



# Curso en Análisis de regresión

Unidad 3 • Selección de variables y construcción del modelo



**LOS LIBERTADORES**  
FUNDACIÓN UNIVERSITARIA

Autor:  
**Juan Carlos Rubrice Cárdenas**

© 2017. Todos los derechos reservados



## Preguntas orientadoras

- ¿Existe alguna prueba para medir la significancia de un subconjunto de variables predictoras en un modelo de regresión lineal múltiple?
- ¿Cuáles son los criterios adecuados para seleccionar un conjunto de variables predictoras para construir el modelo?
- Cuando una aplicación envuelve un conjunto de datos con muchas variables predictoras, ¿cómo podemos seleccionar las más significativas para tener un modelo más simplificado?

## Síntesis

En esta unidad se estudia el problema de selección de variables para un modelo de regresión lineal múltiple. Se presentan algunos criterios para hacer una elección apropiada de las variables predictoras, a saber el criterio del coeficiente de determinación múltiple  $R^2_p$ , el criterio de determinación múltiple ajustado  $R^2_{adj,p}$ , el criterio  $C_p$  de Mallows, el criterio  $AIC_p$  de Akaike y el criterio bayesiano de Schwarz  $SBC_p$ . Finalmente, se consideran tres procedimientos iterativos para seleccionar las variables predictoras que formaran el “mejor” modelo de regresión lineal, conocidos como: la regresión por pasos, la selección hacia adelante y la selección hacia atrás. Estos procedimientos, están basados en una prueba F.

## Palabras clave

- Selección de variables
- Criterio del coeficiente de determinación múltiple  $R^2_p$
- Criterio de determinación múltiple ajustado  $R^2_{adj,p}$
- Criterio  $C_p$  de Mallows
- Criterio  $AIC_p$  de Akaike
- Criterio bayesiano de Schwarz  $SBC_p$
- Regresión por pasos
- Selección hacia adelante
- Selección hacia atrás





## Tema 1

### El problema de la construcción del modelo

En esta sección vamos a abordar el problema de seleccionar de un conjunto de variables predictoras candidatas, un subconjunto de ellas que se consideran adecuadas para construir el modelo de tal manera que hayan suficientes variables predictoras que expliquen significativamente la variable respuesta e incluir la menor cantidad posible de ellas para evitar que la varianza de la respuesta estimada aumente considerablemente.

Encontrar un equilibrio entre estos dos criterios opuestos es la directriz de algunos procedimientos de selección del “mejor” modelo de regresión.

Si hay  $k$  variables predictoras que se pueden incluir en el modelo:  $x_1, x_2, \dots, x_k$ , entonces hay  $2^k$  modelos diferentes que abarcan todas las selecciones posibles de subconjuntos de variables predictoras incluyendo el modelo sin variables predictoras, por ejemplo, para un modelo con tres variables predictoras  $x_1, x_2$  y  $x_3$  se pueden generar  $2^3 = 8$  modelos lineales a saber, los modelos que incluyen 0 variables: {}, 1 sola variable: {  $x_1$  }, {  $x_2$  }, {  $x_3$  }, 2 variables: {  $x_1, x_2$  }, {  $x_1, x_3$  }, {  $x_2, x_3$  }, y el modelo completo de que incluye las 3 variables: {  $x_1, x_2, x_3$  }.

Vamos a considerar algunos procedimientos para seleccionar variables que identifican una colección reducida de modelos que son apropiados de acuerdo a un criterio especificado. Luego, se realiza un análisis detallado de estos modelos para seleccionar el modelo de regresión final que se emplearía. Nos centraremos en cinco procedimientos usados para comparar modelos de regresión.

## Tema 2

### Criterio $R^2 p$ o $SCE_p$

Sean  $p - 1$  el número de variables que se consideran en un modelo reducido con  $p$  parámetros. La idea de este criterio es usar el coeficiente de determinación múltiple  $R^2$  para identificar varios subconjuntos de variables predictores que tienen un  $R^2$  alto. La notación  $R^2_p$  denota el cociente de determinación para un modelo que tiene  $p$  parámetros y  $p - 1$  variables predictoras. Como criterio equivalente, se puede considerar que un





subconjunto con  $p$  variables predictoras es "bueno" si su suma de cuadrados del error, denotada  $SCE_p$ , es pequeña. La relación entre estos dos criterios se hace evidente por la siguiente ecuación:

$$R_p^2 = 1 - \frac{SCE_p}{SCTO},$$

debido a que  $R_p^2$  varía inversamente respecto a  $ESCP$ , ya que la suma total de cuadrados denotada  $SCTO$  es constante independientemente del número de variables involucradas en el modelo.

