Universidad Nacional Autónoma de México

Facultad de Ciencias

Aplicación del Modelo de Regresión Logística en la medición del rendimiento académico de alumnos de primer año de la Licenciatura de Médico Cirujano Generación 2013-2014

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A:

DANAE MIREL MARTÍNEZ VARGAS

TUTORAS:

Dra. Ruth Selene Fuentes García

Dra. María Esther Urrutia Aguilar

2015

Ciudad Universitaria, D. F.

Índice general

Agradecimientos

Agradezco a la Universidad Nacional Autónoma de México (UNAM) y el particular a la Facultad de Ciencias por haberme brindado una formación académica de excelencia, así como haberme abierto las puertas a la realización de proyectos enriquecedores para mi formación académica, personal y profesional.

A la Dirección General de Cooperación e Internacionalización de UNAM (DGECI) por haberme apoyado en el proceso de movilidad estudiantil, así como contribuir de manera activa para la obtención de un financiador que me apoyó económicamente durante mi estancia en el extranjero.

A mis padres Rosa María Vargas y Javier Martínez por el regalo tan grande de la vida, por ser mis guías y mi apoyo. A mi hermano Edgar Iván Martínez, por estar siempre presente en los momentos importantes, por su constante apoyo, cariño y hermandad.

A mis amigos y siempre compañeros de la Facultad, Leonel Hernández, Ariel Oliva, Ricardo López, Néstor Medina y Lizette Lemus por su amistad, compañerismo, confianza y por haber sido un impulso para mí, sobre todo al inicio de la carrera.

A mis profesores de la Facultad y en particular a mi profesor y ejemplo a seguir César Almenara, gracias por las enseñanzas, el apoyo y confianza brindados, por ser fuente de inspiración.

A mi tutora Ruth Selene Fuentes, por inspirarme admiración, por haber compartido sus conocimientos, por brindarme su apoyo en todo momento, por sus sabias observaciones y consejos, por su paciencia y comprensión.

A mi tutora María Esther Urrutia, por sus valiosas aportaciones, por haberme apoyado desde incluso mucho antes de iniciar este proyecto, ya que sin su apoyo este trabajo no hubiera podido ser concretado.

A mis sinodales Jaime Vázquez, Margarita Chávez y a mí también profesor Salvador Zamora, por sus sugerencias y acertados consejos para el enriquecimiento de este trabajo y el tiempo dedicado a la revisión del mismo.

Introducción

El objetivo de la presente tesis es determinar las variables académicas, psicosociales y vocacionales más relevantes para medir y predecir el rendimiento académico de estudiantes de primer año de la licenciatura Médico-Cirujano de la Facultad de Medicina de la Universidad Nacional Autónoma de México (UNAM).

Esta investigación se trata de un estudio longitudinal, ya que la población bajo estudio fue observada a lo largo del ciclo escolar 2013-2014.

La estructura del presente trabajo consiste en 5 capítulos. En el ***primer capítulo*** se presentan los principales resultados teóricos acerca de la construcción del modelo de regresión logística, así como la interpretación y pruebas de hipótesis para los parámetros, pruebas de bondad de ajuste y diagnóstico del modelo.

En el ***segundo capítulo*** se detalla la motivación y objetivo de esta investigación, también se describe el contexto en el cual se desarrolla la problemática y se describe la naturaleza de los datos con los que se cuenta, en la parte final de este capítulo se establecen los criterios de inclusión para seleccionar a los individuos bajo estudio.

En el ***capítulo tercero*** se presenta el análisis exploratorio de los datos, incluyendo el análisis de tablas de contingencia y medidas de dependencia entre las variables categóricas con las que se cuenta. En este capítulo se presentan las variables que están asociadas entre ellas y con la variable de respuesta para nuestro modelo: *El éxito académico*.

En el ***cuarto capítulo*** se presenta la selección del modelo final, y el análisis del mismo: interpretación de parámetros, intervalos de confianza, bondad de ajuste, poder predictivo y diagnóstico. Finalmente en el ***quinto capítulo*** se presentan las conclusiones globales obtenidas a lo largo del desarrollo de la investigación y se proponen técnicas y soluciones para controlar el fenómeno del bajo rendimiento académico en la población de estudiantes de primer año de la Licenciatura de Médico-Cirujano de la UNAM.

Capítulo 1

Introducción al modelo de regresión logística

El objetivo del modelo de regresión logística es describir la relación entre una o más variables independientes y una variable dependiente. Este modelo se diferencia de otros modelos de regresión, por el hecho de que la variable respuesta es dicotómica. Otra característica es que las variables regresoras pueden ser continuas o categóricas. Así, con ayuda de la regresión logística se puede estudiar la probabilidad de ocurrencia o no ocurrencia de algún evento, en función de un conjunto de variables predictoras.

Recientemente, la regresión logística se ha convertido en una socorrida herramienta en aplicaciones de negocios, medicina, epidemiología, sociología, mercadotecnia, etc. ya que es útil para resolver problemas como la predicción del padecimiento o no de una enfermedad, la supervivencia o muerte de pacientes, el otorgamiento o no de un crédito bancario a un determinado individuo, el éxito o fracaso de un negocio, entre muchos otros. Además de que este modelo proporciona una probabilidad estimada de ocurrencia o no del fenómeno bajo estudio.

En este capítulo se presentarán los principales planteamientos teóricos acerca de la regresión logística. En la ***sección 1.1*** se presentan los conceptos de momios y cociente de momios, que servirán de base para la interpretación de los parámetros del modelo.

En la ***sección 1.2*** se presenta la construcción del modelo, así como el método de estimación de los parámetros y la varianza de estos. En la ***sección 1.3*** se aborda la interpretación del vector de parámetros estimados. Mientras que en las ***secciones 1.4 y 1.5*** se toca el tema de bondad de ajuste y significancia de los regresores. Finalmente, la sección ***sección 1.6*** se centra en el diagnóstico del modelo y de las observaciones.

1.1. Momios y cociente de momios.

Consideremos una tabla de contingencia de 2×2 [1], es decir, el caso en el que la variable independiente X también es binaria.

Cuadro 1.1: tabla de contingencia de tamaño 2×2

	Presencia $Y = 1$	Ausencia $Y = 0$	Total
Presencia $X = 1$	a	b	$a + b$
Ausencia $X = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

Un momio es la comparación entre la probabilidad de observar la respuesta $Y = 1$ y de no observarla, para individuos con la misma característica de X. Entonces definimos al momio de presentar la respuesta $Y = 1$ si se tiene la característica $X = i$ de la variable regresora como:

$$\Omega_i = \frac{\pi_i}{1 - \pi_i} = \frac{\mathbb{P}(Y = 1|X = i)}{\mathbb{P}(Y = 0|X = i)} \tag{1.1}$$

A partir de los momios se construye una medida de asociación entre las variables X y Y, el cociente de momios OR (por sus siglas en inglés *Odds Ratio*) definido como:

$$OR = \frac{\Omega_1}{\Omega_0} = \frac{\frac{\mathbb{P}(Y=1|X=1)}{\mathbb{P}(Y=0|X=1)}}{\frac{\mathbb{P}(Y=1|X=0)}{\mathbb{P}(Y=0|X=0)}} = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}} \tag{1.2}$$

Es decir, el OR es el momio de presentar la respuesta $Y = 1$ si se tiene la característica 1 de X sobre el momio de presentar la respuesta $Y = 1$ si se tiene la característica 0 de X. Las estimaciones de estas médidas en términos de los conteos observados en la tabla son:

$$\widehat{OR} = \frac{ad}{cb}$$
$$\widehat{\Omega}_1 = \frac{a}{b}$$
$$\widehat{\Omega}_0 = \frac{c}{d}$$

Si $\widehat{OR} > 1$ diremos que el momio de presentar la respuesta $Y = 1$ si se tiene la característica $X = 1$ es $\widehat{OR}$ veces mayor que el momio de presentar la respuesta $Y = 1$ si se tiene la característica $X = 0$.

[1] ver apéndice A

Por otro lado si $\widehat{OR} < 1$ interpretamos el recíproco diciendo que el momio de presentar la respuesta $Y = 1$ si se tiene la característica $X = 0$ es $1/\widehat{OR}$ veces mayor que el momio de presentar la respuesta $Y = 1$ si se tiene la característica $X = 1$.

Notemos que si el cociente de momios es igual a uno, se sigue que las variables X y Y son independientes y como nuestro objetivo es establecer una asociación entre la variable de respuesta y la variable independiente, es de nuestro interés verificar que el OR tome un valor distinto de uno. Para esto es necesario la construcción del intervalo de confianza. Con ayuda del método delta [2], es posible calcular la varianza del logaritmo del cociente de momios que, en términos de los conteos de la tabla, es: [3]

$$\mathbb{V}(log(\widehat{OR})) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{1.3}$$

Con lo que el intervalo de $(1-\alpha)$x100 % de confianza para el cociente de momios está dado por:

$$exp\left\{log(\widehat{OR}) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right\} \tag{1.4}$$

Si este intervalo de confianza no contiene el uno, podemos afirmar que el $\widehat{OR}$ es estadísticamente distinto de uno, lo que significa que las variables X y Y están asociadas.

1.2. Modelo Logístico

El modelo de regresión logística forma parte de la familia de los llamados *modelos jerárquicos*, que se refiere al hecho de que siempre que se introduzca un término de interacción [3] de orden superior al modelo, deben introducirse también los términos de interacción de orden inferior, así como los términos independientes de las variables que intervienen en dicha interacción (efectos principales).

El objetivo es modelar la probabilidad de éxito ante la presencia de p variables regresoras $\mathbf{X} = (X_1, ..., X_p)$. La variable de respuesta, Y es dicotómica, es decir, toma únicamente los valores $Y = 0$ y $Y = 1$.

$$\mathbb{P}(Y = 1|\mathbf{X})$$

[2] El método delta es un procedimiento técnico que se utiliza para aproximar la esperanza y la varianza de una v.a. utilizando series de Taylor.

[3] Las interacciónes son las asociaciones existentes entre las covariables bajo estudio, o bien, los efectos de una variable de acuerdo a la otra.

$\mathbf{X}$ denota el conjunto de variables regresoras. Denotaremos esta probabilidad como $\pi(\mathbf{X})$ para hacer énfasis en que ésta probabilidad depende de los valores tomados por el conjunto $\{X_1, X_2, ..., X_p\}$. Notemos que $\pi(\mathbf{X})$ es el parámetro para la distribución Bernoulli asociada a Y.

Así pues, buscamos un modelo cuyos valores para la respuesta estén contenidos en el intervalo $(0, 1)$. También es deseable que X tenga un menor efecto cuando $\pi(\mathbf{X})$ este cercano a cero o a uno que cuando este cercano a un rango medio. La función que permite ajustar este tipo de curvas es la función logística, cuyo modelo de regresión asociado es el modelo de regresión logística.

1.2.1. Función logística

La función logística es una función matemática que aparece frecuentemente en la modelación de fenómenos como propagación de enfermedades epidémicas, difusión en redes sociales, etc. Inicia en cero y crece hasta uno en forma sigmoidea, tal como se muestra en la gráfica 1.1. La expresión matemática que describe dicha función es:

$$f(x) = \frac{exp\{x\}}{1 + exp\{x\}} = \frac{1}{1 + exp\{-x\}} \tag{1.5}$$

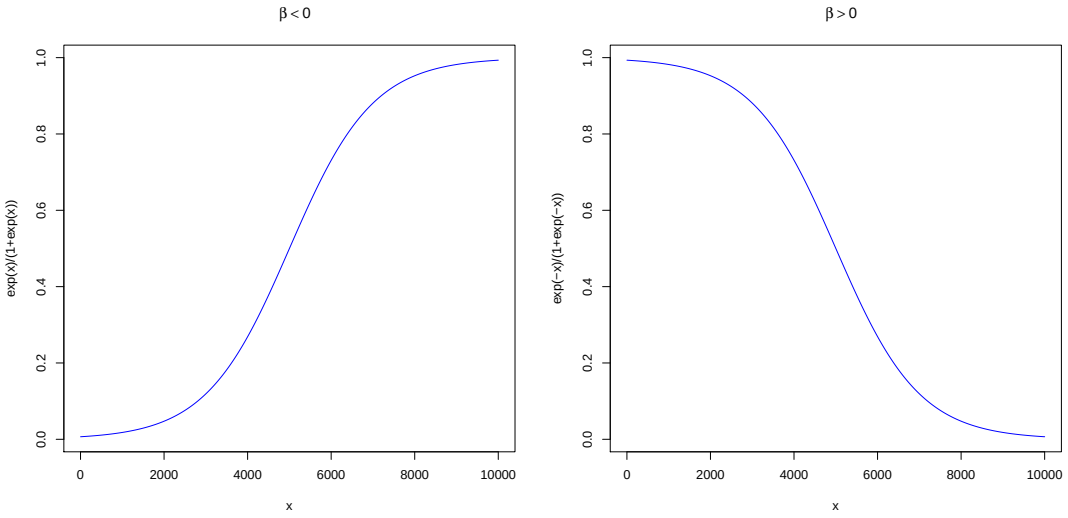


Figura 1.1: Gráfica de la función logística.

La función logística puede ser empleada para establecer el modelo de regresión, que determina la probabilidad de ocurrencia de un determinado evento de acuerdo a la presencia de un conjunto de covariables.

$$\pi(\mathbf{X}) = \mathbb{P}(Y = 1|\mathbf{X}) = \frac{exp\{\beta^t \mathbf{X}\}}{1 + exp\{\beta^t \mathbf{X}\}} \tag{1.6}$$

Donde $\beta = (\beta_0, ..., \beta_p)$ es el vector de parámetros asociados a cada una de las variables predictoras y $\mathbf{X} = (X_0, ..., X_p)$ es el vector de variables predictoras, con $X_0 = 1$

1.2.2. Inferencia acerca del modelo.

La estimación de los parámetros se realiza por el método de máxima verosimilitud. Consideremos que se tiene una muestra de n sujetos, para cada uno de los cuales se obtuvieron los valores observados de la variable Y y del vector de covariables $\mathbf{X} = \{X_1, ..., X_p\}$.

Si suponemos que los datos se pueden agrupar en clases, de acuerdo al valor observado del vector de covariables, entonces las respuestas Y en estas clases tienen una distribución binomial. A cada una de estas clases se les denomina patrón de covariables. [4] Definimos como n_i al número de sujetos pertenecientes a la i-ésima clase, entonces $n = n_1 + ... + n_N$ con N igual al número de valores distintos del vector observado $\mathbf{X}$.

Así pues, la probabilidad de que algún sujeto que pertenezca que al i-ésimo grupo tenga como respuesta asociada $Y = 1$ esta dada por:

$$\pi(\mathbf{X}_i) = \mathbb{P}(Y = 1|\mathbf{X}_i) = \frac{exp\{\beta^t \mathbf{X}_i\}}{1 + exp\{\beta^t \mathbf{X}_i\}} \tag{1.7}$$

Considerando que $\mathbf{X}_i$ denota el valor observado del vector $\mathbf{X}$ para el i-ésimo patrón de covariables. La probabilidad de observar m_i sujetos con respuesta asociada $Y = 1$ de un total de n_i es:

$$\binom{n_i}{m_i} \pi(x_i)^{m_i}(1 - \pi(x_i))^{n_i - m_i} \tag{1.8}$$

Siguiendo el procedimiento estándar para estimar parámetros por máxima verosimilitud, consideramos $\mathbf{m} = (m_1, ..., m_N)$ el vector de valores observados para cada uno de los N grupos (patrones de covariables).

$$\begin{aligned} L(\pi(x_i)) &\propto \prod_{i=1}^{N} \pi(x_i)^{m_i}(1 - \pi(x_i))^{n_i - m_i} \\ &= \pi(x_i)^{\sum_{i=1}^{N} m_i}(1 - \pi(x_i))^{-\sum_{i=1}^{N} m_i} \prod_{i=1}^{N}(1 - \pi(x_i))^{n_i} \end{aligned}$$

[4] un patrón de covariables es un grupo de individuos en el que todos poseen los mismos valores en sus covariables.

$$
\begin{aligned}
&= exp\left\{\sum_{i=1}^{N} m_i log(\pi(x_i))\right\} exp\left\{-\sum_{i=1}^{N} m_i log(1-\pi(x_i))\right\} \prod_{i=1}^{N}(1-\pi(x_i))^{n_i} \\
&= exp\left\{\sum_{i=1}^{N} m_i log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)\right\} \prod_{i=1}^{N}(1-\pi(x_i))^{n_i} \\
&= exp\left\{\sum_{i=1}^{N} m_i(\beta^t \mathbf{X}_i)\right\} (1-\pi(x_i))^{\sum_{i=1}^{N} n_i}
\end{aligned}
$$

Pues el logaritmo del momio puede expresarse de manera lineal en función del vector de parámetros β, tal como se muestra en 1.12. Calculando la log-verosimilitud tenemos:

$$
\begin{aligned}
&\sum_{i=1}^{N} m_i(\beta^t \mathbf{X}_i) + log(1-\pi(x_i))^{\sum_{i=1}^{N} n_i} \\
&= \sum_{i=1}^{N} m_i(\beta^t \mathbf{X}_i) - \sum_{i=1}^{N} n_i(-log(1-\pi(x_i))) \\
&= \sum_{i=1}^{N} m_i(\beta^t \mathbf{X}_i) - \sum_{i=1}^{N} n_i log\left(\frac{1}{1-\pi(x_i)}\right) \\
&= \sum_{i=1}^{N} m_i(\beta^t \mathbf{X}_i) - \sum_{i=1}^{N} n_i log\left(1+\frac{\pi(x_i)}{1-\pi(x_i)}\right) \\
&= \sum_{i=1}^{N} m_i(\beta^t \mathbf{X}_i) - \sum_{i=1}^{N} n_i log(1+exp\{\beta \mathbf{X}_i^t\})
\end{aligned}
$$

Así, las ecuaciones de verosimilitud se obtienen al derivar parcialmente esta última expresión con respecto a β_0 y a β_k para $k = 1, ..., p.$ e igualar a cero, siendo estas de la forma:

$$
\begin{aligned}
\frac{\partial L(\beta)}{\partial \beta_0} &= \sum_{i=1}^{N}(m_i - n_i\pi(x_i)) = 0 \\
\frac{\partial L(\beta)}{\partial \beta_k} &= \sum_{i=1}^{N} x_{ik}(m_i - n_i\pi(x_i)) = 0 \text{ para } k = 1, ..., p.
\end{aligned}
$$

Lo que da lugar a un sistema de $(p+1)$ ecuaciones con variables cuya solución no es cerrada, por lo que para obtener la solución es necesario la aplicación de métodos numéricos como el algoritmo de ***Newton-Raphson*** [5].

[5]método iterativo que sirve para encontrar las raíces de una ecuación dada o la solución aproximada de un sistema de acuaciones.

Después de haber estimado al vector β y haber ajustado el modelo, es deseable verificar la significancia estadística de estos parámetros. Para el cálculo de intervalos de confianza y de los errores estándar para cada una de las $\hat{\beta}_i$ necesitamos estimar la varianza de $\hat{\beta}$. El método de estimación de la matriz de varianzas y covarianzas de $\hat{\beta}$ esta basado en la matriz de segundas derivadas parciales de la función de log-verosimilitud, llamada matriz de información de Fisher. [1] Así pues, las entradas de dicha matriz están dadas por las expresiones siguientes:

$$I(\beta)_{kk} = \frac{\partial^2 L(\beta)}{\partial^2 \beta} = \sum_{i=1}^{N} n_i x_{ik}^2 \pi(x_i)(1 - \pi(x_i))$$

$$I(\beta)_{ku} = \frac{\partial^2 L(\beta)}{\partial \beta_k \partial \beta_u} = \sum_{i=1}^{N} n_i x_{ik} x_{iu} \pi(x_i)(1 - \pi(x_i))$$

Con $X_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ es el valor que toma el conjunto de variables independientes para el i-ésimo patrón de covariables. Una manera útil de factorizar esta matriz de información es [1]

$$I(\widehat{\beta}) = X^t V X$$

Donde la matriz X contiene las observaciones del conjunto de covariables $\mathbf{X}$

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{pmatrix}$$

Y V es la matriz diagonal $V = diag(\widehat{\pi}(x_i)(1 - \widehat{\pi}(x_i)))$.

$$V = \begin{pmatrix} \hat{\pi}(x_1)(1 - \hat{\pi}(x_1)) & 0 & \cdots & 0 \\ 0 & \hat{\pi}(x_2)(1 - \hat{\pi}(x_2)) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\pi}(x_N)(1 - \hat{\pi}(x_N)) \end{pmatrix}$$

La matriz de varianzas y covarianzas para el vector de parámetros estimados, está dada por la inversa de la matriz de información de Fisher. [1]

$$\mathbb{V}(\widehat{\beta}) = (I(\widehat{\beta}))^{-1} \tag{1.9}$$

Los términos de la diagonal de la matriz $\mathbb{V}$ son las varianzas y los elementos fuera de la diagonal son las covarianzas. Siendo así, podemos evaluar la precisión de cada uno de los parámetros estimados de β a través de los errores estándar (SE).

$$SE(\widehat{\beta}_k) = (\mathbb{V}(\widehat{\beta}_k))^{\frac{1}{2}} \tag{1.10}$$

Los intervalos del $(1-\alpha)$x100 % de confianza para β están dados por:

$$\widehat{\beta}_k \mp Z_{1-\frac{\alpha}{2}} SE(\widehat{\beta}_k) \tag{1.11}$$

1.3. Interpretación de los parámetros

La ecuación 1.6 tiene el inconveniente de corresponder a un modelo no lineal, lo que hace difícil su interpretación. Sin embargo, al aplicar la función logaritmo al momio de respuesta obtenemos:

$$\log\left(\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{1.12}$$

puesto que:

$$\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})} = \frac{\mathbb{P}(Y=1|\mathbf{X})}{1-\mathbb{P}(Y=1|\mathbf{X})} = \frac{\frac{\exp\{\beta^t\mathbf{X}\}}{1+\exp\{\beta^t\mathbf{X}\}}}{1-\frac{\exp\{\beta^t\mathbf{X}\}}{1+\exp\{\beta^t\mathbf{X}\}}} = \exp\{\beta^t\mathbf{X}\}$$

La ecuación descrita en 1.12, mejor conocida como *logit*, es un modelo lineal para el logaritmo del momio de respuesta. El vector de parámetros $\beta = (\beta_0, ..., \beta_p)$ determina la tasa de incremento o decremento de la curva para $\pi(\mathbf{X})$, el signo de β_i determina si la curva asciende $(\beta_i > 0)$, o desciende $(\beta_i < 0)$ ante la presencia de la variable X_i.

Así pues, si $\beta_i > 0$ un aumento en X_i ocasiona un aumento de la probabilidad $\pi(\mathbf{X})$, mientras que si $\beta_i < 0$ un aumento en X_i causa la disminución de $\pi(\mathbf{X})$, manteniendo constante el resto de las variables.

