Rev. Fac. Agron. (LUZ). 2006, 23: 196-212

Modelos de regresión lineal múltiple en presencia de variables cuantitativas y cualitativas para predecir el rendimiento estudiantil

M. Rosas, F. Chacín, J. García, M. Ascanio, M. Cobo

Facultad de Agronomía, Universidad Central de Venezuela. Apartado Postal 4579, Maracay 2101, Estado Aragua.

Resumen

El objetivo de esta investigación fue la construcción de modelos de regresión múltiple en presencia de variables cualitativas y cuantitativas, que permitan predecir el rendimiento estudiantil y sugerir al estudiante una alternativa para lograr el éxito en sus estudios en el Instituto Universitario de Tecnología del Yaracuy. Las variables explicativas eran 28, como variable dependiente se usó el índice de rendimiento académico al egresar del Instituto. Se incluyeron variables cualitativas que plantearon la necesidad del uso de varias variables dummy y se hizo estudio del comportamiento de los modelos bajo tales condiciones. Para cada una de las cuatro especialidades, se obtuvo tanto el modelo completo como los modelos reducidos por los métodos de todas las regresiones posibles y paso a paso. Se realizaron pruebas t en el modelo completo y se compararon las variables seleccionadas con las incluidas en los modelos seleccionados por ambos métodos. Se hizo estudio de los coeficientes de regresión de las variables seleccionadas para detectar su estabilidad. Un modelo fue seleccionado para cada una de las cuatro especialidades estudiadas: Agrícola, Conservación de Recursos Naturales Renovables (C.R.N.R.), Alimentos y Pecuaria. Estos modelos explicaron respectivamente el 56,41%; 89,66%, 69.33% y el 73,10% de la variabilidad total del rendimiento y las variables escogidas difierieron de acuerdo a la especialidad.

Palabras clave: Variables dummy, método de todas las regresiones posibles, método Paso a paso, rendimiento académico, variables cualitativas y cuantitativas, comparación de modelos, R^2, Press, Cp de Mallows.

Introducción

En los actuales momentos, Venezuela vive una etapa de expectativas realistas y estimulantes, de retos excitantes para reconstruir el país sobre bases más éticas y firmes a las vividas hasta hace poco. Esta nueva eta-

Recibido el 26-10-2004 ● Aceptado el 14-5-2005
Autor para correspondencia email: garciaj@agr.ucv.ve

pa que vive Venezuela es producto de la actual crisis económica severa, que por lo visto durará todavía algunos años más y que afortunadamente parece que está creando conciencia a todos los niveles y especialmente en el sector educativo. La situación de crisis económica mencionada anteriormente obliga a emprender una búsqueda de soluciones factibles que impidan un deterioro profundo de la Educación Superior. Dentro de esa perspectiva, es dable pensar se redoblaran los esfuerzos para mejorar la Educación Superior, superando los falsos conceptos obsoletos y sus problemas, y preparar la Educación Superior para que cumpla con la comprometedora obligación de ser un instrumento idóneo para hacer frente a la crisis y garantizar un porvenir seguro. El educador, conocedor de que una de las causas primordiales que hacen que la Educación Superior no sea del todo satisfactoria es el bajo rendimiento estudiantil, debe implementar reformas tendientes a mejorarlo; es por esto que a través del presente estudio se pretendió ofrecer un aporte, que sirva de base para mejorar el rendimiento estudiantil en el Instituto Universitario de Tecnología del Yaracuy siendo el objetivo primordial de esta investigación la construcción de modelos de regresión múltiple en presencia de variables cualitativas y cuantitativas, que permitan predecir el rendimiento estudiantil y a la vez sugerir al estudiante una alternativa para lograr el éxito en sus estudios, lo cual se traduciría en mejoras a la Educación Superior. En este estudio se analizó el comportamiento de los modelos con variables explicativas mixtas: cualitativas y cuantitativas, desde el punto de vista de la interpretación de su significación en los modelos. En virtud de la importancia en esta investigación de términos tales como transformaciones, variables falsas y selección de modelos, se desarrollaron brevemente estos aspectos:

Transformaciones: Las transformaciones han sido usadas para encontrar datos que satisfagan los supuestos de un modelo paramétrico conveniente. Barlett (1) señala que el propósito ordinario de la transformación para cualquier tipo de análisis es el de cambiar la escala de mediciones con el objeto de hacerles válidos. El problema consiste en encontrar la transformación adecuada que garantice: 1) la independencia de la media y la varianza, es decir, que la varianza de los datos transformados no se vea afectada por cambios en la media. 2) Que la distribución de la variable transformada sea aproximadamente normal. 3) Que la escala transformada sea una en la cual la media aritmética sea una estimación eficiente del verdadero valor para cualquier grupo de mediciones. 4) Que la escala transformada sea de tal manera que los efectos del modelo sean lineales y aditivos. Las transformaciones más usadas son: el recíproco, logaritmo, raíz cuadrada, arcoseno, entre otras (3).