En vista de que el coeficiente de determinación  $R^2$  aumenta a medida que el número de variables predictoras aumenta en el modelo, la intención de este criterio es encontrar el punto donde agregar más variables al modelo no vale la pena en el sentido de que el incremento en el  $R^2$  no es significativo.

## Tema 3

### Criterio $R^2_{adj,p}$ o $CME_p$

En vista que el coeficiente de determinación crece a medida que aumenta el número de parámetros  $p$  en el modelo, se sugiere utilizar el coeficiente de determinación ajustado que evita sobreestimar el efecto inflacionario que tiene sobre  $R^2$  el aumentar las variables predictoras y que se define como sigue:

$$R_{adj,p}^2 = 1 - \frac{SCE_p/(n-p)}{SCT/(n-1)} = 1 - \frac{CME_p}{SCT/n-1}$$

Observe que el coeficiente de determinación  $R_{adj,p}$  crece si y solo si el cuadrado medio del error  $CME_p$  aumenta. Esto se debe a que la suma de cuadrados total solo depende de los valores  $y$ . Así, de todos los modelos que contienen  $p$  parámetros se considera que el "mejor" es el que tenga el máximo coeficiente de determinación ajustado. Lo interesante es que este máximo no necesariamente crece cuando el número de parámetros aumenta.





Los análisis usualmente eligen el subconjunto de variables que tengan el máximo coeficiente de determinación ajustado o que este coeficiente esté cerca al máximo, que la adición de mas variables predictoras no vale la pena.

## Tema 4

### Criterio $C_p$ de Mallows

Este criterio está basado en el error de estimación cuadrático total estandarizado, es decir:

$$\Gamma_p = \frac{1}{\sigma^2} E \left( \sum_{i=1}^n [\hat{y}_i - E(y_i)]^2 \right) = \frac{E[SCE_p]}{\sigma^2} - n + 2p$$

donde  $\hat{y}_i$  es la respuesta estimada del modelo de subconjunto de  $p$  parámetros,  $SCE_p$  es la suma de cuadrados del error de este modelo, y  $E(y_i)$  es el valor esperado de la respuesta para el modelo correcto. Como  $E(SCE_p)$  y  $\sigma^2$  son desconocidos, podemos estimarlos usando los valores observados mediante los estadísticos  $SCE_p$  y  $\hat{\sigma}^2$ , el último obtenido del modelo completo. Haciendo estas estimaciones el criterio de medida anterior se transforma en:

$$C_p = \frac{SCE_p}{\hat{\sigma}^2} - n + 2p$$

Se considera que un modelo de subconjunto de  $p$  parámetros es “bueno” si su medida  $C_p$  es mínima.

## Tema 5

### Criterio $AIC_p$ y $SBC_p$

Vamos a considerar el criterio de información de Akaike ( $AIC_p$ ) y el criterio bayesiano de Schwarz ( $SBC_p$ ) que penalizan los modelos con un número grande de variables predictoras.





La idea central es buscar modelos que tengan valores pequeños de  $AIC_p$  y  $SBC_p$ , donde estos criterios se definen como:

$$AIC_p = n \ln(SCE_p) - n \ln(n) + 2p$$
$$SBC_p = n \ln(SCE_p) - n \ln(n) + [\ln(n)]p$$

### Ejemplo

La tabla muestra las medidas de los criterios anteriores para los diferentes modelos con p parámetros para investigar la relación de dependencia entre el nivel de Satisfacción (Sat) de los pacientes de un hospital y las variables explicativas: Edad (E), Gravedad (G), Tipo de paciente (TP) y Ansiedad (A).