Como se mencionó al inicio del capítulo, la regresión logística acepta covariables medidas en escala categórica o continua. Si la variable X_i es dicotómica, entonces, el momio de respuesta para $X_i = 1$ con respecto al momio de respuesta para $X_i = 0$ está dado por:

$$OR = \frac{\Omega_1}{\Omega_0} = \frac{\frac{\mathbb{P}(Y=1|X_i=1)}{\mathbb{P}(Y=0|X_i=1)}}{\frac{\mathbb{P}(Y=1|X_i=0)}{\mathbb{P}(Y=0|X_i=0)}} = \frac{\frac{\pi(X_i=1)}{1-\pi(X_i=1)}}{\frac{\pi(X_i=0)}{1-\pi(X_i=0)}} \tag{1.13}$$

donde

$$\pi(X_i = 1) = \frac{exp\{\beta_0 + \beta_i\}}{1 + exp\{\beta_0 + \beta_i\}}$$

y

$$\pi(X_i = 0) = \frac{exp\{\beta_0\}}{1 + exp\{\beta_0\}}$$

de modo que

$$\Omega_1 = \frac{\pi(X_i = 1)}{1 - \pi(X_i = 1)} = exp\{\beta_0 + \beta_i\}$$

$$\Omega_0 = \frac{\pi(X_i = 0)}{1 - \pi(X_i = 0)} = exp\{\beta_0\}$$

de donde se sigue que

$$\log(\Omega_1) - \log(\Omega_0) = \log\left(\frac{\Omega_1}{\Omega_0}\right) = \beta_0 + \beta_i - \beta_0 = \beta_i$$

con lo que tenemos que el cociente de momios resulta ser

$$OR = \frac{\Omega_1}{\Omega_0} = exp\{\beta_i\} \tag{1.14}$$

Que indica que el momio de presentar la respuesta $Y = 1$ si se tiene el valor de la covariable $X_i = 1$ es $exp\{\beta_i\}$ veces mayor que el momio de presentar la respuesta $Y = 1$ si se tiene el valor de la covariable $X_i = 0$, manteniendo constante el resto de las variables.

Por otro lado, si la variable X_i es continua, al tomar los valores $X_i = x$ y $X_i = x + k$, el momio de respuesta resultante está dado por:

$$OR = \frac{\frac{\mathbb{P}(Y=1|X_i=x+k)}{\mathbb{P}(Y=0|X_i=x+k)}}{\frac{\mathbb{P}(Y=1|X_i=x)}{\mathbb{P}(Y=0|X_i=x)}} = \frac{\frac{\pi(X_i=x+k)}{1-\pi(X_i=x+k)}}{\frac{\pi(X_i=x)}{1-\pi(X_i=x)}} \tag{1.15}$$

donde

$$\Omega_{x+k} = \frac{\pi(X_i = x + k)}{1 - \pi(X_i = x + k)} = exp\{\beta_0 + \beta_i(x + k)\}$$

$$\Omega_x = \frac{\pi(X_i = x)}{1 - \pi(X_i = x)} = exp\{\beta_0 + \beta_i(x)\}$$

de donde se sigue que

$$\log(\Omega_{x+k}) - \log(\Omega_x) = \log\left(\frac{\Omega_{x+k}}{\Omega_x}\right) = \beta_0 + x\beta_i + k\beta_i - (\beta_0 + x\beta_i) = k\beta_i$$

De donde se tiene que:

$$OR = \frac{\Omega_{x+k}}{\Omega_x} = exp\{k\beta_i\} \tag{1.16}$$

Que se interpreta como el cambio esperado en el momio de respuesta por k unidades de cambio en X_i, manteniendo constante el resto de las variables. [2]

Finalmente, si la variable explicativa X_i es politómica (acepta más de dos categorías) con h niveles, se elige una categoría de referencia (normalmente la primera) y se construyen $(h-1)$ momios, de manera que comparemos cada una de las categorías con aquella que elijamos como base. Luego la interpretación de los $h-1$ momios es similar a la interpretación para variables dicotómicas.

1.4. Pruebas de hipótesis sobre los parámetros

Adicional a los intervalos de confianza de β expuestos en 1.11 de la *sección 1.2* se recomienda realizar las correspondientes pruebas de hipótesis, con la finalidad de determinar si las variables explicativas son o no estadísticamente significativas para explicar la respuesta.

Para esto, consideremos un modelo con un solo parámetro, es decir, el caso en el que las variables predictoras no tienen impacto alguno en la respuesta, este es el modelo conocido como *modelo nulo*. Por otro lado, consideremos un modelo con tantos parámetros como observaciones, en este caso el valor esperado y el observado de Y resultan ser iguales, este modelo es llamado *modelo saturado*.

Notemos que los dos modelos descritos anteriormente no resultan informativos, sin embargo proporcionan una base para medir la discrepancia[6] del modelo ajustado, debido a que buscamos estar cerca del modelo saturado, pues éste es el que representa el ajuste perfecto, mientras que buscamos estar lejos del nulo, ya que éste indica que no existe relación entre las variables regresoras y la variable respuesta.

1.4.1. Prueba del cociente de verosimilitudes

Dado que hemos logrado ajustar un modelo con k parámetros, se desea probar:

$$H_0 : \beta = 0 \; vs. \; H_a : \beta \neq 0$$

Donde $\beta = (\beta_1, ..., \beta_p)$ es el vector de parámetros del modelo ajustado. Rechazar la hipótesis nula, indica que al menos uno de los regresores es distinto de cero. El estadístico de prueba está dado por:

$$D = -2log\left(\frac{L(\widehat{\beta})}{L(\widehat{\beta}_{sat})}\right) \tag{1.17}$$

[6]Diferencia, desigualdad que resulta de la comparación de las cosas entre sí.

con $L(\widehat{\beta})$ y $L(\widehat{\beta}_{sat})$ las verosimilitudes para el modelo ajustado y el modelo saturado, respectivamente. El estadístico sigue un distribución aproximada Ji-cuadrada con $(n-k)$ grados de libertad, donde n es el número de observaciones y k el número de parámetros del modelo ajustado. [1]

1.4.2. Prueba de Wald

La prueba anterior indica que al menos uno de los parámetros es distinto de cero, pero no indica cuál o cuáles, por tanto, es conveniente aplicar pruebas individuales para verificar por separado la significancia de cada uno de los parámetros del modelo. Las hipótesis nula y alternativa son:

$$H_0 : \beta_i = 0 \; vs. \; H_a : \beta_i \neq 0, \text{ para } i = 1, 2, ..., p$$

El estadístico de prueba está dado por la siguiente expresión:

$$\frac{\widehat{\beta}_i^2}{\widehat{\mathbb{V}}(\widehat{\beta}_i)} \tag{1.18}$$

que sigue una distribución asintótica Ji-cuadrada con un grado de libertad. No rechazar Ho indicaría que β_i podría ser cero, o bien, que la variable independiente X_i no aporta información para predecir a la respuesta Y.

Cuando $|\beta|$ es relativamente grande, la prueba de Wald no resulta tan potente como la prueba del cociente de verosimilitudes [2] por lo que es necesario que esta prueba se aplique tomando en cuenta esta limitante.

1.5. Bondad de Ajuste

Ahora bien, para saber si el modelo se ajusta adecuadamente a los datos, es necesario tomar en cuenta medidas que indiquen la discrepancia entre los valores esperados del modelo y los valores observados en la muestra. Definiremos a continuación algunas medidas que cubren esta necesidad.

1.5.1. Devianza

Una vez que contamos con un candidato a ajustar adecuadamente los datos, una manera de cuantificar qué tan bueno es este ajuste es mediante la devianza, definida como menos dos veces el logaritmo del cociente de verosimilitudes entre modelo saturado y el ajustado. [1]

$$D = -2log\left(\frac{L(\widehat{\beta})}{L(\widehat{\beta}_{sat})}\right) = -2[log(L(\widehat{\beta})) - log(L(\widehat{\beta}_{sat}))] \tag{1.19}$$

Notemos que la devianza es el estadístico de la prueba del cociente de verosimilitudes para comparar el modelo ajustado contra el modelo saturado. Si la diferencia en 1.19 es pequeña, entonces diremos que el modelo propuesto ajusta bien, ya que como mencionamos antes, se busca estar lo más cerca posible del modelo con el ajuste perfecto.

También puede presentarse la situación en la que hayamos encontrado dos modelos que al parecer son adecuados, así que debemos decidir cuál de ellos es mejor. Para compararlos, también usamos la devianza.

Consideremos A_1 y A_2 dos modelos anidados [7] con p y q parámetros respectivamente ($q < p$). Así pues, A_1 contiene todas las variables regresoras de A_2 y más. Según el *principio de parsimonia* [8], el modelo con menos parámetros es preferible sobre el modelo grande, pero debemos asegurarnos que este último no contenga variables que resulten altamente significativas para predecir la respuesta, pues de lo contrario estaríamos perdiendo información valiosa que proporcionan las covariables.

La comparación se hace entonces con la diferencia de devianzas:

$$\begin{aligned} & D_{A_1} - D_{A_2} \\ & = -2log\left(\frac{L(A_1)}{L(\widehat{\beta}_{sat})}\right) + \left(-2log\left(\frac{L(A_2)}{L(\widehat{\beta}_{sat})}\right)\right) \\ & = -2log\left(L(A_1) - L(\widehat{\beta}_{sat}) - ((L(A_2) - L(\widehat{\beta}_{sat}))\right) \\ & = -2log\left(\frac{L(A_1)}{L(A_2)}\right) \end{aligned} \tag{1.20}$$

Si esta cantidad es cercana a cero, esto sugiere que ambos modelos poseen la misma información, por lo que preferiremos el modelo más simple (aquel con menos parámetros). De lo contrario el modelo A_1 contiene información importante que no contienen A_2.

[7] todos los regresores del modelo pequeño deben estar contenidos en el modelo grande.
[8] en igualdad de condiciones, la explicación más sencilla suele ser la correcta.

1.5.2. Prueba Ji-cuadrada de Pearson

Para medir la diferencia entre los valores ajustados por el modelo y los observados, así como probar la hipótesis nula de que el modelo ajusta bien, realizamos la prueba Ji-cuadrada de Pearson.

Nuevamente, consideremos que nuestros datos están conformados por N patrones de covariables, cada uno de ellos de tamaño n_i, con $i = 1, ..., N$. Definimos los residuos de Pearson para el i-ésimo patron de covariables como:

$$r_i = \frac{y_i - n_i\widehat{\pi}_i}{\sqrt{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)}} \tag{1.21}$$

Con y_i el número de éxitos observados en este i-ésimo patrón de covariables y $\widehat{\pi}_i$ la probabilidad de éxito en este mismo. La estadística de prueba se basa en los residuos descritos antes.

$$\sum_{i=1}^{N} \frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)} \tag{1.22}$$

Tiene una distribución asintótica Ji-cuadrada con $N - k$ grados de libertado, donde N es el número total de patrones de covariables en la muestra y k es el número total de parámetros en el modelo ajustado. [1]

1.5.3. Prueba de Hosmer Lemeshow

Cuando una o más de las variables independientes son continuas, lo más probable es tener tantos patrones de covariables como observaciones, en cuyo caso la aplicación de la prueba Ji-cuadrada sería incorrecta, en estos casos se propone aplicar la prueba de Hosmer Lemeshow, cuyas hipótesis nula y alternativa son iguales a las de la prueba anterior.

$$H_0 : \text{El modelo ajusta bien.}$$

Hosmer y Lemeshow [1] proponen agrupar observaciones por medio de las probabilidades estimadas por el modelo y compararlas con las observadas en la muestra. Primero se ordena a los individuos en estudio de acuerdo a sus probabilidades estimadas $\widehat{\pi}_i$, luego se agrupan las observaciones de acuerdo a percentiles, normalmente se generan g=10 grupos (se toman deciles). La estadística de prueba está dada por:

$$\sum_{k=1}^{g} \frac{o_k - n_k\bar{\pi}_k}{n_k\bar{\pi}_k(1-\bar{\pi}_k)} \tag{1.23}$$

con

$$\bar{\pi}_k = \sum_{i=1}^{c_k} \frac{m_i \widehat{\pi}_k}{n_k}$$

$$o_k = \sum_{i=1}^{c_k} y_i$$

Que se distribuye Ji-cuadrada con $g-2$ grados de libertad [1]. Valores grandes de la estadística indican falta de ajuste del modelo.

1.6. Diagnóstico

1.6.1. Poder predictivo y La curva de ROC

La curva de ROC por sus siglas en inglés *Receiver Operating Characteristic* es un método gráfico para evaluar la capacidad predictiva del modelo apoyándose en la tabla de clasificación, definida en el cuadro 1.2

Cuadro 1.2: tabla de clasificación

	$\widehat{Y}=1$	$\widehat{Y}=0$	*Total*
$Y=1$	*Verdaderos positivos* (VP)	*Falsos negativos* (FN)	*Total de positivos* (TP)
$Y=0$	*Falsos positivos* (FP)	*Verdaderos negativos* (VN)	*Total de negativos* (TN)

Donde $\widehat{Y}$ denota el valor predicho a través del modelo. Dependiendo el punto de corte, que llamaremos π_0 la clasificación del modelo será tomada como:

$$\widehat{Y} = 1 \text{ cuando } \widehat{\pi}_i > \pi_0$$

$$\widehat{Y} = 0 \text{ cuando } \widehat{\pi}_i < \pi_0$$

Usualmente se toma π_0 =0.5, sin embargo, se puede elegir arbitrariamente cualquier punto de manera que se maximice el número de verdaderos positivos y de verdaderos negativos.

Definimos la *Sensibilidad* del modelo como $\mathbb{P}(\widehat{Y}=1|Y=1)$, es decir, la probabilidad de clasificar correctamente a los unos como unos. En términos de los conteos de la tabla 1.2

$$Sensibilidad = \frac{VP}{TP} \tag{1.24}$$

Por otro lado, la *Especificidad*, definida como $\mathbb{P}(\widehat{Y} = 0|Y = 0)$, es la probabilidad de clasificar a los ceros correctamente como ceros. En términos de los conteos

$$Especificidad = \frac{VN}{TN} \tag{1.25}$$

Una medida más general, para tratar de cuantificar qué tan buena es la clasificación realizada por el modelo, es la tasa de buena clasificación, definida como:

$$TBC = \frac{VP + VN}{VP + FP + VN + FN} \tag{1.26}$$

Así pues, la curva de ROC no es más que la grafica de la Sensibilidad *vs* 1-Especificidad para distintos puntos de corte. Mientras más cercana esté la curva al borde izquierdo y al borde superior de la gráfica, más preciso es el modelo, ya que esto indica que tiene valores altos en su sensibilidad y su especificidad, mientras que si la curva esta pegada a la identidad, menos preciso es el modelo. [1]

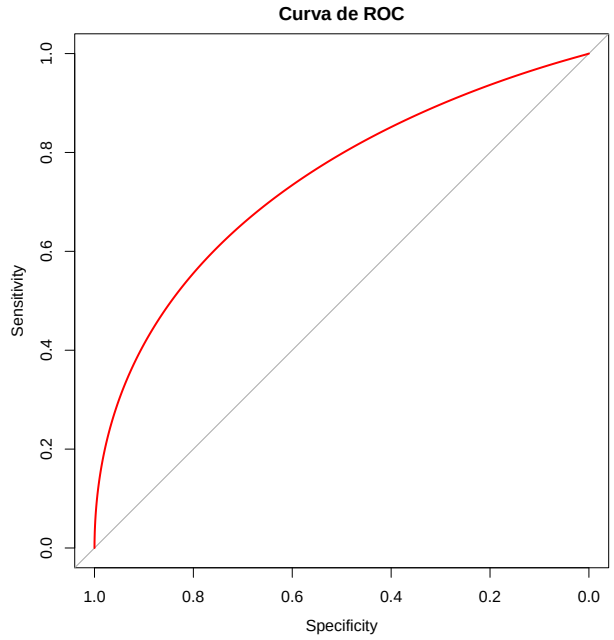


Figura 1.2: Curva de ROC.

El área bajo la curva determina el poder predictivo del modelo ajustado, si el valor de esta área es cercano a 0.5 entonces el modelo no es confiable, ya que predice mal a la respuesta, mientras que, entre más cercano esté a la unidad la capacidad de discriminación del modelo es mejor. No existen pruebas estadísticas que determinen si el poder predictivo es bueno o malo, sin embargo, el juicio depende del área de aplicación y de los objetivos del estudio.

1.6.2. Medidas de influencia

Los datos influyentes pueden ocasionar problemas como sesgo en la estimación de los coeficientes o errores estándar muy grandes, lo que conlleva a la obtención de inferencias estadísticas inválidas. A continuación se presenta una idea general de cómo identificar observaciones que pudieran llegar a tener impacto significativo en el ajuste del modelo.

Palancas: La palancas cuantifican la influencia de cada observación sobre la estimación del vector de parámetros o sobre las predicciones hechas a partir del mismo. Cuanto más grande es una palanca, mayor es la influencia que ejerce la observación asociada en la estimación del modelo.

Dada la matriz de proyección para el modelo de regresión logística [6]

$$H = \hat{W}^{1/2} X (X' \hat{W} X)^{-1} X' \hat{W}^{1/2} \tag{1.27}$$

El apalancamiento para la i-ésima observación está dado por la i-ésima entrada de la diagonal de la matriz H, h_{ii}.

Distancia de Cook: Cuantifica la influencia de una sola observación en la estimación del vector de parámetros $\hat{\beta}$. Valores grandes de esta medida indican que las observaciones en cuestión son influyentes. [6]

La distancia de Cook está dada por:

$$Cook_i = \frac{1}{p} (\hat{\beta} - \hat{\beta}_{(i)})' (X'WX) (\hat{\beta} - \hat{\beta}_{(i)}) \tag{1.28}$$

Con $\hat{\beta}_{(i)}$ el vector de parámetros estimado sin la i-ésima observación.

Dfbetas Cuantifica la influencia de la i-ésima observación en la estimación de alguno de los parámetros $\hat{\beta}_j$

$$Dfbeta_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{SE(\hat{\beta}_j)} \tag{1.29}$$

1.6.3. Multicolinealidad

La Multicolinealidad se produce cuando dos o más variables independientes en el modelo forman una combinación lineal. Por ejemplo, tendríamos un problema de multicolinealidad

si tuviéramos altura medida en centímetros y altura medida en pies en el mismo modelo. El grado de multicolinealidad puede variar y puede tener diferentes efectos en las estimaciones.

Cuando se produce multicolinealidad perfecta, es decir, cuando una variable independiente es una combinación lineal perfecta de otra variable, la estimación de los coeficientes no es única.

La Multicolinealidad moderada es bastante común ya que cualquier correlación entre las variables independientes es un indicador de la colinealidad. Sin embargo cuando se produce la multicolinealidad severa, los errores estándar de los coeficientes tienden a ser muy grandes.

Algunas sugerencias para resolver el problema de la multicolinelidad son:

- Eliminar del modelo la variable menos necesaria implicada en la colinealidad,
- Crear variables sintéticas mediante la aplicación de las técnicas multivariadas de análisis de componentes principales[9] y análisis factorial.

1.6.4. Sobredispersión

Este problema se presenta cuando la varianza observada en los datos es mayor a la varianza esperada del modelo logístico. El problema de sobredispersión puede ser ocasionado por la exclusión de variables explicativas que son necesarias para explicar la variabilidad de la respuesta. Así también, puede ser causada por no incluir términos de interacción, por la presencia de valores atípicos o por emplear una función liga incorrecta. Una forma empírica de detectar la sobredispersión es mediante la devianza: Si la devianza del modelo es mayor a su esperanza teórica, es decir si

$$D > n - p$$

(con n el número de observaciones y p el número de parámetros en el modelo ajustado) entonces tenemos evidencia de sobredispersión en el modelo.

Un alternativa para lidiar con esta problemática es considerar un parámetro de dispersión ϕ, usualmente estimado de la siguiente forma:

$$\phi = \frac{\chi^2}{n - p} \tag{1.30}$$

donde χ^2 es el valor de la estadística en la prueba Ji-cuadrada de Pearson para bondad de ajuste,presentada en 1.22, n es el tamaño de la muestra y p es el número de parámetros que contiene el modelo ajustado.

[9]Ver apéndice A

Posteriormente se considera una función de varianzas para la variable respuesta Y, que contenga dicho parámetro de dispersión, para el caso del modelo de regresión logística, tenemos:

$$\mathbb{V}(Y) = \phi \left(\frac{\mu_i(n_i - \mu_i)}{n_i} \right)$$

con

$$\mu_i = n_i \pi_i = n_i \left(\frac{exp\{\beta^t \mathbf{X}\}}{1 + exp\{\beta^t \mathbf{X}\}} \right)$$

Así pues, con ayuda de las ecuaciones de verosimilitud generadas por estas funciones, se estima el vector de parámetros β empleando el método de máxima verosimilitud.

Capítulo 2

Planteamiento y descripción del problema

2.1. Motivación

En el área de la psicología se han realizado diversos estudios y análisis con la finalidad de establecer los factores que intervienen en el desempeño escolar de los alumnos. El rendimiento académico ha sido definido como el "nivel de conocimientos demostrado en un área o materia de acuerdo con la norma de edad y nivel académico" [4]

Estudios de tipo social acerca del rendimiento escolar, indican que algunas variables relacionadas con éste, son de tipo socioeconómico, la amplitud de los programas de estudio, las metodologías de enseñanza empleadas en las aulas, los incentivos y/o motivaciones que reciben los alumnos, entre otras. Adicionalmente, los expertos afirman que el estrés la ansiedad y la depresión son factores que originan problemas, como el bajo desempeño académico, adicionalmente, el ingreso a la universidad y las dificultades académicas representan un conjunto de situaciones estresantes que pueden descontrolar a los estudiantes.

Partiendo de la premisa de que el rendimiento académico es un fenómeno multifactorial, se pretende identificar específicamente las variables que influyen de manera determinante en el éxito o fracaso de los estudiantes de primer año de la carrera de Médico-Cirujano de la facultad de Medicina de la Universidad Nacional Autónoma de México (UNAM).

2.2. Objetivo

El objetivo del presente análisis es desarrollar un modelo estadístico que permita ampliar el conocimiento sobre las variables asociadas a la reprobación en alumnos de primer año de la licenciatura Médico-Cirujano de la Facultad de Medicina de la UNAM. Mediante la aplicación de un modelo de regresión logística de respuesta binaria y con información acerca

del estado académico, psicológico, de habilidades y aptitudes vocacionales de los alumnos que ingresaron a la carrera en el mes de agosto del 2013, se pretende detectar cuáles son las variables que mejor predicen el desempeño académico de los estudiantes.

2.3. Información disponible

La población que consideraremos en este estudio está conformada por 1205 alumnos de primer ingreso de la carrera de Médico Cirujano, Generación 2013-2014 impartida en la Facultad de Medicina de la UNAM. Las variables regresoras disponibles provienen de cuatro instrumentos tipo examen de opción múltiple que aplica la Facultad para valorar el estado académico, psicológico, de habilidades y de factores asociados a la elección de la carrera de medicina de los alumnos al momento de ingresar, así como de las calificaciones que obtuvieron los estudiantes durante el primer año de la carrera, incluyendo los resultados de los exámenes departamentales.

Descripción de los Instrumentos

1. **Examen de conocimientos generales.**

   Lo aplica la Dirección General de Evaluación Educativa. Permite identificar el grado de nivel académico que poseen los alumnos al ingresar a la Facultad de Medicina, la prueba se estructura en tres partes: una que consiste en 120 reactivos y evalúan las áreas de matemáticas, física, química, biología, historia universal, historia de México, literatura y geografía. Las variables que tomamos en cuenta en este estudio son generadas al asignar una calificación numérica (escala de 0-100) en cada una de las áreas antes mencionadas.

   Otra parte del examen está formada por 60 reactivos de español que evalúan la comprensión de lectura, gramática, redacción, vocabulario y ortografía. La tercera parte de este examen evalúa los conocimientos que tiene el alumno en el idioma extranjero inglés y los clasifica en alguno de los tres niveles: principiante, principiante alto, intermedio bajo.

