durante el análisis univariable en el paso 1 para cada variable continua, en este modelo debemos comprobar el supuesto de que el logit aumenta/disminuye linealmente como una función de la covariable. Hay un número de técnicas y métodos para hacer esto. Nos refirimos a este modelo como modelo de efectos principales.

Paso 6: Una vez que tenemos el modelo de efectos principales, se debe verificar las interacciones entre las variables del modelo. En cualquier modelo, una interacción entre dos variables implica que el efecto de cada variable no es constante en los niveles de la otra variable. Las variables de interacción se crean como el producto de los principales pares de variables de efecto. Esto puede resultar en más de un término de interacción. Por ejemplo, la interacción de dos variables categóricas, cada uno con tres niveles (es decir, dos variables dummy), genera cuatro variables de interacción. Añadimos las interacciones, uno a la vez, en el modelo de efectos principales desde el paso 5 (Esto puede implicar la adición de más de un término al mismo tiempo al modelo).

Paso 7: Antes de cualquier modelo se convierta en el **modelo final** debemos evaluar su idoneidad y comprobar su ajuste. Independientemente de qué método se utiliza para obtener un modelo estadístico multivariable, la selección intencionada o cualquiera de los otros métodos, uno debe realizar el paso 7 antes de usar el modelo para fines de inferencia. Bursac ([16]), estudió las propiedades de la selección intencionada en comparación con la selección paso a paso a través de simulaciones.
Otro enfoque para la selección de variables es utilizar un método paso a paso en el que se seleccionan las variables ya sea para la inclusión o exclusión del modelo de una manera secuencial basada únicamente en criterios estadísticos (ver Apéndice C). Hay dos versiones principales del procedimiento por etapas: (i) de selección hacia adelante con una prueba de eliminación hacia atrás y (ii) eliminación hacia atrás seguido de una prueba de selección hacia adelante. El enfoque progresivo es útil y intuitivamente atractivo ya que construye modelos de una manera secuencial y permite el examen de una colección de modelos que podrían no haber sido examinados.

2.8 Evaluación del ajuste del modelo de Regresión Logística

Supongamos que tenemos los valores de muestra observados de la variable de respuesta, denotados en forma vectorial, como Y, donde $Y = (y_1, y_2, y_3, \cdots, y_n)$. Denotamos los valores estimados por el modelo, o los **valores ajustados**, como $\hat{Y}$ donde $\hat{Y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)$. Concluimos que el modelo se ajusta si: **(1)** las medidas de resumen de la distancia entre Y e $\hat{Y}$ son pequeñas y **(2)** la contribución de cada par, $(y_i, \hat{y}_i)$, $i = 1, 2, 3, \cdots, n$, de estas medidas de resumen no es sistemática y pequeña en relación con la estructura de los errores del modelo. Por lo tanto, una evaluación completa del modelo ajustado implica tanto el cálculo de medidas de resumen y un examen minucioso de los componentes individuales de estas medidas.

Suponga que nuestro modelo ajustado contiene p variables independientes, $X = (x_1, \cdots, x_p)$, y sea J el número de valores distintos de x observados. Si alguno de los sujetos tiene el mismo valor de x entonces $J < n$. Se denota el número de sujetos con $x = x_j$ por $m_j, (0.3cm)j = 1, \cdots, J$. De esto se sigue que $\sum m_j = n$. Sea y_j el número de respuestas , $y = 1$, entre los m_j sujetos con $x = x_j$

2.8.1 El test de Hosmer-Lemeshow

Hosmer y Lemeshow (1980, 1982, véase [16]) propusieron la agrupación basada en los valores de las probabilidades estimadas. En este caso pensamos en las n columnas como correspondiente a los n valores de las probabilidades estimadas, con la primera columna correspondiente al valor más pequeño, y la columna n-ésima con el valor más grande. Dos estrategias de agrupación se propusieron de la siguiente manera: **(i)** el colapso de la Tabla basada en los percentiles de las probabilidades estimadas y **(ii)** el colapso de la Tabla basada en los valores fijos de la probabilidad estimada.

Con el primer método, el uso de $g = 10$ grupos de resultados, el primer grupo que contiene los $n_1'0 = n/10$ sujetos que tienen las probabilidades estimadas más pequeñas, y el último grupo que contiene los $n_1'0 = n/10$ sujetos que tienen las mayores probabilidades estimadas. Con el segundo método, el uso de $g = 10$ grupos de resultados en puntos de corte definidos en los valores de $k/10$, $k = 1, 2, \cdots, 9$ y los grupos contienen todos los sujetos con probabilidades estimadas entre los puntos de corte adyacentes. Por ejemplo, el primer grupo contiene todos los sujetos cuya

probabilidad estimada es menor que o igual a 0.1, mientras que el décimo grupo contiene aquellos sujetos cuya probabilidad estimada es superior a 0.9. Para la fila $y = 1$, las estimaciones de los valores esperados se obtienen sumando las probabilidades estimadas sobre todos los sujetos en un grupo.

Para la fila $y = 0$, el valor esperado estimado se obtiene sumando, sobre todos los sujetos en el grupo, uno menos la probabilidad estimada. Para cualquiera de estas estrategias de agrupación, el estadístico de bondad de ajuste de Hosmer-Lemeshow $\hat{C}$, se obtiene calculando el estadístico chi-cuadrado de Pearson de la Tabla $g \times 2$ de frecuencias esperadas observadas y estimadas. Una fórmula que define el cálculo de $\hat{C}$ es el siguiente:

$$\hat{C} = \sum_{k=1}^{g} [\frac{(o_{ik} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}}] \tag{2.22}$$

donde

$$o_{1k} = \sum_{j=1}^{c_k} y_j,$$

$$o_{0k} = \sum_{j=1}^{c_k} (m_j - y_j),$$

$$\hat{e}_{1k} = \sum_{j=1}^{c_k} m_j \hat{\pi}_j,$$

$$\hat{e}_{0k} = \sum_{j=1}^{c_k} m_j (1 - \hat{\pi}_j)$$

y c_k es número de covariables en el k-ésimo grupo. Con algo de álgebra se demuestra que

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_{1k} - n_k^{'} \bar{\pi}_k)^2}{n_k^{'} \bar{\pi}_k (1 - \bar{\pi}_k)}, \tag{2.23}$$

donde $\bar{\pi}_k$ es el promedio de la probabilidad estimada en el $k-$ésimo grupo,

$$\bar{\pi}_k = \frac{1}{n_k'} \sum_{j=1}^{c_k} m_j \hat{\pi}_j$$

.

Usando un amplio conjunto de simulaciones, Hosmer y Lemeshow demostraron que, cuando $J = n$ y el modelo de regresión logística ajustada es el modelo correcto, la distribución del estadístico $\hat{C}$ está bien aproximada por la distribución chi-cuadrado con $g - 2$ grados de libertad $\chi^2(g-2)$. Aunque no examinado específicamente, es probable que $\chi^2(g-2)$ también se aproxima a la distribución cuando $J \equiv n$. Una alternativa al denominador se muestra en la ecuación (2.23) si consideramos o_{1k} la suma de variables aleatorias distribuidas no idénticamente independientes. Esto sugiere que deberíamos normalizar la diferencia al cuadrado entre las frecuencias observadas y estimadas por:

$$\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j).$$

Es fácil demostrar que:

$$\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j) = n_k' \bar{\pi}_k (1 - \bar{\pi}_k) - \sum_{j=1}^{c_k} m_j (\hat{m}_j (\hat{m}_j - \bar{\pi}_k)^2.$$

En una serie de simulaciones Xu (1996) (véase [16]) demostró que el uso de

$$\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j)$$

resulta en un incremento en el valor de estadístico de prueba. Pigeon y Heyse (1999) propusieron un ajuste que es el radio de dos estimadores

$$\phi_k = \frac{\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j)}{n_k' \bar{\pi}_k (1 - \bar{\pi}_k)}$$

obteniendo el estadístico

$$\hat{C}_p = \sum_{k=1}^{g} \frac{1}{\phi_k} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(0_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right]$$

$$= \sum_{k=1}^{g} \frac{1}{\phi_k} \frac{(o_{1k} - n'_k \bar{\pi}_k)^2}{\hat{e}_{0k}}$$

$$= \sum_{k=1}^{g} \left[\frac{o_{1k} - n'_k \bar{\pi}_k)^2}{\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j)} \right].$$

Pigeon y Heyse reportaron en sus simulaciones que la distribución de $\hat{C}_p$, bajo la hipótesis de que uno tiene ajustar el modelo correcto, con una muestra suficientemente grande, se aproxima por la distribución $\chi^2(g-1)$. Esto parece contradecir a Xu que mostró en sus simulaciones que la distribución de $\hat{C}_p$ podría ser bien aproximada por $\chi^2(g-2)$.

La ventaja de cualquier resumen del estadístico de bondad de ajuste, por ejemplo $\hat{C}$, es que proporciona un único y fácilmente interpretable número que se puede utilizar para evaluar el ajuste. La desventaja es que en el proceso de agrupar podemos perder una desviación importante de ajuste debido a un pequeño número de puntos de datos individuales.

