

UAB_AllSetsReg

All Possible Subsets Regression in Linear, Logistic and Cox Regression

Tabla de contenidos

Presentación	1
Limitaciones	1
Instalar el comando de extensión	2
Comprobación de la instalación	3
Desinstalar un comando de extensión	3
¿Cómo cribar variables para modelos predictivos?	4
Regresión lineal múltiple	5
Estimar el modelo seleccionado	7
Predictores fijos	8
Términos de interacción y modelos jerárquicos	9
Regresión logística	10
Regresión de Cox	12
Variables dependientes del tiempo (VDT)	13
Referencia de sintaxis de UAB_AllSetsReg	17

Presentación

La selección de la mejor ecuación de regresión con finalidad predictora se puede realizar con diferentes procedimientos. Uno de los más potentes consiste en establecer unos criterios de selección, como la C_p de Mallows, el R^2 ajustado, el AIC de Akaike, etc., construir todos los posibles submodelos combinando los términos del modelo máximo, y valorar para cada uno de ellos el grado de cumplimiento del criterio establecido.

El comando de extensión **UAB_AllSetsReg_ML** realiza la selección del mejor modelo predictivo para regresión lineal múltiple y logística y el **UAB_AllSetsReg_Cox** para regresión de Cox. Construyen en primer lugar todos los modelos con 1 término, luego todos los modelos con 2 términos y así sucesivamente hasta el modelo máximo que contiene todos los términos. Para cada subconjunto se presenta información de los diferentes criterios de selección. El resultado se presenta en una nueva ventana de datos en donde cada sujeto es uno de los posibles modelos y las variables contienen los valores de los diferentes criterios.

Limitaciones

El número máximo de términos admisible es 17; sin embargo no conviene definir modelos con más de 15 términos predictores para evitar colapsarlo. Con 15 términos se generan 32767 submodelos con un tiempo de ejecución que varía entre 10 y 15 minutos. Para modelos con muchos predictores es conveniente realizar un primer cribado y quedarse con los 15 mejores.

! La versión actual del comando de extensión (v0.0.8) ha sido comprobada con las versiones de SPSS 19.0, 20.0 y 21.0. Para SPSS 18.0.3 es necesario instalar el comando desde los archivos **UAB_ALLSETSREG_ML_SPSS18.spe** y **UAB_ALLSETSREG_COX_SPSS18.spe**. El comando de extensión no funciona en versiones anteriores a la 18.

Instalar el comando de extensión

Para poder utilizar comandos de extensión debe instalar en su ordenador, en primer lugar, los complementos de integración de Python y R en SPSS Statistics. Esta instalación la realizará sólo una vez y servirá para todos los comandos de extensión que posteriormente instale.

Para ayudarle en el proceso hemos creado un documento titulado *Guía de instalación de los complementos de integración de Python y R en SPSS Statistics*, que debe descargar del campus virtual, en concreto se encuentra en el enlace "Material", pestaña "SPSS Statistics", en el enlace "Guía de instalación de los complementos de integración de Python y R".

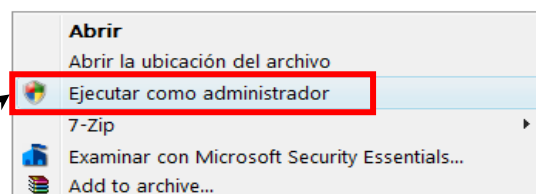
❗ **Para evitar problemas de instalación es fundamental que siga estrictamente los pasos que se indican en la guía de instalación.**

❗ **Recuerde que en SPSS 18 debe instalar desde los archivos *UAB_ALLSETSREG_ML_SPSS18.spe* y *UAB_ALLSETSREG_ML_SPSS18.spe*.**

Una vez instalados con éxito los complementos de integración de Python y R puede proceder a instalar los comandos de extensión. Las últimas páginas de la guía de instalación de Python y R explican cómo instalar, precisamente, el comando de extensión UAB_AllSetsReg_ML. Reproducimos a continuación esa misma explicación.

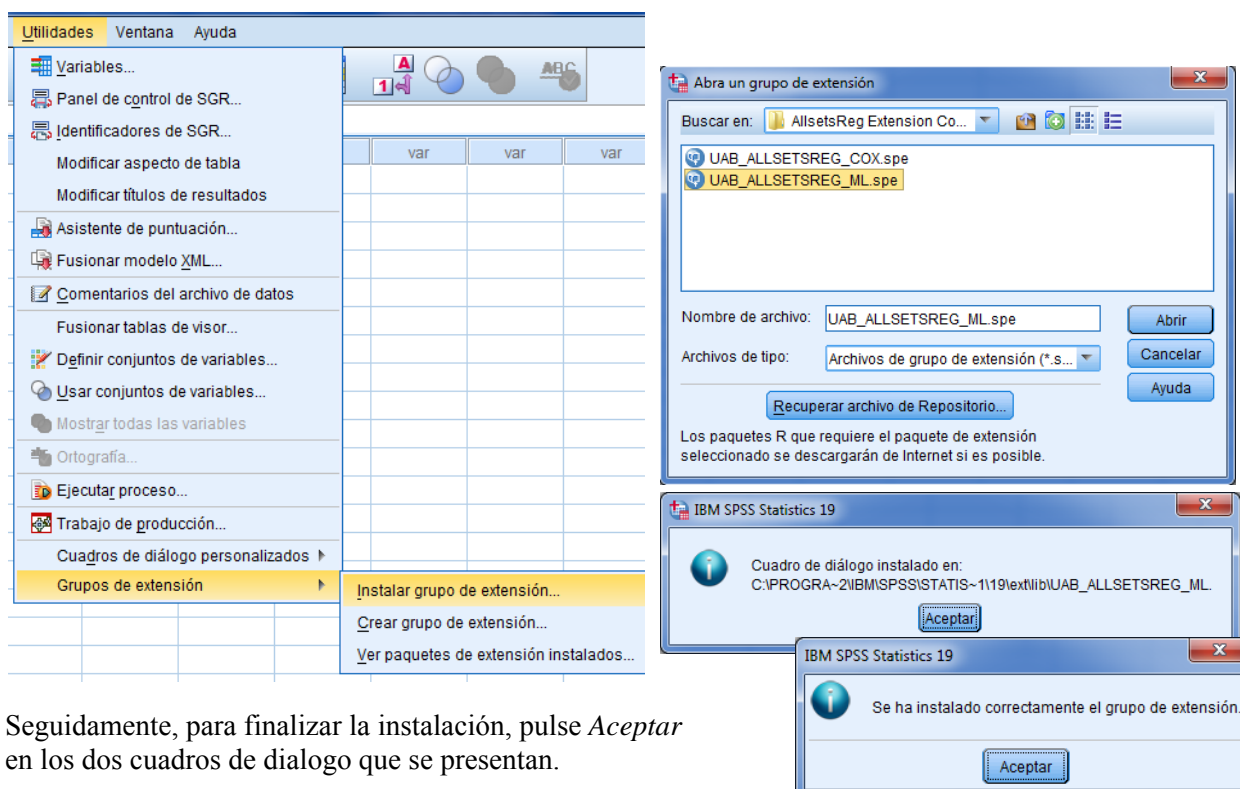
La instalación de un comando de extensión modifica la carpeta de instalación de SPSS y otros componentes protegidos del sistema operativo, por lo que **debe realizarse sobre una sesión de SPSS Statistics abierta con derechos de Administrador**. Además el ordenador **debe estar conectado a Internet** para poder los descargar paquetes de R que necesite el comando.