2. **Inventario de estrategias de estudio y autorregulación (IEEA).**

   Es un cuestionario integrado por 91 reactivos de opción múltiple (cuatro posibles respuestas) que se refieren a lo que los estudiantes piensan acerca de si mismos cuando

adquieren, organizan, recuerdan y aplican lo que aprenden, así como la manera en la cual evalúan, planean y controlan sus maneras de estudiar.

Los ítems fueron diseñados para que el estudiante autovalorara sus estrategias de estudio y autorregulación[1] en cuatro dimensiones principales:

1) Adquisición, que mide los estilos de adquisición de la información, mismos que pueden ser superficiales o de procesamiento profundo;
2) Estilos de recuperación de la información aprendida, ante tareas y ante exámenes;
3) Estilos de procesamiento de información, que puede ser convergente o divergente;
4) De estilos de autorregulación metacognitiva y metamotivacional, con cuatro dimensiones: la dimensión personal, que incluye las escalas de eficacia percibida, de autonomía percibida, de contingencia interna y de orientación a la aprobación externa; la dimensión tarea, que mide la orientación al logro de metas y a la tarea en sí; la dimensión materiales, que registra la adecuación de los mismos. [10]

Mediante este cuestionario se cuantifican los siguientes factores: adquisición selectiva, adquisición generativa, recuperación ante exámenes, recuperación ante tareas, procesamiento convergente, procesamiento divergente, eficacia percibida, contingencia percibida, autonomía percibida, aprobación externa, tarea, logro tarea y materiales.

3. **Aspectos psicosociales.**

Se compone de 155 reactivos. Es una herramienta validada y un proyecto del Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza de la UNAM. Se aplica para tener un perfil inicial en cuanto a los factores psicológicos, de personalidad y de salud de los estudiantes. Por ser un instrumento donde se consideran aspectos personales, el alumno tiene la libertad de contestar o no este examen. La participación fue totalmente voluntaria y los resultados fueron utilizados cuidando la confidencialidad de los mismos.

La prueba se compone en 4 diferentes secciones: La parte correspondiente al inventario de Beck, que contiene reactivos cuya finalidad es detectar la severidad de una depresión. Esta compuesto por ítems relacionados con síntomas depresivos, como la desesperanza e irritabilidad, cogniciones como culpa o sentimientos de estar siendo castigado, así como síntomas físicos relacionados con la depresión. La puntuación total varía de 0 a 63.

La segunda parte corresponde al índice SCL-90 (symptom check list) compuesta por 90 reactivos con respuesta en escala Likert con cinco niveles de puntuación (del 0 al 4), tiene como objetivo evaluar patrones de somatizaciones, obsesiones y compulsiones, sensibilidad interpersonal, depresión, ansiedad, hostilidad, ansiedad fóbica, ideación paranoide, etc. La tercera parte, el índice ASRS se refiere a la detección de problemas del

[1] actividad cognitiva constructiva autorregulada, mejor conocida como estudiar

trastorno por déficit de atención e hiperactividad. Finalmente el Inventario Multifásico de Personalidad es una prueba de personalidad que busca identificar el perfil de personalidad y la detección de psicopatologías.

4. **Factores asociados a la elección de la carrera de Medicina.**

   Es elaborado por la Dirección General de Orientación y Servicios educativos de la UNAM, mediante este instrumento, se evalúan puntos como el razonamiento abstracto, aptitudes (mecánica y de discriminación de figuras), reconocimiento de palabras, intereses y algunos aspectos de personalidad, pero sobre todo la vocación con la que ingresan los alumnos a la facultad de medicina.

   Dicho instrumento incluye los siguientes factores: razonamiento abstracto, aptitud mecánica, ensamble de formas, ciencias físicas, mecánico, matemáticas, ciencias biológicas y de la salud, ecología y medio ambiente, altruismo/servicio social, político, ciencias sociales, administrativo-financiero, organizacional/ persuasivo, artístico plástico visual, expresión musical, expresión oral, expresión escrita, estabilidad y normas, inseguridad personal, incomodidad social, impulsividad, deseabilidad social, mentiras, desajuste, control social, autoestima, autoeficacia, temor al fracaso, evitación del trabajo retador, maestría, orientación al logro, morosidad, liderazgo, negativismo, manipulación social, inseguridad en la elección de carrera, información profesiográfica, seguridad y satisfacción en la elección vocacional, obstáculos familiares y económicos, dificultad en la integración escolar, autoeficiencia en medicina, locus de control externo, prestigio social e interés social.

5. **Calificaciones obtenidas en cada una de las 8 materias del primer año.**

   El plan de estudios vigente entró en vigor con la Generación 2011, es un plan anual, el modelo educativo es mixto ya que está estructurado en base a asignaturas con desarrollo de competencias y tiene como objetivo impartir enseñanza a la vanguardia de las tendencias de la educación médica nacional e internacional, que respondan a la situación cambiante del sistema de salud y a las necesidades y expectativas de la sociedad.

   El Plan de Estudios 2010 de la carrera de Médico Cirujano está integrado por 63 asignaturas que se cursan en cinco año y medio más uno de Servicio Social. En los dos primeros años se cursan las asignaturas de las áreas de bases biomédicas, clínica y sociomédica-humanística. A partir del tercer año los estudiantes asisten a las sedes hospitalarias, donde cursan asignaturas del área clínica y sociomédicas.

   Las materias que conforme al plan de estudios vigente, se cursan durante el primer año de la licenciatura de Médico Cirujano son:

- Anatomía (4 parciales),
- Embriología Humana (4 parciales),
- Bioquímica y Biología Molecular (4 parciales)
- Biología Celular e Histología Médica (3 parciales)
- Integración Básico-Clínica I (2 parciales)
- Informática Biomédica (2 parciales)
- Introducción a la Salud Mental (2 parciales)
- Salud Pública y Comunidad (2 parciales)

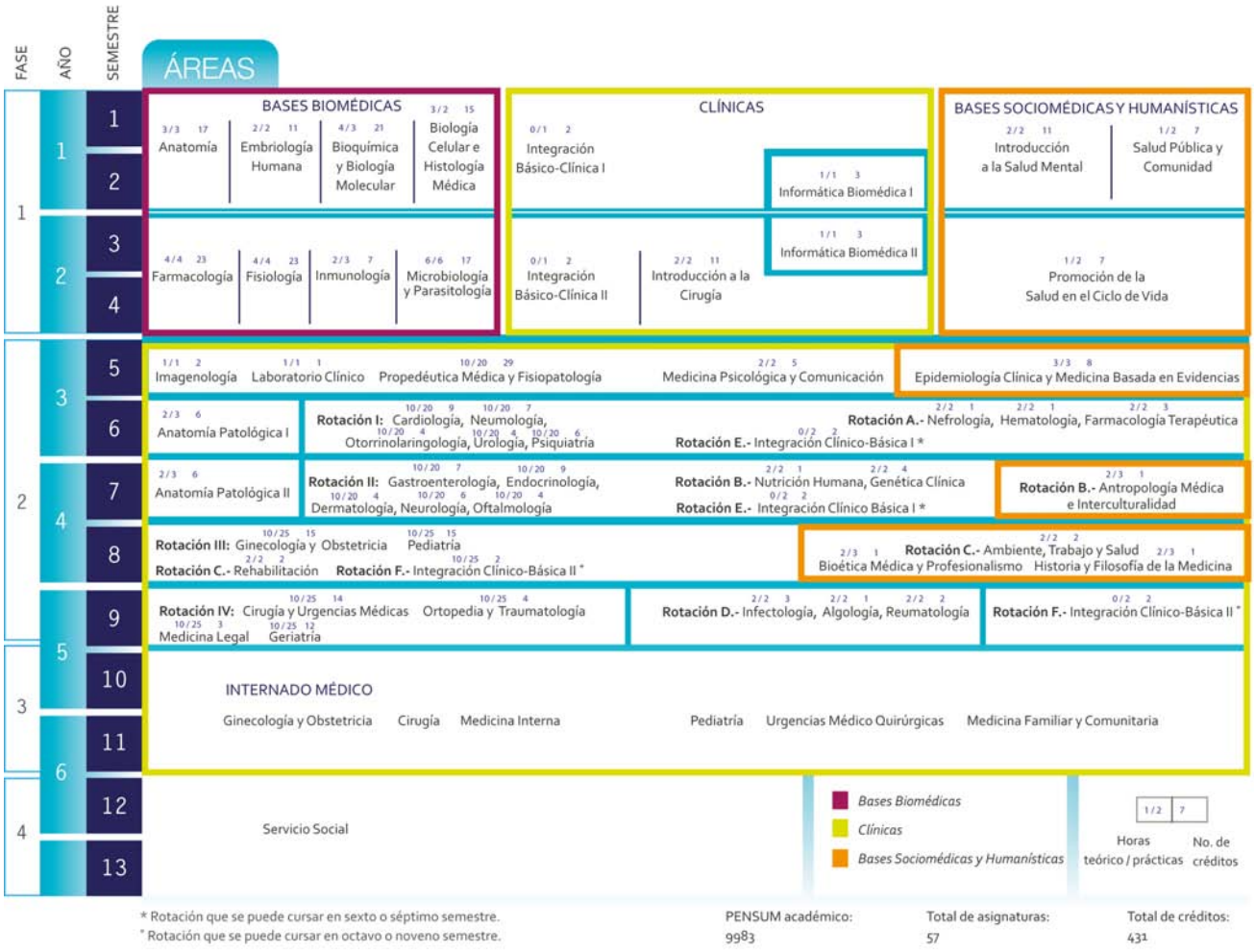


Figura 2.1: Plan de estudios vigente.

Hasta el 2014 los lineamientos de evaluación eran que para excentar cada asignatura, el estudiante debía obtener un promedio mínimo de 8.0, donde la evaluación estaba constituida por 50 % de exámenes departamentales y 50 % de la calificación que otorgaba el profesor. Los exámenes departamentales son estructurados por un cuerpo colegiado y aplicados a todos los estudiantes al mismo tiempo.

Se cuenta con información acerca de las calificaciones de cada una de las asignaturas del primer año durante el ciclo escolar 2013-2014, en éste estudio será considerada la calificación final incluyendo la otorgada por el profesor.

2.4. Contexto de la Facultad de Medicina

Desde hace algunos años, la facultad de Medicina de la UNAM aplica un examen diagnóstico de conocimientos generales a todos los estudiantes de primer ingreso, independientemente del sistema de bachillerato del cuál provengan, este instrumento permite conocer el grado de conocimientos en las áreas de Matemáticas, Física, Química, Biología, Historia Universal, Historia de México, Literatura y Geografía. Al separar a la población de acuerdo al bachillerato de procedencia (ENP, CCH y concurso de selección.) se observó que los estudiantes provenientes del concurso de selección obtienen un mayor porcentaje de aciertos, situación que resulta lógica por el proceso de selección al que fueron sujetos (hasta el 2012, para ser seleccionado era requerido un mínimo de 105 aciertos de un total de 120). En cuanto al plan de estudios, todas las materias del mapa curricular son seriadas, de manera que si el alumno no aprueba alguna de las materias, no podrá continuar cursando ninguna de segundo año, hasta haber acreditado todas y cada una de las de primer año. Se estima que en promedio el 32 % de los alumnos de nuevo ingreso se rezagan por no acreditar una o más materias.

2.5. Criterios de inclusión

Serán considerados en el estudio los alumnos que hayan presentado todos y cada uno de los instrumentos antes mencionados y al menos un parcial de las siguientes asignaturas pertenecientes a las bases biomédicas, a saber:

1. Anatomía
2. Bioquímica y Biología Molecular
3. Biología Celular e Histología Médica
4. Embriología Humana

De los 1085 alumnos que presentaron al menos una de las ocho asignaturas de primer año 18 de ellos no presentarón la prueba de conocimientos generales, 35 más no presentaron la prueba de aspectos psicosociales, 26 no presentaron la prueba de Español, 22 no presentaron el inventario de estrategías de estudio y autorregulación y 68 no presentario la prueba de factores asociados a la carrera de medicina. De modo que despues de aplicar los criterios de inclusión establecidos, la base de datos resultante contiene 925 observaciones de 72 variables.

Capítulo 3

Análisis exploratorio de los datos

Los resultados aquí presentados fueron obtenidos con la ayuda del software estadístico R.

3.1. Información general

Considerando los datos obtenidos después de aplicar los criterios de inclusión mencionados en el *capítulo 2* y con la finalidad de explorar las relaciones entre las variables categóricas con las que se cuenta, realizamos un breve análisis de tablas de contingencia [1]

La variable respuesta para nuestro modelo es llamada *éxito académico* y está codificada como 0 si el alumno reprobó al menos una de las materias de primer año o 1 en caso de haber aprobado todas y cada una de las ocho asignaturas de primer año.

Tabla de clasificación cruzada: Rendimiento académico y Bachillerato de procedencia. El objetivo es explorar la relación entre el bachillerato de procedencia del alumno y el éxito académico.

Cuadro 3.1: Rendimiento académico vs Bachillerato de procedencia

	CCH	OTRO	ENP	Total
Aprobado	103	82	302	487
No Aprobado	218	41	179	438
Total	321	123	481	925

Al desplegar la distribución conjunta y las marginales de esta tabla obtenemos:

[1] ver Apéndice A.

Cuadro 3.2: Rendimiento académico vs Bachillerato de procedencia, conjunta y marginales

	CCH	OTRO	ENP	Total
Aprobado	0.111	0.089	0.326	0.526
No Aprobado	0.236	0.044	0.194	0.474
Total	0.347	0.133	0.520	1

La probabilidad estimada de que un alumno seleccionado aleatoriamente de la muestra provenga de CCH y presente un bajo rendimiento académico es de 0.236, mientras que la probabilidad estimada de que un alumno provenga de otro bachillerato y presente un bajo rendimiento académico es de 0.044, finalmente la probabilidad de que un alumno provenga de Escuela Nacional Preparatoria (ENP) y presente un bajo rendimiento académico es de 0.19.

Por otra parte, las distribuciónes marginales de esta tabla indican que 52.6 % de los estudiantes obtienen notas aprobatorias en todas y cada una de sus materias, mientras que 47.4 % de los estudiantes reprueban por lo menos una de las asignaturas de primer año (bajo rendimiento académico).

La distribución marginal de Y indica que 34.7 % de los alumnos bajo estudio provienen de CCH, 13.3 % provienen de otro bachillerato y 52 % de ENP.

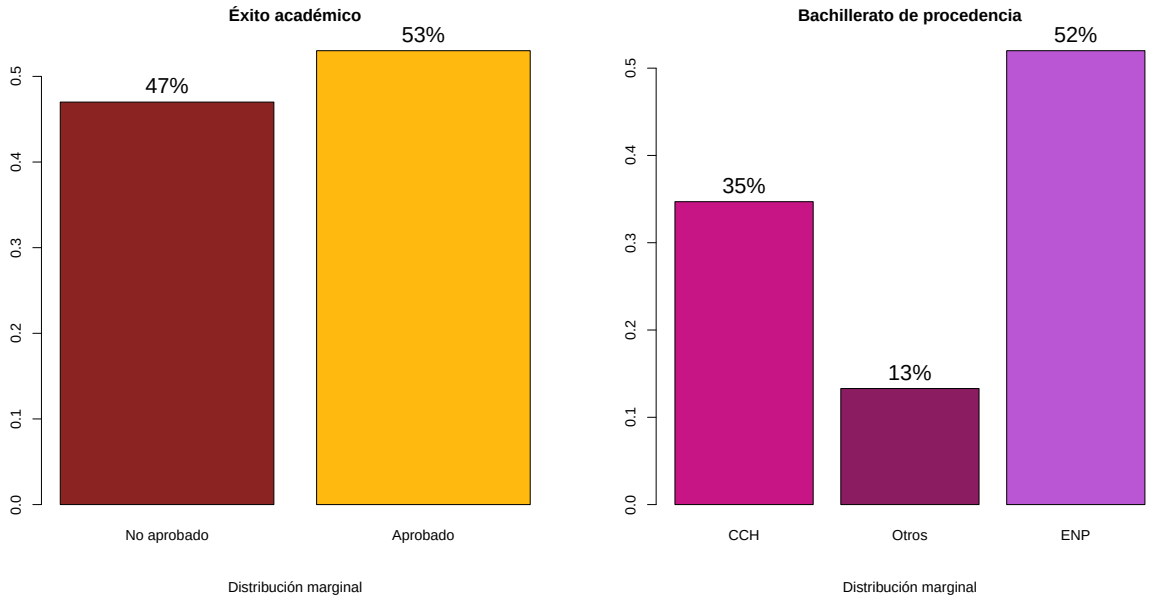


Figura 3.1: Distribuciones marginales asociadas

Ahora bien, al desplegar la distribución condicional del rendimiento académico dado el bachillerato de procedencia, se obtiene que la probabilidad estimada de presentar un bajo rendimiento académico dado que el alumno proviene de CCH es de 0.68, mientras que la

probabilidad estimada de presentar bajo rendimiento académico dado que se proviene de otro bachillerato es de 0.33 y finalmente, la probabilidad estimada de presentar un bajo rendimiento académico dado que se proviene de ENP es 0.37

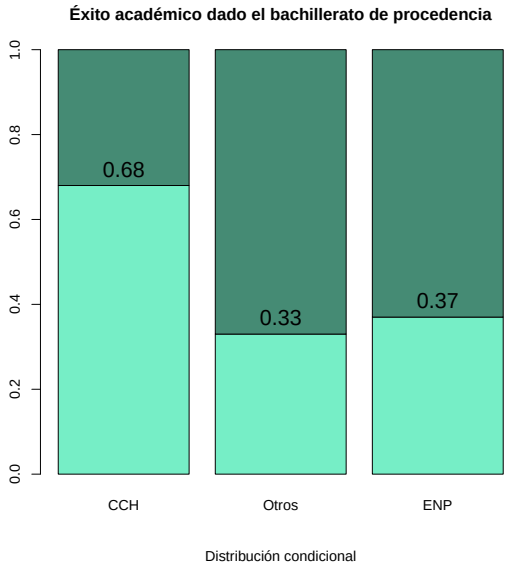


Figura 3.2: Probabilidades estimadas de reprobación dado el bachillerato de procedencia

Así pues, concluimos que los alumnos provenientes de CCH son el sector más vulnerable a presentar un bajo rendimiento académico, para sustentar esta afirmación, aplicamos algunas medidas de asociación y pruebas para contrastar la hipótesis nula de independencia entre variables categóricas[2].

La prueba Ji-cuadrada de independencia, cuyas hipótesis a contrastar son:

$$H_0 : \textbf{Las variables X y Y son independientes}$$
$$vs$$
$$H_a : \textbf{Las variables no son independientes}$$

Arrojó un p-valor de $2.2\text{x}10^{-16}$ por lo que a un nivel de significancia de 95 % se rechaza la hipótesis nula y concluimos que el bachillerato de procedencia y el rendimiento académico están relacionados.

Para medir la fuerza de asociación entre este par de variables empleamos el coeficiente de contingencia que resulto ser de 0.28 y la V de cramér que resultó ser de 0.30, por lo que pese a que la asociación es no nula, si es relativamente baja.

El gráfico de los residuos (figura 3.3) indica que los alumnos provenientes de CCH estan asociados de manera positiva con la condición de reprobación, mientras que los alumnos de otros bachilleratos presetan una asociación negativa (débil) con la condición de reprobación,

[2]ver Apéndice A

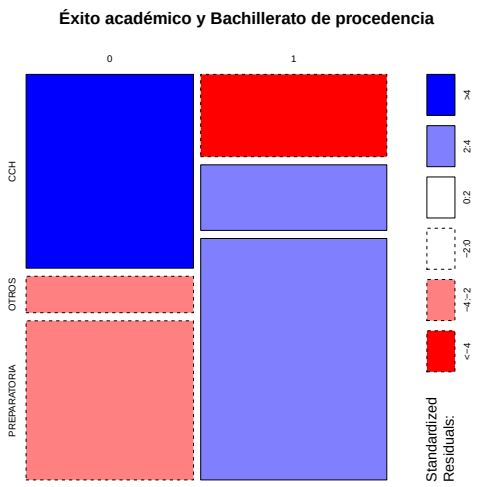


Figura 3.3: Gráfico de mosaico

al igual que los alumnos provenientes de ENP.

Tabla de clasificación cruzada: Género y Bachillerato de procedencia.

A continuación se presenta la tabla de contingencia entre las variables Género (M ó F) y Bachillerato de procedencia (CCH, OTRO, ENP).

Cuadro 3.3: Género vs Bachillerato de procedencia

	CCH	OTRO	ENP	Total
H	102	69	174	345
M	219	54	307	580
Total	321	123	481	925

Las distribuciones conjunta y marginales para esta tabla están dadas por:

Es decir, la probabilidad estimada de que un sujeto elegido aleatoriamente de la muestra sea hombre y provenga de CCH es de 0.11, mientras que la probabilidad estimada de que sea hombre y provenga de otro bachillerato es 0.075 y la probabilidad de que sea hombre y provenga de ENP es de 0.18. Existe una mayor proporción de mujeres que de hombres que provienen de CCH y de ENP, mientras que existe una mayor proporción de hombres que de mujeres que provienen de otros bachilleratos.

Al analizar la distribución condicional del género dado el bachillerato de procedencia obser-

Cuadro 3.4: Género vs Bachillerato de procedencia, conjunta y marginales

	CCH	OTRO	ENP	Total
H	0.110	0.075	0.188	0.373
M	0.237	0.058	0.332	0.627
Total	0.347	0.133	0.520	1

vamos que la probabilidad estimada de ser hombre dado que se proviene de CCH es de 0.32, mientras que la probabilidad estimada de ser mujer dado que se proviene de CCH es de 0.68 (mayor probabilidad para mujeres que provienen de CCH). La probabilidad estimada de ser hombre dado que se proviene de otro bachillerato es de 0.56, mientras que la probabilidad estimada de ser mujer dado que se proviene de otro bachillerato es de 0.43 (mayor probabilidad para hombres provenientes de otros bachilleratos) y finalmente, la probabilidad estimada de ser hombre dado que se proviene de ENP es de 0.36, mientras que la probabilidad estimada de ser mujer dado que se proviene de ENP es de 0.64.

Cuadro 3.5: Género vs Bachillerato de procedencia, distribución condicional

	CCH	OTRO	ENP
H	0.32	0.56	0.36
M	0.68	0.44	0.64

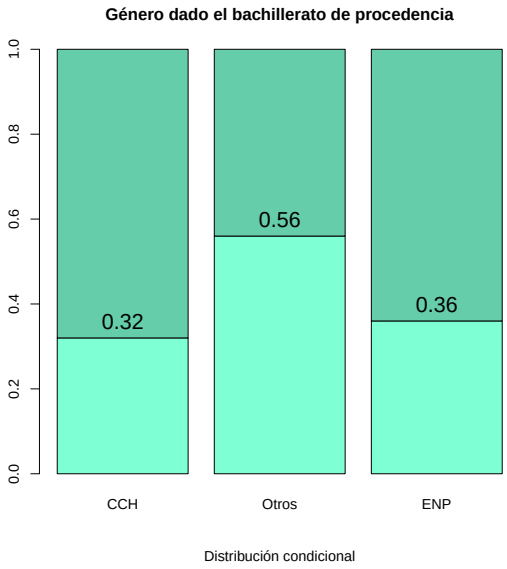


Figura 3.4: Probabilidad de ser hombre o mujer dado el bachillerato de procedencia

Al aplicar la prueba Ji-cuadrada sobre esta tabla obtuvimos un p-valor de $9.962\text{x}10^{-6}$, con lo que se rechaza la independencia y concluimos que el género y el bachillerato de procedencia

están asociados. Sin embargo, el coeficiente de contingencia es de 0.15 al igual que la V de Cramér, por lo que concluimos que la asociación entre este par de variables es positiva débil.

