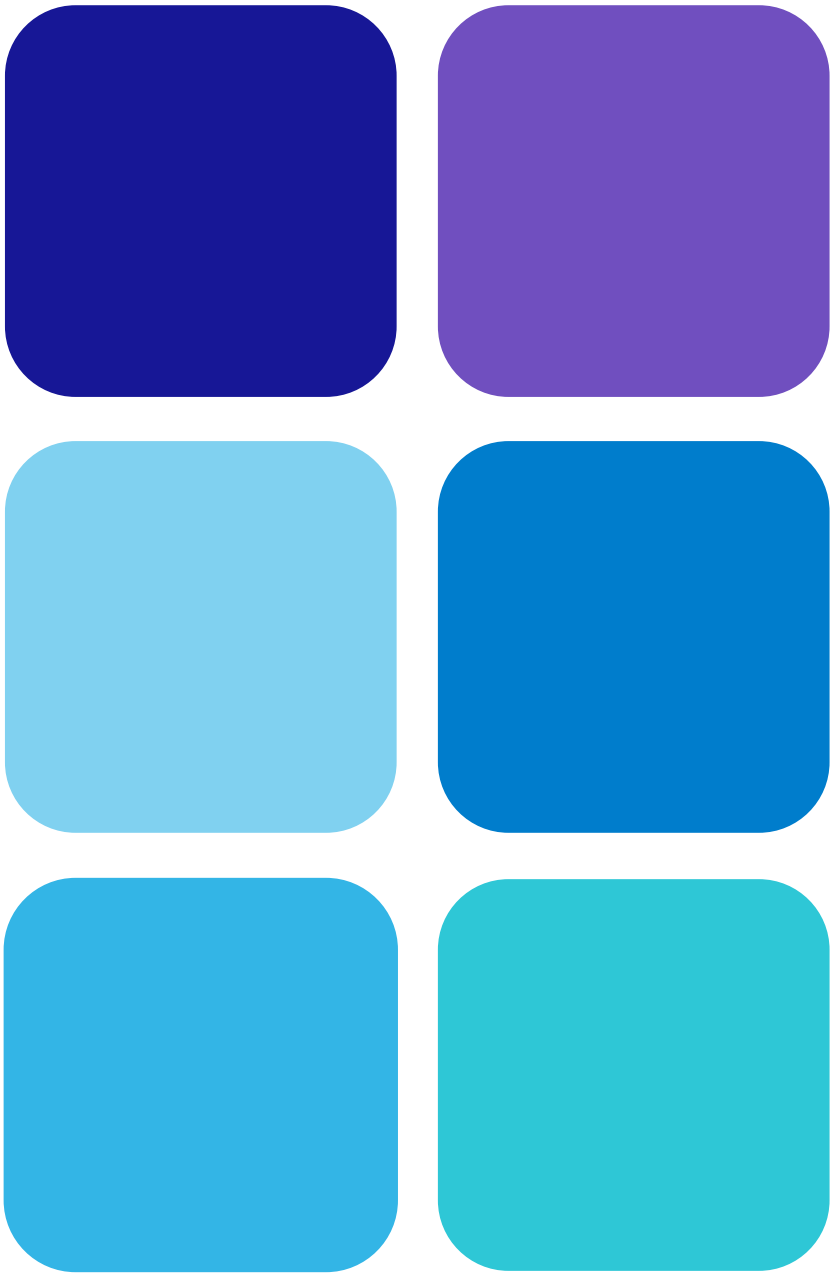


www.sapientia.uji.es | 82

UNIVERSITAT JAUME·I

Pronóstico con interacción de variables categóricas

Jesús F. Rosel Remírez
M.ª Pilar Jara Jiménez
Francisco Herrero Machancoses

Pronóstico con interacción de variables categóricas

Jesús F. Rosel Remírez
M.ª Pilar Jara Jiménez
Francisco Herrero Machancoses

UNIVERSITAT JAUME·I

Departament de Psicologia Evolutiva, Social i Metodologia

■ Codi d'assignatura PS1046

 Pronóstico con interacción de variables categóricas - UJI - DOI: http://dx.doi.org/10.6035/Sapientia82

Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Col·lecció Sapientia 82
www.sapientia.uji.es
Primera edició, 2014

ISBN: 978-84-697-0832-3

Publicacions de la Universitat Jaume I és una editorial membre de l'UNE, cosa que en garanteix la difusió de les obres en els àmbits nacional i internacional. www.une.es

ÍNDICE

Introducción

¿Por qué hacer un texto acerca de la interacción de variables? Solo hay tres libros básicos sobre este tema (Aguinis, 2004; Aiken y West, 1991, y Jaccard, Turrisi y Wan, 2003), y cada uno de ellos tiene sus ventajas e inconvenientes. El libro de Jaccard, Turrisi y Wan (2003) es una breve introducción al tema, aunque centrado sobre todo en interacción entre variables continuas, mientras que el libro de Aiken y West (1991) tiene un planteamiento más amplio, si bien solo dedica un capítulo a la interacción entre una variable continua y una variable categórica. El libro de Aguinis (2004) es el más técnico, y se centra en aspectos de la potencia estadística de la interacción, incluso remite a una página web elaborada por el mismo autor con el fin de calcular la varianza del error en el caso de heterogeneidad de las muestras (http://mypage.iu.edu/~haguinis/mmr/index.html), y de calcular la potencia estadística de los contrastes de pendientes. Ninguno de los anteriores manuales indica cómo introducir los bloques de variables *dummy* (que representan a una variable categórica) en un programa estadístico, cómo realizar el ajuste estadístico del modelo, cómo desglosar la ecuación general del modelo en tantas ecuaciones como grupos haya, ni cómo llevar a cabo una figura de conjunto en la que cada grupo esté perfectamente identificado.

Más recientemente, se han publicado los volúmenes de Hayes (2013) y Jose (2013), pero tienen el inconveniente de que tratan sobre mediación e interacción de variables y que, por lo tanto, asumen conocimientos por parte del lector sobre ambas temáticas (mediación e interacción), lo que no siempre es el caso.

Los motivos por los que se ha hecho este texto son de tipo docente, y tienen como función facilitar el aprendizaje de los alumnos sobre aspectos básicos de la regresión, para ello, se ha seguido un procedimiento paso a paso, indicando:

- Cómo plantear una relación entre las variables con una formulación funcional y en una forma de ecuación estadística.
- Una descripción de cómo se hace para pasar de una variable categórica a otra variable formada por un bloque de variables ficticias.

- Cómo introducir un bloque de variables ficticias en un programa estadístico informático, el SPSS, con el objetivo de conseguir su significación estadística correspondiente.
- Cómo ajustar un modelo con el programa estadístico SPSS.
- Cómo escribir la mejor ecuación global ajustada a los datos.
- La ecuación general puede ser desglosada en el correspondiente número de grupos que contenga; se expone cómo hacerlo y cómo escribir una ecuación de pronóstico para cada grupo diferente.
- Cómo interpretar la ecuación general y cada una de las ecuaciones correspondiente a cada grupo.
- Cómo guardar los valores pronosticados de la variable dependiente.
- Cómo hacer una representación gráfica de conjunto, con una única figura para todos los grupos de la investigación.

Los autores de este manual sobre análisis multivariante, PRONÓSTICO CON INTERACCIÓN DE VARIABLES CATEGÓRICAS, desean agradecer a sus estudiantes la atención prestada a su enseñanza, y sobre todo en la asignatura Análisis Multivariante, donde se han utilizado estas notas como texto para sus clases; esperamos que también puedan ser usadas en los futuros cursos sobre «Técnicas de pronóstico y clasificación».

Los autores hemos procurado seguir, siempre que ha sido posible, las normas de publicación de la American Psychological Association (2010).

Estimado lector, esperamos que disfrutes leyendo y practicando estas páginas, tanto como los autores han disfrutado escribiéndolas.

Referencias

Aguinis, H. (2004). *Regression analysis for categorical moderators*. NY: Guilford Press.

Aiken, L. S., y West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

American Psychological Association (2010). *Publication manual of the APA* (6.ª ed.). Washington, DC: APA.

Hayes, A. F. (2013). *Introduction to mediation, moderation and conditional process analysis*. NY: Guilford.

Jaccard, J., Turrisi, R. y Wan, C. K. (2003). *Interaction effects in multiple regression.* (2ª ed.). Thousand Oaks, CA: Sage.

Jose, P.E. (2013). *Doing statistical mediation and moderation*. NY: Guilford.

Unidad 1. Regresión con una variable independiente de grupos (categórica)

Introducción

Asignar valores a los niveles de una variable categórica (o cualitativa), e incluirla con esta codificación en un análisis de regresión, es tanto como asumir que un valor es el doble del nivel anterior o el triple de dos niveles por encima. Es obvio que este modo de proceder establece valores inadecuados a la hora de incluir este tipo de variables. Esto no implica que no puedan utilizarse variables de este tipo, sino que es necesario que se realicen transformaciones que permitan detectar los diferentes niveles contenidos en las variables cualitativas. Esta solución es posible si transformamos la variable primitiva categórica en variables *dummy* (codificada con valores 1 o 0, según el caso, como explicaremos), en la que un valor cualquiera de dicha variable actúa como referencia del resto, eliminando de este modo la posible colinealidad.

Como ejemplo podríamos considerar la variable comunidad autónoma de procedencia. Concretamente, en España hay 19 comunidades a las que podríamos asignar los 19 primeros números naturales. Si consideramos esta codificación estaríamos asumiendo que la comunidad número 1 es la mitad que la que consideremos con el número 2, así hasta la región número 19, que sería 19 veces más que la primera.

La forma de darle a todos los niveles el mismo valor y que su intervención en la explicación de la variable dependiente sea como presencia o ausencia de cada nivel es transformando dichos niveles en *dummy*. Además, la regresión con variables *dummy* tiene una importante propiedad: la ecuación general de regresión se puede desglosar en la correspondiente ecuación para cada grupo, así si la variable independiente tiene tres niveles (grupos), se puede obtener una ecuación diferente para cada uno de los niveles.

> Una variable *dummy* es una variable binaria, nominal, dicotómica, categórica, que puede tomar únicamente los valores 0 o 1, indicando únicamente la presencia/ausencia del atributo medido.
> En este manual, las variables *dummy* vendrán siempre identificadas por la letra 'D_' al inicio del nombre de la variable.

Objetivos

Cuando el alumno finalice esta unidad sabrá:

1. Transformar las variables independientes cualitativas en variables *dummy*.

2. Incluir variables independientes con dos o tres niveles, transformadas en *dummy*, en una ecuación de regresión.

3. Valorar la ecuación de regresión y establecer la regresión para cada uno de los dos, tres o más grupos representados mediante las variables *dummy*.

4. Interpretar y ser capaz de aplicar los resultados obtenidos.

5. Generar el gráfico que permita comprobar el comportamiento de la variable dependiente para cada uno de los grupos considerados.

1.1. La variable independiente de grupos como variable *dummy*, creación de variables *dummy* para cada grupo

En un modelo de regresión clásica se ha de cumplir que la relación entre una variable independiente y la variable dependiente debe ser lineal: la diferencia de valor entre la variable dependiente en función del valor 5 y el de 6 de la variable independiente es igual al que hay entre 8 y 9 de la variable independiente; dicho de otro modo, siempre que la variable independiente tenga un intervalo de crecimiento de una unidad, el crecimiento de la variable dependiente ha de ser siempre el mismo.

En el caso de variables de grupos (o variables categóricas), esto no tiene por qué ser así; p. ej., si se tiene la variable lugar de nacimiento como variable independiente, y se codifica la provincia A con el valor de 1, la provincia B con el valor de 2 (y así sucesivamente), en regresión lineal con variables continuas se asume que el valor de la variable dependiente es proporcional al de los coeficientes, lo cual es absurdo en regresión con grupos, puesto que el orden de las provincias de nacimiento puede estar hecho por criterios de clasificación, y no implica orden ni equidistancia, sino pertenencia a un grupo.

En el desarrollo de esta unidad indicaremos cómo identificar cada grupo de una determinada variable mediante la creación de variables *dummy* (variables dicotómicas con valores 1 y 0). El motivo de esta codificación (1 y 0) es que el pronóstico de valores y la interpretación resultan mucho más fáciles que si se utiliza cualquier otra.

Hay un principio básico en la creación de variables *dummy*: se necesitan tantas variables *dummy* distintas como grupos haya en esa variable, menos una:

n.º Vs. *dummy* para una variable = (n.º grupos) – 1

Así, para la variable 'género' se necesita una sola variable *dummy*: 2 – 1 = 1.

Si se tiene una variable con 8 grupos, el número de variables *dummy* para identificarla será de 7 (8 – 1).

Si tomamos la variable género, las mujeres se pueden identificar mediante el valor de 1, por lo que los hombres pasarían a tener el valor de 0 (recuérdese que estos valores se usan para identificar a los grupos). No hace falta crear otra variable que identifique con 1 a los hombres y 0 a las mujeres (solo con una de las dos es suficiente).

IMPORTANTE:
El grupo identificado con el valor 0 en las categorías *dummy* es el llamado 'Grupo de Referencia'.

Tabla 1.1.
Ejemplo de transformación de valores de la variable 'Birthplace' (Lugar de nacimiento) en variables dummy.

Nombre de la Variable	Birthplace	D_BP_España	D_BP_Portugal	D_BP_Francia	D_BP_UK
Valores de la Variable	España	0	0	0	0
	Portugal	0	1	0	0
	Francia	0	0	1	0
	UK	0	0	0	1
	...		...	...	...

Si se tiene una variable con lugar de nacimiento ('Birthplace'), y la muestra tiene las siguientes categorías: España, Portugal, Francia y UK (4 en total), entonces solo se necesitan tres variables *dummy*. Se ha de decidir una categoría de referencia (en nuestro caso España). Las tres variables *dummy* serán: 'D_BP_Portugal', 'D_BP_Francia', y 'D_BP_UK', ver ejemplo en la Tabla 1.1.

Obsérvese que la pertenencia al grupo 'España' se sabe en las 'dummies' porque es la que tiene el valor de '0' en todas ellas.

De esta manera se evita: (a) el repetir información, pues se simplifica la información evitando redundancias, (b) la colinealidad estadística, pues si se incluye otra variable *dummy* referida a Spain, entonces, la suma de los valores *dummy* de los cinco países sería igual a 1 para todos los individuos (una de ellas sería combinación lineal de las otras), y la matriz **X'X** de la ecuación de regresión no tendría inversa (resultaría ser una matriz singular, pues su determinante sería igual a 0, y la matriz inversa no existe); recuérdese que los coeficientes de regresión son: $\mathbf{b = (X'X)^{-1}X'y}$.

1.2. Crear categorías *dummy* para variable de dos grupos

Por ejemplo, con los datos del fichero 'Company_data.sav' se desea comprobar si la variable 'Salary' (Salario) es función de la variable 'Gender' (Género), o estadísticamente hablando '*Salary = f(Gender)*'.

Hay un pequeño problema: la variable 'Gender' es de 'cadena', y el SPSS no permite en regresión trabajar con este tipo de variables. Para solventarlo, cambiaremos la variable 'Gender' a *dummy*; y si utilizamos como referencia el grupo 'Male' (Masculino), recodificaremos una nueva variable 'D_gndr_fem' ('*Dummy gender femenino*'), de modo que si en 'Gender' es 'Male', en 'D_gndr_fem' es '0', y si en 'Gender' es 'Female', en 'D_gndr_fem' es '1'. Este cálculo puede realizarse con la siguiente sintaxis:

```
RECODE gender ('m'=0) ('f'=1) INTO D_gndr_fem.
EXECUTE.
```

Conviene añadir en la variable 'Gender' la etiqueta: '*dummy female gender*' y las etiquetas a los valores 1 ('Female') y 0 ('Male'). De este modo, el fichero 'Company_data.sav' queda como en la Figura 1.1.

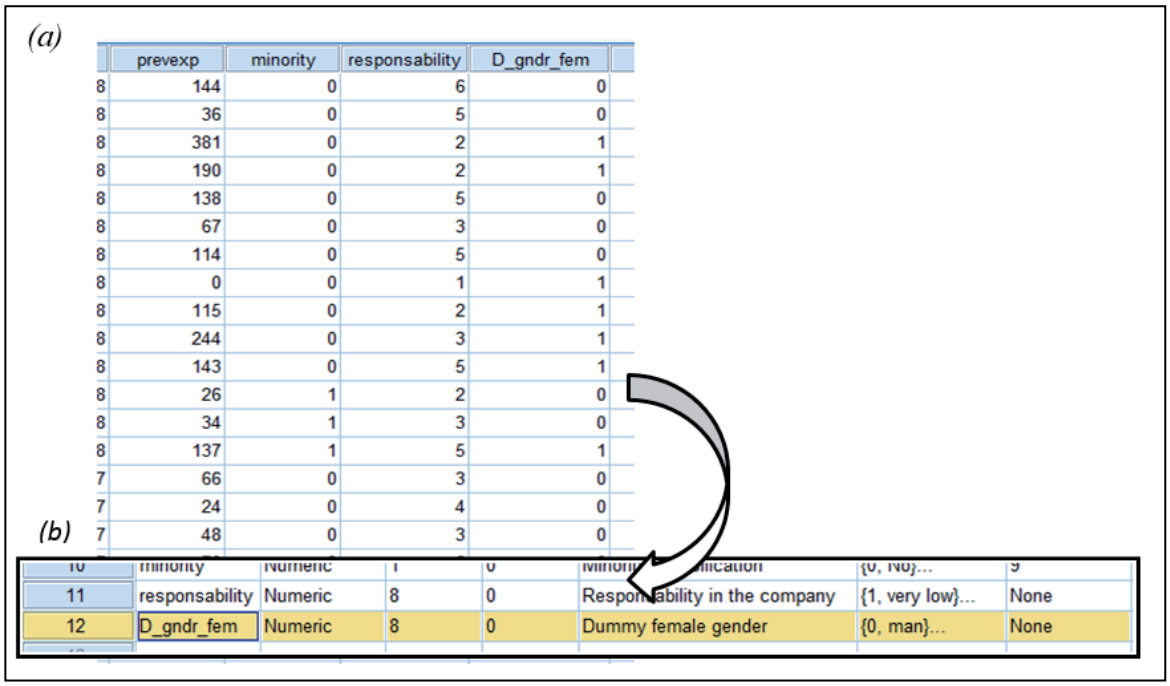


Figura 1.1. Ejemplo del fichero 'Company_data.sav', como queda al añadir la variable 'D_gndr_fem': *(a)* en el 'Data View' y *(b)* en el 'Variable View'

1.3. Regresión con variable de dos grupos. Ecuación de conjunto y ecuación para cada grupo. Interpretación. Representación gráfica

Ahora ya se puede hacer la regresión: *Salary = f(D_gndr_fem)*, pero antes hagamos una figura exploratoria de la relación entre ambas, mediante la sintaxis:

```
GRAPH
   /SCATTERPLOT(BIVAR)=D_gndr_fem   WITH   salary   BY
      gender
   /MISSING=LISTWISE.
```

Con la que se obtiene la Figura 1.2. En ella se incluye la línea de regresión entre ambas variables.

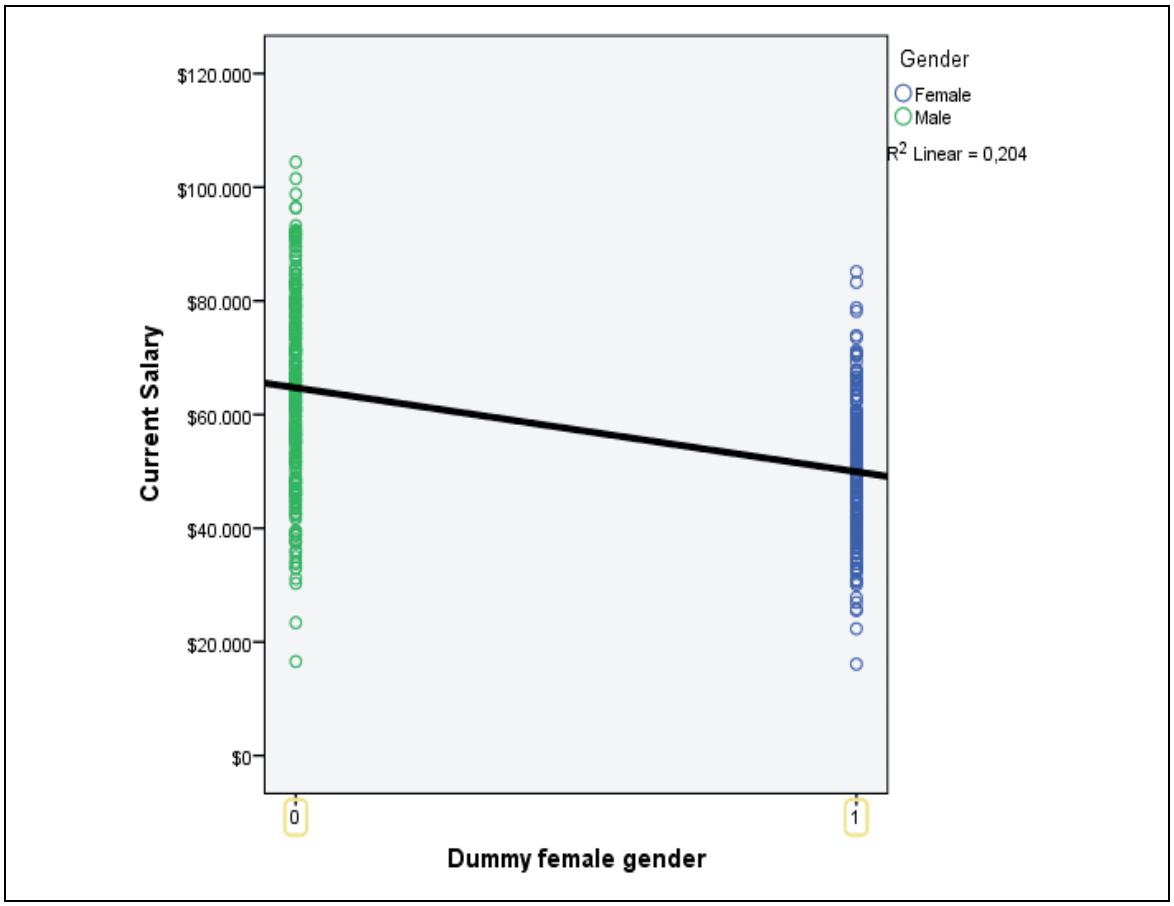


Figura 1.2. Relación entre la variable 'D_gndr_fem' (1: Mujer', 0: 'Hombre') y la variable 'Salary'

Para hacer la ecuación de regresión *Salary = f(D_gndr_fem)*, conviene guardar los valores predichos y los residuales, si se corre la sintaxis:

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA
```

```
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT salary
/METHOD=ENTER D_gndr_fem
/SAVE PRED RESID.
```

En el fichero de datos 'Company_data.sav' conviene guardar los valores predichos con el nombre '*pre_sal_f_D_gndr*' (recomendamos incluir una etiqueta que nos indique cómo hemos obtenido dichos valores predichos, por ejemplo: *Predicted salary = f(dummy_gender)*).

Los resultados obtenidos por el SPSS se incluyen en la Tabla 1.2, donde se ve que el valor de la $F(1,472) = 121.029$ ($p < .001$), por lo que toda la ecuación, en conjunto, explica significativamente, y la significación de la variable 'D_gndr_fem' es $p < .001$. Obsérvese que en una ecuación de regresión simple: $t^2 = F$ (t^2 de la variable independiente = F del conjunto de la ecuación: $-11.001^2 = 121.029$).

Tabla 1.2.
Resultados de la regresión de: Salary = f(Gender)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,452[a]	,204	,202	$14,557.154

a. Predictors: (Constant), D_gndr_fem
b. Dependent Variable: Current Salary

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25647398289	1	25647398289	121,029	,000[b]
	Residual	1,000E+11	472	211910734,1		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary
b. Predictors: (Constant), D_gndr_fem

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	64711,357	906,289		71,403	,000
	D_gndr_fem	-14769,783	1342,545	-,452	-11,001	,000

a. Dependent Variable: Current Salary

Siendo la ecuación de regresión:

$$Salary = 64711.357 - [14769.783 \cdot D\_gndr\_fem] + e \quad (1.1)$$

En este manual seguiremos la convención de poner entre corchetes, dentro de las ecuaciones, las correspondientes variables *dummy* referidas a variables simples o a bloques de variables de interacción; así en la Ecuación 1.1 aparece entre corchetes la variable *dummy* correspondiente a 'Gender'. Ahora bien, la variable 'D_gndr_fem' solo puede tomar el valor '0' si es 'Male' y el '1' si es 'Female', por lo que realmente se tienen dos ecuaciones en valores predichos de 'Salary' (*Salary'*), por tanto, la ecuación de pronóstico no contiene el término de error y, así, para los hombres (grupo de referencia: 'D_gndr_fem' = 0), substituyendo este valor en la Ecuación 1.1:

$$Salary'_M = 64711.357 - 14769.783 \cdot \mathbf{0} = 64711.357 \quad (1.2)$$

En nuestro caso, el grupo de referencia tiene como valor esperado el del término independiente de la ecuación de regresión ($b_0 = 64711.357 = Salary'_M$).

Y para las mujeres:

$$Salary'_F = 64711.357 - 14769.783 \cdot \mathbf{1} = 49941.574 \quad (1.3)$$

Es decir, los hombres ganan, por término medio: *64711.357* dólares al año, mientras las mujeres ganan el valor del término independiente más el del coeficiente de regresión (en este caso *-14769.783*): *49941.574.* La diferencia entre lo que ganan los hombres y las mujeres es estadísticamente significativa a un $\alpha = .05$.

Si se desea comprobar las medias 'Salary' para 'Male' and 'Female', mediante la sintaxis:

```
MEANS TABLES=salary BY gender
     /CELLS MEAN COUNT STDDEV.
```

Se obtienen los valores de la Tabla 1.3.

Tabla 1.3.
Medias de 'Salary' en función de 'Male' and 'Female'

Report

Current Salary

Gender	Mean	N	Std. Deviation
Female	$49,941.57	216	$11,919.638
Male	$64,711.36	258	$16,441.755
Total	$57,980.82	474	$16,299.863

Observar que los valores medios de la Tabla 1.3 coinciden con los valores predichos en las Ecuaciones 1.2 y 1.3, y con los puntos de corte (en valores de la variable dependiente: 'Current salary') en la Figura 1.2 entre la recta de regresión y los valores de X = 0 (Male) y X =1 (Female). Tal vez la figura más comprehensiva en SPSS sea mediante una figura de barras con la sintaxis:

```
GRAPH
     /BAR(SIMPLE)=MEAN(salary) BY D_gndr_fem.
```

Obteniéndose la Figura 1.3. Nótese que el resultado es el mismo que si se pide la figura con la media de los valores predichos:

```
GRAPH
     /BAR(SIMPLE)=MEAN(pre_sal_f_D_gndr) BY
          D_gndr_fem.
```

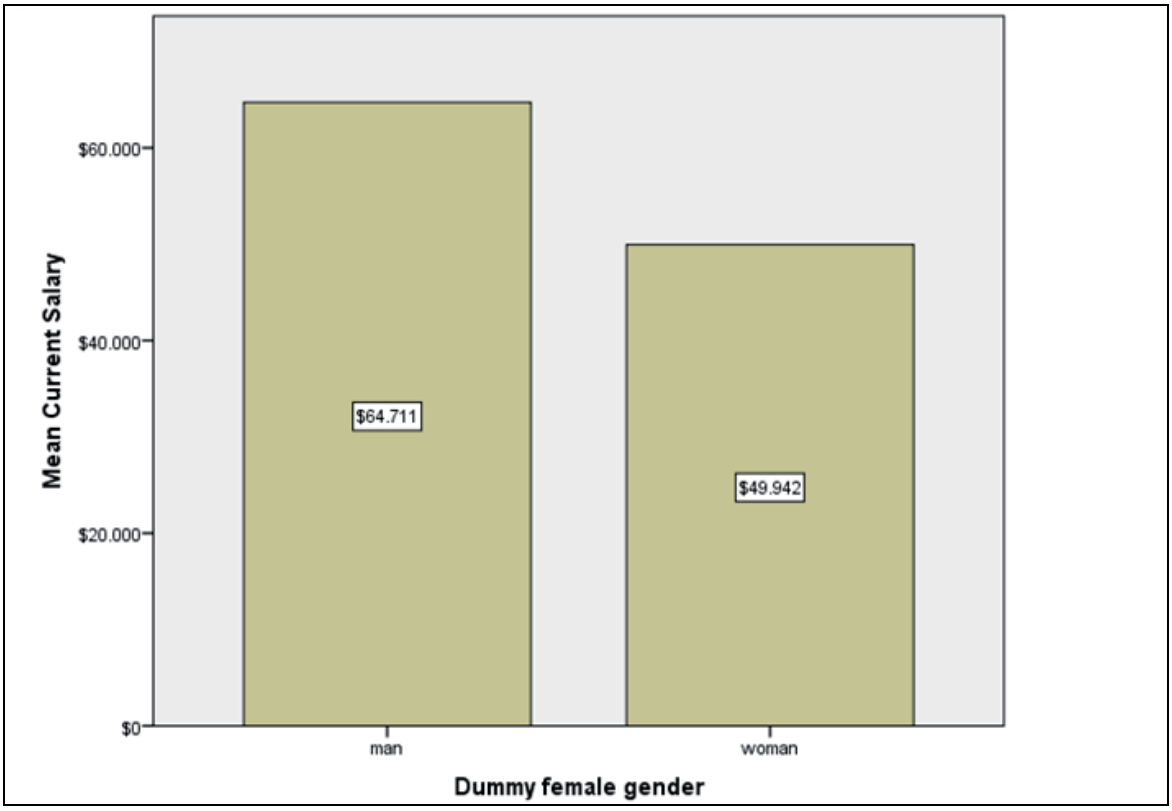


Figura 1.3. Media de los valores predichos de 'Salary' en función de 'Gender'

Como resumen general, se puede afirmar que la regresión con variables *dummy* equivale a una prueba de diferencia de medias, de este modo, el resultado estadísticamente significativo obtenido en la Ecuación 1.1 significa que hay diferencias significativas entre la media de salario actual de los 'Male' (*64711.357)* y la de los 'Female' (*49941.574*).

1.4. Creación de categorías *dummy* para variable de tres grupos

Supongamos que en el fichero 'Company_data.sav' se desea transformar la variable 'Jobcat' (Categoría Laboral) en una variable *dummy*, la variable 'Jobcat' tiene tres categorías ('Clerical' o administrativo, 'Custodial' o seguridad y 'Manager' o directivo), por lo tanto se necesitarán dos variables *dummy*, 'D_JC_custodial' y 'D_JC_manager', es decir, usaremos el grupo 'Clerical' como grupo de referencia.

Fácilmente se puede comprobar que si se hace mediante sintaxis:

```
RECODE jobcat (2=1) (ELSE=0) INTO D_JC_custodial.
    VARIABLE LABELS D_JC_custodial 'Custodial = 1,
            otherwise = 0'.
    EXECUTE.
```

Se crea la variable 'D_JC_custodial' con los valores '1' para la categoría 'Custodial', y a la vez contiene los valores '0' para todas las demás categorías ('Clerical' y 'Manager' en nuestro ejemplo).

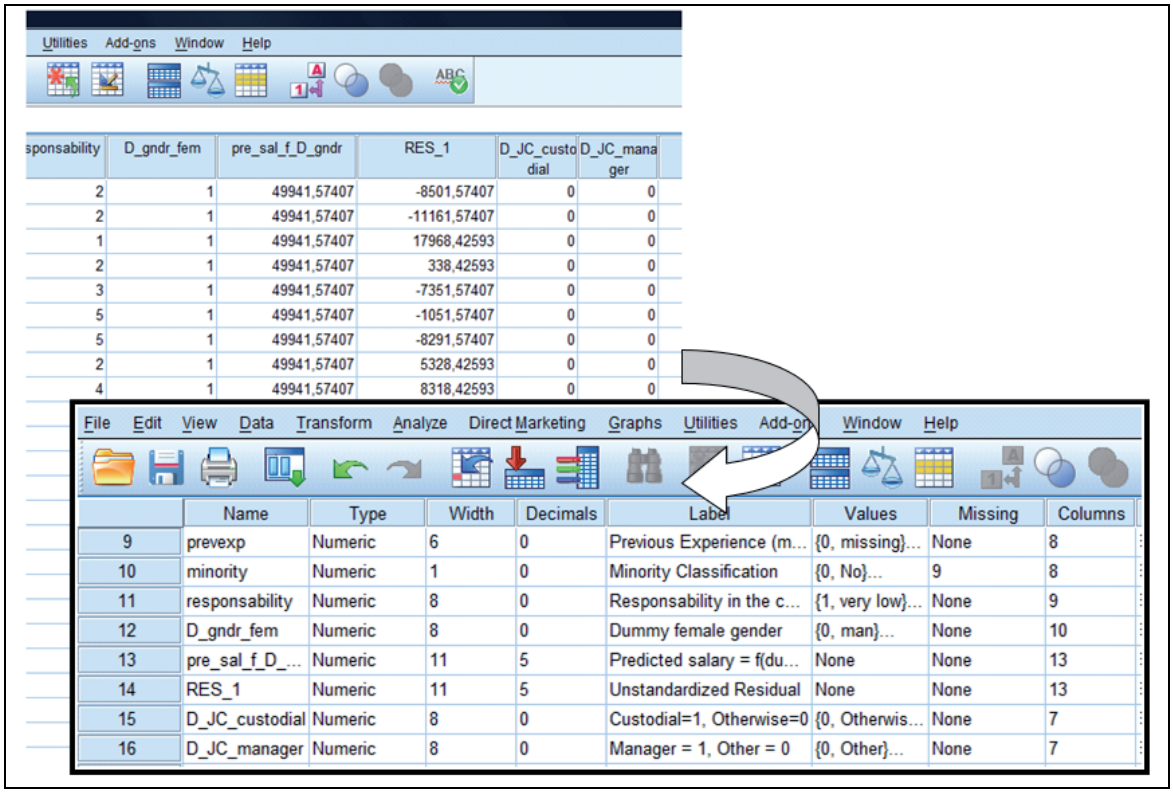


Figura 1.4. Ejemplo del fichero 'Company_data.sav', como queda al añadir las variables 'D_JC_custodial' y 'D_JC_manager'. En la parte superior se incluye la 'Vista de Datos' y en la inferior la 'Vista de Variables'

Del mismo modo, para crear la variable *dummy* 'D_JC_manager':

```
RECODE jobcat (3=1) (ELSE=0) INTO D_JC_manager.
```

```
VARIABLE LABELS D_JC_manager 'Manager = 1,
     otherwise = 0'.
EXECUTE.
```

Así, tendremos las dos nuevas variables añadidas en el 'Data View' y en el 'Variable View' del SPSS tal como aparecen en la Figura 1.4.

Con el fin de comprobar si las nuevas variables ('D_JC_custodial' y 'D_JC_manager') están bien recodificadas pedimos la 'Tabla de frecuencias' de 'Jobcat', y la 'Tabla Cruzada' de 'D_JC_custodial' con 'D_JC_manager':

```
FREQUENCIES VARIABLES=jobcat
     /ORDER=ANALYSIS.
CROSSTABS
     /TABLES=D_JC_custodial BY D_JC_manager
     /FORMAT=AVALUE TABLES
     /CELLS=COUNT
     /COUNT ROUND CELL.
```

Obteniéndose los valores de la Tabla 1.4.

Tabla 1.4.
Tabla de frecuencias de la variable 'Jobcat' y tabla cruzada de 'D_JC_custodial' con 'D_JC_manager'

Employment Category

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Clerical	363	76,6	76,6	76,6
	Custodial	27	5,7	5,7	82,3
	Manager	84	17,7	17,7	100,0
	Total	474	100,0	100,0	

Custodial = 1, otherwise = 0 * Manager = 1, Other = 0 Crosstabulation

Count

		Manager = 1, Other = 0		
		Other	Manager	Total
Custodial = 1, otherwise = 0	Other	363	84	447
	Custodial	27	0	27
Total		390	84	474

Obsérvese que en la Tabla 1.4 las frecuencias de ‘Clerical’ (363) en la tabla de frecuencias son las mismas que en ‘Otherwise’*’Other’ (363) en la ‘Crosstab’ (los valores ‘0’ y ‘0’ corresponden al grupo de referencia, ‘Clerical’), lo mismo ocurre con las frecuencias de las variables ‘Custodial’ (27) y ‘Manager’ (84), nótese como nadie es a la vez ‘Custodial’y ‘Manager’ (frecuencia = 0). Mediante este procedimiento se comprueba que la transformación de la variable ‘Jobcat’ en sus correspondientes variables *dummy* se ha hecho correctamente.

1.5. Regresión con variable de tres grupos. Ecuación de conjunto y ecuación para cada grupo. Interpretación. Representación gráfica

Una vez comprobado que tenemos el fichero de datos perfectamente recodificado con la variable 'Jobcat' en las nuevas variables 'D_JC_custodial' y 'D_JC_manager', ya podemos realizar la regresión de 'Salary' en función de la categoría laboral. Debido a que no se asume linealidad en la clasificación de 'Jobcat', utilizaremos como variables independientes: 'D_JC_custodial' y 'D_JC_manager'.

Haremos la Figura 1.5 exploratoria, mediante un 'Diagrama de Dispersión' de 'Salary' en función de 'Jobcat'.

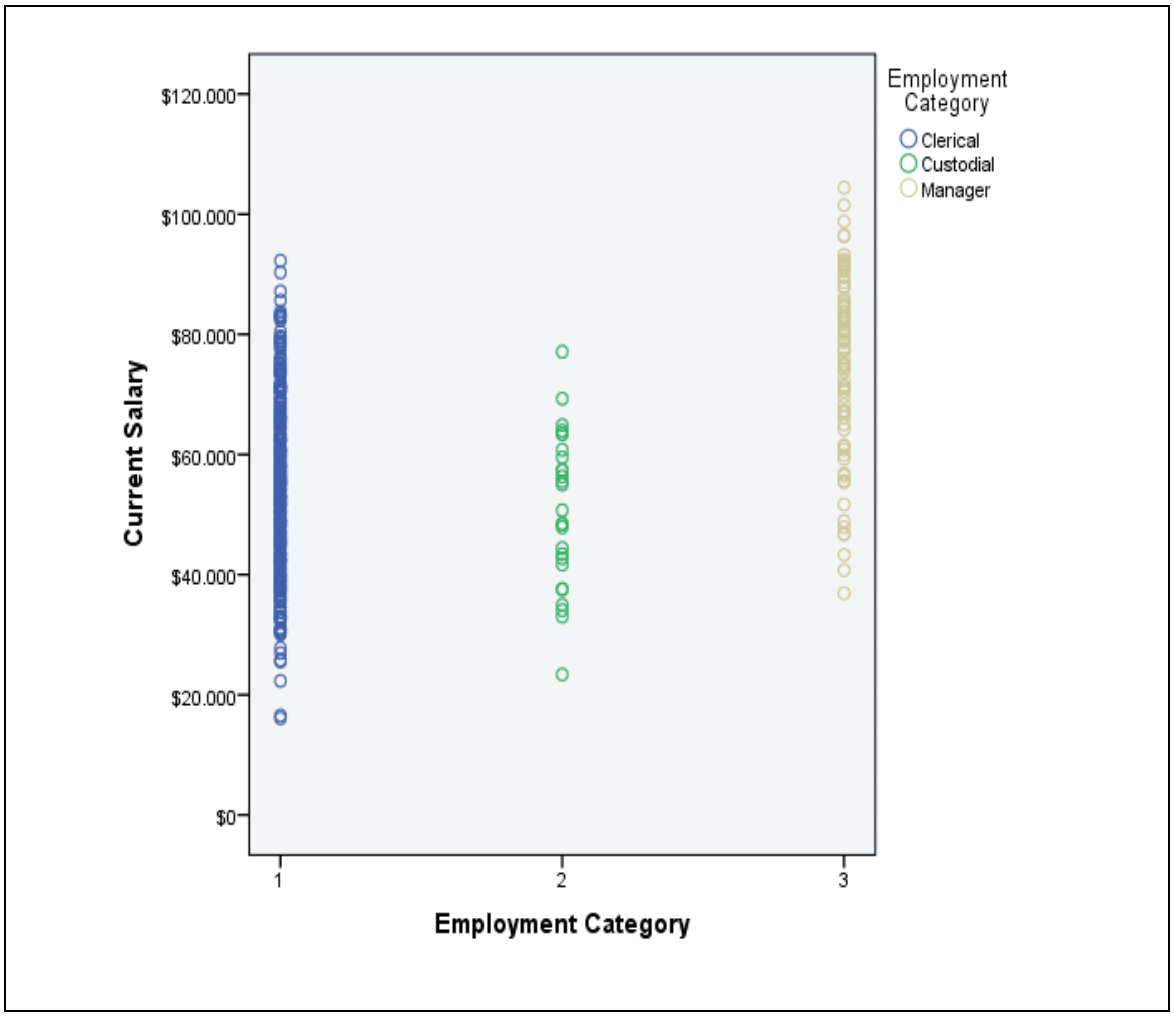


Figura 1.5. Diagrama de dispersión de 'Salary' en función de 'Jobcat' (1: 'Clerical', 2: 'Custodial', 3: 'Manager')

En la Figura 1.5 se observa que el grupo de menor media es el 'Custodial'.

Haciendo la regresión de 'Salary' = f('D_JC_custodial' y 'D_JC_manager'), en sintaxis (guardando los valores pronosticados y los errores de pronóstico):

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT salary
    /METHOD=ENTER D_JC_custodial D_JC_manager
    /SAVE PRED RESID.
```

Se obtienen los resultados de la Tabla 1.5.