Para ello debe situarse sobre el icono de SPSS, hacer clic con el botón derecho y en el menú contextual escoger la opción *Ejecutar como administrador*.



Tras lanzar *SPSS Statistics* con derechos de Administrador puede instalar (o reinstalar) un comando de extensión con la opción de menú *Utilidades | Grupos de extensión | Instalar grupo de extensión...*

Las siguientes imágenes muestran la instalación del procedimiento de selección de la mejor ecuación con todos los subconjuntos para regresión lineal y logística (UAB_AllsetsReg_ML). Se abre una ventana para indicar donde se encuentra el archivo *.spe* con el comando de extensión:



Seguidamente, para finalizar la instalación, pulse *Aceptar* en los dos cuadros de dialogo que se presentan.

📌 Si un comando de extensión utiliza un paquete de R que no está en el ordenador la instalación intentará descargar e instalar el paquete, por lo que es necesario **tener una conexión a Internet activa** cuando se instala un comando de extensión. Si la instalación automática falla, será necesario descargar e instalar manualmente los paquetes de R requeridos por el comando de extensión siguiendo las instrucciones del apéndice de la guía *Cómo instalar manualmente un paquete de R*.

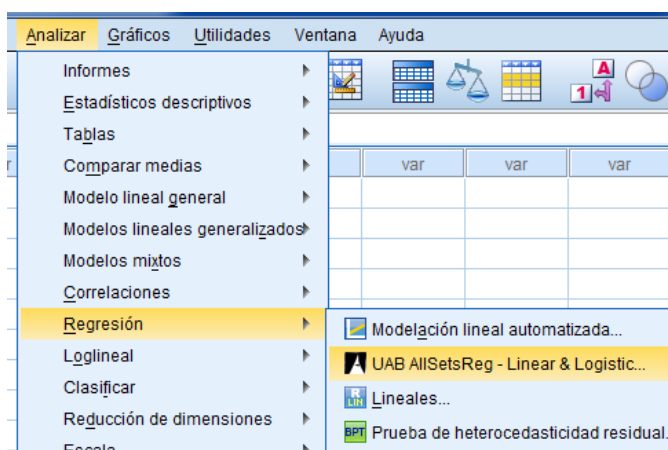
📌 Para que la operativa del nuevo comando de extensión esté totalmente disponible, hay que cerrar y volver a abrir *SPSS Statistics* tras la instalación.

Comprobación de la instalación

Es necesario cerrar SPSS y volverlo a abrir para que aparezca el comando de extensión en los menús.

Abra el menú *Analizar / Regresión y* además de los procedimientos propios de SPSS encontrará el procedimiento *UAB AllSetsReg - Linear & Logistic...* que acaba de instalar.

La mejor prueba para asegurarse de que ha instalado correctamente un comando de extensión es realizar un análisis con dicho comando.



Desinstalar un comando de extensión

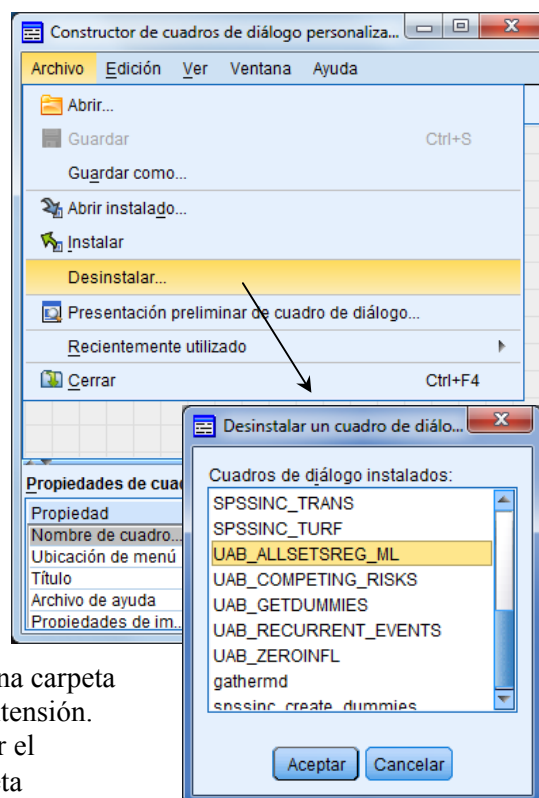
Es posible desinstalar el cuadro de diálogo (y la opción de menú asociada) desde la opción de menú *Utilidades | Cuadros de diálogo personalizados | Constructor de cuadros de diálogo personalizados*.

En la ventana del constructor de cuadros de diálogo escoja la opción de menú *Archivo / Desinstalar*, y en la lista que se abre a continuación escoja el comando de extensión que desea desinstalar.

Cierre *SPSS Statistics* y vuelva a abrirlo para que la desinstalación del cuadro de diálogo tenga efecto.

El proceso indicado desinstala la opción de menú y el cuadro de diálogo, pero el comando de extensión **sigue estando disponible en sintaxis**. Para desinstalar completamente un comando de extensión, de forma que no sea posible utilizarlo en sintaxis, es necesario localizar la carpeta *extensions* dentro de la carpeta de instalación de *SPSS Statistics* (*C:\Program Files\IBM\SPSS\Statistics\19\extensions* para SPSS19 en Windows 7) y borrar el contenido asociado al comando de extensión: una carpeta y los archivos con el mismo nombre que el comando de extensión. Por ejemplo, para el *UAB_ALLSETSREG_ML* hay que borrar el *UAB_ALLSETSREG.R*, *UAB_ALLSETSREG.xml* y la carpeta *UAB_ALLSETSREG_ML*.

⚠ Esta operación hay que realizarla con *SPSS Statistics* **cerrado**.



¿Cómo cribar variables para modelos predictivos?

Cuando el número inicial de predictores disponibles es tan elevado que supera el máximo admitido por el comando de extensión, que recordemos es de aproximadamente 15 términos, debe realizar un cribado previo de predictores para seleccionar los más relevantes.

La mejor estrategia es utilizar los métodos automáticos de inclusión y exclusión por pasos. Las variables a desechar se seleccionan entre las **primeras** que va excluyendo una regresión con el método BACKWARD y entre las **últimas** que va incluyendo una nueva regresión STEPWISE con los siguientes criterios de inclusión/exclusión para obligar a que se incorporen al modelo todos términos:

CRITERIA= PIN(0.999) POUT(1)

La elección no será difícil porque los términos de ambos conjuntos acostumbran a presentar numerosas coincidencias. Lógicamente, los términos a excluir deberán ser claramente no significativos.

El inconveniente de esta propuesta reside en las variables con más de dos categorías que se introducen descompuestas en variables ficticias, ya que los métodos de regresión por pasos no las consideran como una sola variable. De las siguientes alternativas podemos usar la más adecuada para cada variable concreta: 1) recodificar la variable en una variable binaria, 2) si es una variable con categorías ordenadas introducirla como cuantitativa, 3) las variables ficticias se obligan a permanecer en el modelo con un ENTER y se va valorando su significación a lo largo de los pasos, y 4) la variable no se introduce en las regresiones por pasos y se incluye directamente en el modelo máximo que evaluará UAB_AllSetsReg.

En los próximos apartados se explica detalladamente el funcionamiento del comando de extensión para la regresión lineal múltiple, y a continuación se detallan las particularidades de funcionamiento para los modelos de regresión logística y de Cox.