p	R <sub>p</sub> <sup>2</sup>	R <sub>adj,p</sub> <sup>2</sup>	C <sub>p</sub>	AIC <sub>p</sub>	SBC <sub>p</sub>	E	G	TP	A
1	0	0	184.484	153.6602	154.8791				
2	0.8124	0.8043	17.91603	113.819	116.2568	X			
2	0.5227	0.502	78.02193	137.1674	139.6051		X		
2	0.05473	0.01363	175.1286	154.2531	156.6909			X	
2	0.2876	0.2567	126.804	147.181	149.6188				X
3	0.8966	0.8872	2.455308	100.9332	104.5898	X	X		
3	0.8126	0.7956	19.8816	115.7969	119.4535	X		X	
3	0.8133	0.7964	19.72713	115.6974	119.354	X			X
3	0.5343	0.4919	77.63292	138.5568	142.2135		X	X	
3	0.5795	0.5412	68.25411	136.0045	139.6611		X		X
3	0.3187	0.2567	122.3625	148.067	151.7237			X	X
4	0.8966	0.8818	4.452171	102.9295	107.805	X	X	X	
4	0.9035	0.8897	3.019026	101.2009	106.0764	X	X		X
4	0.8135	0.7868	21.70115	117.6806	122.5561	X		X	X
4	0.5893	0.5306	68.21347	137.4129	142.2884		X	X	X
5	0.9036	0.8843	5	103.1772	109.2715	X	X	X	X

Tabla 1. Criterios

Las líneas de color azul en la tabla muestran las variables que se deben incluir en el mejor modelo con p parámetros, por ejemplo el mejor modelo con tres variables (p = 4) es el que incluya a las variables predictoras: Edad (E), Gravedad (G) y Ansiedad (A).





## Tema 6

### Procedimientos de selección de variables

En esta sección vamos a estudiar tres procedimientos para seleccionar las variables predictoras que formarán el "mejor" modelo de regresión lineal, a saber: la regresión por pasos, la selección hacia adelante y la selección hacia atrás. Estos procedimientos están basados en una prueba F que se describe a continuación. Considera un modelo de regresión lineal múltiple reducido que contiene las j variables predictoras:  $X_1, X_2, \dots, X_j$ , es decir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j + \varepsilon \quad (1)$$

Si al modelo anterior le agregamos las variables predictoras  $X_{j+1}, X_{j+2}, \dots, X_k$ , donde  $j < k$ , obtenemos el modelo completo de k variables predictoras:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j + \beta_{j+1} X_{j+1} + \beta_{j+2} X_{j+2} + \dots + \beta_k X_k + \varepsilon$$

Surge la cuestión de si la adición de las  $k - j$  variables  $X_{j+1}, X_{j+2}, \dots, X_k$  al modelo (1) es estadísticamente significativa. Para ello, se deben probar las siguientes hipótesis estadísticas:

$$H_0: \beta_{j+1} = \beta_{j+2} = \dots = \beta_k = 0$$
$$H_1: \text{Uno o más de estos parámetros no es igual a cero}$$

El estadístico que se utiliza para realizar esta prueba es:

$$F = \frac{\frac{SCE(X_1, X_2, \dots, X_j) - SCE(X_1, X_2, \dots, X_j, X_{j+1}, X_{j+2}, \dots, X_k)}{k-j}}{\frac{SCE(X_1, X_2, \dots, X_j, X_{j+1}, X_{j+2}, \dots, X_k)}{n-k-1}}$$

Este estadístico tiene una distribución de probabilidad F con  $k - j$  y  $n - k - 1$  grados de libertad en el numerador y en el denominador, respectivamente. Se rechaza  $H_0$  si  $F > F_\alpha$  o si valor  $-p < \alpha$ .

#### Ejemplo





Considere el modelo reducido que incluye las variables predictoras: Edad (E) y Gravedad (G) para el problema de la satisfacción de pacientes. Veamos si incluir las variables Tipo de paciente (TP) y Ansiedad (A) es estadísticamente significativa para explicar la variable respuesta Satisfacción (S).