Tabla 1.5.
Resultados de la regresión: 'Salary' = f('D_JC_custodial' y 'D_JC_manager')

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,517[a]	,267	,264	$13,984.809

a. Predictors: (Constant), Manager = 1, Other = 0, Custodial = 1, otherwise = 0

b. Dependent Variable: Current Salary

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33553494470	2	16776747235	85,782	,000[b]
	Residual	92115770309	471	195574883,9		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), Manager = 1, Other = 0, Custodial = 1, otherwise = 0

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	54336,942	734,012		74,027	,000
	Custodial = 1, otherwise = 0	-3506,942	2789,675	-,050	-1,257	,209
	Manager = 1, Other = 0	21689,129	1693,235	,509	12,809	,000

a. Dependent Variable: Current Salary

Una vez hecha la regresión, guardamos la variable pronosticada como: 'pre_sal_f_jobcat', y ponemos como etiqueta: 'Predicted Salary = f(jobcat)'.

Si observamos la significación de conjunto en la Tabla 1.5, con $F(2, 471) = 85.782$, $p < .001$, este resultado nos indica que la ecuación explica en conjunto de manera significativa los datos, pero como los valores de las variables son medias, equivale a indicar que hay diferencias significativas en la variable dependiente entre las medias de los grupos ('Clerical', 'Custodial' y 'Manager').

En cuanto al efecto del grupo 'Custodial', podemos observar que este no resulta significativo (*b = -3506.942, t = -1.257, p = .209*), indicando que la diferencia de salario entre el grupo 'Custodial' y el de referencia ('Clerical') no es significativa, pero este efecto debe dejarse en la ecuación, puesto que el de conjunto es significativo.

> En regresión de variables independientes formadas por grupos, si la variable independiente es significativa, han de dejarse todos los grupos, aunque haya alguno no significativo.

Del mismo modo, el efecto del grupo 'Manager' es significativo (*b = 21689.129, t = 12.809, p < .001*), indicando que hay diferencia estadísticamente significativa entre el salario del grupo 'Manager' y el de los 'Clerical'.

Siendo la ecuación de regresión general, en valores pronosticados:

$$Salary' = 54336.942 + [-3506.942 \cdot D\_JC\_custodial + 21689.129 \cdot D\_JC\_manager] \quad (1.4)$$

Y, como hemos dicho, es toda ella estadísticamente significativa.

La ecuación de pronóstico para cada grupo será:

(a) Para el grupo de 'Clerical' (en *D_JC_custodial = 0,* y en *D_JC_manager* = 0):

$$Salary'_{Cler} = 54336.942 + [-3506.942 \cdot \mathbf{0} + 21689.129 \cdot \mathbf{0}] = \mathbf{54336.942} \quad (1.5)$$

Como en la Ecuación 1.1, el grupo de referencia tiene como valor esperado el del término independiente de su ecuación de regresión (*54336.942*).

(b) Para el grupo 'Custodial' (en *D_JC_custodial = 1,* y en *D_JC_manager* = 0):

$$Salary'_{Cust} = 54336.942 + [-3506.942 \cdot \mathbf{1} + 21689.129 \cdot \mathbf{0}] = \mathbf{50830.000} \quad (1.6)$$

(c) Para el grupo 'Manager' (en *D_JC_custodial = 0,* y en *D_JC_manager* = 1):

$$Salary'_{Mngr} = 54336.942 + [-3506.942 \cdot \mathbf{0} + 21689.129 \cdot \mathbf{1}] = \mathbf{76026.071} \quad (1.7)$$

Que coincide con las medias de cada respectivo grupo (y con los valores pronosticados guardados en el fichero 'Company_data.sav'), como puede verse en la Tabla 1.6, mediante la sintaxis:

```
MEANS TABLES=salary BY jobcat
     /CELLS MEAN COUNT STDDEV.
```

Tabla 1.6.
Media de 'Salary' para cada grupo de 'Jobcat'

Report

Current Salary

Employment Category	Mean	N	Std. Deviation
Clerical	$54,336.94	363	$13,822.362
Custodial	$50,830.00	27	$12,814.845
Manager	$76,026.07	84	$15,003.277
Total	$57,980.82	474	$16,299.863

La representación de las medias es la de la Figura 1.6, obtenida a partir de la sintaxis:

```
GRAPH
     /BAR(SIMPLE)=MEAN(salary) BY jobcat.
```

Con el fin de comprobar las medias de cada valor pronosticado, se puede pedir la tabla de medias mediante la sintaxis:

```
MEANS TABLES=pre_sal_f_jobcat BY jobcat
     /CELLS MEAN STDDEV COUNT.
```

En resumen, la regresión con una variable independiente de tres grupos, transformada en sus dos correspondientes variables *dummy*, nos permite comprobar si existen globalmente, diferencias de medias entre los grupos.

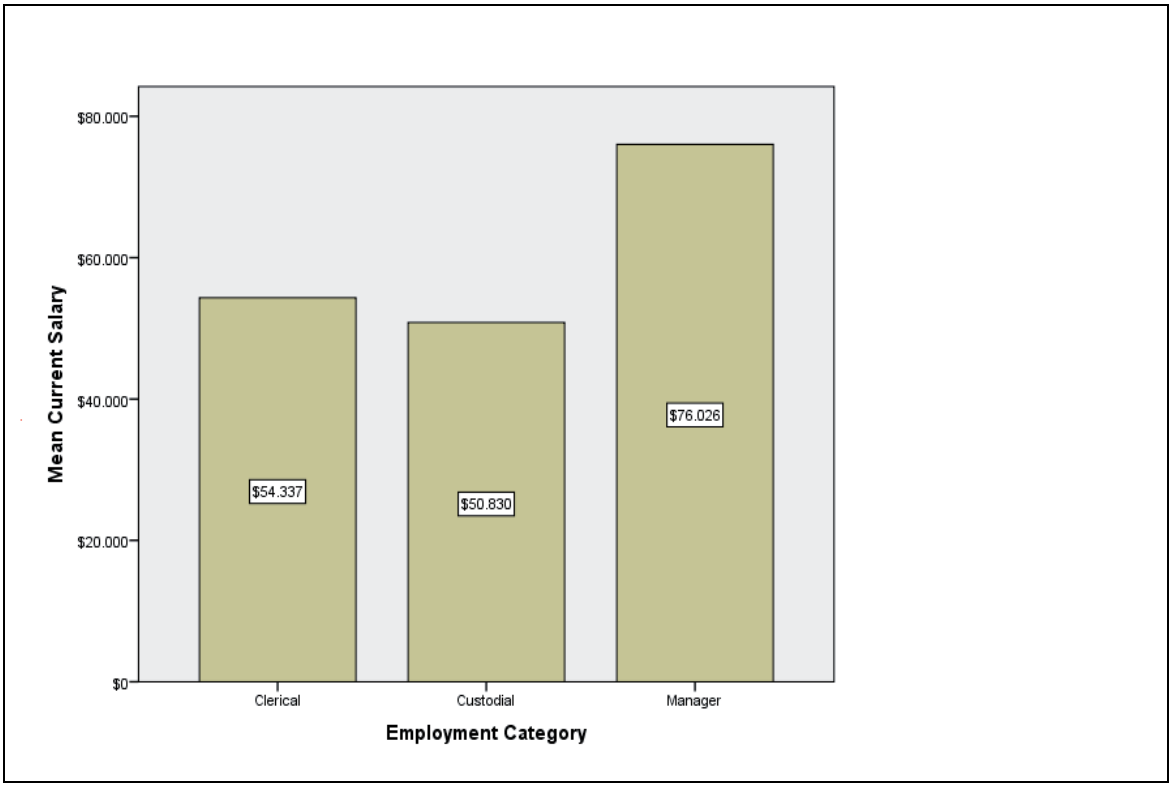


Figura 1.6. Medias de 'Salary' en función de 'Jobcat'

1.6. Comparaciones «a posteriori» de dos grupos en variables independientes de tres o más grupos

En el caso de que se estuviese interesado en la comparación de dos grupos, siempre que uno de ellos no sea el grupo de referencia, podría calcularse mediante uno de estos dos procedimientos: (a) poniendo uno de los grupos de comparación como grupo de referencia, y volviendo a correr la regresión con las nuevas *dummy*, y (b) por la fórmula (Hardy, 1993):

$$t = \frac{(b_j - b_k)}{\sqrt{\left(Var_{b_j} + Var_{b_k} - 2Cov_{b_j b_k}\right)}} \qquad 1.8)$$

Siendo b_j y b_k los respectivos coeficientes de regresión, $Var(b_j)$ y $Var(b_k)$ las varianzas de los coeficientes y $Cov(b_j b_k)$ la covarianza entre ambos.

En nuestro caso, cabe peguntarse si hay diferencia entre la media de salario del grupo 'Custodial' y el 'Manager', para ello tendríamos que aplicar la Ecuación 1.8 y buscar en unas tablas de *t* el resultado, con el fin de comprobar si es significativo o no el resultado obtenido.

Un sistema más complejo de comparación de medias, desde una perspectiva del análisis de la varianza, se puede encontrar en Kirk (2012), en Edwards (1985) y en Pedhazur y Pedhazur (1991). Los sistemas de estos manuales son más adecuados y de más amplio espectro, puesto que también sirven para comparar un conjunto de medias versus otro conjunto de medias, mediante comparaciones «a la medida», pero la ampliación de estos apartados se sale del objetivo del presente manual.

1.7. Actividad dirigida: Crear categorías *dummy* para variable de cuatro o más grupos. Ecuación de conjunto y ecuación para cada grupo. Representación gráfica

Cuestiones:

Con los datos del fichero 'Company_data.sav' se desea hacer la regresión 'Salary' = f('Center'), para ello, seguir los siguientes pasos,

- Transforma la variable categórica 'Center', que tiene 6 categorías (1: 'Center 1', 2: 'Center 2', … y 6: 'Center 6'), en 5 variables *dummy* usando como grupo de referencia el grupo 'Center 1'. Tendrás que hacer 5 variables *dummy*, llámalas: 'D_Ctr_2', 'D_Ctr_3', 'D_ Ctr_4', 'D_ Ctr_5' y 'D_ Ctr_6' (que son los acrónimos de '*Dummy*_Center_2', '*Dummy*_Center_3', ..., y '*Dummy*_Center_6').
- Comprueba que haces bien las anteriores transformaciones, para ello, pide una tabla de frecuencias de 'Center', y de 'D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6', y comprueba que las frecuencias se corresponden correctamente.
- Haz la ecuación de regresión de 'Salary' = f('Center'), usando las 5 variables *dummy* como Variables independientes: 'D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6'.
- Comenta la significación del estadístico F. Interpreta los resultados de conjunto.
- Escribe la ecuación de regresión de conjunto de 'Salary' = f('Center').
- Escribe las ecuaciones para cada Centro. Interpreta cada una de ellas (¿difiere la media del grupo respecto de la del grupo de referencia?).
- Haz una figura de conjunto que represente los resultados.

Pautas para la correcta realización:

(a) Corre la sintaxis:

```
RECODE Center (2=1) (ELSE=0) INTO D_Ctr_2.
     VARIABLE LABELS D_Ctr_2 'Low Center '.
RECODE Center (3=1) (ELSE=0) INTO D_Ctr_3.
     VARIABLE LABELS D_Ctr_3 ' Medium low Center'.
RECODE Center (4=1) (ELSE=0) INTO D_Ctr_4.
     VARIABLE LABELS D_Ctr_4 'Medium high Center'.
RECODE Center (5=1) (ELSE=0) INTO D_Ctr_5.
```

```
     VARIABLE LABELS D_Ctr_5 'High Center '.
RECODE Center (6=1) (ELSE=0) INTO D_Ctr_6.
     VARIABLE LABELS D_Ctr_6 'Very high Center'.
EXECUTE.
```

(b) Escribir y correr la sintaxis:

```
FREQUENCIES VARIABLES=Center D_Ctr_2 D_Ctr_3 D_Ctr_4
D_Ctr_5 D_Ctr_6.
```

(c) Sintaxis:

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA
     /CRITERIA=PIN(.05) POUT(.10)
     /NOORIGIN
     /DEPENDENT salary
     /METHOD=ENTER D_Ctr_2 D_Ctr_3 D_Ctr_4 D_Ctr_5
          D_Ctr_6
     /SAVE PRED RESID.
```

(d) Se obtienen los resultados de conjunto:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,598[a]	,358	,351	$13,133.276

a. Predictors: (Constant), Center 6 , Center 3, Center 2, Center 5, Center 4

b. Dependent Variable: Current Salary

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44947243868	5	8989448774	52,118	,000[b]
	Residual	80722020911	468	172482950,7		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), Center 6 , Center 3, Center 2, Center 5, Center 4

(g) Sintaxis:

```
GRAPH
     /BAR(SIMPLE)=MEAN(salary) BY Center.
```

ATENCIÓN: Si lo deseas, puedes guardar el fichero con las variables añadidas (las *dummy* creadas y las generadas con los pronósticos y los residuales de la regresión); llama a este fichero 'Company_data_Unit1.sav', también puedes guardar el fichero de sintaxis y los resultados con las figuras.

IMPORTANTE
Guardar el fichero de datos con las variables *dummy* generadas para 'Gender' ('D_gndr_fem'), para 'Jobcat' ('D_JC_custodial' y 'D_JC_manager'), y para 'Center' ('D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6'), para ello solo hay que borrar las variables pronosticadas y sus residuales calculados mediante la regresión; **llámalo *'Comp_dat_Dms.sav'*** (acrónimo de '*Comp*any *dat*a with *D*u*m*mie*s*'), así tendrás las variables *dummy* que se han usado en esta unidad ya guardadas para cuando hagas otros análisis.

1.8. Adenda 1: Relación entre la prueba 't' de Student-Fisher, la prueba F del ANOVA de un factor y la regresión con una variable de grupos

La idea de esta 'Adenda' es mostrar cómo el análisis de la varianza (ANOVA) de un factor (o una variable independiente) es coincidente con la regresión en la que la variable independiente está formada por grupos. Este apartado lo dividiremos en dos subapartados: (a) para una variable independiente de dos grupos, (b) una variable independiente de tres o más grupos.

(a) En el caso de que se haga una prueba 't' de Student-Fisher para dos grupos, por ejemplo, 'Salary' en función de 'Gender', a partir de la sintaxis:

```
T-TEST GROUPS=gender('f' 'm')
     /MISSING=ANALYSIS
     /VARIABLES=salary.
```

Se obtienen los resultados de la Tabla 1.7.

Tabla 1.7.
Resultados de la prueba de 't' de Student-Fisher de 'Salary' en función de 'Gender'

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Current Salary	Equal variances assumed	25,871	,000	-11,001	472	,000	-$14,769.783	$1,342.545	-$17,407.887	-$12,131.678
	Equal variances not assumed			-11,309	462,894	,000	-$14,769.783	$1,305.972	-$17,336.151	-$12,203.414

Que coinciden con los ya vistos en el apartado 3, Tabla 1.2, de la que seleccionamos el estadístico *'t': -11.001*, como se muestra en las Tablas 1.7 y 1.8.

Tabla 1.8.
Resultados del análisis de regresión de 'Salary' en función de 'Gender' transformada en 'dummy'

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	64711,357	906,289		71,403	,000
	D_gndr_fem	-14769,783	1342,545	-,452	-11,001	,000

a. Dependent Variable: Current Salary

Se observa que en ambos casos (el del análisis de regresión y el de la comparación de dos medias) el valor de la 't' de Student-Fisher coincide; de la misma manera, también coincide el valor de la probabilidad del estadístico 't' ($p < .001$). En la Tabla 1.6 aparece el valor F de 'Levene's Test for Equality of Variances' ($F = 25.871$, $df = 1$,

472, p < .001) lo cual significa que los dos grupos de la variable 'Gender' ('Male' and 'Female') difieren en varianzas (es decir, en sus perfiles de variación, un grupo tiene mayor varianza, que el otro); en contrapartida, el test 't' o el ANOVA comparan las medias de los respectivos grupos teniendo en cuenta la variabilidad de los datos.

Si pedimos el ANOVA de 'Salary' en función de 'Gender', por medio de la sintaxis:

```
ONEWAY salary BY D_gndr_fem
     /MISSING ANALYSIS.
```

Se obtienen los resultados de la Tabla 1.9.

Tabla 1.9.
ANOVA de 'Salary' en función de 'Gender'

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	25647398289	1	25647398289	121,029	,000
Within Groups	1,000E+11	472	211910734,1		
Total	1,257E+11	473			

Si se comparan estos resultados con los de la Tabla 1.2, se observa que son los mismos.

En resumen, la significación estadística de la regresión de una variable independiente de dos grupos es coincidente con la del estadístico 't' y la del ANOVA, además, se comprueba que $t^2 = F$.

(b) En el caso de una variable independiente de tres o más grupos, los resultados de conjunto de la regresión y los del ANOVA son los mismos.

Veíamos que al hacer la actividad 1, apartado (d), al pedir 'Salary' = f('Center'), lo correcto es llevar a cabo la regresión con variables *dummy*:

'Salary' = f('D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5', 'D_Ctr_6'),

que da el resultado de la Tabla 1.10.

Tabla 1.10.
Tabla del ANOVA de la regresión de 'Salary' en función de 'D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5', 'D_Ctr_6'

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44947243868	5	8989448774	52,118	,000[b]
	Residual	80722020911	468	172482950,7		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), Center 6 , Center 3, Center 2, Center 5, Center 4

Y si se hace el ANOVA de 'Salary' en función de 'Center', mediante la sintaxis:

```
ONEWAY salary BY Center
     /MISSING ANALYSIS.
```

se obtiene la Tabla 1.11.

Tabla 1.11.
Resultados del ANOVA de 'Salary' en función de 'Center'

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	44947243868	5	8989448774	52,118	,000
Within Groups	80722020911	468	172482950,7		
Total	1,257E+11	473			

Se comprueba que los resultados de la regresión de 'Salary' en función de 'Center', mediante las cinco variables *dummy* (*'D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5', 'D_Ctr_6'*) es lo mismo que el ANOVA de un factor de 'Salary' en función de 'Center', y ello es debido a que el programa SPSS automáticamente establece internamente las categorías *dummy* correspondientes a la variable independiente 'Center' y da los resultados finales.

Resumiendo, la ventaja de hacer comparaciones de medias mediante el análisis de regresión (si la variable independiente tiene más de tres grupos) es que el analista «diseña» sus propios valores de variable independiente y puede pronosticar de manera «transparente» los valores de cada grupo, pero tiene el inconveniente de que ha de dedicar tiempo a elaborar las variables *dummy* de cada grupo; en contrapartida, las comparaciones de medias mediante el análisis de la varianza (cuando la variable

independiente tiene más de tres grupos) tiene la gran ventaja de que es más rápido (el SPSS hace automáticamente su propio sistema de codificación), pero para pronosticar valores el sistema es más «opaco».

Desde una perspectiva histórica, es interesante reseñar que Karl Pearson, quien vivió de 1857 a 1936 (uno de cuyos trabajos más representativos sobre la regresión es el del año 1896), desarrolló la teoría de la regresión estadística en un contexto de investigación «de campo», midiendo eventos tal como se daban en contextos psicológicos, sociales o naturales; su interés era detectar la magnitud de los coeficientes de regresión ($b_0, b_1, \ldots, b_k$) y su correspondiente significación estadística (mediante la probabilidad del estadístico '*t*', para ello desarrolló la teoría de los tests de hipótesis estadísticas); mientras que la teoría del ANOVA fue elaborada por Ronald Fisher (1890-1962), quien trabajaba en un contexto de investigación «experimental», comprobando la efectividad de tratamientos para mejora de resultados, sus desarrollos estadísticos sobre el ANOVA se produjeron en las publicaciones de los años 1925 y 1935, donde elaboró el sistema de la partición de la varianza mediante las sumas de cuadrados y al consiguiente cálculo del estadístico de conjunto *F*.

Otros estadísticos (Mann, 1945; Cohen, 1968; Tatsuoka, 1975) han mostrado que ambos modelos tienen una raíz común, puesto que en el ANOVA de Fisher los efectos de las variables de grupo (categóricas, experimentales…) se resuelven mediante el procedimiento del análisis de la regresión (introduciendo variables categóricas para cada grupo), y que también en el análisis de regresión de Pearson se estima la significación de conjunto mediante el cálculo del estadístico *F* de Fisher.

Por otro lado, el ANOVA fue desarrollado para la realización de experimentos planificados en los cuales hay doble sistema de azar: de los sujetos de la muestra respecto de la población, y de la asignación de los sujetos de la muestra a los respectivos tratamientos; mientras que en la regresión lo más frecuente es tomar una muestra «incidental» accesible al investigador, si bien lo ideal es que los sujetos fuesen también seleccionados al azar y, sin embargo, esto en muchos estudios «de campo» no puede llevarse a cabo por razones obvias: los sujetos ya están autoseleccionados (por la institución a la que pertenecen, o por las variables implícitas a los mismos...), pero como principio general, en estudios de regresión tómese una muestra completa (en este caso la población objeto de estudio) o una muestra representativa al azar.

Estos dos submodelos estadísticos (el de la regresión y el del ANOVA) han servido a veces de justificación para defender que en ciencias humanas, sociales y biológicas, hay dos grandes tendencias de investigación: la de campo y la experimental (véase Cronbach, 1957, 1975), es precisamente el mismo Cronbach quien propugna la unificación y la integración de ambos enfoques mediante el modelo de interacción en regresión.

1.9. Adenda 2: Regresión con variable de tres o más grupos ¿regresión simple o múltiple?

Al realizar la ecuación de regresión de una variable dependiente sobre otra variable independiente de con tres o más grupos, se observa la paradoja que una ecuación de regresión simple pasa a ser de regresión múltiple, pues al transformar la variable original de grupos en otras variables *dummy*, la regresión pasa a ser múltiple (pues ya hay dos o más variables independientes).

Por ejemplo, revisando lo hecho hasta ahora, al hacer la ecuación de regresión de 'Salary' en función de 'Jobcat', se tienen tres categorías ('Clerical', 'Custodial' y 'Manager'), por lo tanto se necesitarán dos variables *dummy* ('D_JC_custodial' y 'D_JC_manager'), es decir, usaremos el grupo 'Clerical' como grupo de referencia, si se desea realizar la ecuación de regresión: 'Salary' = f('Jobcat'), ha de llevarse a cabo la siguiente transformación:

$$Salary = b0 + b1 \cdot Jobcat + e$$
$$[NO\ CORRECTO]$$

$$Salary = b0 + [b1 \cdot D\_JC\_custodial + b2 \cdot D\_JC\_manager] + e \quad (1.9)$$
$$[CORRECTO]$$

Volvemos a insistir en el hecho de que en la Ecuación 1.9 se escriben dentro de corchetes las variables *dummy* correspondientes a la variable de grupos 'Jobcat'.

Y en el caso de la regresión de 'Salary' en función de 'Center' (con seis categorías: 'Very low', 'Low', 'Medium low', 'Medium high', 'High' y 'Very high'), para la que se hacen cinco variables *dummy*: 'D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5', 'D_Ctr_6' (siendo 'Very low' la categoría de referencia), para estimar los valores de la ecuación de regresión: 'Salary' = f('Center'), habría de hacerse:

$$Salary = b0 + b1 \cdot Center + e$$
$$[NO\ CORRECTO]$$

$$Salary = b0 + [b1 \cdot D\_Ctr\_2 + b2 \cdot D\_Ctr\_3 + b3 \cdot D\_Ctr\_4 + b4 \cdot D\_Ctr\_5 + b5 \cdot D\_Ctr\_6] + e \quad (1.10)$$
$$[CORRECTO]$$

Es decir, en la Ecuación 1.9, la ecuación simple se ha desarrollado con dos variables independientes, mientras que en la Ecuación 1.10 han sido cinco las variables independientes utilizadas, con lo cual, una aparente regresión simple se ha transformado en otra ecuación de regresión múltiple.

Ahora bien, tanto en la Ecuación 1.9 como en la 1.10, no interesa la significación de un determinado efecto de grupo (comparando este grupo con el de referencia), sino que lo más importante es el efecto conjunto de todos ellos, para lo cual solo se ha de

comprobar la significación de la *F* global, respondiendo a la pregunta ¿hay diferencias significativas, en conjunto, entre las medias de la variable dependiente, en función de los grupos?, lo cual también equivale a preguntar ¿hay por lo menos un grupo que difiere estadísticamente en su media de la de los demás grupos?, es decir, para que la *F* de conjunto sea significativa, por lo menos un grupo (en media de variable dependiente) ha de diferir de otro grupo (en su respectiva media de la variable dependiente). En resumen, la regresión con una variable independiente de tres o más grupos, es una regresión simple en su planteamiento, pero es como una regresión múltiple en su resolución.

1.10. Conclusiones

En esta unidad, el lector habrá aprendido:

- En regresión de una variable dependiente sobre otra variable independiente de grupos, se ha de hacer variables *dummy* para los grupos de la variable independiente, tanto si la variable independiente tiene dos grupos, o tres o más grupos.
- Lo que significa el 'grupo de referencia' de una variable independiente con 'n' grupos, y la necesidad de hacer tantas variables *dummy* como grupos menos una unidad (n.º 'dummies' = n.º grupos – 1).
- Llevar a cabo la correspondiente ecuación de regresión con variables *dummy*.
- Interpretar adecuadamente el estadístico F del ANOVA de conjunto y cada uno de los coeficientes mediante la probabilidad del estadístico 't' de cada coeficiente.
- Representar la ecuación global, para todos los grupos, y la ecuación de regresión para cada uno de ellos. Saber interpretarlo.
- Hacer una representación gráfica de conjunto para los pronósticos de la regresión por grupos.
- La regresión simple, en la que hay una sola variable independiente, con grupos, o variables categóricas, tiene cuatro propiedades importantes:

  (a) si la variable independiente solo tiene dos grupos, analíticamente, coincide con las pruebas de comparación de medias: de 't' de Student-Fisher, o con la 'F' del ANOVA (para dos o más grupos).

  (b) si la variable independiente es de dos grupos, entonces el valor de la 'p' es la misma en la regresión y en el contraste de medias, y se observa que $t^2 = F$.

  (c) los valores de pronóstico para cada grupo son las medias de cada respectivo grupo (solo se pronostican niveles).

  (d) el ANOVA es un tipo especial de regresión, con variables *dummy*, aunque hay otros tipos de codificación para la identificación de los grupos (por los efectos y el ortogonal).

Lecturas recomendadas

El libro más asequible para poder actualizarse en esta unidad es el de Hardy (1993), capítulos dos y tres, en él aparecen ejemplos desarrollados, y es de fácil lectura.

Para otros sistemas de codificación de grupos (además del *dummy*) y de contrastes de medias, ya hemos hablado del libro de Hardy (1993) en sus capítulos 4 y 5, el de Kirk (2012) en el capítulo 7, el de Edwards (1985) en el capítulo 25, y el de Pedhazur y Pedhazur (1991) en el capítulo 19.

Bibliografía

Cohen J. (1968). «Multiple regression as a general data-analytic system». *Psychollogical Bulletin, 70*, 426-443.

Cronbach, L. J. (1957). «The two disciplines of scientific psychology». *American Psychologist*, *12*, 671-684. [http://psychclassics.yorku.ca/Cronbach/Disciplines/]

—— (1975). «Beyond the two disciplines of scientific psychology». *American Psychologist, 30*, 671-84.

Edwards, A. L. (1985). *Experimental Design in Psychological Research*. Nueva York: Harper and Row.

Fisher, R. (1925). *Statistical methods for research workers.* Edimburgo: Oliver y Boyd. [http://psy.ed.asu.edu/~classics/Fisher/Methods/]

—— (1935). *The design of experiments.* Edimburgo: Oliver y Boyd.

Hardy, M. A. (1993). *Regression with dummy variables.* Newbury Prak, CA: Sage.

Kirk, R. E. (2012). *Experimental design: procedures for the behavioral sciences*. Londres: Sage.

Mann, H. B. (1949). *Analysis and design of experiments: Analysis of variance and analysis of variance designs*. Nueva York, N. Y.: Dover Publications.

Pearson, K. (1896). «Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia». *Philosophical Transactions of the Royal Society of London*, *187*, 253-318.

Pedhazur, E. J.; Pedhazur, L. (1991). *Measurement, design*, *and analysis*: *An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.

Tatsuoka, M. M. (1975). *The general linear model: A 'new' trend in analysis of variance*. Champaign, IL: The Institute for Personality and Ability Testing.

Actividades

1. Con los datos del fichero 'Comp_dat_Dms.sav' se desea hacer la regresión 'Salbegin' = f('Center'), para ello, llevar a cabo los siguientes pasos:

- Haz la ecuación de regresión de 'Salbegin' = f('Center'), usando correctamente las cinco variables *dummy* como variables independientes: 'D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6'.
- Comenta la significación del estadístico F de conjunto. Interpreta los resultados de conjunto.
- Escribe la ecuación de regresión de conjunto de 'Salbegin' = f('Center'), con sus correspondientes variables *dummy*.
- Escribe las ecuaciones para cada grupo. Comenta cada una de ellas (¿difiere la media de un determinado grupo respecto de la del grupo de referencia?).
- Haz una figura de conjunto que represente los resultados.

2. Se desea hacer la ecuación de regresión lineal de 'Salary' = f('Minority'). Como verás, 'Minority' tiene dos categorías: 0: 'No minority' y 1: 'Yes minority'. Observa que ya es *dummy*. Responde:

- Haz la ecuación de regresión de 'Salary' = f('Minority').
- Comenta la significación del estadístico F de conjunto. Interpreta los resultados de conjunto.
- Escribe la ecuación de regresión de conjunto de 'Salary' = f('Minority'), con sus correspondientes variables *dummy*.
- Escribe las ecuaciones para cada grupo. Comenta cada una de ellas (¿difiere la media del grupo respecto de la del grupo de referencia?).
- Haz una figura de conjunto que represente los resultados.

Unidad 2. Regresión con dos variables independientes de grupos (categóricas), solo 'efectos principales'

Introducción

En esta unidad haremos la regresión de una variable dependiente sobre dos variables independientes de grupos (categóricas), pero utilizando solo los efectos principales de las dos variables independientes. Este sistema tiene dos inconvenientes:

- La diferencia de medias en la variable dependiente entre subcategorías es siempre la misma (posteriormente veremos con más detalle este aspecto en una figura).
- Los valores pronosticados no suelen reproducir las medias reales de los respectivos grupos.

Estos inconvenientes son debidos a que solo se utilizan los efectos principales de cada variable, sin calcular la interacción entre las mismas; en la Unidad 3 estudiaremos cómo resolver estos inconvenientes incluyendo la interacción de variables.

El lector ha de saber previamente cómo transformar variables categóricas en variables *dummy*, tomando uno de los grupos como referencia, así como realizar la regresión mediante el sistema de 'bloques'.

Una vez resueltos estos aspectos, se realizará la ecuación de regresión lineal múltiple en la que interviene la variable independiente de tres niveles como un solo 'bloque', obteniendo así la ecuación de conjunto y la ecuación correspondiente para cada grupo. Posteriormente veremos la interpretación detallada de la ecuación de conjunto, así como de los diferentes estimadores obtenidos. Se comprobará asimismo la representación gráfica sin interacción y quedará patente la necesidad de considerar la interacción.

Objetivos

Cuando el estudiante finalice esta unidad sabrá:

- Organizar ‘bloques’ mediante codificación *dummy*.
- Plantear la ecuación de regresión considerando los ‘bloques’, estimar sus parámetros, interpretar tanto de los estadísticos de conjunto de cada modelo con variables dummies, como la significación del cambio de R^2 para cada bloque de variables de grupos.
- Interpretar los coeficientes de la ecuación de regresión, y escribir la ecuación global de regresión, y los valores pronosticados para cada grupo. Asimismo, representar gráficamente los resultados obtenidos.
- Relacionar el análisis de regresión con variables independientes principales de grupos y el análisis ANOVA con las mismas variables, así como las ventajas e inconvenientes del uso de la regresión y del ANOVA (solo con variables principales) para variables independientes de grupos.

2.1. Regresión con una variable independiente de dos grupos y otra variable independiente de tres grupos, solo 'efectos principales'. Ecuación de conjunto y ecuación para cada grupo. Cambio de R^2. Interpretación. Representación gráfica

Supongamos que estamos interesados en comprobar cuál es el efecto de la variable 'Minority' y de la variable 'Jobcat' sobre 'Salary'. Ya hemos visto anteriormente que la variable 'Minority' tiene dos categorías: 0: 'No minority' y 1: 'Yes minority', mientras que hay tres grupos en la variable 'Jobcat', 1: 'Clerical', 2: 'Custodial' y 3: 'Manager'.

Recordemos que en la Unidad 1 guardamos el fichero 'Comp_dat_Dms.sav' (acrónimo de '*Com*pany *dat*a with *D*u*m*mie*s*') tenemos ya las variables *dummy* de cada grupo guardadas, por lo que no volveremos a generar las variables *dummy* para cada variable de grupo. Trabajaremos con los datos de este fichero.

Como indicábamos, si se desea analizar 'Salary' = f('Minority', 'Jobcat'), pero solo con efectos principales, hemos de tener en cuenta:

a) La variable 'Minority' es una variable de grupos que ya es *dummy* (con valores '0' y '1'),
b) que al ser 'Jobcat' una variable con 3 categorías, hemos de introducir las correspondientes variable *dummy*: 'D_JC_custodial' y 'D_JC_manager', es decir, usaremos el grupo 'Clerical' como grupo de referencia.

Con lo cual la función anterior quedaría:

$$\text{'Salary'} = \text{f('Minority', ['D\_JC\_custodial', 'D\_JC\_manager'])} \quad (2.1)$$

En la Ecuación 2.1, hemos puesto las dos variables *dummy* de 'Jobcat' entre corchetes para indicar que son dummies de una variable primitiva ('Jobcat'), así la función inicial y la ecuación 2.1 son similares, pero la Ecuación 2.1 es la correcta para estimar los valores de la regresión. La función 2.1, en una ecuación de regresión lineal sería:

$$\text{Salary} = b_0 + b_1 \cdot \text{Minority} + [b_2 \cdot \text{D\_JC\_custodial} + b_3 \cdot \text{D\_JC\_manager}] + e \quad (2.2)$$

En la Ecuación 2.2 también hemos escrito entre corchetes los términos correspondientes a la variable 'Jobcat' desglosada en sus variables *dummy*, con el fin de destacar que es la misma variable (en bloque), pero desplegada en cada una de sus categorías.

Para llevar a cabo la Ecuación 2.2, guardando los valores predichos y pronosticados, corremos la sintaxis:

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA CHANGE /(*)
     /CRITERIA=PIN(.05) POUT(.10)
     /NOORIGIN
     /DEPENDENT salary
     /METHOD=ENTER minority   /(**)
     /METHOD=ENTER D_JC_custodial D_JC_manager /(**)
     /SAVE PRED RESID.
```

En la anterior sintaxis hay dos novedades importantes:

a) En la línea con la anotación `*` aparece la opción `CHANGE`, que es el código para calcular el 'cambio de R^2'.
b) Con la anotación `**` se indica que las variables, por medio de la opción `METHOD=ENTER`, se introducen en 'bloques' (pensados como conjuntos de variables diferentes pertenecientes a la misma variable de grupos): el bloque formado por 'Minority' (primera línea con `**)`, y el otro bloque formado por 'D_JC_custodial', 'D_JC_manager' (segunda línea con `**`).