Al analizar las tablas de contingencia de género *vs* rendimiento académico, pareciera ser que las mujeres son un sector más propenso a la reprobación, sin embargo, esto se debe a que el género está a su vez relacionado con el bachillerato de procedencia, esto sugiere que una interacción en nuestro modelo final podría ser adecuada.

3.2. Examen de Conocimientos generales

Como se mencionó con anterioridad, esta prueba evalúa los conocimientos generales de los alumnos en las áreas de Matemáticas, Física, Química, Biología, Historia Universal, Historia de México, Literatura y Geografía. La escala de calificación para cada una de las asignaturas evaluadas va del cero al cien, donde puntajes altos indican un mayor grado de conocimientos en la asignatura correspondiente. Mil ciento setenta y seis estudiantes de nuevo ingreso presentaron esta prueba obteniendo los resultados mostrados a continuación.

Resultado global de la prueba de conocimientos generales

Mín	1er cuartil	Mediana	Media	3er cuartil	Máx
25.42	48.31	57.63	59.17	68.64	96.61

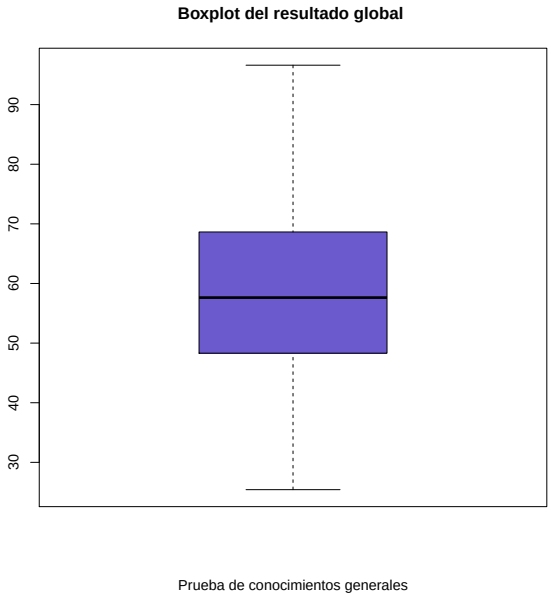
Al analizar la variable que denota el resultado global de la prueba notamos que el rango intercuartil[3] es de 20.33, no se tienen datos atípicos. El promedio mínimo resultó ser de 25.42, mientras que el máximo fue de 96.61; sin embargo, la media fue de 59.17. Es decir, los alumnos de primer ingreso que presentaron el examen de conocimientos generales obtuvieron en promedio una calificación de 5.9 sobre 10.

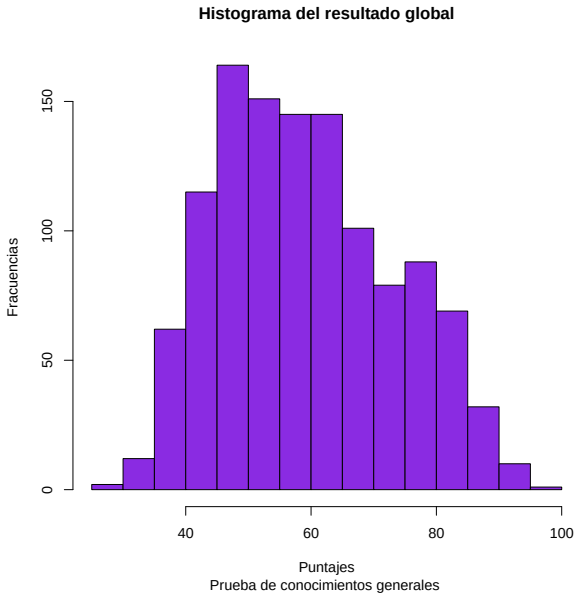
El histograma muestra que la mayor frecuencia se concentra en alumnos que obtuvieron entre 45 y 50 aciertos de un total de 100. La distribución empírica de nuestros datos es ligeramente sesgada a la izquierda.

Distinguiendo los resultados por género, del número total de alumnos que presentaron la prueba, 438 eran hombres y 738 eran mujeres, lo que representa el 37.24 % y el 62.76 % respectivamente.

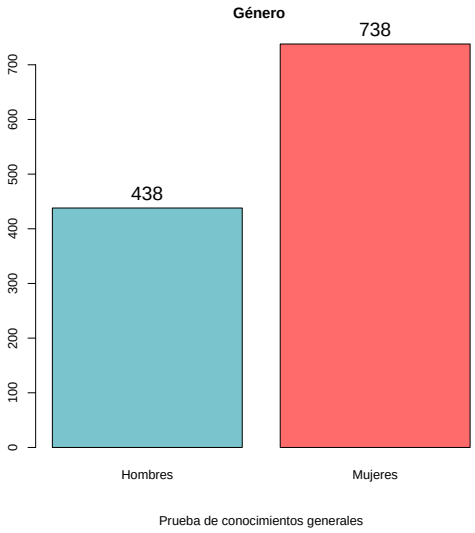
Observamos que los hombres obtuvieron 63 aciertos en promedio, mientras que las mujeres únicamente 57. Sin embargo, los datos para las mujeres están ligeramente menos dispersos que los de los hombres. Cabe destacar que esta diferencia es ocasionada por el hecho de que

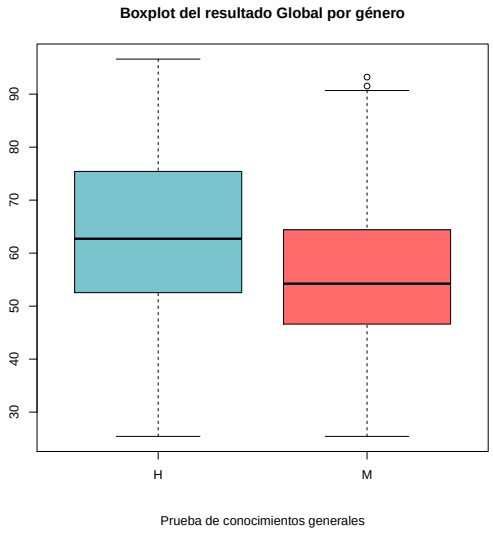
[3]el tercer cuartil menos el primer cuartil. $Q_3 - Q_1$





el bachillerato de procedencia está asociado con el resultado de la prueba y a su vez el género y el bachillerato de procedencia también son un par de variables asociadas. [4]



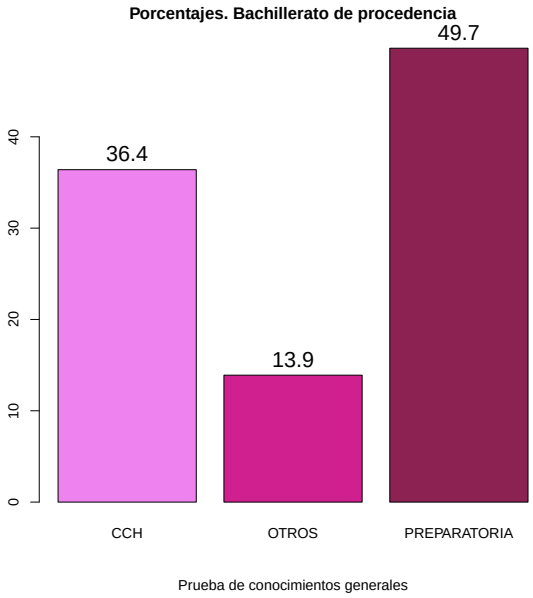


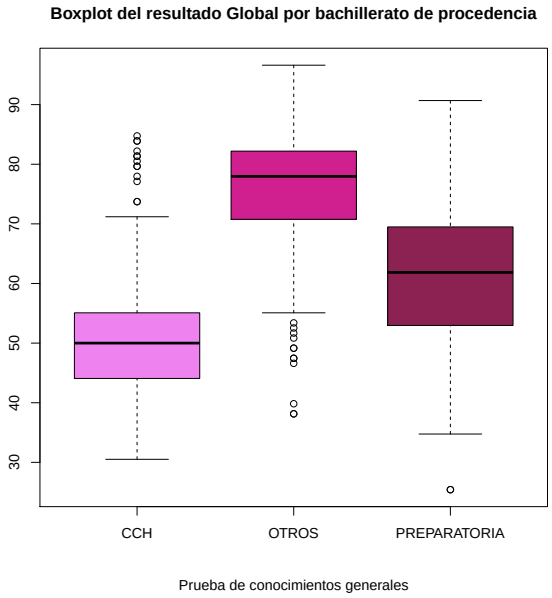
Ahora bien, distinguiendo por bachillerato de procedencia, observamos que 36.4 % provienen de CCH, 49.7 % provienen de ENP y únicamente el 13.9 % provienen de otro bachillerato [5].

Para exalumnos de CCH observamos que la distribución de sus puntajes está ligeramente sesgada a la izquierda, de hecho la media del resultado Global para este grupo es de 50.37 y el rango intercuartil es de 11.01. Por otra parte para exalumnos de la Escuela Nacional Preparatoria (ENP) tenemos una distribución centrada, con media 61.5 y un rango intercuertil

[4] pareciera que las mujeres obtienen resultados inferiores, sin embargo esto se observa por que tenemos una mayor cantidad de mujeres provenientes de CCH y de ENP.

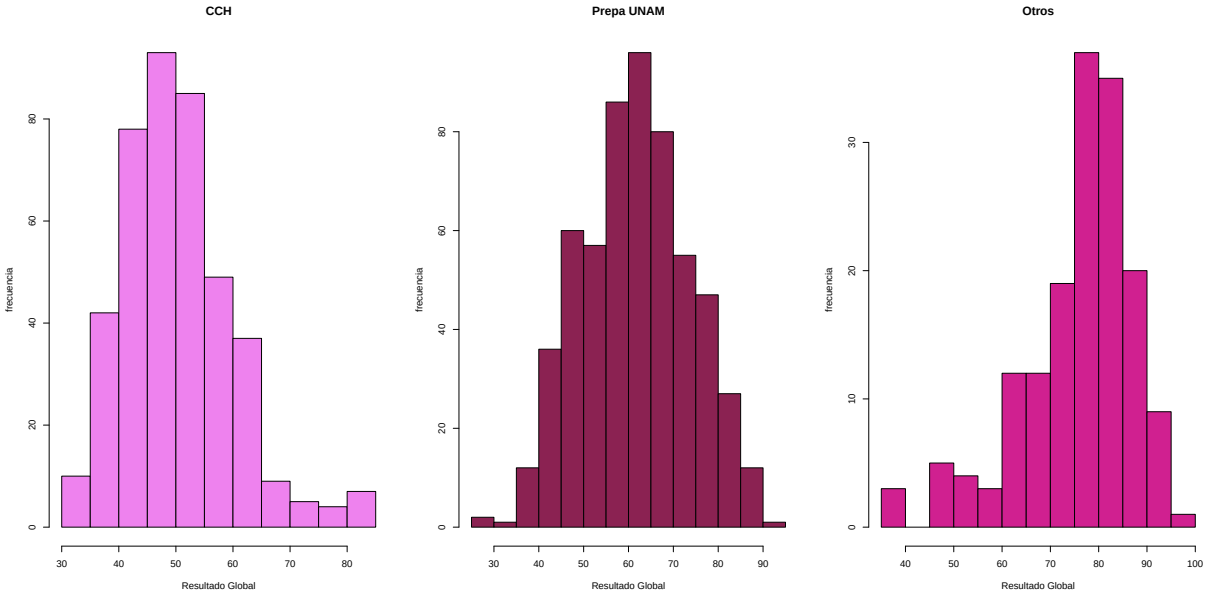
[5] estos últimos son los alumnos que ingresaron a la universidad mediante el concurso de ingreso.





de 16.31. Finalmente para exalumnos de otros bachilleratos tenemos una distribucón sesgada a la derecha, la media del resultado global de la prueba de conocimientos generales para esta población es de 75.5 y un rango intercuartil de 11.22.

El hecho de que los alumnos provenientes de otros bachilleratos obtengan puntajes considerablemente más altos que los alumnos que provienen de bachilleratos de la UNAM, se debe a la condición de que estos últimos no son sometidos a ningun examen de admisión para ingresar a la Facultad de Medicina, a diferencia de alumnos que provienen de otros bachilleratos, quienes deben presentar un examen de admisión para poder ingresar a la Facultad.



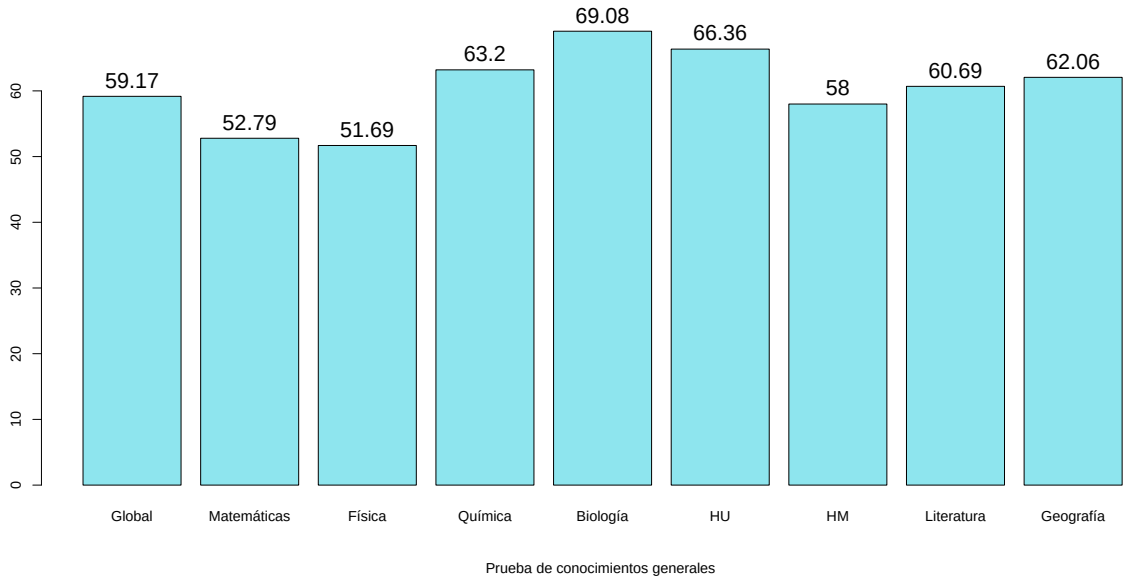


Figura 3.5: Número promedio de aciertos por asignatura.

En cuanto al promedio de calificaciones las mejores notas se obtuvieron en la asignatura de biología (promedio de 6.9 sobre 10) mientras que la materia en la cual los estudiantes recibieron peores notas fue Física (promedio de 5.2 sobre 10). En la siguiente gráfica se muestran el número promedio de aciertos por asignatura.

3.3. Inventario de Estregias de Estudio y Autorregulación (IEEA)

Como se aclaró antes, las afirmaciones del cuestionario de estrategias de estudio y autorregulación se refieren a lo que los estudiantes piensan acerca de si mismos cuando adquieren, organizan, recuerdan y aplican lo que aprenden. A través de este instrumento se logran cuantificar los factores siguientes:

1. adquisición selectiva,
2. adquisición generativa,
3. recuperación ante tareas,
4. recuperación ante exámenes,
5. procesamiento convergente,
6. procesamiento divergente,
7. eficacia percibida,
8. contingencia percibida,
9. autonomía percibida,
10. aprobación externa,
11. tarea,
12. logro tarea,
13. materiales.

La escala de medición de cada uno de los factores mencionados antes va del 0 al 12 donde 0 a 4 puntos indican niveles bajos, de 5 a 8 puntos indican niveles medios, y 9 a 12 puntos indican niveles altos [6].

Asignando los umbrales anteriores, los factores pueden ser tratados como variables categóricas ordinales. [7] De los mil ciento setenta y dos alumnos de primer ingreso que realizaron la prueba se obtuvieron los resultados siguientes:

[6] índices altos corresponden a perfiles de alumnos que se interesan más por sus actividades académicas
[7] los niveles de la variable tienen un orden, alto, medio y bajo en este caso.

Adquisición Selectiva y Adquisición Generativa,

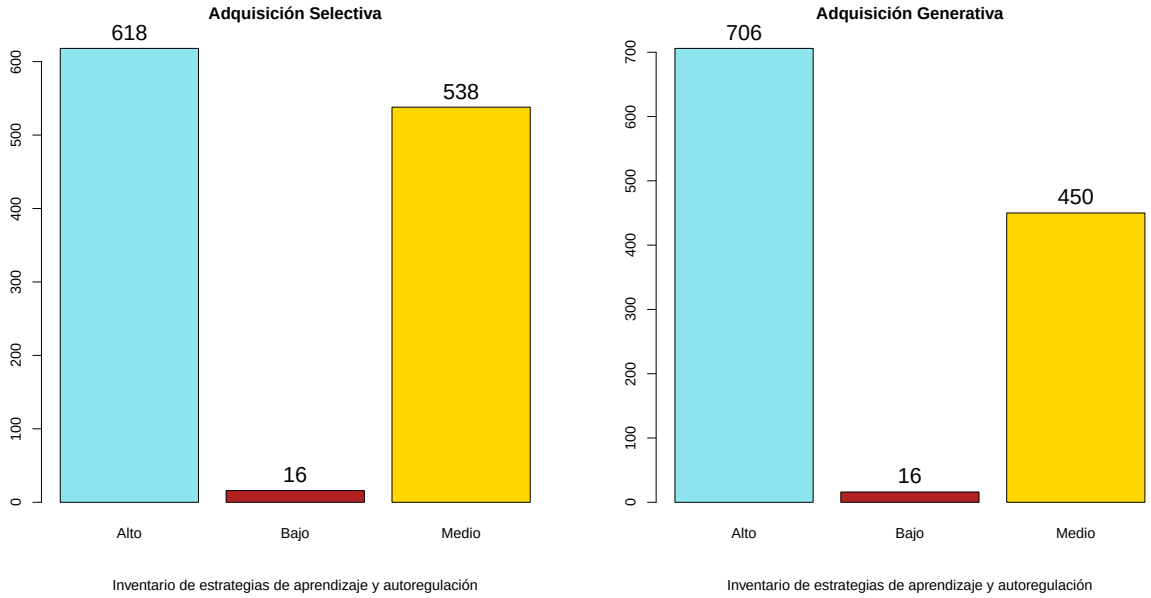
Se refiere al estilo de adquisición de información ante diferentes tareas académicas y para presentar exámenes, en este caso se distinguen dos estilos: el **selectivo** que se refiere al procesamiento superficial [8] y el **generativo** que se refiere al procesamiento profundo[9]. En ambos factores la gran mayoria de los individuos bajo estudio presentaron niveles altos.

Cuadro 3.6: Adquisición Selectiva

Nivel	Frecuencia	Porcentaje
Bajo	16	1.36 %
Medio	538	45.9 %
Alto	618	52.7 %

Cuadro 3.7: Adquisición Generativa

Nivel	Frecuencia	Porcentaje
Bajo	16	1.36 %
Medio	450	38.4 %
Alto	706	60.2 %



Con el objetivo de explorar la relación entre estos factores y el *éxito académico*[10] realizamos el análisis de la tabla de clasificación cruzada[11] con la base de datos resultante después de haber aplicado los criterios de inclusión establecidos en el *capítulo 2*.

[8] poco esfuerzo, recuerdo breve.
[9] organizar y modificar información. Mayor tiempo.
[10] la variable respuesta para el modelo.
[11] ver apéndice A.

Cuadro 3.8: Rendimiento académico vs Aquisición Generativa

	Alto	Bajo	Medio	Total
Aprobado	286	6	195	487
No Aprobado	211	6	221	438
Total	497	12	416	925

Las distribuciones marginal y conjunta están dadas por:

Cuadro 3.9: Rendimiento académico vs Aquisición Generativa, conjunta y marginales

	Alto	Bajo	Medio	Total
Aprobado	0.31	0.01	0.21	0.52
No Aprobado	0.23	0.01	0.24	0.48
Total	0.54	0.02	0.45	1

cuya interpretación es similar a las dadas al inicio del capítulo. Ahora bien, al desplegar la distribución condicional del *éxito académico* dado el factor *adquisición selectiva*, observamos que la probabilidad estimada de tener un alto rendimiento académico dado que se obtuvo un nivel alto en adquisición selectiva es de 0.57, mientras que la probabilidad estimada de tener un alto rendimiento académico dado que se obtuvo un nivel medio en adquisición selectiva es de 0.47 y finalmente, la probabilidad estimada de tener un alto rendimiento académico dado que se obtuvo un resultado bajo en el factor adquisicion selectiva es 0.06

Al aplicar la prueba Ji-cuadrada sobre esta tabla obtuvimos un p-valor de 0.005582, por lo que concluimos que el factor *adquisición selectiva* y el *éxito académico* están relacionados, aunque esta relación es débil, ya que la V de Cramér y el coeficiente de contingencia reportan un valor de 0.106 y 0.105 respectivamente.

No se encontró asociación entre el factor *adquisición generativa* y el *éxito académico*, sin embargo, los dos factores: *adquisición selectiva* y *adquisición generativa*, sí están asociados. Las medidas de asociación entre este par de variables son 0.304 para la V de Cramér y 0.395 para el coeficiente de contingencia.

Recuperación ante Tareas y Recuperación ante exámenes,

Se refiere al estilo para **recuperar** la información aprendida ante la realización de tareas académicas y pruebas tipo examen. No se encontró asociación entre el éxito académico de los estudiantes y este par de factores.

Cuadro 3.10: Recuperación ante tareas

Nivel	Frecuencia	Porcentaje
Bajo	15	1.27 %
Medio	395	33.7 %
Alto	762	65.03 %

Cuadro 3.11: Recuperación ante exámenes

Nivel	Frecuencia	Porcentaje
Bajo	28	2.38 %
Medio	546	46.6 %
Alto	598	51.02 %

Procesamiento Convergente y Divergente,

Es el estilo de procesamiento de la información, en términos de reproducir la información aprendida (convergente) y de crear y pensar críticamente sobre lo aprendido (divergente).

