Escuela Superior Politécnica del Litoral
Materia de Graduación: Regresión Lineal Avanzada

"Construcción de Software para Regresión: El Caso de Selección de Modelos y Pruebas de Homocedasticidad"

Previa a la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Graduandos:

Macías Cabrera Sindy Victoria

Pincay Chiquito César Alfonso

ERLA

Contenido

- Introducción
  1. *Modelos de Regresión*
  2. *Selección de Variables de Predicción*
  3. *Acerca de ERLA*
  4. *Validación del Modelo en el Software ERLA*
- Conclusiones y Recomendaciones

Introducción

- Análisis de Regresión.
- Medidas de bondad de Ajuste
- Desarrollo de ERLA.

Modelos de Regresión

- **Regresión Polinómica**
  - se tiene una variable dependiente y una variable de explicación, que se relacionan por un modelo polinómico.

$$y = \beta_0 + \beta_1 x^2 + \beta_2 x^3 + \varepsilon$$

- **Regresión Lineal Simple**
  - En este caso se tiene una variable independiente, una variable dependiente y una relación rectilínea entre ellos.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

...viene Modelos de Regresión

- **Regresión Lineal Múltiple**
  - Para este caso se tiene a una variable dependiente y varias variables de explicación o independientes.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i, \quad i = 1, 2, 3, \ldots, n$$

- **Supuestos:**

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases} \qquad E(\varepsilon_i) = 0 \qquad \varepsilon \sim N(0, \sigma^2)$$

…viene Modelos de Regresión

- **Representación Matricial del Modelo de Regresión Lineal Múltiple**
  - El modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i$ para i=1, 2, 3,…, n, con *p* parámetros ó (p-1) variables de explicación, se lo puede representar matricialmente de la siguiente manera:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2p-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \Rightarrow \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

…viene Modelos de Regresión

- Donde:
  - El vector de observaciones $\mathbf{Y} \in R^n$
  - La matriz de diseño $\mathbf{X} \in M_{nxp}$
  - El vector de parámetros $\boldsymbol{\beta} \in R^p$
  - El vector de errores $\boldsymbol{\varepsilon} \in R^n$
- Además hay tener en cuenta que:
  - $E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$ ya que $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
  - La Matriz de Varianzas y Covarianzas del Error es: $\Sigma_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbf{I}$

Estimación de los Parámetros

- De acuerdo con el modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ y las condiciones previamente mencionadas, se tiene el vector $\boldsymbol{\beta}$ y $\Sigma_{\varepsilon} = \sigma^2\mathbf{I}$ son parámetros desconocidos pero estadísticamente estimables.
- Como métodos de estimación de parámetros se identifican: Mínimos Cuadrados y Máxima Verosimilitud.

…viene Estimación de los Parámetros

- **Estimación por Mínimos Cuadrados**

  Este es un método de ajuste de curvas que a principios del siglo XIX sugirió el matemático francés Adrien Legendre.

$$S(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum (y_i - \mu_i)^2$$

$$= \sum \left(y_i - \beta_0 - \beta_1 x_{i\,1} - \ldots - \beta_{p-1} x_{i\,p-1} \right)^2$$

…viene Estimación de los Parámetros

- Aplicando el criterio de las derivadas

$$\frac{\partial S(\beta_0,\beta_1,\ldots,\beta_{p-1})}{\partial \beta_0}=0$$

$$\frac{\partial S(\beta_0,\beta_1,\ldots,\beta_{p-1})}{\partial \beta_1}=0$$

$$\vdots$$

$$\frac{\partial S(\beta_0,\beta_1,\ldots,\beta_{p-1})}{\partial \beta_{p-1}}=0$$

$$X^T Y = (X^T X)^{-1} b$$

$$b = (X^T X)^{-1} X^T Y$$

…viene Estimación de los Parámetros

- **Estimación por Máxima Verosimilitud**

  Este método se basa, en la distribución del error. De acuerdo a líneas previas se dijo que el error tiene distribución Normal, por lo que la distribución de Y_i es también Normal:

$$Y_i \sim N\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_1 x_{ip-1}, \sigma^2\right)$$

…viene Estimación de los Parámetros

- La expresión de la función de densidad conjunta para el vector $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ es la siguiente:

$$f\left(\mathbf{Y}\right) = f\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \prod_{i=1}^{n} f(y_i) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \beta_1 x_i + \ldots + \beta_{p-1} x_{i\,p-1}\right)\right)^2}$$