Variables Falsas: El uso de variables falsas es un método para cuantificar características de tipo cualitativo (que no son susceptibles de ser cuantificadas) o que presentan la conveniencia de separar categorías dis-

cretas. En el análisis de regresión se utilizan variables falsas cuando se cumplen las siguientes condiciones: 1) las observaciones originales pueden ser agrupadas en clases o grupos de tipo cualitativo. 2) el efecto de esta agrupación es alterar la ordenada al origen sin alterar la pendiente (4 y 2).

Ruiz-Maya *et al.* (7) establecen que si los datos originales pueden separarse en dos o más grupos significativos, habría que estudiar los efectos de los diferentes grupos. Por ejemplo si una variable respuesta se hace depender de dos variables explicativas X_1 y X_2 y suponemos que la función que relaciona estas variables explicativas con la variable respuesta es lineal. La ecuación del modelo es $y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}$.Puede presentarse que en el conjunto de variables explicativas contemplamos tres grupos: cuantitativas, cualitativas y mixtas. Si las variables explicativas son cuantitativas (continuas o discretas) y se supone que se han efectuado cuatro observaciones; el sistema de ecuaciones que da lugar puede ser planteado en forma matricial

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \\ 1 & x_{14} & x_{24} \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

La matriz formada por los valores observados de las variables explicativas recibe el nombre de matriz de diseño y será designada por X. Si las variables explicativas X_1 y X_2, son cualitativas y las suponemos dicotómicas se le pueden asignar los valores 1 y 0; matricialmente la representación del sistema de ecuaciones sería:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

No resulta imprescindible codificar las variables mediante 0 y 1, sino que pueden asignarse códigos según las necesidades, pero es necesario tener presente que el tipo de codificación elegido influye en las estimaciones del modelo; sin embargo, se obtienen las mismas estimaciones de la variable respuesta así como los mismos valores de los estadísticos de bondad de ajuste y los mismos resultados de los contrastes de hipótesis (7).

La introducción de una variable explicativa categórica cuando el número de niveles de la variable es superior a dos, lleva consigo cambios en la formulación del modelo debido a la codificación. La exposición del procedimiento sigue a Hosmer-Lemeshow (5) y Wrigley (8). Si se tiene el caso de una sola variable explicativa X, presentando cinco niveles, a partir de ella se definen cuatro variables "ficticias"

Codificación de la variable	Variables ficticias X_{11}	X_{12}	X_{13}	X_{14}
Nivel 1	1	0	0	0
Nivel 2	0	1	0	0
Nivel 3	0	0	1	0
Nivel 4	0	0	0	1
Nivel 5 o Nivel de referencia	0	0	0	0

:X_{11} , X_{12} X_{13} y X_{14}, una menos que el número de niveles, atribuyéndoles a cada una valores 1 ó 0, según se halle presente o no el correspondiente nivel. En este caso sería:

Una vez introducidas el modelo queda: $y=\beta_0+ \beta_{11}x_{11}+ \beta_{12}x_{12}+\beta_{13}x_{13}+ \beta_{14}x_{14}$. Los coeficientes b_{1i} cuantifican el efecto producido por la presencia del correspondiente nivel de la variable explicativa X. Matricialmente el sistema se plantea:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{14} \end{bmatrix}$$

Si solo el coeficiente β_{11} es significativo el modelo quedaría $y=\beta_0+\beta_{11}x_{11}$, de la misma manera si la significación es de β_{12} el modelo sería $y=\beta_0+\beta_{12}x_{12}$ y así sucesivamente, por supuesto si solo β_{15} es significativo el modelo quedaría $y=\beta_0$, pudiendo presentarse otros modelos si varios de ellos son simultáneamente significativos.

La razón de definir una variable ficticia menos que el número de niveles de X es que de no hacerlo así, la matriz de diseño no conduciría a una inversa. La complicación del cálculo es manifiesta, debido al elevado número de variables ficticias que es preciso introducir cuando el número de variables explicativas cualitativas es alto e igualmente cuando se incrementa el número de niveles que cada una posee.

Cuando las variables explicativas son mixtas, como por ejemplo si la variable X_1 es cuantitativa y X_2 cualitativa multinomial y presenta cinco niveles, el modelo es:

$y_i=\beta_0+\beta_1x_{1i}+\beta_{21}x_{21}+\beta_{22}x_{22}+\beta_{23}x_{23}+\beta_{24}x_{24}$

La matriz de diseño X del modelo, en este caso es

X_1 X_{21} X_{22} X_{23} X_{24}

$$\begin{pmatrix} 1 & x_{11} & 1 & 0 & 0 & 0 \\ 1 & x_{12} & 0 & 1 & 0 & 0 \\ 1 & x_{13} & 0 & 0 & 1 & 0 \\ 1 & x_{14} & 0 & 0 & 0 & 1 \\ 1 & x_{15} & 0 & 0 & 0 & 0 \end{pmatrix}$$

En el modelo lineal la pendiente de la recta viene dada por el parámetro b_1, y los coeficientes que afectan a los diferentes niveles de la variable cualitativa suponen un desplazamiento paralelo de la función en cada uno de los cuatro casos. La estimación de y_i se obtiene fácilmente sin más que sustituir en el modelo el correspondiente valor x_{1i} (de la variable explicativa cualitativa) y para cada nivel de la variable ficticia la codificación establecida. Por ejemplo la estimación de y_i para el nivel 3 de la variable ficticia es igual a

$$y_i=\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_{21i}[x_{21i}=0]+\hat{\beta}_{22}[x_{22i}=0]+\hat{\beta}_{23}[x_{23i}=1]+\hat{\beta}_{24}[x_{24i}=0]$$

$$y_i=\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_{23i}$$

Selección de Modelos:

Pruebas de hipótesis usando t:

Martínez (6) establece que una vez verificada la significación de la prueba F en el análisis de regresión múltiple, cuando ocurre el caso de una

hipótesis sencilla, digamos Ho: $\beta_i = 0$ versus $\beta_i \neq 0$ esta puede realizarse alternativamente usando una prueba de t de dos colas que se denominan pruebas parciales. Así si $t_{1-\alpha/2}$ es la t tabulada al nivel α de significancia con n grados de libertad, los del error; Ho se rechaza si el valor absoluto de la t calculada es igual o mayor que $t_{1-\alpha/2}$. La equivalencia de esta prueba con la correspondiente prueba de F se deriva del hecho que el cuadrado de una variable con una distribución t de Student con n grados de libertad, se distribuye como una F con 1 y n grados de libertad.

Criterios de comparación de modelos:

Los criterios de comparación más usados son:

a) el cuadrado medio del error (CME)
b) el coeficiente de determinación (R^2)
c) la suma de cuadrados de predicción (PRESS)
d) el estadístico Cp de Mallows

Los criterios CME y R^2 están relacionados por la ecuación $R^2 = 1 - \frac{(n-p-1)CME}{SCTotal}$

En este trabajo se consideró solo el coeficiente de determinación. El uso del R^2 se ha popularizado porque toma valores de 0 a 1, esto permite apreciar la calidad del ajuste sin utilizar tablas. El criterio PRESS se usó con el objetivo de predecir el valor y_i de la observación i que no se incluyó en el ajuste y se tiene un error estimado $y_i - \hat{y}_i$. Al repetir este procedimiento omitiendo cada vez una de las n observaciones, se puede calcular la suma de cuadrados de la predicción:

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

donde y_i se calcula con las n-1 observaciones que quedan al omitir la iésima observación. El Cp es un estimador que mide la eficiencia de las variables, en términos de la suma de los cuadrados medios de residual estandarizado de la predicción o error total y viene dado por:

$$Cp = \frac{SCEp}{s^2} + 2p - n$$

donde n es el número de variables del modelo completo, p el número de variables del modelo reducido, s^2 el CME en el modelo completo y SCEp la suma de cuadrados para un modelo con p parámetros incluyendo el intercepto. Las desviaciones de Cp con respecto a p, se puede tomar como medida del sesgo, siendo E(Cp/sesgo = 0) = p. El sesgo cero es casi ideal, entonces lo que se necesita es un modelo adecuado. Los criterios mencionados se complementan con las pruebas totales sobre todos los coeficientes del modelo y con las pruebas parciales sobre cada coeficiente.

Métodos de selección de variables:

Existe la tendencia a incluir en modelos de regresión todas las variables científicamente relevantes, independientemente de su contribución al modelo. El problema con esta posición es que el modelo puede sobredimensionarse y producir estimados numéricamente inestables. Esta sobredimensión se manifiesta en coeficientes estimados y/o desviaciones estándar demasiado grandes, ello es particularmente problemático en problemas donde el número de variables es grande en relación con el número de sujetos. Es por ello que en algunos casos los métodos de selección son de mucha

importancia y se aplican de acuerdo el caso. En este trabajo se compararon solo dos métodos de selección de variables y no se consideran los métodos de selección hacia atrás (backward) y hacia delante (forward) en favor del paso a paso, ya que este realiza una reconsideración de las variables incluidas al realizar una inclusión nueva y además no se presentaron problemas de multicolinealidad lo cual haría que el backward fuese una mejor opción

Materiales y métodos

La muestra estuvo constituida por el total de bachilleres en Ciencias egresados en cinco promociones, la cual fue de 233 alumnos. Se prefirió trabajar con este número de alumnos, por ser relativamente accesible su estudio. Se encontraban distribuidos de acuerdo a las siguientes especialidades:

Agrícola	103 alumnos
Alimentos	56 alumnos
Pecuaria	43 alumnos

Conservación de los Recursos Naturales Renovables 31 alumnos

Definición de variables: Variable dependiente : Indice de rendimiento académico, la cual se define conceptualmente como la valoración cuantitativa del progreso del estudiante y se obtiene multiplicando la calificación dada en cada asignatura por el número de créditos que le corresponden, se suman los productos obtenidos y este resultado se divide entre la suma de los créditos computados. La escala de valorización es del 1 al 9. Su definición operacional establece que el puntaje requerido para egresar está comprendido entre 6 y 9 , de acuerdo a lo establecido en los reglamentos de la institución, cuyas categorías son las siguientes:

Bueno	6
Distinguido	7
Sobresaliente	8
Excelente	9

Variables explicativas o regresoras: Se utilizaron 28, algunas son cuantitativas y otras cualitativas, para las últimas se utilizaron variables falsas cuya codificación aparece en la columna 5 del cuadro 1.

