Construcción de Software para Regresión El caso de Selección del Modelo y Pruebas de Homocedasticidad

Autores:
Sindy Macías Cabrera, César Pincay Chiquito
Coautor:
Gaudencio Zurita Herrera
Instituto de Ciencias Matemáticas (ICM)
Escuela Superior Politécnica del Litoral (ESPOL)
Km 30.5 vía Perimetral, Guayaquil-Ecuador
svmacias@espol.edu.ec, cpincay@espol.edu.ec, gzurita@espol.edu.ec

Resumen

Resulta importante dentro del Análisis de Regresión determinar qué variables predictoras o de explicación son óptimas para explicar a la variable de interés, es por esto que existen indicadores de Selección del Modelo que permiten la determinación de estas variables. Uno de ellos y el más utilizado es el R^2_{adj}*, pero existen otras medidas de bondad de ajuste tales como el Criterio Akaike, estadístico Cp deMallows y PRESS; cada uno proporciona las posibles combinaciones de las (p-1) variables de explicación y están basados principalmente en la minimización de las medidas de variación del modelo de regresión, utilizando Suma y Media Cuadrática. Debido a que estas medidas de ajuste no son comunes en los softwares más usuales, se ha desarrollado ERLA (Estadística de Regresión Lineal Avanzada) el cual ayuda en la comparación conjunta de los valores y variables, quedando a decisión del investigador la elección de las mismas. Este software está constituido desde las técnicas más básicas hasta las más avanzadas como Regresión Rigde, Regresión Logística y por supuesto Selección del Modelo.*

Palabras Claves: *Selección del Modelo, Análisis de Regresión, Medidas de Bondad de Ajuste, ERLA, indicadores*

Abstract

Is important in the regression analysis to determine which predictors or explanations variables are optimal to explain to the variable of interest, which is why there are Model Selection indicators that allow the determine these variables. One of them, and the most used is R^2_{adj}*, but there are other measures of goodness of fit such as the Akaike criterion, Mallows Cp and PRESS statistic, each providing the possible combinations of the p-1 variables of explanation and are based primarily on minimization measures regression model variation, using Sum and Mean Square. Because of these adjustment measures are not common in most common software, has been developed ERLA (Advanced Linear Regression Statistics) which helps in comparing the values and variables and being so decision of the investigator's election thereof. This software is made from the most basic techniques to advance as Rigde Regression and Logistic Regression and of course Model Selection.*

Keywords*: Model Selection, Regression analysis, Goodness of fit measures, ERLA, indicators.*

1. Introducción

En la actualidad se encuentran en el mundo un sin número de paquetes o aplicaciones estadísticas los cuales permiten efectuar el análisis descriptivo, inferencial, de un conjunto de datos. Estos paquetes para llegar al mercado pasan por un proceso de transición en el cual se logran corregir errores o fallas. Día tras día se busca que los programas sean cada vez más amigables a la vista del usuario, sin perder por supuesto el propósito del mismo, es por todo esto que como proyecto de graduación en las aulas del Instituto de Ciencias Matemáticas de la ESPOL, nace la idea de desarrollar un programa que cumpla con lo antes propuesto, el cual es "ERLA".

El desarrollo de "ERLA" ha sido realizado en dos plataformas informáticas estas fueron Matlab[1] R2010a y Visual Net 2008[2], lográndose una conexión basados en una estructura cliente/servidor; esta conexión en el ambiente informático es administrada por el componente conocido como Middleware (COM) éste es un software de conectividad que ofrece un conjunto de servicios que hacen posible el funcionamiento de aplicaciones distribuidas sobre plataformas heterogéneas y COM es el tipo de Middleware que permite la conexión específica entre las dos plataformas usadas en nuestro caso.

"ERLA" es un paquete computacional direccionado a resolver problemas estadísticos utilizando Regresión Lineal. Este "paquete" contiene desde estadística básica como Tablas de Frecuencias, Estadísticas Descriptivas hasta Regresión de Ridge, Regresión Logística, Selección de Modelos, Puntos de Influencia y más. Siendo los indicadores de calidad de Selección de Modelos la contribución específica que se detallará en este reporte.