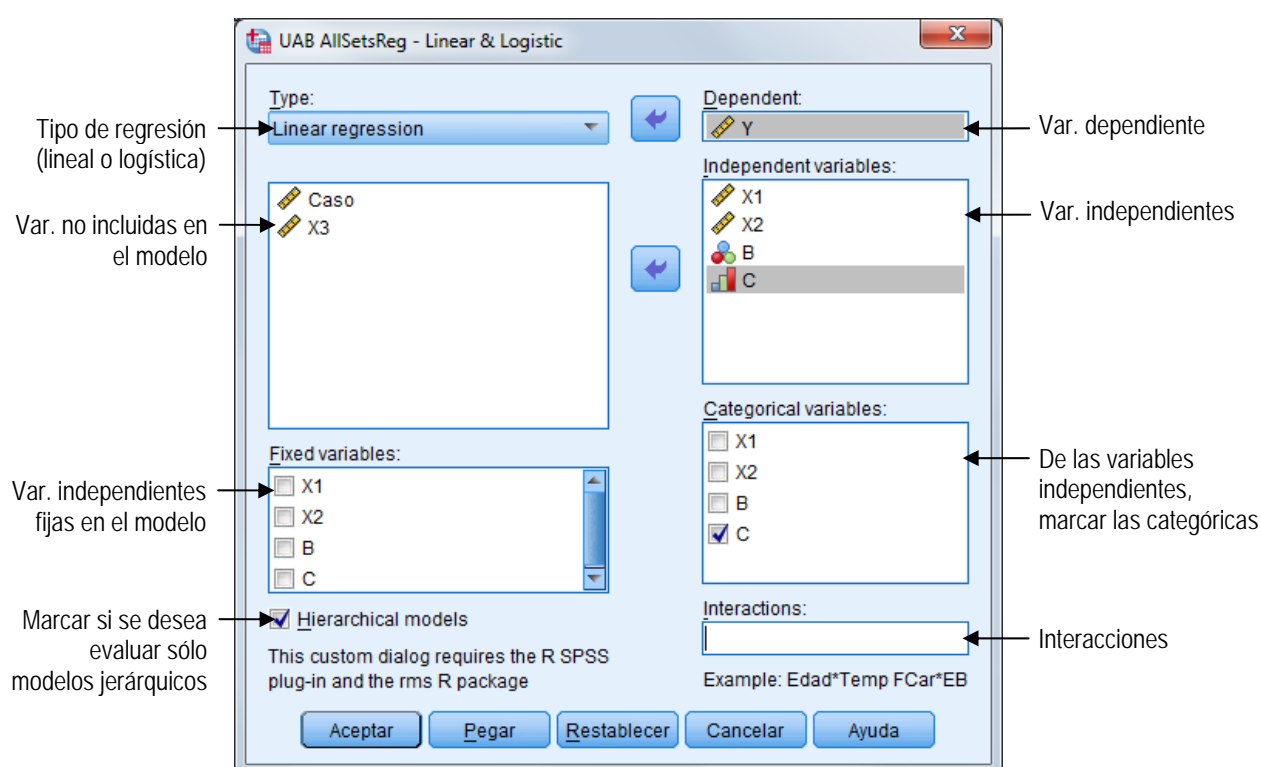
Regresión lineal múltiple

Para ilustrar el funcionamiento del comando de extensión con regresión lineal múltiple emplearemos el archivo de datos de prueba **DataTest_AllSetsReg_MR.sav**. Abra este archivo, que contiene las variables cuantitativas Y, X1, X2 y X3, la variable binaria B y la variable C con 3 categorías. En total hay 100 casos. El caso 4 no tiene valor X1 y el 10 carece de valor en la variable C.

Ejecute la opción de menú *Analizar | Regresión | UAB AllSetsReg - Linear & Logistic...* Si ha realizado correctamente todo el proceso de instalación aparecerá un cuadro de diálogo con la lista de todas las variables de la ventana de datos.

Para comenzar suponga que quiere hallar el mejor modelo predictivo de Y a partir de los predictores cuantitativos X1 y X2, del predictor binario B y del predictor C con 3 categorías (la variable X3 no se usa en este ejemplo).

En primer lugar elija **Linear regression** como modelo de regresión a estimar. Pase la variable dependiente Y a **Dependent**. Todas las variables independientes se arrastran al cuadro **Independent Variables**, y para los predictores que sean categóricos se marca su casilla de verificación en el cuadro **Categorical variables**. El proceso las descompone automáticamente en variables ficticias respecto a una categoría de referencia (el tipo de codificación no es relevante en un contexto predictivo), aunque este proceso es invisible para el usuario.



Para conseguir directamente los resultados puede hacer click en **Aceptar**, si bien recomendamos que haga uso de la sintaxis, que consigue con el botón **Pegar**:

```
UAB ALLSETSREG TYPE=LINEAR DEPENDENT=Y
/VARIABLES INDEPENDENT=X1 X2 B C
CATEGORICAL=C.
```

Tras ejecutar la anterior sintaxis obtendrá un resumen del proceso realizado en la ventana de resultados, indicando los términos del modelo de referencia y el número total de submodelos comparados. Puesto que el comando de extensión trabaja con la opción LISTWISE, se genera una tabla con los valores válidos y desconocidos de cada variable y se indica el número de sujetos (con valores válidos en todas las variables) que se han utilizado para efectuar los análisis, en el ejemplo 98, ya que se eliminan los casos 4 y 10.

```

Extension Command UAB AllSetsReg v0.0.8 (2013.01.18)
(c) JM Domenech & JB Navarro
Programmer: R Sesma
Laboratori d'Estadística Aplicada - Universitat Autònoma de Barcelona

This Extension Command uses functions of the Design and stats R packages.
stats R package (c) R Development Core Team, http://www.R-project.org.
rms R package (c) Frank E Harrell Jr, http://CRAN.R-project.org/package=rms.

```

ALL VARIABLES

Dependent: Y
Continuous: X1, X2, B
Categorical: C
Type: linear

Total number of submodels estimated: 15
A new dataset named @0.8146Results has been created with the results.

Case Processing Summary

	Y	X1	X2	B	C
Valid	100	99	100	100	99
Missing	0	1	0	0	1

Valid number of cases (listwise): 98

También se genera una nueva matriz de datos que presenta los resultados de todos los subconjuntos estimados. Por defecto los subconjuntos están ordenados por el criterio C_p de Mallows, aunque pueden ser reordenados por cualquiera de los otros criterios:

- R^2 Adj (coeficiente de determinación ajustado)
- AIC (Criterio de información de Akaike)
- BIC (Criterio de información bayesiano de Schwarz)
- R^2 (coeficiente de determinación)

	NVar	Variables	C_p	R^2 Adj	AIC	BIC	R^2
1	2	X1, X2	,29	,493	245	253	,503
2	3	X1, X2, B	2,29	,488	247	258	,503
3	2	X1, B	3,67	,474	249	257	,485
4	3	X1, X2, C	4,00	,484	249	262	,505
5	4	X1, X2, B, C	6,00	,478	251	266	,505
6	1	X1	6,70	,453	252	257	,458
7	3	X1, B, C	7,24	,466	252	265	,488
8	2	X1, C	10,06	,445	255	266	,462
9	1	X2	87,02	,016	309	314	,026
10	1	B	87,36	,014	309	315	,024
11	2	X2, B	88,56	,008	311	319	,029
12	2	X2, C	89,25	,005	312	323	,036
13	2	B, C	89,94	,001	313	323	,032
14	1	C	90,95	-,005	312	320	,016
15	3	X2, B, C	91,07	-,005	314	327	,037
16							

La primera fila de la nueva matriz de datos indica que el subconjunto formado por las 2 variables (Nvar=2) X1 y X2 es el mejor modelo de acuerdo al criterio C_p de Mallows ($C_p=0.29$).