…viene Estimación de los Parámetros

- Basados a la expresión anterior se tiene que la función de verosimilitud en forma matricial y en termino de los parámetros $\boldsymbol{\beta}, \sigma^2$ es la siguiente:

$$L\left(\mathbf{Y};\boldsymbol{\beta},\sigma\right)=\frac{1}{\sqrt{2\pi\boldsymbol{\sigma}}}\exp^{-\frac{1}{2\boldsymbol{\sigma}^2}\left(\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)^{T}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)\right)}$$

…viene Estimación de los Parámetros

- Por lo que los betas por estimación de máxima verosimilitud se los define como sigue:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix}$$

- Cuya matriz de varianzas y covarianzas es:

$$\sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \Sigma_{\mathbf{b}}$$

Matriz "HAT"

- La "*Matriz Hat*", "*H*", relaciona los valores ajustados con los valores observados , lo cual indica la influencia que cada valor observado tiene sobre cada valor ajustado.
- Pues bien, suponiendo un modelo de regresión lineal, se tiene que:

$$\hat{Y} = Xb \Rightarrow \hat{Y} = X(X^T X)^{-1} X^T Y \Rightarrow H = X\left(X^T X\right)^{-1} X^T$$

$$\hat{Y} = HY$$

Análisis de Varianza

- Tabla Anova

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMAS CUADRÁTICAS	MEDIAS CUADRÁTICAS	F
Regresión	p-1	$SCR=\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MCR=SCR/p-1	$\frac{MCR}{MCE}$
Error	n-p	$SCE=\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MCE=SCE/n-p	
Total	n-1	$SCT=\sum_{i=1}^{n}(y_i - \bar{y})^2$		

- En vista de que $F=\frac{MCR}{MCE}$ tiene distribución $F_{(p-1,\,n-p)}$, con $(1-\alpha)100\%$ de confianza se debe rechazar H_0 a favor de H_1, si el estadístico F_0 es mayor que el percentil $(1-\alpha)100$ de $F(\nu_1,\nu_2)$ con $\nu_1=(p-1)$ grados de libertad en el numerador y $\nu_2=(n-p)$ grados de libertad en el denominador.

Análisis de Varianza

- Tabla Anova en forma Matricial:

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMAS CUADRÁTICAS	MEDIAS CUADRÁTICAS	F
Regresión	p-1	$SCR = \mathbf{y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{y}$	MCR=SCR/p-1	$F_o = \frac{\text{MCR}}{\text{MCE}}$
Error	n-p	$SCE = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$	MCE=SCE/n-p	
Total	n-1	$SCT = \mathbf{y}'\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\,\mathbf{y}$		

Selección de variables de predicción

- Se supone que el número de variables explicativas que pueden haber en el modelo es (p -1), el número de observaciones es n; y, si se ajusta un modelo de regresión lineal con estas variables explicativas, el número de parámetros del modelo es p. Entonces se definen las siguientes medidas de bondad de ajuste:

...viene Selección de variables de predicción

- *Coeficiente de Determinación (R^2)*
- *R^2-Ajustado*
- *Varianza Residual (S_R^2)*
- *Estadístico de Mallows*
- *Criterio de Información de Akaike (AIC)*
- *Suma de Cuadrados de Predicción (PRESS)*

...viene Selección de variables de predicción

- *Coeficiente de Determinación (R^2)*

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

- *R2-Ajustado*

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

...viene Selección de variables de predicción

- R^2_{adj} en términos del Coeficiente de Determinación R^2

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

- Dicha expresión en términos de varianzas se tiene que:

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

...viene Selección de variables de predicción

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 este indicador es una mejor medida que R^2 para la selección del modelo

...viene Selección de variables de predicción

- ***Varianza Residual*** (s_R^2)

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

El criterio de minimizar la varianza residual es equivalente al criterio de maximizar el coeficiente de determinación ajustado.

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.

...viene Selección de variables de predicción

- ***Estadístico de Mallows***

  Este criterio toma en cuenta la Media Cuadrática del Error, 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 $E\left(\hat{\beta}_i\right) \neq \hat{\beta}_i$ lo cual indica que el objetivo de este indicador es minimizar la MCE.

...viene Selección de variables de predicción

- ***Estadístico de Mallows***

  C_P de Mallows está definido como:

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

El valor en el que el C_p es el mejor es cuando este se aproxima al número de parámetros.

...viene Selección de variables de predicción

- ***Criterio de Información Akaike (AIC)***

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

- Este criterio es similar al C_p una medida de bondad de ajuste, pero el AIC considera la función verosimilitud.
- Seleccionamos el modelo que tenga el menor valor de AIC.

...viene Selección de variables de predicción

- ***Suma de Cuadrados de Predicción (PRESS)***
  - 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*".