Técnicas para el análisis estadístico: Se utilizó el análisis de regresión lineal múltiple, en donde uno de los supuestos básicos es que las variables independientes no estén fuertemente correlacionadas en cuyo caso los coeficientes generados pueden tener graves errores muestrales, lo que afectaría las predicciones (2). El plan a seguir en la construcción del modelo comprendió tres etapas: planificación, desarrollo y validación.

Planificación: En esta etapa se siguieron los pasos consistentes en definir el problema, seleccionar la respuesta y sugerir las variables de mayor importancia y evaluarlas. Posteriormente se obtuvo una muestra de las observaciones (muestra piloto) y se calcularon los valores estadísticos básicos. Se realizó un examen de los residuales con la finalidad de detectar el incumplimiento de los supuestos del análisis de regresión. Se estudió también la matriz de correlación para verificar si existían problemas de multicolinealidad y se analizaron los coeficientes de determinación de los modelos (R^2). Se aplicaron trans-

Cuadro 1. Definición y características de las variables explicativas.

Nº	Variable	Notación	Definición operacional	Codificación
1	Sexo	X_6	Sexo del estudiante	0: masculino 1: femenino
2	Estado civil	X_2 X_3	Estado civil del estudiante	00: soltero 10: casado 01: otro
3	Procedencia del estudiante	X_7 X_8 X_9 X_{10}	Sitio donde vivía el estudiante antes de iniciar sus estudios	0000: Yaracuy 1000 Lara 0100 Portuguesa 0010 Región de los llanos 0001 Otras regiones
4	Tipo de institución	X_4 X_5	Tipo de institución donde el estudiante culminó sus estudios de Bachillerato	00 Oficial 10 Privado 01 Otro
5	Régimen de estudios	X_{11} X_{12}	Régimen de estudios durante el Bachillerato	00: Regular 10: Parasistema 01 Otro
6	Turno	X_{13}	Turno de la institución donde el estudiante realizó sus estudios	0: Diurno 1: Nocturno
7	Lugar al cual pertenece el liceo	X_{14} X_{15}	Lugar al cual pertenece el liceo donde el Bachiller obtuvo el título	00: Capital 10: Distrito 01: Otro
8	Lugar de residencia	X_{16} X_{17} X_{18}	Lugar de residencia del estudiante durante sus estudios Universitarios Yaracuy	000: San Felipe 100: Otra región de 010 Barquisimeto 001 Otro

Cuadro 1. Definición y características de las variables explicativas (Continuación).

Nº	Variable	Notación	Definición operacional	Codificación
9	Tipo de alojamiento	X_{19} X_{20}	Tipo de alojamiento durante sus estudios universitarios 01: otro	00: con familia 10: residencia
10	Edad	X_{21}	Edad del estudiante al iniciar sus estudios universitarios	17,18,...
11	Duración del bachillerato	X_{22}	Años que tardó el estudiante en realizar sus estudios de Bachillerato	5,6,7,...
12	Tiempo para ingresar al instituto	X_{23}	Diferencia entre la edad al graduarse el estudiante y la edad al iniciar sus estudios	Menos de 1 año,1,2,...
13 . . . 28	Calificación del Bachillerato	X_{24} X_{25} . . . X_{38} X_{39}	Nota obtenida en las asignaturas: Historia Universal (X_{24}), Dibujo (X_{25}), Geografía Universal(X_{26}), Historia de Venezuela(X_{27}), Geografía de Venezuela(X_{28}),Geografía Económica(X_{29}),Filosofía(X_{30}), Ciencias de la Tierra(X_{31}), Castellano(X_{32}), Matemática(X_{33}), Biología(X_{34}), Inglés(X_{35}), Física(X_{36}), Educación Artística(X_{37}), Sociales(X_{38}) y Química(X_{39})	10,11,...,20

formaciones a algunas de las variables regresoras y ello aumentó la determinación de los modelos. En el presente trabajo se utilizó la raíz cuadrada, la cual se apropia en los casos en los cuales los datos estadísticos son números enteros positivos, tal como es el caso presente, ya que edad, duración del bachillerato, tiempo para ingresar y las calificaciones en algunas de las asignaturas estudiadas, fueron reportadas en esta manera. Las variables transformadas fueron denotadas: edad (RX_{21}), duración del bachillerato (RX_{22}), tiempo para ingresar (RX_{23}) y las calificaciones en las asignaturas : Historia Universal (RX_{24}), Dibujo (RX_{25}), Geografía Universal (RX_{26}), Historia de Venezuela (RX_{27}), Geografía de Venezuela (RX_{28}), Geografía Económica (RX_{29}), Filosofía(RX_{30}) y Ciencias de la Tierra(RX_{31}).

Desarrollo: En esta etapa se recolectó la totalidad de los datos, se verificó su calidad y se aplicaron modelos tentativos. Luego se sometieron los modelos a la consideración de especialistas en la materia, se realizaron los análisis gráficos y estudio de residuales y se verificó si los modelos satisfacían las metas propuestas con relación a los coeficientes de determinación de los mismos. Se compararon también dos métodos de selección de variables: el paso a paso (stepwise) y el de todas las regresiones posibles.