Para ordenar los mejores subconjuntos según otro criterio simplemente se ordena la matriz de datos de acuerdo a dicho criterio. Tenga en cuenta que de acuerdo a C_p , AIC y BIC el mejor modelo es el que tiene un valor menor, pero según R^2 y R^2 Adj el mejor modelo es el que tiene un valor mayor.

Por ejemplo, si ordenamos la matriz según el valor BIC observamos que el mejor modelo sigue siendo el que incluye los términos X1 y X2 (BIC=253), pero el segundo y tercer mejores modelos cambian.

	NVar	Variables	Cp	R2Adj	AIC	BIC	R2
1	2	X1, X2	,29	,493	245	253	,503
2	2	X1, B	3,67	,474	249	257	,485
3	1	X1	6,70	,453	252	257	,458
4	3	X1, X2, B	2,29	,488	247	258	,503
5	3	X1, X2, C	4,00	,484	249	262	,505
6	3	X1, B, C	7,24	,466	252	265	,488
7	2	X1, C	10,06	,445	255	266	,462
8	4	X1, X2, B, C	6,00	,478	251	266	,505
9	1	X2	87,02	,016	309	314	,026
10	1	B	87,36	,014	309	315	,024
11	2	X2, B	88,56	,008	311	319	,029
12	1	C	90,95	-,005	312	320	,016
13	2	X2, C	89,25	,005	312	323	,036
14	2	B, C	89,94	,001	313	323	,032
15	3	X2, B, C	91,07	-,005	314	327	,037
16							

Nota: El valor C_p del modelo máximo siempre es el número de términos que incluye más 1. En el resultado del proceso se obtiene una $C_p=6$ para el modelo máximo que contiene los predictores X1, X2, B y C, porque la variable C se analiza descompuesta en dos variables ficticias, por lo que en realidad el modelo máximo contiene 5 términos: X1, X2, X3, @C21, @C31.

Estimar el modelo seleccionado

Una vez seleccionado el mejor modelo deberá estimarlo mediante el procedimiento REGRESSION con la opción LISTWISE para el tratamiento de los valores faltantes. Si el mejor modelo incluye variables categóricas deberá generar previamente las variables ficticias correspondientes. En el ejemplo, para el mejor modelo predictivo de Y de acuerdo al criterio C_p , se obtiene:

```
REGRESSION /MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/DEPENDENT Y
/METHOD=ENTER X1 X2.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,711 ^a	,505	,495	3,435

a. Predictors: (Constant), X2, X1

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	31,602	2,703		11,692	,000	26,236	36,967
	X1	,327	,034	,691	9,605	,000	,259	,395
	X2	,211	,069	,221	3,069	,003	,075	,348

a. Dependent Variable: Y

Observe que los valores de ajuste no coinciden exactamente entre los resultados del comando de extensión ($R^2=0.503$, $R^2\text{Adj}=0.493$) y el resultado del procedimiento REGRESSION ($R^2=0.505$, $R^2\text{Adj}=0.495$). Ello se debe a que en el comando de extensión se ha aplicado la opción LISTWISE a todos los predictores disponibles, por tanto se ha eliminado el caso 10 que no tiene valor en el predictor C ($n=98$). Por el contrario, como en el procedimiento REGRESSION no se incluye el predictor C, el caso 10 se incluye en el análisis ($n=99$).

Para evitar este tipo de incoherencias, que pueden dar lugar a resultados más diferentes y sesgados cuando el número de valores desconocidos es elevado, es recomendable que antes de estimar el modelo finalmente seleccionado se establezca un filtro seleccionando los casos que no tengan ningún valor faltante en el conjunto de predictores incluidos en el modelo máximo.

```
COUNT NumMis=X1 X2 B C (SYSMIS).
TEMPORARY.
SELECT IF (NumMis=0).
REGRESSION /MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/DEPENDENT Y
/METHOD=ENTER X1 X2.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,710 ^a	,503	,493	3,443

a. Predictors: (Constant), X2, X1

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	31,853	2,731		11,664	,000	26,431	37,274
	X1	,326	,034	,693	9,557	,000	,258	,394
	X2	,205	,070	,213	2,943	,004	,067	,343

a. Dependent Variable: Y

Predictores fijos

Por cuestiones de diseño o por motivos teóricos es posible que ciertas variables deban estar presentes en todos los modelos analizados. Para forzar que un determinado predictor esté en todos los modelos debe marcar su casilla de verificación en el cuadro **Fixed variables**.

Suponga que por motivos de diseño el término X1 debe estar presente en todos los modelos evaluados. La imagen siguiente muestra la especificación de este nuevo modelo máximo, junto con la sintaxis que genera y el resultado del proceso que ordena los subconjuntos según C_p de Mallows. Observe que sólo se presentan 8 modelos, ya que de todos los modelos con 1 único término sólo se estudia el que incluye el término X1; de los modelos con 2 términos sólo se incluyen los tres modelos que contienen X1, etc.

Fixed variables:

☒ X1

☐ X2

☐ B

☐ C

```
UAB ALLSETSREG TYPE=LINEAR DEPENDENT=Y
/VARIABLES INDEPENDENT=X1 X2 B C
CATEGORICAL=C
FIXED X1.
```

	NVar	Variables	Cp	R2Adj	AIC	BIC	R2
1	2	X1, X2	,29	,493	245	253	,503
2	3	X1, X2, B	2,29	,488	247	258	,503
3	2	X1, B	3,67	,474	249	257	,485
4	3	X1, X2, C	4,00	,484	249	262	,505
5	4	X1, X2, B, C	6,00	,478	251	266	,505
6	1	X1	6,70	,453	252	257	,458
7	3	X1, B, C	7,24	,466	252	265	,488
8	2	X1, C	10,06	,445	255	266	,462

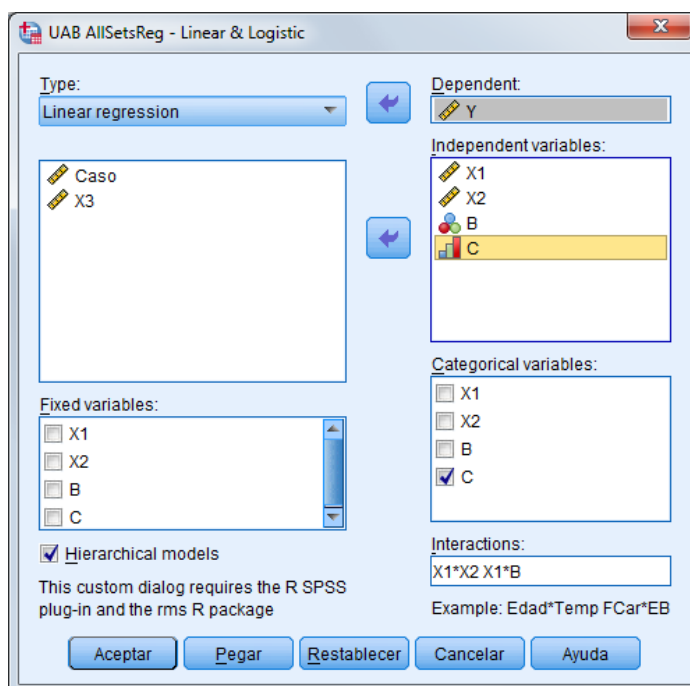
Términos de interacción y modelos jerárquicos

El modelo máximo puede contener términos de interacción entre los predictores seleccionados. La siguiente imagen ilustra cómo añadir las interacciones $X1 \times X2$ y $X1 \times B$ al modelo máximo del ejemplo.