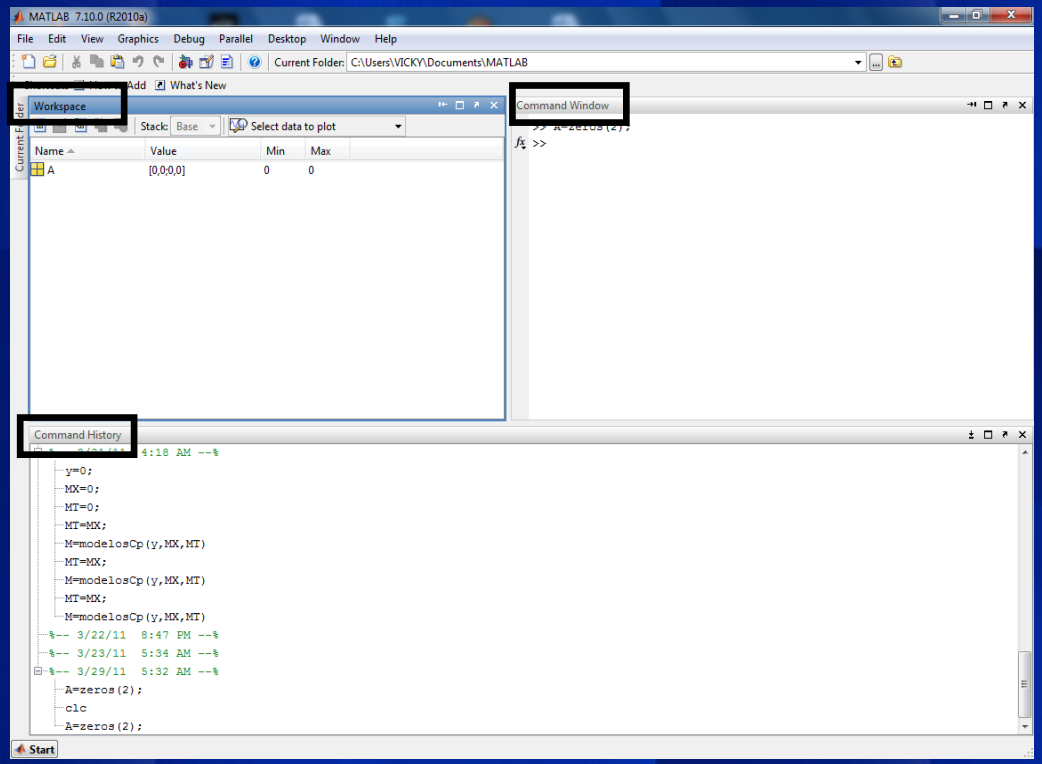
$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 \quad \Rightarrow \quad \text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

Acerca de ERLA

- 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.

Acerca de ERLA

- ***MATLAB***(Laboratorio de Matrices)



Command Window.- Es la ventana de comandos para interactuar.

Command History.- Contiene el registro de los comandos que han sido ingresados.

Workspace.-Contiene la descripción de las variables usadas en cada sección.

Acerca de ERLA

- Se presenta el algoritmo utilizado para construir la Función "Regresión Lineal" :

```
function R1=RegressionCoefficients(y,MX)
%El primer argumento debe ser la variable
a ser explicada
%El segundo argumento debe ser la matriz
con variables de explicación
%Devuelve una matriz con las inferencias
sobre los betas
paramat long g;
d=size(MX);
n=d(1);
p=d(2)+1;
j=ones(n,1);
X=[j,MX];
I=eye(n);
J=ones(n);
A=inv(X'*X);
H=X*A*X';
SCE=y'*(I-H)*y;
MCE=SCE/(n-p);
b=A*X'*y;
Sb=MCE*A;
R1=zeros(p,4);
para i=1:p
   R1(i,1)=b(i);
   R1(i,2)=sqrt(Sb(i,i));
   R1(i,3)=R1(i,1)/R1(i,2);
   R1(i,4)=abs(R1(i,3));
   R1(i,4)=tcdf(R1(i,4),n-p);
   R1(i,4)=(1-R1(i,4))*2;
 fin
```