Validación: Una vez comprobado que los modelos cumplían las metas propuestas se estableció la etapa de validación, la cual es una etapa muy útil y necesaria y a veces puede conducir a una reconsideración total del problema. Existen varios métodos de validación, en esta investigación se utilizó el criterio Press el cual consiste en la suma de cuadrados de los residuales calculados eliminando la i-ésima observación del total de los datos.

Resultados y discusión

Estimacion de los modelos de regresion

Para cada una de las cuatro especialidades, se obtuvo tanto el modelo completo como los modelos reducidos por dos métodos de selección de variables. Se verificó el cumplimiento del supuesto de normalidad con la prueba de Wilk y Shapiro y se consideró que valores de $W \geq 0,9$ tenían una buena aproximación a la distribución normal y la homogeneidad de la varianza de los residuales, en cuyo caso al graficarlos se encontró que se formaron bandas horizontales, no visualizándose ningún patrón que indicara la violación a dicho supuesto. Se estudió la matriz de correlación, no observándose ninguna evidencia de problemas de multicolinealidad; ello se confirmó con el estudio de las raíces características de cada modelo en donde $K=\lambda_1/\lambda_2$ fue el indicador de la no existencia de problemas de esta naturaleza ya que en todos los casos estuvo por debajo de 100.

En la especialidad Agrícola (cuadro 2), las pruebas de t del modelo completo resultaron significativas solo las variables X_6, X_{16}, RX_{21} y X_{39}.

Cuadro 2. Comparación del modelo completo con los modelos seleccionados usando dos métodos de selección de variables en la especialidad agrícola.

Modelo		R^2	Cp	Nº VAR
Completo	$\hat{y}$=4,945+0,192X_4-0,147X_6+0,186X_7+0,011X_8+0,016X_{10}+0,080X_{13}+0,002X_{14} -0,180X_{16}-0,003X_{19}-0,412RX_{21}-0,051RX_{22}+0,165RX_{23}-0,165RX_{24}+0,181RX_{25} -0,078RX_{26}+0,077RX_{27}+0,142RX_{28}-0,001RX_{29}+0,042RX_{30}+0,253RX_{31}+0,065X_{32} -0,067X_{33}-0,020X_{34}+0,030X_{35}-0,051X_{36}-0,013X_{37}+0,007X_{38}+0,085X_{39}	60,05	29	28
Paso a paso	$\hat{y}$=5,593+0,205X_4-0,163X_6+0,180X_7-0,184X_{16}-0,440RX_{21}+0,135RX_{23}+0,184RX_{24} +0,173RX_{25}+0,219RX_{31}+0,062X_{32}-0,072X_{33}+0,030X_{35}+0,051X_{36}+0,089X_{39}	58,16	6,86	14
Todas las Reg.	$\hat{y}$=5,278+0,195X_4-0,152X_6+0,177X_7-0,197X_{16}-0,462RX_{21}+0,143RX_{23} +0,170RX_{25}+0,245RX_{31}+0,064X_{32}-0,073X_{33}+0,051X_{36}+0,091X_{39}	56,41	6,20	12

Todas ellas fueron incluidas en los modelos seleccionados por ambos métodos. Se hizo un estudio de los coeficientes de las variables seleccionadas, detectándose que eran relativamente estables al compararlos con el modelo completo. En este caso fue preferible el modelo seleccionado por el método de todas las regresiones posibles ya que contenía dos variables regresoras menos con muy poca disminución en el R^2 y un menor Cp. De las 28 variables estudiadas (sin considerar las ficticias) se incluyeron: sexo X_6, procedencia del estudiante X_7 , tipo de institución X_4, lugar de residencia X_{16} edad del estudiante RX_{21} , tiempo de ingreso RX_{23} y 6 de las 16 calificaciones en asignaturas de bachillerato. En las variables cualitativas se consideró que si una de las categorías presentaba significación la variable era significativa. En las variables cuantitativas se pueden analizar los signos de los coeficientes: edad del estudiante RX_{21} es negativa, por ello el rendimiento disminuye en estudiantes de mayor edad; tiempo de ingreso RX_{23} es positivo, por ello el rendimiento es mayor en estudiantes que tienen un período mayor de espera antes de comenzar sus estudios universitarios. En cuanto a las asignaturas: el rendimiento en la especialidad agricola es mayor en estudiantes con altas calificaciones en Dibujo RX_{25}, Ciencias de la Tierra RX_{31}, Castellano X_{32}, Física X_{36} y Química X_{39} pero bajas en Matemáticas X_{33}. Por ello en esta especialidad los estudiantes de bachillerato destacados en asignaturas científicas (sin considerar matemáticas) y buen manejo del idioma, se espera que obtengan buenos índices de rendimiento académicos en sus estudios en esta especialidad.