Incluir interacciones incrementa sustancialmente el número de modelos evaluados, como puede comprobar en la matriz de datos que se genera con todos los submodelos evaluados.

Un modelo como el de este ejemplo, con 6 términos (4 variables y 2 interacciones), genera un total de $2^6 - 1 = 63$ submodelos. No obstante, puesto que 2 de estos términos son interacciones y se ha activado la casilla **Hierarchical models**, se generan un total de 25 submodelos ya que sólo se valoran los jerárquicos. La sintaxis generada es:

```
UAB ALLSETSREG TYPE=LINEAR DEPENDENT=Y
/VARIABLES INDEPENDENT=X1 X2 B C
CATEGORICAL=C
INTERACTIONS=X1*X2 X1*B.
```



	NVar	Variables	Cp	R2Adj	AIC	BIC	R2
1	2	X1, X2	,24	,493	245	253	,503
2	3	X1, X2, X1*X2	1,59	,491	247	257	,507
3	3	X1, X2, B	2,24	,488	247	258	,503
4	4	X1, X2, B, X1*B	2,44	,492	247	260	,513
5	4	X1, X2, B, X1*X2	3,59	,486	249	262	,507
6	2	X1, B	3,62	,474	249	257	,485
7	3	X1, X2, C	3,95	,484	249	262	,505
8	3	X1, B, X1*B	4,25	,476	249	260	,493
9	5	X1, X2, B, X1*X2, X1*B	4,32	,487	249	265	,514
10	4	X1, X2, C, X1*X2	5,42	,481	250	266	,508
11	4	X1, X2, B, C	5,95	,478	251	266	,505
12	5	X1, X2, B, C, X1*B	6,21	,482	251	269	,514
13	1	X1	6,64	,453	252	257	,458
14	3	X1, B, C	7,18	,466	252	265	,488
15	5	X1, X2, B, C, X1*X2	7,42	,475	252	271	,508
16	4	X1, B, C, X1*B	7,88	,467	253	269	,495
17	6	X1, X2, B, C, X1*X2, X1*B	8,00	,478	253	274	,516
18	2	X1, C	10,00	,445	255	266	,462
19	1	X2	86,92	,016	309	314	,026
20	1	B	87,25	,014	309	315	,024
21	2	X2, B	88,46	,008	311	319	,029
22	2	X2, C	89,14	,005	312	323	,036
23	2	B, C	89,83	,001	313	323	,032
24	1	C	90,84	-,005	312	320	,016
25	3	X2, B, C	90,97	-,005	314	327	,037
26							

Aún incluyendo interacciones, el mejor modelo de acuerdo al criterio C_p de Mallows sigue siendo el que incluye los predictores $X1$ y $X2$.

Regresión logística

Como en regresión lineal múltiple, la selección de la mejor ecuación de regresión logística con finalidad predictora se puede realizar con diferentes procedimientos. Uno de los mejores consiste en establecer unos criterios de selección, como la AIC o el área bajo la curva ROC (AUC), construir todos los posibles submodelos combinando los términos del modelo máximo, y valorar para cada uno de ellos el grado de cumplimiento del criterio establecido.

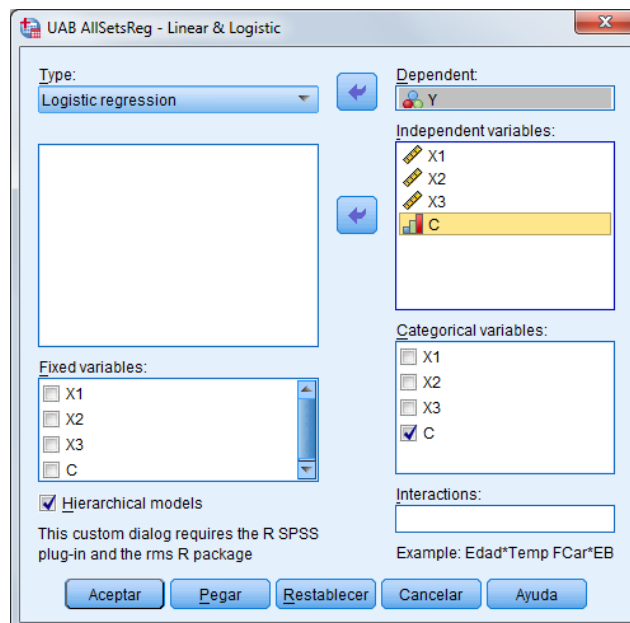
Para practicar abra el archivo de datos de prueba **DataTest_AllSetsReg_LR.sav**, que contiene las variables cuantitativas X1, X2 y X3, la variable C con 3 categorías y la variable Y binaria. En total hay 200 casos. El caso 10 no tiene valor en X2 y el caso 31 carece de valor en la variable X3.

Suponga que queremos seleccionar el mejor subconjunto de regresión para predecir la variable binaria Y tomando como modelo máximo el que contiene los predictores cuantitativos X1, X2, X3 y el predictor C con 3 categorías.

Ejecute la opción de menú *Analizar | Regresión | UAB AllSetsReg - Linear & Logistic...*. Se presenta el cuadro de diálogo ya conocido, en el que debe seleccionar **Logistic regression** como modelo de regresión. Pase la variable dependiente Y a **Dependent** y los 4 predictores al cuadro **Independent Variables**. No olvide indicar que C es un predictor categórico.

La sintaxis que obtiene sigue las mismas pautas que hemos presentado para la regresión lineal:

```
UAB ALLSETSREG TYPE=LOGISTIC DEPENDENT=Y
/VARIABLES INDEPENDENT=X1 X2 X3 C
CATEGORICAL=C.
```



Al ejecutar la sintaxis obtiene en la ventana de resultados un resumen del proceso.

```
Extension Command UAB AllSetsReg v0.0.8 (2013.01.18)
(c) JM Domenech & JB Navarro
Programmer: R Sesma
Laboratori d'Estadística Aplicada - Universitat Autònoma de Barcelona

This Extension Command uses functions of the Design and stats R packages.
stats R package (c) R Development Core Team, http://www.R-project.org.
rms R package (c) Frank E Harrell Jr, http://CRAN.R-project.org/package=rms.
```

```
-----
ALL VARIABLES
Dependent: Y
Continuous: X1, X2, X3
Categorical: C
Type: logistic
```

```
Total number of submodels estimated: 15
A new dataset named @0.3807Results has been created with the results.
```

Case Processing Summary

	Y	X1	X2	X3	C
Valid	200	200	199	199	200
Missing	0	0	1	1	0

Valid number of cases (listwise): 198

Se genera también una nueva ventana de datos con todos los modelos estimados.