Como parte del análisis de regresión, se realiza una investigación básica a las variables objeto de estudio, todo esto con el fin de observar el comportamiento y las fortalezas de la relación entre ellas. Dicho de otra manera, se realiza el análisis descriptivo y determinamos las correlaciones entre dichas variables, para de esta manera observar qué variables son las que aportarían en proporción significativa a los modelos de regresión.

Ante esto nos vemos obligados a realizar empíricamente la selección de las variables explicativas, aquellas combinaciones de variables que de acuerdo con la matriz de correlación determinamos tienen mayor fortaleza con la variable respuesta. Existen métodos de selección de las variables explicativas, pero no son comunes en los softwares estadísticos más usuales.

Como tema específico en este trabajo se detallarán las técnicas que permiten determinar las posibles regresiones de un conjunto de variables explicativas, para una variable a ser explicada Y. Dichas técnicas, son las que utilizan R^2, R^2_{aj}, Criterio de Akaike, estadístico Cp de Mallows y PRESS.

2. Selección del modelo

Para decidir entre dos o más subconjuntos de variables explicativas en el estudio de un modelo de regresión múltiple es necesario disponer de indicadores que midan la bondad del ajuste del modelo construido. Para un modelo de p parámetros se supone que el número de variables explicativas que pueden haber en el modelo es (p -1), el número de observaciones es n. Entonces se definen las siguientes medidas de bondad de ajuste: R^2; R^2_{aj}; Criterio de Akaike; Estadístico Cp de Mallows; y, PRESS.

2.1. Coeficiente de determinación (R^2)

R2, Coeficiente de Determinación definido como:

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Este indicador intenta medir la calidad del modelo utilizado y aumenta al ir introduciendo nuevas variables en el modelo. Se denota R^2_j j=1,..., (p-1), el máximo valor posible de R^2 cuando en el modelo hay "j" variables explicativas, se puede probar que $R^2_{j-1} < R^2_j$, (R^2_j es monótona creciente) y las diferencias $R^2_j - R^2_{j-1}$ decrecen. En base a esto, al crecer "j" un criterio sería considerar un número pequeño que por conveniencia es denotado por "Δ" y elegir el modelo con "j" más pequeño y tal que $R^2_{p-1} - R^2_j < \Delta$; siendo R^2_{p-1} el coeficiente de determinación del modelo con las "p-1" variables explicativas.

A medida que se introducen variables en el modelo, la potencia de explicación aumenta y además R^2 tiene el inconveniente de no considerar el número de variables explicativas, lo que hace que tienda a sobre ajustar y utilizar demasiadas variables.

[1] El fabricante de Matlab es MathWorks

[2] Visual Net fue creado por Microsoft

El R^2_{p-1} es el coeficiente de determinación R^2 para un modelo con (p-1) variables de explicación "p" coeficientes de regresión, en líneas previas se dijo que:

$$R^2_{p-1} = \frac{SCR_{p-1}}{SCT}$$

Debido a que la SCT = SCR + SCE, manipulando algebraicamente se obtiene:

$$R^2_{p-1} = 1 - \frac{SCE_{p-1}}{SCT}$$

Donde SCE_{p-1} es la Suma Cuadrática del Error para el modelo con (p-1) variables de explicación, y $SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$ es la Suma Cuadrática Total que es la misma para todos los modelos donde "p-1" no cambia.

Es preferible tener modelos con R^2_{p-1} de mayor tamaño. Habrá varios modelos con "p-1" variables y cada uno tendrá un Coeficiente de Determinación (R^2) diferente. Esto tendría sentido para seleccionar el mejor o los mejores R^2 de los modelos de "p-1" variables.

2.2. R^2-Ajustado

El R^2_{adj} ajustado, tiene como principal importancia determinar la variabilidad explican las variables independientes, con respecto a la variable respuesta cuando se introduce una variable adicional al modelo.