En la especialidad C.R.N.R. (cuadro 3), las pruebas de t del modelo completo resultaron significativas las variables X_4, X_8, X_{16}, RX_{21} , RX_{22}, RX_{23} , RX_{24} ,RX_{27} ,RX_{28} ,RX_{29}, RX_{30}, X_{34}, X_{36} ,X_{39} . No todas ellas fueron incluidas en los modelos seleccionados por ambos métodos. La diferencia entre los modelos seleccionados fue considerable. Hubo una diferencia del 10% en el R^2 y en cuanto al número de variables de 10. Los coeficientes de regresión de los modelos seleccionados al compararlos con los del modelo completo no eran estables, esto es hubo algunos cambios de magnitud e inclusive de signos. Considerando el principio de parsimonia, se escogió el modelo con menor número de variables (paso a paso) el cual incluyó: procedencia del estudiante X_{10}, duración del bachillerato RX_{22}, tiempo para ingresar RX_{23}, y 9 de las 16 calificaciones en asignaturas de bachillerato. Analizando los signos de los coeficientes en las variables cuantitativas se pudo observar: duración del bachillerato RX_{22} y tiempo para ingresar RX_{23} son positivos, por ello los estudiantes que obtuvieron el título de bachiller en mayor tiempo y tardan más tiempo para ingresar tienen mejor rendimiento en esta especialidad. En cuanto a las asignaturas, el rendimiento es mayor en estudiantes con altas calificaciones en Historia Universal RX_{24} , Geografía de Venezuela RX_{28} , Castellano X_{32} y Educación Artística X_{37} y el rendimiento es menor en aquellos que tenían altas notas en Histo-

Cuadro 3. Comparación del modelo completo con los modelos seleccionados usando dos métodos de selección de variables en la especialidad C.R.N.R.

Modelo		R^2	Cp	Nº VAR
Completo	$\hat{y}$=12,398-0,096X_4-0,527X_6-0,702X_8-0,600X_{10}+0,457X_{13}+0,013X_{14} -0,378X_{16}+0,297X_{19}-3,454RX_{21}+1,073RX_{22}+0,606RX_{23}+2,163RX_{24} -0,520RX_{27}+1,659RX_{28}+0,856RX_{29}-1,107RX_{30}+0,336RX_{31}+0,060X_{32} +0,073X_{33}-0,330X_{34}+0,044X_{35}-0,120X_{36}-0,014X_{37}-0,052X_{38}-0,173X_{39}	99,12	27	26
Paso a paso	$\hat{y}$=0,564+0,349X_{10}+0,322RX_{22}+0,754RX_{23}+0,441RX_{24} -0,433RX_{27}+1,788RX_{28}-0,627RX_{30}+0,248X_{32}-0,092X_{33} -0,179X_{34}-0,054X_{35}+0,087X_{37}	89,66	18,00	12
Todas las Reg.	$\hat{y}$=0,632-0,513X_4-0,294X_6-0,442X_7+0,623X_8+0,286X_{19} -0,586RX_{21}+0,663RX_{22}+0,493RX_{23}+1,584RX_{24}+0,774RX_{25} -0,858RX_{27}+2,034RX_{28}+0,487RX_{29}-0,992RX_{30}+0,320RX_{31} +0,240X_{32}-0,154X_{33}-0,278X_{34}-0,045X_{35}+0,052X_{36}+0,072X_{38}-0,223X_{39}	99,57	20,88	22

ria de Venezuela RX_{27} , Filosofía RX_{30} , Matemática X_{33}, Biología X_{34} e Inglés X_{35} . Esto condujo a pensar que los estudiantes inclinados a las asignaturas científicas no serán los de mejor desempeño en esta especialidad ,sino aquellos inclinados a las humanísticas.

En la especialidad Alimentos (cuadro 4), las pruebas de t del modelo completo resultó significativa solo la variable X_{39} (Calificación en Química durante el bachillerato). Ella fue incluida en los modelos seleccionados por ambos métodos. En relación a los coeficientes de regresión de las variables seleccionadas en ambos modelos al compararlos con los del modelo completo, se observó que hay leves variaciones, pudiendo decirse que eran relativamente estables. En esta especialidad ambos métodos de selección produjeron modelos que explicaron muy poco acerca de la variabilidad total; para los propósitos de predicción no son convenientes y por ello se propuso el modelo completo el cual tenía una determinación de aproximadamente un 70%. Este es un modelo con numerosas variables regresoras, de allí que la ganancia en determinación fue a costa de mayor complejidad. Se considera que todas las variables en estudio son de importancia para la predicción del rendimiento estudiantil en esta especialidad.

En la especialidad Pecuaria (cuadro 5), en las pruebas de t del modelo completo resultó significativa solo la variable X_{17}. Ella fue incluida en los modelos seleccionados por ambos métodos. Hubo variaciones en cuanto a los coeficientes de regresión

Cuadro 4. Comparación del modelo completo con los modelos seleccionados usando dos métodos de selección de variables en la especialidad alimentos.