Cuadro 3.12: Procesamiento Convergente

Nivel	Frecuencia	Porcentaje
Bajo	10	0.85 %
Medio	543	46.33 %
Alto	619	52.82 %

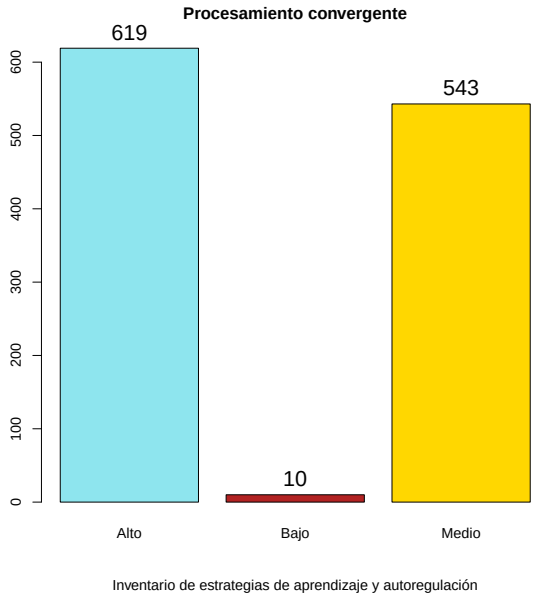
Cuadro 3.13: Procesamiento Divergente

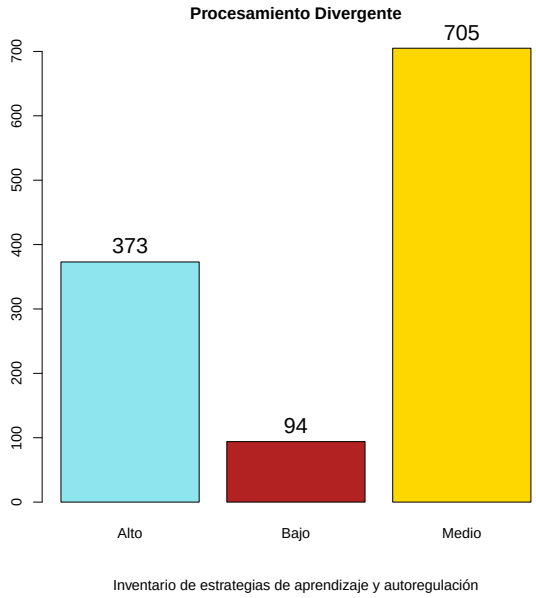
Nivel	Frecuencia	Porcentaje
Bajo	94	8.02 %
Medio	705	60.15 %
Alto	373	31.82 %

No se encontraron asociaciones entre el procesamiento divergente y el éxito académico ni entre el procesamiento convergente y el éxito académico. Sin embargo, estos dos factores si están relacionados entre sí, la V de Cramér reportó un valor de 0.33, mientras que el coeficiente de contingencia resultó ser de 0.43.

Cuadro 3.14: procesamiento convergente vs procesamiento divergente

	Alto	Bajo	Medio	Total
Alto	255	16	226	497
Bajo	0	4	4	8
Medio	42	56	322	420
Total	297	76	552	925





Eficacia percibida, Contingencia percibida y Autonomía percibida

Se refiere al estilo de Autorregulación[12] constituido por tres componentes: los del estudiante como aprendiz, en cuanto a su eficacia (Eficacia Percibida) [13], contingencia interna (Contingencia Percibida), autonomía percibida.[14]

Cuadro 3.15: Eficacia percibida

Nivel	Frecuencia	Porcentaje
Bajo	28	2.39 %
Medio	671	57.25 %
Alto	473	40.36 %

La tabla de clasificación cruzada del factor *eficacia Percibida* con la variable *éxito académico* está dada por:

Cuadro 3.16: Éxito académico vs Eficacia Percibida

	Alto	Bajo	Medio	Total
No aprobado	151	12	275	438
Aprobado	232	11	244	487
Total	383	23	519	925

[12] actividad cognitiva constructiva autorregulada, mejor conocida como estudiar.
[13] que tan eficaz y seguro se siente al alumno al estudiar y adquirir conocimiento
[14] como se siente el estudiante al adquirir conocimientos por si mismo, sin apoyo de un instructor o profesor.

El p-valor al aplicar la prueba de independencia G^2 [15] es de 0.0002644, por lo que a un nivel del 95 % de significancia, concluimos que el éxito académico y el factor *eficacia percibida* están asociados, sin embargo, nuevamente esta asociación es débil al reportar una V de Cramér y un coeficiente de contingencia de 0.13 (en ambos casos).

Cuadro 3.17: Contingencia percibida

Nivel	Frecuencia	Porcentaje
Bajo	11	0.94 %
Medio	337	28.75 %
Alto	824	70.31 %

También existe asociación entre el factor *contingencia percibida* y el *éxito académico*, aunque como en el caso anterior esta asociación es débil. El p-valor de la prueba G^2 resulto ser de 0.010266, por lo que rechazamos la independencia entre variables, sin embargo, las medidas de asociación apenas alcanzan el 0.01

A su vez, las variables *contingencia percibida* y *eficacia percibida* guardan una relación, al aplicar la prueba G^2 de independencia, se rechaza H_0 y las medidas de asociación resultan ser 0.25 para la V de Cramér y 0.34 para el coeficiente de contingencia.

Cuadro 3.18: Autonomía percibida

Nivel	Frecuencia	Porcentaje
Bajo	16	1.37 %
Medio	372	31.74 %
Alto	784	66.9 %

Aprobación externa, Tarea y Logro Tarea

Cuadro 3.19: Aprobación externa

Nivel	Frecuencia	Porcentaje
Bajo	113	9.64 %
Medio	505	43.09 %
Alto	554	47.29 %

Se encontró asociación entre el *éxito académico* y el factor de *tarea*, el p-valor al aplicar la prueba G^2, resultó ser de 0.0025362 y las medidas de asociación V de Cramér y el coeficiente de contingencia toman ambas el valor de 0.104.

[15] prueba asintóticamente equivalente a la prueba Ji-cuadrada, que es generada a través del cociente de verosimilitudes.

Cuadro 3.20: Tarea

Nivel	Frecuencia	Porcentaje
Bajo	6	0.51 %
Medio	390	33.28 %
Alto	776	66.21 %

Cuadro 3.21: Logro tarea

Nivel	Frecuencia	Porcentaje
Bajo	14	1.19 %
Medio	512	43.7 %
Alto	646	55.12 %

Materiales

Se refiere a los instrumentos de evaluación y regulación que emplean los profesores como parte de las evaluaciones que realizan.

Cuadro 3.22: Materiales

Nivel	Frecuencia	Porcentaje
Bajo	7	0.6 %
Medio	398	33.96 %
Alto	767	65.44 %

En resumen, para este instrumento se encontraron asociaciones entre la variable respuesta *éxito académico* y los factores:

1. adquisición selectiva,
2. eficacia percibida,
3. contingencia percibida.

(Después de análizar las tablas de clásificación cruzada correspondientes.)

3.4. Aspectos Psicosociales

Esta prueba se encarga de clasificar a los alumnos como casos probables o no casos, de acuerdo al resultado que obtuvieron en el cuestionario. De los mil ciento cuarenta y seis alumnos que presentaron la prueba se obtuvieron los siguientes resultados:

Tabla de clasificación cruzada: Severidad por Sexo,

Con los datos disponibles generamos la siguiente tabla de contingencia y analizamos las distribuciones asociadas a la misma.

Cuadro 3.23: tabla de contingencia, Severidad de acuerdo al género

	No caso	Probable Caso	Total
Hombre	382	45	427
Mujer	615	104	719
Total	997	149	1146

La probabilidad estimada de ser hombre y tener un diagnóstico de *no caso* es de 0.33, mientras que la probabilidad estimada de ser mujer y tener un diagnóstico de *no caso* es de 0.54. Por otra parte, la probabilidad estimada de ser hombre y tener un diagnóstico de *probable caso* es de 0.04, mientras que la probabilidad estimada de ser mujer y tener un diagnóstico de *probable caso* es de 0.09.

37.25 % de los individuos que presentaron las prueba son hombres y el 62.75 % mujeres. 87 % de los individuos que presentaron la prueba obtuvieron un diagnóstico de *no caso* y el 13 % obtuvieron un diagnóstico de *probable caso*.

Cuadro 3.24: Conjunta y Marginales, Severidad por Sexo

	No caso	Probable Caso	Marginal por renglón
Hombre	0.33	0.04	0.37
Mujer	0.54	0.09	0.63
Marginal por columna	0.87	0.13	1

Adicionalmente aplicamo la prueba Ji-cuadrada de independencia, para explorar la relación entre este par de variables.

H_0:**el sexo y el diagnóstico de la prueba son independientes.**

vs

H_a:**el sexo y el diagnóstico de la prueba estan asociados.**

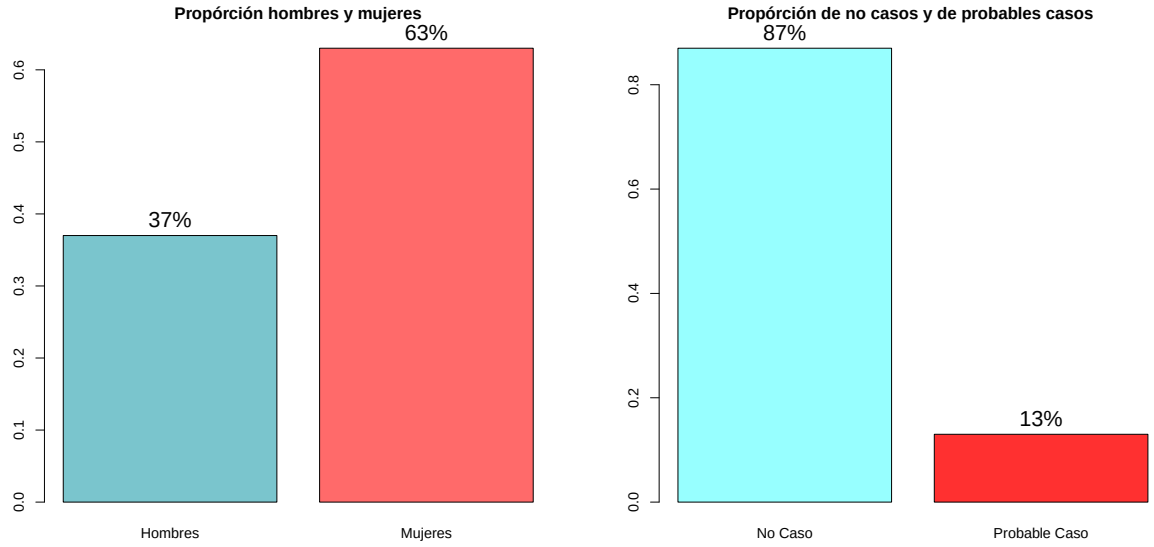


Figura 3.6: Aspectos Psicosociales

A un nivel del 95 % de significancia, no se rechaza la hipótesis de independencia y concluimos que las variables severidad y sexo son independientes.

Tabla de clasificación cruzada: severidad vs bachillerato de procedencia,

La tabla de contingencia asociada al diagnóstico arrojado por la prueba y al bachillerato de procedencia está dada por:

Cuadro 3.25: tabla de contingencia, severidad de acuerdo al bachillerato de procedencia

	No caso	Probable Caso	Total
CCH	361	62	423
ENP	505	72	577
Otro	131	15	146
Total	997	149	1146

La probabilidad estimada de tener un diagnóstico de *no caso* y provenir de CCH es de 0.31, la probabilidad estimada de presentar un diagnóstico de *no caso* y provenir de ENP es de 0.44 y finalmente la probabilidad estimada de tener un diagnóstico de *no caso* y provenir de otro bachillerato es de 0.11.

37 % de los individuos que presentaron las prueba provienen de CCH, 50 % de ENP y 13 %

de otro bachillerato. Como bien se mencionó en el análisis de la tabla anterior, 87 % de los individuos que presentaron la prueba obtuvieron un diagnóstico de *no caso* y el 13 % obtuvieron un diagnóstico de *probable caso*.

Cuadro 3.26: Conjunta y Marginales, Severidad por Bachillerato de Procedencia

	No caso	Probable Caso	Marginal por renglón
CCH	0.31	0.05	0.37
ENP	0.44	0.06	0.50
Otro	0.11	0.01	0.13
Marginal por columna	0.87	0.13	1

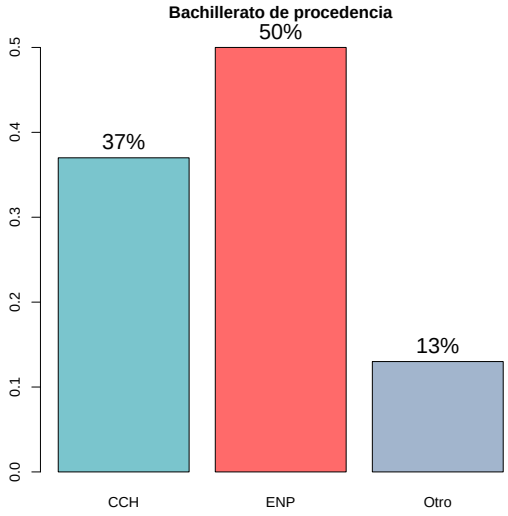


Figura 3.7: Distribución marginal: Aspectos psicosociales

Al aplicar la prueba Ji-cuadrada de independencia, no rechazamos la hipótesis nula y concluimos que el bachillerato de procedencia y el resultado arrojado por la prueba (severidad) son independientes. No se encontró asociación entre el éxito académico y la severidad.

3.5. Factores asociados a la elección de la carrera de medicina

El manual consta de diferentes partes, cada una compuesta por varias secciones con un conjunto de preguntas y afirmaciones ante las que se debe emitir una opinión o punto de vista.

Esta prueba cuantifica los siguientes factores:

1. razonamiento abstracto (TPRANT),
2. aptitud mecánica (TPAMNT),
3. ensamble de formas (TPEFNT),
4. ciencias físicas (TPIFIS),
5. mecánico (TPIMEC),
6. matemáticas (TPICAL),
7. ciencias biológicas y de la salúd (TPIBIO),
8. ecología y medio ambiente (TPIECO),
9. altruismo/servicio social (TPISER),
10. político (TPIPOL),
11. ciencias sociales (TPISOC),
12. administrativo-financiero (TPIADM),
13. organizacional/persuasivo (TPIPER),
14. artístico plástico visual (TPIART),
15. expresión musical (TPIMUS),
16. expresión oral (TPIORA),
17. expresión escrita (TPILIT),
18. estabilidad y normas (TPEYN),
19. inseguridad personal (TPINP),
20. incomodidad social (TPINS),
21. impulsividad (TPIMP),

22. deseabilidad social (TPDES),
23. mentiras (TPMEN),
24. desajuste (TPDPP),
25. control social (TPCSO),
26. autoestima (TMAES),
27. autoeficacia (TMAEF),
28. temor al fracaso (TMTEF),
29. evitación del trabajo retador (TMETR),
30. maestría (TMMAE),
31. orientación al logro (TMOAL),
32. morosidad (TMMOR),
33. liderazgo (TMLID),
34. negativismo (TLDYP),
35. manipulación social (TLMAQ),
36. inseguridad en la elección de carrera (TEIET),
37. información profesiográfica (TEIPT),
38. seguridad y satisfacción en la elección vocacional (TESST),
39. obstáculos familiares y económicos (TEOBT),
40. dificultad en la integración escolar (TEDET),
41. autoeficacia en medicina (TEAFT),
42. locus de control externo (TECET),
43. prestigio social (TEPST),
44. interés social (TEIST).

Cada uno de los factores antes mencionados puede clasificarse como: pruebas de aptitudes, intereses, autopercepción según normas referenciales, perfil de motivación y perfil vocacional. Los resultados obtenidos por los novecientos treinta y tres alumnos que presentaron la prueba se muestran a continuación.

Pruebas de Aptitudes

Conformadas por los factores: razonamiento abstracto, aptitud mecánica y ensamble de formas. Esta parte de la prueba consta de series de imágenes que deben ser completadas por el alumno (razonamiento abstracto), también contiene series de dibujos en los cuales se plantean situaciones y preguntas, que buscan evaluar la habilidad para identificar los principios físicos que explican el movimiento de los cuerpos, el funcionamiento de los aparatos y el comportamiento de algunos fenómenos (aptitud mecánica) y finalmente se evalúa la habilidad para armar figuras a partir de piezas que se encuentran impresas en la prueba (ensamble de formas).

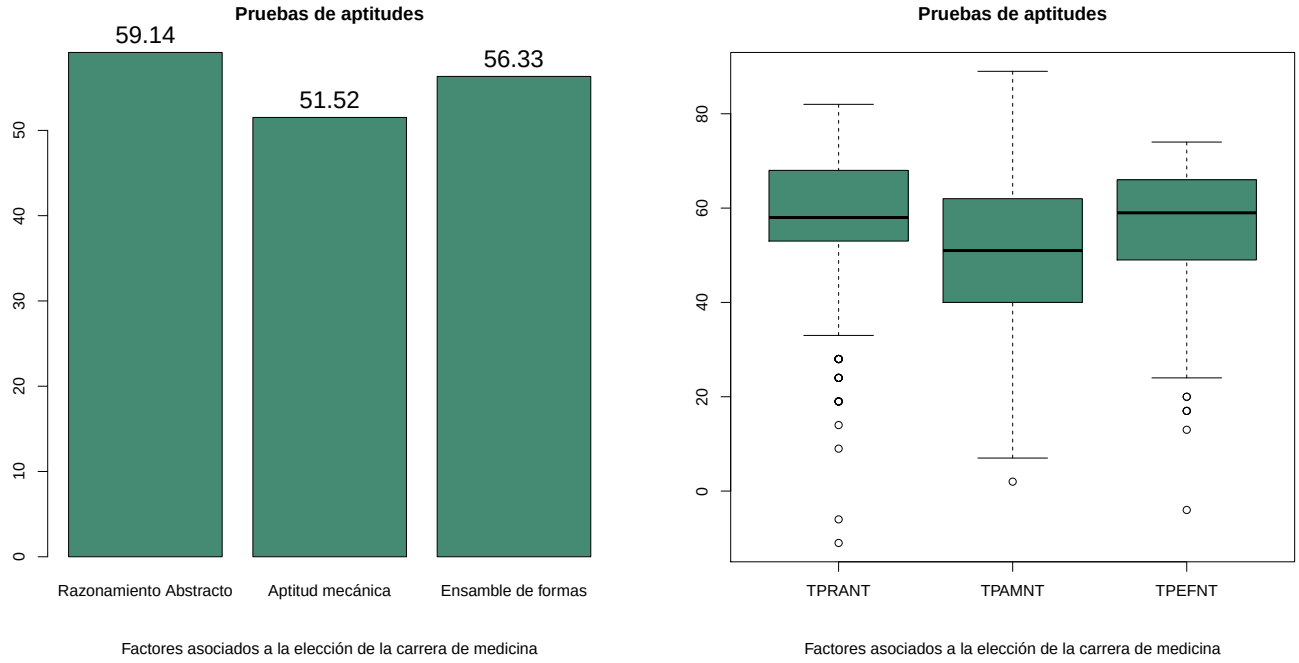


Figura 3.8: Pruebas de Aptitudes

En todos lo casos notamos la presencia de datos atípicos, que corresponden a alumnos cuyos resultados fueron más bajos de los esperados. El factor en el cual se obtuvo una media mayor y datos menos dispersos fue razonamiento abstracto. A diferencia del factor aptitud mecánica, cuyo promedio fue de 51.5 y cuyos datos se muestran más dispersos en relación al resto de los factores.

Se generó una nueva variable llamada *Aptitudes* que resulta de sumar los puntajes de los tres factores anteriores. Naturalmente, las correlaciones entre esta nueva variable y los tres factores son altas.

Intereses

Esta parte de la prueba busca conocer las actividades que los estudiantes prefieren realizar. Los intereses están cuantificados a través de los factores siguientes: ciencias físicas, mecánico, matemáticas, ciencias biológicas y de la salud, ecología y medio ambiente, altruismo/servicio Social, político, ciencias sociales, administrativo-financiero, organizacional/persuasivo, artístico plástico visual, expresión musical, expresión oral, expresión escrita.

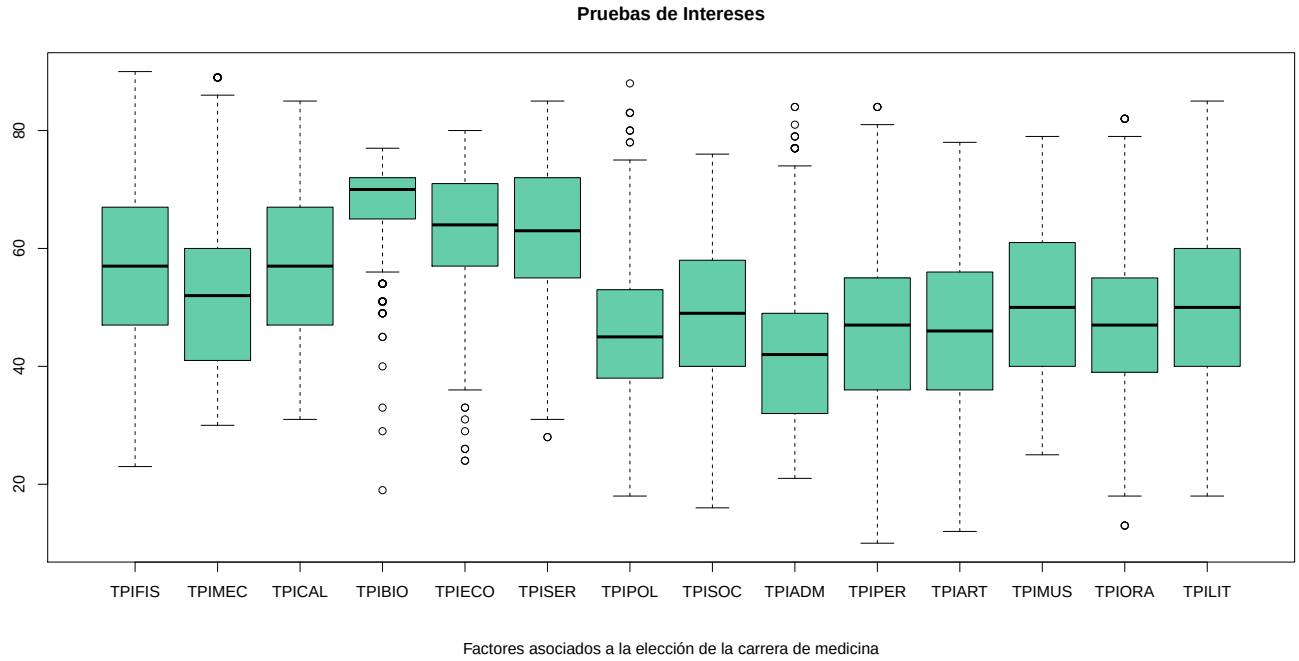


Figura 3.9: Pruebas de Intereses

Con ayuda de las gráficas de caja notamos que el factor de ciencias biológicas y de la salud resulta ser el factor con una mediana mayor al resto de los factores y la variable cuyos datos están menos dispersos; (hecho que tiene mucho sentido, dado que los individuos que presentan la prueba son alumnos de nuevo ingreso de la carrera de Medicina.) Sin embargo esta variable presenta observaciones atípicas, que corresponden a alumnos que obtuvieron bajos puntajes en este factor. Asimismo los factores de altruismo/servicio social y ecología y medio ambiente, son los que presentan una mediana superior a los demás factores.

Por otra parte, las actividades que parecen menos interesarles a los estudiantes de nuevo ingreso son aquellas relacionadas con la administración, finanzas y política.Con los factores ciencias físicas, mecánico y matemáticas se creó una variable en escala aditiva llamada *Ciencias*.

Autopercepción Según Normas Referenciales

Consta de una serie de afirmaciones que el alumno contesta dando su opinión. Este rubro compuesto por los factores: deseabilidad social, estabilidad y normas, mentiras, control social, desajuste, impulsividad, negativismo, manipulación social.