Debemos probar las siguientes hipótesis:

$$H_0 : \beta_{TP} = \beta_A = 0$$

$H_1$  : Uno o más de estos parámetros no es igual a cero

Para determinar el valor del estadístico F, primero calculamos las sumas de cuadrados del error:

$$SCE(\text{Edad; Gravedad}) = 1114,546$$

$$SCE(\text{Edad; Gravedad; Tipo - paciente; Ansiedad}) = 1038,947$$

Así, el valor del estadístico de prueba F es:

$$F = \frac{(1114,546 - 1038,947)/(4 - 2)}{1038,947/(25 - 4 - 1)} = 0,7276502$$

El valor-p asociado es 0,4953977. Por lo tanto, no se puede rechazar la hipótesis nula a un nivel de significancia del 5%, y no es estadísticamente significativa la inclusión de las variables Tipo de paciente (TP) y Ansiedad (A) en el modelo.

## Tema 7

### Regresión por pasos

Este es un procedimiento iterativo que consiste en determinar paso a paso si alguna de las variables que ya están incluidas en un modelo inicial dado debe ser eliminada de este, usando la prueba F para cada uno de las variables predictoras que intervienen en el modelo para un nivel de significancia  $\alpha$  dado. Una vez que se eliminan todas las variables





que no pasen la prueba  $F$ , se analiza si se puede incluir alguna de las variables predictoras que no estén incluidas en el modelo y se agrega la variable que tenga el mínimo valor-p menor que  $\alpha$ . Luego, se vuelve a revisar si alguna de las variables del modelo se puede eliminar, de lo contrario se analiza si alguna de las variables externas al modelo se puede incluir y así sucesivamente. El procedimiento finaliza cuando ninguna variable predictora interna puede ser eliminada del modelo y ninguna variable predictora externa pueda ser incluida en este.

**Advertencia:** debido a su naturaleza este procedimiento puede generar un modelo cuyo coeficiente de determinación ajustado no sea mínimo.

### Ejemplo

Vamos a aplicar el procedimiento de regresión paso a paso para el problema de regresión múltiple con los datos de los pacientes de un hospital bajo consideración. Iniciemos considerando el modelo lineal múltiple con las variables predictoras  $E$  (Edad) y  $G$  (Gravedad), es decir:

$$Y = \beta_0 + \beta_E E + \beta_G G + \varepsilon$$

Probemos si la variable  $G$  debe ser o no excluida del modelo inicial para un nivel de significancia  $\alpha = 0,01$ , es decir:

$$\begin{aligned} H_0: \beta_G &= 0 \\ H_1: \beta_G &\neq 0 \end{aligned}$$

Aplicando la prueba  $F$ , obtenemos:

$$\begin{aligned} F &= \frac{[SCE(E) - SCE(E,G)]/1}{SCE(E,G)/(25-2-1)} \\ &= \frac{(2021,584 - 1114,546)/1}{1114,546/22} \\ &= 17,904 \end{aligned}$$

Este valor observado de  $F$  tiene un valor-p =  $3,43 \times 10^{-4}$  que es menor al nivel de significancia  $\alpha = 0,01$  dado. Por lo tanto, no se excluye la variable  $G$  del modelo inicial.





Ahora, probemos si la variable  $E$  debe ser excluida del modelo, es decir:

$$H_0: \beta_E = 0$$
$$H_1: \beta_E \neq 0$$

Aplicando la prueba  $F$  obtenemos:

$$F = \frac{[SCE(G) - SCE(E,G)]/1}{SCE(E,G)/(25 - 2 - 1)}$$
$$= \frac{(5143,925 - 1114,546)/1}{1114,546/22}$$
$$= 79,53583$$

Este valor observado de  $F$  tiene un valor-p =  $9,284962 \times 10^{-9}$  que es menor al nivel de significancia  $\alpha = 0,01$  dado. Por lo tanto, no se excluye la variable  $E$  del modelo inicial.

En esta primera parte del procedimiento, concluimos que las variables  $E$  y  $G$  no pueden ser excluidas del modelo inicial. Ahora, vamos a ver si alguna de las variables predictoras que no han sido incluidas en este modelo se pueden incorporar a este, a saber: Tipo de paciente ( $TP$ ) y Ansiedad ( $A$ ). Para ello, se plantean las hipótesis:

$$H_0: \beta_{TP} = 0$$
$$H_1: \beta_{TP} \neq 0$$

y

$$H_0: \beta_A = 0$$
$$H_1: \beta_A \neq 0$$

Aplicando la prueba  $F$  para probar estas hipótesis, obtenemos:

$$F = \frac{[SCE(E,G) - SCE(E,G,TP)]/1}{SCE(E,G,TP)/(25 - 3 - 1)}$$
$$= \frac{(1114,546 - 1114,383)/1}{1114,383/21}$$
$$= 0,003071655$$

y

$$F = \frac{[SCE(E,G) - SCE(E,G,A)]/1}{SCE(E,G,A)/(25 - 3 - 1)}$$
$$= \frac{(1114,546 - 1039,935)/1}{1039,935/21}$$
$$= 1,506662$$