Más tarde explicaremos, sobre los mismos resultados obtenidos, la implicación de estas dos instrucciones complementarias.

En la Tabla 2.1 aparecen los resultados de la ecuación de regresión, esta tabla contiene en el apartado (a) el sumario de los modelos 1 y 2, y el cambio de R^2, en el (b) el ANOVA de cada modelo, y en el (c) las correspondientes ecuaciones de regresión para cada modelo; ampliaremos estos aspectos a continuación.

IMPORTANTE:
Si aplicas la sintaxis anterior, no olvides eliminar "/(*)" y "/(**)".

En el fichero de datos, guarda PRE_1 con el nombre 'PRE_Sal_f_min_jbct', y añade la etiqueta a esa misma variable: 'PRE_'Salary' = f('Minority', ['D_JC_custodial', 'D_JC_manager'])', con el fin de identificar de dónde provienen esos resultados.

Tabla 2.1
Resultados de la regresión: 'Salary' = f('Minority', ['D_JC_custodial', 'D_JC_manager'])

(a)

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,165[a]	,027	,025	$16,093.082	,027	13,233	1	472	,000
2	,520[b]	,271	,266	$13,963.988	,243	78,452	2	470	,000

a. Predictors: (Constant), Minority Classification
b. Predictors: (Constant), Minority Classification, Custodial = 1, otherwise = 0, Manager = 1, Other = 0
c. Dependent Variable: Current Salary

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3427267305	1	3427267305	13,233	,000[b]
	Residual	1,222E+11	472	258987282,8		
	Total	1,257E+11	473			
2	Regression	34022572657	3	11340857552	58,160	,000[c]
	Residual	91646692122	470	194992962,0		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary
b. Predictors: (Constant), Minority Classification
c. Predictors: (Constant), Minority Classification, Custodial = 1, otherwise = 0, Manager = 1, Other = 0

(c)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	59406,432	836,639		71,006	,000
	Minority Classification	-6497,490	1786,121	-,165	-3,638	,000
2	(Constant)	54929,650	826,562		66,456	,000
	Minority Classification	-2473,020	1594,463	-,063	-1,551	,122
	Custodial = 1, otherwise = 0	-2908,936	2812,079	-,041	-1,034	,301
	Manager = 1, Other = 0	21214,185	1718,221	,498	12,347	,000

a. Dependent Variable: Current Salary

En la Tabla 2.1 apartado (a) aparecen los valores de la correlación (R) y la correlación a cuadrado (R^2) de cada uno de los apartados de la ecuación que hemos pedido: 'Minority' (primer bloque), más 'D_JC_custodial', 'D_JC_manager' (segundo bloque); por tanto, ahora tenemos dos modelos de regresión:

El modelo 1:

$$\text{'Salary'} = f(\text{'Minority'}) \qquad (2.3)$$

el modelo 2:

$$\text{'Salary'} = \text{f('Minority', ['D\_JC\_custodial', 'D\_JC\_manager'])} \quad \textbf{(2. 4)}$$

El modelo 1 está formado por el primer 'bloque' de variables introducido en la ecuación, y el modelo 2 está formado por los dos primeros (y en nuestro caso, únicos) bloques de variables de la ecuación (si hubiese más bloques, saldrían tantos modelos acumulados como bloques).

Así, se comprueba que la correlación al cuadrado (o coeficiente de determinación) de 'Salary' en función de 'Minority', de acuerdo con la Ecuación 2.3, es de *.027*, pero la correlación al cuadrado de 'Salary' con 'Minority', 'D_JC_custodial', 'D_JC_manager', según la Ecuación 2.4, es de *.271* (al introducir el segundo bloque, se ha mejorado la correlación al cuadrado en *.243* puntos). El apartado más interesante es el que se ha destacado en un marco de color rojo, y que es el cambio en R^2, suministrando la significación estadística del segundo bloque de variables ['D_JC_custodial', 'D_JC_manager'], $F(2, 470) = 78.452\ (p < .001)$, lo cual indica que el segundo bloque de variables es significativo (tratadas como si fuesen una sola variable).

En la Tabla 2.1 apartado (b) se comprueba que la ecuación con el modelo 1 obtiene un ajuste de conjunto con un valor de $F(1, 472) = 13.233\ (p < .001)$, mientras que con el modelo 2, Ecuación 2.5, tiene: $F(3, 470) = 58.160\ (p < .001)$, lo cual indica que ambos modelos son estadísticamente significativos, con lo cual podemos aceptar ambos como explicativos de la variable dependiente; pero de momento, solo centraremos la atención en el modelo 2, indicando que es significativo.

En la Tabla 2.1 apartado (c), se puede observar que en el modelo 2, el efecto de la variable simple 'Minority' es no-significativo ($b = .2473.0,\ t = -1.551,\ p = .122$), por lo que podría quitarse de la Ecuación 2.3 (salvo que se tuviesen razones de tipo teórico para dejar este coeficiente en la ecuación). Dejamos a la consideración del lector qué hacer con esta variable; aunque nosotros la dejaremos para continuar con la exposición didáctica del modelo.

En cuanto a la probabilidad de las variables *dummy* de 'Jobcat' (recordemos que su categoría de referencia es el de 'Clerical'), se observa que la *dummy* 'D_JC_custodial' tiene una $p < .301$, lo cual indica que la diferencia en variable dependiente en función de esta categoría y la de referencia ('Clerical') es no significativa (las respectivas medias en 'Salary' no difieren significativamente entre sí); pero el efecto de

'D_JC_manager' es significativo a un α = .05 ($p < .001$), indicando que las respectivas medias de 'Salary', en función de que sea 'Manager' o 'Clerical', difieren entre sí.

Obsérvese que la variable 'Minority' cuando está sola como variable independiente (como en el modelo 1) es estadísticamente significativa, pero al añadir las variables *dummy* 'D_JC_custodial' y 'D_JC_manager', deja de ser significativa, seguramente porque hay una elevada correlación entre las tres variables *dummy*, pero en conjunto, es mayor la correlación de las variables 'D_JC_custodial' y 'D_JC_manager' con 'Salary', que la que hay entre 'Minority' y 'Salary'.

Los resultados de la Tabla 2.1 indican que el modelo más simple y el de mejor ajuste sería: 'Salary' = f('Jobcat') = f(['D_JC_custodial', 'D_JC_manager']), puesto que al introducirse en la ecuación hacen que la otra variable independiente, 'Minority', sea no significativa y, para ello, tendríamos que estimar los parámetros de la regresión: 'Salary' = f(['D_JC_custodial', 'D_JC_manager']), obteniendo un resultado de conjunto significativo (véase Ecuación 2.5), y mejor que el de 'Salary' = f('Minority'), puesto que en la ecuación conjunta, 'Minority' es no significativa. En regresión con bloques de variables, cuando se tienen dudas sobre si alguna variable simple o algún bloque es o no significativo, la recomendación es volver a correr el modelo dejando en último lugar la variable simple o el bloque de variables sobre el que se tienen dudas acerca de su significación estadística, de esta manera se obtiene su aportación final al modelo y se evitan equívocos sobre los diferentes parámetros de información multivariada.

Seguiremos desarrollando, por motivos didácticos, el modelo propuesto en las Ecuaciones 2.1 a 2.3.

En la Tabla 2.1 apartado (c) se ponen los coeficientes de regresión para cada variable, así nuestro modelo vendría representado por:

$$Salary = 54929.650 - 2473.020 \cdot Minority + \left[-2908.936 \cdot D_{JC_{custodial}} + 21214.185 \cdot D_{JC_{manager}}\right] + e \quad (2.5)$$

Indicando la Ecuación 2.5 que es, en conjunto, estadísticamente significativa (tal como se ve en la Tabla 2.1(b): $F(3, 470) = 58.160$ $(p < .001)$); además, las dos variables

dummy de 'Jobcat' ('D_JC_custodial', 'D_JC_manager') como bloque, son significativas (*F(2, 470) = 78.452 (p < .001)*).

> Insistimos en que no es interesante la significación de cada variable *dummy* de 'Jobcat' por separado, sino su significación de conjunto; es decir, no se puede quitar la categoría 'D_JC_custodial' aunque sea no significativa, porque forma parte de un bloque de variables que en conjunto, sí es todo él significativo.

Así, la significación particular de la variable 'D_JC_custodial' es *p = .301* (*b= -2908,936, t = -1.034*), mientras que la de 'D_JC_manager' es *p = < .001* (*b = 21214,185, t = 12.347)*. Estos dos valores nos dan si los valores pronosticados de cada una de estas variables difieren del valor pronosticado del grupo de referencia ('Clerical'), como a continuación aclararemos.

Para calcular la significación estadística del cambio de R^2, se usa la fórmula:

$$F = \frac{(R_2^2 - R_1^2)/(k_2 - k_1)}{(1 - R_2^2)/(N - k_2 - 1)} \qquad (2.\ 6)$$

donde *F* es el estadístico *F* de Fisher (con $k_2 - k_1$ grados de libertad en el numerador y $N - k_2 - 1$ gados de libertad en el denominador), R_2^2 es el valor de la R^2 en el modelo 2 (ampliado), y R_1^2 es el valor de la R^2 en el modelo 1 (reducido), k_2 es el número de variables independientes introducidos en la regresión del modelo 2, mientras k_1 es el número de variables independientes del modelo 1, y *N* es el número de casos del análisis. Este estadístico F da la significación de conjunto de las variables añadidas al modelo 2 respecto del 1.

Así, en la Tabla 2.1 (a), se puede comprobar la mejora en el cambio de R^2 del modelo 2 respecto del modelo 1 mediante la Ecuación 2.6:

$$F = \frac{(R_2^2 - R_1^2)/(k_2 - k_1)}{(1 - R_2^2)/(N - k_2 - 1)} = \frac{(.271 - .027)/(3 - 1)}{(1 - .271)/(474 - 3 - 1)} = \frac{.244/2}{.729/470} = \frac{.122}{.0016} = 78.45 \qquad (2.\ 7)$$

Dicho valor del estadístico *F*, con 2 grados de libertad en el numerador y 470 en el denominador da una *p < .001*. Como se puede observar en la Tabla 2.1 (a), indicando que la aportación de las dos variables *dummy*, 'D_JC_custodial' y 'D_JC_manager' es significativa, por tanto la capacidad de explicación de ambas variables sobre la variable dependiente, una vez extraída la capacidad explicativa del modelo 1, es aceptable a un nivel α = .05.

Para hacer el pronóstico de cada grupo, vemos que el cruce de 'Minority' (2 grupos o categorías) con 'Jobcat' (3 grupos), produce que haya 2*3 = 6 grupos:

- El grupo 'No minority'-'Clerical' (codificación: 'Minority' = 0, 'D_JC_custodial' = 0, 'D_JC_manager' = 0), obsérvese que es el grupo de referencia.
- El grupo 'No minority'-'D_JC_custodial' (codificación: 'Minority' = 0, 'D_JC_custodial' = 1, 'D_JC_manager' = 0).
- El grupo 'No minority'-'D_JC_manager' (codificación: 'Minority' = 0, 'D_JC_custodial' = 0, 'D_JC_manager' = 1).
- El grupo 'Minority'-'Clerical' (codificación: 'Minority' = 1, 'D_JC_custodial' = 0, 'D_JC_manager' = 0).
- El grupo 'Minority'-'D_JC_custodial' (codificación: 'Minority' = 1, 'D_JC_custodial' = 1, 'D_JC_manager' = 0).
- El grupo 'Minority'-'D_JC_manager' (codificación: 'Minority' = 1, 'D_JC_custodial' = 0, 'D_JC_manager' = 1).

Por lo tanto, a partir de la Ecuación general 2.2, el pronóstico para cada submuestra será:

a) grupo 'No minority'-'Clerical':

$$\boldsymbol{Salary_{NoMin*Cler}} = 54929.650 - 2473.020 \cdot 0 - 2908.936 \cdot 0 + 21214.185 \cdot 0 = \mathbf{54929.650} \quad (2.8)$$

Nótese que cuando todas las variables de la ecuación de regresión son dummies, el valor pronosticado del grupo de referencia es el del término independiente de la ecuación (*54929.650*).

b) grupo 'No minority'-'D_JC_custodial':

$$\boldsymbol{Salary_{NoMin*Cust}} = 54929.650 - 2473.020 \cdot 0 - 2908.936 \cdot 1 + 21214.185 \cdot 0 = \mathbf{52020.714} \quad (2.9)$$

c) grupo 'No minority'-'D_JC_manager':

$$\boldsymbol{Salary_{NoMin*Mang}} = 54929.650 - 2473.020 \cdot 0 - 2908.936 \cdot 0 + 21214.185 \cdot 1 = \mathbf{76143.835} \quad (2.10)$$

d) grupo 'Minority'-'Clerical':

$$\boldsymbol{Salary_{Min*Cler}} = 54929.650 - 2473.020 \cdot 1 - 2908.936 \cdot 0 + 21214.185 \cdot 0 = \mathbf{52456.630} \quad \textbf{(2. 11)}$$

e) grupo 'Minority'-'D_JC_custodial':

$$\boldsymbol{Salary_{Min*Cust}} = 54929.650 - 2473.020 \cdot 1 - 2908.936 \cdot 1 + 21214.185 \cdot 0 = \mathbf{49547.694} \quad \textbf{(2. 12)}$$

f) grupo 'Minority'-'D_JC_manager':

$$\boldsymbol{Salary_{Min*Mang}} = 54929.650 - 2473.020 \cdot 1 - 2908.936 \cdot 0 + 21214.185 \cdot 1 = \mathbf{73670.815} \quad \textbf{(2. 13)}$$

Observa dos aspectos:

I. En el fichero de datos del SPSS, en la variable de valores pronosticados 'PRE_Sal_f_min_jbct', solo aparecen los 6 valores de las Ecuaciones 2.8 a 2.13,

II. para comprobar con SPSS los valores pronosticados, si se pide la sintaxis:

```
SUMMARIZE
     /TABLES=PRE_Sal_f_min_jbct BY minority BY jobcat
     /FORMAT=NOLIST TOTAL
     /TITLE='Case Summaries'
     /MISSING=VARIABLE
     /CELLS=MEAN STDDEV COUNT.
```

Aparecen los resultados de la Tabla 2.2, donde se comprueba que éstos coinciden con los obtenidos en nuestro análisis. Además, la desviación típica de todas las categorías es '0' porque todos los valores dentro de cada subgrupo son iguales.

Aunque resulte redundante con los datos de la Tabla 2.2, representaremos estos mismos resultados en la Figura 2.1, como suele hacerse en la literatura científica, para ello, corremos la sintaxis:

```
GRAPH
     /LINE(MULTIPLE)=MEAN(PRE_Sal_f_min_jbct) BY
          minority BY jobcat.
```

En la Figura 2.1 se puede observar que en el eje horizontal se ha representado la VI 'Minority' y dentro de la figura aparecen tres líneas: las correspondientes a las tres

categorías de la variable 'Jobcat' ('Clerical', 'Custodial' y 'Manager'); los valores pronosticados aparecen en los respectivos vértices de las líneas (marcados con puntos). Obsérvese que los valores pronosticados son los mismos que en la Tabla 2.2 (teniendo en cuenta el redondeo de decimales).

En la Figura 2.1 se observa que las tres líneas son paralelas (equidistantes), debido a la aditividad del modelo de efectos principales. Esto es porque las diferencias de valores 'homólogos' son siempre las mismas, pues si se calcula el valor pronosticado de 'Minority yes' menos el de 'Minority No' para cada grupo de 'Jobcat', siempre se obtiene el mismo resultado: para 'Clerical': 52456.629 - 54929.650, para 'Custodial': 49547,693 - 52020,714, y para los 'Manager: 73670,814 - 76143,834 (en los tres casos da como resultado: *2473.020),* que es el coeficiente de la variable 'Minority' en la Ecuación 2.5.

Tabla 2.2.
Media, desviación típica y número de personas (N) de la variable 'Predicted salary' de acuerdo con la Ecuación 9.3

Case Summaries

'PRE_'Salary' = f('Minority', ['D_JC_custodial', 'D_JC_manager'])'

Minority Classification	Employment Category	Mean	Std. Deviation	N
No	Clerical	54929,64952	0E-8	276
	Custodial	52020,71352	0E-8	14
	Manager	76143,83430	0E-8	80
	Total	59406,43243	8820,156004	370
Yes	Clerical	52456,62912	0E-8	87
	Custodial	49547,69313	0E-8	13
	Manager	73670,81391	0E-8	4
	Total	52908,94231	4282,380118	104
Total	Clerical	54336,94215	1057,144287	363
	Custodial	50830,00000	1259,200350	27
	Manager	76026,07143	529,8146924	84
	Total	57980,82278	8481,116028	474

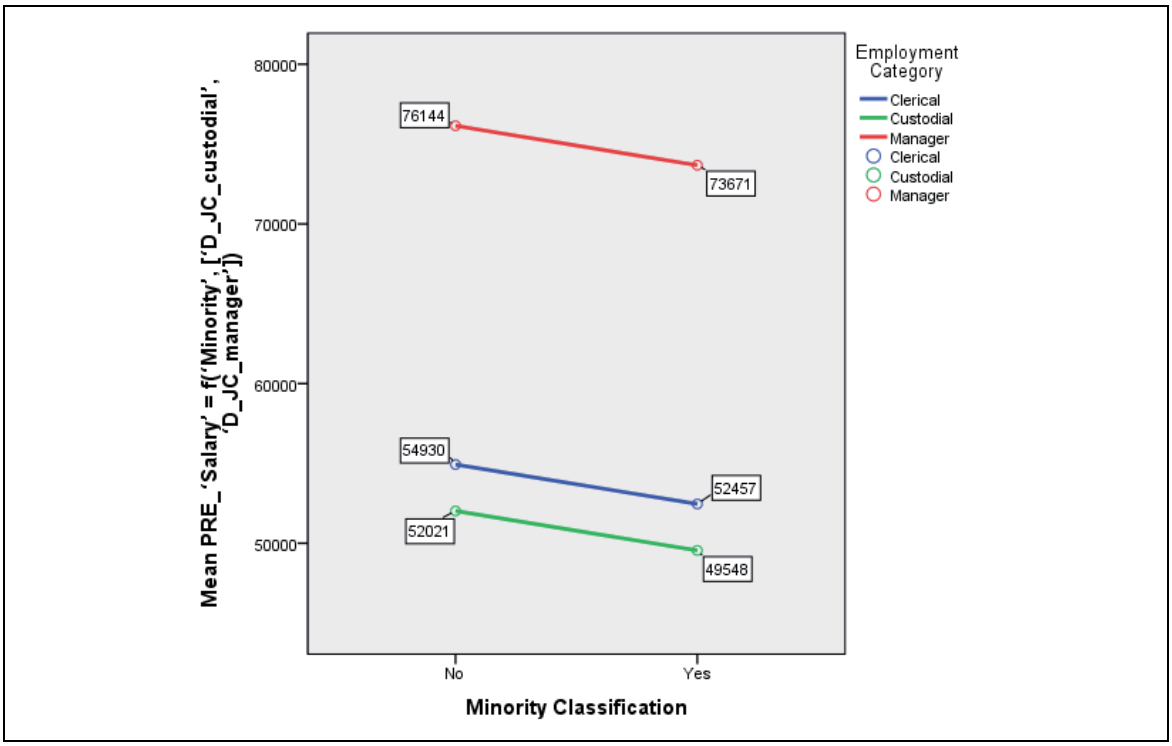


Figura 2.1. Valores pronosticados del modelo de las Ecuaciones 2.3 y 2.6

Como vemos, la geometría de los valores pronosticados es muy sencilla (siempre da rectas paralelas), pero la realidad suele ser más compleja; así, para ver cuáles son las medias reales de la variable 'Salary' de cada respectivo grupo, podemos hacer la Figura 2.2, por medio de la sintaxis:

```
GRAPH
     /LINE(MULTIPLE)=MEAN(salary) BY minority BY
          jobcat.
```

En la figura 2.2 se comprueba cómo las medias reales de 'Salary' no son las pronosticadas, y que las líneas no son estrictamente paralelas; con el fin de reproducir los valores reales de cada respectivo grupo, necesitaríamos hacer interacción de variables, como veremos en la Unidad 3.

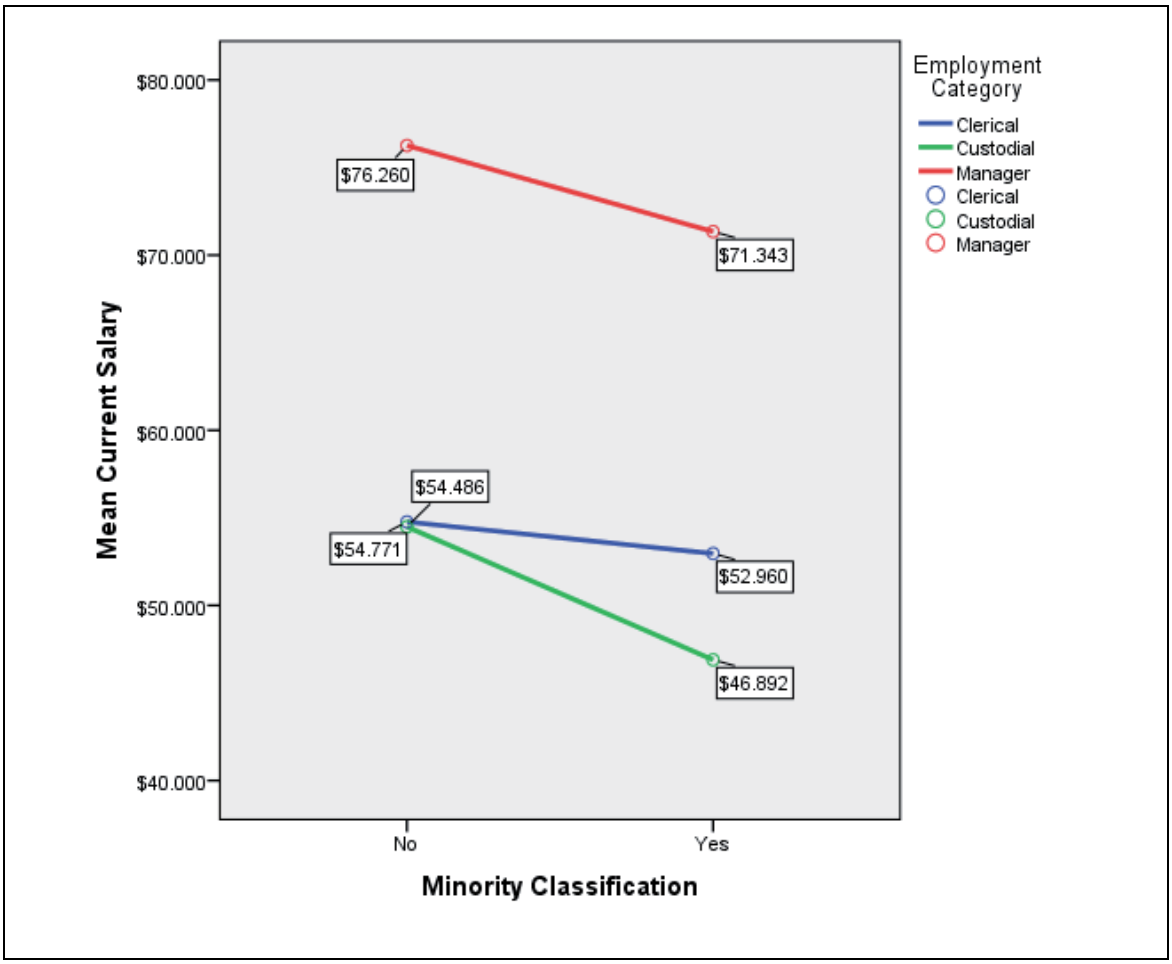


Figura 2.2. Medias reales de la variable 'Salary' en función de 'Minority' y 'Jobcat'

2.2. Regresión con una variable independiente de dos grupos y otra de seis grupos, solo 'efectos principales'. Ecuación de conjunto y ecuación para cada grupo. Interpretación. Representación gráfica

En el apartado 2.1 hemos visto la regresión de una variable dependiente sobre una variable independiente de dos grupos y otra de tres, en este apartado expondremos un diseño algo más complejo: la regresión de una variable independiente de dos grupos y otra variable independiente de seis grupos, con el fin de comprobar su efecto sobre una variable dependiente. Para ello, abriremos el fichero 'Comp_dat_Dms.sav', y tomaremos como variables independientes: 'Gender' (con dos grupos), y 'Center' (con seis grupos), insistiendo, una vez más, que solo haremos el cálculo de los efectos principales.

Desde una perspectiva funcional, el modelo a analizar sería 'Salary' = f('Gender', 'Center'), pero como hemos visto, 'Gender' no está dicotomizada, siendo 'D_gndr_fem' su equivalente *dummy*, mientras 'Center' está desglosada en las siguientes variables *dummy*: 'D_Ctr_2', 'D_Ctr_3', 'D_ Ctr_4', 'D_ Ctr_5' y 'D_ Ctr_6' (cada una de ellas representa desde '*dummy* center 2' hasta '*dummy* center 6', siendo el 'Center 1' el grupo de referencia; si es preciso, repasar apartado 5.7), así la expresión anterior, en forma funcional, para poderse hacer correctamente se transforma en:

$$\text{'Salary'} = f(\text{'}D_{gndr_{fem}}\text{'}, [\text{'}D_{Ctr_2}\text{'}, \text{'}D_{Ctr_3}\text{'}, \text{'}D_{Ctr_4}\text{'}, \text{'}D_{Ctr_5}\text{' y '}D_{Ctr_6}\text{'}]) \quad (2.\ 14)$$

O lo que es lo mismo, pasando la Ecuación funcional 2.14 a la Ecuación 2.15, de álgebra de regresión lineal, se ha de estimar:

$$\text{Salary} = b_0 + b_1 \cdot D_{gndr_{fem}} + [b_2 \cdot D_{Ctr_2} + b_3 \cdot D_{Ctr_3} + b_4 \cdot D_{Ctr_4} + b_5 \cdot D_{Ctr_5} + b_6 \cdot D_{Ctr_6}] + e \quad (2.\ 15)$$

En la Ecuación 2.15 hemos escrito entre corchetes los términos correspondientes a la variable 'Center' desglosada en variables *dummy*.

Por lo tanto, correremos la sintaxis de análisis de la regresión para la Ecuación 2.15 como en el apartado 2.1, haciendo dos 'bloques de variables' (en sintaxis: `METHOD=ENTER`), guardando los valores pronosticados y los residuales (`SAVE PRED RESID`), y pidiendo el cálculo de la diferencia de R^2 para cada bloque de variables (`CHANGE`):

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA CHANGE
```

```
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT salary
/METHOD=ENTER D_gndr_fem
/METHOD=ENTER D_Ctr_2 D_Ctr_3 D_Ctr_4 D_Ctr_5
     D_Ctr_6
/SAVE PRED RESID.
```

Que da los resultados de la Tabla 2.3; esta tabla contiene en el apartado (a) el sumario de los modelos 1 y 2, y el cambio de R^2, en el (b) el ANOVA de cada modelo, y en el apartado (c) las correspondientes ecuaciones de regresión.

Tabla 2.3
Resultados de la ecuación 9.17

(a)

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,452[a]	,204	,202	$14,557.154	,204	121,029	1	472	,000
2	,626[b]	,392	,384	$12,795.412	,188	28,785	5	467	,000

a. Predictors: (Constant), D_gndr_fem
b. Predictors: (Constant), D_gndr_fem, Center 3, Center 5, Center 6 , Center 2, Center 4
c. Dependent Variable: Current Salary

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25647398289	1	25647398289	121,029	,000[b]
	Residual	1,000E+11	472	211910734,1		
	Total	1,257E+11	473			
2	Regression	49210823787	6	8201803964	50,096	,000[c]
	Residual	76458440992	467	163722571,7		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary
b. Predictors: (Constant), D_gndr_fem
c. Predictors: (Constant), D_gndr_fem, Center 3, Center 5, Center 6 , Center 2, Center 4

Tabla 2.3 *(Continuación)*
(c)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	64711,357	906,289		71,403	,000
	D_gndr_fem	-14769,783	1342,545	-,452	-11,001	,000
2	(Constant)	54902,625	2369,970		23,166	,000
	D_gndr_fem	-8431,355	1652,207	-,258	-5,103	,000
	Center 2	2667,706	2217,843	,064	1,203	,230
	Center 3	3510,025	2427,393	,076	1,446	,149
	Center 4	2255,224	2573,165	,057	,876	,381
	Center 5	10360,364	2437,832	,251	4,250	,000
	Center 6	24188,311	2811,759	,518	8,603	,000

a. Dependent Variable: Current Salary

IMPORTANTE
En el fichero de datos, guarda PRE_1 con el nombre 'PRE_Sal_f_gend_cntr' ('predicted values of salary in function of gender and center), y añade la etiqueta a esa misma variable: 'PRE_'Salary' = f('D_gndr_fem', ['D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6'])', con el fin de identificar de dónde provienen esos resultados.

En el apartado (a) de la Tabla 2.3 se comprueba que el cambio de R^2 (entre el modelo 1 y el 2) es significativo ($F(5, 467) = 28.785$, $p < .001$), por lo que aceptamos que las cinco variables *dummy* añadidas en el modelo 2 aportan capacidad explicativa al pronóstico de 'Salary' mediante la Ecuación 2.15 puesta a prueba. En la misma Tabla 2.3, apartado (b), se comprueba que todo el modelo de la Ecuación 2.15 es significativo. En el apartado (c) aparecen los valores de los distintos coeficientes, donde se observa que el coeficiente de la variable 'D_gndr_fem' es significativo, por tanto, aceptamos que tanto la variable 'D_gndr_fem' como todas las *dummy* que integran la variable 'Center' ('D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6') son significativas y aportan capacidad de pronóstico al modelo, con lo cual aceptamos estadísticamente la Ecuación 2.15 como explicativa de los datos.

Nótese que las variables *dummy* 'Center 2', 'Center 3' y 'Center 4' no son estadísticamente significativas (las tres con una $p > .05$), es decir, no difieren significativamente del grupo de referencia ('Center 1') pero no pueden quitarse, pues

forman parte de una variable más amplia ('Center') y, si se suprimieran, se cometerían errores de estimación, de pronóstico y de interpretación en el modelo.

La ecuación de conjunto con sus respectivos coeficientes queda así:

$$\text{Salary} = 54902.6 - 8431.4 \cdot \text{D\_gndr\_fem} + [2667.7 \cdot \text{D\_Ctr\_2} + 3510.0 \cdot \text{D\_Ctr\_3} + 2255.2 \cdot \text{D\_Ctr\_4} + 10360.4 \cdot \text{D\_Ctr\_5} + 24188.3 \cdot \text{D\_Ctr\_6}] + e \quad (2.16)$$

Siendo las ecuaciones de los valores pronosticados (Salary') para cada respectivo grupo:

a) Para hombres que trabajan en el 'Center 1' ('D_gndr_fem' = 0; 'D_Ctr_2' = 0, 'D_Ctr_3' = 0, 'D_Ctr_4'= 0, 'D_Ctr_5' = 0, y 'D_Ctr_6' = 0):

$$Salary'_{M,Cr1} = 54902.6 - 8431.4 \cdot 0 + [2667.7 \cdot 0 + 3510.0 \cdot 0 + 2255.2 \cdot 0 + 10360.4 \cdot 0 + 24188.3 \cdot 0] = \mathbf{54902.6} \quad (2.17)$$

Por tanto, el valor esperado de salario para los hombres que trabajan en 'Center 1' es de 54902.6 dólares al año. Obsérvese, una vez más, que al ser este el grupo de referencia, su valor esperado es el del término independiente.

b) Para hombres que trabajan en el 'Center 2' ('D_gndr_fem' = 0; 'D_Ctr_2' = 1, 'D_Ctr_3' = 0, 'D_Ctr_4'= 0, 'D_Ctr_5' = 0, y 'D_Ctr_6' = 0):

$$Salary'_{M,Cr2} = 54902.6 - 8431.4 \cdot 0 + [2667.7 \cdot 1 + 3510.0 \cdot 0 + 2255.2 \cdot 0 + 10360.4 \cdot 0 + 24188.3 \cdot 0] = 54902.6 + 2667.7 = \mathbf{57570.3} \quad (2.18)$$

c) Para los hombres que trabajan en el 'Center 3' ('D_Ctr_3' = 1, otherwise = 0):

$$Salary'_{M,Cr3} = 54902.6 - 0 + [0 + 3510.0 + 0 + 0 + 0] = 54902.6 + 2667.7 = \mathbf{58412.6} \quad (2.19)$$

d) Para hombres que trabajan en 'Center 4':

$$Salary'_{M,Cr4} = 54902.6 - 0 + [0 + 0 + 2255.2 + 0 + 0] = 54902.6 + 2255.2 = \mathbf{57157.8} \quad (2.20)$$

e) Para hombres que trabajan en 'Center 5':

$$Salary'_{M,Cr5} = 54902.6 - 0 + [0 + 0 + 0 + 10360,4 + 0] = 54902.6 + 10360,4 = \mathbf{65263.0} \quad (2.21)$$

f) En hombres que trabajan en 'Center 6':

$$Salary'_{M,Cr6} = 54902.6 - 0 + [0 + 0 + 0 + 0 + 24188.4] = 54902.6 + 24188{,}4$$
$$= \mathbf{79091.0}$$

(2. 22)

g) Para mujeres que trabajan en el 'Center 1' ('D_gndr_fem' = 1; 'D_Ctr_2' = 0, 'D_Ctr_3' = 0, 'D_Ctr_4'= 0, 'D_Ctr_5' = 0, y 'D_Ctr_6' = 0):

$$Salary'_{F,Cr1} = 54902.6 - 8431.4 \cdot 1 + [2667.7 \cdot 0 + 3510.0 \cdot 0 + 2255.2 \cdot 0 + 10360.4 \cdot 0 + 24188.3 \cdot 0] = \mathbf{46471.2}$$

(2. 23)

h) Para mujeres que trabajan en el 'Center 2' ('D_gndr_fem' = 1; 'D_Ctr_2' = 1, 'D_Ctr_3' = 0, 'D_Ctr_4'= 0, 'D_Ctr_5' = 0, y 'D_Ctr_6' = 0):

$$Salary'_{F,Cr2} = 54902.6 - 8431.4 + [2667.7 \cdot 1 + 0 + 0 + 0 + 0] = 46471.2 + 2667.7 = \mathbf{49138.9}$$

(2. 24)

i) Para mujeres que trabajan en el 'Center 3' ('D_Ctr_3' = 1, otros = 0):

$$Salary'_{F,Cr3} = 54902.6 - 8431.4 + [0 + 3510.0 + 0 + 0 + 0] = 46471.2 + 2667.7 = 49138.9$$

(2. 25)

j) Para mujeres que trabajan en 'Center 4':

$$Salary'_{F,Cr4} = 54902.6 - 8431.4 + [0 + 0 + 2255.2 + 0 + 0] = 46471.2 + 2255.2 = \mathbf{48726.4}$$

(2. 26)

k) Para mujeres que trabajan en 'Center 5':

$$Salary'_{F,Cr5} = 54902.6 - 8431.4 + [0 + 0 + 0 + 10360.4 + 0] = 46471.2 + 10360.4 = \mathbf{56831.6}$$

(2. 27)

l) En mujeres que trabajan en 'Center 6':

$$Salary'_{F,Cr6} = 54902.6 - 8431.4 + [0 + 0 + 0 + 0 + 24188.4] = 46471.2 + 24188.4 = \mathbf{70659.5}$$

(2. 28)

Se pueden comprobar las medias de los valores esperados mediante la sintaxis:

```
SUMMARIZE
     /TABLES=PRE_Sal_f_gend_cntr BY gender BY Center
     /FORMAT=NOLIST TOTAL
     /TITLE='Case Summaries'
     /MISSING=VARIABLE
     /CELLS=MEAN VAR COUNT.
```

Los valores obtenidos desde la Ecuación 2.17 hasta la 2.28 y los de la tabla de esta sintaxis son iguales (salvo los errores debidos al redondeo de los resultados).

Para mayor aclaración, representaremos en conjunto los resultados obtenidos en la Ecuación 2.16, mediante una figura de líneas, con la sintaxis:

```
GRAPH
     /LINE(MULTIPLE)=MEAN(PRE_Sal_f_gend_cntr) BY
          gender BY Center.
```

Apareciendo la Figura 2.3, nótese que los valores esperados son los de los vértices de las líneas, que las líneas representan los grupos de diferente 'Center', y que las líneas son paralelas (como siempre que se incluyan variables independientes de grupos, con solo los efectos principales).

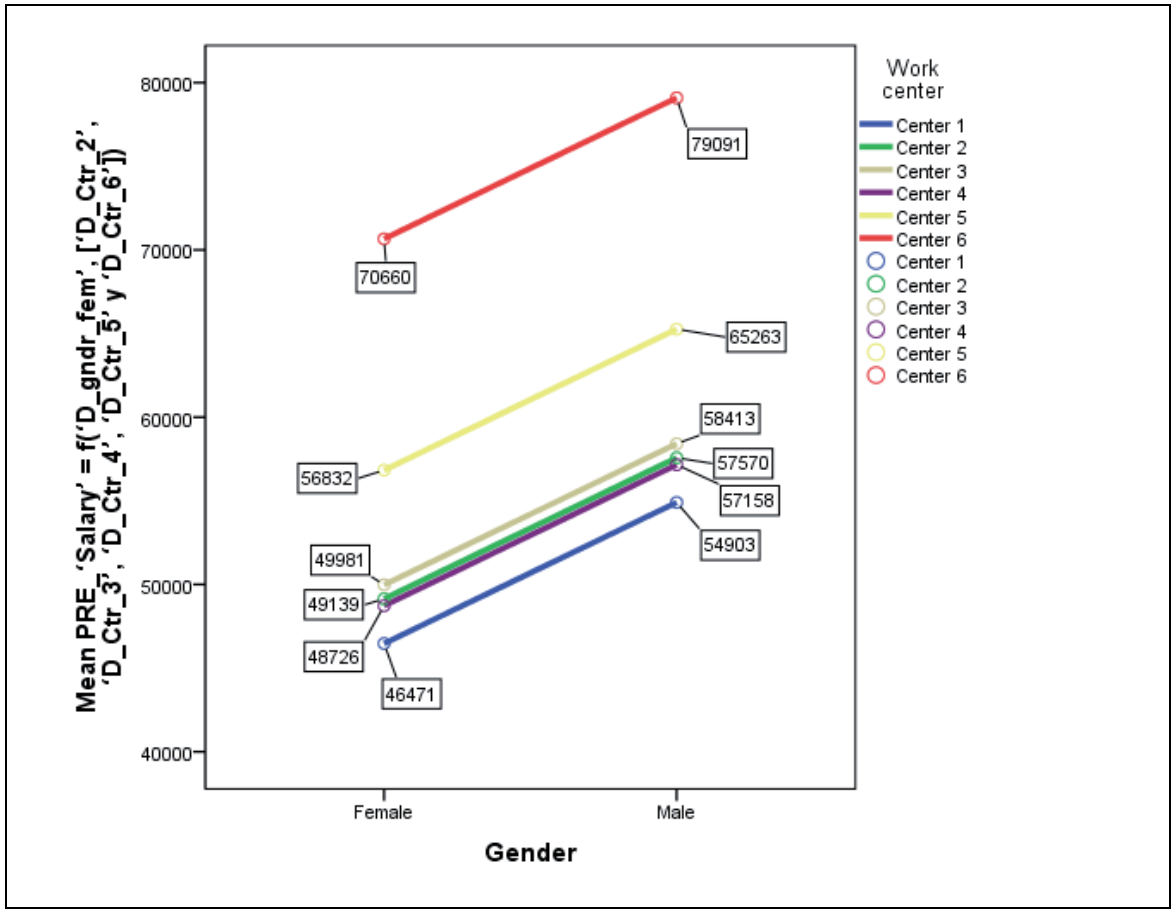


Figura 2.3. Valores esperados de la variable 'Salary' en función de 'Gender' y 'Center', conforme a la Ecuación 2.16

De cualquier forma, como indicábamos al final del apartado 2.1, los valores esperados de una ecuación de regresión con variables independientes de grupos (aunque sea toda

la ecuación significativa y los efectos principales también los sean) no suelen reproducir los las medias reales; como ejemplo, si se desea hacer una figura homóloga con las medias reales de 'Salary' en función de 'Gender' y de 'Center', se puede escribir la sintaxis:

```
GRAPH
     /LINE(MULTIPLE)=MEAN(salary) BY gender BY Center.
```

Que al correrla con el SPSS daría como resultado la Figura 2.4, en donde se aprecia que las líneas ya no son paralelas, y que hay una gran disparidad entre los valores reales y los esperados de 'Female' con 'Center 6'. Para reproducir los valores reales en las medias de los grupos (cuando se usan solo variables independientes de grupos), se ha de hacer la interacción entre las variables de grupos (ver la Unidad 3).

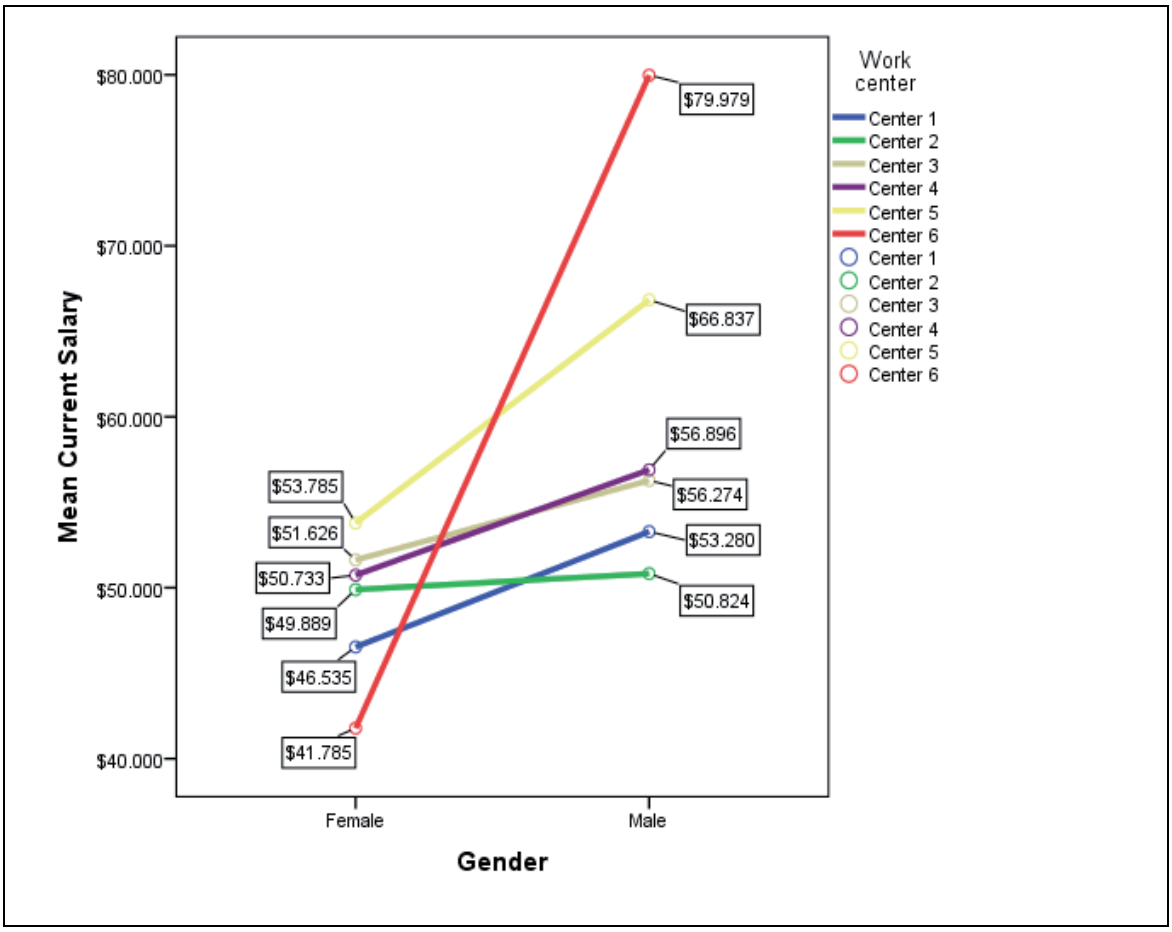


Figura 2.4. Medias de la variable 'Salary' en función de 'Gender' y 'Center'

2.3. Adenda: Relación entre la prueba F del análisis de la varianza (ANOVA) de dos factores (solo efectos principales) y la regresión con dos variables independientes de grupos

Hemos visto los resultados obtenidos en la regresión con variables independientes de grupos cuando solo se incluyen los efectos principales. Los resultados que obtendremos en el caso de que se analicen los mismos datos con el modelo de ANOVA de efectos principales, son idénticos, por ejemplo, si se desea hacer el contraste de medias de 'Salary' en función de 'Minority' y de 'Jobcat', se corre la sintaxis (en la cual no se incluye la interacción de 'Minority'*'Jobcat'):

```
UNIANOVA salary BY minority jobcat
    /METHOD=SSTYPE(3)
    /INTERCEPT=INCLUDE
    /SAVE=PRED RESID
    /CRITERIA=ALPHA(.05)
    /DESIGN=minority jobcat.
```

Obteniéndose los resultados de la Tabla 2.4.

Tabla 2.4.
Resultados del ANOVA de 'Salary' en función de 'Minority' y de 'Jobcat'

Between-Subjects Factors

		Value Label	N
Minority Classification	0	No	370
	1	Yes	104
Employment Category	1	Clerical	363
	2	Custodial	27
	3	Manager	84

Tests of Between-Subjects Effects

Dependent Variable: Current Salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3,402E+10[a]	3	11340857552	58,160	,000
Intercept	5,485E+11	1	5,485E+11	2812,875	,000
minority	469078186,9	1	469078186,9	2,406	,122
jobcat	30595305352	2	15297652676	78,452	,000
Error	91646692122	470	194992962,0		
Total	1,719E+12	474			
Corrected Total	1,257E+11	473			

a. R Squared = ,271 (Adjusted R Squared = ,266)

En la Tabla 2.4 se comprueba que en 'Corrected Model' el valor de *F(3, 470) = 58.160 (p < .001),* es coincidente con el valor de todo el modelo de la ecuación de regresión representado en la Tabla 2.1, apartado (b). Del mismo modo, el efecto de la variable 'Minority' en la Tabla 2.4 (*F(1, 470) = 2.406 (p = .122*), indica que este efecto es no significativo estadísticamente, y podría quitarse del modelo; el valor de la *p* de esta variable, que aparece en la Tabla 2.4, es el mismo que el de la Tabla 2.1 apartado (c) para el coeficiente de regresión, y nuevamente, al ser dos grupos: $t^2 = F$ ($-1.551^2 = 2.406$). Asimismo, la significación del efecto de la variable 'Jobcat' en la Tabla 2.4 del ANOVA (*F(2, 470) = 78.452 (p < .001)*), coincide con la del valor en la Tabla 2.1, apartado (a) de la significación del bloque formado por las variables de 'Jobcat'.

Dejamos al lector que haga el ANOVA de 'Salary' en función de 'Gender' y de 'Center' solo con los efectos principales de las dos variables independientes, con el fin de que compare los resultados y se compruebe que son iguales en los aspectos más importantes (en el valor de la F de conjunto, y en el efecto de cada variable independiente principal), los resultados en el valor de la probabilidad ('Signification') del 'intercept' (en el ANOVA), o de la 'Constant' (en la regresión) pueden cambiar de valor en función del sistema de codificación utilizado (según sea *dummy*, por los efectos o mediante contrastes). Como detalle adicional, se puede comprobar cómo los valores pronosticados y los residuales son los mismos en el fichero de datos tanto para la regresión de: 'Salary' = f('D_gndr_fem', ['D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6']), como los del ANOVA para: 'Salary' = f('Gender', 'Center'). Con el fin de comprobar estos resultados puede correrse la sintaxis:

```
UNIANOVA salary BY gender Center
     /METHOD=SSTYPE(3)
     /INTERCEPT=INCLUDE
     /SAVE=PRED RESID
     /CRITERIA=ALPHA(.05)
     /DESIGN=gender Center.
```

Una vez más, la ventaja de la regresión sobre el ANOVA es que se comprueba mejor el planteamiento del modelo y es más fácil hacer el correspondiente pronóstico de los valores esperados para cada grupo; mientras la principal ventaja del ANOVA es la comodidad de análisis (no es preciso hacer la correspondiente codificación de los grupos dentro de cada variable), aunque se pierde la capacidad de pronóstico (se ha de

conocer qué sistema de codificación utiliza el programa informático utilizado, en nuestro caso el SPSS, y qué valores se asignan a cada grupo), pero los resultados son los mismos.

2.4. Conclusiones

Las conclusiones de esta unidad, referidas al aprendizaje de los aspectos más importantes son:

- Cómo organizar 'bloques' de variables de grupos (mediante codificación *dummy*) correspondientes a cada variable principal.
- De qué forma se ha de plantear la ecuación de regresión con los respectivos 'bloques' de las variables independientes de grupos.
- Cómo estimar los parámetros de la ecuación de regresión.
- Cuál es la correcta interpretación de los estadísticos de conjunto de cada modelo de regresión introduciendo variables dummies mediante 'bloques'.
- Cómo interpretar la significación del cambio de R^2 para cada bloque de variables de grupos.
- Cuál es la interpretación de cada coeficiente de la ecuación de regresión.
- Cómo se ha de escribir la ecuación global de regresión, y los valores pronosticados para cada grupo.
- Elaborar las tablas de medias de los valores esperados conforme a los resultados del modelo estadístico obtenido.
- De qué forma se puede hacer una representación gráfica de los resultados obtenidos.
- Cuál es la relación entre el análisis de regresión con variables principales de grupos y el análisis ANOVA con las mismas variables, y cómo son los correspondientes resultados, así como su interpretación.
- Ventajas e inconvenientes del uso de la regresión y del ANOVA (ambos solo con variables principales) para variables independientes de grupos.

Lecturas recomendadas

Realmente, ningún libro trae un apartado sobre la regresión solo con efectos principales, pues todos los manuales lo tratan como un caso particular de la regresión o del ANOVA cuando la interacción de las variables es no-significativa. Como complemento útil a lo que se ha expuesto en esta unidad, puede resultar útil la lectura del apartado «Simple main effects» del capítulo 20 del libro de Pedhazur y Pedhazur (1991: 523-530), pero está más enfocado desde el cálculo del ANOVA que desde el de la regresión.

Bibliografía

Pedhazur, E. J.; Pedhazur, L. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.

Actividades

1. Con los datos del fichero 'Comp_dat_Dms.sav' se desea hacer la regresión 'Salbegin' = f('Gender', 'Center'), para ello, realiza los siguientes pasos:

- Haz la ecuación de regresión de 'Salbegin' = f('Gender', 'Center'), usando correctamente las correspondientes variables *dummy*, para 'Gender': 'D_gndr_fem', y para 'Center': 'D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5' y 'D_Ctr_6'.
- Comenta la significación del estadístico *F* de conjunto. Interpreta los resultados de conjunto.
- Interpreta la significación de cada variable principal. ¿Qué significa la mejora de R^2 en la tabla de resultados?
- Escribe la ecuación de regresión de conjunto de 'Salbegin' = f('Gender', 'Center').
- Escribe las ecuaciones para cada grupo. Comenta cada una de ellas (¿difiere la media del grupo respecto de la de su grupo de referencia?).
- Haz una figura de conjunto que represente los resultados.

2. Se desea hacer la ecuación de regresión lineal de 'Salbegin' = f('Gender', 'Jobcat'). Responde:

- Haz la ecuación de regresión de 'Salbegin' = f('Gender', 'Jobcat').
- Comenta la significación del estadístico *F* de conjunto. Interpreta los resultados de conjunto.
- Interpreta la significación de cada variable principal. ¿Qué significa la mejora de R^2 en la tabla de resultados?
- Escribe la ecuación de regresión de conjunto de 'Salbegin' = f('Gender', 'Jobcat').
- Escribe las ecuaciones para cada grupo. Comenta cada una de ellas.
- Haz una figura de conjunto que represente los resultados.

Unidad 3. Regresión con interacción de dos variables independientes de grupos

Introducción

En esta unidad nos ocuparemos de la interacción entre variables de grupos desde la perspectiva del análisis de regresión. En ocasiones nos encontramos ante la necesidad de comprobar cuál es el efecto de la interacción entre dos variables independientes, concretamente consideraremos la interacción cuando una de ellas está medida con dos niveles (grupos) y la otra con tres. Esta interacción, cuando resulta significativa, proporciona información privilegiada mucho más rica y compleja que la de los efectos principales de las variables independientes.

Para conseguir la interacción entre dos variables independientes de grupos, primeramente, hemos de transformar cada una de ellas en bloques de variables *dummy*, y una vez transformada cada una de ellas en variables *dummy*, se ha de multiplicar cada *dummy* de la primera variable (bloque de variables *dummy*) por cada *dummy* de la segunda variable. Aprenderemos a formar dichos bloques, a llevar a cabo las interacciones entre bloques de variables *dummy*, a estimar la ecuación de regresión, a conocer el cambio en R^2 para cada bloque de variables, a valorar la ecuación y usarla adecuadamente.

Objetivos

Cuando el estudiante finalice esta unidad sabrá:

- Transformar cada variable independiente de grupos en bloques de variables *dummy*.
- Generar la interacción de las variables independientes de grupos.
- Plantear la ecuación de regresión que incluya los 'bloques' y las variables principales anidadas en las interacciones.
- Estimar e interpretar los parámetros de ecuaciones de regresión con términos de interacción 'bloques'.
- Interpretar los estadísticos de conjunto de toda la ecuación de regresión, para cada 'bloque' de variables *dummy* y para la interacción de las variables.
- Llevar a cabo el correspondiente pronóstico para cada grupo y hacer la representación gráfica de los resultados obtenidos.

3.1. Regresión con interacción de una variable independiente de dos grupos y otra variable independiente de tres. Ecuación de conjunto y ecuación para cada grupo. Interpretación. Representación gráfica

Poner a prueba la hipótesis de la interacción entre variables de grupos es muy conveniente por dos motivos: (a) los valores pronosticados (medias de cada respectivo grupo) reproducen las medias reales de la variable dependiente en cada grupo, y (b) la interacción entre variables proporciona una información privilegiada acerca de la naturaleza de los datos, y es que hay ocasiones en las cuales la relación entre los grupos de las variables independientes es más compleja que la aditiva proporcionada por la regresión solo con los efectos principales de las variables independientes. De hecho, muchos programas informáticos de análisis de la varianza (ANOVA) proporcionan por defecto la interacción de las variables; nosotros haremos en esta unidad la interacción entre variables de grupos desde una perspectiva del análisis de regresión.

Supongamos que estamos interesados en comprobar cuál es el efecto de la interacción entre las variables independientes 'Minority' y 'Jobcat' sobre 'Salary'. Ya sabemos que la variable 'Minority' tiene dos grupos: 0: 'No minority' y 1: 'Yes minority', mientras que hay tres grupos en la variable 'Jobcat', 1: 'Clerical', 2: 'Custodial' y 3: 'Manager'.

Utilizaremos el fichero 'Comp_dat_Dms.sav', puesto que allí tenemos ya las variables *dummy* de cada grupo guardadas.

Como indicábamos, se desea analizar:

$$\text{'Salary'} = f(\text{'Minority'} * \text{'Jobcat'}) \qquad (3.1)$$

La forma de representar la Ecuación 3.1 tiene dos características importantes:

a. Aparece el símbolo * entre las dos variables independientes, indicando que se pone a prueba la interacción (entendida como producto) de las mismas, pero además,
b. Cada vez que aparece una interacción entre variables, se han de incluir también todas las posibles interacciones de nivel inferior a las mismas hasta llegar a las correspondientes variables principales.

Así, en la Ecuación 3.1 se ha de poner a prueba, por defecto, la regresión de 'Salary' sobre 'Minority', sobre 'Jobcat' y sobre 'Minority'*'Jobcat', aunque con sus correspondientes variables *dummy*; es decir, la Ecuación 3.1 se transformaría en (desde una perspectiva de ecuación funcional):

$$\text{'Salary'} = \text{f('Minority', 'Jobcat', 'Minority'} * \text{'Jobcat')} \quad (3.2)$$

Pero para desarrollar correctamente esas dos ecuaciones, hemos de tener en cuenta que las variables 'Minority' y 'Jobcat' son variables de grupo. La variable 'Minority' es una variable de dos grupos que ya es *dummy* (con valores '0' y '1'), y 'Jobcat' es una variable con 3 categorías, por lo que está formada por dos variables *dummy*: 'D_JC_custodial' y 'D_JC_manager' (con el grupo 'Clerical' como grupo de referencia).

Para conseguir la interacción entre dos variables de grupos (formada cada una de ellas por variables *dummy*), hemos de multiplicar cada *dummy* de la primera variable por cada *dummy* de la segunda variable. Así pues, para conseguir las variables de interacción de 'Minority'*'Jobcat' hemos de multiplicar cada una de las variables *dummy* que forma la variable 'Minority' por cada una de las que forma la variable 'Jobcat' (dos *dummy*: ('D_JC_custodial' y 'D_JC_manager'), por tanto:

$$\text{'Minority'} * \left['\text{D}_{\text{JC}_{\text{custodial}}}', '\text{D}_{\text{JC}_{\text{manager}}}'\right]$$

$$= [\text{'Minority'} * \text{'D\_JC\_custodial'} + \text{'Minority'} * \text{'D\_JC\_manager'}] \quad (3.3)$$

Con lo cual, el desarrollo completo de la Ecuación 3.1, en álgebra de regresión quedaría:

$$\text{Salary} = \text{b0} + \text{b1} \cdot \text{Minority} + \left[\text{b2} \cdot \text{D}_{\text{JC}_{\text{custodial}}} + \text{b3} \cdot \text{D}_{\text{JC}_{\text{manager}}}\right]$$

$$+ [\text{b4} \cdot \text{'Minority'} * \text{'D\_JC\_custodial'} + \text{b5} \cdot \text{'Minority'} * \text{'D\_JC\_manager'}] + \text{e} \quad (3.4)$$

En la Ecuación 3.4, el término 'Minority' representa el efecto principal de dicha variable (que ya es *dummy*), mientras los términos del corchete 'D_JC_custodial' y 'D_JC_manager' representan el efecto principal de la variable 'Jobcat', y la interacción 'Minority'*'Jobcat' está representada en la Ecuación 3.4 por los términos dentro del segundo corchete: 'Minority'*'D_JC_custodial' y 'Minority'*'D_JC_manager'.

Por tanto, para poder estimar los parámetros de la Ecuación 3.4, se han de generar las interacciones 'Minority'*'D_JC_custodial' y 'Minority'*'D_JC_manager', para ello, en el fichero 'Comp_dat_Dms.sav', se corre la sintaxis:

```
COMPUTE MinXDcust=minority * D_JC_custodial.
    EXECUTE.
COMPUTE MinXDmngr=minority * D_JC_manager.
    EXECUTE.
```

Generándose las variables: ‘MinXDcust’ (acrónimo de ‘Minority’*‘D_JC_custodial’) y ‘MinXDmngr’ (‘Minority’*‘D_JC_manager’), como se muestra en la Figura 3.1.

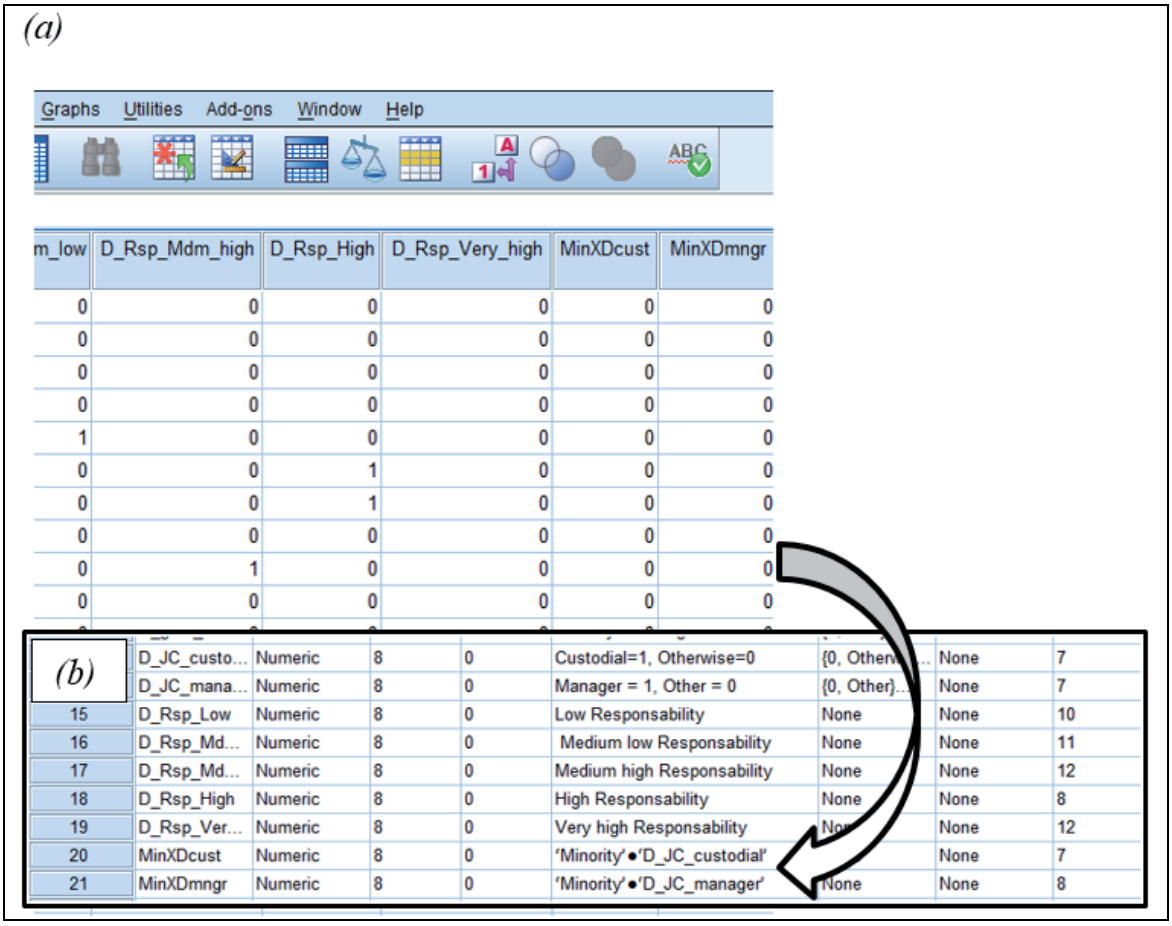


Figura 3.1. Variables de la interacción de ‘Minority’*‘D_JC_custodial’ y ‘Minority’*‘D_JC_manager’, (a) en el ‘Data View’ y (b) en el ‘Variable View’, tal como aparecen en el SPSS

IMPORTANTE
Guarda el nuevo fichero de datos, con las interacciones, con el nombre: ‘Comp_dat_Dms_U3.sav’, de este modo ya tenemos guardadas las variables *dummy* con las interacciones de ‘Minority’*’Jobcat’.

Una vez que ya tenemos las variables principales y las interacciones, podemos llevar a cabo la estimación de los parámetros de la regresión, para ello, se corre la sintaxis:

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA CHANGE
     /CRITERIA=PIN(.05) POUT(.10)
```

```
/NOORIGIN
/DEPENDENT salary
/METHOD=ENTER minority
/METHOD=ENTER D_JC_custodial D_JC_manager
/METHOD=ENTER MinXDcust MinXDmngr
/SAVE PRED RESID.
```

Donde se puede comprobar:

a. Que se ha pedido el cambio en R^2 (`CHANGE`) para cada bloque de variables,

b. se han introducido tres bloques de variables (`ENTER`) y

c. se guardan los valores pronosticados y los residuales (`SAVE PRED RESID`).

Los resultados de la regresión aparecen en la Tabla 3.1.

Tabla 3.1.
Resultados de la regresión: 'Salary' = f('Minority''Jobcat')*

(a)

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,165[a]	,027	,025	$16,093.082	,027	13,233	1	472	,000
2	,520[b]	,271	,266	$13,963.988	,243	78,452	2	470	,000
3	,522[c]	,273	,265	$13,976.331	,002	,585	2	468	,557

a. Predictors: (Constant), Minority Classification
b. Predictors: (Constant), Minority Classification, Custodial = 1, otherwise = 0, Manager = 1, Other = 0
c. Predictors: (Constant), Minority Classification, Custodial = 1, otherwise = 0, Manager = 1, Other = 0, MinXDmngr, MinXDcust
d. Dependent Variable: Current Salary

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3427267305	1	3427267304,649	13,233	,000[b]
	Residual	1,222E+11	472	258987282,785		
	Total	1,257E+11	473			
2	Regression	34022572657	3	11340857552,231	58,160	,000[c]
	Residual	91646692122	470	194992961,963		
	Total	1,257E+11	473			
3	Regression	34251167467	5	6850233493,395	35,069	,000[d]
	Residual	91418097312	468	195337814,770		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary
b. Predictors: (Constant), Minority Classification
c. Predictors: (Constant), Minority Classification, Custodial = 1, otherwise = 0, Manager = 1, Other = 0
d. Predictors: (Constant), Minority Classification, Custodial = 1, otherwise = 0, Manager = 1, Other = 0, MinXDmngr, MinXDcust

Tabla 3.1. *(Continuación)*
(c)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	59406,432	836,639		71,006	,000
	Minority Classification	-6497,490	1786,121	-,165	-3,638	,000
2	(Constant)	54929,650	826,562		66,456	,000
	Minority Classification	-2473,020	1594,463	-,063	-1,551	,122
	Custodial = 1, otherwise = 0	-2908,936	2812,079	-,041	-1,034	,301
	Manager = 1, Other = 0	21214,185	1718,221	,498	12,347	,000
3	(Constant)	54770,833	841,276		65,104	,000
	Minority Classification	-1810,374	1718,432	-,046	-1,054	,293
	Custodial = 1, otherwise = 0	-284,405	3828,896	-,004	-,074	,941
	Manager = 1, Other = 0	21489,417	1774,674	,504	12,109	,000
	MinXDcust	-5783,747	5650,814	-,058	-1,024	,307
	MinXDmngr	-3107,376	7364,047	-,017	-,422	,673

a. Dependent Variable: Current Salary

Antes de continuar, guardamos los pronósticos de la ecuación de regresión con el nombre de variable 'PRE_Sal_f_mntyXjbct', y con la etiqueta: 'Predicted Value of: 'Salary' = f('Minority'*'Jobcat')', como se muestra en la Figura 3.2.

(a)

D_Rsp_High	D_Rsp_Very_high	MinXDcust	MinXDmngr	PRE_Sal_f_mntyXjbct	RES_1
0	0	0	0	54770,83333	[illegible]
0	0	0	0	54770,83333	-15990,[illegible]
0	0	0	0	54770,83333	13139,16[illegible]
0	0	0	0	54770,83333	-4490,833[illegible]

(b)

17	D_Rsp_Mdm_low	Numeric	8	0	Medium low Responsability	None	None	11
18	D_Rsp_Mdm_high	Numeric	8	0	Medium high Responsability	None	None	12
19	D_Rsp_High	Numeric	8	0	High Responsability	None	None	8
20	D_Rsp_Very_high	Numeric	8	0	Very high Responsability	None	None	12
21	MinXDcust	Numeric	8	0	'Minority'●'D_JC_custodial'	None	None	7
22	MinXDmngr	Numeric	8	0	'Minority'●'D_JC_manager'	None	None	8
23	PRE_Sal_f_mntyXjbct	Numeric	11	5	Predicted Value of: 'Salary' = f('Minority'*'Jobcat')	None	None	15
24	RES_1	Numeric	11	5	Unstandardized Residual	None	None	13

Figura 3.2. Valores guardados en el SPSS de los valores pronosticados y de los errores de la regresión: 'Salary' = f('Minority'*'Jobcat'), (a) en el 'Data View' y (b) en el 'Variable View'

Comentaremos los resultados de la Tabla 3.1. En el apartado (a) se comprueba que la interacción de 'Minority' y 'Jobcat' es no significativa (corresponde al tercer bloque de variables), puesto que el cambio en $R^2 = .002$, con un valor de $F(2, 468) = .585$ $(p = .557)$; por lo

tanto, estadísticamente no es explicativo incluir la interacción de las dos variables en el modelo (pese a que toda la ecuación sí es significativa, como se indica en la Tabla 3.1, apartado (b), $F(5, 468) = 35.069$, $p < .001$).

> Pese a que la interacción es no significativa, continuaremos desarrollando el ejemplo a efectos expositivos. Por tanto, como se indica en la Unidad 2, el modelo estadísticamente más adecuado y parsimonioso es 'Salary' = f('Jobcat'), es decir, 'Salary' es función de 'Jobcat'.

En el caso de que aceptáramos la ecuación del modelo 3 de la Tabla 3.1 (con ecuación funcional: 'Salary' = f('Minority'*'Jobcat'), la ecuación de regresión con sus coeficientes sería la de la Ecuación 3.4, substituyendolos por los correspondientes coeficientes de la Tabla 3.1, apartado (c):

$$\text{Salary} = 54770.8 - 1810.4 * \text{Minority} + \left[-284.4 \cdot \text{D}_{\text{JC}_{\text{custodial}}} + 21489.4 \cdot \text{D}_{\text{JC}_{\text{manager}}}\right] + [-5783.7 \cdot \text{Minority} * \text{D\_JC\_custodial} - 3107.4 \cdot \text{Minority} * \text{D\_JC\_manager}] + e \quad (3.5)$$

En la Tabla 3.1 apartado (b) se ha comprobado cómo toda la Ecuación 3.4 del modelo 3 es significativa, pese a que no son significativas la variable 'Minority' ni la interacción 'Minority'*'Jobcat', lo cual indica que la capacidad explicativa de todo el modelo 3 viene dada por la significación del efecto de 'Jobcat' sobre 'Salary'.

En la Tabla 3.1 apartado (c) se comprueba que en el modelo 3 (como conjunto) la diferencia de medias entre 'Minority' y su grupo de referencia ('No minority') es no significativa (se podría quitar del modelo, salvo por razones teóricas muy consistentes). Veíamos en el cambio de R^2 que la variable 'Jobcat', en conjunto, es significativa, pero en esta Tabla 3.1 (c) se observa que el grupo 'Custodial' es estadísticamente no significativo ($p = .941$), indicando que las medias en 'Salary' de 'Custodial' y del grupo de referencia de esta variable ('Clerical') no difieren. Recuérdese que:

a. Este grupo se ha de dejar dentro de la ecuación (pese a no ser significativo) porque toda la variable es significativa (como se indica en su cambio de R^2 de la Tabla 3.1 *(a)*), y
b. no es correcto borrar este grupo y volver a estimar los valores de la regresión sin ese grupo (se cometería un sesgo de pronóstico de 'Custodial' y además la

estimación del modelo sería incorrecta al no tener en cuenta los verdaderos grados de libertad al eliminar un grupo).

La diferencia de medias de la variable dependiente 'Salary' entre los grupos 'Manager' y el de referencia ('Clerical') es significativa (t = 12.109, p < .001). Por último, no hay diferencias en las respectivas pendientes de 'Salary' según la variable 'Jobcat', pues todo el bloque de la interacción es no significativo estadísticamente, como tampoco lo son sus términos simples de interacción: 'Minority'*'Custodial' (p = .307), ni tampoco 'Minority'*'Manager' (p = .673), con respecto a la pendiente de referencia ('No minority'*'Clerical').

Aceptando (a efectos didácticos) el modelo de la Ecuación 3.4, calcularemos los valores pronosticados para cada grupo, y teniendo en cuenta que es un modelo de 2*3 grupos, ha de haber seis valores pronosticados:

I. Para el grupo 'No Minority' y 'Clerical' ('Minority' = 0, 'D_JC_custodial' = 0, 'D_JC_manager' = 0), la Ecuación 3.4 queda:

$$Salary'_{NM,Clrl} = 54770.8 + -1810.4 \cdot 0 + [-284.4 \cdot 0 + 21489.4 \cdot 0] + [-5783.7 \cdot 0 \cdot 0 - 3107.4 \cdot 0 \cdot 0] = \mathbf{54770.8} \quad (3.6)$$

II. Para el grupo 'No Minority' y 'Custodial' ('Minority' = 0, 'D_JC_custodial' = 1, 'D_JC_manager' = 0), la Ecuación 3.4 quedaría:

$$Salary'_{NM,Cust} = 54770.8 + -1810.4 \cdot 0 + [-284.4 \cdot 1 + 21489.4 \cdot 0] + [-5783.7 \cdot 0 \cdot 1 - 3107.4 \cdot 0 \cdot 0] = \mathbf{54486.4} \quad (3.7)$$

III. Para el grupo 'No Minority' y 'Manager'

$$Salary'_{NM,Mngr} = 54770.8 + -1810.4 \cdot 0 + [-284.4 \cdot 0 + 21489.4 \cdot 1] + [-5783.7 \cdot 0 \cdot 0 - 3107.4 \cdot 0 \cdot 1] = \mathbf{76260.2} \quad (3.8)$$

IV. Grupo 'Yes minority' y 'Clerical',

$$Salary'_{YM,Clrl} = 54770.8 + -1810.4 \cdot 1 + [-284.4 \cdot 0 + 21489.4 \cdot 0] + [-5783.7 \cdot 1 \cdot 0 - 3107.4 \cdot 1 \cdot 0] = \mathbf{52960.2} \quad (3.9)$$

V. El grupo 'Yes Minority' y 'Custodial',

$$Salary'_{YM,Cust} = 54770.8 + -1810.4 \cdot 1 + [-284.4 \cdot 1 + 21489.4 \cdot 0] + [-5783.7 \cdot 1 \cdot 1 - 3107.4 \cdot 1 \cdot 0] = \mathbf{46892.3} \quad (3.10)$$

VI. Para el grupo 'Yes Minority' y 'Manager' ('Minority' = 1, 'D_JC_custodial' = 0, 'D_JC_manager' = 1):

$$Salary'_{YM,Mngr} = 54770.8 + -1810.4 \cdot 1 + [-284.4 \cdot 0 + 21489.4 \cdot 1]$$
$$+[-5783.7 \cdot 1 \cdot 0 - 3107.4 \cdot 1 \cdot 1] = \mathbf{71342.4} \quad (3.11)$$

Se puede comprobar que los resultados son correctos pidiendo la tabla de valores pronosticados, guardados en el fichero de datos del SPSS como 'PRE_Sal_f_mntyXjbct', en función de 'Minority' y de 'Jobcat', para ello se corre la sintaxis:

```
SUMMARIZE
     /TABLES=PRE_Sal_f_mntyXjbct BY minority BY jobcat
     /FORMAT=NOLIST TOTAL
     /TITLE='Case Summaries'
     /MISSING=VARIABLE
     /CELLS=MEAN STDDEV COUNT.
```

Que genera la Tabla 3.2, donde se ven los valores pronosticados de 'Salary' en función de 'Minority' en interacción con 'Jobcat'. Nuevamente, se observa que los valores de las desviaciones típicas de los valores predichos son igual a cero, porque para cada grupo solo aparece un valor pronosticado.