Notamos que todos los factores que lo integran tienen una media de entre 50 y 60 puntos, hecho que se aprecia también en las gráficas de caja, los cuartiles para cada una de las variables están cercanos entre ellos y las principales diferencias se encuentran en los datos atípicos. La única variable que no presenta datos atípicos es impulsividad. Intuitivamente, esto nos indica que existe un acuerdo generalizado acerca de las normas referenciales entre los estudiantes de primer ingreso. Con la ayuda de los factores deseabilidad social, mentiras y control social, se generó una nueva variable llamada *Normas*.

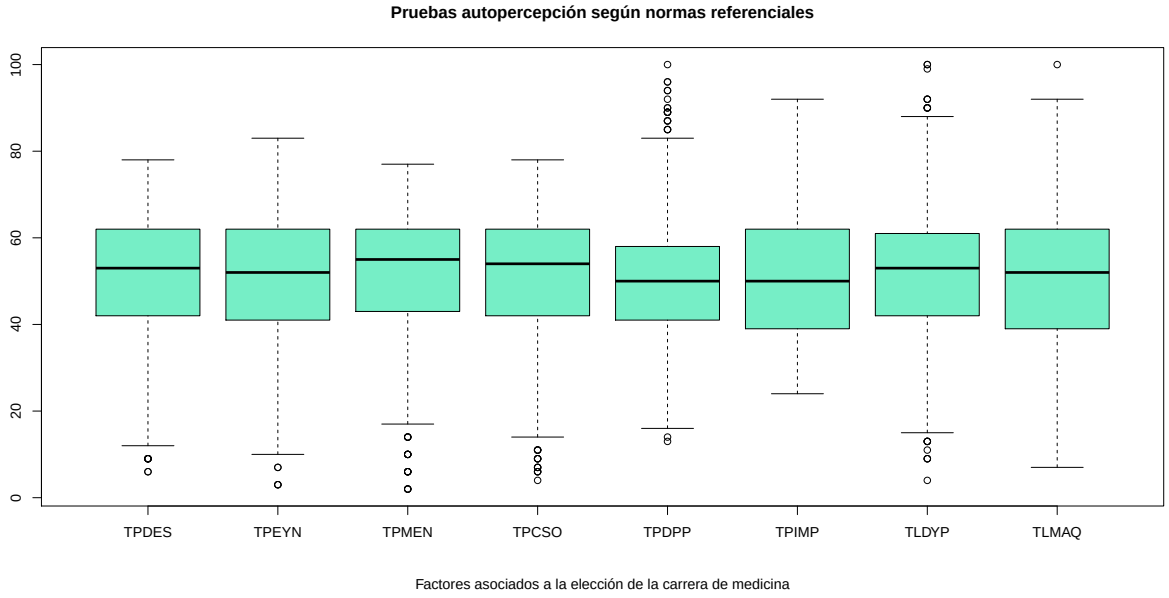


Figura 3.10: Pruebas de Autopercepción según normas referenciales

Perfil de motivación

Este rubro está integrado por los factores: orientación al logro, autoeficiencia en medicina, maestría, liderazgo, autoestima, temor al fracaso, evitación del trabajo retador, morosidad, incomodidad social, inseguridad personal y autoeficacia. Los resultados son los siguientes:

El factor con un mayor promedio corresponde a liderazgo, mientras que el factor con un menor promedio corresponde a temor al fracaso, sin embargo este último presenta varios datos atípicos. A partir de los factores orientación al logro y maestría, se generó una nueva variable llamada *Motivación*

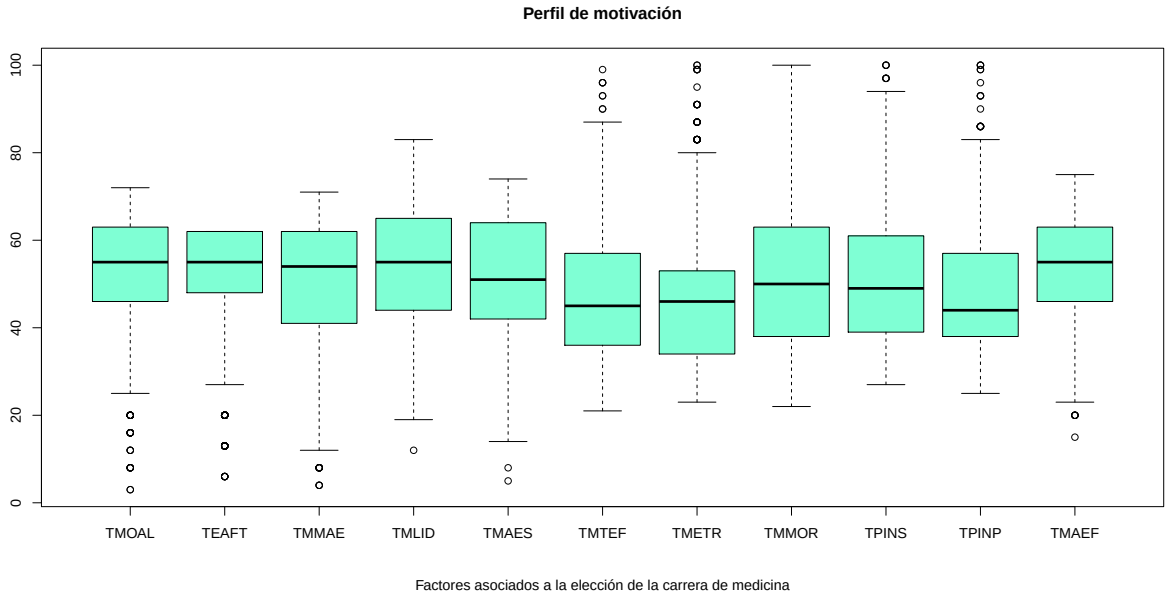


Figura 3.11: Perfil de motivación

Perfil de vocacional

Integrado por los factores: inseguridad en la elección de carrera, información profesiográfica, seguridad y satisfacción en la elección vocacional, obstáculos familiares y económicos, dificultad en la integración escolar, locus de control externo, **prestigio social**, interés social.

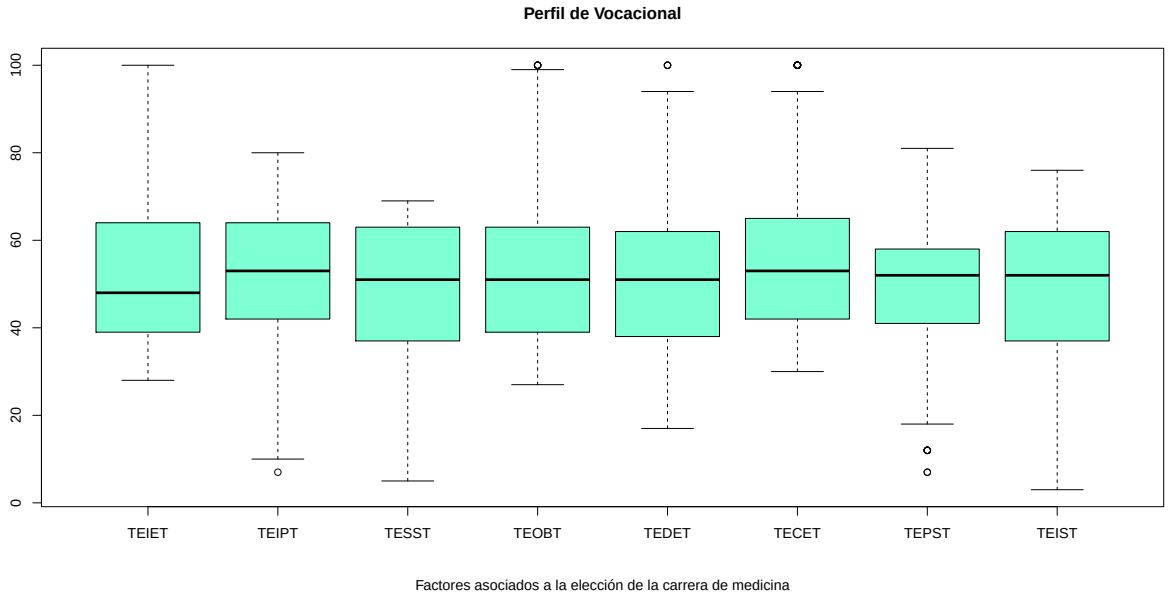


Figura 3.12: Perfil de motivación

Se encontró asociación entre las variables *éxito académico* y prestigio social.

3.6. Calificaciones obtenidas en cada una de las ocho asignaturas de primer año

Como se mencionó en el *capítulo 2*, las ocho materias que integran el primer año de la licenciatura de Médico-Cirujano de la facultad de Medicina de la UNAM son: Anatomía, Embriología Humana, Bioquímica y Biología Molecular, Biología Celular e Histología Médica, Integración Básico-Clínica I, Informática Biomédica, Introducción a la Salud Mental y Salud Pública y Comunidad.

Las asignaturas que presentan un mayor índice de reprobación y abandono son las que pertenecen al bloque de bases bomédicas, a saber: Anatomía, Bioquímica y Biología Molecular, Biología Celular e Histolgía Médica y Embriología Humana.

Los resultados presentados a continuación, consideran que un alumno tuvo éxito en cada asignatura si su promedio final es superior a 6, sin embargo, se debe considerar que los alumnos que obtuvieron una calificación inferior a 8 debieron presentar el examen final de la asignatura correspondiente.

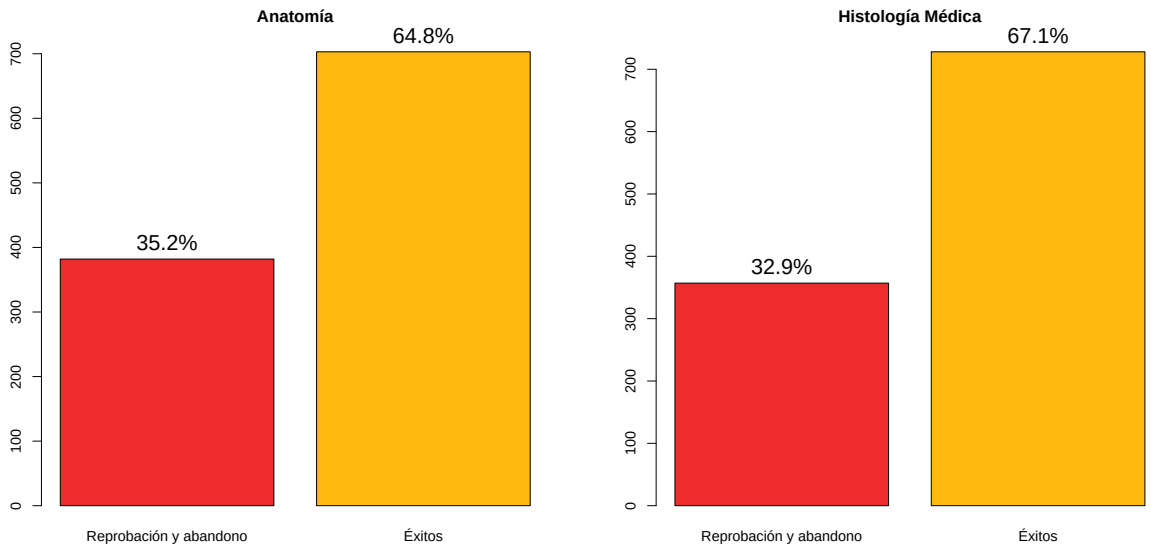


Figura 3.13: Porcentaje de reprobación y abandono: Anatomía y Biología Celular e Histología Médica

Notamos que al tratarse de estas cuatro asignaturas pertenecientes al bloque de bases biomédicas, casi una tercera parte del total de alumnos desertan o bien, obtienen una calificación inferior a seis, esto sin considerar a los alumnos que no excentan (obtienen una nota inferior a ocho) quienes deben obtener una calificación aprobatoria en el examen final, de lo contrario se considera que no aprobaron la asignatura correspondiente. La materia con mayor porcentaje de reprobación y abandono es Anatomía (35.2 %), seguida de Bioquímica y Biología Molecular (34.2 %).

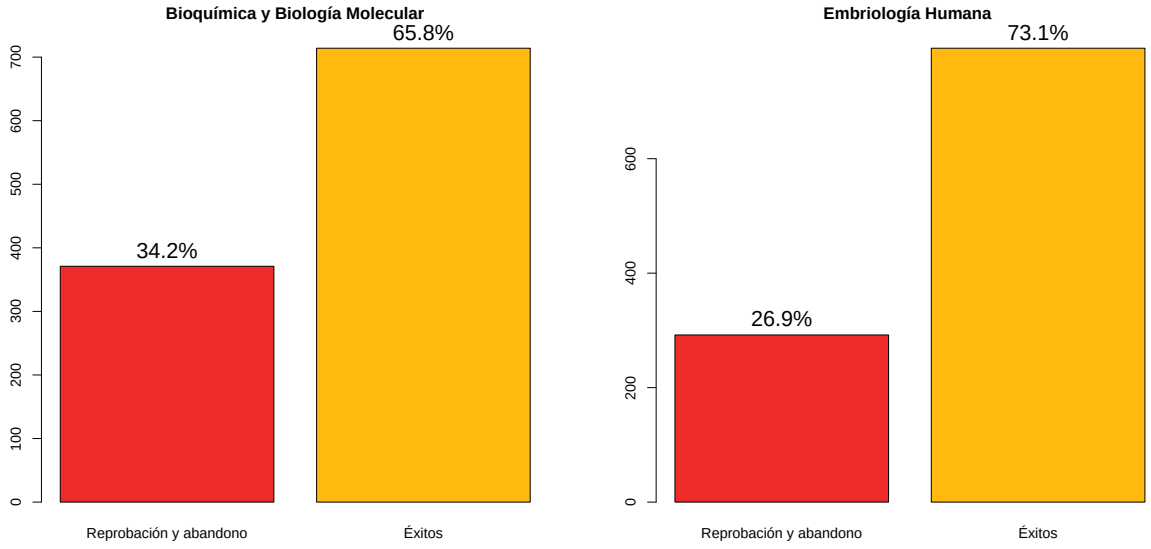


Figura 3.14: Porcentaje de reprobación y abandono Bioquímica y Embriología

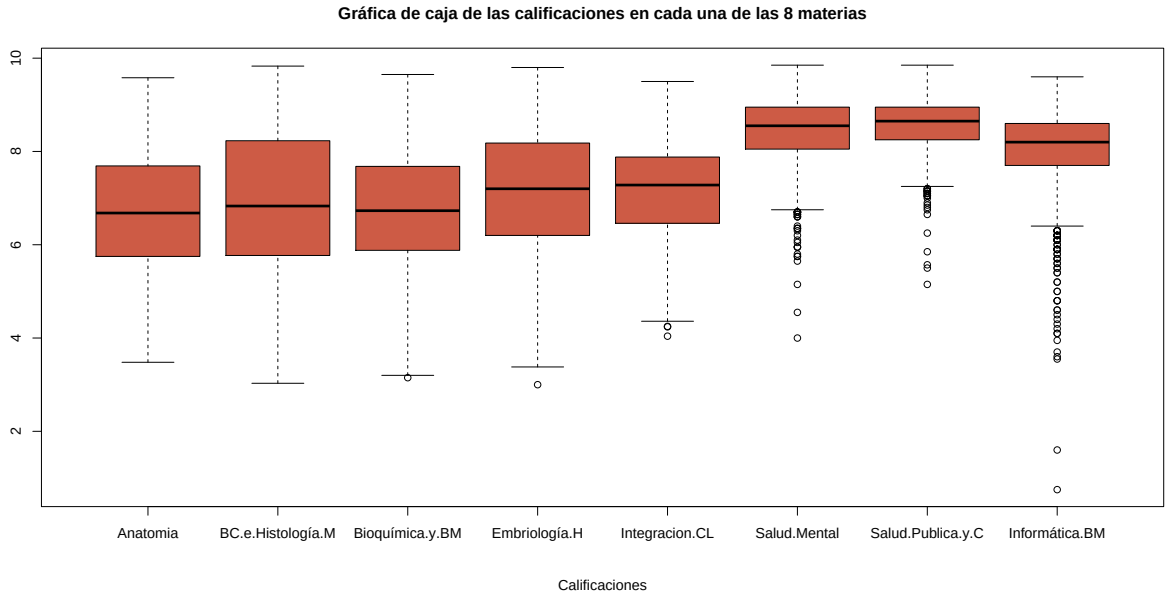


Figura 3.15: Gráfica de caja de cada una de las asignaturas de primer año

La asignatura en la cual los alumnos presentarón mejores resultados es Salud Pública y Comunidad, seguida de Salud Mental e Informática Biomédica, sin embargo, en la gráfica de caja observamos la existencia de datos atípicos, correspondientes a alumnos cuyas calificaciones fueron no satisfactorias.

Capítulo 4

Ajuste del modelo

En el análisis exploratorio encontramos importantes asociaciones entre el éxito académico y variables como el bachillerato de precedencia, el factor adquisición selectiva, la prueba de conocimientos generales, etc. y a su vez también se observaron múltiples asociaciones entre las variables pertenecientes al conjunto de regresoras, lo anterior sugiere que probablemente sería adecuado incluir interacciones en el modelo final. La variable respuesta está balanceada (47.2 % de observaciones codificadas como cero[1] contra 52.6 % de observaciones codificadas como uno[2]) por lo que la aplicación de la liga logit es adecuada en este caso.

Después de haber analizado un considerable número de modelos llegamos a dos ajustes que parecen ser adecuados y con la aprobación del experto de área procedemos a hacer la comparación entre estos, para decidir de forma definitiva cuál es el mejor. Las covariables que se consideran para la construcción de los modelos son:

1. bachillerato de procedencia,
2. resultado de la prueba de conocimientos,
3. adquisición selectiva,
4. adquisición generativa,
5. eficacia percibida (IEEA),
6. aptitudes,[3]
7. ciencias exactas, [4]
8. ciencias biológicas y de la salud,

[1] correspondientes a alumnos con bajo rendimiento académico.

[2] correspondiente a alumnos que obtuvieron notas aprovatorias en todas y cada una de sus asignaturas.

[3] covariable generada a partir de los factores razonamiento abstracto, aptitud mecánica y ensamble de formas.

[4] generada con los factores ciencias físicas, mecánico y matemático.

9. normas,[5]
10. motivación, [6]
11. prestigio social (factores asociados a la carrera de medicina).

El modelo uno está integrado por las covariables: bachillerato de procedencia (categórica con tres niveles), prueba de conocimientos generales, eficacia percibida (categórica con tres niveles), aptitudes, ciencias biológicas y de la salud, normas, motivación y prestigio social.

Mientras que el modelo dos está integrado únicamente por las covariables: bachillerato de procedencia, conocimientos generales, eficacia percibida, aptitudes, ciencias biológicas y de la salud y prestigio social, de modo que estos dos modelos son anidados, es decir el modelo uno contiene todas las covariables del modelo dos y más. El logit 1.12 para el modelo uno se expresa como:

$$logit(\pi(X)) = \beta_1 * Bachillerato + \beta_2 * Conocimientos +$$
$$\beta_3 * EficaciaPercibida + \beta_4 * Aptitudes + \beta_5 * CienciasBio +$$
$$\beta_6 * Normas + \beta_7 * Motivación + \beta_8 * PrestigioSocial$$

Para el caso de las variables categóricas, existe un parámetro para cada uno de los niveles de la variable. La salida en r se muestra a continuación:

```
> modelo1<-glm(factor(Respuesta)~Bachillerato+Conocimientos+EficPer+
+ Aptitud+CienciasBio+Normas+Motivacion+PrestSoc,data=datos,family=binomial)
> summary(modelo1)

Call:
glm(formula = factor(Respuesta) ~ Bachillerato + Conocimientos +
    EficPer + Aptitud + CienciasBio + Normas + Motivacion + PrestSoc,
    family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5090  -0.9143   0.4659   0.9059   2.1912

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.272413   0.980665  -5.376 7.60e-08 ***
BachilleratoOTROS       -0.165430   0.303902  -0.544 0.586199
BachilleratoPREPARATORIA  0.683756   0.177305   3.856 0.000115 ***
```

[5] generadas por los factores deseabilidad social, control social y mentiras.
[6] formada por los factores orientación al logro y maestría.

```
Conocimientos             0.059841   0.007640   7.833 4.77e-15 ***
EficPerbajo              -0.036914   0.478934  -0.077 0.938563
EficPermedio             -0.376855   0.164072  -2.297 0.021625 *
Aptitud                   0.002766   0.002663   1.039 0.298969
CienciasBio               0.028213   0.012107   2.330 0.019788 *
Normas                   -0.003451   0.001809  -1.907 0.056471 .
Motivacion                0.004873   0.003234   1.507 0.131820
PrestSoc                 -0.013079   0.005428  -2.410 0.015974 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1279.7  on 924  degrees of freedom
Residual deviance: 1059.2  on 914  degrees of freedom
AIC: 1081.2

Number of Fisher Scoring iterations: 4
```

Observamos que el par de variables con mayor significancia son el bachillerato de procedencia y la prueba de conocimientos generales. Además por las relaciones exploradas en la parte descriptiva de este estudio, sabemos que este par de variables están asociadas, por lo que probamos un modelo con las mismas covariables añadiendo una interacción entre estas dos variables.

```
> modelo1.1<-glm(factor(Respuesta)~Bachillerato*Conocimientos+EficPer+
+ Aptitud+CienciasBio+Normas+Motivacion+PrestSoc,data=datos,family=binomial)
> summary(modelo1.1)

Call:
glm(formula = factor(Respuesta) ~ Bachillerato * Conocimientos +
    EficPer + Aptitud + CienciasBio + Normas + Motivacion + PrestSoc,
    family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4701  -0.9157   0.4806   0.9065   2.2136

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -5.475315   1.180752  -4.637 3.53e-06 ***
BachilleratoOTROS            -0.486224   1.667406  -0.292   0.7706
BachilleratoPREPARATORIA      1.098385   0.950012   1.156   0.2476
Conocimientos                 0.063852   0.014597   4.374 1.22e-05 ***
```

```
EficPerbajo                    -0.047929   0.480369  -0.100   0.9205
EficPermedio                   -0.378440   0.164236  -2.304   0.0212 *
Aptitud                         0.002751   0.002664   1.033   0.3018
CienciasBio                     0.028098   0.012105   2.321   0.0203 *
Normas                         -0.003418   0.001812  -1.886   0.0593 .
Motivacion                      0.004820   0.003235   1.490   0.1363
PrestSoc                       -0.012923   0.005440  -2.376   0.0175 *
BachOTROS:Conocimientos         0.003003   0.024256   0.124   0.9015
BachPREPARATORIA:Conocimientos -0.007479   0.017245  -0.434   0.6645
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1279.7  on 924  degrees of freedom
Residual deviance: 1058.9  on 912  degrees of freedom
AIC: 1084.9

Number of Fisher Scoring iterations: 4
```

Se observa que los parámetros asociados a la interacción *Bachillerato*Conocimientos* no son significativos, adicionalmente el aic de modelo muestra un incremento al añadir dicha interacción. Por otra parte, el logit para el modelo dos está dado por:

$$logit(\pi(X)) = \beta_1 * Bachillerato + \beta_2 * Conocimientos + \beta_3 * EficaciaPercibida$$
$$+\beta_4 * Aptitudes + \beta_5 * CienciasBio + \beta_6 * PrestigioSocial$$

```
modelo2<-glm(factor(Respuesta)~Bachillerato+Conocimientos+EFPERC+
    Aptitudes+TPIBIO+TEPST,data=datos,family=binomial)
> summary(modelo2)

Call:
glm(formula = factor(Respuesta) ~ Bachillerato + Conocimientos +
    EFPERC + Aptitudes + TPIBIO + TEPST, family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3629  -0.9236   0.4730   0.9037   2.1272

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -5.832020   0.933845  -6.245 4.23e-10 ***
BachilleratoOTROS      -0.164016   0.300042  -0.547 0.584623
```

```
BachilleratoPREPARATORIA  0.671296   0.175738   3.820 0.000134 ***
Conocimientos             0.062048   0.007578   8.188 2.66e-16 ***
EFPERCbajo                0.025675   0.470837   0.055 0.956512
EFPERCmedio              -0.384554   0.155536  -2.472 0.013419 *
Aptitudes                 0.002774   0.002642   1.050 0.293813
TPIBIO                    0.032886   0.011285   2.914 0.003567 **
TEPST                    -0.011048   0.005298  -2.085 0.037052 *

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1279.7  on 924  degrees of freedom
Residual deviance: 1064.2  on 916  degrees of freedom
AIC: 1082.2

Number of Fisher Scoring iterations: 4
```

Nuevamente, para este segundo ajuste, observamos que el bachillerato de procedencia, la prueba de conocimientos generales y el nivel de interés de los alumnos en el área de ciencias biológicas y de la salud son las variables más significativas. El criterio de información de Akaike[7] y el criterio de información Bayesiano [8] son muy parecidos para ambos ajustes.