2.8.2 Área dentro de la Curva Característica Operativa del Receptor COR

La sensibilidad y la especifidad, tan excelentes como otras medidas de actuación y clasificación calculadas de Tablas 2×2, dependen del punto de corte usado para clasificar un resultado de hipótesis como positivo. Una mejor y más completa descripción es el área dentro de la curva **Característica de Operación de Recepción** (en inglés ROC: the Receiver Operating Characteristic Curve). Esta curva se origina de la teoría de detección de señales, demuestra como la recepción detecta la existencia de señales en la presencia de ruido. En ésta gráfica la probabilidad de detectar una señal verdadera (sensitividad) y una señal falsa (1-especifidad) para un rango entero de posibles puntos de corte. Esta medida está entre los estándares para evaluar un modelo ajustado a asignar en general, probabilidades grandes de los resultados a los

subgrupos, que no adquiere la salida $(y = 0)$.
El área dentro de la curva COR (se denotará por A esta área), cuyo rango va de 0.5 a 1.0, provee una medida de la capacidad del modelo para la discriminación entre los sujetos quienes experimentan la salida de interés contra quienes no. Así que, ¿cuál es el área dentro la curva COR que describe buena discriminación? (véase [16]). Desafortunadamente, no hay un número mágico, solo directrices. En general, se usan las siguientes reglas:

$$Si \begin{cases} A = 0.5 & \text{No hay discriminación.} \\ 0.5 < A < 0.7 & \text{Pobre discrimación.} \\ 0.7 \leq A < 0.8 & \text{Aceptable discriminación.} \\ 0.8 \leq A < 0.9 & \text{Excelente discriminación.} \\ A \geq 0.9 & \text{Excepcional discriminación.} \end{cases}$$

Definición. Sensibilidad: Es la proporción de individuos que son correctamente clasificados, calculada como el complemento de los falsos negativos.

Definición. Especifidad: Es la proporción de individuos que son correctamente clasificados, calculada como el complemento de los falsos positivos.

2.8.3 Supuestos del modelo

Cuando no se cumplen los supuestos del modelo logístico los resultados pueden ser sesgados (los coeficientes son demasiado altos o bajos) o ineficientes (el error estándar es demasiado alto para la dimensión del coeficiente). En el origen de estos problemas ponemos señalar:

1. **Incorrecta especificación del modelo.** Por ejemplo, falta alguna variable explicativa importante.

2. **Omisión de casos.** No existen casos en alguna de las celdas de la Tabla de Contingencia que cruza la variable dependiente (Y=0 o Y=1) y una variable independiente categórica.

3. **No normalidad de los residuos.** Al igual que en la regresión lineal múltiple, los residuos deben seguir una distribución determinada. Mientras en el primer caso deben seguir una distribución normal, en el segundo, la regresión logística, deben seguir una distribución binomial (que puede aproximarse a una normal para muestras de gran tamaño).

4. **Multicolinealidad.** Alguna de las variables independientes tiene un grado de correlación alto.

Según esto el primero y el segundo son fácilmente comprobables, y en el tercero no hay mucho problema si no se cumple. Pero el problema de la multicolinealidad, si debe tenerse en cuenta en todo análisis de regresión logística [2]. Para ver más acerca del tercer punto [22].

Capítulo 3

Análisis de Datos: de los factores académicos de los alumnos de la FCFM

Este es un análisis de tipo inferencial donde se usa la teoría antes expuesta, pero se ha hecho un análisis descriptivo de la información y se observó que 20% egresa, así en este estudio nos enfocaremos a determinar los factores que determinan la eficiencia terminal por medio de la regresión logística, cuya teoría se explicó en el capítulo anterior.

Es conocido por los profesores que dan los cursos Matemáticas Elementales (ME), Cáculo Diferencial (CD), Cálculo Integral (CI), Cálculo Diferencial en Varias Variables (CDV) y Cálculo Integral en Varias Variables (CIV) que los primeros dos años, los alumnos desertan más que en cualquier otro periodo.
Los alumnos que ya se titularon tienen mejores promedios que los alumnos que todavía están tomando cursos, y los promedios de los alumnos que desertan son menores en general a los otros dos (egresados y en cursos). Lo que sí se ha observado es que el promedio que los alumnos tienen en la preparatoria no es un indicador de que si los alumnos serán capaces de terminar la carrera. El promedio que tienen en tercer año de la carrera está en relación al promedio final del alumno.

3.0.4 Descripción de la información

El Universo consta de 1047 alumnos inscritos en las licenciaturas Matemáticas y Matemáticas Aplicadas, y 990 alumnos de las licenciaturas Física y Física Aplicada, que se ofrecen en la FCFM de la BUAP, la información se obtuvo de la Dirección de

Administración Escolar por medio de sus departamentos de Cómputo y de Titulación y de la Dirección General de Profesiones. El cohorte generacional fue hecho al periodo verano 2013; la información se analizó en SPSS 19 [35]. En la Tabla 3.1 se observan al total de alumnos de las licenciaturas de Matemática y Matemáticas Aplicadas, clasificados por Generación y Sexo. De la misma manera se pueden observar a los alumnos de Física y Física Aplicada en la Tabla 3.2.

Tabla 3.1: 1 Distribución de los alumnos de MAT y LMA por Generación y Sexo

Licenciatura	MAT			LMA		
Generación	Hombres	Mujeres	Total	Hombres	Mujeres	Total
2000	45	38	83	12	7	19
2001	40	43	83	15	8	23
2002	58	26	84	17	14	31
2003	50	35	85	14	15	29
2004	35	29	64	12	15	27
2005	47	46	93	23	13	36
2006	54	35	89	24	12	36
2007	52	36	88	20	18	38
2008	67	54	121	28	24	52
Total	448	342	790	165	126	291

Tabla 3.2: 2 Distribución de los alumnos de FIS y LFA por Generación y Sexo

Licenciatura	FIS			LFA		
Generación	Hombres	Mujeres	Total	Hombres	Mujeres	Total
2000	36	23	59	17	4	21
2001	57	33	90	17	6	23
2002	45	21	66	16	10	26
2003	45	21	73	24	12	36
2004	51	23	64	11	12	23
2005	65	23	88	30	5	35
2006	58	20	68	28	6	34
2007	71	24	85	43	12	55
2008	57	28	75	27	7	34
Total	485	216	701	215	74	289

De la información obtenida de todos los alumnos de las cuatro licenciaturas, estos se clasificaron como: alumnos que desertaron (alumnos que no se inscribieron en dos semestres consecutivos [36]), alumnos no egresados (alumnos que estan inscritos actualmente, pero que de acuerdo a los planes de estudio de la facultad podrían haberse titulado pues por el cohorte generacional considerado tienen al menos 5 años tomando cursos), y alumnos egresados (alumnos que han alcanzado el 100 por ciento de créditos y estan titulados). Así los datos con los que se trabajaron fueron con los alumnos no egresados y egresados; el total de estos alumnos son 308 para las licenciaturas de Matématicas y Matemáticas Aplicadas, y 327 alumnos en Física y Física Aplicada.

Se usó la siguiente codificación. (Matemáticas y Matemáticas Aplicadas): la variable respuesta fue egreso que tuvo dos valores: egresado (Y=1) y no egresado (Y=0). Las variables independientes consideradas fueron: Carrera (Licenciaturas de Matemáticas (MAT) y Matemáticas Aplicadas (LMA)), Periodo (esto se refiere al periodo de ingreso de los alumnos, es decir al año en que ingresaron a alguna de las licenciaturas); las calificaciones en las materias ME, CD, CI, CDV, CIV; y otras variables como Años Cursados (AÑOS), Promedio Primer año (PROM1), Promedio Segundo año (PROM2), Promedio Tercer Año (PROM3), Promedio final (PROMAC), Promedio Preparatoria o Bachiller (PROMPREP), Puntaje exámen de admisión (PUNTAJE) y sexo (SEXO).

(Física y Física Aplicada) : en este caso las variables PROM1, PROM2, PROM3, PR0M4, PROM5, PROM6, PROM7, PROM8, PROM9 y PROM10, PROM11, PROM12, PROM13, PROM14, PROM15, PROM16, PROM17, PROM18, PROM20, se refieren a los promedios de los alumnos de Física, tomados por semestre, del primero al vigésimo semestre. La variable Carrera ahora será: FIS (Física) y LFA (Física Aplicada). Por lo tanto para los alumnos de Física y Física Aplicada se tienen en un inicio 22 variables independientes.

En las Tablas 3.3 y 3.4 se observan las frecuencias de las calificaciones de las materias tomadas como covariables en el análisis de regresión para las cuatro licenciaturas, en éstas se observa que la calificaciones más frecuentes para los alumnos de Física y Física Aplicada en ME, CD, CI, CDV y CIV, son 9, 8, 8, 8, y 10 respectivamente; y para los alumnos de Matemáticas y Matemáticas Aplicadas son 8 en las primeras cuatro materias y 10 en CIV.

Tabla 3.3: Frecuencias de las calificaciones en los alumnos de Matemáticas y Matemáticas Aplicadas

Calificación	ME	%	CD	%	CI	%	CDV	% CIV	%	
5	2	0.6	4	1.3	11	3.6	15	4.9	7	2.5
6	16	5.2	21	6.8	16	5.2	12	3.9	4	1.3
7	38	12.3	53	17.2	42	13.6	36	18.2	40	13.0
8	103	33.4	96	31.2	98	31.8	82	26.6	65	21.1
9	71	23.1	70	22.7	73	23.7	59	19.2	71	23.1
10	78	25.3	64	20.8	67	21.8	81	26.3	99	32.1
perdidos	0	0	0	0	1	0.3	3	1.0	22	7.1

En la Tabla 3.4, se puede observar que la variable ME, tiene 223 valores perdidos, el 68% del total de los datos, y por tanto ME ya no se tomará como variable explicatoria del fenómeno, este mismo análisis se realizó para todas las variables disponibles y como resultado se obtuvo que los promedios del 11 al 20, ya no serán consideradas como variables explicativas.

Tabla 3.4: Frecuencias de las calificaciones en los alumnos de Física y Física Aplicada

Calificación	ME	%	CD	%	CI	%	CDV	%	CIV	%
5	3	0.9	4	1.2	13	4.0	11	3.4	9	2.8
6	13	4.0	26	8.0	20	6.1	9	2.8	10	3.1
7	23	7.0	68	20.8	50	15.3	72	22.0	36	11.0
8	14	4.3	95	29.1	95	29.1	102	31.2	61	18.7
9	30	9.2	79	24.2	70	21.4	63	19.3	91	27.8
10	21	6.4	53	16.2	76	23.2	63	19.3	102	31.2
perdidos	223	68.2	2	0.6	3	0.9	7	2.1	18	5.5

En la Tabla 3.5, observamos que de los 308 alumnos de MAT y LMA, 235 (76.3%) han egresado y 73 aún toman cursos.