El Coeficiente de Determinación Ajustado (R^2_{adj}) se define: por los grados de libertad asociados a la sumas cuadráticas; la SCE y la SCT son ajustados por (n-p-1) y por (n-1) que son sus grados de libertad respectivamente.

En términos de sumatorias R^2_{adj} se define por la expresión

$$R^2_{adj} = 1 - \frac{\frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Simplificando

$$R^2_{adj} = 1 - \frac{\frac{1}{n-(p+1)} SCE}{\frac{1}{n-1} SCT}$$

Quedando R^2_{adj} en términos del Coeficiente de Determinación R^2, definido por la siguiente expresión

$$R^2_{adj} = 1 - \frac{(n-1)}{(n-p-1)} (1 - R^2)$$

Dicha expresión en términos de varianzas se tiene que:

$$R^2_{adj} = 1 - \frac{s^2}{SCT/(n-1)} = 1 - \frac{s^2}{s^2_y}$$

Donde $s^2 = \frac{SCE}{(n-p-1)}$ es la Media Cuadrática de los Residuos, y s^2_y es la varianza de la muestra, sin ningún ajuste por variables de regresión. La ecuación anterior muestra que R^2_{adj} no aumenta necesariamente con una variable de explicación más. Si no hay mejoría en R^2_{adj} por la adición de una variable, que el término $\frac{(n-1)}{(n-p-1)}$ en realidad baja el R^2_{adj}. Por esta razón, se postula que el R2 ajustado es una mejor medida que R2 para la selección del modelo.

$$R^2_{adj} = 1 - \frac{(n-1)}{n-(p+1)} (1 - R^2) \Leftrightarrow R^2_{adj} \leq R^2$$

2.3. Varianza Residual (s^2_R)

Para cada valor x_i de X, se obtiene una diferencia (el residuo) entre el valor observado de Y y el correspondiente valor teórico obtenido en el modelo de regresión. Por lo tanto se define la VARIANZA RESIDUAL como la media de todos los residuos elevados al cuadrado:

$$s^2_R = \frac{1}{n-(p+1)} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = MCE$$

Donde MCE es la media cuadrática del error; un buen criterio de selección de variables explicativas es elegir el subconjunto de "j" variables que minimice el valor de MCE, siendo esta la varianza residual obtenida con el modelo de "j" variables de explicación.

Teniendo en cuenta que:

$$R^2_{adj} = 1 - \frac{1}{s_y^2} MCE$$

Se puede deducir que:

$$R^2_{adj(p-1)} > R^2_{adj.\,j} \Leftrightarrow MCE_{p-1} < MCE_j$$

Por lo tanto el criterio de minimizar la varianza residual es equivalente al criterio de maximizar el coeficiente de determinación ajustado.

El R^2_{adj} representa la reducción (proporcional) en la varianza residual obtenidos por el modelo de regresión. Es así que en el momento de considerar la selección del mejor modelo, no solo se deben observar los indicadores sino que además el valor de la varianza residual la cual. Es conveniente enfatizar que la varianza residual no se la considera como un indicador de selección de modelos, sino más bien como una guía para así determinar cuál de los indicadores es el que más conviene en el estudio de Regresión.

Se ha mencionado anteriormente que habrá más de un modelo fijo para (p-1) variables de explicación, en lugar de examinar todos estos modelos, se fijará la atención al mejor, por ejemplo, los mejores tres o cuatro modelos con mayores valores de $R^2_{adj.}$ y menores valores de s^2_R .

2.4. Estadístico C_p de Mallows"

Los criterios previos se basan en la Suma Cuadrática del Error "SCE", ahora se explicará un criterio que toma en cuenta la Media Cuadrática del Error (MCE, es decir la varianza del error) en la selección del modelo, lo que conlleva a que si se omite una variable explicativa importante que influya en la predicción, los estimadores de los coeficientes de regresión serían sesgados, es decir $E(\hat{\beta}_i) \neq \hat{\beta}_i$ lo cual indica que el objetivo de este indicador es minimizar la MCE, C_P de Mallows está definido como:

$$C_p = \frac{SCR_p}{s^2} - (n - 2p)$$

Donde p es el número de parámetros en un modelo de Regresión Lineal Múltiple, con (p – 1) el número de variables explicativas, es la varianza del error con todas las variables y SCE_p es la suma cuadrática del error al ir ajustando el modelo con p parámetros.