Ahora, los valores-p asociados a estos valores  $F$  observados son 0.9563258 y 0.2332316, respectivamente. Por lo tanto, ni la variable  $TP$  ni la variable  $A$  se incluyen en el modelo inicial a un nivel de significancia  $\alpha = 0,01$ . El procedimiento de regresión por pasos en este





caso nos dice que el “mejor” modelo lineal múltiple es el que incluye las variables predictoras Edad ( $E$ ) y Gravedad ( $G$ ).

## Tema 8

### Selección hacia adelante

En este procedimiento se inicia sin ninguna variable predictora y se van incluyendo variables predictoras al modelo de una en una, usando el procedimiento de regresión por pasos para determinar si estas deben ser incluidas al modelo. Una vez una variable se incluya en el modelo, esta no puede ser eliminada en pasos posteriores. Este procedimiento culmina cuando el valor-p asociado a las variables predictoras que no están en el modelo sea mayor que el valor de significancia fijado para la prueba  $F$ .

#### Ejemplo

Vamos a aplicar la selección de variables predictoras hacia adelante para el problema de modelar linealmente la satisfacción de los pacientes de un hospital en términos de las posibles variables predictoras: Edad ( $E$ ), Gravedad ( $G$ ), Tipo de paciente ( $TP$ ) y Ansiedad ( $A$ ).

#### Paso 1

Iniciamos con un modelo sin variables predictoras de un solo parámetro  $\beta_0$  cuya ecuación de regresión estimada es  $\hat{y} = y$ . Ahora, veamos si podemos incluir a este modelo la variable  $E$  (Edad) usando la prueba  $F$  para las hipótesis:

$$\begin{aligned} H_0: \beta_E &= 0 \\ H_1: \beta_E &\neq 0 \end{aligned}$$

El valor observado del estadístico de prueba  $F$  es:





$$\begin{aligned} F &= \frac{[SCE(\text{modelo sin predictoras}) - SCE(E)]/1}{SCE(E)/(25-1-1)} \\ &= \frac{(10778,24 - 2021,584)/1}{2021,584/23} \\ &= 99,62638 \end{aligned}$$

El valor-p asociado a esta prueba es  $7,919317 \times 10^{-10}$ . Se concluye que la variable  $E$  es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y debe incluirse en el modelo.

Continuamos, analizando si la variable  $TP$  (Tipo de paciente) se debe incluir en el modelo aplicando la prueba  $F$  para las hipótesis:

$$\begin{aligned} H_0: \beta_{TP} &= 0 \\ H_1: \beta_{TP} &\neq 0 \end{aligned}$$

El valor observado del estadístico de prueba  $F$  es:

$$\begin{aligned} F &= \frac{[SCE(\text{modelo sin predictoras}) - SCE(TP)]/1}{SCE(TP)/(25-1-1)} \\ &= \frac{(10778,24 - 10188,36)/1}{10188,36/23} \\ &= 1,331641 \end{aligned}$$

El valor- $p$  asociado a esta prueba es  $0,260364$ . Se concluye que la variable  $TP$  no es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y en congruencia no se debe incorporar al modelo.

Ahora, se examina si la variable  $A$  (Ansiedad) se debe incluir en el modelo aplicando la prueba  $F$  para las hipótesis:

$$\begin{aligned} H_0: \beta_A &= 0 \\ H_1: \beta_A &\neq 0 \end{aligned}$$

El valor observado del estadístico de prueba  $F$  es:





$$\begin{aligned} F &= \frac{[SCE(\text{modelo sin predictoras}) - SCE(A)]/1}{SCE(A)/(25-1-1)} \\ &= \frac{(10778,24 - 7678,023)/1}{7678,023/23} \\ &= 9,286895 \end{aligned}$$

El valor-*p* asociado a esta prueba es 0,005715364. Se concluye que la variable *A* es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y en congruencia se debe incorporar al modelo.