En cambio para los alumnos de FIS y LFA, 266 de los 327 han egresado ver Tabla 3.6, esto es el 81.3% del total de alumnos.

Tabla 3.5: Número de alumnos egresados de MAT y LMA

Y	número	%
0	73	23.7
1	235	76.3

Tabla 3.6: Número de alumnos egresados de FIS y LFA

Y	número	%
0	61	18.7
1	266	81.3

Del total de alumnos 97 son de la licenciatura de Matemáticas, 211 de Matemáticas Aplicadas, 91 de Física y 143 de Física Aplicada, esto se observa en las Tablas 3.7 y 3.8.

Tabla 3.7: Frecuencias de alumnos de MAT y LMA por Carrera

Carrera	número	%
MAT	97	31.5
LMA	211	68.5

Tabla 3.8: Frecuencias de alumnos de FIS y LFA por Carrera

Carrera	número	%
FIS	91	18.7
LFA	143	81.3

En las Tablas 3.9 y 3.10, se observa la distribución de los alumnos de la cuatro licenciaturas por Sexo, en éstas se tiene que el número de hombres es mayor al número de mujeres, sobre todo en las licenciaturas de Física y Física aplicada con un 63% de hombres.

A continuación se realizaron la pruebas: comparación de medias para la variable respuesta Y (categórica) y en las variables explicativas cuantitativas, para cada

Tabla 3.9: Frecuencias de alumnos de MAT y LMA por Sexo

Sexo	número	%
M	161	52.2
F	147	47.7

Tabla 3.10: Frecuencias de alumnos de FIS y LFA por Sexo

Sexo	número	%
M	206	63.0
F	121	37.0

grupo de alumnos, y la prueba chi-cuadrado para las variables cualitativas, en este caso Carrera, Sexo y Periodo.

3.1 Modelación del egreso en las carreras de Matemáticas y Matemáticas Aplicadas

El resultado fue que para los alumnos de MAT y LMA, para la variable CD, no se rechaza la hipótesis nula de igualdad de medias y por tanto ésta variable ya no se considerará como variable explicativa; y en la prueba chi-cuadrado las variables Carrera y Sexo resultaron independientes de la variable respuesta (Y), y en el caso de Periodo la prueba no es concluyente, por lo tanto ninguna de estas variables cualitativas o categóricas se tomará como variable explicativa, para el análisis de regresión logística de MAT y LMA.

El siguiente paso fue la selección de variables hacia adelante y hacia atras considerando las siguientes: Carrera (MAT y LMA), ME, CI, CDV, CIV, AÑOS, PROM1, PROM2, PROM3, PROMAC, PROMPREP, PUNTAJE y SEXO en SPSS, para MAT y LMA. Las variables que resultaron en el modelo final fueron ME, CIV y PROMAC que como puede observarse tienen alta significancia (Tabla 3.11).

Tabla 3.11: Variables en la ecuación; salida de SPSS

Variables	$\hat{\beta}$	ET	Wald	gl	Sig	Exp($\hat{\beta}$)
ME	-.396	.179	4.872	1	.029	0.673
CIV	.344	.144	5.682	1	.017	1.410
PROMAC	2.695	.587	21.060	1	.000	14.800
Constante	-21.074	4.386	23.085	1	.000	0.000

Razón de Momios (RM): Matemáticas y Matemáticas Aplicadas:

La razón de momios se obtuvo exponenciando la estimación del parámetro y corresponde a la columna $Exp(\hat{\beta})$ en la Tabla 3.11 por lo tanto:

- La RM para ME es 0.673, lo que indica que un incremento de la calificación de ME disminuye la probabilidad de que un alumno egrese de alguna de las licenciaturas (MAT y LMA).
- La RM para CIV es 1.416, esto indica que un incremento de la calificación de CIV incrementa la probabilidad de que un alumno egrese de alguna de las lilcenciaturas (LMA y MAT).
- La RM para PROMAC es 14.8, lo que indica que un incremento en el PROMAC incrementa la probabilidad de que un alumno egrese de alguna de las licenciaturas (LMA y MAT).

Modelo de regresión logística ajustado:

$$\hat{\pi}(x_1, x_2, x_3) = \frac{exp(-21.074 - 0.396x_1 + 0.344x_2 + 2.695x_3)}{1 + exp(-21.074 - 0.396x_1 + 0.344x_2 + 2.695x_3)}.$$

donde
$x_1 = ME, x_2 = CIV, x_3 = PROMAC$

Bondad de Ajuste
Prueba de Hosmer y Lemeshow

En el modelo obtenido se debe verificar que efectivamente este modelo se ajusta correctamente a los datos. En la Tabla 3.12 tenemos los valores calculados del estadístico de Hosmer y Lemeshow de bondad de ajuste para el modelo de MAT y LMA el c valor es igual a 8.22, con un p valor de 0.394, y por tanto podemos concluir que no se rechaza la hipótesis nula,ver Tabla 3.12.

Tabla 3.12: Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Significancia
3	8.420	8	.394

Tabla de Clasificación

En la Tabla 3.13 observamos la Tabla de Clasificación para MAT y LMA tenemos que: 40% de los alumnos no egresados se pronosticaron correctamente y 60% se clasificaron incorrectamente. El 96% de los alumnos egresados se pronosticaron correctamente, y por tanto esto nos da un 85 de porcentaje global pronosticado correctamente. Con lo cual el modelo se puede considerar bueno.

Tabla 3.13: Tabla de clasificación de valores observados y pronosticados

	Observado	Pronosticado	
	0	1	Porcentaje Correcto
0	20	30	40.0%
1	8	202	96.2 %
Porcentaje global			85.4%

El valor de corte es .500

Curva COR

SPSS permite probar la hipótesis de que $A = 0.5$ contra $A \neq 0.5$ (A= Área bajo la curca COR), esto es, que el modelo no tiene capacidad discriminante contra de que si la tiene. En la Tabla 3.14 vemos que $A = 0.776$ con intervalo de confianza del 95% de 0.713 a 0.839 y un valor $p = 0.000$ por tanto el modelo tiene capacidad discriminante,

es decir clasificará de manera correcta a los alumnos de MAT y LMA. Aún más, el modelo tiene una capacidad discriminante aceptable.

Tabla 3.14: Área bajo la curva COR

Variables resultado de contraste: Probabilidad pronosticada

Área	Error		Límite inferior	Límite superior
.776	.032	0.000	.713	.839

a. Bajo el supuesto no parámetrico

b. Ho: área verdadera = 0.5

En la Figura 3.1 se puede observar la gráfica de la curva COR para los alumnos de Matemáticas y Matemáticas Aplicadas.

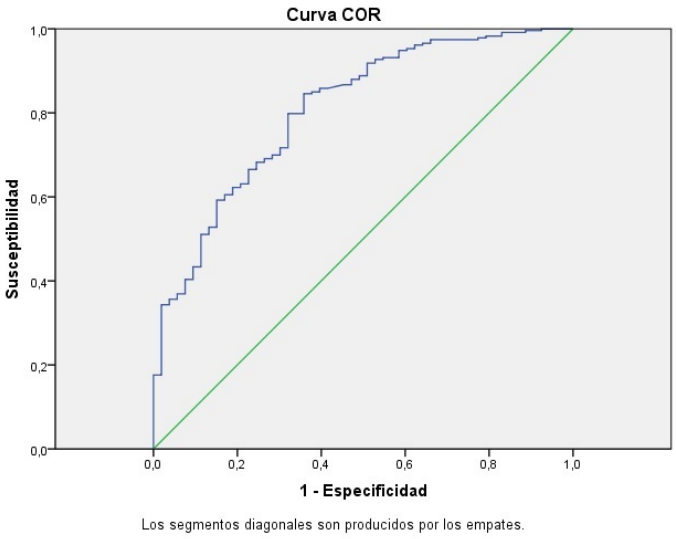

Figura 3.1: Área bajo la curva COR

Ésta gráfica se obtuvo de los pares ordenados $1 - Especifidad - Sensibilidad$ (i.e., la sensibilidad en función de los falsos positivos o complemento de la especifidad) obtenidos. Entre más alejada esté la curva COR (en azul) de la identidad (en verde), el modelo ajustado será mejor.

Análisis de los supuestos del modelo

En cuanto a la incorrecta elección de las variables se puede decir que las variables que se consideraron se basa en las observaciones de los estudios realizados anteriormente. En cuanto al punto 2 es poco probable que pase.

No se realizó el análisis de la normalidad de los residuos pues se ha encontrado que esto no afecta la interpretación de los modelos. Se analizó la existencia de colinealidad entre las variables independientes del modelo final usando sus correlaciones yvel número de condición escalado (NCE). Se encontró que hay colinealidad severa (NCE=60) y correlación moderada entre las variables Promedio Final y Matemáticas Elementales (r=0.50) y entre Promedio Final y Cálculo Integral en varias Variables (r=0.44), pero las pruebas de bondad de ajuste nos indican que los modelos obtenidos son buenos para explicar la probabilidad de egreso de los alumnos de MAT-LMA, y de FIS y LFA.

Interpretación de los coeficientes

La interpretación de los coeficientes en un modelo de regresión logística comúnmente se realiza mediante cambios en la escala logit, a través de razones de odds condicionales. Sin embargo dichos coeficientes no tienen una interpretación directa en términos de probabilidad, que podría considerarse como la más frecuente para la mayoría de los usuarios [19].

Otra forma de presentar los resultados del modelo de regresión logística es a través de una estimación ajustada de las probabilidades asociadas a cada covariable. El cálculo de probabilidades ajustadas puede ser mas conveniente para una persona no experta en las nociones de estimación de parámetros de regresión y probabilidad asociada [18].