Para interpretar este estadístico, se define el Error Cuadrático Medio de predicción "ECMP" para los puntos observados cuando se utiliza un modelo con "p" parámetros como:

$$ECMP_p = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\hat{y}_{p,i} - m_{p,i}\right)^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\hat{y}_{p,i} - E\left(\hat{y}_{p,i}\right) + E\left(\hat{y}_{p,i}\right) - m_{p,i}\right)^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} var\left(\hat{y}_{p,i}\right) + Sesgo^2\left(\hat{y}_{p,i}\right)$$

Donde $\hat{y}_{p,i}$ es el valor ajustado cuando se utiliza el modelo con p parámetros y $m_{p,i} = E[y \mid X = x_{p,i}]$ siendo un buen criterio de selección del modelo el de elegir el modelo que tenga el ECMP (Error Cuadrático Medio de Predicción) mínimo.

También se puede probar que en los modelos sin sesgo $C_p = p$. Por lo tanto, aquellos subconjuntos de "p-1" variables explicativas que tengan un $C_p \cong p = j + 1$ son los mejores. Se puede construir una gráfica de Cp para los diferentes subconjuntos que se quieren analizar frente a p. Y se considerarán buenos a aquellos subconjuntos que tienen Cp pequeño que Cp = p.

En la "Figura 1" se puede observar el gráfico Cp para dos puntos de variables explicativas y se observa que el punto A tiene un sesgo mucho mayor que el del subconjunto B, pero éste tiene menor Cp.

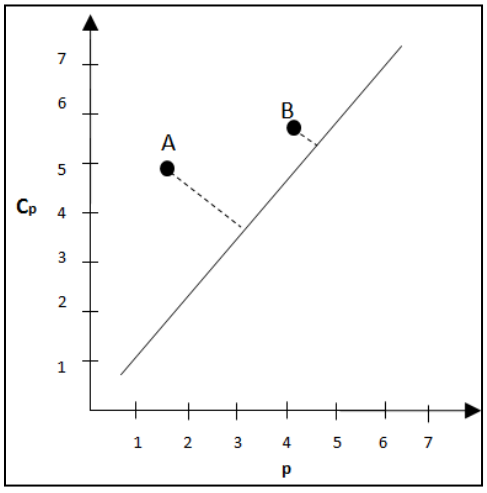

Figura 1 Representación Gráfica del Indicador C_P Mallows.

En estadística, Cp Mallows, llamado así por Colin Mallows, se utiliza a menudo como una regla de identificación para diversas formas de regresión paso a paso. Un punto a considerar es la colinealidad la cual en el análisis de regresión consiste en que las variables de explicación del modelo están relacionadas constituyendo así una combinación lineal.

Este inconveniente resulta ser muy frecuente en los modelos de regresión. A menudo muchas de las variables independientes se esperaría que tengan efectos que son altamente correlacionados y no se puede estimar por separado. Cuando hay demasiadas variables explicativas muchas de ellas cuyos coeficientes deben ser estimados, se han incluido en un modelo de regresión que se dice que está "sobre-ajustado."

El peor caso es cuando el número de parámetros a estimar es mayor que el número de observaciones, por lo que no pueden ser estimadas en absoluto. El estadístico "Cp" se puede utilizar en la selección de un modelo reducido sin problema, tanto tiempo como "S2" Error cuadrático Medio, es distinto de cero, lo que permite calcular "Cp".

El modelo con parámetros p. Denotemos el error cuadrático medio de este modelo por "S2". Nosotros suponemos que el modelo más grande da una descripción adecuada, y por lo tanto $E(S^2)= \sigma^2$.