Modelo		R^2	Cp	Nº VAR
Completo	$\hat{y}$=6,230+0,500X_4+0,422X_6-0,956X_7-0,990X_8-1,155X_9 -0,794X_{10} +0,107X_{12} +1,063X_{13} +0,107X_{14}-0,064X_{15}-0,159X_{16} -0,396X_{18}+0,813X_{19}-0,987X_{20}-0,252RX_{21}-0,628RX_{22}+0,530RX_{23} +0,571RX_{24}-0,108RX_{25} +0,181RX_{26}-0,298RX_{27}+0,200RX_{28} +0,184RX_{29}+0,618RX_{30}-0,186RX_{31}-0,220X_{32}-0,072X_{33} +0,085X_{34} -0,071X_{35}+0,085X_{36}-0,094X_{37}+0,059X_{38} +0,181X_{39}	69,33	35	33
Paso a paso	$\hat{y}$=4,565+0,043X_{35}-0,037X_{37} +0,064X_{38} +0,096X_{39}	37,60	10,70	4
Todas las Reg,	$\hat{y}$=4,966+0,870X_{13}-1,686X_{20}-0,328RX_{21} +0,438RX_{23} +0,359RX_{30} +0,108X_{39}	46,98	6,18	6

Cuadro 5. Comparación del modelo completo con los modelos seleccionados usando dos métodos de selección de variables en la especialidad pecuaria.

Modelo		R^2	Cp	Nº VAR
Completo	$\hat{y}$=4,243-0,037X_2-0,184X_4-0,103X_6-0,059X_{10}-0,280X_{14} -0,041X_{16}-1,100X_{17}+0,985X_{18}-0,092X_{19}-0,211RX_{21} -0,4321RX_{22}+0,297RX_{23}+0,275RX_{24}-0,202RX_{25} -0,133RX_{26}+0,581RX_{27}+0,319RX_{28}+0,162RX_{29} -0,192RX_{30}+0,248RX_{31}+0,018X_{32}+0,006X_{33}+0,020X_{34} +0,035X_{35}-0,004X_{36}+0,009X_{37}-0,022X_{38}+0,036X_{39}	85,88	29	28
Paso a paso	$\hat{y}$=4,633-0,241X_4-0,087X_6-0,304X_{14}-0,881X_{17}+0,917X_{18} -0,405RX_{21}-0,240RX_{22}+0,370X_{23}+0,276RX_{24}-0,407RX_{25} -0,156RX_{26}+0,595RX_{27}+0,195RX_{28}+0,286RX_{29}+0,241RX_{31}+0,025X_{34}	84,22	18,8	16
Todas las Reg,	$\hat{y}$=5,653-0,727X_{17}+0,797X_{18}-0,568RX_{21}+0,445RX_{23} +0,473RX_{27}+0,292RX_{28}+0,009X_{35}-0,006X_{39}	73,10	3,18	8

de las variables incluidas en los modelos seleccionados pero se consideró relativamente estable. En este caso fue preferible (considerando el criterio de parsimonia) el modelo seleccionado por el método de todas las regresiones posibles ya que contenía solo ocho variables regresoras lo cual se tradujo en ganancia en precisión y menor complejidad en el modelo a costa de pérdida en determinación ya que la explicación de la variabilidad total de este modelo difirió en un 10% del modelo paso a paso, pero aún así la explicación del 73% fue bastante buena. Las variables que se incluyeron fueron: lugar de residencia X_{17} y X_{18}, edad X_{21}, tiempo para ingresar X_{23}, y 4 de las 16 calificaciones en asignaturas de bachillerato. Considerando el signo de las variables cuantitativas se pudo observar que edad RX_{21} es negativo y tiempo para ingresar RX_{23} es positivo, por ello el rendimiento será superior en estudiantes más jóvenes pero que tardan más tiempo en ingresar, las asignaturas de bachillerato que influyeron en esta carrera indicaron que cuanto mayor fue su calificación en las asignaturas Historia de Venezuela RX_{27}, Geografía de Venezuela RX_{28} e Inglés X_{35} es de esperarse que rendimiento de los estudiantes sea mejor en la especialidad Pecuaria, mientras que en relación a la asignatura Química X_{39} a medida que la nota sea menor en bachillerato el rendimiento será mejor.

Validacion de los modelos seleccionados por los métodos en estudio:

Con la finalidad de examinar la precisión de los modelos seleccionados por los métodos Paso a Paso y de Todas las regresiones posibles en cada una de las especialidades se utilizó el método PRESS. En el cuadro 6 se presenta el resumen de tales resultados:

En todos los casos resultaron ser más precisos los modelos seleccionados por el método de Todas las regresiones posibles; sin embargo, las diferencias fueron muy pequeñas y ambos modelos pudieron considerarse buenos.

Cuadro 6. Valores PRESS para los modelos seleccionados.