El signo de los coeficientes estimados del modelo de regresión logística en la Tabla 3.11 dan una explicación de las variables predictoras usadas, luego los signos de CIV y PROMAC son positivos y por consiguiente producen un crecimiento en la probabilidad para egresar de alguna de las licenciaturas de Matemáticas o Matemáticas Aplicadas. En cambio el signo de la variable ME es negativo con lo cual produce

un decrecimiento en la misma probabilidad.

En las siguientes Figuras se muestran las proyecciones del comportamiento del modelo de regresión logística ajustado. Las gráficas se realizaron en el programa R [29]. En la Figura 3.2 se observan seis gráficas, estas corresponden a las proyecciones del modelo ajustado para MAT y LMA, realizadas de la siguente forma, en las primeras tres la variable CIV tiene el valor fijo 10 y PROMAC toma los valores 9, 7 y 5 respectivamente. En las siguentes tres PROMAC tiene el valor fijo 9 y ME los valores 9, 8 y 6 respectivamente. En las seis gráficas la variable ME varía de 5 a 10.

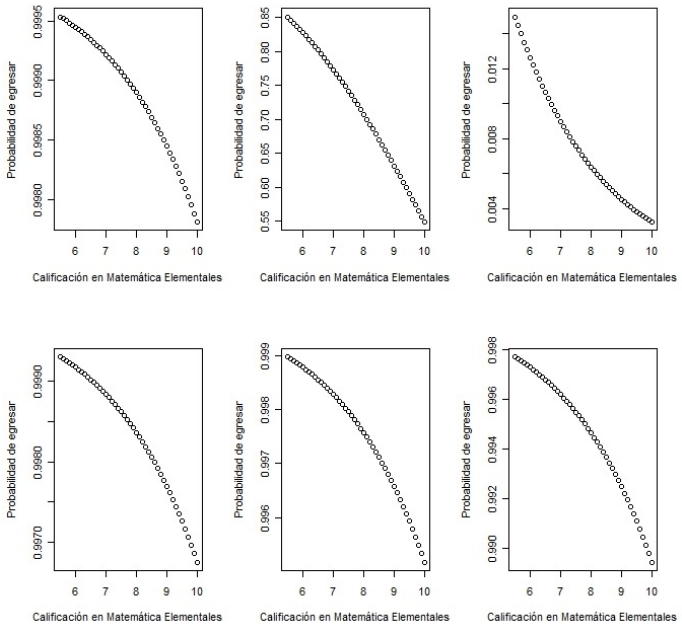

Figura 3.2: Probabilidad de egresar

Se observa que la probabilidad de egresar disminuye notablemente conforme la calificación de ME aumenta y la calificación de PROMAC se mantine fija, ésto debido a que el coeficiente de ME en el modelo ajustado es menor a 1. En cambio si PROMAC se mantiene con el valor 9 y la calificación de CIV va disminuyendo la probabilidad no decrece en gran medida.

En ésta Figura (3.3), se observan las gráficas realizadas de forma similar a la Figura anterior, para éstas la variable CIV varía de 5 a 10, mientras que las otras variables toman los valores $ME = 10$ en las primeres tres y PROMAC los valores 9, 7 y 5 respectivamente. En las siguentes PROMAC toma el valor 9 y ME los valores $9, 7$ y 5.

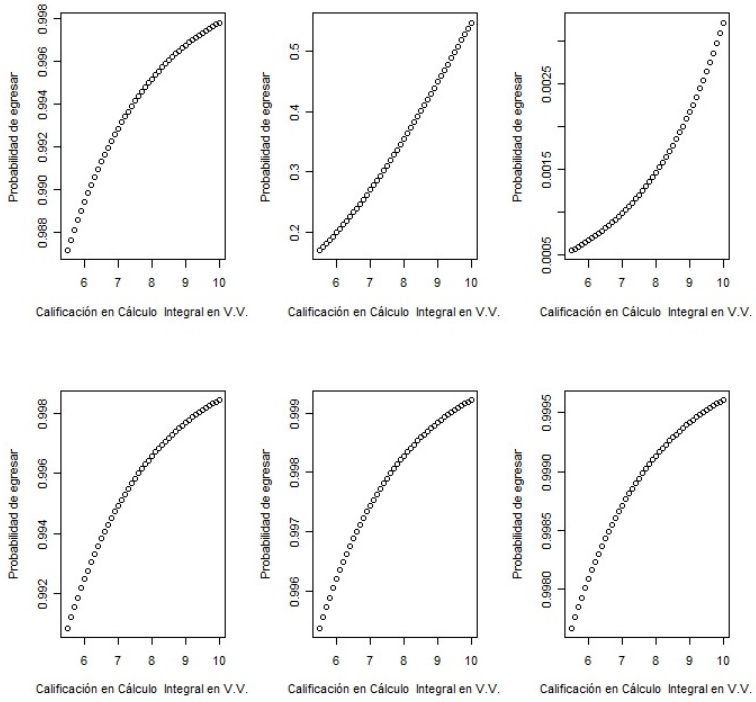

Figura 3.3: Probabilidad de egresar

Para estas gráficas se observa que la probabilidad de egresar disminuye conforme PROMAC decrece, ésta probabilidad pasa del rango de 0.98-0.99 en la primera al rango de 0.000- 0.002 en la tercera. En cambio en las gráficas de abajo la probabilidad de egresar no cambia si PROMAC se mantiene con el valor fijo de 9.

En la Figura 3.4, (realizada de forma semejante a las anteriores, en las tres primeras $ME = 10$y $CIV = 10, 8, 6$ y en las siguientes $CIV = 10$ y $ME = 9, 7, 5$.) se observa

más claro la naturaleza dicotómica de la variable probabilidad de egreso, la cual toma el valor cero si un alumno no egresa de alguna de las licenciaturas MAT o LMA o el valor 1 si egresa.

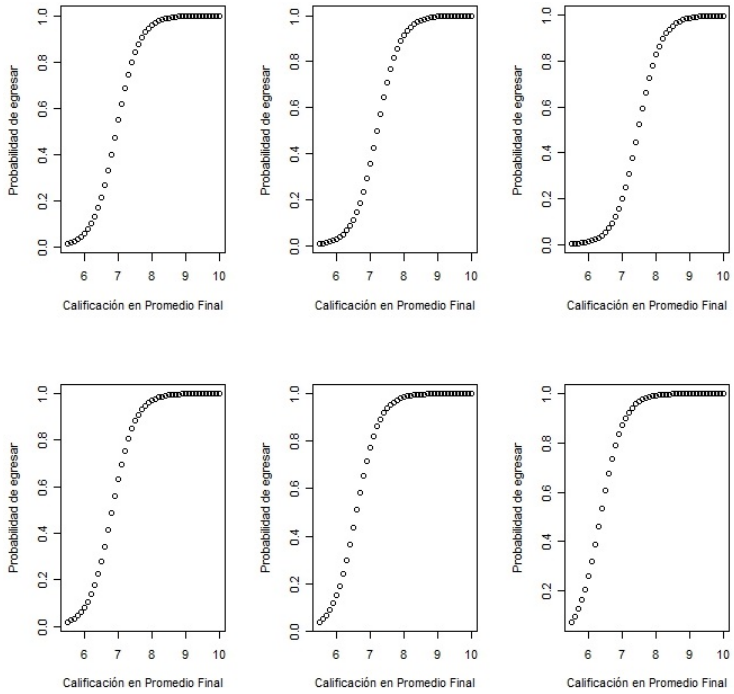

Figura 3.4: Probabilidad de egresar

En estas gráficas se observa más claramente que la variable PROMAC es la que más afecta el comportamiento de la probalibilidad de egresar

3.2 Modelación del egreso en las carreras de Física y Física Aplicada

Para FIS y LFA, las variables que se consideraron para la selección fueron Carrera (FIS y LFA), CI, CDV, CIV, PROM1, PROM2, PROM3, PROM4, PROM5, PROM6, PROM7, PROM8, PROM9, PROM10, PROMAC, PROMPREP, PUNTAJE y SEXO. Como resultado de la selección, las variables CIV y PROM9 quedaron en el modelo

final, cabe decir que en este caso se realizó dos veces más la selección de variables, en la primera con las seis variables CIV, PUNTAJE, PROM2, PROM5, PROM7, PROM9. Después de esta selección quedaron sólo las variables: CIV, PUNTAJE, PROM5, PROM7 y PROM9; de donde finalmente las variables CIV y PROM9 quedaron en el modelo final (ver Tabla 3.15).

Para los alumnos de FIS y LFA, ninguna de las variables cuantitativas resultó excluida en este análisis, y para las variables cualitativas se obtuvó un resultado similar al anterior, esto es ninguna de las variables Carrera, Sexo o Periodo, se tomaran como variables explicativas .

Razón de Momios (RM): Física y Física Aplicada :

- La RM para CIV es 1.372, lo que indica que un incremento de la calificación de CIV incrementa la probabilidad de que un alumno egrese de alguna de las licenciaturas (FIS y LFA).
- La RM para PROM9 es 0.949, lo que indica que un incremento de la calificación de PROM9 incrementa la probabilidad de que un alumno egrese de alguna de las licenciaturas (FIS y LFA).

Tabla 3.15: Variables en la ecuación; salida de SPSS

Variables	$\hat{\beta}$	ET	Wald	gl	Sig	Exp($\hat{\beta}$)
CIV	.317	.140	5.123	1	.024	1.372
Prom 9	2.393	.479	24.926	1	.000	10.949
Constante	-21.040	3.818	30.368	1	.000	0.000

Modelo de regresión logística ajustado:

$$\hat{\pi}(x_1, x_2) = \frac{exp(-21.040 + 0.317x_1 + 2.393x_2)}{1 + exp(-21.040 + 0.317x_1 + 2.393x_2)}.$$

donde

$x_1 = CIV, x_2 = PROM9$

Bondad de Ajuste
Prueba de Hosmer y Lemeshow

Similarmente al análisis hecho para MAT y LMA tenemos que: el valor del estadístico de Hosmer y Lemeshow para FIS y LFA es $c = 4.963$, con p valor de 0.762, y así podemos concluir que tampoco se rechaza la hipóteis nula (ver Tabla 3.16).