	NVar	Variables	AIC	AUC	Se	Sp	@2LL	p_fit_HL	p_fit_CH
1	3	X1, X2, X3	252,4	,721	68.7%	64.6%	244,4	,711	,738
2	2	X2, X3	253,6	,711	69.7%	65.7%	247,6	,433	,456
3	4	X1, X2, X3, C	254,4	,720	67.7%	65.7%	242,4	,691	,295
4	3	X2, X3, C	255,4	,709	69.7%	65.7%	245,4	,142	,557
5	2	X1, X2	258,8	,688	62.6%	65.7%	252,8	,886	,601
6	2	X1, X3	260,2	,680	64.6%	65.7%	254,2	,138	,702
7	1	X2	260,4	,679	61.6%	64.6%	256,4	,525	,067
8	1	X3	261,3	,662	63.6%	58.6%	257,3	,857	,836
9	3	X1, X2, C	261,4	,688	64.6%	64.6%	251,4	,430	,762
10	3	X1, X3, C	262,3	,679	61.6%	63.6%	252,3	,459	,098
11	2	X2, C	262,8	,680	62.6%	67.7%	254,8	,789	,434
12	2	X3, C	263,3	,668	60.6%	64.6%	255,3	,848	,290
13	1	X1	274,8	,573	58.6%	57.6%	270,8	,516	,335
14	2	X1, C	277,9	,573	52.5%	54.5%	269,9	,755	,191
15	1	C	279,4	,538	59.6%	47.5%	273,4	1,000	,000

Los criterios de selección del mejor modelo para regresión logística son:

- AIC (Criterio de información de Akaike)
- AUC (Área bajo la curva ROC)
- Se Sp (sensibilidad y especificidad para un punto de corte 0.50.)
- @2LL (valor de -2 veces el logaritmo de la verosimilitud)
- p_fit_HL (grado de significación del índice de ajuste de Hosmer-Lemeshow)
- p_fit_CH (grado de significación del índice de ajuste de le Cessie-van Houwelingen)

⚠ Los índices de ajuste de *Hosmer-Lemeshow* y de *le Cessie-van Houwelingen* pueden dar resultados muy diferentes en determinados modelos. Si para el modelo seleccionado uno de ellos es estadísticamente significativo y el otro no, es recomendable realizar un análisis de residuales para garantizar el cumplimiento de los supuestos del modelo.

De acuerdo al criterio AIC, por el que inicialmente se ordenan los subconjuntos, el mejor modelo predictivo incluye X1, X2 y X3 (AIC=252.4). Se trata además de un modelo con una sensibilidad (68.7%) y especificidad (64.6%) similares, aunque bajas, con el valor más alto del área bajo la curva ROC (0.721), aunque también es un valor bajo, y con una prueba de ajuste global claramente no significativa ($p=0.711$). El funcionamiento para predictores fijos y términos de interacción es idéntico a la regresión lineal.

Regresión de Cox

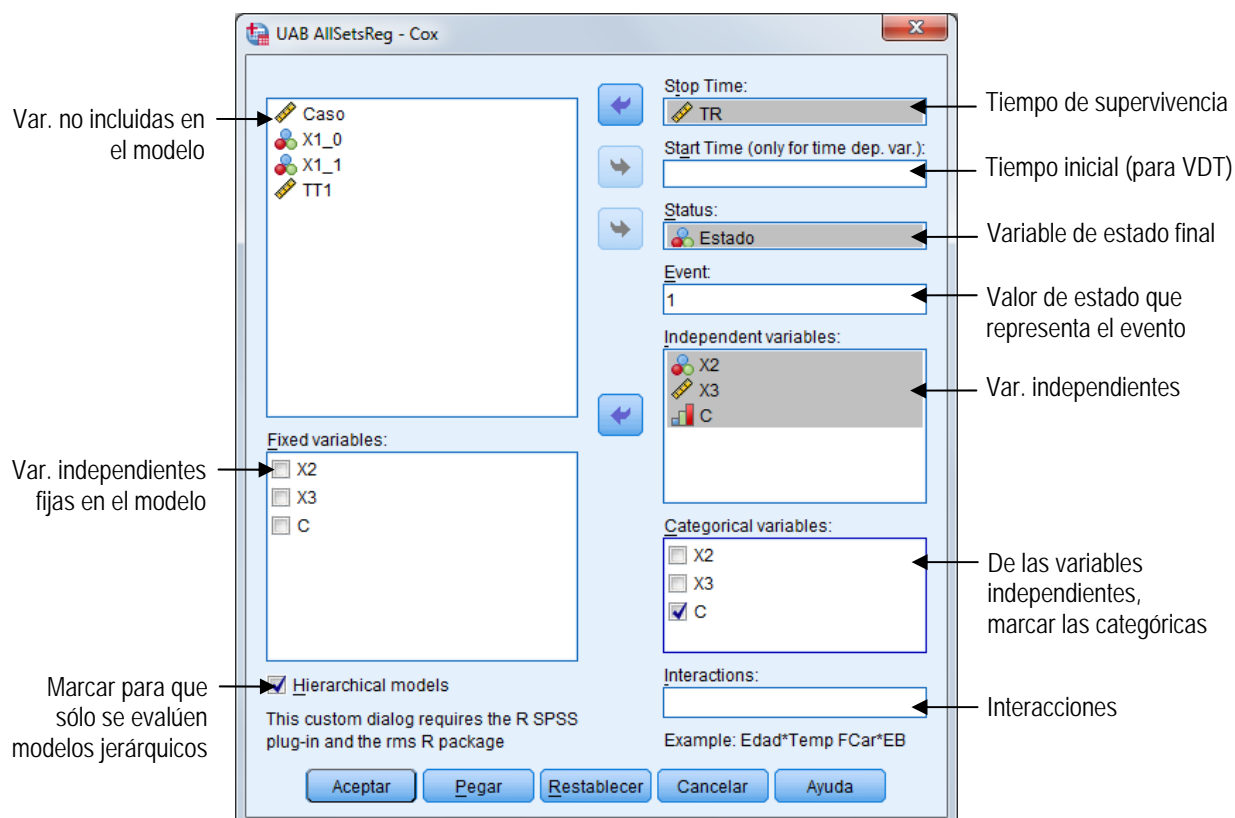
Para realizar la selección del mejor modelo predictivo en regresión de Cox debe **instalar en primer lugar el comando de extensión UAB_AllSetsReg_Cox** mediante el archivo Uab_AllSetsReg_Cox.spe que ha descargado junto a esta documentación (en el apartado 'Instalar el comando de extensión' de este mismo documento se explica el procedimiento).

Para practicar emplearemos el archivo de datos de prueba **DataTest_AllSetsReg_Cox.sav**, que contiene las variables:

- FE; FU: Fecha de entrada y Fecha de último seguimiento.
- Estado: Estado en el último seguimiento (0:Sin recaída, 1:Recaída, 2:Muerto). El evento de interés es la recaída, mientras que los sujetos sin recaída y los muertos se consideran datos censurados, ya que en el momento del último seguimiento no habían presentado el evento de interés.
- X1_0: Predictor binario dependiente del tiempo. Esta variable registra el valor inicial.
- X1_1: Predictor binario dependiente del tiempo. Esta variable registra los cambios que se han producido respecto al valor inicial, que se dan en los casos 1, 4, 12, 22 y 30.
- FX1_1: Fecha en que se produce un cambio en el predictor binario dependiente del tiempo X1.
- X2: Predictor binario.
- X3: Predictor cuantitativo.
- C: Predictor categórico con 3 categorías.
- TR: Tiempo de supervivencia. Se ha generado con la instrucción:
`COMPUTE TR= CTIME.DAYS(FU-FE).`
- TT1: Tiempo desde FE hasta FX1_1 (sólo tiene valor para los 5 casos que cambian en X1). Se ha generado con la instrucción:
`COMPUTE TT1= CTIME.DAYS(FX1_1 - FE).`

Como puede comprobar en la lista anterior, se incluyen todas las variables necesarias para realizar una regresión de Cox junto a las asociadas a una variable dependiente del tiempo (VDT) como es X1.