	Modelo 1	Modelo 2
AIC	1081.21	1134.34
BIC	1082.24	1125.71

Adicionalmente probamos otros dos modelos que incluyen interacciones, la salida en r se muestra a continuación:

```
> modelo2.1<-glm(factor(Respuesta)~Bachillerato*Conocimientos+EficPer+
Aptitudes+CienciasBio+PrestSoc,data=datosMod,family=binomial)
> modelo2.1$aic
[1] 1085.786

> modelo2.2<-glm(factor(Respuesta)~Bachillerato+Conocimientos+
EficPer+AdquSel*AdquGen+CienciasBio+PrestSoc,data=datosMod,family=binomial)
> modelo2.2$aic
[1] 1088.061
```

[7] es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. El AIC proporciona un medio para la selección del modelo.

[8] criterio para la selección de modelos entre un conjunto finito de modelos.

Aparentemente, en este caso el incluir interacciones no necesariamente mejora el ajuste del modelo, por lo que nos concentraremos en la comparación de los modelos uno y dos presentados al inicio. Dado que nuestros dos modelos ajustados son anidados, es posible realizar la comparación mediante la diferencia de devianzas 1.20

Las hipótesis nula y alternativa de esta prueba son:

$$H_0 : \textbf{Los dos modelos son igual de buenos.}$$
$$vs$$
$$H_a : \textbf{El modelo grande contiene información que el modelo pequeño no contiene.}$$

Obtuvimos un p-valor de 0.02484192, por lo que a un nivel del 95 % de significancia se rechaza la hipótesis nula y concluimos que el modelo con más parámetros, es decir el modelo uno, contiene información relevante que no esta contenida en el modelo dos. Consideramos que el mejor modelo para predicir el éxito académico es el modelo uno.

4.1. Interpretación para los parámetros del modelo

Para la variable *bachillerato de procedencia* se tomó como categoria basal CCH, la interpretación de β_{11}=-0.16 resulta de considerar $1/exp\{\beta_{11}\}$=1.18. El momio de presentar un rendimiento académico satisfactorio si se proviene de CCH es 1.18 veces mayor que el momio de presentar un rendimiento académico satisfactorio si se proviene de otro bachillerato, manteniendo constante el resto de las variables.

Para el parámetro asociado a la categoría ENP, β_{12}=0.68, tenemos que $exp\{\beta_{12}\}$=1.98, es decir, el momio de presentar un rendimiento académico satisfactorio si se proviene de ENP es 1.98 veces mayor que el momio de presentar un rendimiento académico satisfactorio si se proviene de CCH, manteniendo constante el resto de las variables.

β_2=0.059 el coeficiente asociado a la prueba de *conocimientos generales* indica que $exp\{10 * \beta_2\}$=1.82, es decir, el cambio esperado en el momio de respuesta por 10 unidades de incremento en la prueba de conocimientos generales es de 1.82, manteniendo constante el resto de las variables.

β_{31}=-0.037 el parámetro asociado a la variable *eficacia percibida* (categoria bajo) se interpreta considerando $1/exp\{\beta_{31}\}$=1.037, el momio de presentar un rendimiento académico satisfactorio si se presenta la categoría *alto* es 1.04 veces mayor que el momio de presentar un rendimiento académico satisfactorio si se presenta la categoría *bajo* de *eficacia percibida*, manteniendo constante el resto de las variables.

β_{32}=-0.37 es el parámetro asociado a la variable *eficacia percibida* (categoria medio), de modo que el momio de presentar un rendimiento académico satisfactorio dado que se presenta la categoría *alto* de la variable *eficacia percibida* es 1.45 veces mayor que el momio de presentar un rendimiento académico satisfactorio dado que se presenta la categoría *medio* de la variable *eficacia percibida*, manteniendo constante el resto de las variables.

En cuanto al parámetro asociado a la covariable *aptitudes*, β_4=0.0027 tenemos que $exp\{50 * \beta_4\}$=1.148 es el cambio esperado en el momio de respuesta por 50 unidades de cambio en la variable *aptitudes*, manteniendo constante el resto de las variables.

Para el parámetro asociado al factor *ciencias biológicas y de la salud* β_5=0.028 tenemos que $exp\{10 * \beta_5\}$=1.32 que es el cambio esperado en el momio de respuesta por 10 unidades de cambio en el factor *ciencias biológicas y de la salud*, manteniendo constantes el resto de las variables.

Para la covariable *normas*, cuyo parámetro asociado β_6=-0.0034, concluimos que el cambio esperado en el momio de respuesta por 50 unidades de cambio en esta covariable es de 1.18.

Para la covariable *motivación* se tiene que el cambio esperado en el momio de respuesta por 50 unidades de cambio en esta covariable es de 1.27 y finalmente el cambio esperado en el momio de respuesta por 10 unidades de cambio en la covariable *prestigio social* es de 1.14, manteniendo constante el resto de las variables.

Con el objetivo de visualizar el efecto de cada parámetro, a continucación se presentan las

probabilidades de éxito estimadas para distintos valores de las covariables continuas presentes en el modelo.

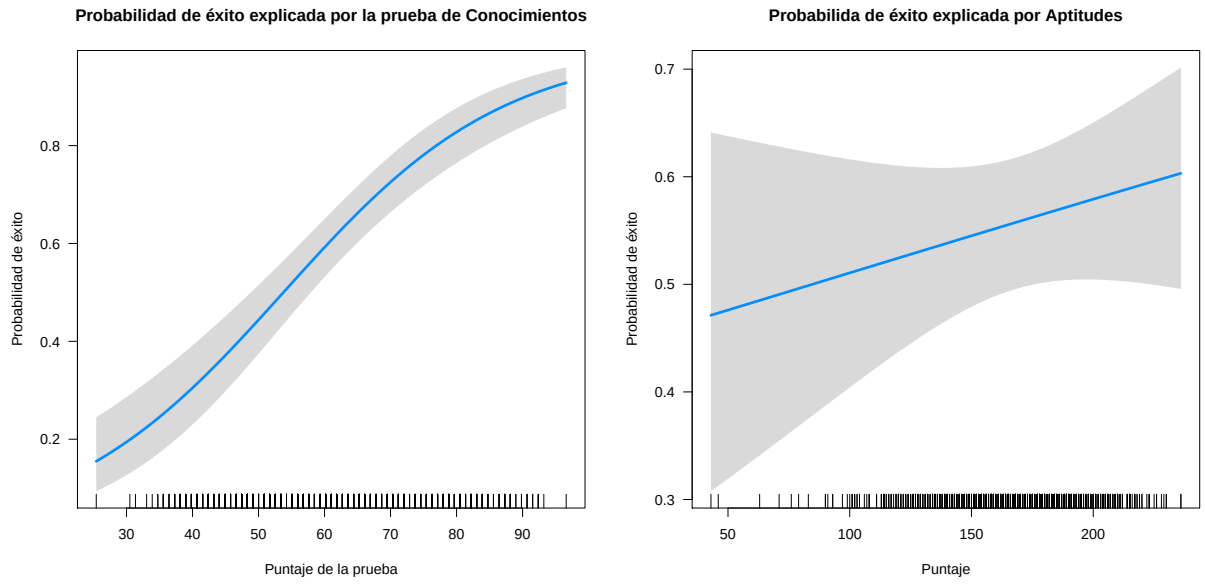


Figura 4.1: Probabilidad de éxito explicada por las variables Conocimientos y Aptitudes.

Como es de esperarse un incremento en el puntaje que cuantifica el nivel de conocimientos generales del alumno produce un incremento en la probabilidad de éxito académico. [9] La variable Aptitudes (integrada por los factores razonamiento abstracto, aptitud mecánica y ensamble de formas), también muestra que un incremento en el puntaje de esta variable genera una mayor probabilidad de éxito académico, manteniendo el resto de las variables contantes.

De manera similar a lo observado con el par de variables anteriores, existe una asociación positiva entre altos puntajes de las variables Ciencias Biológicas y de la Salud y Motivación (integrada por los factores orientación al logro y maestría) con probabilidades altas de éxito académico, es decir entre mayor sea el puntaje para estas variables, la probabilidad de éxito académico tenderá a aumentar.

Finalmente, para el factor Prestigio Social, se indica que mayores puntajes de este factor, se asocian con probabilidades de éxito menores. Esto probablemente indique que para algunos estudiantes de primer ingreso, la elección de carrera se basó en gran medida a la popularidad o prestigio que pudiera aportarles el hecho de ser estudiantes de medicina, más que otras motivaciones.

[9] recordemos que la escala de medición de ésta variable va del 0 al 100.

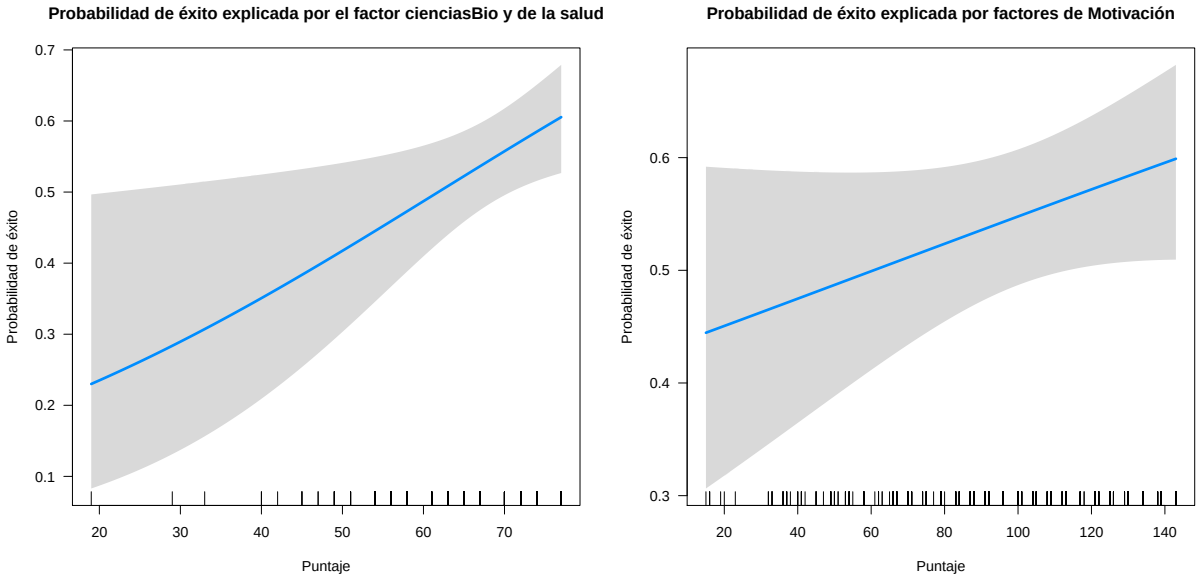


Figura 4.2: Probabilidad de éxito explicada por las variables Ciencias Biológicas y Motivación.

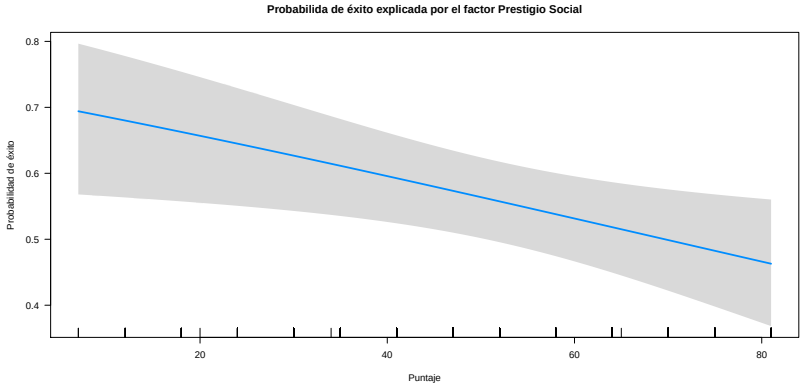


Figura 4.3: Probabilidad de éxito explicada por la variable Prestigio Social.

4.2. Pruebas de hipótesis sobre los parámetros

Al aplicar la prueba general del cociente de verosimilitudes 1.17 cuyas hipótesis nula y alternativa están dadas por:

$$H_0 : \beta = 0 \; vs. \; H_a : \beta \neq 0$$

rechazamos la hipótesis nula y concluimos que almenos uno de los coeficientes estimados es distinto de cero. También, aplicamos la prueba de Wald 1.18 para todos y cada uno de los parámetros antes mencionados y obtuvimos los resultados siguientes:

$$H_0 : \beta_i = 0 \; vs. \; H_a : \beta_i \neq 0, \text{ para } i = 1, 2, ..., p$$

Se rechazó la hipótesis nula para los parámetros asociados a las variables: bachillerato de procedencia (otros), conocimientos, eficacia percibida (medio), ciencias biológicas y de la salud y prestigio social, con lo que concluimos que los parámetros asociados a dichas covariables son estadísticamente distintos de cero.

4.3. Bondad de ajuste y poder predictivo del modelo

Debido a que nuestro ajuste contiene variables continuas, es más adecuado aplicar la prueba de Hosmer Lemeshow, 1.23 cuyas hipótesis a contrastar son:

$$H_0 : \textbf{El modelo ajusta bien a los datos .}$$
$$vs$$
$$H_a : \textbf{El modelo no ajusta bien.}$$

Obtuvimos un p-valor de 0.1754, por lo que no se rechaza la hipótesis nula y concluimos que el modelo ajusta adecuadamente.

En cuanto al poder predictivo, hemos codificado la variable respuesta como 1 si el alumno obtiene una calificación aprobatoria en todas y cada una de sus 8 asignaturas de primer año y como 0 si reprueba al menos una, de modo que para fines de esta investigación será más valioso tener una mayor *especificidad*1.24(ceros bien clasificados como ceros) que *sensibilidad*1.25(unos bien clasificados como unos), ya que nuestro objetivo es identificar a la población en riego de presentar un bajo rendimiento académico.

A continuación se presenta la tabla de clasificación 1.2 para este ajuste,

Cuadro 4.1: tabla de clasificación

	$\widehat{Y} = 1$	$\widehat{Y} = 0$	*Total*
$Y = 1$	355 (VP)	132 (FN)	487 (TP)
$Y = 0$	139 (FP)	299 (VN)	438 (TN)

Tomando como punto de corte π_0 =0.5, la *sensibilidad* y la *especificidad* diagnósticadas son:

$$Sensibilidad = \frac{355}{355 + 132} = 0.73$$

$$Especificidad = \frac{299}{299 + 139} = 0.68$$

Con lo que la tasa de buena clasificación1.26 esta dada por:

$$TBC = \frac{355 + 299}{355 + 132 + 139 + 299} = 0.71$$

concluimos que el poder predictivo del modelo es moderadamente bueno. Al realizar la curva de ROC encontramos que el área bajo la curva es de 0.7736. Para fines de esta investigación, consideramos que es adecuado.

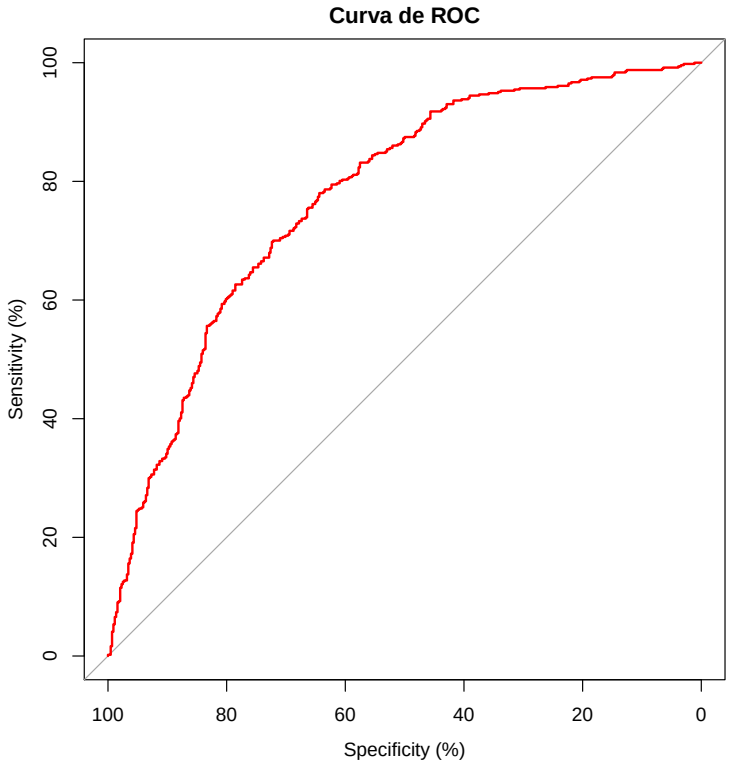


Figura 4.4: Curva de ROC.

4.4. Medidas de influencia y sobredispersión

Las medidas que usamos para evaluar las observaciones influyentes son las paláncas, la distancia de Cook y las Dfbetas, que pueden ser visualizadas mediante la gráfica 4.5

El ancho de los circulos representa la distancia de Cook, mientras que las observaciones pegadas al borde inferior derecho de la gráfica, (como lo es la observación 261) representan las palancas. Es necesario tener precausión en esta parte, ya que la prevalencia de todas esas observaciones puede traer consecuencias negativas en las estimaciones involucradas en el ajuste.

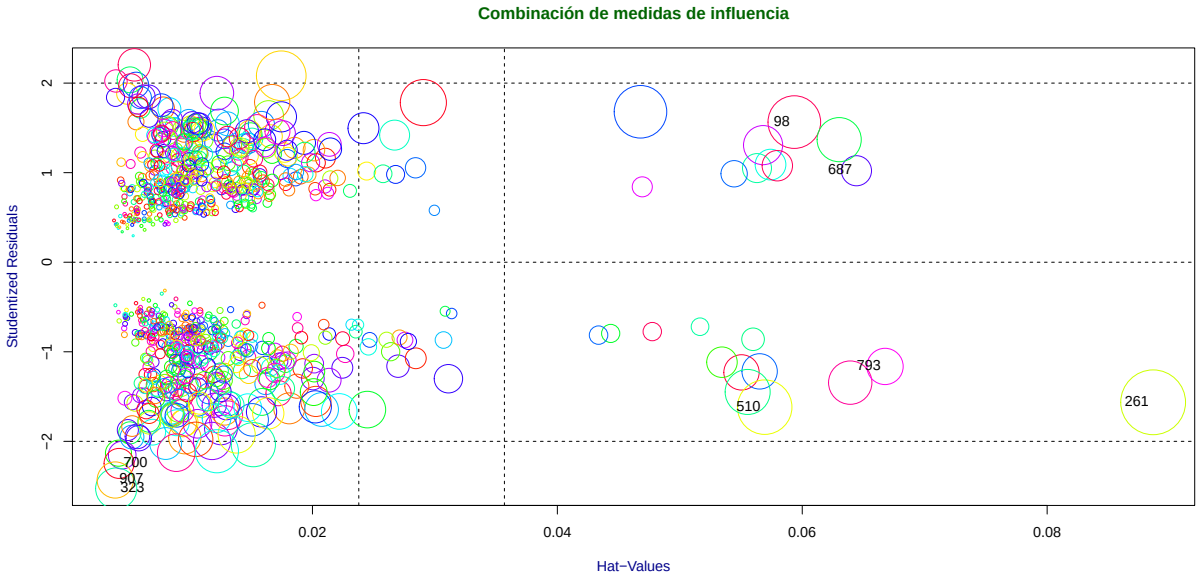


Figura 4.5: Observaciones influyentes.

En cuanto a la sobredispersión, como se discutió en el *capítulo 1*, si la devianza del modelo es mayor a su esperanza teórica, entonces tenemos evidencia de sobredispersión en el modelo. En este caso, la devianza del modelo ajustado resulta ser de 1059.21, mientras que la esperanza $n - p$ es igual a 914, por lo que concluimos que el modelo presenta problemas de sobredispersión, podemos plantear la hipótesis de que este problema se presenta como consecuencia de no incluir alguna o algunas covariables necesarias para explicar la variabilidad de la respuesta, sin embargo, pueden existir otras causas como la existencia de valores atípicos en la muestra, tal como se muestra en el figura4.5.

Una alternativa para dar solución a dicho problema es estimar un parámetro de dispersión 1.30 que en este caso es calculado como:

$$\phi = \frac{\chi^2}{n-p} = \frac{951.27}{925-11} = 1.041$$

y considerar la función de medias y la función de varianzas considerando este parámetro, para posteriormente estimar el vector de parámetros β a través de las ecuaciones de verosimilitud generadas con la ayuda de dichas funciones. Luego se estima el vector de parámetros por máxima verosimilitud.

Capítulo 5

Conclusiones y Comentarios Finales

La aplicación del modelo de regresión logística es una herramienta eficaz, al permitir llegar a conclusiones válidas para la toma de decisiones favorables, en este caso al desarrollo académico de alumnos de la licenciatura Médico-Cirujano de la Facultad de Medicina de la UNAM.

Se ha demostrado también que el conocimiento estadístico, al interactuar con otras áreas del conocimiento es funcional, al permitir enriquecer el conocimiento previo y dar certidumbre a los resultados. La aplicación de técnicas estadísticas tiene un alcance multidisciplinario y permite contar con un criterio confiable para caracterizar y predecir fenómenos sociales, económicos, físicos, educativos, etc.

La literatura sustenta evidencias teóricas que permiten definir al desempeño académico como un comportamiento multidimensional. Variables tales como el pensamiento abstracto, las estructuras de conocimiento y las estrategias de aprendizaje son clasificadas como variables cognoscitivas determinantes del aprendizaje en las aulas [10].

En el presente ejercicio, se logró construir un modelo que indica que el éxito académico está influenciado, entre varias otras cosas, por las estrategias de autorregulación y la auto-eficacia, factores que son de suma importancia al considerar que los estudiantes de recién ingreso al medio universitario se caracterizan por tener un perfil de estrategias de estudio deficiente, lo que dificulta su eficacia al estudiar y por ende que su desempeño académico sea óptimo.

Debido a lo anterior, se propone fomentar las estrategias de autorregulación en los estudiantes universitarios a través de supervisión de actividades escolares y aprendizaje estratégico, así mismo implantar estrategias que refuercen el autoconocimiento y autocontrol de los alumnos.