Acerca de ERLA

- Se presenta el algoritmo utilizado para el calculo de los indicadores de calidad del modelo :

```
función M=modelosR2(y,MX)
t1=size(MX);
v=t1(2);
SCT=R2Ajustado2_SCT(y,MX);
para i=1:v
   c(i)=nchoosek(v,i);
fin
p=1;
i=1;
k=c(1);
t=0;
si v==1
   M(t+1)=R2 Ajustado2(y,MX,SCT);
   M=M';
Si no
   mientras i<v
cc=1;
      vr=combinacion(v,i,'c');
      para j=p:k
         M(j)=R2 Ajustado2(y,MX(:,vr(cc,:)),SCT);
         t=j;
         cc=cc+1;
      fin
   p=t+1;
   i=i+1;
   k=t+c(i);
   fin
   vr=combinator(v,v,'c');
   M(t+1)=R2 Ajustado2(y,MX,SCT);
   M=M';
Fin
```

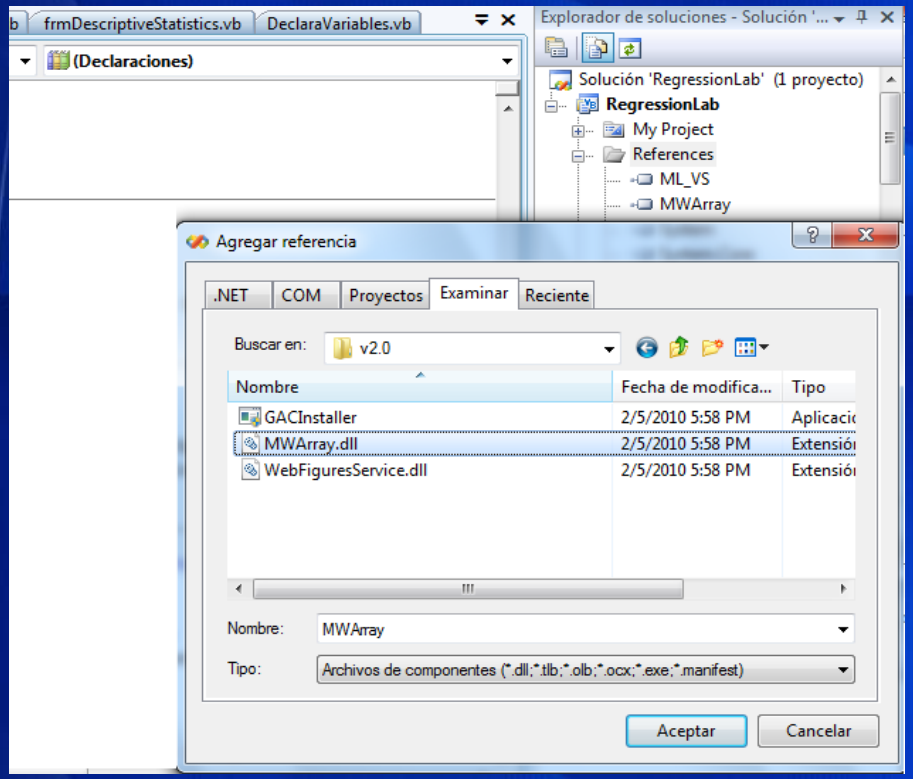
Acerca de ERLA

- ***Conexión entre VISUAL BASIC.NET y MATLAB***

  La conexión entre estos dos programas comienza en Matlab con la creación de las librerías respectivas, ya que ésta es la base para la creación de las funciones que proporcionaran los resultados esperados. Para ello inicialmente se crean funciones, se comprueba los resultados de las funciones creadas para luego crear librerías (archivos *.dll).

Acerca de ERLA

Ya desde Visual Basic.NET, se añade una referencia hacia la librería principal de Matlab MWArray.dll, para con esto poder acceder a las funciones creadas en Matlab convertidas en librerías.



Acerca de ERLA

- El proyecto desarrollado en Visual Studio.NET se lo compila para luego poder tener un archivo ejecutable (*.exe), con el cual este software podrá ser instalado en sistemas operativos Windows.

Validación del Modelo en el Software ERLA

- Se considera el caso de una "Central Eléctrica". Las variables que se consideran son:
  - ***C:*** Costo en dólares
  - ***D:*** Fecha de expedición permiso de construcción
  - ***T1:*** Tiempo entre la solicitud de permiso y la expedición o permiso
  - ***T2:*** Tiempo entre la emisión de la licencia de funcionamiento y permiso de construcción
  - ***S***: Capacidad de Energía neta de la planta
  - ***PR:*** Existencia previa de un reactor en el mismo sitio.
  - ***NE:*** Planta construida en la región noreste
  - ***CT:*** Uso de la torre de enfriamiento
  - ***BW:*** Sistema de suministro de vapor nuclear
  - ***N:*** Número acumulado de plantas de energía
  - ***PT:*** Llave de plantas

…viene Validación del Modelo en el Software ERLA

- De acuerdo con la ejecución de ERLA, basados en el ejemplo antes mencionado se determinó el valor del R^2 Ajustado, C_p Mallows, Akaike y PRESS de las 1024 combinaciones de las 10 variables de explicación (11 parámetros).