Tabla 3.16: Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Significancia
1	4.963	8	.762

Tabla de Clasificación

Ahora en la Tabla 3.17 observamos la Tabla de Clasificación para FIS y LFA, obtenemos que: 29.5% de los alumnos no egresados se pronosticaron correctamente y un 98.4% de los alumnos egresados se pronosticaron correctamente, por lo tanto se obtiene un 88% global de alumnos pronosticados correctamente. Con lo cual el modelo bueno, incluso un poco mejor (en esta prueba) que el modelo para MAT y LMA.

Tabla 3.17: Tabla de Clasificación de valores observados y pronosticados

	Observado	Pronosticado	
	0	1	Porcentaje Correcto
0	13	31	29.5%
1	4	247	98.4 %
Porcentaje global			88.1%

El valor de corte es .500

Curva COR

En la Tabla 3.18 vemos que $A = 0.818$, con un intervalo del 95% de confianza de 0.751 a 0.884, con lo cual se rechaza la hipótesis Ho: $A = 0.5$, por lo tanto el modelo para los alumnos de Física y Física Aplicada tiene alta capacidad discriminante.

Tabla 3.18: Área bajo la curva COR

Variables resultado de contraste: Probabilidad pronosticada

Área	Error			
			Límite inferior	Límite superior
.818	.034	0.000	.751	.884

a. Bajo el supuesto no parámetrico

b. Ho: área verdadera = 0.5

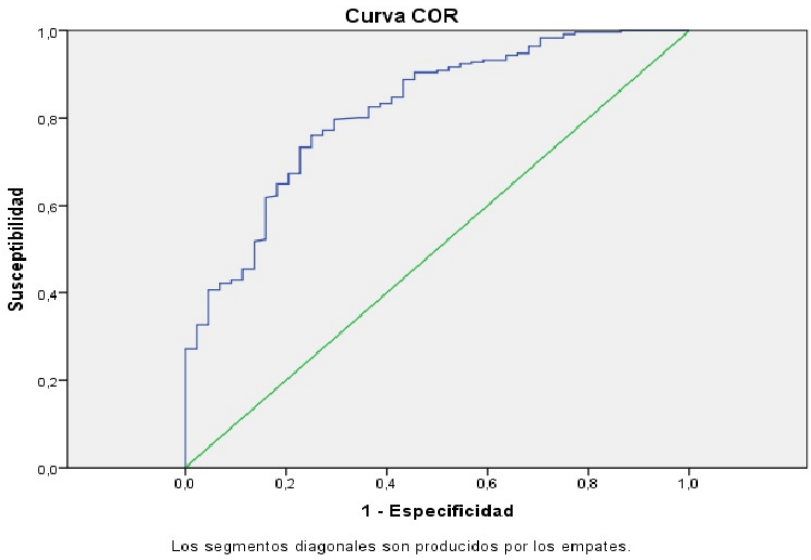

Figura 3.5: Área bajo la curva COR

Análisis de los supuestos del modelo

Para FIS y LFA se realizaron análisis semejantes que para MAT y LMA, en los cuales se obtuvieron resultados similares en cuanto a que se observa colinealidad en los datos, pero las pruebas de bondad de ajuste nos indican que se tiene un buen modelo.

Interpretación de los coeficientes

En cuanto a la probabilidad de egresar de alguna de las licenciaturas de Física y Física Aplicada, los coeficientes de las variables que quedaron en el modelo final son positivos y por tanto producen un crecimiento en la probabilidad de egresar.

En las siguientes Figuras se muestran las proyecciones del comportamiento del modelo de regresión logística ajustado para los alimnos de FIS y LFA. En este caso fue posible realizar las 6 gráficas cuando alguna de las variables CIV o PROM9 varían de 10 a 5. La primera muestra el comportamiento del probabilidad de egresar cuando la calificación de PROM9 varía de 10 a 5. En la segunda la variable de varía es CIV de 10 a 5.

En la Figura 3.6 observamos las seis gráficas que resultan de asignar las calificaciones de 10 a 5 a la variable CIV, vemos que conforme aumenta la calificación del noveno semestre y la de CIV, aumenta la probabilidad de egresar de FIS o LFA, por ejemplo en la primera ésta probabilidad es igual a 0.5 si PROM9 es 8, en cambio en la sexta gráfica la probabilidad es 0.2 si PROM9 es 8, este cambio se debe a que CIV cambia de 10 a 5.

En la Figura 3.7 se observan las gráficas que resultan de variar PROM9 de 10 a 6, en estas la probabilidad disminuye de manera evidente si la calificación del noveno semestre también disminuye, si se observa la primera gráfica la probabilidad de egresar va de 0.9992 a 0.9996, en cambio en la sexta va de 0.0010 a 0.0025.