En primer lugar practicaremos la selección del mejor modelo predictivo **sin tener en cuenta la variable dependiente del tiempo (VDT)** X1. El modelo máximo incluirá los predictores X2, X3 y C. Elija la opción de menú *Anализar | Regresión | UAB AllSetsReg - Cox...* Si ha realizado correctamente el proceso de instalación aparecerá un cuadro de diálogo con la lista de todas las variables de la ventana de datos:



Al pegar aparece la siguiente sintaxis:

```
UAB ALLSETSREG TYPE=COX StopTime=TR Status=Estado Event=1  
/VARIABLES INDEPENDENT=X2 X3 C CATEGORICAL=C.
```

Y al ejecutarla se obtiene un resumen del proceso en la ventana

```
Extension Command UAB AllSetsReg v0.0.8 (2013.01.18)  
(c) JM Domenech & JB Navarro  
Programmer: R Sesma  
Laboratori d'Estadística Aplicada - Universitat Autònoma de Barcelona  
  
This Extension Command uses functions of the Design and stats R packages.  
stats R package (c) R Development Core Team, http://www.R-project.org.  
rms R package (c) Frank E Harrell Jr, http://CRAN.R-project.org/package=rms.
```

```
-----  
ALL VARIABLES  
Stop Time: TR  
Status: Estado (Event value= 1)  
Continuous: X2, X3  
Categorical: C  
Type: cox
```

```
Total number of submodels estimated: 7  
A new dataset named @0.5248Results has been created with the results.
```

Case Processing Summary

	TR	Estado	X2	X3	C
Valid	32	32	32	32	32
Missing	0	0	0	0	0

Valid number of cases (listwise): 32

Junto a una nueva matriz de datos con los criterios de selección del mejor modelo, que en el caso de la regresión de Cox son:

- AIC (Criterio de información de Akaike)
- R^2 de Atkinson
- @2LL (valor de -2 veces el logaritmo de la verosimilitud)

Según el criterio AIC el mejor modelo predictivo de la tasa de riesgo de sufrir el evento es el que incluye X3 y C, si bien el modelo que además incluye X2 tiene un AIC muy próximo, y además este modelo tiene un menor valor de $-2LL$. Ante estos resultados se puede afirmar que en la práctica ambos modelos son igualmente adecuados.

	NVar	Variables	AIC	R2	@2LL
1	2	X3, C	83,4	,2580	77,449
2	3	X2, X3, C	83,5	,2574	75,514
3	2	X2, X3	85,9	,2359	81,933
4	1	X3	88,2	,2154	86,244
5	1	C	100,9	,1032	96,862
6	2	X2, C	101,8	,0951	95,774
7	1	X2	111,0	,0131	108,990

Variables dependientes del tiempo (VDT)

Para introducir en el modelo máximo un predictor dependiente del tiempo, es decir cuyos valores puedan cambiar durante el seguimiento, es **necesario cambiar la estructura de la matriz de datos de manera que contenga un registro para cada cambio de valor en la VDT**.

En la matriz de datos de ejemplo hay un total de 32 registros, con 5 sujetos que cambian su valor en el predictor X1. La matriz de datos reestructurada tendrá 37 registros, con los identificadores de los 5 sujetos duplicados.

🔗 Esta forma de estructurar la matriz de datos es la empleada por el programa de análisis estadístico R, con el que internamente se realizan los cálculos de selección del mejor modelo predictivo. Además es la estructura que se obtiene cuando la entrada de datos se ha realizado con un gestor de bases de datos relacional como Microsoft Access.

Vemos con detenimiento los cambios realizados a partir del caso 1, el primero que cambia su valor en X1. Eliminando las 3 fechas que sirvieron para el cálculo de los tiempos TR y TT1, los datos originales de este caso son:

Caso	Estado	X1_0	X1_1	X2	X3	C	TR	TT1
1	1	1	0	0	18	1	48	22

Por tanto se trata de un sujeto que desde el momento inicial (día 0) hasta el día 21 (TT1-1) tiene el valor X1=1, y en ese intervalo de tiempo no se ha producido el evento. El día 22 cambia al valor X1=0, permaneciendo en el estudio hasta el día TR=48 en el que se produce el evento (Estado=1).

En la matriz reestructurada el caso 1 se convertirá en dos registros con los siguientes valores:

Caso	Estado	X1	X2	X3	C	Start	Stop
1	0	1	0	18	1	0	21
1	1	0	0	18	1	22	48

Observe que las dos variables que registraban los valores de X1_0 y X1_1 se han agrupado en una única variable X1, y que los tiempos de supervivencia se indican con las variables Start y Stop. El primer registro contempla desde el día 0 al día 21, con X1=1 y Estado=0 porque no se produce el evento. El segundo registro contempla desde el día 22 al día 48, con X1=0 y Estado=1 porque se produce el evento. Los valores en el resto de predictores que no cambian (X2, X3, C) se mantienen idénticos en los dos nuevos registros.

De la misma manera se opera con los otros 4 casos cuyo valor en X1 cambia. Observe el caso 30, es similar al caso 1 con la diferencia de que el cambio en X1 es pasar del valor 0 al 1, y de que el Estado en los dos registros reestructurados tiene valor 0 porque en este sujeto no se produce el evento de interés.

Caso	Estado	X1_0	X1_1	X2	X3	C	TR	TT1
30	0	0	1	0	23	0	190	40



Caso	Estado	X1	X2	X3	C	Start	Stop
30	0	0	0	23	0	0	39
30	0	1	0	23	0	40	190

En los casos que no cambian de valor en X1 se debe realizar un pequeño cambio, poniendo el valor Start=0 y el valor Stop en el del tiempo de supervivencia. Por ejemplo, para el caso 2:

Caso	Estado	X1_0	X1_1	X2	X3	C	TR	TT1
2	2	1	.	1	21	1	85	.



Caso	Estado	X1	X2	X3	C	Start	Stop
2	2	1	1	21	1	0	85

En el .zip que ha descargado del campus, la matriz de datos reestructurada se encuentra almacenada en el archivo **DataTest_AllSetsReg_CoxVDT.sav**, ábralo y ejecute la opción de menú *Analizar | Regresión | UAB AllSetsReg - Cox...*

Estimaremos el mejor subconjunto predictivo de la tasa de riesgo del evento a partir de un modelo máximo que incluya los predictores X1, X2, X3, C y la interacción X1*X2.