Por lo tanto, se incluyen al modelo las variables predictoras *E* (Edad), *G* (Gravedad) y *A* (Ansiedad).

## Paso 2

En este paso, se prueba si la variable *TP* que no fue incluida respecto al modelo sin predictoras, ahora se puede incluir en relación al modelo con las variables predictoras incluidas en el paso anterior. Aquí, las hipótesis son:

$$\begin{aligned} H_0: \beta_{TP} &= 0 \\ H_1: \beta_{TP} &\neq 0 \end{aligned}$$

El valor observado del estadístico de prueba *F* es:

$$\begin{aligned} F &= \frac{[SCE(E, G, A) - SCE(E, G, TP, A)]/1}{SCE(E, G, TP, A)/(25-4-1)} \\ &= \frac{(1039,935 - 1038,947)/1}{1038,947/20} \\ &= 0,01901926 \end{aligned}$$

El valor-*p* asociado a esta prueba es 0,8916903. Se concluye que la variable *TP* no es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y en congruencia no se debe incorporar al modelo.





El procedimiento de selección hacia delante finaliza aquí, ya que no se puede agregar ninguna variable al modelo.

Concluimos que el “mejor” modelo de regresión lineal múltiple para este problema por el procedimiento selección hacia adelante es el que incluye a las variables predictoras:  $E$  (Edad),  $G$  (Gravedad) y  $A$  (Ansiedad).

## Tema 9

### Eliminación hacia atrás

En este procedimiento se inicia con un modelo que incluye todas las variables predictoras. Luego, de una en una, se van eliminando las variables predictoras que no pasen la prueba  $F$  usada en el método de regresión por pasos. No obstante, en este método no se permite que una variable que ya fue eliminada vuelva a ingresar al modelo. Este procedimiento finaliza cuando ninguna de las variables predictoras que están en el modelo puedan ser eliminadas del mismo.

#### Ejemplo

Vamos a aplicar el procedimiento de eliminación hacia atrás de variables predictoras para el problema de modelar linealmente la satisfacción de los pacientes de un hospital en términos de las posibles variables predictoras: Edad ( $E$ ), Gravedad ( $G$ ), Tipo de paciente ( $TP$ ) y Ansiedad ( $A$ ).

#### Paso 1

Iniciamos con un modelo completo que incluya todas las variables predictoras relacionadas. Probemos si alguna de las variables predictoras deben ser eliminadas examinándolas una a una. Empecemos con la variable  $E$  (edad) usando la prueba  $F$  para las hipótesis:

$$H_0: \beta_E = 0$$
$$H_1: \beta_E \neq 0$$

El valor observado del estadístico de prueba  $F$  es:





$$F = \frac{[SCE(G, TP, A) - SCE(E, G, TP, A)]/1}{SCE(E, G, TP, A)/(25-4-1)}$$
$$= \frac{(4426,612 - 1038,947)/1}{1038,947/20}$$
$$= 65,21343$$

El valor-p asociado a esta prueba es  $7,919317 \times 10^{-10}$ . Se concluye que la variable  $E$  es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y debe incluirse en el modelo.

Luego, veamos si podemos excluir la variable predictora  $G$  (Gravedad) aplicando la prueba  $F$  para las hipótesis:

$$H_0: \beta_G = 0$$
$$H_1: \beta_G \neq 0$$

El valor observado del estadístico de prueba  $F$  es:

$$F = \frac{[SCE(E, TP, A) - SCE(E, G, TP, A)]/1}{SCE(E, G, TP, A)/(25-4-1)}$$
$$= \frac{(2010,421 - 1038,947)/1}{1038,947/20}$$
$$= 18,70113$$

El valor-p asociado a esta prueba es  $0,0003294745$ . Se concluye que la variable  $G$  es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y por lo tanto no se puede eliminar de este.