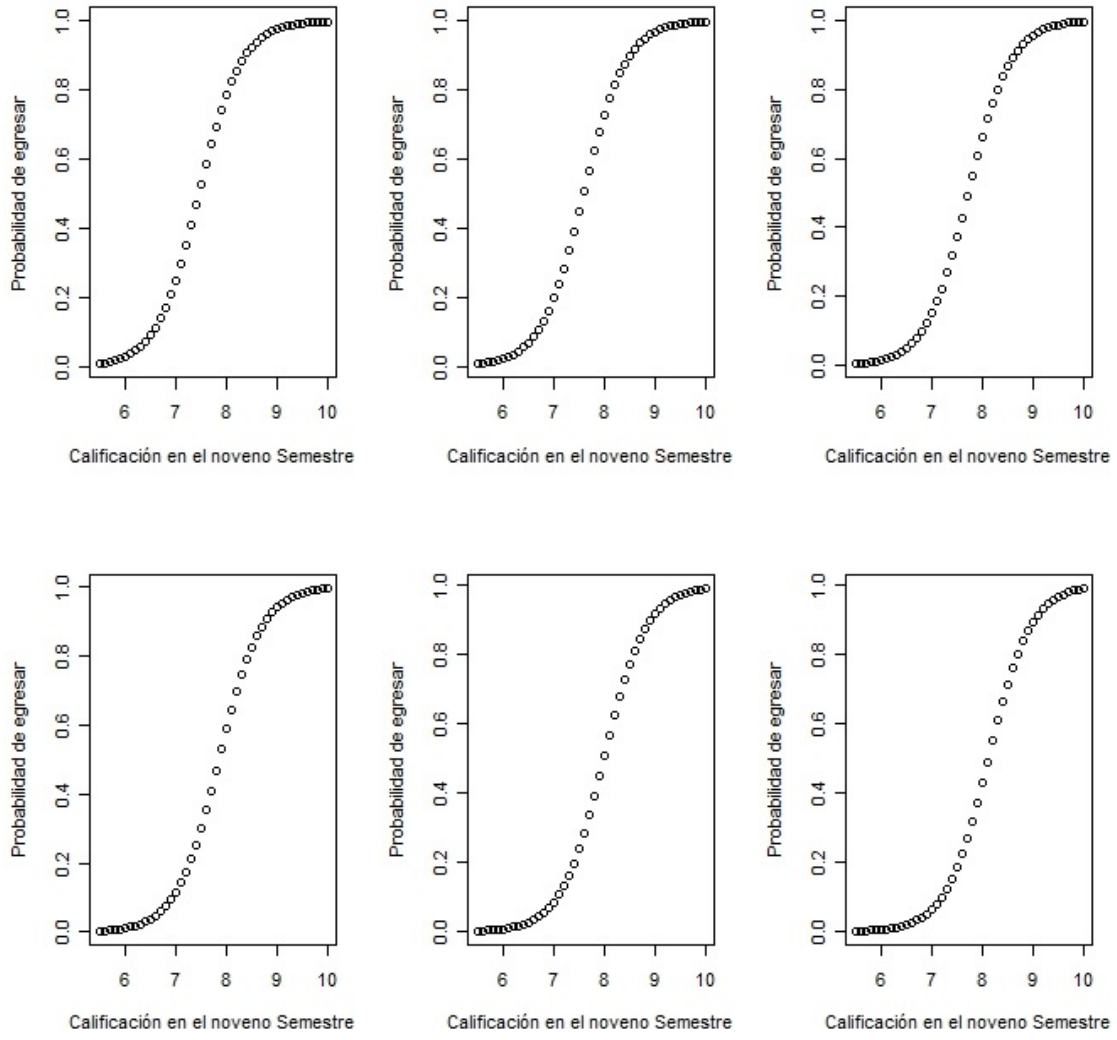

Figura 3.6: Probabilidad de egresar, $CIV = 10, 9, 8, 7, 6, 5$

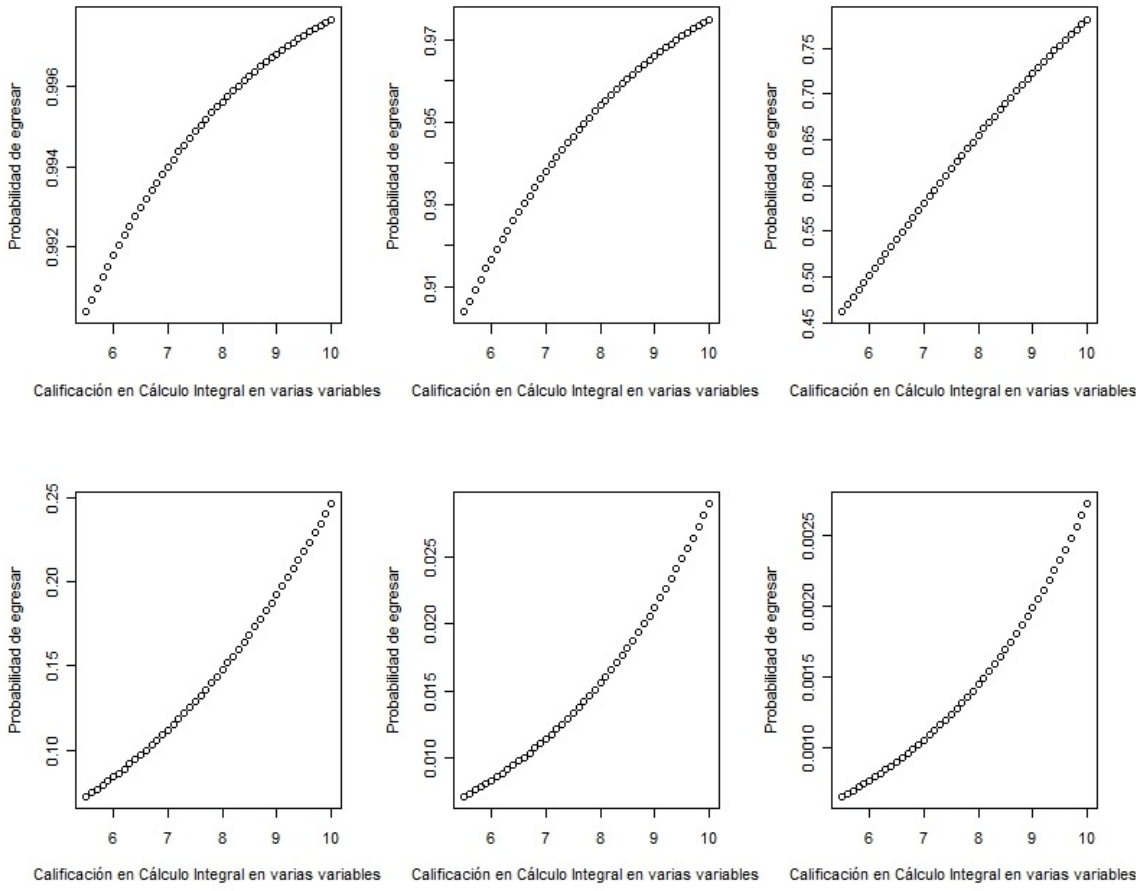

Figura 3.7: Probabilidad de egresar, $PROM9 = 10, 9, 8, 7, 6, 5$
=8

Capítulo 4

Conclusiones

En este análisis, se estudiaron los factores académicos para que un alumno egrese o no de alguna de las licenciaturas Matemáticas, Matemáticas Aplicadas, Física o Física Aplicada usando el modelo de regresión logística.
Las variables usadas fueron Matemáticas Elementales, Cálculo Diferencial, Cálculo Integral, Cálculo Diferencial en Varias Variables, Cálculo Integral en Varias Variables, Años cursados, Promedio primer año, Promedio segundo año, Promedio tercer año, Promedio final, Promedio Preparatoria o Bachiller, Puntaje examen de admisión, Carrera y Sexo para el caso de Matemáticas y Matemáticas Aplicadas.

Para Física y Física Aplicada los factores usados fueron Matemáticas Elementales, Cálculo Diferencial,Cálculo Integral, Cálculo Diferencial en Varias Variables, Cálculo Integral en Varias Variables, Años cursados, Promedio primer semestre, Promedio segundo semestre, Promedio tercer semestre, Promedio cuarto semestre, Promedio quinto semestre, Promedio sexto semestre, Promedio séptimo semestre, Promedio octavo semestre, Promedio noveno semestre, Promedio décimo semestre, Promedio final, Periodo, Promedio Preparatoria o Bachiller, Puntaje examen de admisión, Carrera y Sexo.

El análisis concluye que las materias Matemáticas Elementales y Cálculo Integral en Varias Variables producen un crecimiento positivo en la probabilidad de que un alumno egrese de las licenciaturas de Matemáticas o Matemáticas Aplicadas. Mientras la variable Promedio Actual aumenta de manera muy significativa esta probabilidad.

Y en el caso de las licenciaturas de Física y Física Aplicada Cálculo Integral en

Varias Variables y Promedio del noveno semestre son los factores que aumentan significativamente en ésta probabilidad. En este análisis también se pudo observar la baja eficiencia terminal en los dos grupos de licenciaturas, la cual es mayor en el caso de los alumnos de Matemťicas y Matemáticas Aplicadas. Observamos que aún sigue habiendo mayoría en cuanto a hombres y mujeres sobre todo en las licenciaturas de Física y Física aplicada.

El modelo de regresión logística es usado para estimar sí un alumno egresa o no, las pruebas de bondad de ajuste nos indican que el modelo obtenido es un buen modelo para poder clasificar u obtener las probabilidades de los alumnos para egresar de las licenciaturas analizadas.

En las gráficas realizadas se observan como cambian las respectivas probabilidades de egresar de alguna de las licenciaturas.

Apéndice A

Comparación de medias

Prueba t para dos muestras independientes.

Esta opción debe utilizarse cuando la comparación se realice entre las medias de dos poblaciones independientes (los individuos de una de las poblaciones son distintos a los individuos de la otra) como, por ejemplo, en el caso de la comparación de las poblaciones de hombres y mujeres. Por lo tanto, compara las medias de una variable para dos grupos de casos. La matriz de datos debe estar configurada como es habitual, es decir, existe una columna para los datos de la variable de interés y una segunda columna con los códigos que definen las poblaciones objeto de comparación. La prueba calcula estadísticos descriptivos para cada grupo además de la prueba de Levene para la igualdad de varianzas, así como los valores de t para varianzas iguales y desiguales y el intervalo de confianza del 95% para la diferencia de medias véase [32].

El contraste de hipótesis para muestras independientes divide los casos en dos grupos y compara las medias de los grupos respecto a una variable. En una situación ideal los sujetos deberían asignarse aleatoriamente a los grupos, de forma que cualquier diferencia pueda atribuirse al efecto del tratamiento y no a otros factores. Dicho de otro modo, debe asegurarse que las diferencias en otros factores no enmascaren o resalten una diferencia significativa entre las medias. El SPSS (véase [12]) permite introducir más de una variable de contraste y calcula una prueba t para cada variable. En cambio, la variable de agrupación solamente puede ser una y requiere definir los grupos que se desee comparar. Los grupos de la variable de agrupación se pueden definir de dos formas: a) mediante valores especificados (se escribe un valor para el Grupo 1 y otro para el Grupo 2, quedando los casos con otros valores

excluidos del análisis); o b) con un punto de corte (se establece un número que divida los valores de la variable de agrupación en dos partes. Los casos con valores menores que el punto de corte forman un grupo y los casos con valores mayores o iguales que el punto de corte forman el otro grupo). Antes de analizar los resultados del contraste de la diferencia de medias, es conveniente detenerse para valorar la comparación de las varianzas de ambos grupos (basándose en el estadístico F de Snedecor) a través de la prueba de Levene. La prueba de Levene debe arrojar una significación mayor de $0,05$ para que se cumpla el requisito de homocedasticidad (expresado en la Tabla como se han asumido varianzas iguales a través del estadístico F).

A.1 Test de Levene

En estadística, el test de Levene es una estadística inferencial utilizado para evaluar la igualdad de las varianzas para una variable calculada para dos o más grupos. Algunos procedimientos estadísticos comunes asumen que las varianzas de las poblaciones de las que se extraen diferentes muestras son iguales. La Prueba de Levene evalúa este supuesto. Pone a prueba la hipótesis nula de que las varianzas poblacionales son iguales (llamado homogeneidad de varianza o homocedasticidad). Si el P-valor resultante de la prueba de Levene es inferior a un cierto nivel de significación (típicamente 0,05), las diferencias obtenidas en las variaciones de la muestra no es probable que se han producido sobre la base de un muestreo aleatorio de una población con varianzas iguales. Por lo tanto, la hipótesis nula de igualdad de varianzas se rechaza y se concluye que hay una diferencia entre las variaciones en la población. Algunos de los procedimientos que asumen normalmente homocedasticidad, para lo cual uno puede utilizar las pruebas de Levene, incluyen análisis de varianza y pruebas t. Cuando la prueba de Levene muestra significación, se debe cambiar a pruebas generalizadas, libre de supuestos homocedasticidad. El test de Levene también puede ser utilizado como una prueba principal para responder a una pregunta independiente de si dos sub-muestras en una población dada tienen varianzas iguales o diferentes.

Apéndice B

Términos Estadísticos

Definición: *Una estadística es cualquier función de las variables aleatorias que se observaron en la muestra de manera que esta función no contiene cantidades desconocidas*. Si se emplea una estadística T para estimar un parámetro desconocido θ, entonces T recibe el nombre **estimación** de θ. Esto es, un estimador es una estadística que identifica el mecanismo funcional por medio del cual, una vez que las observaciones en la muestra se realizan se obtiene una estimación (véase [5]).

B.1 Tablas de Contingencia

Una Tabla de contingencia es una de las formas más comunes de resumir datos categóricos. Es decir, el interés se centra en estudiar si existe alguna asociación entre una variable fila y otra variable columna y/o calcular la cantidad de dicha asociación.
Sean X e Y dos variables categóricas con I y J categórias, respectivamente. Un sujeto puede venir clasificado en una de las $I \times J$ categorías, que es el número posible de categorías que existe.

Cuando las casillas de la Tabla contienen las frecuencias observadas, la Tabla se denomina **Tabla de Contingencia**, término que fue introducido por Pearson en 1904. Una Tabla de contingencia (o Tabla de Clasificación) con I filas y J columnas se denomina Tabla $I \times J$.
La distribución conjunta de dos variables categóricas determina su relación. Esta distribución también determina las distribuciones marginales y condicionales [1].