UAB AllSetsReg - Cox

Caso

Stop Time: Stop

Start Time (only for time dep. var.): Start

Status: Estado

Event: 1

Independent variables: X1, X2, X3, C

Fixed variables: X1, X2, X3, C

☒ Hierarchical models

This custom dialog requires the R SPSS plug-in and the rms R package

Acceptar Pegar Restablecer Cancelar Ayuda

Interactions: X1*X2

Example: Edad*Temp FCar*EB

```
UAB ALLSETSREG TYPE=COX StartTime=Start
StopTime=Stop Status=Estado Event=1
/VARIABLES INDEPENDENT=X1 X2 X3 C
CATEGORICAL=C INTERACTIONS=X1*X2.
```

	NVar	Variables	AIC	R2	@2LL
1	3	X1, X3, C	73,2	,3489	65,166
2	4	X1, X2, X3, C	75,1	,3315	65,114
3	5	X1, X2, X3, C, X1*X2	76,7	,3175	64,693
4	2	X1, X3	80,1	,2870	76,117
5	3	X1, X2, X3	80,5	,2837	74,493
6	4	X1, X2, X3, X1*X2	81,8	,2721	73,792
7	2	X3, C	83,4	,2581	77,362
8	3	X2, X3, C	83,4	,2577	75,413
9	2	X1, C	84,9	,2440	78,946
10	2	X2, X3	85,9	,2356	81,894
11	3	X1, X2, C	86,9	,2263	78,943
12	1	X3	88,2	,2149	86,217
13	4	X1, X2, C, X1*X2	88,6	,2113	78,623
14	1	X1	95,5	,1505	93,450
15	2	X1, X2	96,9	,1372	92,950
16	3	X1, X2, X1*X2	97,4	,1335	91,370
17	1	C	100,6	,1044	96,639
18	2	X2, C	101,5	,0966	95,517
19	1	X2	110,8	,0136	108,839

Según el criterio AIC el mejor modelo predictivo es el que incluye las variables X1, X3 y C. Para estimar este modelo con SPSS debe emplear la estructura original, en la que cada caso ocupa un único registro, definiendo el predictor X1 dependiente del tiempo tal y como SPSS requiere.

```
GET FILE=' DataTest_AllSetsReg_Cox.sav'.
TIME PROGRAM.
COMPUTE X1= (T_ < TT1)*X1_0 + (T_ >= TT1)*X1_1.
IF (MISSING(TT1)) X1= X1_0.
COXREG TR
  /STATUS=Estado(1) /CONTRAST (C)=Indicator(1)
  /METHOD=ENTER X1 X3 C /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

Omnibus Tests of Model Coefficients^a

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
65,187	42,436	4	,000	47,282	4	,000	47,282	4	,000

a. Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
X1	-2,655	,818	10,522	1	,001	,070	,014	,350
X3	-,455	,139	10,761	1	,001	,634	,483	,833
C			11,193	2	,004			
C(1)	2,341	,705	11,016	1	,001	10,397	2,609	41,436
C(2)	1,209	,880	1,884	1	,170	3,349	,596	18,806

! Las estimaciones obtenidas directamente con SPSS o mediante el comando de extensión (que emplea funciones de R) difieren ligeramente ($-2LL=65.187$ y $-2LL=65.166$ respectivamente). Ello es debido a que la función de verosimilitud maximizada por ambas aplicaciones no es idéntica. En la práctica las diferencias no tienen ninguna relevancia.

La forma de reestructurar la matriz de datos para el análisis con el comando de extensión UAB_AllSetsReg_Cox en presencia de VDT sirve para cualquier número de cambios que se produzcan en el predictor. Supongamos unos datos nuevos en los que X1 tiene 4 categorías y cambia dos veces de valor en el caso 1 y tres veces en el caso 2. Tenemos los valores X1_0, X1_1, X1_2 y X1_3, junto a los tiempos TT1, TT2 y TT3, que indican el tiempo transcurrido con cada valor de X1:

Caso	Estado	X1_0	X1_1	X1_2	X1_3	X2	X3	C	TR	TT1	TT2	TT3
1	0	2	1	3	.	0	18	1	125	22	34	.
2	1	1	3	4	1	1	21	1	254	101	75	19

El análisis de estos datos con SPSS requiere la siguiente sintaxis:

```
TIME PROGRAM.
COMPUTE X1= (T_ < TT1)*X1_0 + (T_ >= TT1 AND T_ < TT2)*X1_1 +
            (T_ >= TT2 AND T_ < TT3)*X1_2 + (T_ >= TT3)*X1_3.
IF (MISSING(TT1)) X1= X1_0.
COXREG TR
  /STATUS=Estado(1) /CONTRAST (X1)=Indicator(1)
  /METHOD=ENTER X1 /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

En la matriz de datos reestructurada, estos dos casos se transforman en siete registros, tres para el caso 1 y cuatro para el caso 2:

Caso	Estado	X1	X2	X3	C	Start	Stop
1	0	2	0	18	1	0	21
1	0	1	0	18	1	22	55
1	0	3	0	18	1	56	125
2	0	1	1	21	1	0	100
2	0	3	1	21	1	101	175
2	0	4	1	21	1	176	194
2	1	1	1	21	1	195	254

El caso 1 está desde el día 0 hasta el día 21 (TT1=22 días) con el valor X1=2, después está desde el día 22 al día 55 (TT2=34 días) con el valor X1=1, y finalmente está desde el día 56 al día 125 (día de la última observación) con el valor X1=3. Este caso no presenta el evento de interés, por lo que su variable Estado=0 en los 3 registros.

El caso 2 está desde el día 0 hasta el día 100 (TT1=101 días) con el valor X1=1, después está desde el día 101 al día 175 (TT2=75 días) con el valor X1=3, después pasa desde el día 176 al día 194 (TT3=19 días) con el valor X1=4, y finalmente está desde el día 195 al día 254 (día de la última observación) con el valor X1=1. Este caso presenta el evento de interés, por lo que su variable Estado=1 en el último registro.

Referencia de sintaxis de UAB_AllSetsReg

```
UAB ALLSETSREG  TYPE={LINEAR  } DEPENDENT=varname
                  {LOGISTIC}
                  TYPE={COX      } StopTime=varname [StartTime=varname]
                                      Status=varname Event=integer
                  [NONHIERARCHICAL]
/VARIABLES      INDEPENDENT = varlist
                  [CATEGORICAL = varlist]
                  [INTERACTIONS = varlist]
                  [FIXED = varlist]

[HELP]
```

- **TYPE**: tipo de regresión. **LINEAR** realiza una regresión lineal, **LOGISTIC** realiza una regresión logística y **COX** realiza una regresión de Cox.
- **DEPENDENT**: variable dependiente (si **TYPE** = **LINEAR** o **LOGISTIC**). Si **TYPE** = **LOGISTIC**, esta variable debe ser binaria (0,1).
- **StopTime**: En modelos sin variables dependientes del tiempo es el tiempo de supervivencia. En modelos con variables dependientes del tiempo es el tiempo final de cada intervalo (si **TYPE** = **COX**).
- **StartTime**: En modelos sin variables dependientes del tiempo este parámetro debe quedar vacío. En modelos con variables dependientes del tiempo es el tiempo de inicio de cada intervalo (si **TYPE** = **COX**). Requiere que la matriz de datos haya sido preparada por el usuario de acuerdo a la estructura requerida para modelos con variables dependientes del tiempo.
- **Status**: variable de estado final (si **TYPE** = **COX**).
- **Event**: valor (entero ≥ 0) de la variable de estado que indica el evento (si **TYPE** = **COX**).
- **NONHIERARCHICAL**, NO se aplican modelos jerárquicos. Si no aparece este parámetro, por defecto se utilizan modelos jerárquicos.
- **INDEPENDENT**: lista de variables independientes. Esta lista incluye las variables categóricas, pero NO las interacciones.
- **CATEGORICAL**: lista de variables categóricas entre las independientes.
- **INTERACTIONS**: lista de interacciones, en el formato `var1*var2 var3*var4...`. Los componentes de las interacciones deben estar en la lista de variables **INDEPENDENT**.
- **FIXED**: lista de variables fijas entre las independientes.