Continuamos, analizando si la variable  $TP$  (Tipo de paciente) se debe eliminar del modelo aplicando la prueba  $F$  para las hipótesis:

$$H_0: \beta_{TP} = 0$$
$$H_1: \beta_{TP} \neq 0$$

El valor observado del estadístico de prueba  $F$  es:

$$F = \frac{[SCE(E, G, A) - SCE(E, G, TP, A)]/1}{SCE(E, G, TP, A)/(25-4-1)}$$
$$= \frac{(1039,935 - 1038,947)/1}{1038,947/20}$$
$$= 0,01901926$$





El valor-p asociado a esta prueba es 0,8916903. Se concluye que la variable *TP* no es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y en congruencia se debe eliminar del modelo.

Finalmente, se examina si la variable *A* (ansiedad) se debe eliminar del modelo aplicando la prueba *F* para las hipótesis:

$$\begin{aligned} H_0: \beta_A &= 0 \\ H_1: \beta_A &\neq 0 \end{aligned}$$

El valor observado del estadístico de prueba *F* es:

$$\begin{aligned} F &= \frac{[SCE(E, G, TP) - SCE(E, G, TP, A)]/1}{SCE(E, G, TP, A)/(25-4-1)} \\ &= \frac{(1114,383 - 1038,947)/1}{1038,947/20} \\ &= 1,452163 \end{aligned}$$

El valor-p asociado a esta prueba es 0,2422469. Se concluye que la variable *A* es significativa para el modelo a un nivel de significancia  $\alpha = 0,01$  y en congruencia se debe eliminar del modelo. Así, en este primer paso se eliminaron las variables *TP* (Tipo de paciente) y *A* (ansiedad).

## Paso 2

Probar si alguna de las variables que no fueron eliminadas en el paso anterior ahora pueden ser eliminadas en el modelo reducido. Empecemos con la variable *E* (Edad) usando la prueba *F* para las hipótesis:

$$\begin{aligned} H_0: \beta_E &= 0 \\ H_1: \beta_E &\neq 0 \end{aligned}$$

El valor observado del estadístico de prueba *F* es:

$$\begin{aligned} F &= \frac{[SCE(G) - SCE(E, G)]/1}{SCE(E, G)/(25-2-1)} \\ &= \frac{(5143,925 - 1114,546)/1}{1114,546/22} \\ &= 79,53583 \end{aligned}$$





El valor-*p* asociado a esta prueba es  $9,284962 \times 10^{-9}$ . Se concluye que la variable *E* es significativa para el modelo reducido a un nivel de significancia  $\alpha = 0,01$  y por lo tanto no se puede eliminar de este.

Ahora, veamos si debemos eliminar la variable *G* (Gravedad) del modelo reducido usando la prueba *F* para las hipótesis:

$$\begin{aligned} H_0: \beta_G &= 0 \\ H_1: \beta_G &\neq 0 \end{aligned}$$

El valor observado del estadístico de prueba *F* es:

$$\begin{aligned} F &= \frac{[SCE(E) - SCE(E,G)]/1}{SCE(E,G)/(25-2-1)} \\ &= \frac{(2021,548 - 1114,546)/1}{1114,546/22} \\ &= 17,904 \end{aligned}$$

El valor-*p* asociado a esta prueba es  $0,0003429368$ . Se concluye que la variable *G* es significativa para el modelo reducido a un nivel de significancia  $\alpha = 0,01$  y por lo tanto no se puede eliminar de este. El procedimiento de eliminación hacia atrás finaliza, ya que no se pudo eliminar ninguna variable del modelo reducido.

Concluimos que el "mejor" modelo de regresión lineal múltiple para este problema por el procedimiento de eliminación hacia atrás es el que incluye a las variables predictoras: *E* (Edad) y *G* (Gravedad).

En los dos ejemplos anteriores observamos que el procedimiento de selección hacia adelante y eliminación hacia atrás generan modelos de regresión diferentes. Así que es posible encontrar modelos diferentes usando estos dos procedimientos de selección de variables predictoras. En esta situación, el analista debe tomar una decisión respecto a cual modelo es preferible.





## Bibliografía

- Montgomery D., Peck E. y Vining G. (2006). Introducción al análisis de regresión lineal. Tercera edición. México: CECSA
- Kutner M., Nachtsheim C., Neter J. y Li W. (2005). Applied Linear Statistical Models. Quinta edición. New York: McGraw-Hill.
- Sheather S. (2009). A Modern Approach to Regression with R. New York: Springer.
- Weisberg S. (2005). Applied Linear Regression. Tercera edición. New Jersey: Wiley.