B.1.1 Distribución Conjunta

La distribución conjunta viene dada por:

$$\pi_{ij} = P(X = i, Y = j) \quad \text{con} \quad i = 1, 2, \ldots, I \quad \text{y} \quad j = 1, \ldots, J. \tag{B.1}$$

Es la probabilidad de (X, Y) en la casilla de la fila i y la columna j.

B.1.2 Distribución Marginal

Las distribuciónes marginales son los totales de los renglones y columnas obtenidos por la suma de las probabilidades conjuntas, estas son:

$$\begin{aligned}
\pi_{i+} = P(X = i) &= \sum_{i=1}^{J} P(X = i, Y = j) \\
&= \sum_{j=1}^{J} \pi_{ij} \\
\pi_{+j} = P(Y = j) &= \sum_{i=1}^{I} P(X = i, Y = j) \\
&= \sum_{i=1}^{I} \pi_{ij}.
\end{aligned}$$

Es la probabilidad marginal o probabilidad de Y por última columna (ver Tabla B.1). Es decir, el simbolo + indica la suma de las casillas correspondientes, a un índice dado. Estas expresiones cumplen que la suma sobre todos sus índices, π_{++} vale uno. Se verifica que

$$\sum_i \pi_{+j} = \sum_i \pi_{i+} = \sum_i \sum_j \pi_{ij} = 1. \tag{B.2}$$

Las distribuciones marginales son sólo variables de información, y no pertenecen a los vínculos de asociación entre las variables.

Tabla B.1: Tabla de Contingencia de orden $I \times J$

	Columnas Y				
Renglón $R(X)$	1	2	...	J	Total
1	π_{11}	π_{12}	...	π_{1j}	π_{1+}
2	π_{21}	π_{22}	...	π_{2j}	π_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
i	π_{i1}	π_{i2}		π_{ij}	π_{i+}
⋮	⋮	⋮	⋮	⋮	⋮
I	π_{I1}	π_{I2}	...	π_{Ij}	π_{I+}
Total	π_{+1}	π_{+2}	...	π_{+J}	π_{++}=π

B.1.3 Distribución Condicional

En la mayor parte de las Tablas de contingencia, una de las variables, digamos Y es una variable respuesta y la otra variable X es una variable explicatoria o predictiva. En esta situación no tiene sentido hablar de distribución conjunta. Cuando se considera una categoría fija de X, entonces Y tiene una distribución de probabilidad que se expresa como una probabilidad condicionada. Así se puede estudiar el cambio de esta distribución cuando van cambiando los valores de X.
Distribución condicionada de Y respecto de X:

$$P(Y = j \mid X = i) = \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}. \tag{B.3}$$

Se tiene que

$$\sum_j \pi_{j|i} = 1. \tag{B.4}$$

Y el vector de probabilidades:

$$(\pi_{1|i}, \pi_{2|i}, \ldots, \pi_{j|i}) \tag{B.5}$$

forma la distribución condicional de Y en la categoría i de X.

La mayor parte de los estudios se centran en la comparación de las distribuciones condicionadas de Y para varios niveles de las variables explicativas.

B.1.4 Independencia y Homogeneidad

Cuando las variables que se consideran son de tipo respuesta, se pueden usar distribuciones conjuntas o bien distribuciones condicionales para describir la asociación entre ellas. Dos variables son independientes sí:

$$\pi_{ij} = \pi_{i+} \cdot \pi_{+j}. \tag{B.6}$$

Lo cual implica que la distribución condicionada es igual a la marginal, es decir:

$$\pi_{j|i} = \pi{+}j \quad \text{para} j = 1, \ldots, J \quad \text{dado que} \quad \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \text{para todo i y j.} \tag{B.7}$$

Si X e Y son variables de respuesta entonces se habla de independencia. Si Y es variable respuesta y X es variable explicativa entonces se habla de homogeneidad. Cada distribución condicional de Y es idéntica a la marginal de Y siempre y cuando exista independencia. Así, dos variables son independientes cuando la probabilidad de la columna respuesta j es la misma en cada fila, para $j = 1, \ldots, J$.

La Tabla B.2 muestra la notación para las distribuciones conjuntas, marginales y condicionales para el caso 2×2. El método de notación es similar para distribuciones muestrales, con la letra P en lugar de π. Por ejemplo, $\{P_{ij}\}$ denota la distribución conjunta muestral en una Tabla de contingencia. Las frecuencias son denotadas por $\{n_{ij}\}$, con $n = \sum_i \sum_j n_{ij}$, el tamaño total de la muestra por lo que:

$$P_{ij} = \frac{n_{ij}}{n_{i+}} \tag{B.8}$$

la proporción de veces que los sujetos en la fila i dieron la respuesta j es

$$P_{i|j} = \frac{P_{ij}}{P_{i+}} = \frac{n_{ij}}{n_{i+}} \text{donde} \quad n_{i+} = nP_{i+} = \sum_j n_{ij}. \tag{B.9}$$

Tabla B.2: Probabilidades Conjuntas, Codicionales y Marginales

	Columna		
Fila	1	2	Total
1	π_{11}	π_{12}	π_{1+}
	($\pi_{1\|1}$)	($\pi_{1\|2}$)	(1.0)
2	π_{21}	π_{22}	π_{2+}
	($\pi_{2\|1}$)	($\pi_{2\|2}$)	(1.0)
Total	π_{+1}	π_{+2}	1.0

B.1.5 Maneras de Comparar Proporciones

Las variables de respuesta que tienen dos categorías son llamadas binarias. Los estudios son frecuentemente comparados en varios grupos sobre una respuesta binaria, Y. Cuando hay I grupos, se muestran los resultados en una tabla de Contingencia de orden $I \times 2$, en la cual las columnas son los niveles de Y.

Diferencia de Proporciones

En la columna i, $i = 1, \ldots, I$, $\pi_{1|i}$ es la probabilidad de la respuesta 1 y $(\pi_{1|i}, \pi_{2|i}) = (\pi_{1|i}, 1 - \pi_{1|i})$ es la distribución condicional de la respuesta binaria. Podemos comparar dos columnas digamos h e i, usando la **diferencia de proporciones**, $\pi_{1|h} - \pi_{1|i}$. Comparar sobre la respuesta 2 es equivalente de comparar sobre la respuesta 1, donde:

$$\pi_{2|h} - \pi_{2|i} = (1 - \pi_{1|h}) - (1 - \pi_{1|i}) = \pi_{1|i} - \pi_{1|h}. \tag{B.10}$$

La diferencia de proporciones esta entre -1.0 y $+1.0$.Esta es igual a cero cuando las columnas h y i tienen idéntica distribución condicional. La respuesta Y es estadísticamente independiente de la clasificación de columnas cuando $\pi_{1|h} - \pi_{1|i} = 0$ para

todos los pares de las columnas h e i.

Para Tablas de Contingencia de orden $I \times J$, podemos comparar la probabilidad condicional de la respuesta j para las columnas h e i usando la diferencia $\pi_{j|h} - \pi_{j|i}$. Las variables son independientes cuando ésta diferencia es igual a cero para todos los pares de columnas h e i y todas las posibles respuestas j; equivalentemente, cuando las $(I-1)(J-1)$ diferencias $\pi_{j|h} - \pi_{j|I} = 0, i = 1, \ldots, I-1, j = 1, \ldots, J-1$. Cuando ambas variables son respuesta y hay una distribución conjunta (π_{ij}), la comparación de proporciones dentro de las columnas h e i satisface:

$$\pi_{1|h} - \pi_{1|i} = \pi_{h1}/\pi_{h+} - \pi_{i1}/\pi_{i+}. \tag{B.11}$$

B.1.6 Riesgo Relativo

Una diferencia en proporciones de tamaño fijo tiene más importancia cuando las proporciones están entre 0 y 1, que cuando están cerca de la mitad del rango. En tales casos, el radio de proporciones es también una medida de uso descriptivo.
Para Tablas de 2×2, el **riesgo relativo** es el radio:

$$\pi_{1|1}/\pi_{1|2}. \tag{B.12}$$

Este radio puede ser cualquier número real no negativo. Un riesgo relativo de 1.0 corresponde a independencia. Comparando sobre la segunda respuesta se obtiene un riesgo relativo diferente,

$$\pi_{2|1}/\pi_{2|2} = (1 - \pi_{1|1})(1 - \pi_{1|2}). \tag{B.13}$$

B.1.7 Odds Ratio

En referencia a la Tabla de 2×2 (Tabla B.2). Dentro de la columna 1, el **odds** que la respuesta esta en la columna 1 en lugar de la columna 2 está definido por:

$$\Omega_1 = \pi_{1|1}/\pi_{2|1}. \tag{B.14}$$

Dentro de la fila 2, el correspondiente odds es igual a:

$$\Omega_2 = \pi_{1|2}/\pi_{2|2}. \tag{B.15}$$

Para distribuciones conjuntas, la definición equivalente es:

$$\Omega_i = \pi_{i1}/\pi_{i2}, \qquad i = 1, 2. \tag{B.16}$$

Cada Ω_i es no negativo, con valor mayor que 1.0 cuando la respuesta es mas probable que la respuesta 2. Por ejemplo, cuando $\Omega_1 = 4.0$, en la primera fila, la respuesta 1 es 4 veces más probable que la respuesta 2. La distribuciones condicionales dentro de la fila son idénticas, por lo tanto las variables son independientes, si y sólo si $\Omega_1 = \Omega_2$.
El radio de los odds Ω_1 y Ω_2:

$$\theta = \Omega_1/\Omega_2 \tag{B.17}$$

es llamado el **odds ratio**. De la definición de odds usamos la probabilidad conjunta,

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{B.18}$$

El odds ratio puede ser igual a cualquier número no negativo. Cuando todas las probabilidades de las celdas son positivas, la independencia de X y Y es equivalente a $\theta = 1$. Cuando $1 < \theta < \infty$, los sujetos en la fila 1 son más probables a contestar la primera respuesta que los que están en la fila 2; esto es, $\pi_{1|1} > \pi_{1|2}$. Si $0 < \theta < 1$, la primera respuesta es menos probable en la fila 1 que en la fila 2; es decir, $\pi_{1|1} < \pi_{1|2}$. Cuando una celda tiene probabilidad cero, θ es igual a 0 o ∞.