Tabla 3.2.
Medias de los valores predichos de 'Salary' = f('Minority''Jobcat')*

Case Summaries

'Predicted Value of: 'Salary' = f('Minority'*'Jobcat')'

Minority Classification	Employment Category	Mean	Std. Deviation	N
No	Clerical	54770,83333	0E-8	276
	Custodial	54486,42857	0E-8	14
	Manager	76260,25000	0E-8	80
	Total	59406,43243	8864,209952	370
Yes	Clerical	52960,45977	0E-8	87
	Custodial	46892,30769	0E-8	13
	Manager	71342,50000	0E-8	4
	Total	52908,94231	4215,106554	104
Total	Clerical	54336,94215	773,8820414	363
	Custodial	50830,00000	3866,737085	27
	Manager	76026,07143	1053,568426	84
	Total	57980,82278	8509,560283	474

Si hacemos la representación gráfica de los valores pronosticados mediante figura de líneas, hemos de correr la sintaxis:

```
GRAPH
     /LINE(MULTIPLE)=MEAN(PRE_Sal_f_mntyXjbct) BY
          minority BY jobcat.
```

Lo que genera la Figura 3.3, en la que se observa que las tres líneas que representan a las categorías de 'Jobcat' (es decir, 'Clerical', 'Custodial' y 'Manager') no son geométricamente paralelas, aunque esta falta de paralelismo no es estadísticamente significativa (no se rechaza un relativo paralelismo entre ellas, más allá de lo que cabría esperar por azar). En esta figura se observa que los valores pronosticados coinciden con las medias reales de cada respectivo grupo, como puede verse en la Figura 3.3, o incluso pidiendo la tabla de medias y desviaciones típicas por medio de la sintaxis:

```
SUMMARIZE
     /TABLES=salary BY minority BY jobcat
     /FORMAT=NOLIST TOTAL
     /TITLE='Case Summaries'
     /MISSING=VARIABLE
     /CELLS=MEAN STDDEV COUNT.
```

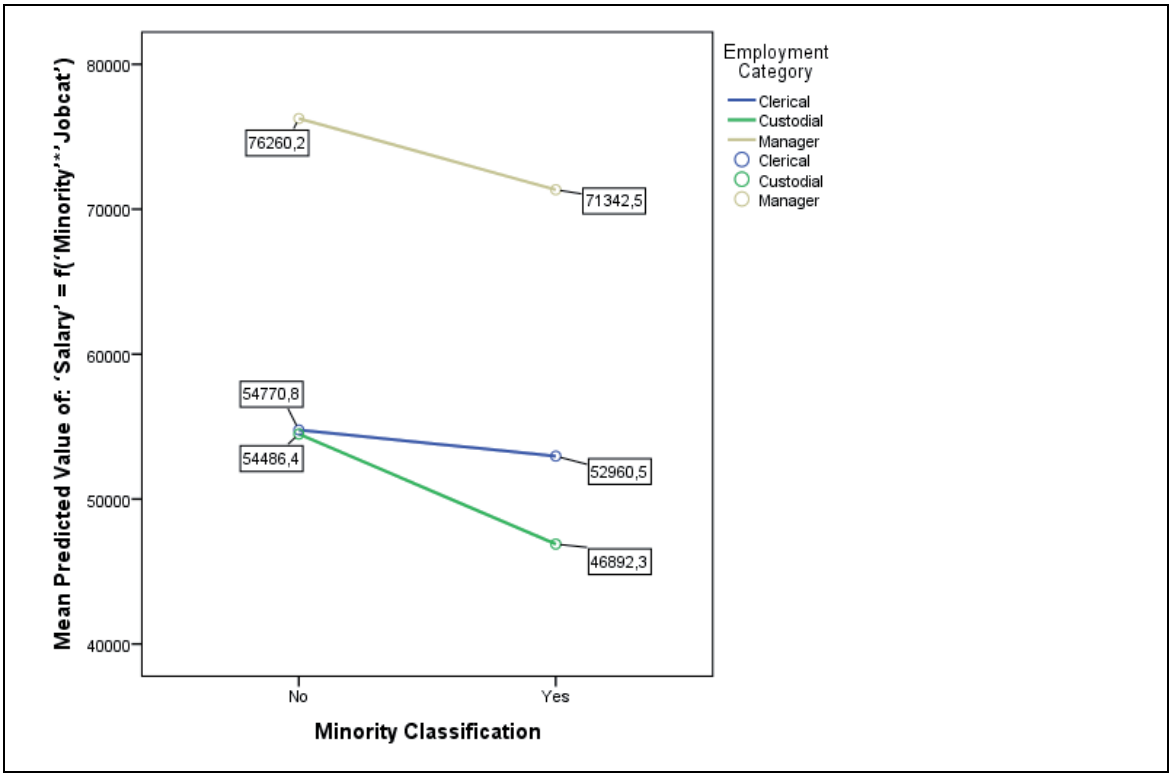


Figura 3.3. Medias de los valores predichos de 'Salary' = f('Minority'*'Jobcat')

Se obtiene la Tabla 3.3, donde se comprueba que las medias de los valores originales coinciden con las medias de los valores pronosticados (ver Tabla 3.2).

Tabla 3.3.
Medias de los valores reales de 'Salary' = f('Minority''Jobcat')*

Case Summaries

Current Salary

Minority Classification	Employment Category	Mean	Std. Deviation	N
No	Clerical	$54,770.83	$13,341.785	276
	Custodial	$54,486.43	$11,611.235	14
	Manager	$76,260.25	$15,116.967	80
	Total	$59,406.43	$16,275.969	370
Yes	Clerical	$52,960.46	$15,248.025	87
	Custodial	$46,892.31	$13,317.636	13
	Manager	$71,342.50	$13,388.267	4
	Total	$52,908.94	$15,420.091	104
Total	Clerical	$54,336.94	$13,822.362	363
	Custodial	$50,830.00	$12,814.845	27
	Manager	$76,026.07	$15,003.277	84
	Total	$57,980.82	$16,299.863	474

Resumiendo, cuando se utiliza un modelo con dos variables independientes de grupos con interacción de variables, los valores pronosticados coinciden con los valores originales, aunque haya algún efecto del modelo de interacción que no sea significativo, y que estadísticamente hayan de quitarse del modelo los efectos no significativos, porque no aporta capacidad explicativa a los resultados del modelo final, como es nuestro caso, donde una vez puesto a prueba el modelo de interacción: 'Salary' = f('Minority'*'Jobcat'), se ha comprobado que el modelo estadísticamente más parsimonioso de acuerdo a los resultados sería el de: 'Salary' = f('Jobcat').

Téngase en cuenta que si la interacción 'Minority'*'Jobcat' hubiese sido significativa, también tendríamos que haber dejado en el modelo 'Minority' y 'Jobcat', aunque una o ambas hubiesen sido no significativas estadísticamente, respondiendo al principio de jerarquía de variables en la interacción.

3.2. Regresión con interacción de una variable independiente de dos grupos y otra de seis grupos. Ecuación de conjunto y ecuación para cada grupo

En el supuesto de que se desee poner a prueba la hipótesis de la interacción:

$$\text{'Salary'} = \text{f('Gender'} * \text{'Center')} \tag{3. 12}$$

Se tendría una regresión por grupos de dimensión 2*6, es decir, la primera variable independiente tiene dos grupos, mientras la segunda es de seis grupos, lo que da un total de doce grupos. Recuérdese que la Ecuación funcional 3.12 se desglosa en la siguiente:

$$\begin{aligned}\text{'Salary'} = \text{f('D\_gndr\_fem', ['D\_Ctr\_2', 'D\_Ctr\_3', 'D\_Ctr\_4', 'D\_Ctr\_5', 'D\_Ctr\_6'],}\\ \text{['D\_gndr\_fem'} * \text{'D\_Ctr\_2', 'D\_gndr\_fem'} * \text{'D\_Ctr\_3', 'D\_gndr\_fem'}\\ * \text{'D\_Ctr\_4', 'D\_gndr\_fem'} * \text{'D\_Ctr\_5', 'D\_gndr\_fem'} * \text{'D\_Ctr\_6'])}\end{aligned} \tag{3. 13}$$

En donde se comprueba que la primera línea es la variable *dummy* de 'Gender', la segunda y tercera líneas, dentro de un solo corchete, corresponden a la variable 'Center' desarrollada en dummies, y el resto, las tres últimas líneas dentro de un corchete, serían las variables de interacción de 'Gender' con 'Center'. Téngase en cuenta que para conseguir las variables de interacción 'Gender'*'Center' se ha de multiplicar cada variable *dummy* de 'Gender' por cada una de las variables *dummy* de 'Center'.

La Ecuación 3.13, si se escribe mediante álgebra de ecuaciones lineales se desarrolla del siguiente modo:

$$\begin{aligned}\text{Salary} = b_0 + b_1 \cdot D_{\text{gndr}_{\text{fem}}}\\ + \left[b_2 \cdot D_{\text{Ctr}_2} + b_3 \cdot D_{\text{Ctr}_3} + b_4 \cdot D_{\text{Ctr}_4} + b_5 \cdot D_{\text{Ctr}_5} + b_6 \cdot D_{\text{Ctr}_6}\right]\\ +[b_7 \cdot D_{\text{gndr}_{\text{fem}}} * D_{\text{Ctr}_2} + b_8 \cdot D_{\text{gndr}_{\text{fem}}} * D_{\text{Ctr}_3} + b_9 \cdot D_{\text{gndr}_{\text{fem}}} * D_{\text{Ctr}_4}\\ + b_{10} \cdot \text{D\_gndr\_fem} * \text{D\_Ctr\_5} + b_{11} \cdot \text{D\_gndr\_fem} * \text{D\_Ctr\_6}] + e\end{aligned} \tag{3. 14}$$

Para llevar a cabo las interacciones de las variables *dummy* que componen la interacción de 'Gender' con 'Center', se han de generar las variables de las tres últimas líneas de la Ecuación 3.14, dentro del último corchete, mediante la sintaxis:

```
COMPUTE D_femXD_Ctr_2=D_gndr_fem * D_Ctr_2.
     VARIABLE LABELS  D_femXD_Ctr_2
     'COMPUTE D_femXD_Ctr_2=D_gndr_fem * D_Ctr_2'.
COMPUTE D_femXD_Ctr_3=D_gndr_fem * D_Ctr_3.
     VARIABLE LABELS  D_femXD_Ctr_3
     'COMPUTE D_femXD_Ctr_3=D_gndr_fem * D_Ctr_3'.
COMPUTE D_femXD_Ctr_4=D_gndr_fem * D_Ctr_4.
```

```
     VARIABLE LABELS  D_femXD_Ctr_4
     'COMPUTE D_femXD_Ctr_4=D_gndr_fem * D_Ctr_4'.
COMPUTE D_femXD_Ctr_5=D_gndr_fem * D_Ctr_5.
     VARIABLE LABELS  D_femXD_Ctr_5
     'COMPUTE D_femXD_Ctr_5=D_gndr_fem * D_Ctr_5'.
COMPUTE D_femXD_Ctr_6=D_gndr_fem * D_Ctr_6.
     VARIABLE LABELS  D_femXD_Ctr_6
     'COMPUTE D_femXD_Ctr_6=D_gndr_fem * D_Ctr_6'.
EXECUTE.
```

Con lo cual se han generado cinco nuevas variables *dummy*, que se corresponden al producto de las dummies de 'Gender'*'Center', que aparecen el 'Data View' y en el 'Variable View' del SPSS como se muestra en la Figura 3.4.

IMPORTANTE
Guarda el fichero de datos con las interacciones como: 'Comp_dat_Dms_U3.sav', de este modo ya tenemos guardadas las variables *dummy* con las interacciones de 'Gender'*'Center'.

(a)

D_femXD_Ctr_2	D_femXD_Ctr_3	D_femXD_Ctr_4	D_femXD_Ctr_5	D_femXD_Ctr_6
0	0	0	0	0
0	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
1	0	0	0	0

(b)

	MinXDcust	Numeric	8	0		
	MinXDmngr	Numeric	8	0		None
22	PRE_Sal_f_mntyXjbct	Numeric	11	5	Predicted Value of: 'Salary' = f('Minority'*'Jobcat')	None
23	RES_1	Numeric	11	5	Unstandardized Residual	None
24	D_femXD_Ctr_2	Numeric	8	0	COMPUTE D_femXD_Ctr_2=D_gndr_fem * D_Ctr_2	None
25	D_femXD_Ctr_3	Numeric	8	0	COMPUTE D_femXD_Ctr_3=D_gndr_fem * D_Ctr_3	None
26	D_femXD_Ctr_4	Numeric	8	0	COMPUTE D_femXD_Ctr_4=D_gndr_fem * D_Ctr_4	None
27	D_femXD_Ctr_5	Numeric	8	0	COMPUTE D_femXD_Ctr_5=D_gndr_fem * D_Ctr_5	None
28	D_femXD_Ctr_6	Numeric	8	0	COMPUTE D_femXD_Ctr_6=D_gndr_fem * D_Ctr_6	None

Figura 3.4. Variables *dummy* de la interacción de generadas para la interacción de 'Gender'*'Center', como aparecen (a) en el 'Data View' y (b) en el 'Variable View' del SPSS

Para llevar a cabo la estimación de los parámetros de la regresión de la Ecuación 3.14, se ha de hacer un bloque por cada variable primitiva, tal como aparece en dicha ecuación, mediante la sintaxis:

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA CHANGE
     /CRITERIA=PIN(.05) POUT(.10)
     /NOORIGIN
     /DEPENDENT salary
     /METHOD=ENTER D_gndr_fem
     /METHOD=ENTER D_Ctr_2 D_Ctr_3 D_Ctr_4 D_Ctr_5
          D_Ctr_6
     /METHOD=ENTER D_femXD_Ctr_2 D_femXD_Ctr_3
          D_femXD_Ctr_4 D_femXD_Ctr_5 D_femXD_Ctr_6
     /SAVE PRED.
```

Dando los resultados de la Tabla 3.4, que contiene: en el apartado (a) el sumario de los modelos 1, 2 y 3, y el cambio de R^2, en el (b) el ANOVA de cada modelo, y en el (c) las correspondientes ecuaciones de regresión para cada modelo, aunque hemos puesto solo los parámetros del modelo 3, suprimiendo los parámetros correspondientes a los modelos 1 y 2, con el fin de ahorrar espacio. Ampliaremos estos aspectos a continuación.

Tabla 3.4.
Resultados de la regresión de 'Salary' = f('Gender''Center'), de acuerdo con la Ecuación 3.14*

(a)

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,452[a]	,204	,202	$14,557.154	,204	121,029	1	472	,000
2	,626[b]	,392	,384	$12,795.412	,188	28,785	5	467	,000
3	,644[c]	,415	,401	$12,617.007	,023	3,660	5	462	,003

a. Predictors: (Constant), D_gndr_fem

b. Predictors: (Constant), D_gndr_fem, Center 3, Center 5, Center 6 , Center 2, Center 4

c. Predictors: (Constant), D_gndr_fem, Center 3, Center 5, Center 6 , Center 2, Center 4, COMPUTE D_femXD_Ctr_6=D_gndr_fem * D_Ctr_6, COMPUTE D_femXD_Ctr_4=D_gndr_fem * D_Ctr_4, COMPUTE D_femXD_Ctr_5=D_gndr_fem * D_Ctr_5, COMPUTE D_femXD_Ctr_3=D_gndr_fem * D_Ctr_3, COMPUTE D_femXD_Ctr_2=D_gndr_fem * D_Ctr_2

d. Dependent Variable: Current Salary

Tabla 3.4. *(Continuación)*
(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25647398289,106	1	25647398289,11	121,029	,000[b]
	Residual	100021866490,008	472	211910734,089		
	Total	125669264779,114	473			
2	Regression	49210823786,672	6	8201803964,445	50,096	,000[c]
	Residual	76458440992,442	467	163722571,718		
	Total	125669264779,114	473			
3	Regression	52124006966,842	11	4738546087,895	29,767	,000[d]
	Residual	73545257812,272	462	159188869,724		
	Total	125669264779,114	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), D_gndr_fem

c. Predictors: (Constant), D_gndr_fem, Center 3, Center 5, Center 6 , Center 2, Center 4

d. Predictors: (Constant), D_gndr_fem, Center 3, Center 5, Center 6 , Center 2, Center 4, COMPUTE D_femXD_Ctr_6=D_gndr_fem * D_Ctr_6, COMPUTE D_femXD_Ctr_4=D_gndr_fem * D_Ctr_4, COMPUTE D_femXD_Ctr_5=D_gndr_fem * D_Ctr_5, COMPUTE D_femXD_Ctr_3=D_gndr_fem * D_Ctr_3, COMPUTE D_femXD_Ctr_2=D_gndr_fem * D_Ctr_2

(c)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
.... //	Constant	.... //	.... //	.... //	.... //	.... // ....
3	Constant	53280,000	8921,571		5,972	,000
	D_gndr_fem	-6745,098	9094,822	-,206	-,742	,459
	Center 2	-2455,556	9863,168	-,059	-,249	,804
	Center 3	2994,000	9214,159	,065	,325	,745
	Center 4	3616,087	9018,024	,092	,401	,689
	Center 5	13557,167	9069,045	,328	1,495	,136
	Center 6	26699,385	9057,786	,571	2,948	,003
	COMPUTE D_femXD_Ctr_2=D_gndr_fem * D_Ctr_2	5809,172	10117,743	,134	,574	,566
	COMPUTE D_femXD_Ctr_3=D_gndr_fem * D_Ctr_3	2097,508	9597,075	,035	,219	,827
	COMPUTE D_femXD_Ctr_4=D_gndr_fem * D_Ctr_4	582,344	9884,930	,006	,059	,953
	COMPUTE D_femXD_Ctr_5=D_gndr_fem * D_Ctr_5	-6307,230	9513,362	-,096	-,663	,508
	COMPUTE D_femXD_Ctr_6=D_gndr_fem * D_Ctr_6	-31449,287	12835,859	-,125	-2,450	,015

a. Dependent Variable: Current Salary

IMPORTANTE
En el fichero de datos, guarda PRE_1 con el nombre 'PRE_Sal_f_GndrXCntr' ('Predicted values of Salary in function of 'Gender'*'Center'), y añade la etiqueta a esa misma variable: 'Predicted Values of 'Salary' = f('Gender'*'Center')', con el fin de identificar de dónde provienen estos resultados.

La Tabla 3.4 (a) indica que tanto el bloque 1, como el 2 y el 3, cuando son introducidos en la ecuación de cada respectivo modelo, son significativos, como indica el cambio de R^2. En el apartado (b) se comprueba cómo todo el modelo 3 es significativo. Pero en el apartado (c), si nos fijamos en el coeficiente de la variable 'Gender', este resulta no

significativo (*b = -6745,098, se(b) = 9094,822, p = .459*), y tal vez tampoco sea significativa en esta ecuación el segundo bloque de variables, formado por las categorías de la variable 'Center', pero han de dejarse estas variables, puesto que tanto 'Gender' como 'Center' están anidadas bajo la interacción de 'Gender'*'Center', y esta interacción es significativa.

Para llevar a cabo una regresión con interacción de variables rigen, al menos, dos principios:

1. Si se introduce una interacción de variables, han de incluirse, necesariamente, todas las variables e interacciones de menor nivel que estén anidadas en la interacción de mayor nivel.

   Es decir, supongamos la interacción de dos variables A y B, tal que la nueva variable sea A*B; en el modelo se han de incluir las variables A, B y A*B.

   Si consideramos la interacción A*B*C, en el modelo se han de incluir las variables A, B, C y A*B, A*C, B*C, A*B*C.

2. La significación de la interacción viene dada por la interacción de nivel más alto, las demás se incluyen en el pronóstico, aunque sean no significativas.

   Supongamos que en una regresión se plantea la interacción A*B*C, y esta es significativa, el modelo ha de incluir, además, las variables A, B, C, A*B, A*C, B*C, aunque estas (ninguna, alguna o todas) sean no significativas.

Si no se consideran las premisas que acaban de indicarse, el pronóstico de la variable dependiente no sería correcto.

Por lo tanto, al ser 'Gender'*'Center' estadísticamente significativa, se han de dejar también en la ecuación final las variables 'Gender' y 'Center', aunque fuesen estadísticamente no significativas.

Y la ecuación de regresión correspondiente a los resultados de la Tabla 3.4 es:

$$\begin{aligned}\text{Salary} = {} & 53280{,}0 - 6745{,}1 \cdot \text{D\_gndr\_fem} + [-2455{,}6 \cdot \text{D\_Ctr\_2} + 2994{,}0 \cdot \text{D\_Ctr\_3} \\ & + 3616{,}1 \cdot \text{D\_Ctr\_4} + 13557{,}2 \cdot \text{D\_Ctr\_5} + 26699{,}4 \cdot \text{D\_Ctr\_6}] \\ & + [5809{,}2 \cdot \text{D\_gndr\_fem} * \text{D\_Ctr\_2} + 2097{,}5 \cdot \text{D\_gndr\_fem} * \text{D\_Ctr\_3} \\ & + 582{,}3 \cdot \text{D\_gndr\_fem} * \text{D\_Ctr\_4} - 6307{,}2 \cdot \text{D\_gndr\_fem} \\ & * \text{D\_Ctr\_5}, - 31449{,}2 \cdot \text{D\_gndr\_fem} * \text{D\_Ctr\_6}] + e\end{aligned} \quad \textbf{(3. 15)}$$

El hecho de que la interacción sea significativa ha de interpretarse como que los grupos no siguen un modelo aditivo, por tanto, la figura de conjunto, si se unen las líneas correspondientes a los grupos, presentará una o varias líneas no paralelas. De momento, haremos los pronósticos para cada grupo y después la figura de conjunto.

Así, si se desea pronosticar el 'Salary' de una persona 'Male' y que trabaja en 'Center 1' (ambos grupos son de referencia en las varaibles dummy), tendrá como valores *dummy*:

'D_gndr_fem' = 0
'D_Ctr_2' = 0
'D_Ctr_3' = 0
'D_Ctr_4' = 0
'D_Ctr_5' = 0
'D_Ctr_6' = 0

'D_gndr_fem'*'D_Ctr_2' = 0*0 = 0
'D_gndr_fem'*'D_Ctr_3' = 0*0 = 0
'D_gndr_fem'*'D_Ctr_4' = 0*0 = 0
'D_gndr_fem'*'D_Ctr_5' = 0*0 = 0
D_gndr_fem'*'D_Ctr_6' = 0*0 = 0

Por tanto, en la Ecuación 3.15:

$$Salary'_{M,Cntr1} = 53280{,}0 - 6745{,}1 \cdot 0 + [-2455{,}6 \cdot 0 + 2994{,}0 \cdot 0 + 3616{,}1 \cdot 0 + 13557{,}2 \cdot 0 + 26699{,}4 \cdot 0] + [5809{,}2 \cdot 0 * 0 + 2097{,}5 \cdot 0 * 0 + 582{,}3 \cdot 0 * 0 - 6307{,}2 \cdot 0 * 0, -31449{,}2 \cdot 0 * 0] = \mathbf{53280{,}0} \quad (3.16)$$

Comprobándose que el grupo de referencia tiene el valor de la ordenada en el origen.

Si se desea pronosticar el valor del grupo 'Male' con 'Center 5', se tendrá que sustituir el valor de 0 en todas las categorías, excepto en: 'D_Ctr_5' = 1, por tanto:

$$Salary'_{M,Cntr5} = 53280{,}0 - 6745{,}1 \cdot 0 + [-2455{,}6 \cdot 0 + 2994{,}0 \cdot 0 + 3616{,}1 \cdot 0 + 13557{,}2 \cdot \mathbf{1} + 26699{,}4 \cdot 0] + [5809{,}2 \cdot 0 \cdot 0 + 2097{,}5 \cdot 0 * 0 + 582{,}3 \cdot 0 \cdot 0 - 6307{,}2 \cdot 0 \cdot \mathbf{1}, -31449{,}2 \cdot 0 \cdot 0] = 53280{,}0 + 13557{,}2 = \mathbf{66837{,}2} \quad (3.17)$$

En el caso de pronosticar el valor de 'Female' con 'Center 3', entonces los valores *dummy* serán:

'D_gndr_fem' = 1
'D_Ctr_2' = 0
'D_Ctr_3' = 1
'D_Ctr_4' = 0
'D_Ctr_5' = 0
'D_Ctr_6' = 0

'D_gndr_fem'*'D_Ctr_2' = 0*0 = 0
'D_gndr_fem'*'D_Ctr_3' = 1*1 = 1
'D_gndr_fem'*'D_Ctr_4' = 0*0 = 0
'D_gndr_fem'*'D_Ctr_5' = 0*0 = 0
D_gndr_fem'*'D_Ctr_6' = 0*0 = 0

y el pronóstico será:

$$Salary'_{F,Cntr3} = 53280{,}0 - 6745{,}1 \cdot \mathbf{1} + [-2455{,}6 \cdot 0 + 2994{,}0 \cdot \mathbf{1} + 3616{,}1 \cdot 0 + 13557{,}2 \cdot 0 + 26699{,}4 \cdot 0] + [5809{,}2 \cdot 0 \cdot 0 + 2097{,}5 \cdot \mathbf{1} \cdot \mathbf{1} + 582{,}3 \cdot 0 \cdot 0 - 6307{,}2 \cdot 0 \cdot 0, -31449{,}2 \cdot 0 \cdot 0] = 53280{,}0 - 6745{,}1 + 2994{,}0 + 2097{,}5 = \mathbf{51626{,}4} \quad (3.18)$$

Los valores pronosticados pueden representarse en una tabla mediante la sintaxis:

```
SUMMARIZE
     /TABLES= PRE_Sal_f_GndrXCntr BY gender BY Center
     /FORMAT=NOLIST TOTAL
     /TITLE='Case Summaries'
     /MISSING=VARIABLE
     /CELLS=MEAN STDDEV COUNT.
```

Que da como resultado la Tabla 3.5.

Tabla 3.5.
Valores pronosticados de 'Salary' = f('Gender''Center') , conforme a la Ecuación 3.15*

Case Summaries

'Predicted Values of 'Salary' = f('Gender'*'Center')'

Gender	Work center	Mean	Std. Deviation	N
Female	Center 1	46534,90196	0E-8	51
	Center 2	49888,51852	0E-8	81
	Center 3	51626,41026	0E-8	39
	Center 4	50733,33333	0E-8	12
	Center 5	53784,83871	0E-8	31
	Center 6	41785,00000	0E-8	2
	Total	49941,57407	2460,181105	216
Male	Center 1	53280,00000	0E-8	2
	Center 2	50824,44444	0E-8	9
	Center 3	56274,00000	0E-8	30
	Center 4	56896,08696	0E-8	92
	Center 5	66837,16667	0E-8	60
	Center 6	79979,38462	0E-8	65
	Total	64711,35659	9897,396407	258
Total	Center 1	46789,43396	1297,623631	53
	Center 2	49982,11111	282,3507748	90
	Center 3	53647,10145	2320,821365	69
	Center 4	56185,00000	1978,445359	104
	Center 5	62390,76923	6220,173500	91
	Center 6	78839,25373	6548,801460	67
	Total	57980,82278	10497,55906	474

Nótese que coinciden los tres valores esperados obtenidos en las Ecuaciones 3.16 a 3.18 con los de la Tabla 3.5. Del mismo modo, hemos indicado que al pronosticar valores esperados de la regresión con interacción de variables, se reproducen los valores medios de cada respectivo grupo, se puede comprobar fácilmente si en la sintaxis utilizada para generar la Tabla 3.5 se usa 'Salary' en lugar de 'PRE_Sal_f_GndrXCntr'.

Se puede observar la interacción gráfica de las variables independientes corriendo la sintaxis:

```
GRAPH
     /LINE(MULTIPLE)=MEAN(PRE_Sal_f_GndrXCntr) BY
          Center BY gender.
```

Que produce la Figura 3.5, donde se comprueba que no hay un 'paralelismo' en segmentos homólogos, sobre todo en 'Center 6', se observa que el grupo 'Male' tiene el valor más alto de 'Predicted salary', dándose la paradoja de que en el grupo 'Female' se dan los ingresos esperados más bajos cuando están en el 'Center 6' (los datos son ficticios, pero reflejan la interacción 'Gender'*'Center').

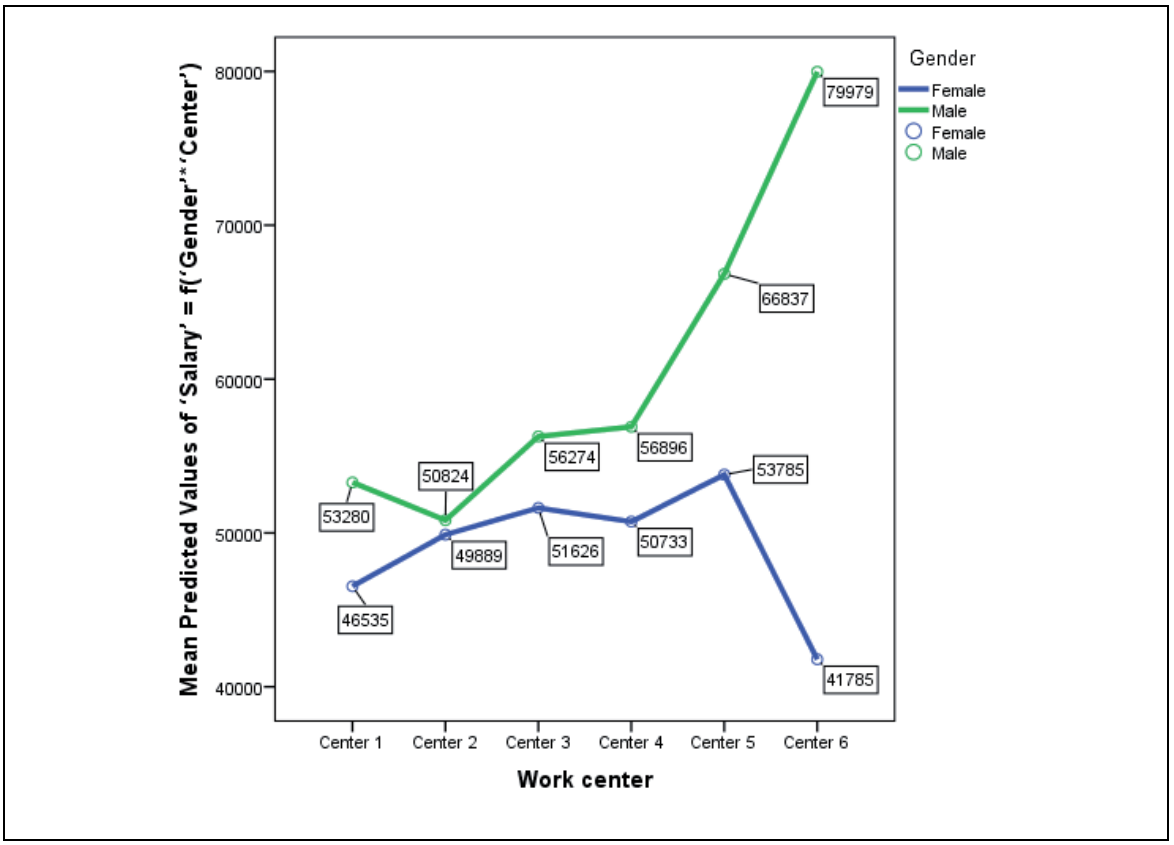


Figura 3.5. Valores pronosticados de 'Salary' = f('Gender'*'Center') , conforme a la Ecuación 3.15

IMPORTANTE
Guarda el fichero de datos (con las nuevas dummies y con los valores pronosticados y los residuales de las regresiones en el fichero), ponle de nombre 'Comp_dat_Dms_U3_1.sav'.
Borra las variables con los valores pronosticados, de modo que queden las variables primitivas y las variables dummies correspondientes a los grupos y a sus interacciones, nombra este fichero 'Comp_dat_Dms_U3_2.sav'.

3.3. Adenda: Relación entre la prueba F del análisis de la varianza (ANOVA) de dos factores, con interacción de los mismos, y la regresión con interacción de dos variables de grupos

En el caso de llevar a cabo una comparación de medias de dos factores mediante el ANOVA, en SPSS se puede llevar a cabo por medio de la sintaxis:

```
UNIANOVA salary BY gender Center
    /METHOD=SSTYPE(3)
    /INTERCEPT=INCLUDE
    /CRITERIA=ALPHA(0.05)
    /DESIGN=gender Center gender*Center.
```

Se obtendrían los resultados de la Tabla 3.6, en la que se observa que la significación de conjunto $F(11, 474) = 29.767$, $p < .001$ es la misma que la obtenida mediante la regresión, como se comprueba en la Tabla 3.4 (b) modelo 3.

Tabla 3.6.
Resultados del análisis de la varianza de 'Salary' = f('Gender''Center')*

Tests of Between-Subjects Effects

Dependent Variable: Current Salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5,212E+10[a]	11	4738546088	29,767	,000
Intercept	3,187E+11	1	3,187E+11	2001,752	,000
gender	3574580402	1	3574580402	22,455	,000
center	3488729129	5	697745825,7	4,383	,001
gender * center	2913183180	5	582636636,0	3,660	,003
Error	73545257812	462	159188869,7		
Total	1,719E+12	474			
Corrected Total	1,257E+11	473			

a. R Squared = ,415 (Adjusted R Squared = ,401)

Del mismo modo, en la Tabla 3.6, el valor de la interacción 'Gender'*'Center' da: $F(5, 474) = 3.660$, $p = .003$, que coincide con el resultado de la mejora mediante el bloque 3 de la Tabla 3.4 en el apartado (a). No tienen por qué coincidir en la Tabla 3.4 el valor de la significación de la 'Constant' ($p < .001$) con el valor de la significación del 'Intercept' ($p < .001$) de la Tabla 3.6, así como la significación de las variables principales 'Gender' y 'Center' en ambas tablas porque son sensibles tanto a la co-dificación utilizada en la regresión, como a los grupos de referencia utilizados; en la regresión hemos hecho codificación *dummy*, con grupos de referencia de la categoría

‘Male’ en la variable ‘Gender’ y de ‘Center 1’ en la variable ‘Center’, pero tendríamos que mirar el manual del programa estadístico utilizado, el SPSS en nuestro caso, para comprobar qué grupos de referencia y qué codificación usa el módulo `UNIANOVA` del SPSS.

Resumiendo, la regresión con interacción de variables de grupos y la comparación de medias mediante el ANOVA de Fisher dan la misma información básica, sobre todo en los parámetros más importantes: en el estadístico F de conjunto y en el de la interacción de las variables, los otros parámetros pueden variar según los grupos de referencia y del sistema de codificación utilizados para identificar cada grupo.

Ambos sistemas tienen ventajas e inconvenientes, como hemos indicado; la gran ventaja de la regresión (módulo `REGRESSION` en el SPSS) es el control para el pronóstico (recuérdese que en la interacción, los valores esperados para cada grupo son los de las medias reales de cada respectivo grupo), y el inconveniente es el de la realización de los códigos *dummy* por parte del usuario y los productos de las respectivas interacciones. En cambio, la ventaja del análisis de la varianza de Fisher (módulo `UNIANOVA` del SPSS) es la de la rapidez del análisis (el programa elabora automáticamente los códigos de cada grupo con algoritmos incorporados por defecto y también las interacciones correspondientes), mientras que para poder hacer pronósticos ha de conocerse el sistema de codificación.

3.4. Conclusiones

En esta unidad se ha introducido la interacción de dos variables independientes de grupos, poniendo como ejemplos la interacción de dos variables con 2*3 grupos y la interacción de otras dos variables de 2*6 grupos, donde el lector habrá aprendido:

- La forma de organizar la interacción de variables independientes de grupos, preparando las variables de interacción para introducirlas como 'bloques' en la regresión.
- De qué forma se ha de plantear la ecuación de regresión con los respectivos 'bloques' de las variables independientes de grupos y su correspondiente interacción.
- El principio de jerarquía de la interacción en regresión establece que si se introduce la interacción entre dos o más variables independientes, entonces se han de incluir en la regresión todas las interacciones y variables principales que haya anidadas bajo la interacción de orden superior.
- Cómo estimar los parámetros de la ecuación de regresión.
- Cuál es la correcta interpretación de los estadísticos de conjunto de cada modelo de regresión introducido por 'bloques'.
- Los dos parámetros más importantes en la ecuación de regresión son la significación de: (a) el valor de la F de conjunto de la ecuación final, y (b) de la interacción con más variables (mediante la significación del cambio de R^2 para la interacción de orden superior, tomada como un bloque de variables).
- Cuál es la interpretación de cada coeficiente de la ecuación de regresión.
- Cómo se ha de escribir la ecuación global de regresión, y la correspondiente ecuación para cada grupo.
- De qué modo se calculan los valores pronosticados para cada grupo.
- La interacción de variables reproduce las medias de los grupos originales.
- Lo anterior no implica que se haya de dejar la interacción estadística en la ecuación de regresión final, pues si los datos son estadísticamente aditivos, en lugar de multiplicativos (interacción no significativa estadísticamente), entonces no es necesario incluir la interacción.
- Hacer una representación gráfica de los resultados obtenidos.

- Cuál es la relación entre el análisis de regresión con interacción de variables de grupos y el ANOVA de Fisher, con la interacción de las mismas variables, y cómo son los correspondientes resultados, así como la interpretación de los mismos.
- Ventajas e inconvenientes del uso de la regresión (con interacción de variables) para variables independientes de grupos: la ventaja de la regresión es la capacidad de pronosticar mediante las codificaciones establecidas por el analista, y el inconveniente es el de la laboriosidad de la codificación y elaboración de las variables *dummy*.
- Ventajas e inconvenientes del uso del ANOVA de Fisher (con interacción de variables de grupos): tiene como ventaja la rapidez del análisis (cada programa estadístico realiza automáticamente la codificación y la interacción de variables), pero tiene el inconveniente del establecimiento del pronóstico, a partir de la ecuación utilizada, mediante los códigos utilizados para cada grupo.

Lecturas recomendadas

No hay libros específicos que estudien la interacción de variables de grupos desde la regresión, solo se centran en los efectos principales de los grupos o en el ANOVA; tan solo el libro de Cohen, Cohen, West y Aiken (2003), desde una perspectiva del ANOVA de Fisher, dedica el capítulo 9 (pp. 354-389) a la interacción con variables categóricas, aunque con un nivel más técnico que el usado aquí en la exposición del método. En el manual de Kirk (2012), en el capítulo 7, sobre «Aproximación del ANOVA al modelo lineal general» se exponen los distintos sistemas de codificación para su posterior análisis, pronóstico e interpretación. En Cohen (1968) y Tatsuoka (1975) se desarrolla la relación entre la regresión y el ANOVA.

Bibliografía

Cohen J. (1968). «Multiple regression as a general data-analytic system». *Psychollogical Bulletin, 70,* 426-443.

Cohen, J.; Cohen, P.; West, S. G. y Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Kirk, R. E. (2012). *Experimental design: procedures for the behavioral sciences*. Londres: Sage.

Tatsuoka, M. M. (1975). *The general linear model: A 'new' trend in analysis of variance.* Champaign, IL: The Institute for Personality and Ability Testing.

Actividades

1. Con los datos del fichero 'Comp_dat_Dms_U3.sav' se desea hacer la regresión 'Salbegin' = f('Minority'*'Center'), para ello, efectúa los siguientes pasos:

- Genera las variables *dummy* de la interacción 'Minority'*'Center'.
- Haz la ecuación de regresión de 'Salbegin' = f('Minority'*'Center'), usando correctamente las correspondientes variables *dummy*.
- Comenta la significación del estadístico F de conjunto. Interpreta los resultados de conjunto.
- Interpreta la significación de la interacción. ¿Qué significa la mejora de R^2 en la tabla de resultados?
- Escribe la ecuación de regresión de conjunto de 'Salbegin' = f('Minority'* 'Center').
- Escribe las ecuaciones para cada grupo. Comenta cada una de ellas.
- Haz una figura de conjunto que represente los resultados.

2. Se desea hacer la ecuación de regresión lineal de 'Salbegin' = f('Minority'*'Jobcat'). Responde:

- Haz la ecuación de regresión de 'Salbegin' = f('Minority'*'Jobcat'). (Para ello, comprueba si tienes las variables 'Minority' y 'Jobcat' como variables *dummy*, y si ya tienes en tu fichero de datos la interacción de ambas variables *dummy*).
- Comenta la significación del estadístico F de conjunto. Interpreta los resultados de conjunto.
- Interpreta la significación de la interacción de 'Minority'*'Jobcat'.
- Escribe la ecuación de regresión de conjunto de 'Salbegin' = f('Minority'*'Jobcat').
- Escribe las ecuaciones para cada grupo. Comenta cada una de ellas.
- Haz una figura de conjunto que represente los resultados.

Unidad 4. Regresión lineal con interacción de una variable independiente continua con otra variable independiente dicotómica

Introducción

En esta unidad abordamos el estudio de la regresión múltiple en el caso de tener una variable independiente de 'k=2' grupos (por ejemplo, la variable 'Género', con dos grupos, 'Masculino' y 'Femenino') y otra variable independiente continua. Desde esta perspectiva, una primera aproximación sería mediante la regresión con variables principales, en la que la variable independiente dicotómica se transforma en *dummy*, con lo cual se obtienen dos rectas paralelas, una para cada grupo.

Una segunda aproximación sería mediante el estudio de la interacción de ambas variables independientes, de este modo se comprueba, y se modeliza estadísticamente de manera más adecuada si las pendientes de los dos grupos son distintas entre sí. Así se consigue una ecuación de regresión de conjunto para los dos grupos, pero que se adapta mejor a cada uno de los grupos, comprobándose si las pendientes de ambos grupos son distintas (e indirectamente, si las intersecciones también difieren).

En la siguiente unidad se verá cómo generalizar la interacción de una variable independiente continua con otra variable independiente de tres o más grupos, pero antes se ha de desarrollar la teoría y la aplicación con una variable independiente de dos grupos.

Objetivos

Cuando el alumno finalice la unidad sabrá:

- Transformar las variables categóricas en variables *dummy*, para incluirlas en el modelo de regresión.
- Obtener las ecuaciones de regresión con grupos, mediante variables *dummy*, incluyendo solo las variables principales, lo que supone asumir igualdad de pendientes en los distintos grupos.
- Representar gráficamente las rectas de regresión, paralelas, en el caso de incluir solo variables principales mediante transformaciones *dummy*, para cada grupo.
- Realizar la interacción entre las variables independientes objeto de estudio.
- Representar gráficamente el diagrama de dispersión y las rectas de interacción entre una variable independiente continua y otra variable independiente categórica con dos niveles.
- Calcular las rectas de regresión con interacción entre una variable independiente continua y otra categórica con dos niveles.
- Obtener la ecuación de pronóstico, y su validación estadística, de una variable dependiente y dos variables independientes que interaccionen: una continua y otra categórica con dos niveles.
- Interpretación de la interacción de variables (intersección y pendiente) entre una variable independiente continua y otra categórica con dos niveles.

4.1. Regresión lineal de una variable dependiente sobre una variable independiente dicotómica y otra variable independiente continua: Análisis exploratorios y regresión solo con variables principales

Hasta aquí hemos trabajado con análisis de regresión simple y múltiple en el sentido clásico del modelo: mediante variables continuas 'principales', sin transformarlas, si bien hemos aprendido cómo transformar variables categóricas en variables dummies, y hemos hecho interacción entre variables categóricas. No obstante, es frecuente que las muestras de datos contengan grupos basados en variables como género, color de la piel, categoría laboral o lugar de nacimiento, además de variables independientes continuas. En este sentido, hemos de recordar que la regresión múltiple clásica con grupos asume, y da como resultado, que todos los grupos tienen la misma pendiente (b_1), pero diferente origen (b_0). Vamos a comprobar paso a paso que esto es exactamente así.

Para ello, teniendo en cuenta los datos de la matriz de datos que hemos guardado en la unidad anterior: Comp_dat_Dms_U3.sav, podemos obtener la regresión de la variable 'Salary' en función de las variables 'Educ', medida como cuantitativa, y 'Gender', medida nominalmente, y transformada a variable *Dummy*, que hemos renombrado como 'D_gndr_fem', donde el valor 0 = 'Male' y el valor 1 = 'Female'. En la unidad anterior se comprobaba que si la variable independiente de grupos se transformaba en *dummy*, efectivamente producía dos rectas paralelas. La ecuación funcional sería:

$$\text{Salary} = f(\text{'Educ'}, \text{'Gender'}) \tag{4. 1}$$

Al transformar 'Gender' en dicotómica, quedaría:

$$\text{Salary} = f(\text{'Educ'}, \text{'D\_gndr\_fem'}) \tag{4. 2}$$

Así pues, en formato algebraico de regresión lineal:

$$\text{Salary} = b0 + b1 \cdot \text{Educ} + [b2 \cdot \text{D\_gndr\_fem}] + e \tag{4. 3}$$

Para ello utilizamos la siguiente sintaxis:

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA
     /CRITERIA=PIN(.05) POUT(.10)
```

```
/NOORIGIN
/DEPENDENT salary
/METHOD=ENTER educ D_gndr_fem
/SAVE PRED RESID.
```

En la sintaxis se resaltan los párrafos: ENTER Y SAVE, escritos en negritas, que son los que permiten que, tras estimar la ecuación de regresión, el programa guarde los valores de pronóstico encontrados. En la Tabla 4.1 encontramos los resultados de la regresión planteada y en la Figura 4.1, vemos los valores predichos que genera la ecuación para el 'salary'.

Tabla 4.1.
Resultados de la regresión: 'Salary=f('Educ', 'D_gndr_fem')

(a)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,700[a]	,490	,488	$11,665.009

a. Predictors: (Constant), D_gndr_fem, Educational Level (years of study)

b. Dependent Variable: Current Salary

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	61579151443	2	30789575722	226,273	,000[b]
	Residual	64090113336	471	136072427,5		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), D_gndr_fem, Educational Level (years of study)

(c)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	18057,887	2961,406		6,098	,000
	Educational Level (years of study)	3233,037	198,956	,572	16,250	,000
	D_gndr_fem	-8110,173	1151,230	-,248	-7,045	,000

a. Dependent Variable: Current Salary

La Tabla 4.1(a) muestra el ajuste del modelo; los valores de R^2 indican que el modelo lineal explicaría, en conjunto, el 49 % de la varianza del 'Salary' en función de 'Educ ' y 'D_gnd_fem'; (b) también podemos comprobar que el p-value del estadístico F para el ajuste de conjunto del modelo es significativo; asimismo, (c) nos indica que todos los coeficientes implicados en el modelo son significativos; por tanto, dichos coeficientes obtenidos permiten plantear la siguiente ecuación de conjunto:

$$\text{Salary} = 18057{,}9 + 3233{,}0 \cdot \text{Educ} - [8110{,}2 \cdot \text{D\_gndr\_fem}] + e \quad \textbf{(4. 4)}$$

Pero una ecuación de conjunto con grupos puede desglosarse en sus correspondientes ecuaciones, una para cada grupo del modelo, al haber dos grupos (masculino y femenino), habrá dos ecuaciones; así pues, la ecuación para hombres, que corresponde al valor de la variable *Dummy* igual a cero ('D_gndr_fem' = 0), es:

$$Salary'_{Male} = 18057{,}9 + 3233{,}0 \cdot Educ - 8110{,}2 \cdot 0$$

$$\boldsymbol{Salary'_{Male} = 18057{,}9 + 3233{,}0 \cdot Educ} \quad \textbf{(4. 5)}$$

Para mujeres ('D_gndr_fem' = 1):

$$Salary'_{Female} = 18057{,}9 + 3233{,}0 \cdot Educ - 8110{,}2 \cdot 1$$

$$\boldsymbol{Salary'_{Female} = 9947{,}7 + 3233{,}0 \cdot Educ} \quad \textbf{(4. 6)}$$

Por tanto, queda claro que se mantiene para ambos grupos ('Male' y 'Female') el mismo valor en la pendiente, pero no el valor de la intersección.

La Figura 4.1 muestra los pronósticos obtenidos en el 'Data View' del SPSS y si los representamos gráficamente, la Figura 4.2 hace patente dicha conclusión: ambas rectas de regresión son paralelas. Y su aplicación permite obtener un sistema de pronóstico detallado por género.

Para obtener la representación gráfica de los valores pronosticados con las Ecuaciones 4.5 y 4.6, se corre la sintaxis:

```
GRAPH
      /SCATTERPLOT(BIVAR)=educ WITH Y_Pred BY gender
      /MISSING=LISTWISE.
```

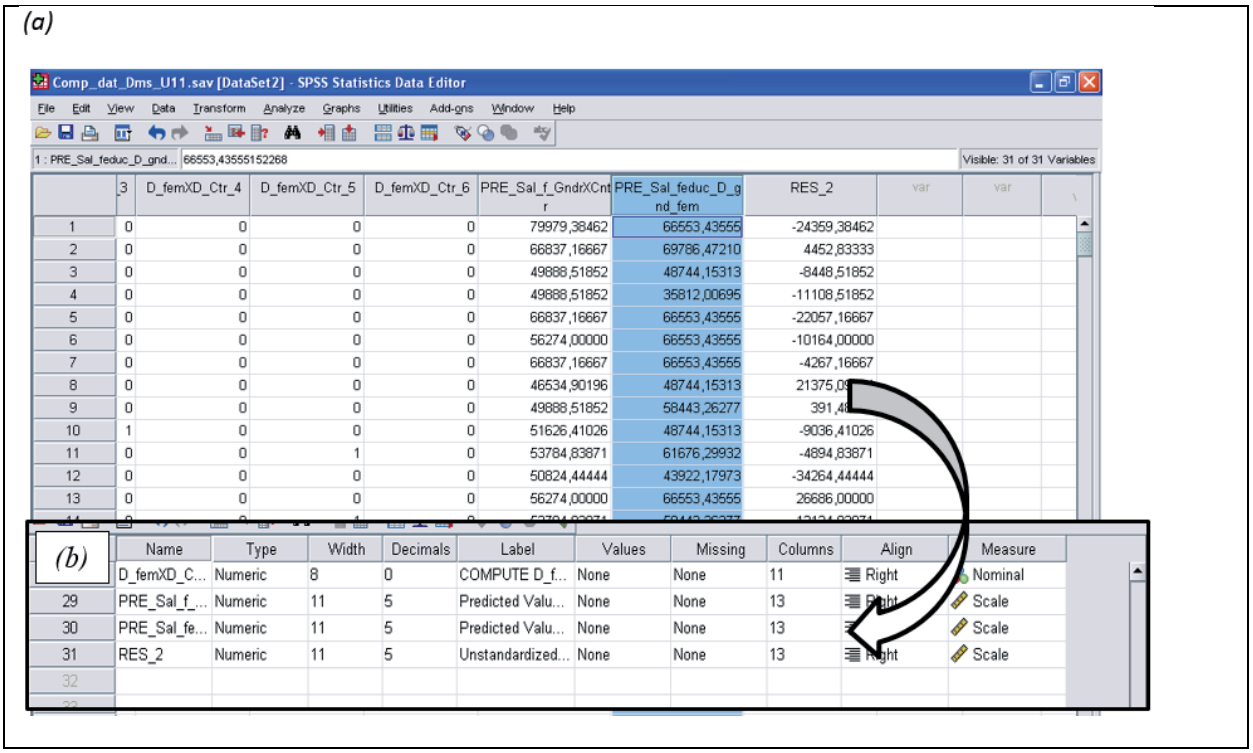


Figura 4.1. Valores predichos de 'Salary', marcada en azul, en función de 'Educ' y 'D_gndr_fem' (a) en modo 'Data View' y (b) en modo 'Variable View´, como aparecen en SPSS

La Figura 4.2 permite comprobar cómo ambas pendientes son iguales, en tanto que los puntos de corte con la ordenada en el origen coinciden con los encontrados en las rectas de las Ecuaciones 4.5 y 4.6, cuando se considera uno u otro género.

Visto lo anterior, hemos comprobado cómo el hecho de considerar los dos grupos de la variable 'Gender', nos realiza pronósticos lineales paralelos para cada grupo por separado. Lo que genera distintos valores de intercepto pero pendientes iguales. No obstante, es lógico pensar que no siempre el crecimiento en la variable dependiente haya de ser constante para cada unidad de la variable independiente en ambos grupos de la segunda variable independiente. Esta situación nos lleva a considerar la posibilidad de que exista alguna interacción entre las variables independientes que pudiera explicar de modo más preciso el comportamiento de la variable dependiente.

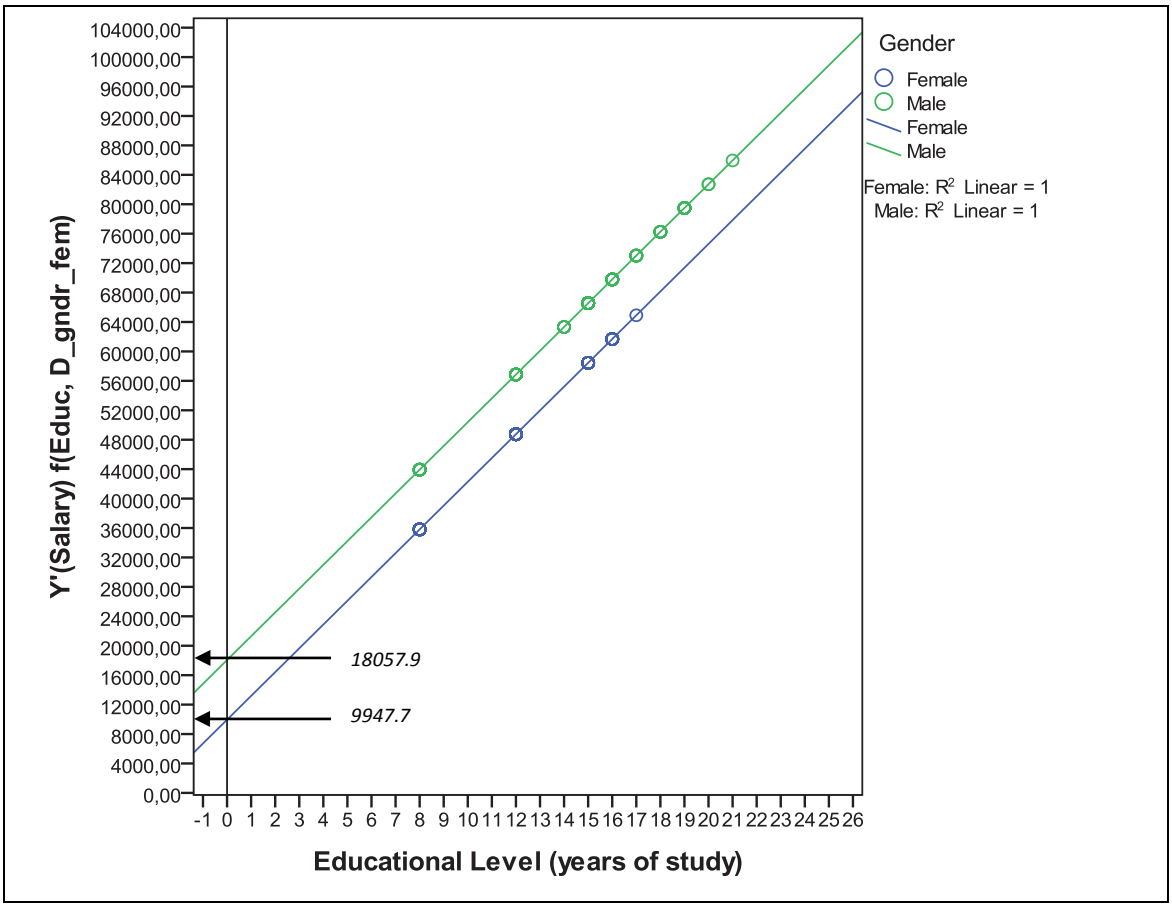


Figura 4.2. Rectas de regresión con valores pronosticados para dos grupos, sin interacción, conforme a la Ecuación 4.4

4.2. Interacción: Regresión lineal de una variable dependiente sobre una variable independiente continua en interacción con otra variable independiente de dos categorías

Las rectas incluidas en la Figura 4.2 indican que la ganancia obtenida en 'Salary' por cada año de estudio es la misma, independientemente del género; es decir, que cada año de nivel educativo permitiría incrementar en valor pronosticado el salario igualmente si eres hombre o mujer; o sea, por cada año de estudio se gana 3233 dólares más, para todas las personas de la muestra, en cambio, lo que varía en el grupo de 'Male' y de 'Female' es su punto de partida, su intersección, siendo mayor en el de 'Male'. Vamos a comprobar si dicha aseveración se mantiene cuando incluimos en la ecuación la variable de interacción entre 'Educ' y 'D_gndr_fem'. Centrémonos ahora en qué pasaría si en lugar de incluir solo las variables independientes descritas en la Ecuación 4.3, considerásemos, además, la interacción entre 'Education' y 'Gender'. Para ello, previamente, hemos de generar la nueva variable de interacción multiplicando ambas en la matriz de datos del SPSS.

Lo que en términos de sintaxis se obtendría con:

```
COMPUTE EducXD_gndr_fem=educ * D_gndr_fem.
     VARIABLE LABELS EducXD_gndr_fem
     'COMPUTE EducXD_gndr_fem=educ * D_gndr_fem'.
     EXECUTE.
```

Donde la variable ('D_ gndr_fem') es la variable *Dummy* generada desde 'Gender' y está compuesta por los valores 0 ('Male') y 1 ('Female'). De manera que en la variable de interacción 'Educ'X'Gen_D_fem', aparecerá cero para las personas de la muestra 'Male' ('Educ'*0= 0), y serán los mismos valores de 'Educ' cuando estemos en un dato de la muestra 'Female' ('Educ'*1 = 'Educ'). De este modo habremos obtenido una nueva variable que es la interacción de las dos independientes y que se incluye ahora en el modelo de regresión, como nueva variable independiente, además de 'Gender' y 'Educ'. Así, la Ecuación 4.7, estará formada por tres variables independientes de las que una se obtiene por multiplicación de las otras dos.

$$\text{Salary} = b_0 + b_1 \cdot \text{Educ} + [b_2 \cdot \text{D\_gndr\_fem}] + [b_3 \cdot \text{EducXGen\_D\_fem\_1}] + e \quad (4.7)$$

La sintaxis de SPSS que permite obtener la nueva ecuación quedaría como sigue:

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA
      /CRITERIA=PIN(.05) POUT(.10)
      /NOORIGIN
      /DEPENDENT salary
      /METHOD=ENTER educ D_gndr_fem EducXD_gndr_fem
      /SAVE PRED RES.
```

Si hacemos pasar la sintaxis por el programa, podemos verificar paso a paso la significación y la estimación de los parámetros del modelo de regresión propuesto. La Tabla 4.2, como resalta el recuadro en rojo en el apartado (b) del ANOVA, permite considerar que el modelo que incluye la interacción es significativo (el conjunto de variables independientes incluidas en el modelo ajustan adecuadamente, explicando correctamente la variable dependiente). Por tanto, procedemos a conocer el coeficiente de determinación (R^2), los correspondientes valores estimados y su significación, para los coeficientes que conforman la Ecuación 4.7.

Tabla 4.2.
Resultados de la regresión: 'Salary=f('Educ', 'D_gndr_fem', 'EducX D_gndr_fem'). (a) Ajuste de conjunto, (b) significación global del modelo de regresión, y (c) significación de los coeficientes

(a)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,726[a]	,527	,524	$11,240.289

a. Predictors: (Constant), COMPUTE EducXD_gndr_fem=educ * D_gndr_fem, Educational Level (years of study), D_gndr_fem

b. Dependent Variable: Current Salary

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	66287537638	3	22095845879	174,886	,000[b]
	Residual	59381727141	470	126344100,3		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), COMPUTE EducXD_gndr_fem=educ * D_gndr_fem, Educational Level (years of study), D_gndr_fem

Tabla 4.2. *(Continuación)*
(c)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6034,042	3467,330		1,740	,082
	Educational Level (years of study)	4066,276	235,338	,720	17,278	,000
	D_gndr_fem	24247,894	5415,423	,742	4,478	,000
	EducXD_gndr_fem	-2477,024	405,762	-,967	-6,105	,000

a. Dependent Variable: Current Salary

La Tabla 4.2 (a) muestra que el modelo que incluye la interacción explicaría el 52,7 % de la varianza del 'Salary' (R^2 = ,527), recordemos en el modelo de la Tabla 4.1 la capacidad explicativa era del 49 %, esta diferencia puede parecer pequeña, pero es estadísticamente significativa. En la Tabla 4.2 (b) se comprueba que el p-value del contraste de la *F* para los parámetros del modelo también es significativo, indicándonos la significación conjunta de los coeficientes que hemos incluido en la ecuación, y (c) permite comprobar que el coeficiente de la interacción propuesta es significativa.

IMPORTANTE
Guarda el fichero de datos con las interacciones como: 'Comp_dat_Dms_U4.sav', de este modo ya tenemos guardadas las variables *dummy* con las interacciones de 'Gender'*'Educ'. Guarda los valores predichos como 'Pre_Sal_f_EducXD_gnd_fem'.
Si la interacción entre dos variables es significativa, los términos de las variables principales han de dejarse en la ecuación aunque resulten ser no significativos.

Una vez conocidos los valores de los coeficientes ya podemos completar la ecuación de regresión como aparece en la Ecuación 4.8:

$$\text{Salary} = 6034{,}0 + 4066{,}3 \cdot \text{Educ} + [24247{,}9 \cdot \text{D\_gndr\_fem}] - [2477{,}0 \cdot \text{EducXD\_gndr\_fem}] + e \quad (4.8)$$

Igualmente, podemos guardar los pronósticos que genera la Ecuación 4.8, como muestra la Figura 4.3. Lo que, a su vez, nos permite plantear la ecuación de pronóstico para cada grupo, pues la Ecuación 4.8 contiene, implícitamente, dos ecuaciones, según se haga el pronóstico para hombres o para mujeres.

Para hombres (Valor de 'D_gndr_fem=0):

$$Salary'_{Male} = 6034{,}0 + 4066{,}3 \cdot Educ + 24247{,}9\,(0) - 2477{,}0\,(0)$$
$$= \mathbf{6034{,}0 + 4066{,}3 \cdot Educ} \qquad \textbf{(4. 9)}$$

Es decir, a los hombres, grupo de referencia (*dummy* = 0), se les puede predecir el 'Salary' mediante la constante y la pendiente de la variable predictora continua, 'Educ'; dado que, en lo que se refiere la variable 'Gender' su valor es '0'.

En cuanto al grupo de mujeres, el pronóstico en 'Salary' vendría determinado por la Ecuación 4.8, teniendo en cuenta que 'D_gndr_fem=1', lo que daría lugar a la ecuación:

$$Salary'_{Female} = 6034{,}0 + 4066{,}3 \cdot Educ + 24247{,}9\,(1) - 2477{,}0\,(Educ \cdot 1)$$
$$= (6034{,}0 + 24247{,}9) + (4066{,}3 \cdot Educ - 2477{,}0\,Educ)$$
$$= \mathbf{30281{,}9 + 1589{,}3 \cdot Educ} \qquad \textbf{(4. 10)}$$

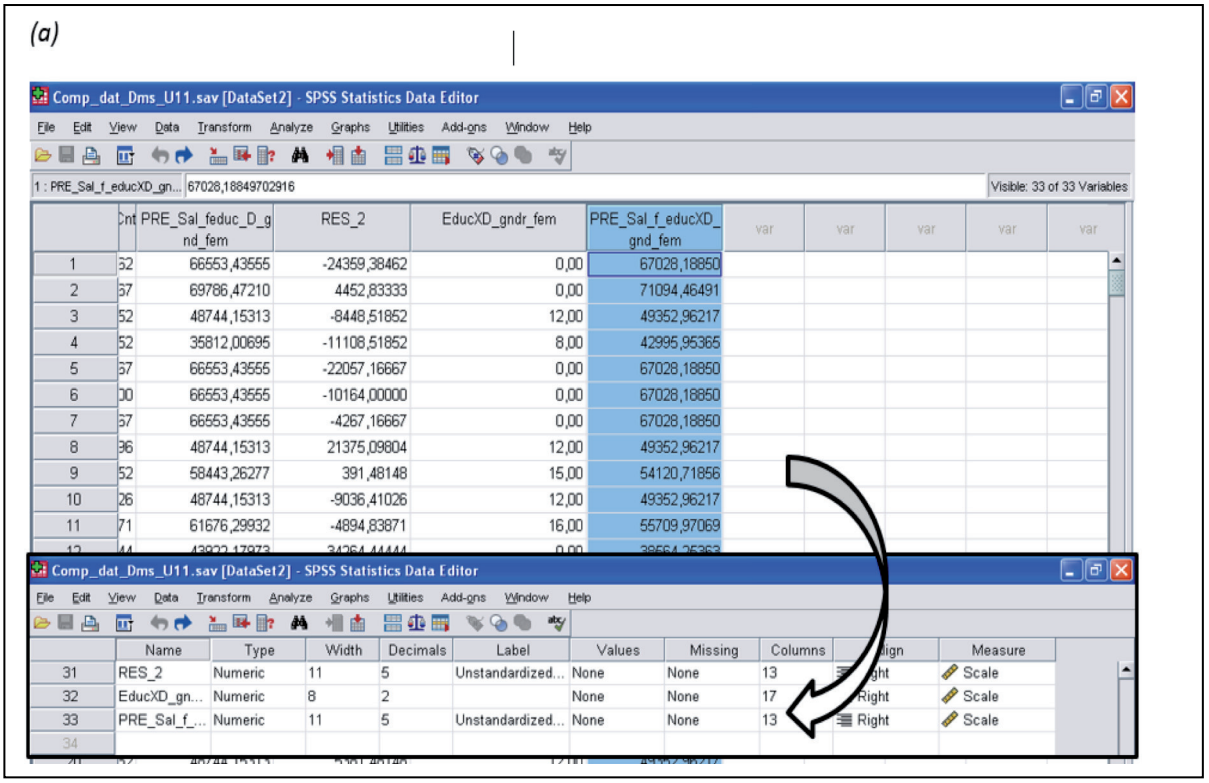


Figura 4.3. Variable predicha de 'Salary', en función de 'Educ', 'D_gndr_fem' y 'Educ* D_gndr_fem' en modo 'Data view' y (b) en modo 'Variable view´, como aparecen en el programa SPSS

Este procedimiento de regresión permite que obtengamos una ecuación para cada grupo, en nuestro caso una para hombres y otra para mujeres. Ha quedado de manifiesto que cuando el grupo de la variable a pronosticar es el de referencia (*dummy* = 0), es más fácil de obtener, dado que mantiene la constante y en la pendiente solo queda implicada la variable continua considerada en la predicción. No obstante, en el grupo *dummy* = 1, el coeficiente de la *dummy* se añade a la constante, y el término de interacción se añade al coeficiente de la variable continua 'Educ'; es decir, se obtienen dos ecuaciones con ordenadas en el origen y pendientes distintas.

Con los datos predichos de este modo que, como muestra la Figura 4.3, quedan anexados a la matriz de datos añadiéndolos al final de las variables que ya teníamos, podemos obtener el diagrama de dispersión de Y': 'PRE_Sal_f_educXD_gnd_fem' con X: 'Educ', estableciendo marcas por 'Gender', con 'línea de ajuste total' y con 'línea de ajuste por grupos', para la nueva estrategia de regresión. En este caso, la sintaxis del diagrama de dispersión será:

```
GRAPH
      /SCATTERPLOT(BIVAR)=educ WITH
            PRE_Sal_f_educXD_gnd_fem BY gender
      /MISSING=LISTWISE.
```

La Figura 4.4 muestra el diagrama de dispersión, en el que se aprecia el comportamiento de la variable dependiente 'PRE_Sal_f_educXD_gnd_fem BY 'Gender' en función de los años de 'Educ' para los grupos generados por la variable 'Gender'. Por tanto si, tal y como hemos indicado, consideramos los valores predichos, observamos que su comportamiento, cuando se considera la interacción de variables, ya no sigue el trazado de rectas paralelas; por el contrario, cada grupo establecido por 'Gender' muestra su comportamiento particular tanto en intersección como en pendiente. Parece clara la precisión estadística y substantiva que ofrece un modelo que contempla una interacción significativa.

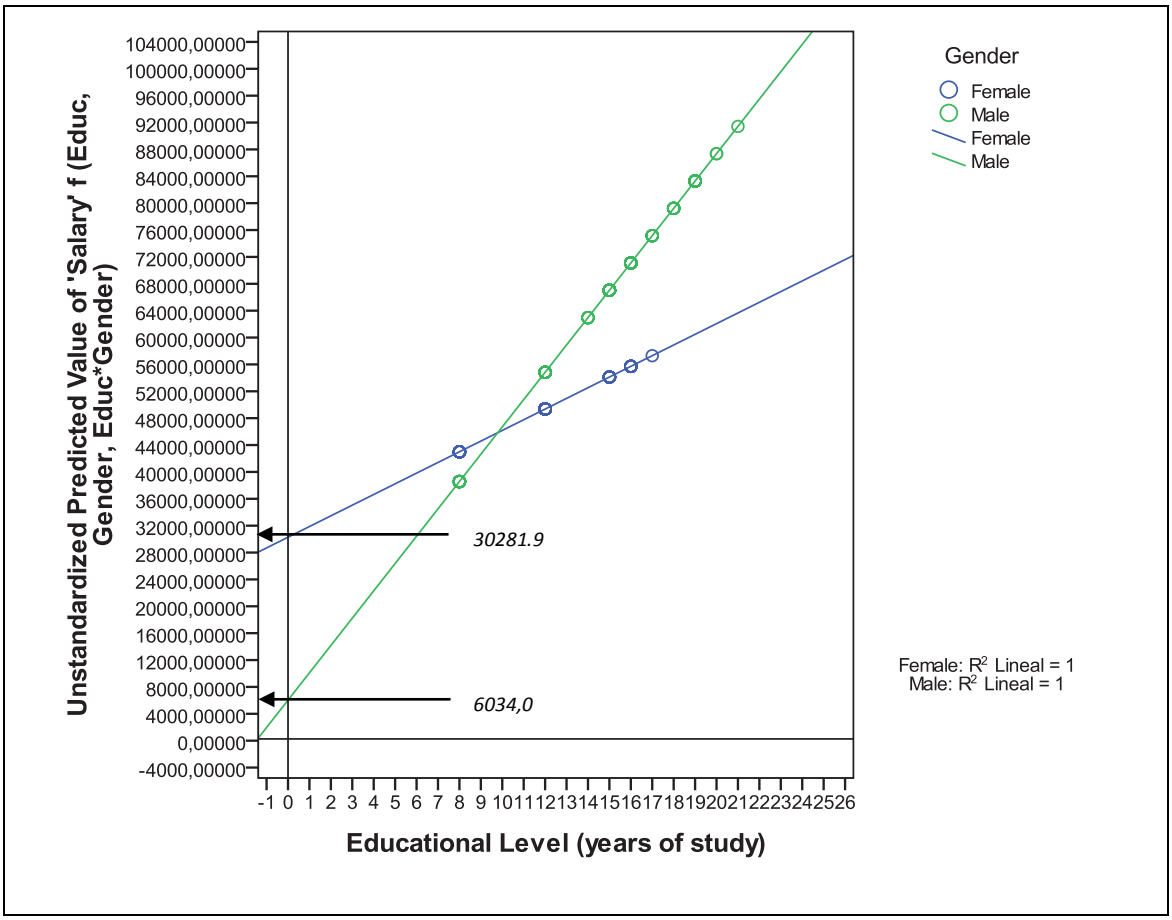


Figura 4.4. Diagrama de dispersión, con rectas de regresión, de 'PRE_Sal_f_educXD_gnd_fem' por niveles educativos para hombres y mujeres, indicando las respectivas intersecciones con el eje de la variable dependiente para el grupo de hombres y el de mujeres

4.3. ¿Por qué usar la interacción?

La Figura 4.4 deja claro que el hecho de incluir en el modelo el término de interacción permite una aproximación detallada y más ajustada al comportamiento de cada grupo. En este punto nos ocuparemos de profundizar en la interacción, tanto en las ventajas que produce, como en el procedimiento para obtener el modelo de regresión, así como su interpretación desde SPSS. En lo que a la interpretación de la interacción se refiere es sencilla, ya se ha dicho que cuando es significativa: Las pendientes de las rectas de regresión son diferentes entre sí, variando para cada nivel.

Por tanto, si la interacción, cualquiera que sea, es significativa, estaría indicando que el proceso no es 'aditivo' entre las variables; es decir, gráficamente, no serían rectas paralelas o casi-paralelas, sino que lo que se está produciendo es un proceso 'multiplicativo'. Y en este caso la representación gráfica no produciría rectas paralelas.

Como se aprecia en la Figura 4.4, los hombres y las mujeres parecen no seguir rectas paralelas cuando se considera la interacción ('Educ'*'D_gndr_fem'), es evidente que varían tanto en intersección como en pendiente.

Por otro lado, si comparásemos las conclusiones que se derivan de las Ecuaciones 4.5 y 4.6 (solo con variables principales), y las de 4.9 y 4.10 (con interacción de variables), llegaríamos a pronósticos bien dispares cuando comparamos hombres y mujeres. Para hacer un análisis más pormenorizado, consideremos, por ejemplo, el sujeto número 1 de la matriz de datos, es una mujer que tiene 12 años de nivel educativo y para la ecuación con hombres, podemos considerar el sujeto número 217 que tiene 15 años de nivel educativo. Si hacemos sus pronósticos desde las ecuaciones obtenidos encontramos:

Para el grupo de hombres (D_gndr_fem = 0), sin y con interacción:

Desde la Ecuación 4.5:

$$Salary' = 18057{,}9 + 3233{,}0 \cdot (15) = 66552{,}9 \quad (4.\ 11)$$

A partir de la Ecuación 4.9:

$$Salary' = 6034{,}0 + 4066{,}3 \cdot (15) = 67028{,}5 \quad (4.\ 12)$$

Es decir, en el primer caso, si no se tiene en cuenta la interacción, cuando consideramos solo el grupo de hombres, lo que inicialmente cobraría un hombre sin ningún año de

estudios (teniendo en cuenta solo las intersecciones) es casi el triple de lo que sucede cuando se tienen en cuenta la interacción. En cuanto a la posibilidad de incrementar el 'Salary' en los hombres (teniendo en cuenta las respectivas pendientes), por cada año de educación es claro que el salario aumenta en mayor medida que se aumenta en nivel educativo si consideramos el modelo con interacción. Concretamente, el incremento en 'Salary' por año pasa a ser de 3223,0 a 4066,3 cuando se tiene en cuenta la interacción.

Para mujeres (D_gndr_fem = 1), sin y con interacción:

Según la Ecuación 4.6:

$$Salary' = 9947{,}7 + 3233{,}0 \cdot (12) = 48743{,}7 \quad \textbf{(4. 13)}$$

Si se pronostica desde la Ecuación 4.10:

$$Salary' = 30281{,}9 + 1589{,}3 \cdot (12) = 49353{,}5 \quad \textbf{(4. 14)}$$

En el grupo de mujeres sucede al contrario que en el grupo anterior; es decir, si no se considera la interacción, para las mujeres con cero años de nivel educativo, su 'Salary' actual sería de 20334,2 dólares menos (30281,9 - 9947,7), o aproximadamente una tercera parte, que si consideramos la interacción; sin embargo, la posibilidad de incrementar su salario a medida que incrementan los años de educación es casi la mitad si se considera la interacción que si no se considera ésta (en la Ecuación 4.6 es de 3233,0 y pasa a ser en la Ecuación 4.11 de 1589,3 por año de estudios).

Comparando ahora hombres y mujeres en las Ecuaciones 4.5 y 4.6, podemos observar que aunque el punto de partida es diferente cuando unos y otras no tiene ningún año de estudios, sin embargo, la posibilidad de incrementar su salario por cada año de incremento en educación es exactamente el mismo en ambos grupos, cuando, realmente, si consideramos las Ecuaciones 4.9 y 4.10, comprobamos que las mujeres empiezan cobrando casi cinco veces más que ellos y sus expectativas de incremento de salario por año de educación es, aproximadamente, 2,56 veces menor para ellas que para ellos.

El valor de la variable 'Educ' en el que los hombres y las mujeres cobran lo mismo, será cuando se igualen los valores pronosticados en las Ecuaciones 4.9 y 4.10, es decir:

$$Salary'_{Male} = Salary'_{Female} \quad \textbf{(4. 15)}$$

o lo que es lo mismo:

$$6034{,}4 + 4066{,}3 \cdot \text{'Educ'} = 30281{,}9 + 1589{,}3 \cdot \text{'Educ'} \quad \textbf{(4. 16)}$$

de donde despejando ‘Educ’, tenemos: ’Educ’ = 9,79; por tanto, cuando los hombres o las mujeres tienen 9,79 años de educación, ambos cobrarían exactamente el mismo ‘Salary’ (en ambos caso se cobra 45842 dólares). Este valor puede comprobarse en la Figura 4.5 (ver flechas de líneas discontinuas), ése es el punto en el ambos grupos se cruzan. Como conclusión, podría decirse que a niveles más bajos de 9,79 años de estudios las mujeres cobran más en esta empresa que los hombres; sin embargo, a nivel educativo por encima de 9,79 años, ellas cobran menos salario, a pesar de que incrementen en la misma proporción que ellos su nivel educativo.

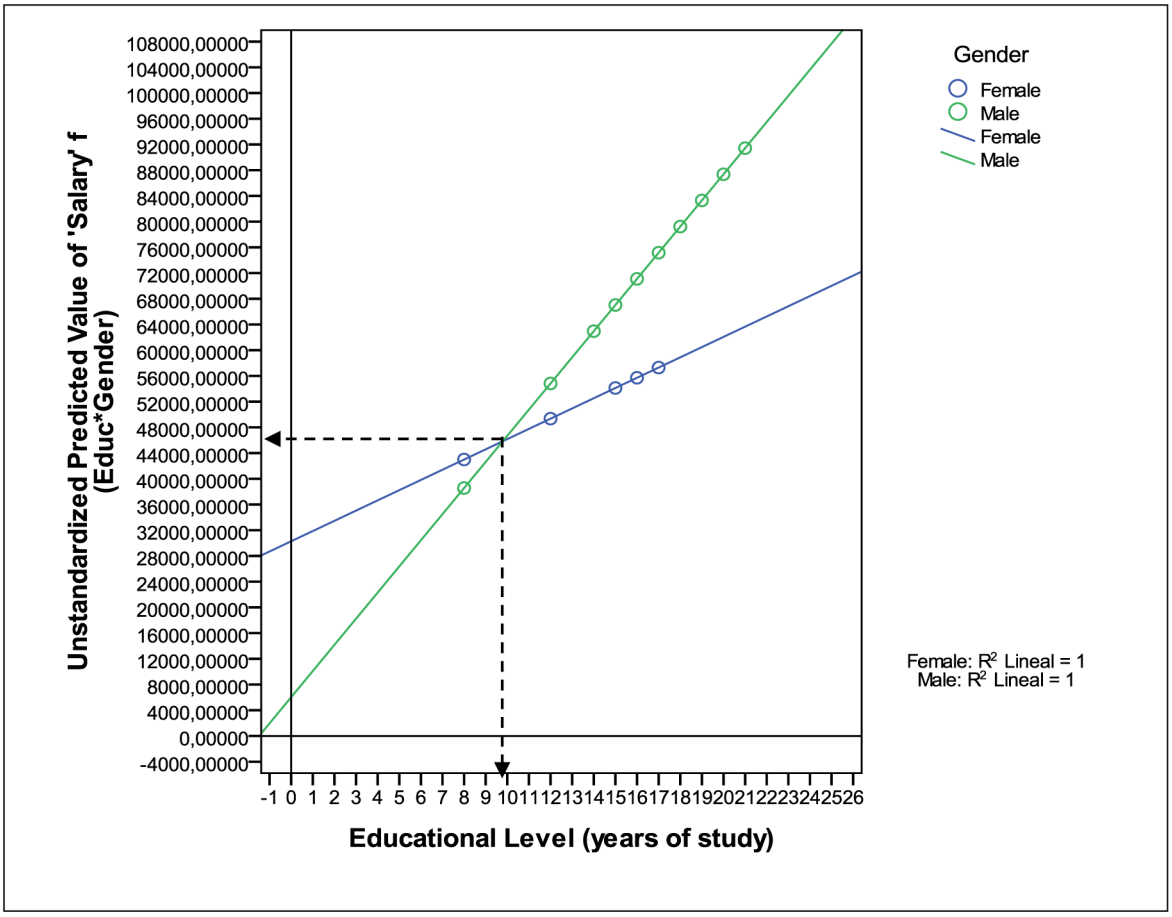


Figura 4.5. Diagrama de dispersión, con rectas de regresión, de ‘PRE_Sal_f_educXD_gnd_fem’ por niveles educativos para hombres y mujeres, mostrando el punto donde hombres y mujeres cobran el mismo salario

Además, en nuestro caso, al ser significativa la regresión con interacción de las variables, la capacidad explicativa es mejor que cuando solo se usa la regresión con variables principales; los investigadores que trabajan con grupos deberían poner a prueba la hipótesis de la interacción de variables, pues es un procedimiento lógico pensar que existe un comportamiento diferencial para los diferentes grupos de la muestra estudiada.

4.4. Adenda: El análisis de la covarianza (ANCOVA) y la regresión con interacción de una variable cuantitativa y otra dicotómica

Si consideramos la posibilidad de utilizar un análisis de covarianza para expresar el modelo que hemos venido desarrollando en esta unidad, podríamos pensar en explicar la variable 'Salary' (cuantitativa) desde la covariable (o variable continua) 'Educ' y un factor o variable de grupos que vendría representado por la variable dicotómica 'Gender'. Podemos en este caso, pedir al programa que incluya el término de interacción entre las variables 'EducX Gender'. De manera que la sintaxis a correr en el programa quedaría detallada como sigue:

```
UNIANOVA salary BY Gender WITH educ
      /METHOD=SSTYPE(3)
      /INTERCEPT=INCLUDE
      /SAVE=PRED
      /CRITERIA=ALPHA(0.05)
      /DESIGN=educ Gender Gender*educ.
```

Resaltamos en negrita el diseño para que se compruebe el modelo que se contrasta y también se pide que guarde los pronósticos, con objeto de que cada lector pueda comprobar cómo los valores obtenidos por uno u otro método son iguales.

Al correr la anterior sintaxis en SPSS, nos lleva a los resultados que se muestran en la Tabla 4.3, en la que se observa que la significación de conjunto F(3, 474) = 174,886, p<.001 es idéntica a la que obteníamos en el modelo de regresión que incluye la interacción que se detalla en la Tabla 4.2, apartado (b).

Tabla 4.3.
*Resultados del ANCOVA de 'Salary'= f('Educ, Gender, Educ*Gender')*

Tests of Between-Subjects Effects

Dependent Variable: Current Salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6,629E+10[a]	3	22095845879	174,886	,000
Intercept	5681796811	1	5681796811	44,971	,000
educ	24544712371	1	24544712371	194,269	,000
gender	2533017826	1	2533017826	20,049	,000
gender * educ	4708386195	1	4708386195	37,266	,000
Error	59381727141	470	126344100,3		
Total	1,719E+12	474			
Corrected Total	1,257E+11	473			

a. R Squared = ,527 (Adjusted R Squared = ,524)

Si nos fijamos en la significación de la interacción 'Gender*educ', *F(1, 470) = 37,266, p<.001*, que coincide con el valor del cuadrado de la 't', recuérdese que $t^2 = F$, cuando hay un grado de libertad en el numerador del estadístico *F*, que encontramos para el valor de la misma interacción en la Tabla 4.2 apartado (c).

Otro dato interesante es resaltar los valores de la R^2 (.527) y de la *F(3,474) = 174,886, p<.001*, que, obviamente, son los mismos en ambas Tablas 4.2 y 4.3 (de regresión con interacción y del ANCOVA).

Como hemos visto, la regresión con interacción entre una variable continua y una dicotómica y el análisis mediante ANCOVA, ambos dan la misma información tanto para la *F* de conjunto como para la correspondiente al término de interacción. Sin embargo, el control que ejerce el analista desde el modelo de regresión es mucho mayor, dado que conoce los coeficientes de cada término en la regresión y puede elaborar todas las ecuaciones que explican la variable dependiente para cada grupo. El análisis ANCOVA lleva a cabo automáticamente sus propios términos internos de codificación de grupos y es más difícil el acceso directo a la confección de las ecuaciones que generan los valores predichos.

4.5. Conclusiones

- Para hacer regresión por grupos necesitamos variables *dummy*.
- Pero la regresión mediante variables principales con variables *dummy* da la misma pendiente para todos los grupos, lo que significa que la representación gráfica da rectas paralelas (una para cada grupo),
- Puede ser que los grupos no sean paralelos (diferentes pendientes): hay que poner siempre a prueba la interacción de variables con grupos, planteando las correspondientes hipótesis de nulidad sobre la igualdad de pendientes entre los grupos.
- Convendría que en cualquier investigación se plantease la hipótesis de la interacción, pues proporciona una información que no da ningún otro procedimiento estadístico.
- El investigador ha de saber plantear la ecuación de conjunto del modelo y la ecuación particular para cada grupo de la ecuación general.
- Ha de saber interpretar adecuadamente cada ecuación.
- También ha de poder llevar a cabo la correspondiente representación gráfica de conjunto.

Lecturas recomendadas

Hay varios libros que tratan sobre la interacción de variables desde una vertiente aplicada a la investigación, aunque no exponen cómo desarrollar gráficos a partir de los valores esperados (o pronosticados), estos son el de Hardy (1993), el de Aiken y West (1991) y el de Jaccard y Turrisi (2003); y un interesante libro que tiene un capítulo (el número 8) sobre la interacción de las variables titulado «Modelos de regresión para la predicción cuantitativa y cualitativa» es el de Kutner, Nachtsheim y Neter (2004).

Bibliografía

Aiken, L. S.; West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.

Jaccard, J.; Turrisi, R. (2003). *Interaction effects in multiple regression*. (2ª ed.). Thousand Oaks, CA: Sage.

Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). *Applied linear regression models*. Nueva York: McGraw-Hill.

Actividades

Se quiere hacer la regresión lineal con interacción: Salini = f(‘edu’*’minority’).

1. Comprueba si ‘minority’ es *dummy*, si no lo es, hazla *dummy*.

2. Haz el producto (interacción) de ‘edu’*’minority’, llama a esta nueva variable ‘eduxmin’.

3. Haz la ecuación de regresión: Salini = f(‘edu’, ’minority’, ‘eduXmin’).

4. Interpreta la significación de los estadísticos de conjunto (sobre todo para el conjunto de la ecuación y para la interacción).

5. Escribe la ecuación de regresión de conjunto.

6. Interpreta los coeficientes obtenidos.

7. Escribe la ecuación de regresión para cada grupo.

8. Haz una figura con los pronósticos obtenidos en la regresión.

9. Interpreta los resultados obtenidos.

Unidad 5. Regresión lineal con interacción de una variable independiente continua y otra de tres grupos

Introducción

Como venimos observando en anteriores unidades, cualquier investigador puede tener como objetivo la predicción de una variable dependiente a partir de dos variables independientes en las que, al menos una de ellas, de métrica cuantitativa, se comporta de modo diferente cuando se la considera desde los diferentes grupos que marca la otra independiente; por ejemplo, si la variable categórica tiene tres niveles (o grupos), en los que dentro de cada grupo cambia la pendiente (o la intersección, o ambas) de la relación entre la variable independiente continua y la variable dependiente. En estos casos, puede ser muy adecuado incluir un término de interacción entre ambas independientes; para ello, inicialmente realizamos el producto entre las dos variables independientes que interactúan, pero este producto se ha de realizar una vez que se ha transformado la variable categórica en variables dummies. En esta unidad nos ocuparemos de cómo realizar e interpretar dicha interacción cuando una de las variables independientes implicadas en la interacción categoriza la muestra en tres grupos.

Objetivos

Cuando el alumno finalice el tema sabrá:

- Representar gráficamente el diagrama de dispersión entre una variable independiente continua y otra categórica con tres grupos, recodificada esta variable independiente en otras dos variables equivalentes de tipo *dummy*.
- Calcular las rectas de regresión con interacción entre una variable independiente continua y otra categórica con tres grupos, para pronosticar los valores de la variable dependiente.
- Obtener la ecuación de pronóstico, y su validación, de una variable dependiente y dos independientes: una continua y otra categórica con tres grupos.
- A partir de la ecuación general de regresión, calcular la ecuación de pronóstico para cada uno de los tres grupos.
- Representar gráficamente el diagrama de dispersión y las rectas de interacción entre una variable independiente continua y otra variable independiente categórica con tres niveles.
- Extraer la adecuada interpretación de la interacción de variables (intersección y pendiente) entre una variable independiente continua y otra categórica con tres grupos.

5.1. Regresión de una variable independiente continua y otra categórica con tres grupos (solo efectos principales)

En el tema anterior tratamos la conveniencia de incluir en el modelo de regresión múltiple el término de interacción como variable explicativa, desde la consideración de que una variable dicotómica establecía dos grupos cuyos comportamientos en cuanto a que sobre la variable explicada configuraban dos sistemas diferentes (uno para cada grupo), en ordenada en el origen y/o en pendiente.

Supongamos ahora que pretendemos ajustar una ecuación de regresión múltiple en la que una de las variables independientes, la categórica, está conformada por tres grupos; por ejemplo, podemos considerar la variable: 'Employment Category', que encontramos en la matriz de datos 'Comp_dat_Dms_U4.sav' como 'Jobcat'. En este caso, el planteamiento es que pretendemos explicar la variable 'Salary' a partir de las variables independientes 'Educ', y 'Jobcat'. Pero sería un error considerar directamente la variable Jobcat, tal como está codificada en la matriz de datos ('Clerical' = 1, 'Custodial' = 2, 'Manager' = 3), porque en ese caso lo que estaríamos suponiendo es que el nivel 'Custodial' vale el doble que 'Clerical' o que 'Manager' vale el triple que 'Clerical', lo que no es correcto en ningún sentido (este error sería el mismo que si en lugar de categorías laborales tuviésemos la variable 'lugar de nacimiento', con tres países del mismo entorno, codificados con los valores 1, 2 y 3). Por tanto, el primer paso que hemos de dar es el de considerar los tres grupos en el mismo plano de igualdad teórica entre los tres grupos y, para ello, generaremos las variables *dummy*, con el fin de incluirlas en la ecuación que inicialmente, con fines didácticos, no va a considerar la interacción como estrategia analítica. Así, vamos a plantear, desde una perspectiva funcional, que:

$$\text{'Salary'} = f(\text{'Educ' y 'Jobcat'}) \qquad (5.1)$$

Tomando la matriz de datos 'Comp_dat_Dms_U4.sav', desde la que venimos implementando las variables de nueva confección y/o recodificación, encontramos que se generaron las variables *dummy*: 'D_JC_custodial' y 'D_JC_manager', por lo que en nuestros análisis la categoría de referencia es 'Clerical'. Simplemente, a título de recuerdo se remite al lector al apartado 4 de la Unidad 1.

Lo que vamos a comprobar, en este punto, es lo que queda plasmado en la Ecuación 5.2:

$$Salary = b_0 + b_1 \cdot Educ + [b_2 \cdot \text{D\_JC\_custodial'} + b_3 \cdot \text{D\_JC\_manager}] + e \quad (5.2)$$

Donde 'Educ' es una variable cuantitativa que recoge el número de años de estudio que cada empleado tiene en su formación, y la variable 'Jobcat' está recodificada como variables *dummy*, que permite organizar la muestra en tres categorías laborales de los empleados, consideradas en el mismo plano, concretamente son: 'Clerical', 'Custodial' y 'Manager'. Obsérvese en la Ecuación 5.2 que el grupo de referencia de la variable 'Jobcat' es el de 'Clerical'.

Antes de realizar el ajuste de la ecuación, hemos de recordar que ya contamos, desde una unidad anterior (Unidad 3), con la variable categorizada en sus variables *dummy*. En este punto, desde una perspectiva exploratoria, vamos a observar las relaciones que se producen entre las variables dependiente e independientes; para ello, realizaremos su diagrama de dispersión y comprobaremos, gráfica y conceptualmente, qué ocurre al plantear la Ecuación 5.2 (es decir, incluyendo solo las variables principales, 'Educ' y 'Jobcat' transformada en dummies, antes de incluir término de interacción). Así pues, realizamos la ecuación de regresión haciendo correr la sintaxis:

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA CHANGE
      /CRITERIA=PIN(.05) POUT(.10)
      /NOORIGIN
      /DEPENDENT salary
      /METHOD=ENTER educ
      /METHOD=ENTER D_JC_custodial D_JC_manager
      /SAVE PRED.
```

Nótese que las variables *dummy* integrantes de 'Jobcat' ('D_JC_custodial' y 'D_JC_manager') se han introducido como un bloque (mediante la sintaxis: /METHOD=ENTER D_JC_custodial D_JC_manager), por lo que también se pide al SPSS que proporcione el cambio de R^2 para cada bloque (mediante la sintaxis: CHANGE).

En la Tabla 5.1 se muestran los resultados obtenidos, en ella podemos comprobar que el ajuste global del modelo aditivo, solo con las variables principales, es estadísticamente adecuado para explicar la variable 'Salary', en función de las variables independientes:

'Educ' y '[D_JC_custodial, D_JC_manager]' (nótese que estos dos grupos representan a 'Jobcat').

Tabla 5.1.
Significación del modelo aditivo:
Salary = f('Educ', ['D_JC_custodial', 'D_JC_manager'])

(a)

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,661[a]	,436	,435	$12,251.190	,436	365,284	1	472	,000
2	,680[b]	,462	,459	$11,993.328	,026	11,257	2	470	,000

a. Predictors: (Constant), Educational Level (years of study)
b. Predictors: (Constant), Educational Level (years of study), Custodial = 1, otherwise = 0, Manager = 1, Other = 0
c. Dependent Variable: Current Salary

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
2		58064503530	3	19354834510	134,558	,000[c]
		67604761249	470	143839917,6		
		1,257E+11	473			

a. Dependent Variable: Current Salary
c. Predictors: (Constant), Educational Level (years of study), Custodial = 1, otherwise = 0, Manager = 1, Other = 0

(c)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
2		12452,644	3269,728		3,808	,000
		3254,978	249,349	,576	13,054	,000
	Custodial = 1, otherwise = 0	5224,807	2484,166	,074	2,103	,036
	Manager = 1, Other = 0	7425,064	1817,315	,174	4,086	,000

a. Dependent Variable: Current Salary

En el apartado (a) de Tabla 5.1, podemos observar que el modelo de variables principales explica el 46.2 % de la varianza de 'Salary', y que el 'bloque formado por las variables *dummy* 'Custodial' y 'Manager' es significativo. En (b) queda patente el valor significativo de ajuste del modelo, el p-value del estadístico F, para la ecuación de conjunto, es significativo. En (c) se muestra que los coeficientes estimados son todos significativos. El coeficiente de 'Educational Level' es, redondeando, de 3255 (t =

13.05, p < .001), lo que indica que posee un efecto positivo sobre el 'Salary'; es decir, a mayor nivel de años de estudio, mayor salario cobra también actualmente cualquier empleado de esa empresa de la muestra y que, manteniendo constantes los demás valores de la Ecuación 5.2, por cada año más de estudios, se cobraría en esa empresa 3255 dólares más al año. Asimismo, el coeficiente de 'D_JC_custodial' es de 5224.8 (*t = 2.103, p = .036*), lo cual indica que la ordenada en el origen (y por ser rectas paralelas, a lo largo de toda la distribución de datos) es significativamente mayor el 'Salary' de los 'Custodial' que el de los 'Clerical'; y el coeficiente de 'D_JC_manager' es de 7425.1 (*t = 4.086, p = 0.000*), indicando que para cualquier valor de 'Educ' el 'Salary' es significativamente mayor en los 'Manager' que en el grupo de referencia, 'Clerical'.

El conjunto de los apartados que acabamos de comentar, nos llevan a plantear, redondeando a las décimas, la Ecuación de conjunto 5.2, para los tres grupos de 'Jobcat':

$$\text{Salary'} = 9380.5 + 3255 \cdot \text{Educ} + [5224.8 \cdot \text{D\_JC\_custodial} + 7425.1 \cdot \text{D\_JC\_manager}] \quad \textbf{(5. 3)}$$

En este punto, puede ser interesante que el lector ponga a prueba lo que ya ha aprendido en las unidades 3 y 4 y que plantee los pronósticos que pueden derivarse para cada nivel de 'Jobcat' considerando las variables *dummy*.

Así pues, supongamos que la variable 'D_JC_custodial' es cero y 'D_JC_manager' también es cero, por lo que nos centramos en el grupo 'Clerical', el salario pronosticado para el grupo de 'Clerical' será:

$$Salary'_{clerical} = 9380.5 + 3255 \cdot Educ \quad \textbf{(5. 4)}$$

El salario pronosticado para 'Custodial', se obtiene cuando 'D_JC_custodial' = 1, y 'D_JC_manager' =0, y será:

$$\begin{aligned} Salary'_{custodial} &= 9380.5 + 3255 \cdot Educ + 5224.8 \cdot (1) + 7425.1 \cdot (0) \\ &= 9380.5 + 5224.8 + 3255 \cdot Educ = 14605.3 + 3255 \cdot Educ \end{aligned} \quad \textbf{(5. 5)}$$

El salario pronosticado para los 'Manager', se obtiene cuando 'D_JC_custodial' = 0, y 'D_JC_manager' = 1, cuya ecuación será:

$$Salary'_{manager} = 9380.5 + 3255 \cdot Educ + 5224.8 \cdot (0) + 7425.1 \cdot (1)$$
$$= 9380.5 + 7425.1 + 3255 \cdot Educ = 16805.6 + 3255 \cdot Educ \quad \textbf{(5. 6)}$$

Es obvio, como se deriva desde la Ecuación 5.4 hasta la 5.6, que las tres líneas anteriores son paralelas, puesto que el valor de sus pendientes no varía, pero sí cambia la ordenada en el origen, que da inicio a un valor diferente para cada nivel de categoría.

Podemos corroborar lo anterior atendiendo a los valores de los pronósticos obtenidos con la ecuación de regresión (guardados en el fichero de datos con el nombre de variable 'PRE_1'), realizando el diagrama de dispersión mediante la sintaxis:

```
GRAPH
      /SCATTERPLOT(BIVAR)=educ WITH PRE_Sal_f_educ_jobcat
           BY jobcat
      /MISSING=LISTWISE.
```

Que permite obtener la Figura 5.1, en la que se puede comprobar que existen tres rectas paralelas, según el grupo de referencia que consideremos.

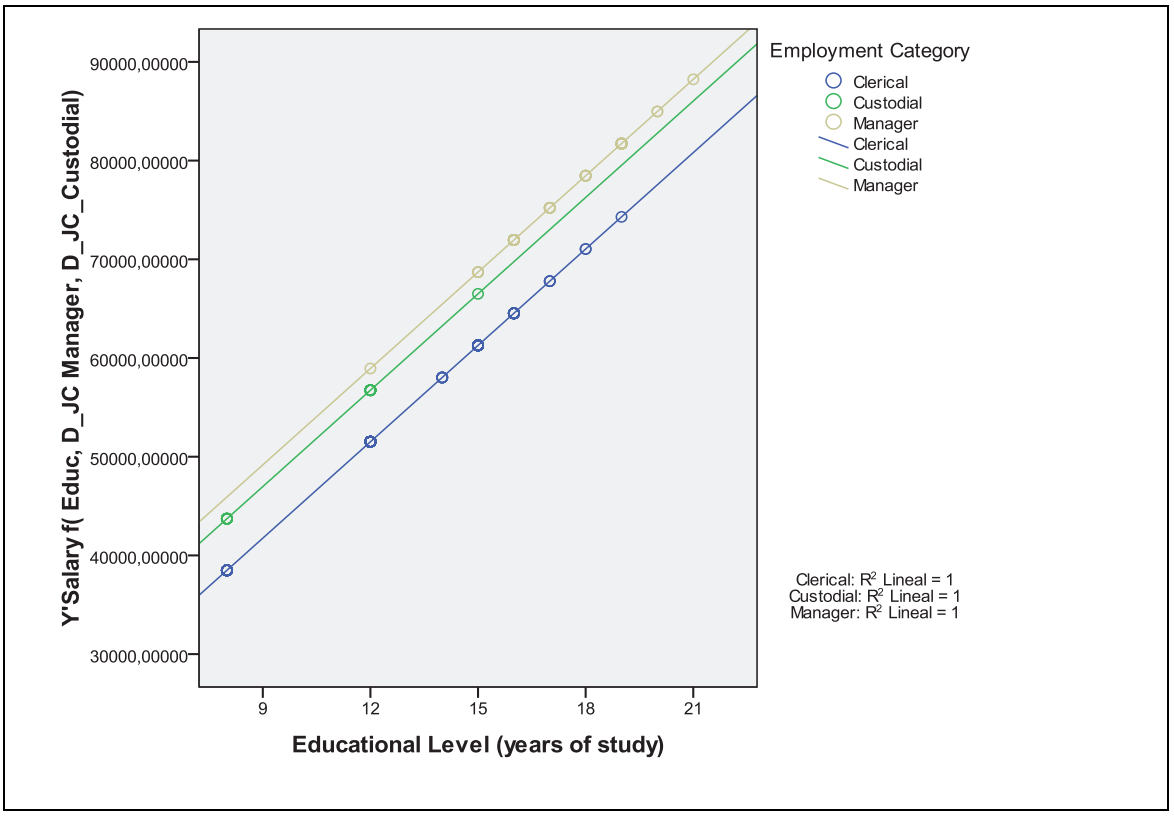


Figura 5.1. Diagrama de dispersión de los pronósticos de 'Salary' en función de 'Educ' para cada uno de los tres grupos de 'Jobcat'

Es decir, la predicción de 'Salary' en función de 'Educ' es diferente en intersección, b_0, si se pertenece a un grupo de categoría laboral u otro, pero la pendiente, b_1, es la misma para los tres grupos. Además, el grupo que a igualdad de nivel de años de 'Educ' más ingresos tiene es el de los 'Manager', siguiéndole el de los 'Custodial', y por último el de los 'Clerical'; además, la distancia existente entre las tres líneas es la misma (medida sobre el eje de ordenadas, paralelamente al eje Y, a lo largo de todo el recorrido de las rectas de pronóstico de la regresión); de este modo, la distancia entre los 'Clerical' y los 'Manager' es de 7425.1 (el valor del coeficiente de la variable 'D_JC_manager' en la Ecuación 5.2), la existente entre los 'Clerical' y los 'Custodial' es de 5224.8 (Ecuación 5.2), y la distancia entre los 'Clerical' y los 'Manager' será de 2200.3 (7425.1 – 5224.8).

5.2. Interacción: una variable independiente continua con otra variable independiente de tres categorías

En el modelo de la Ecuación 5.3 se comprueba cómo los tres grupos de la variable 'Jobcat' siguen líneas paralelas (la misma pendiente) con diferentes ordenadas en el origen. Se plantea ahora la hipótesis de si los grupos siguen líneas no paralelas, es decir, se ha de estimar y valorar un modelo de regresión lineal en el que tengamos en cuenta el término de interacción entre ambas variables independientes; para ello, hemos de considerar que la variable categórica 'Jobcat' posee tres niveles (o grupos).

Una vez comprobado que la matriz de datos contiene las variables *dummy* correspondientes a los tres niveles de la variable 'Jobcat', definidos con dos variables: 'D_JC_custodial' y 'D_JC_manager', procederemos a poner a prueba el modelo con interacción, que en forma compacta funcional es: Salario = f('Educ'*'Catlab'), que desarrollado funcionalmente queda: Salario = f('Educ', 'Catlab', 'Educ'*'Catlab), y teniendo en cuenta que 'Catlab' ahora está formada por dos variables *dummy* ('D_JC_custodial' y 'D_JC_manager'), la ecuación completa expresada de manera funcional con variables *dummy* quedará:

$$\text{Salario} = \text{f}(\text{'Educ'}, [\text{'D\_JC\_custodial'}, \text{'D\_JC\_manager'}], [\text{'Educ} * \text{D\_JC\_ custodial'}, \text{'Educ} * \text{D\_JC\_manager}]') \quad (5.7)$$

Para plantear la ecuación de regresión, es necesario que, previamente, realicemos los productos de las variables: 'Edu* D_JC_ custodial' y 'Edu* D_JC_manager'. Corriendo la sintaxis:

```
COMPUTE EducXD_JC_custodial=educ * D_JC_custodial.
     EXECUTE.
COMPUTE EducXD_JC_manager=educ * D_JC_manager.
     EXECUTE.
```

La Figura 5.2 contiene los valores obtenidos para las dos nuevas variables que se han generado tras correr la sintaxis COMPUTE.

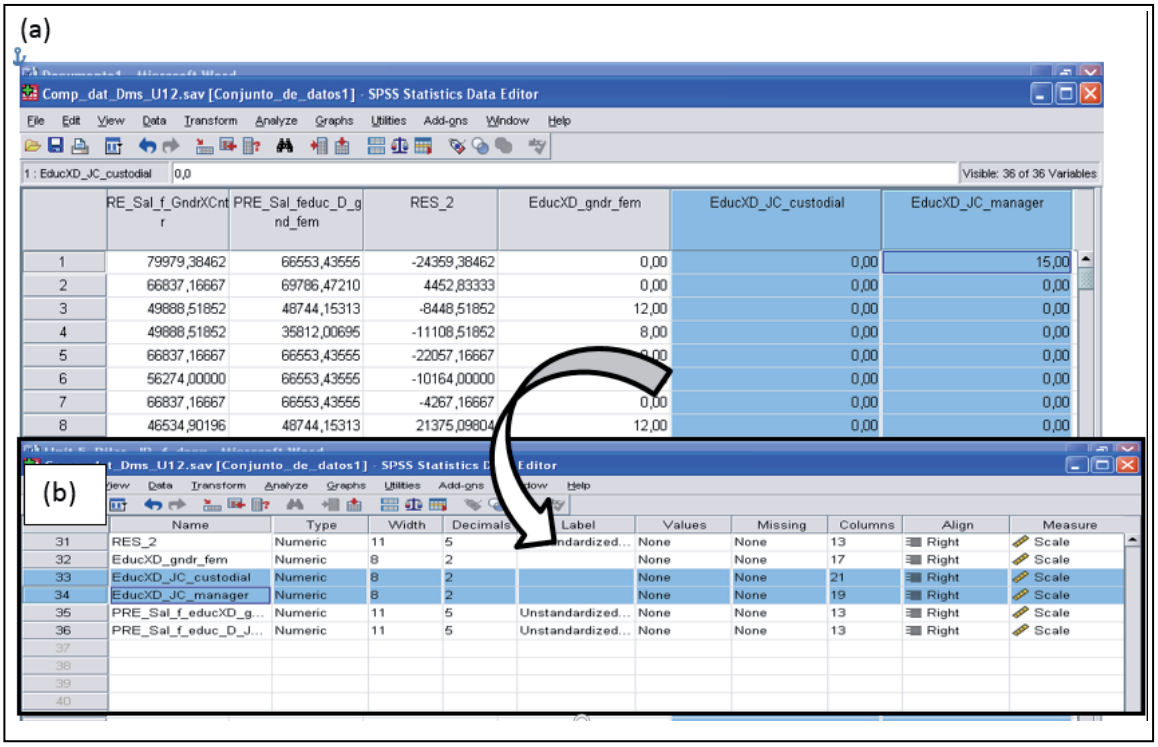


Figura 5.2. Variables 'EducXD_JC_custodial' y 'EducXD_JC_manager' en modo (a) 'Data view' y (b) en modo 'Variable view', como aparecen en la matriz de datos del SPSS

En este punto conviene que comprobemos, en la matriz de datos, que se han generado adecuadamente tanto las variables recodificadas, como las nuevas variables producto y procedamos a hacer el análisis de regresión con los términos multiplicativos que, en forma de algebra de regresión, aparece en la Ecuación 5.8.

$$\text{Salary}' = b_0 + b_1 \cdot \text{Edu} + \left[b_2 \cdot \text{D}_{\text{JC}_{\text{custodial}}} + b_3 \cdot \text{D}_{\text{JC}_{\text{manager}}}\right] + \left[b_4 \cdot \text{Edu} * \text{D}_{\text{JC}_{\text{custodial}}} + b_5 \cdot \text{Edu} * \text{D\_JC\_manager}\right] \quad (5.8)$$

IMPORTANTE
Guarda el fichero de datos como: 'Comp_dat_Dms_U5.sav', desde el que seguiremos los análisis en esta unidad y que contendrá las interacciones 'Edu*D_JC_custodial' y 'Edu*D_JC_manager'.
Si la interacción entre dos variables es significativa, los términos de las variables principales han de dejarse en la ecuación aunque resulten ser no significativos.