Especialidad	Paso a paso	Todas las regresiones posibles
Agrícola	8,23	8,20
Alimentos	10,31	9,95
Pecuaria	4,42	2,90
C,R,N,R,	1,76	0,33

Conclusiones

Los resultados obtenidos en los análisis anteriores permiten establecer resultados generalizables hasta el límite definido por las características de la muestra y a medida que las exigencias y condiciones sean semejantes. Las técnicas de análisis de regresión múltiple en presencia de varia-

bles cualitativas y cuantitativas permitieron establecer modelos con el fin de sugerir al estudiante que ingrese a la carrera, donde alcanzará el mayor nivel de expectativas permitiendo también incrementar el interés del estudiante hacia una profesión relacionada con el mayor éxito posible. Los modelos seleccionados de acuerdo a su determinación, precisión, número de variables y el criterio del educador, fueron:

Para la especialidad agricola:

$\hat{y}$= 5,278 + 0,195X_4 -0,152X_6 + 0,177X_7 -0,197X_{16} -0,462RX_{21} + 0,143RX_{23} + 0,170RX_{25} + 0,245RX_{31} + 0,064X_{32} -0,073X_{33} + 0,051X_{36} + 0,091X_{39}

Para CRNR:

$\hat{y}$=0,564+ 0,349X_{10}+ 0,322RX_{22}+ 0,754RX_{23}+ 0,441RX_{24}- 0,433RX_{27}+1,788RX_{28}-0,627RX_{30}+ 0,248X_{32}-0,092X_{33}-0,179X_{34}-0,054X_{35}+ 0,087X_{37}

Para alimentos:

$\hat{y}$=6,230+0,500X_4+0,422X_6- 0,956X_7- 0,990X_8- 1,155X_9- 0,794X_{10}+0,107X_{12}+1,063X_{13}+0,107X_{14}- 0,064X_{15}-0,159X_{16}-0,396X_{18}+0,831X_{19}- 0,987X_{20}-0,252RX_{21}-0,628RX_{22}+ 0,530RX_{23}+0,571RX_{24}-0,108RX_{25}+ 0,181RX_{26}-0,298RX_{27}+ 0,200RX_{28}+ 0,184RX_{29}+ 0,618RX_{30}-0,186RX_{31}- 0,220X_{32}- 0,072X_{33}+0,085X_{34}- 0,071X_{35}+0,085X_{36}- 0,094X_{37}+ 0,059X_{38}+ 0,181X_{39}

Para pecuaria:

$\hat{y}$=5,653-0,727X_{17}+0,797X_{18}- 0,568RX_{21}+ 0,445RX_{23}+ 0,473RX_{27}+ 0,292RX_{28}+ 0,009X_{35} -0,006X_{39}

Estos modelos explicaron respectivamente el 56,41%; 89,66%, 69.33% y el 73,10% de la variabilidad total del rendimiento y las variables escogidas difieren de acuerdo a la especialidad. Los dos métodos de selección de variables discutidos en este trabajo fueron instrumentos muy útiles, sin embargo, en este trabajo no se pudo establecer que un método sea mejor que el otro. Es de notar que en presencia de variables cualitativas y cuantitativas, las pruebas de t como pruebas parciales se comportaron mucho más estrictas que cualquiera de los métodos de selección estudiados.

Cuando se usa en forma mixta variables cuantitativas y cualitativas, la complicación del cálculo es manifiesta, debido al elevado número de variables ficticias que es preciso introducir cuando el número de variables cualitativas es alto e igualmente cuando se incrementa el número de niveles que cada una posee. Tal como fue mencionado, los coeficientes de regresión asociados a las variables dummy cuantifican el efecto producido por la presencia del correspondiente nivel de la variable explicativa, pero solo es posible establecer, si algún nivel es significativo y en cuyo caso la variable es significativa, pero no se puede a ciencia cierta establecer cual es la categoría más importante ya que el nivel de referencia codificado por 0 afecta el intercepto y este es único, por ende, en presencia de varias variables cualitativas es imposible su interpretación. Cuando la intención es determinar en una variable cualitativa cual(es) es el nivel(es) más influyente(s) sobre la variable respuesta, se recomienda fijar modelos de regresión para cada variable cualitativa o modelos de regresión múltiple donde se incluya solo una variable cualitativa y una cuantitativa.

Se recomienda validar este modelo en otras instituciones similares y de esa forma se podría ofrecer como aporte a todas las instituciones del país. Se recomienda además, continuar este estudio incluyendo otras variables que no fueron incluidas, con la finalidad de que la explicación de la variabilidad total de estos modelos se acerque más al 100%; así como también reducir el número de variables regresoras, haciendo uso de métodos multivariados tal como el de componentes principales, para luego realizar los análisis de regresión múltiple.

Literatura citada

1. Barlett, M.S. 1974. "The use of transformations" Biometrics, 3.1.

2. Chacín, F. 1998. Análisis de Regresión y Superficie de Respuesta. Maracay. Revista de la Facultad de Agronomía. U.C.V.

3. Chacín, F. 1999. Avances Recientes en el Diseño y Análisis de Experimentos. Maracay. Revista de la Facultad de Agronomía. U.C.V.

4. Faber, R. 1971. "Use of Dummy Variables in Regression Analysis". Mimeo ECIEL.

5. Hosmer, D.A. y S. Lemeshow. 1989.Applied Logistic Regression, John Wiley and Sons. New York. 307 p.

6. Martinez G. A. 1988. Teoría de la regresión con aplicaciones agronómicas Editorial Trillas. Primera edición. 490 p.

7. Ruiz-Maya L., F.J. Martín, J.M. Montero y P. Uriz. 1995. Análisis Estadístico de Encuestas: datos cualitativos. Colección Plan Nuevo. Editorial AC. Madrid. España.722 p.

8. Wrigley, N. 1985. Categorical data analysis for geographers and environmental scientists. Longman. London. 231p.