Deteniéndose especialmente un modelo candidato con $q = p\text{-}1$ variables explicativas, $p \leq q$ y p escrito como parámetros $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Cuando $\mathbf{X}_1$ contiene 1 (la columna de unos) y los vectores (p-1) variables explicativas. Si este modelo más pequeño ya es adecuado, entonces:

$$\frac{SCE_p}{\sigma^2} \sim \chi^2_{n\text{-}p\text{-}1}$$

Los modelos bajo supuestos de normalidad e independencia estocástica, que se consideran más opcionales son aquellos con pocas variables y $C_p \cong p$. Una vez se haya encontrado ese modelo, no hay necesidad de emplear un modelo más complicado que involucra a más de (p-1) variables.

Se concluye que el mejor modelo es aquel que no tiene falta de ajuste ("underfitting") ni alto sobreajuste ("overfitting") en los datos.

Falta de ajuste, se da cuando el estimado del valor predicho de la variable de respuesta tiene alto sesgo y poca varianza,

Sobreajuste, se da cuando la varianza del estimado del valor predicho es alta, pero el sesgo es bajo.

2.5. Criterio de información Akaike (AIC)

El indicador AIC derivado del denominado Criterio de Información Akaike, otra medida de bondad de ajuste y de un modelo de Regresión; fue desarrollado por el científico Japonés Hirotsugu Akaike y publicado por primera vez bajo el nombre de "criterio de información", se basa en la entropía de la información, el cual ofrece una medida relativa de la pérdida de información cuando un determinado modelo se utiliza para describir la realidad.

El AIC no es una prueba del modelo en el sentido de las pruebas de hipótesis, sino que proporciona un medio para la comparación entre modelos, un criterio para la selección del modelo.

Dado un conjunto de datos, varios posibles modelos pueden ser clasificados de acuerdo a su AIC, los modelos con valores más pequeños de la AIC son los preferidos.

Así se define el AIC como:

$$AIC_p = n\left[\ln\left(\frac{SCE_p}{n}\right)\right] + 2(p+1)$$

El primer término en la expresión anterior es, como en la Cp de Mallows, una medida de bondad de ajuste (disminuye al crecer el de la estimación por máxima de la verosimilitud); el segundo penaliza el número de parámetros.
El segundo término, 2(p+1), representa una función que aumenta, con el número de parámetros estimados.

2.6. Suma de Cuadrados de predicción (PRESS)

Este indicador de calidad de los modelos de regresión fue propuesto por Allen en 1974, de una combinación de todas las regresiones posibles, basado en el análisis de residuales y validación cruzada, la cual consiste en estimar los modelos con una muestra (muestra de entrenamiento o aprendizaje) y evaluarlos examinando su comportamiento en la predicción de otra diferente (muestra de validación). Supongamos que hay p parámetros en el modelo y que tenemos "n" observaciones disponibles para estimar los parámetros del modelo, en cada paso se deja de lado la i-ésima observación del conjunto de datos y se calculan todas las regresiones posibles; se calcula la predicción y el residual correspondiente para la observación que no fue incluida, el cual es llamado el residual "PRESS".

Se puede expresar esta medida:

$$e_i = \frac{e_i}{1-h_{ii}}$$

como una función de los residuales ordinarios $e_i = y_i - \hat{y}_i$ y los términos de apalancamiento hij del modelo de regresión original.

Siendo $1-h_{ii}$ parte de la Suma cuadrática del error, visto en el capítulo anterior.

Donde la medida de Sumas Cuadradas de Predicción "PRESS" para el modelo de regresión que contiene "p" parámetros se define por:

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2$$

O equivalente a

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}} \right)^2$$

En conclusión se dice que el mejor modelo entre varios es aquel que tiene el menor valor del índice "PRESS".

3. Software ERLA (Estadística de Regresión Lineal Avanzada)

ERLA es un software desarrollado para ser implementado en Microsoft Windows, para el cual se utilizó Visual Basic.NET y Matlab.