Otro aspecto considerable es que las creencias de auto-eficiencia determinan las percepciones de los alumnos sobre su capacidad para desempeñar las tareas requeridas en el curso. La idea que tenga el alumno sobre sus propias capacidades influye en las tareas que elige, las metas propuestas, la planificación, esfuerzo y persistencia de las acciones encaminadas a dichas metas. Por lo tanto también resulta valioso reforzar la seguridad de los alumnos en sí mismos

a la hora de analizar y aprender conceptos nuevos.

En cuanto a las componentes motivacionales, el aprendizaje efectivo en las aulas, se concibe como un proceso constructivo, acumulativo, auto-regulado, orientado a metas e individualmente diferente, [10] hecho por el cual los aspectos relacionados con las motivaciones e incentivos con los que cuentan los estudiantes resultan ser de vital importancia para mantener un rendimiento académico satisfactorio. Estudios anteriores han demostrado la existencia de relaciones significativas entre motivación, valoración de la tarea y creencias de autoeficacia con el uso de estrategias. Por tal motivo, se proponen atender las componentes motivacionales de los estudiantes, en este respecto, el papel del profesor es de suma importancia.

En la literatura se propone hacer del conocimiento de los profesores cuales son las variables motivacionales relacionadas con el aprendizaje. [1]Proponer tareas, intentando activar la curiosidad e interés del alumno y mostrando claramente su relevancia para el aprendizaje, promover la participación de los alumnos es clase con la finalidad de afianzar sus conocimientos y seguridad en si mismos.

Tal como se observó en el análisis exploratorio de los datos, existe una relación entre el bachillerato de procedencia y el éxito académico y hacemos énfasis en el hecho de que las probabilidades condicionales de reprobación dado el bachillerato de procedencia son más de dos veces mayores para alumnos provenientes de CCH, que para exalumnos de ENP y otros bachilleratos.

Otra evidencia a destacar es que la media del resultado global de la prueba de conocimientos generales es de 50.37 para exalumnos de CCH, de 61.5 para exalumnos de ENP y de 75.5 para alumnos que ingresaron mediante concurso de selección, lo que coloca a los exalumnos de CCH en una notoria desventaja frente al resto de los estudiantes. Ante esta circunstancia, se propone la creación y mantenimiento de programas de ayuda para los alumnos vulnerables, cuya población ha sido bien identificada con ayuda del modelo.

Finalmente, cabe mencionar que el bachillerato de procedencia y la prueba de conocimientos generales son variables que no es posible controlar,[2] de modo que se debe trabajar sobre aquellas variables que si podemos manejar, como es la motivación, la implementación de estrategias de aprendizaje, generar que las aptitudes mecánicas y matemáticas de los estudiantes se eleven, plantear las actividades académicas de manera que estás sean orientadas al logro y a incentivar el éxito académico de los alumnos.

[1]Frente a esta situación, sería valioso y de suma utilidad ofrecer al futuro profesor universitario una formación que incluya aspectos pedagógicos.

[2]por ejemplo, no sería factible dejar de aceptar a ciertos alumnos solo por que provienen de un determinado bachillerato

Apéndice A

Otros resultados teóricos de utilidad

A.1. Tablas de contingencia de dos vías

Las tablas de contingencia son útiles para analizar la asociación entre variables categóricas. Supongamos que se cuenta con dos variables X y Y, sea I el número de categorias de X y J el número de categorias de Y, una tabla con I renglones y J columnas que contiene los conteos cruzados de cada variable es llamada tabla de contingencia o **tabla de clasificación cruzada**. Las celdas de esta tabla representan las IxJ posibles combinaciones de respuesta.

Cuadro A.1: tabla de contingencia de tamaño 2X2

	Presencia $Y=1$	Ausencia $Y=0$	Total
Presencia $X=1$	n_{11}	n_{12}	n_{1+}
Ausencia $X=0$	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

A.1.1. Distribuciones asociadas a una tabla de contingencia

Distribución conjunta. Definimos

$$\pi_{ij} = \mathbb{P}(X = i, Y = j) \tag{A.1}$$

como la probabilidad de que un determinado individuo seleccionado aleatoriamente de la muestra pertenezca a la categoría i y X y a la categoría j de Y, donde la forma de estimar estas probabilidades está dada por:

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n} \tag{A.2}$$

Las probabilidades $\{\hat{\pi}_{ij}\}$ forman la distribución conjunta de X y de Y y por supuesto satisfacen

$$\sum_{ij} \hat{\pi} = 1 \tag{A.3}$$

Distribución marginal
Las distribuciones marginales son los totales por renglón y por columna de las probabilidades conjuntas. Las denotaremos como $\{\pi_{i+}\}$ para la variable renglón X y $\{\pi_{+j}\}$ para la variable columna Y, es decir $\hat{\pi}_{i+} = P(X = i)$ y $\hat{\pi}_{+j} = P(Y = j)$, de modo que para tablas de 2X2 tenemos:

$$\pi_{1+} = \pi_{11} + \pi_{12}$$
$$\pi_{+1} = \pi_{11} + \pi_{21}$$

Cuadro A.2: distribuciones conjunta y marginales

	Presencia $Y = 1$	Ausencia $Y = 0$	Total
Presencia $X = 1$	π_{11}	π_{12}	π_{1+}
Ausencia $X = 0$	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	1

Al tratarse de funciones de distribución, se cumple que:

$$\pi_{+1} + \pi_{+2} = 1$$

$$\pi_{1+} + \pi_{2+} = 1$$

Distribución condicional
Cuando se busca explicar una variable en función de otra es conveniente conocer la distribución condicional de la variable que consideramos de respuesta Y, ante la presencia de la otra variable X. Se definen dichas probabilidades como:

$$\pi_{i|j} = \mathbb{P}(Y = j|X = i), i, j = 1, 2 \tag{A.4}$$

Cuya estimación en términos de los conteos de la tabla esta dada por:

$$\hat{\pi}_{j|i} = \frac{n_{ij}}{n_{i+}} \tag{A.5}$$

El conjunto de probabilidades $\{\pi_{1|i}, \pi_{2|i}\}$ constituyen la distribución condicional de Y en cada categoria i de X.

A.1.2. Prueba Ji-cuadrada de independencia.

En esta prueba se contrastan las hipótesis:

$$H_0 : \textbf{Las variables X y Y son independientes.}$$
$$vs$$
$$H_a : \textbf{Las variables no son independientes}$$

El estadístico de prueba esta dado por:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{A.6}$$

donde $e_{ij} = \frac{n_{i+} n_{+j}}{n}$ representa el número de conteos esperados en la celda (i, j) bajo el supuesto de independecia. Si el valor de esta estadística es grande, entonces los valores observados en la tabla y los espereados bajo H_0 difieren mucho, de modo que la hipótesis nula es falsa.

Para garantizar que la prueba arroje resultados veridicos, es necesario considerar algunas restricciones sobre el tamaño de muestra y algunas reglas de uso, a saber:

1. Para muestras menores a 20 y cuya tabla de clasificación asociada es de 2X2, se recomienda la aplicación de la prueba exacta de Fisher, cuyas hipótesis nula y alternativa son las mismas que la prueba Ji-cuadrada.

2. Usar la prueba si al menos el 80 % de las celdas contienen conteos esperados $e_{ij} > 5$, pero ninguno de ellos es menor a uno. Cuando no se cumpla esta regla, es posible agrupar categorias de la tabla, siempre y cuando esto tenga sentido.

A.2. Análisis de componentes principales

El análisis de componentes principales es una técnica de Análisis Multivariado que tiene como objetivo reducir la dimensión de un conjunto datos, es decir, se busca recabar la información que aportan p variables en únicamente q componentes principales ($q < p$). Ésta técnica es considerada como un paso intermedio en el ajuste de modelos de regresión, ya que cuando se cuenta con un gran número de variables asociadas a un mismo individuo, resulta complicado ajustar de forma adecuada un modelo, debido a posibles problemas de multicolinealidad o bien, que se viole el *principio de parsimonia*.

Con éste análisis se pretende explicar la estructura de varianzas y covarianzas de un conjunto de p variables a través de un número reducido de combinaciones lineales de esas mismas variables, de forma que la información contenida en las p variables, pueda ser remplazada con las primeras q componentes principales.

Otra utilidad de esta técnica es que transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas, eliminando así posibles problemas de multicolinealidad.

Si las variables originales no están correlacionadas o están muy poco correlacionadas esta técnica no tiene ninguna utilidad y la dimensión real de los datos es la misma que el número de variables medidas.

Algebraicamente, dado un conjunto de p variables aleatorias $X_1, X_2, ..., X_p$ con matriz de covarianzas asociada Σ y con eigen-valores $\lambda_1, \lambda_2, ..., \lambda_p$ se consideran las combinaciones lineales de dichas variables:

$$\begin{array}{rl} Y_1 = & a_1'X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 = & a_2'X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ \vdots & \qquad\qquad \vdots \\ Y_p = & a_p'X = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{array} \tag{A.7}$$

Nuestro proposito es obtener Y_1 tal que la varianza sea la mayor posible, esto es, buscamos maximizar

$$\mathbb{V}(Y_1) = \mathbb{V}(a_1'X) = a_1'\Sigma a_1$$

sujeto a $a_1'a_1 = 1$(A.8)

Para la segunda componente, necesitamos maximizar

$$\mathbb{V}(Y_2) = \mathbb{V}(a_2'X) = a_2'\Sigma a_2$$

sujeto a

$$a_2'a_2 = 1$$

y a

$$\mathbb{C}ov(Y_1, Y_2) = 0 \tag{A.9}$$

por la condición de no correlación entre las componentes generadas. Procediendo de manera similar con las componentes subsecuentes. Para la i-ésima componente maximizamos

$$\mathbb{V}(Y_i) = \mathbb{V}(a_i' X) = a_i' \Sigma a_i$$

sujeto a

$$a_i' a_i = 1$$

y a

$$\mathbb{C}ov(Y_i, Y_k) = 0 \text{ para } k < i \tag{A.10}$$

Bajo un argumento de algebra lineal se puede mostrar que el valor de a_i que maximiza la varianza de la i-ésima componente Y_i, es el eigen-vector asociado al i-ésimo eigenvalor de la matriz Σ, (denotado como e_i)$\forall i < p$. De forma que la magnitud de e_{ik}, la k-ésima entrada del vector e_i mide la importancia de la k-ésima variable en la i-ésima componente principal.

De hecho se puede también mostrar que la correlación entre la i-ésima componente principal y la k-ésima variable esta dada por:

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \tag{A.11}$$

Siendo esta medida de utilidad para conocer cual es la influencia de la k-ésima variable en la i-ésima componente principal. De esta forma obtenemos un conjunto de p combinaciones lineales (componentes principales) cuyas características se enlistan a continuación:

1. Ambos conjuntos contienen la misma proporción de varianza

$$\sum_{i=1}^{p} \mathbb{V}(X_i) = \sum_{i=1}^{p} \mathbb{V}(Y_i) \tag{A.12}$$

2. Las componentes principales (c.p.) generadas por esta técnica no estan correlacionadas.

$$\mathbb{C}ov(Y_i, Y_j) = 0 \text{ para } i \neq j \tag{A.13}$$

3. La primera componente principal explica la mayor cantidad de varianza posible de los datos.

4. Cada componente subsecuente explica la mayor cantidad posible de varianza restante.

5. La proporción de varianza explicada por la k-ésima componente esta dada por:

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \tag{A.14}$$

Para que el análisis de componentes principales tenga sentido, debemos tener una fuerte correlación entre las variables involucradas.

Verificamos este supuesto con las variables incluidas en el **Inventario de Estrategias de Estudio y Autorregulación (IEEA)**

Las correlaciones resultaron no ser altas, de forma que la primera componente concentra únicamente el 60 % de la varianza.

Apéndice B

Códigos en r

```
#########################
##TABLAS DE CONTINGENCIA##
#########################

#####################################################
##éxito académicos vs bachillerato de procedencia##
#####################################################

###tablas de dos vías
t1<-table(factor(Respuesta),factor(Bachillerato))

##Incluyendo totales por renglón y por columna
t1.1<-mar_table(t1)

##Distribución Conjunta
conjunta<-prop.table(t1)
conjunta

##Marginales
mar.renglon<-apply(conjunta ,1,sum)
mar.col<-apply(conjunta ,2,sum)
mar.renglon<-round(mar.renglon,2)
mar.col<-round(mar.col,2)

ConMar<-mar_table(conjunta)
ConMar<-round(ConMar,3)

x11()
par(mfrow=c(1,2))
b<-barplot(mar.renglon,main="Éxito académico",horiz=FALSE,
```

```
col=c("brown4","darkgoldenrod1"),names.arg=c("No aprobado",
"Aprobado" ), sub="Distribución marginal")
text(x=b,y=c(mar.renglon[1],mar.renglon[2]),labels=c("47%",
"53%"),pos=3,col="black",cex=1.5,xlim=c(0,900),xpd=TRUE)


b<-barplot(mar.col,main="Bachillerato de procedencia",
horiz=FALSE,col=c("mediumvioletred","maroon4", "mediumorchid"),
names.arg=c("CCH","Otros", "ENP" ), sub="Distribución marginal")
text(x=b,y=c(mar.col[1],mar.col[2], mar.col[3]),labels=c("35\%",
"13%","52%"),pos=3,col="black",cex=1.5,xlim=c(0,900),xpd=TRUE)

##Condicionales
##condicionando a la variable X
##Rendimiento académico explicada mediante el bach de procedencia
condicional<-prop.table(t1,2)
cond<-round(condicional,2)
cond<-cond[1,]

x11()
par(mfrow=c(1,1))
b<-barplot(cond, main="Éxito académico dado el bachillerato de
procedencia",horiz=FALSE,col="aquamarine4",names.arg=c("CCH",
"Otros", "ENP" ), sub="Distribución condicional")
text(x=b,y=c(cond[1],cond[2], cond[3]),labels=c(cond[1],cond[2],
cond[3]),pos=3,col="black",cex=1.5,xlim=c(0,900),xpd=TRUE)

#Prueba Ji-cuadrada de independencia
chisq.test(t1)

#Medidas de asociación
assocstats(t1)

x11()
par(mfrow=c(1,2))
assoc(t1,shade=T, main="Respuesta y Bachillerato de procedencia")
mosaic(t1,shade=T, main="Éxito académico y Bachillerato de procedencia")

#####################################
##Prueba de conocimientos generales##
#####################################

#Resultados por género.
T1<-table(Sexo)
```

```
x11()
par(mfrow=c(1,2))
b<-barplot(T1,main="Género ",horiz=FALSE,col=c("cadetblue3","indianred1"),
names.arg=c("Hombres", "Mujeres" ), sub="Prueba de conocimientos generales")
text(x=b,y=c(T1[1],T1[2]),labels=c(T1[1],T1[2]),pos=3,col="black",
cex=1.5,xlim=c(0,900),xpd=TRUE)

#Boxplot
boxplot(Global~Sexo, main="Boxplot del resultado Global por género",
col=c("cadetblue3","indianred1"), sub="Prueba de conocimientos generales")

muj<-Global[Sexo=="M"]
hom<-Global[Sexo=="H"]

x11()
par(mfrow=c(1,2))
hist(muj, main="Resultado Global de las mujeres",
col="indianred1", ylab="frecuencia", xlab="Puntajes")

hist(hom, main="Resultado Global de los Hombres",
col="cadetblue3", ylab="frecuencia", xlab="Pntajes")

#Resultados por bachillerato de procedencia
BP<-table(BachilleratoCat)
BP<-(BP*100)/1151
bp<-round(BP,2)

x11()
par(mfrow=c(1,2))
b<-barplot(bp,main="Porcentajes. Bachillerato de procedencia",
horiz=FALSE,col=c("violet","violetred", "violetred4"),
names.arg=c("CCH", "OTROS", "PREPARATORIA" ),
sub="Prueba de conocimientos generales")
text(x=b,y=c(bp[1],bp[2], bp[3]),labels=c(bp[1],bp[2], bp[3]),
pos=3,col="black",cex=1.5,xlim=c(0,900),xpd=TRUE)


bchC<-GlobalCat[BachilleratoCat=="CCH"]
bchO<-GlobalCat[BachilleratoCat=="OTROS"]
bchP<-GlobalCat[BachilleratoCat=="PREPARATORIA"]

x11()
par(mfrow=c(1,3))
hist(bchC, main="CCH",breaks=10, col="violet",
ylab="frecuencia", xlab="Resultado Global")
```

```
hist(bchP, main="Prepa UNAM", breaks=10, col="violetred4",
ylab="frecuencia", xlab="Resultado Global")
hist(bchO, main="Otros", breaks=10, col="violetred",
ylab="frecuencia", xlab="Resultado Global")

summary(bchC)
summary(bchP)
summary(bchO)

x11()
boxplot(GlobalCat~BachilleratoCat, main="Boxplot
del resultado Global por bachillerato de procedencia",
col=c("violet","violetred", "violetred4"),
sub="Prueba de conocimientos generales")

########################
#######Factores#########
########################

#####Aptitudes######
Aptitudes<-cbind(TPRANT, TPAMNT, TPEFNT)
attach(Aptitudes)
summary(Aptitudes)
medias<-cbind(mean(TPRANT), mean(TPAMNT), mean(TPEFNT))
medias<-round(medias,2)

x11()
par(mfrow=c(1,2))
d<-barplot(medias,col=c("aquamarine4", "aquamarine4", "aquamarine4"),
main="Pruebas de aptitudes", names.arg=c("Razonamiento Abstracto",
 "Aptitud mecánica", "Ensamble de formas"), sub="Factores asociados a
la elección de la carrera de medicina")
text(x=d,y=c(medias[1],medias[2], medias[3]),labels=c(medias[1],
medias[2], medias[3]),pos=3,col="black",cex=1.5,xlim=c(0,900),xpd=TRUE)

boxplot(Aptitudes, main="Pruebas de aptitudes",sub="Factores asociados
 a la elección de la carrera de medicina", col="aquamarine4")
cor(Aptitudes)

###################
#####Intereses#####
###################

Intereses<-cbind(TPIFIS, TPIMEC, TPICAL, TPIBIO, TPIECO, TPISER,
TPIPOL, TPISOC, TPIADM, TPIPER, TPIART, TPIMUS, TPIORA, TPILIT)
```

```
x11()
boxplot(Intereses, main="Pruebas de Intereses",sub="Factores
asociados a la elección de la carrera de medicina", col="aquamarine3")
cor(Intereses)

#####################################################
#####Autopercepción según normas referenciales#####
#####################################################

Autop<-cbind(TPDES, TPEYN, TPMEN, TPCSO,TPDPP,TPIMP,TLDYP, TLMAQ)
attach(Autop)
summary(Autop)

x11()
boxplot(Autop, main="Pruebas autopercepción según
normas referenciales",sub="Factores asociados a la
elección de la carrera de medicina", col="aquamarine2")

###############################
#####Pérfil de motivación#####
###############################

Moti<-cbind(TMOAL, TEAFT,TMMAE, TMLID, TMAES, TMTEF,TMETR,
TMMOR, TPINS, TPINP,TMAEF)

attach(Moti)
summary(Moti)

x11()
boxplot(Moti, main="Perfil de motivación",sub="Factores
asociados a la elección de la carrera de medicina", col="aquamarine")

###############################
#####Aspectos vocacionales#####
###############################

Voca<-cbind(TEIET,TEIPT, TESST, TEOBT, TEDET, TECET, TEPST, TEIST)

attach(Voca)
summary(Voca)

x11()
boxplot(Voca, main="Perfil de Vocacional",sub="Factores asociados
a la elección de la carrera de medicina", col="aquamarine")
```

```
#Ajuste del modelo
modelo1<-glm(factor(Respuesta)~Bachillerato+Conocimientos+EFPERC+
    Aptitudes+TPIBIO+Normas+Moti+TEPST,data=datos,family=binomial)
summary(modelo1)


modelo2<-glm(factor(Respuesta)~Bachillerato+Conocimientos+EFPERC+
    Aptitudes+TPIBIO+TEPST,data=datos,family=binomial)
summary(modelo2)
summarise(glmlist(modelo1,modelo2))

AIC(modelo1)
AIC(modelo2)

BIC(modelo1)
BIC(modelo2)

#Mediante la diferencia de devianzas
anova(modelo2,modelo1,test="Chisq")
1-pchisq(5.0349,1)
#Se rechaza la hipótesis nula de que ambos modelos son igual de buenos


##Iinterpretación de los parámetros del modelo##
summary(modelo1)
Coeficientes<-coef(modelo1)

#Intervalos de confianza
cbind(coef=coef(modelo1),confint(modelo1))
exp(cbind(coef=coef(modelo1),confint(modelo1)))

#Pruebas de significancia para los parámetros
length(coef(modelo1))
wald.test(b=coef(modelo1),Sigma=vcov(modelo1), Terms=1:11)

#Probabilidades
x11()
visreg(modelo1,"Conocimientos",xlab="Puntaje de la prueba",
scale="response",ylab="Probabilidad de éxito", main="Probabilida
de éxito explicada por la prueba de Conocimientos")

##Prueba de Hosmer-Lemeshow
library(ResourceSelection)
hoslem.test(modelo1\$y,fitted(modelo1),g=10)
```

```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  modelo1$y, fitted(modelo1)
X-squared = 11.4909, df = 8, p-value = 0.1754

#La curva de ROC
install.packages("pROC")
library(pROC)

x11()
roc1<-roc(modelo1$y,modelo1$fitted.values, percent=TRUE,
smoooth=T, plot=T, main="Curva de ROC", col="red")

> roc1
Call:
roc.default(response = modelo1$y, predictor = modelo1$fitted.values,     percent = TRUE,

Data: modelo1$fitted.values in 438 controls (modelo1$y 0) < 487 cases (modelo1$y 1).
Area under the curve: 77.36%

##Diagnóstico de las observaciones
influencia<-influence.measures(modelo1)
influ<-influencia[[1]]
x11()
influencePlot(modelo1,id.n=3,col=rainbow(50),
main="Combinación de medidas de influencia",
col.main="darkgreen",col.lab="darkblue")
```

Bibliografía

[1] David W. Hosmer Stanley Lemeshow; Applied Logistic Regresion ; Wiley; 2000.

[2] Alan Agresti; Categorical Data Analysis; Wiley; 2002.

[3] Alan Agresti; An Introduction to Categorical Data Analysis; Segunda Edición; 2007

[4] Jiménez, M. ; Competencia social: intervención preventiva en la escuela; Universidad de Alicante.; 2000.

[5] Agresti; A survey of exact inference for contingency tables.; Statist. Sci. ;1992.

[6] Daryl Pregibon; Logistic Regression Diagnostics, Annals of Statistics ; 1981.

[7] Bradley Efron; The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise; 1975.

[8] Jonhson y Wichern; Applied Multivariate Statistical Analysis; 1992.

[9] Anderson, T.W.; An Introduction to Multivariate Statistical Analisys; New York:Jonh Wiley; 2003.

[10] Castañeda, Sandra; Inventario de Estrategias de Aprendizaje y Autorregulación; UNAM; 2003.

[11] Belsley and Welsch; Regression Diagnostics: Identifying Influencial Data and Sources of Collineality; Wiley.

[12] María Cristina Rinaudo, Analía Chiecher y Danilo Donolo; Motivación y uso de estrategias en estudiantes universitarios. Su evaluación a partir del Motivated Strategies Learning Questionnaire

[13] Choosing Between Logistic Regression and Discriminant Analysis; S. James Press; Sandra Wilson. Journal of the American Statistial Association.