Apéndice C

SPSS: Métodos de Selección de variables en SPSS

Existen distintos métodos según se utilicen todas las variables sin eliminar las no significativas, se seleccionen las variables hacia adelante (es decir, se van incluyendo en el modelo las variables más significativas hasta que todas la que no han sido seleccionadas no son significativas) o se seleccionan hacia atrás (es decir, se incluyen en el modelo todas las variables y se van eliminando las menos significativas y así, sucesivamente, hasta que todas las variables en el modelo sean significativas). Los métodos de introducción posibles son:

Introducir. Procedimiento para la selección de variables en el que todas las variables de un bloque se introducen en un solo paso. Incluye todas las variables aunque no sean significativas.

Adelante: Condicional. Contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación de acuerdo a la probabilidad de un estadístico de la razón de verosimilitud que se basa en estimaciones condicionales de los parámetros.

Adelante: RV. Contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación en relación al estadístico de la razón de verosimilitud, que se basa en estimaciones de la máxima verosimilitud parcial.

Adelante: Wald. Método de selección por pasos hacia adelante que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del estadístico de Wald. El estadístico de Wald permite una prueba χ^2 para contrastar la hipótesis nula de que el coeficiente de cada variable independiente es cero.

Atrás: Condicional. Selección por pasos hacia atrás. El contraste para la eliminación se basa en la probabilidad del estadístico de la razón de verosimilitud, el cuál se basa a su vez en las estimaciones condicionales de los parámetros.

Atrás: RV. Selección hacia atrás por pasos. El contraste para la eliminación se fundamenta en la probabilidad del estadístico de la razón de verosimilitud, el cual se basa en estimaciones de máxima verosimilitud parcial.

Atrás: Wald. Selecćon por pasos hacia atrás. El contraste para la eliminación se basa en la probabilidad del estadístico de Wald.
Véase [12] para observar los métodos en SPSS.

Apéndice D

Definiciones de las Instituciones Educativas

COPAES: Consejo para la Acreditación de la Educación Superior. Es la instancia capacitada y reconocida por el Gobierno Federal, a través de la SEP, cuyo objetivo es conferir reconocimiento formal a favor de organizaciones cuyo fin sea acreditar programas académicos de Educación Superior que ofrezcan instituciones públicas y partículares, previa valoración de su capacidad organizativa, técnica y operativa, de sus marcos de evaluación para la acreditación de programas académicos, de la administración de sus procedimientos y de la imparcialidad del mismo.

CIEES: Comités Interinstitucionales para la Evaluación de la Educación Superior.

Referencias

[1] Agresti A., *An Introduction to Categorical Data Analysis*, 2a. Edition, Jonh Wiley and Sons, 2007.

[2] Arriaza M.,*Guía práctica de análisis de datos*, Instituto de Investigación y Formación Agraria y Pesquera, 2006.

[3] Aparicio L. E., *Un estudio sobre factores que obstaculizan la permanencia, logro educativo y eficiencia terminal en las áreas de matemáticas del nivel superior: el caso de la Facultad de Matemáticas de la Universidad Autónoma de Yucatán*, 2004.

[4] Bracho T., *Capital cultural: impacto en el rezago educativo, Revista Latinoamericana de Estudios Educativos*, Vol XX, No 2, 1990.

[5] Canavos G. (trad. Urbina E.), *Probabilidad y Estadística: Aplicaciones y métodos*, McGraw-Hill, 1988.

[6] Cantoral R., *Enseñaza de la matemática en la educación superior*, Revista Electrónica Sinectica núm. 19, Instituto Tecnológico y de Educación Superior de Occidente, México, 2001.

[7] Korshunov Y. (trad. Lanier R.), *Fundamentos matemáticos de la cibernética*, Editorial Mir, 1974.

[8] Cruz G. M., *Deserción escolar, factores dependientes de la institución educativa, en estudiantes de la carrera de químico farmacéutico biólogo de la Facultad de Ciencias Químicas de la Universidad Autónoma de Nuevo León*, Tesis de Maestría en Enseñanza Superior, Nuevo León, 2003.

[9] Díaz J., *Análisis Descriptivo de los Egresados y Titulados de las Licenciaturas de Matemáticas y Matemáticas Aplicadas de las Generaciones 2000 a 2004*, Tesis de Licenciatura FCFM-BUAP, 2013.

[10] De la Fuente S., *Regresión Logística, Facultad de Ciencias Económicas y Empresariales*, UAM, Madrid, 2011.

[11] Gonzalez J., Galindo N., Galindo J. & Gold M., *Los paradigmas de la calidad educativa. De la autoevaluación a la acreditación. Unión de Universidades de América Latina*, A.C., 2004.

[12] Guisande C., Barreiro A., Maneiro I., Riveiro I., Vergara A. & Vaamonde A., *Tratamiento de datos*, Ediciones Díaz de Santos, Universidad de Vigo, 2006.

[13] Hernández S., Reyes H. & Linares G., *Análisis Estadístico de algunos factores que afectan el proceso de enseñanza aprendizaje en la FCFM-BUAP, usando técnicas estadísticas multivariadas*, VI Encuentro Participación de la Mujer en la Ciencia, CIO, 2009.

[14] Hernández S., Reyes H., Ibarra M. & Linares G., *Proceso de enseñanza aprendizaje*, VII encuentro Participación de la mujer en la ciencia, CIO, 2010.

[15] Hernández, S. *Uso del modelo de Regresión Logística para estudiar la aprobación de la materia de Matemáticas Básicas de la FCFM en las generaciones 2010 y 2011*, Tesis de licenciatura, México, 2013.

[16] Hosmer D., Lemeshow S. & Sturdivant R., *Applied Logistic Regression*, Jonh Wiley & Sons, 2013.

[17] Lindsey J., *Applying Generalized Linear Models*, Springer, 1997.

[18] Iglesias T., *Métodos de Bondad de ajuste en regresión logística*, Trabajo fin de Máster, 2013.

[19] Lopéz J. García J., *Eventos por variable en Regresión Logística y Redes Bayesianas para Predecir Actitudes Emprendedoras*, Universidad de Almeria, 2011.

[20] Long S., *Regression Models for Categorical and Limited Dependent Variables*, Sage, 1997.

[21] Maldonado A., *Identificación de los Factores que intervienen en la reprobación del curso de Matemáticas Básicas de la FCFM de la BUAP*, Tesis de Licenciatura, México, 2012.

[22] Montgomery D., Peck E., Vining G. (trad. González V.), *Introducción al Análisis de Regresión Lineal*, 3a. edición, Cecsa, 2006.

[23] Montoya S., *Análisis de la eficiencia terminal de la generación 1999-2003 de la maestría en ciencias con especialidad en administración pública*, campus virtual politécnico de la escuela superior de comercio y administración, unidad Santo Tomás, IPN, Tesis de Maestría, 2006.

[24] Navarro E., Duarte V., Hernández S., *La eficiencia terminal en la educación superior privada en México: estudio del caso de la Universidad Cristóbal Colón*, Revista de la Universidad Cristóbal Colón, www.eumed.net/rev/rvec/19/, 2004.

[25] Nieto A, Reyes H., Godínez F., Tajonar F. y Vázquez V., *Descripción de las generaciones 2000-2008 de la Facultad de Físico-Matemáticas de la Benemérita Universidad Autónoma de Puebla* , IV Encuentro sobre Didáctica de la Estadística, la Probabilidad y el Análisis de Datos, 2014.

[26] Nieto A., Godínez F. y Reyes H., *Variables que influyen para que un alumno egrese de dos licenciaturas de la Buap*, XXIX Foro Internacional de Estadística, Cartel 2014.

[27] Niño E., *La relación estilos de aprendizaje y rendimiento escolar académico en alumnos de una facultad de la UANL*, Universidad Autónoma de Nuevo León, Tesis de Doctorado, Noviembre 2003.

[28] Ontiveros I., *Seguimiento de Egresados de la Licenciatura en Artes Visuales de la Escuela de Pintura, Escultura y Artesanías de la UJED*, Tesis de Maestría, 2006.

[29] R (2011), R versión 2.13.1 Copyright (C) 2011, The Foundation for Statical Computing, for Windows.

[30] Radhakrishna Rao C., Toutenburg H. and Shalabh, Heumann C., *Linear Models and Generalizations: Least Squares and Alternatives*, 3a. Edition, Springer, 2008.

[31] Reyes J., Canizo J., Meza E., Herrera A., Cruz H., Nieto A. y Godínez J.; *Descripción de las generaciones 2000-2008 para los alumnos que desertan en dos licenciaturas de la FCFM-BUAP*, Encuentro Participación de la Mujer en la ciencia, CIO 2014.

[32] Rubio H. y Berlanga S., *Como aplicar las pruebas parámetricas bivariadas t de Student y ANOVA en SPSS. Caso práctico*, REIRE, 2012.

[33] Ryan T. *Modern Regression Methods*, Wiley, second edition, 2008.

[34] Santos de los E., *Los procesos de permanencia y abandono escolar en educación superior*, Universidad de Colima, Revista Iberoamericana de Educación, 2000.

[35] SPSS(2010), IBM SPSS Statistics 19 para Windows.

[36] Tinto V., *Una consideración de las teorías de la deserción estudiantil en la trayectoria escolar en la educación superior*, México: ANUIES, 1987.

[37] Vittinghoff E., Glidden D., Shboski S., McCulloch C., *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models*, second edition, Springer, 2012.