La utilización básica de estos dos programas es Visual Basic.NET para la presentación de la interfaces de interacción con el usuario y Matlab para el desarrollo de las funciones matemáticas y estadísticas.

Figura 2 Inicio ERLA

3.1 Selección del Modelo en ERLA

El ítem selección de modelos, permite al usuario determinar el mejor modelo de regresión, basado en los indicadores R Ajustado, Cp Mallows, Akaike y el estadístico PRESS. De acuerdo con el ejemplo propuesto, "Imagen de la ESPOL en Guayaquil", para este caso se considera el conjunto de proposiciones de las cuales como variable a ser explicada se escoge la proposición 3 y como variables de explicación las proposiciones 2, 6, 11, 15, 10, 12, 17 y 18.

Los pasos a seguir en ERLA son:

1. **Barra de menú➢ Análisis de datos ➢Selección de modelos**

2. En el cuadro de diálogo Selección de modelos, seleccione la variable a ser explicada y las variables de explicación y pulse el botón Opciones.

3. En el cuadro de diálogo Indicadores seleccione el estadístico mediante el cual desea hacer la comparación y pulse el botón Aceptar. Los resultados se mostraran en la ventana de reporte. Véase Figura 3

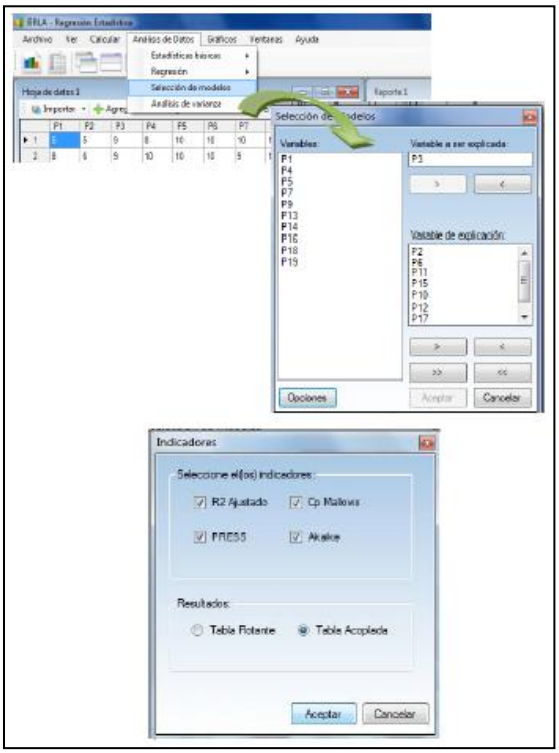

Figura 3 Selección del Modelo

4. Conclusiones y Recomendaciones

4.1 Conclusiones

Las tecnologías de la información (TI) ofrecen grandes posibilidades al mundo de la educación. Pueden facilitar el aprendizaje de conceptos y materias, ayudar a resolver problemas y contribuir a desarrollar las habilidades cognitivas.
Se enuncian las principales conclusiones derivadas del Trabajo Especial de Grado expuesto.

• Existen numerosas técnicas para la construcción de un software estadístico, por lo que es importante escoger y determinar las que mejor se adapten al contexto y a las necesidades que se deseen satisfacer, así como a las características de la población objetivo.

• Asimismo el lenguaje de programación Microsoft Visual Basic 8.0 de la familia de Microsoft Visual Studio 8.0 permitió el desarrollo de un software con una interface amigable con el usuario la cual satisface el requerimiento de ser apto para fines educativos; además de que el usuario final fue un programa computacional con características profesionales y que permiten su fácil entendimiento, entre las cuales se pueden mencionar cuadros de dialogo, consejos como ayuda. Menú emergente para el manejo de resultados, etc.

• Si bien hay en el mercado diversas opciones de software estadísticos, su utilización se limita en gran parte a la parte básica de la técnica de regresión, por lo que es importante fomentar a "ERLA" en su desarrollo e implementación para que se incremente su uso en las aulas de clase, así como en los diferentes niveles de investigación.