…viene Validación del Modelo en el Software ERLA

- Resultados:

# Parámetros	R^2 Ajustado	C_p Mallows	AIC	PRESS	# Variables Explicativas
2	0.4364	55.91	-78.68	4.38	**1**
3	0.6314	27.04	-91.36	2.76	**2**
4	0.7326	13.16	-100.75	1.81	**3**
5	0.7814	7.29	-106.36	1.60	**4**
6	0.7980	6.05	-108.10	1.60	**5**
7	0.8068	5.97	-108.77	1.67	**6**
8	0.8065	7.04	-108.03	1.75	**7**
9	0.8149	8.49	-108.81	1.91	**8**
10	0.8072	9.05	-106.93	2.05	**9**
11	0.7985	11.00	-105.014	2.32	**10**

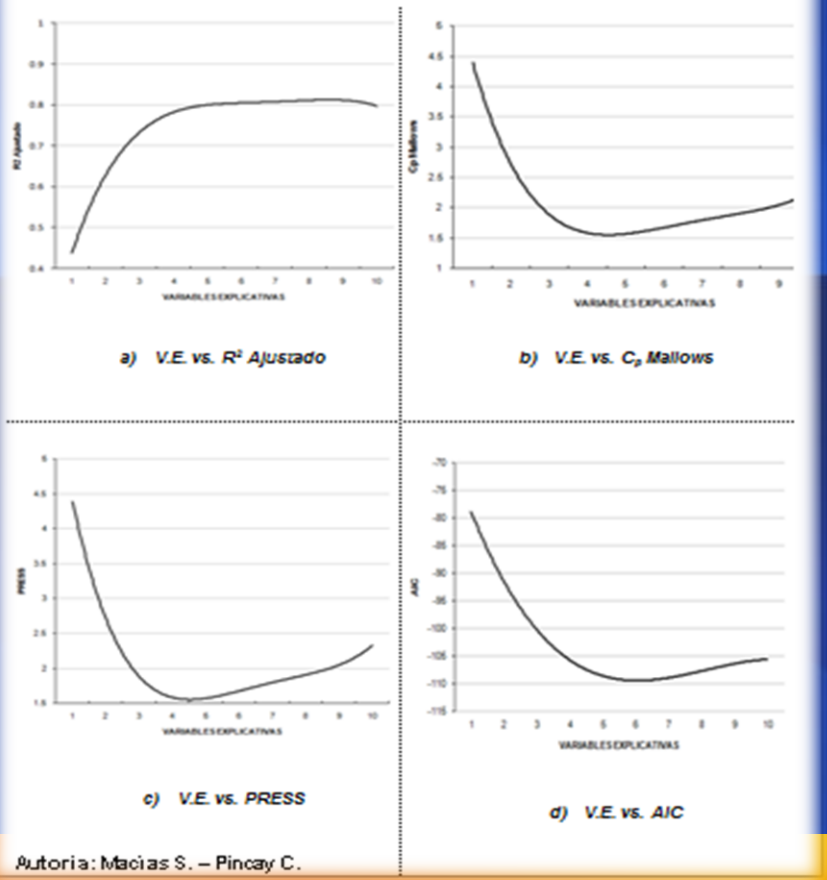
…viene Validación del Modelo en el Software ERLA

- Resultados:
  - **R^2 Ajustado: 8 V.E. (**0.8149)
  - **C_p Mallows: 5 V.E. (**6.0500)
  - **AIC: 8 V.E. (**-108.81)
  - **PRESS: 4 V.E. (1.6000)**

$$C = -11.68 + 0.24\,D + 0.006\,T_2 + 0.001\,S$$
$$- 0.11\,PR + 0.26\,NE + 0.11\,CT - 0.01\,N - 0.21\,PT$$

…viene Validación del Modelo en el Software ERLA

Figura 14: Graficas de Tendencia de los Indicadores de Selección de Modelos: R^2 Ajustado, C_p Mallows, Akaike y PRESS.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



a) V.E. vs. R^2 Ajustado

b) V.E. vs. C_p Mallows

c) V.E. vs. PRESS

d) V.E. vs. AIC

Autoria: Macias S. – Pincay C.

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:

CONCLUSIONES

- 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.

- 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.

CONCLUSIONES

- 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 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”.

CONCLUSIONES

- 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.

RECOMENDACIONES

- 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.