• El sistema de software presentado está asentado en los principios de las teorías constructivistas, ya que se basa la construcción del conocimiento en la capacidad de cada individuo, apoyando así la construcción inicial de modelos predictivos. Sin embargo es importante señalar que un software estadístico basado en un sólo enfoque estaría incompleto, por lo que es necesario involucrar aspectos de las demás teorías existentes, como se lo ha realizado con "ERLA".

• El desarrollo de un software estadístico incluye profesionales y/o expertos, por lo que a una primera instancia fue necesario considerar un número de graduandos, en el proceso para determinar, de manera más completa, los aspectos que influyen en el proceso de construcción y aprendizaje, para así lograr un mejor desarrollo y uso de "ERLA".

• La Cátedra de Regresión Lineal Avanzada tiene como uno de sus objetivos "Relacionar los conocimientos adquiridos de Ingeniería Clásica con aplicaciones avanzadas y recientemente descubiertas por especialistas en el tema, mediante la elaboración de simulaciones de problemas con la ayuda del computador". Sin embargo esto poco se lleva a la práctica, ya que las actividades o tareas orientadas a cumplir con este objetivo no se han mantenido ni aprovechado de la manera más eficiente con el paso del tiempo, por lo que es vital desarrollar aplicaciones que permitan lograr el objetivo citado.

• El presente Reporte Especial de Grado puede servir de base para su expansión y adaptación a otros tópicos o temas y/o para futuros proyectos en ésta y otras áreas de conocimiento.

• Todo sistema de software depende del apoyo que reciba, de Entidades ya sean Públicas o Privadas; y de la utilización del mismo, por lo que el éxito de este proyecto depende del uso, impulso y aplicación de la Escuela Superior Politécnica del Litoral "ESPOL" y profesionales.

4.2. Recomendaciones

Desde la concepción del desarrollo de un sistema de software surgen ideas que deben ser descartadas para poder determinar el alcance del proyecto, sin embargo, dichas ideas pueden servir de base para la expansión y mejoramiento del proyecto.
Algunas de las recomendaciones se exponen en las líneas siguientes:

• Disminuir la incertidumbre en la administración del software en los distintos módulos, usando el manual de usuario.

• Elaborar módulos de estadísticas, donde los usuarios pueden consultar el rendimiento del Software (individual o por sección) y los usuarios puedan consultar su rendimiento de forma personal o global con respecto al Software.

5. Referencias Bibliográficas

[1] **Bovas A. y Johannes L.** (2006) *Introduction to Regression Modeling*, Primera Edición, Thomson Brooks/Cole, USA.

[2] **Zurita G.** (2010) *Probabilidad y Estadística, Segunda Edición*, Centro de Difusión y Publicaciones - ESPOL, Guayaquil, Ecuador.

[3] **Rencher A.** *Methods of Multivariate Analysis*, Segunda Edición, Wiley Interscience.

[4] **Freund J., Miller I., Miller M.** (2000) *Estadística Matemática con Aplicaciones*, Sexta Edición, Prentice Hall, México.

[5] **Timm N.** (2002) *Applied Multivariate Analysis*, Springer, New York, USA.

[6] **Mallows, C.** (1973) *Some comments on Cp*, *Techmetrics*, 15: 661 – 664.

[7] **Contreras Juana, Del Pino Claudio** (2011) *Matemática interactiva*, http://matesup.utalca.cl

[8] **Universidad de Málaga. (2011)** *Bioestadística: Métodos y Aplicaciones*, http://www.bioestadistica.uma.es/libro/node97.htm

[9] **Universidad Nacional de Colombia.** (2011) *Métodos de Regresión*, http://www.virtual.unal.edu.co/cursos/ciencias

[10] **Galton F.** (1889) *Natural Inheritance*, Primera Edición, Macmillan, Londres.

[11] **ReliaSoft Corporation.** (2011) *Hypothesis Tests in Multiple Linear Regression*, http://www.weibull.com

[12] **Lopez, E.** (1998) *Tratamiento De La Colinealidad en Regresión Múltiple*, 10: 491 – 507.