Desde SPSS, podemos estimar la Ecuación 5.6 corriendo la sintaxis:

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA CHANGE
```

```
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT salary
/METHOD=ENTER educ
/METHOD=ENTER D_JC_custodial D_JC_manager
/METHOD=ENTER EducXD_JC_custodial EducXD_JC_manager
/SAVE PRED RESID.
```

La sintaxis incluye tres veces el METHOD=ENTER, para indicar al programa que estime el modelos de regresión con tres bloques; de modo que, el primer bloque incluye la variable independiente 'Educ', en un segundo bloque se incluyen las variables *dummy*: 'D_JC_custodial' y 'D_JC_manager' (correspondientes a la variable 'Jobcat'), y el tercer bloque contiene: 'EducXD_JC_custodial' y 'EducXD_JC_manager' (que corresponde a la interacción de 'Educ'*'Jobcat'). En realidad, para llevar a cabo la regresión con interacción necesitamos conocer la significación de 'EducXD_JC_custodial' y 'EducXD_JC_manager' en bloque, puesto que si es significativo este bloque de interacción, se han de dejar las variables anidadas bajo este producto, aunque fuesen no significativas estadísticamente.

Además, mediante la línea: /STATISTICS COEFF OUTS R ANOVA CHANGE, pedimos al programa que haga una valoración de si el efecto de cada bloque, al añadirse al bloque anterior, es significativo.

Con la línea: /SAVE PRED RESID indicamos al programa que, al finalizar el análisis, guarde los valores predichos y los residuales en la matriz de datos. Ambos se implementarán al final de la matriz de datos como nuevas variables calculadas para la regresión considerada.

IMPORTANTE
Guarda los valores predichos con el nombre 'PRE_Sal_f_EducXD_JobCat'.

En la salida de resultados podemos comprobar la estimación del modelo. La Tabla 5.2 contiene la información más importante sobre la estimación del modelo.

El apartado a) de la Tabla 5.2 muestra el valor de R^2 del tercer bloque, que es el que incluye la interacción, indicando que el modelo lineal explica el 47 % de la variable 'Salary'; se puede ver en la tabla que la mejora del R^2 del segundo al tercer bloque es de

.8 % (47.0 % - 46.2 %), pero esta mejora es estadísticamente significativa (la mejora de *F* del modelo 3 respecto del modelo 2 es de 3677, con df: 2, 468, p = .026), es decir, el bloque de las dos variables de la interacción es estadísticamente significativo.

Tabla 5.2.
Resultados de la Ecuación de regresión 5.6. (a) Ajuste del cambio en el valor de F para cada bloque de variables, (b) Significación global del modelo de regresión, y (c) Significación de los coeficientes

(a)

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,661[a]	,436	,435	$12,251.190	,436	365,284	1	472	,000
2	,680[b]	,462	,459	$11,993.328	,026	11,257	2	470	,000
3	,686[c]	,470	,465	$11,925.606	,008	3,677	2	468	,026

a. Predictors: (Constant), Educational Level (years of study)
b. Predictors: (Constant), Educational Level (years of study), Custodial = 1, otherwise = 0, Manager = 1, Other = 0
c. Predictors: (Constant), Educational Level (years of study), Custodial = 1, otherwise = 0, Manager = 1, Other = 0, EducXD_JC_custodial, EducXD_JC_manager
d. Dependent Variable: Current Salary

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
3		59110263758	5	11822052752	83,125	,000[d]
		66559001021	468	142220087,7		
		1,257E+11	473			

a. Dependent Variable: Current Salary
d. Predictors: (Constant), Educational Level (years of study), Custodial = 1, otherwise = 0, Manager = 1, Other = 0, EducXD_JC_custodial, EducXD_JC_manager

(c)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
3		15912,481	3513,910		4,528	,000
		2986,101	268,711	,528	11,113	,000
		-7249,412	11524,367	-,103	-,629	,530
		-29314,767	14497,954	-,687	-2,022	,044
	EducXD_JC_custodial	1153,925	1087,501	,171	1,061	,289
	EducXD_JC_manager	2198,151	855,216	,894	2,570	,010

a. Dependent Variable: Current Salary

Como se resalta en el apartado b), comprobamos que toda la ecuación del modelo 3 (con la interacción) es significativa. En el apartado c) resaltamos los valores de significación

que alcanza cada uno de los coeficientes que integran la ecuación. Teniendo en cuenta lo anterior, podemos desarrollar las siguientes ecuaciones:

La Ecuación 5.8 es la de la regresión para todos los grupos, que después permitirá establecer una diferente para cada nivel de 'Jobcat':

$$\begin{aligned}\text{Salary}' = {} & 15912.5 + 2986.1 \cdot \text{Educ} \\ & + \left[(-7249.4) \cdot \text{D}_{\text{JC}_{\text{custodial}}} + (-29314.8) \cdot \text{D}_{\text{JC}_{\text{manager}}}\right] \\ & + [1153.9 \cdot \text{EduccXD\_JC\_custodial} + 2198.1 \cdot \text{EducXD\_JC\_manager}]\end{aligned} \quad \textbf{(5. 9)}$$

Obsérvese que se ha incluido el término 'EduXD_JC_custodial' (aunque su coeficiente es no significativo) porque es inseparable del otro término de la interacción y porque entre los dos son significativos, con p = .026.

Llegados a este punto, se ha de observar que la Ecuación 5.9 (con interacción de 'Educ'*'Jobcat') ajusta mejor a los datos de nuestra muestra que la Ecuación 5.3 (solo con variables principales), pues la Ecuación 5.9, correspondiente a los resultados de la Tabla 5.2, en donde se comprueba que la interacción de las variables (modelo 3) hace una aportación significativa a la explicación de la variable dependiente 'Salary', con respecto a lo que hace solo la regresión con variables principales (modelo 2 de la Tabla 5.1 y modelo 2 de la Tabla 5.2).

5.3. Interpretación de la interacción

Si interpretamos la Ecuación 5.7 encontramos que el coeficiente de 'D_JC_custodial', con valor –7249.4, indica que la ordenada en el origen es 7249.4 unidades menos que la del grupo de referencia, en nuestro caso 'Clerical', y que no hay diferencias significativas entre ambas (p = .530), pese a no ser significativa, ha de dejarse en la ecuación debido a que la interacción es significativa. Al mismo tiempo, el valor –29314.8, como coeficiente de D_JC_manager, significa que la ordenada en el origen es 29314.8 unidades menor que la del grupo 'Clerical', y hay diferencias significativas entre ambas ordenadas en el origen.

En cuanto al coeficiente de 'EduXD_JC_custodial' (con valor 1153.9) indica que la pendiente es 1153.9 unidades mayor que la del grupo de referencia 'Clerical', aunque no hay diferencias significativas entre ambas (p = .289). Del mismo modo, el coeficiente de 'EduXD_JC_manager' (2198.1) indica que la pendiente es 2198.1 unidades mayor que la del grupo de referencia 'Clerical', y, en este caso sí hay diferencias significativas entre ambas pendientes (p = .010). La probabilidad de estas interacciones es de .026 (Tabla 5.2 (a)), indicando que esta interacción es en conjunto estadísticamente significativa.

Desde de la Ecuación 5.9, podemos plantear las ecuaciones de regresión para cada grupo.

Para el grupo de referencia: 'Clerical' (D_JC_custodial = 0; D_JC_manager = 0):

$$\begin{aligned}\text{Salary}'_{Clerical} = {} & 15912.5 + 2986.1 \cdot \text{Edu} + \left[(-7249.4 \cdot 0] + (-29314.8 \cdot 0]\right] \\ & + [1153.9 \cdot \text{Edu} * 0 + 2198.1 \cdot \text{Edu} * 0] = 15912.5 + 2986.1 \cdot \text{Edu}\end{aligned} \quad \textbf{(5. 10)}$$

Por tanto, si pronosticásemos 'Salary' en el grupo 'Clerical', cuando se tiene Edu = 0, el grupo de referencia tendrá el valor de la ordenada en el origen de la Ecuación de regresión de conjunto 5.9, y la pendiente incrementará en 2986.1 unidades por cada año de incremento en educación.

Para el grupo 'Custodial' (D_JC_custodial = 1; D_JC_manager = 0):

$$\begin{aligned}\text{Salary}'_{Custodial} = {} & 15912.5 + 2986.1 \cdot \text{Edu} + [(-7249.4 \cdot 1) + (-29314.8 \cdot 0)] \\ & + [1153.9 \cdot \text{Edu} * 1 + 2198.1 \cdot \text{Edu} * 0] = 8663.1 + 4140 \cdot \text{Edu}\end{aligned} \quad \textbf{(5. 11)}$$

En este caso, comprobamos que la ordenada en el origen es de 8663.1, que es la que tiene el grupo de referencia, 15912.5, más el valor del coeficiente de la variable *dummy* 'D_JC_custodial' (-7249.4), es decir, 8663.1; o lo que es lo mismo, 7249.4 unidades menos que el grupo de referencia. Mientras la pendiente es de 2986.1 + 1153.9; es decir, por término medio, se incrementará en 4140 unidades de la variable dependiente ('Salary') por cada año de incremento en Educación.

Para el grupo 'Manager' (D_JC_custodial = 0; D_JC_manager = 1):

$$\begin{aligned}\text{Salary}'_{\text{Manager}} = {} & 15912.5 + 2986.1 \cdot \text{Edu} + [(-7249.4\ 0) + (-29314.8 \cdot 1)] \\ & + [1153.9\ \text{Edu} * 0 + 2198.1 \cdot \text{Edu} * 1] = -13402.3 + 5184.2 \cdot \text{Edu}\end{aligned} \quad \textbf{(5. 12)}$$

Se comprueba que la ordenada en el origen es de -13402.3, que puede obtenerse mediante la suma 15912.5 + (-29314.8) = -13402.3; es decir 29314.8 unidades menos que el grupo de referencia. En cuanto a la pendiente, por cada año de educación, con respecto al grupo de referencia ('Clerical'), incrementa 2198.1 unidades más; es decir, para 'Manager', por cada año de educación, 'Salary' incrementará, por término medio, 2986.1 + 2198.1 = 5184.2 dólares.

Para representar los valores estimados desde la regresión con interacción podemos correr la sintaxis:

```
GRAPH
      /SCATTERPLOT(BIVAR)=educ WITH PRE_Sal_f_EducXD_JobCat
            BY jobcat
      /MISSING=LISTWISE.
```

Y obtendremos el diagrama de dispersión de la Figura 5.3, en el que hemos incluido la línea de ajuste en subgrupos por 'Jobcat', además de marcar las ordenadas en el origen en cada uno de ellos.

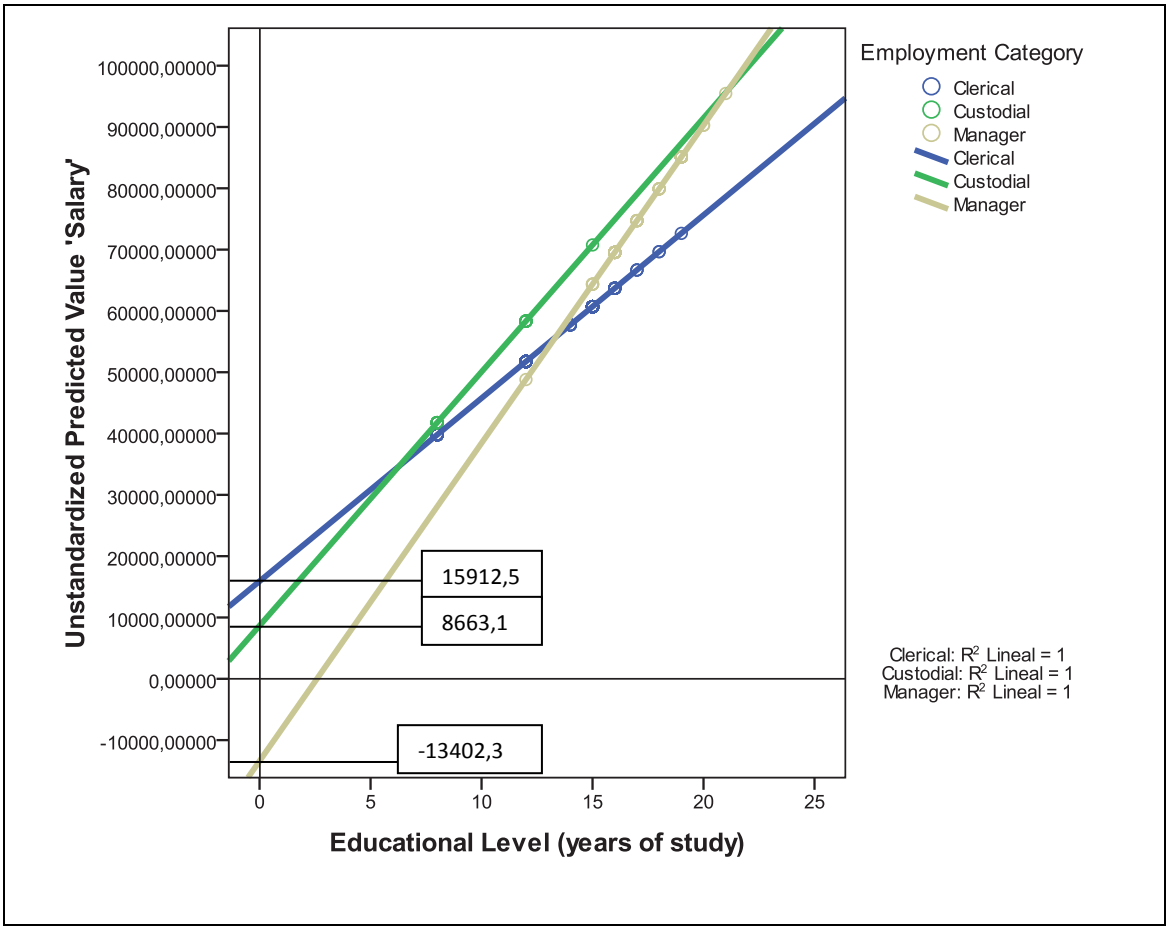


Figura 5.3. Líneas de regresión de 'Salary' en función de 'Educ*Jobcat'

5.4. Conclusiones

- La interacción de variables indica que los tres grupos tienen pendientes significativamente diferentes (en conjunto), lo que no significa que, necesariamente, todas las pendientes sean distintas entre sí, para comprobarlo tendríamos que hacer «comparaciones a la medida» entre cada par de grupos, pero sí indica que hay dos o más pendientes que difieren entre sí.
- Si fijamos nuestra atención en la Tabla 5.2, apartado c) solo podemos afirmar que:
    a) Las pendientes de 'Clerical' y 'Custodial' no difieren significativamente entre sí ($t = 1.061$, $p = .289$).
    b) Las pendientes de 'Clerical' y 'Manager' difieren significativamente entre sí ($p < .05$), efectivamente, $t = 2.570$, $p = .010$.
- Si quisiéramos comprobar si hay diferencias significativas entre las pendientes de 'Custodial' y de 'Manager', tendríamos 2 opciones: a) hacer comparaciones a la medida (prueba de t de Student-Fisher), o b) más fácil: utilizar o bien 'Custodial' o bien 'Manager' como grupo de referencia y volver a hacer la regresión con interacción (pero con uno de esos dos grupos de referencia).
- Respecto a las ordenadas en origen, mirando nuevamente la Tabla 5.2, apartado c), también podemos afirmar:
    a) Las intersecciones de 'Clerical' y de 'Manager' difieren entre sí ($p < .05$), $t = -2.022$, $p = .044$.
    b) Las intersecciones de 'Clerical' y de 'Custodial' **NO** difieren entre sí ($p > .05$), $t = -0.629$, $p = .530$.
- Es muy importante señalar que la interacción es significativa y, por tanto, las rectas de pronóstico no son paralelas, es decir, es más correcta la Ecuación 5.7 que la 5.3. Por tanto, si un investigador se hubiese conformado con hacer solo la ecuación con variables principales, habría asumido que las pendientes son paralelas, lo cual no es del todo cierto, ya que es más correcto asumir el no paralelismo de las rectas de regresión, como indica la significación estadística de la interacción de variables.

Lecturas recomendadas

Esta unidad es un desarrollo de las unidades anteriores, y se han de tener claros los conceptos que hemos desarrollado hasta ahora para poder seguir profundizando en la interacción de variables. De todos modos, para poder afianzar los conocimientos adquiridos, se recomienda la lectura del libro de Hardy (1993) sobre regresión con variables *dummy*, y la de los libros de Aiken y West (1991) y Jaccard y Turrisi (2003) sobre interacción de variables. Debido a que estos libros no explican cómo hacer representaciones a partir de los resultados obtenidos, recomendamos el repaso de esta unidad y de las unidades anteriores de estos apuntes, con el fin de llevar a cabo la correcta representación del modelo de conjunto.

Bibliografía

Aiken, L. S.; West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.

Jaccard, J.; Turrisi, R. (2003). *Interaction effects in multiple regression.* (2.ª ed.). Thousand Oaks, CA: Sage.

Actividades

Se quiere hacer la regresión lineal con interacción para 'Educ*Center', de modo que la regresión será: Salary= f('Educ*Center). Utiliza para ello el fichero: 'Comp_dat_Dms_U5.sav'.

Al recodificar 'Center' en sus variables dummies, obtendremos las siguientes variables, y utilizando 'Center' = 1 como grupo de referencia, tendremos:

'Center' ↔ ['D_Ctr_2', 'D_Ctr_3', 'D_Ctr_4', 'D_Ctr_5', 'D_Ctr_6']

1. Haz los productos (interacción) de 'Edu*Center', llama a estas nuevas variables 'EduXCtr_2', 'EduXCtr_3', 'EduXCtr_4','EduXCtr_5,'EduXCtr_6' (de 'Educ*Centro 2', 'Educ*Centro 3'…).

2. Haz la ecuación de regresión:

   Salary = f('Edu', ['Ctr_2', 'Ctr_3', 'Ctr_4' , 'Ctr_5' , 'Ctr_6'],
   ['EduXCtr_2', 'EduXCtr_3', 'EduXCtr_4','EduXCtr_5,'EduXCtr_6']).

3. Interpreta la significación de los estadísticos de conjunto (sobre todo para el conjunto de la ecuación y para la interacción).

4. Escribe la ecuación de regresión de conjunto.

5. Interpreta los coeficientes obtenidos.

6. Escribe la ecuación de regresión para cada grupo.

7. Haz una figura con los pronósticos obtenidos en la regresión.

8. Interpreta los resultados obtenidos.

Unidad 6. Regresión lineal con interacción de tres variables independientes: una continua y dos categóricas

Introducción

En esta unidad exploraremos y cuantificaremos la relación lineal de una variable dependiente en función de tres variables independientes, donde una de las variables independientes será continua, la variable 'Educ', y las otras dos serán categóricas, 'Gender' y 'Jobcat'. Para ello, teniendo en cuenta que las dos variables independientes categóricas se han de convertir en variables *dummy*, incluiremos todas las variables principales independientes, las posibles interacciones dos a dos entre las variables independientes y la interacción de las tres variables independientes. Aprenderemos a plantear la ecuación de regresión completa, a valorar si hay diferencias en las respectivas ordenadas en el origen y en las pendientes, y a utilizar el modelo con fines aplicados.

Objetivos

Cuando el estudiante finalice esta unidad sabrá:

- Diseñar el modelo completo con la interacción de tres variables independientes: dos categóricas y una continua.
- Ajustar el modelo correcto, eliminando las interacciones no significativas.
- Valorar y explicar la interacción de variables independientes categóricas, sabiendo que si tal interacción es significativa habrá diferencias en las respectivas ordenadas en el origen.
- Valorar y explicar que cuando en la interacción significativa hay una, o alguna, variable independiente continua, las pendientes (de la/s recta/s o del plano o planos) están afectadas según los niveles de la otra variable, o de las otras variables.
- Formular la ecuación resultante de conjunto y las ecuaciones de ajuste para cada grupo del modelo (tantas ecuaciones como grupos hay).
- Realizar la figura que apoye las explicaciones del modelo propuesto.
- Interpretar adecuadamente el modelo (en conjunto, cada ecuación final y cada uno de sus coeficientes e interacciones).

6.1. Regresión lineal con interacción de tres variables independientes: una variable independiente continua y dos categóricas

Supongamos que nuestro interés se centra en comprobar la existencia de interacción entre una variable independiente continua, como por ejemplo 'Educ', con otras dos categóricas, 'Gender' y 'Jobcat', para explicar linealmente la variable dependiente 'Salary'. Este interés puede representarse de forma resumida con la relación funcional:

$$\text{Salary} = \text{f}(\text{'Educ} * \text{Gender} * \text{Jobcat'}) \tag{6. 1}$$

Ahora bien, no hemos de olvidar el principio de anidamiento, según el cual en el modelo anterior se deben incluir todas las variables simples e interacciones de orden inferior, hasta llegar a la interacción de mayor orden ('Educ*Gender*Jobcat'), pudiendo distinguir en el modelo tres apartados diferenciados (variables principales independientes, las posibles interacciones entre cada dos variables independientes y la interacción de las tres variables independientes):

$$\begin{aligned}\text{Salary} = \text{f}(\text{'Educ} * \text{Gender} * \text{Jobcat'}) = \\ \begin{array}{rl} \text{f}(\text{'Educ'}, [\text{'Gender'}], [\text{'Jobcat'}], & \text{Variables principales} \\ \hline [\text{'Educ} * \text{Gender'}], [\text{'Educ} * \text{Jobcat'}], [\text{'Gender} * \text{Jobcat'}], & \text{Interacción de dos variables independientes} \\ \hline [\text{'Educ} * \text{Gender} * \text{Jobcat'}]) & \text{Interacción de las tres variables independientes} \end{array}\end{aligned} \tag{6. 2}$$

Como podemos observar en la matriz de datos de trabajo que utilizaremos en esta unidad ('Comp_dat_Dms_U6.sav'), la variable 'Educ' es una variable continua, por lo que se representará con una única variable; la variable 'Gender' es dicotómica, por lo que se representará con una única variable *dummy*, 'D_gndr_fem', donde el grupo de referencia será 'Male'; y la variable 'Jobcat' es una variable de tres categorías que se representará por dos variables *dummy*, 'D_JC_Custodial' y 'D_JC_Manager', donde el grupo de referencia será 'Clerical'. Así pues, el modelo anteriormente presentado debería reformularse con las variables *dummy* utilizadas, resultando el siguiente modelo equivalente al anterior:

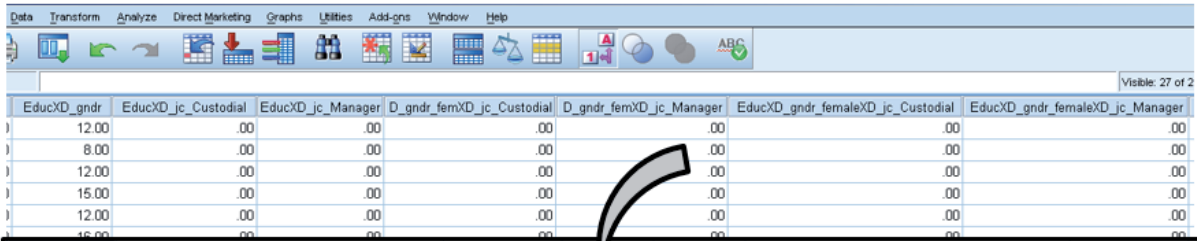
$$\begin{array}{ll} \text{Salary} = \text{f}(\text{'Educ'}, [\text{'D}_{\text{gndr}_{\text{fem}}}\text{'}], \left[\text{'D}_{\text{jc}_{\text{Custodial}}}\text{'}, \text{'D}_{\text{jc}_{\text{Manager}}}\text{'}\right], & \text{Variables principales} \\ \hline \begin{array}{l}[\text{'Educ} * \text{D}_{\text{gndr}_{\text{fem}}}\text{'}], \\ \left[\text{'Educ} * \text{D}_{\text{jc}_{\text{Custodial}}}\text{'}, \text{'Educ} * \text{D}_{\text{jc}_{\text{Manager}}}\text{'}\right], \\ \left[\text{'D}_{\text{gndr}_{\text{fem}}} * \text{D}_{\text{jc}_{\text{Custodial}}}\text{'}, \text{'D}_{\text{gndr}_{\text{fem}}} * \text{D}_{\text{jc}_{\text{Manager}}}\text{'}\right]\end{array} & \text{Interacción de dos variables independientes} \\ \hline [\text{'Educ} * \text{'D\_gndr}_{\text{fem}} * \text{D\_jc}_{\text{Custodial}}\text{'}, \text{'Educ} * \text{'D\_gndr}_{\text{fem}} * \text{D\_jc}_{\text{Manager}}\text{'}]) & \text{Interacción de las tres variables independientes} \end{array} \tag{6. 3}$$

Por lo tanto, en nuestro fichero de datos tendremos que tener la variable dependiente 'Salary', y las variables independientes 'Educ', 'D_gndr_female', 'D_jc_Custodial', 'D_jc_Manager', 'Educ*D_gndr_female', 'Educ*D_jc_Custodial', 'Educ*D_jc_Manager', 'D_gndr_female* D_jc_Custodial', 'D_gndr_female*D_jc_Manager', 'Educ*D_gndr_female* D_jc_Custodial', 'Educ*D_gndr_female*D_jc_Manager', de las cuales, algunas ya tenemos creadas de unidades anteriores, y otras tendremos que crearlas en este momento, de la misma forma que lo hemos venido haciendo hasta ahora, corriendo la siguiente sintaxis:

```
COMPUTE EducXD_gndr= Educ * D_gndr_fem.
     VARIABLE LABELS  EducXD_gndr 'Educ*D_gndr_fem'.
     EXECUTE.
COMPUTE EducXD_jc_Custodial = Educ*D_jc_Custodial.
     VARIABLE     LABELS     EducXD_jc_Custodial     'Educ*
          D_jc_Custodial'.
     EXECUTE.
COMPUTE EducXD_jc_Manager= Educ*D_jc_Manager.
     VARIABLE LABELS EducXD_jc_Manager 'Educ*D_jc_Manager'.
     EXECUTE.
COMPUTE              D_gndr_femXD_jc_Custodial              =
D_gndr_fem*D_jc_Custodial.
     VARIABLE        LABELS        D_gndr_femXD_jc_Custodial
          'D_gndr_fem*D_jc_Custodial'.
     EXECUTE.
COMPUTE D_gndr_femXD_jc_Manager = D_gndr_fem*D_jc_Manager.
     VARIABLE         LABELS         D_gndr_femXD_jc_Manager
          'D_gndr_fem*D_jc_Manager'.
     EXECUTE.
COMPUTE                    EducXD_gndr_femaleXD_jc_Custodial=
Educ*D_gndr_fem*D_jc_Custodial.
     VARIABLE    LABELS    EducXD_gndr_femaleXD_jc_Custodial
          'Educ*D_gndr_fem*D_jc_Custodial'.
     EXECUTE.
COMPUTE  EducXD_gndr_femaleXD_jc_Manager  =  Educ*D_gndr_fem*
D_jc_Manager.
     VARIABLE     LABELS     EducXD_gndr_femaleXD_jc_Manager
          'Educ*D_gndr_fem*D_jc_Manager'.
     EXECUTE.
```

Los resultados de estas transformaciones de datos aparecen en la Figura 6.1.

(a)



EducXD_gndr	EducXD_jc_Custodial	EducXD_jc_Manager	D_gndr_femXD_jc_Custodial	D_gndr_femXD_jc_Manager	EducXD_gndr_femaleXD_jc_Custodial	EducXD_gndr_femaleXD_jc_Manager
12.00	.00	.00	.00	.00	.00	.00
8.00	.00	.00	.00	.00	.00	.00
12.00	.00	.00	.00	.00	.00	.00
15.00	.00	.00	.00	.00	.00	.00
12.00	.00	.00	.00	.00	.00	.00

(b)

	EducXD_gndr	Numeric	8	2	Educ * D_gndr_fem	None	Nor
	EducXD_jc_Custodial	Numeric	8	2	Educ* D_jc_Custodial	None	Nor
23	EducXD_jc_Manager	Numeric	8	2	Educ* D_jc_Manager	None	Nor
24	D_gndr_femXD_jc_Custodial	Numeric	8		'D_gndr_fem * D_jc_Custodial'	None	Nor
25	D_gndr_femXD_jc_Manager	Numeric	8	2	'D_gndr_fem * D_jc_Manager'	None	Nor
26	EducXD_gndr_femaleXD_jc_Custodial	Numeric	8	2	'Educ*D_gndr_fem* D_jc_Custodial'	None	Nor
27	EducXD_gndr_femaleXD_jc_Manager	Numeric	8	2	'Educ*D_gndr_fem* D_jc_Manager'	None	Nor

Figura 6.1. Variables de la interacción de 'Educ*D_gndr', 'Educ*D_jc_Custodial', 'Educ*D_jc_Manager', 'D_gndr_fem*D_jc_Custodial', 'D_gndr_fem*D_jc_Manager', 'Educ*D_gndr_fem*D_jc_Custodial', 'Educ* D_gndr_fem*D_jc_Manager', (a) en el 'Data View' y (b) en el 'Variable View', tal como aparecen en el SPSS

IMPORTANTE
Guarda el fichero de datos como: 'Comp_dat_Dms_U6.sav', desde el que seguiremos los análisis en esta unidad y que contendrá las nuevas variables que hemos creado.

6.2. Estudio exploratorio

Con el fin de observar cómo resultarán los gráficos de regresión que plantearemos, elaboraremos la Tabla de contingencia 6.1.a con las variables categóricas empleadas, 'Jobcat' y 'Gender', siendo la sintaxis:

```
CTABLES
     /VLABELS VARIABLES=gender jobcat DISPLAY=DEFAULT
     /TABLE gender [C][COUNT F40.0, TOTALS[COUNT F40.0]] BY
          jobcat [C]
     /CATEGORIES VARIABLES=gender jobcat ORDER=A KEY=VALUE
          EMPTY=INCLUDE TOTAL=YES POSITION=AFTER
     MISSING=EXCLUDE.
```

Donde podemos observar que no existe ningún caso en la categoría cruzada Female-Custodial, lo que conllevará a que en vez de tener seis rectas de regresión (una para cada combinación de categorías), tengamos solo cinco a priori, es decir, si se pide una figura, no aparecerá la correspondiente a Female-Custodial.

Tabla 6.1. Tablas de contingencia con las variables empleadas: (a) 'Gender' - 'Jobcat' y (b) 'Gender' – 'Jobcat' – 'Educ'

(a) Gender * Employment Category Crosstabulation

		Employment Category			
		Clerical	Custodial	Manager	Total
		Count	Count	Count	Count
Gender	Female	206	0	10	216
	Male	157	27	74	258
	Total	363	27	84	474

(b) Gender * Employment Category * Educational Level (years of study)

				Educational Level (years of study)										
				8	12	14	15	16	17	18	19	20	21	Total
				Count	Count	Count	Count	Count	Count	Count	Count	Count	Count	Count
Gender	Female	Employment Category	Clerical	30	128	0	33	14	1	0	0	0	0	206
			Custodial	0	0	0	0	0	0	0	0	0	0	0
			Manager	0	0	0	0	10	0	0	0	0	0	10
			Total	30	128	0	33	24	1	0	0	0	0	216
	Male	Employment Category	Clerical	10	48	6	78	10	2	2	1	0	0	157
			Custodial	13	13	0	1	0	0	0	0	0	0	27
			Manager	0	1	0	4	25	8	7	26	2	1	74
			Total	23	62	6	83	35	10	9	27	2	1	258
	Total	Employment Category	Clerical	40	176	6	111	24	3	2	1	0	0	363
			Custodial	13	13	0	1	0	0	0	0	0	0	27
			Manager	0	1	0	4	35	8	7	26	2	1	84
			Total	53	190	6	116	59	11	9	27	2	1	474

Y si pedimos la Tabla personalizada 6.1.b, mediante la sintaxis:

```
CTABLES
     /VLABELS VARIABLES=gender jobcat educ DISPLAY=DEFAULT
     /TABLE gender [C] > jobcat [C][COUNT F40.0] BY educ [C]
     /CATEGORIES VARIABLES=gender jobcat educ ORDER=A
          KEY=VALUE EMPTY=INCLUDE TOTAL=YES POSITION=AFTER
          MISSING=EXCLUDE.
```

Dando los totales marginales de las variables 'Gender', 'Employment Category' y 'Educational Level (years of study)', pudiéndose observar cómo para la categoría Female-Manager todos los casos (diez en total) tienen 16 años de estudio.

6.3. Ecuación general y ecuación para cada grupo (interacción)

La función mediante la cual expresamos anteriormente la dependencia de la variable 'Salary' a partir de las variables 'Educ', 'Gender' y 'Jobcat', debemos expresarla en términos de ecuación estadística con el fin de proceder posteriormente a la estimación y estudio de cada uno de los parámetros que la integran:

$$\begin{aligned} \text{Salary} = {} & b_0 + b_1 \cdot \text{'Educ'} + \left[b_2 \cdot \text{'}D_{\text{gndr}_{\text{fem}}}\text{'}\right] + \left[b_3 \cdot \text{'}D_{\text{jc}_{\text{Custodial}}}\text{'} + b_4 \cdot \text{'}D_{\text{jc}_{\text{Manager}}}\text{'}\right] \\ & + \left[b_5 \cdot \text{'Educ} * D_{\text{gndr}_{\text{fem}}}\text{'}\right] \\ & + \left[b_6 \cdot \text{'Educ} * D_{\text{jc}_{\text{Custodial}}}\text{'} + b_7 \cdot \text{'Educ} * D_{\text{jc}_{\text{Manager}}}\text{'}\right] \\ & + \left[b_8 \cdot \text{'}D_{\text{gndr}_{\text{fem}}} * D_{\text{jc}_{\text{Custodial}}}\text{'} + b_9 \cdot \text{'}D_{\text{gndr}_{\text{fem}}} * D_{\text{jc}_{\text{Manager}}}\text{'}\right] \\ & + \left[b_{10} \cdot \text{'Educ} * \text{'}D_{\text{gndr}_{\text{fem}}} * D_{\text{jc}_{\text{Custodial}}}\text{'} + b_{11} \cdot \text{'Educ} * \text{'}D_{\text{gndr}_{\text{fem}}} * D_{\text{jc}_{\text{Manager}}}\text{'}\right] \\ & + e \end{aligned}$$

(6. 4)

A la hora de proceder con el ajuste del modelo, con el fin de poner a prueba las interacciones, en primer lugar debemos empezar realizando la regresión lineal por bloques con el método de análisis 'Introducir', incluyendo el modelo completo, donde en el último bloque pondremos a prueba la interacción más alta y en los bloques previos habremos introducido todas sus correspondientes interacciones anidadas y las variables principales, comprobando la significación de la interacción de mayor orden; si es significativa a nivel global, hemos finalizado el ajuste del modelo ya que todo ajusta, por lo que deberíamos calcular la ecuación global, la ecuación para cada grupo y tratar de realizar el gráfico correspondiente. Para ello, podemos correr la siguiente sintaxis:

```
REGRESSION
     /MISSING LISTWISE
     /STATISTICS COEFF OUTS R ANOVA CHANGE
     /CRITERIA=PIN(.05) POUT(.10)
     /NOORIGIN
     /DEPENDENT salary
     /METHOD=ENTER     educ     D_gndr_fem     D_JC_custodial
          D_JC_manager
     /METHOD=ENTER Educ_X_D_gndr_fem
     /METHOD=ENTER Educ_X_D_jc_Custodial Educ_X_D_jc_Manager
     /METHOD=ENTER                D_gndr_fem_X_D_jc_Custodial
          D_gndr_fem_X_D_jc_Manager
```

```
/METHOD=ENTER      Educ_X_D_gndr_female_X_D_jc_Custodial
     Educ_X_D_gndr_female_X_D_jc_Manager.
EXECUTE.
```

Si esta interacción más alta no ajustara, deberíamos probar el modelo con las interacciones anidadas dentro de la ecuación. Si alguna de estas interacciones anidadas no fuera significativa deberíamos de ir eliminándolas una a una, quitando en primer lugar aquellas que tengan el valor de p más grande, ya que alguna interacción puede convertirse en significativa si eliminamos su colinealidad con otras. Este proceso finalizará cuando nos den significativas las interacciones de nivel más alto (o en su defecto, las variables principales), momento en el que elaboraremos la ecuación global y para cada grupo, además de la realización de los gráficos e interpretación conjunta de gráficos y ecuaciones.

Tabla 6.2.
Resultados de la regresión por bloques 'Salary' = f('Educ''Gender'*Jobcat').*

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	,851[a]	,724	,721	$9,012.577	,724	307,231	4	469	,000
2	,854[b]	,729	,726	$8,932.295	,005	9,468	1	468	,002
3	,860[c]	,740	,736	$8,770.688	,011	9,703	2	466	,000
4	,862[d]	,743	,739	$8,727.704	,003	5,601	1	465	,018

a. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years)

b. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ * D_gndr_fem

c. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ * D_gndr_fem, Educ* D_jc_Custodial , Educ* D_jc_Manager

d. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ * D_gndr_fem, Educ* D_jc_Custodial , Educ* D_jc_Manager , 'D_gndr_fem * D_jc_Manager'

En la Tabla 6.2 de resumen del modelo resultante, el SPSS nos da la significación de mejora para el modelo de bloque en bloque, para los cuatro bloques. Aunque hayamos pedido cinco, el SPSS omite el cálculo del quinto bloque, indicándonos la salida del SPSS que se han eliminado del análisis las variables 'D_gndr_fem* D_jc_Custodial', 'Educ*D_gndr_fem* D_jc_Custodial', ya que el SPSS da el mensaje de que éstas 'are constants or have missing correlations'. Para el resto de bloques en los que se ha realizado el análisis podemos observar en la Tabla 13.2 cómo va modificándose el estadístico R^2 en los diferentes modelos propuestos. Hemos omitido la tabla de ANOVA de significación de conjunto de cada modelo, puesto que los valores de *F* para cada modelo dan una $p < .001$.

Si observamos además la Tabla 6.3, de coeficientes, podemos determinar que para los diferentes modelos existen interacciones que no resultan significativas, observando además

que en el último bloque se ha eliminado la interacción 'Educ*D_gndr_fem*D_jc_manager', indicándonos esto que esa variable de interacción es colineal con las otras variables.

Tabla 6.3.
Coeficientes de la regresión por bloques 'Salary' = f('Educ''Gender'*Jobcat')*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
4	(Constant)	9476,710	4480,532		2,115	,035
	Educational Level (years)	1605,737	321,858	,271	4,989	,000
	D_gndr_fem	1006,269	5618,617	,029	,179	,858
	D_jc_Custodial	22503,524	9197,564	,306	2,447	,015
	D_jc_Manager	-7539,151	11739,896	-,169	-,642	,521
	Educ * D_gndr_fem	-414,946	422,376	-,155	-,982	,326
	Educ* D_jc_Custodial	-1707,978	835,672	-,242	-2,044	,042
	Educ* D_jc_Manager	2085,978	698,769	,810	2,985	,003
	'D_gndr_fem * D_jc_Manager'	-8158,625	3447,222	-,069	-2,367	,018

a. Dependent Variable: Current Salary

Esto es, las variables del bloque 5, 'Educ*D_gndr_fem*D_jc_custodial' y 'Educ*D_gndr_fem*D_jc_manager', no aportan capacidad explicativa al modelo una vez introducidas las anteriores variables, por lo que podríamos perfectamente eliminar ese bloque 5 de variables independientes (el de interacción de las tres variables principales), dejando los cuatro restantes y la variable dependiente igual.

Si prestamos atención a los coeficientes del último bloque vemos que la interacción 'D_gndr_fem*D_jc_custodial' sigue sin aparecer, ya que no había ningún sujeto en ésta; además podemos observar también que el coeficiente de 'Educ*D_gndr_fem' es no significativo (aunque en la significación del bloque 2, da una p = .002, esta probabilidad es la de la interacción de 'Educ*D_gndr_fem' después de ser añadida a las variables principales). Para comprobar su correcto nivel de significación, en caso de duda, habría que pasarlo al final del análisis, como último bloque, dado que esta variable podría ser colineal con 'Educ*Catlab' y/o con 'Gndr*Catlab', por lo que si existiera colinealidad habría que eliminarlo del modelo. Realmente no haría falta pasar la variable 'Educ*D_gndr_fem' al final, porque al ser una sola variable de interacción, su significación viene ya dada por la de esa variable dentro de la ecuación final para todo el bloque 4 (p=.326).

Con el fin de que el alumno lo compruebe, correremos la anterior sintaxis modificando la posición del último bloque, obteniendo los resultados de la Tabla 6.4.

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA CHANGE
      /CRITERIA=PIN(.05) POUT(.10)
      /NOORIGIN
      /DEPENDENT salary
      /METHOD=ENTER     educ     D_gndr_fem     D_JC_custodial
           D_JC_manager
      /METHOD=ENTER Educ_X_D_jc_Custodial Educ_X_D_jc_Manager
      /METHOD=ENTER                D_gndr_fem_X_D_jc_Custodial
           D_gndr_fem_X_D_jc_Manager
      /METHOD=ENTER Educ_X_D_gndr.
      EXECUTE.
```

En la Tabla 6.4 podemos observar que el último bloque (‘Educ * D_gndr_fem’) resulta ser no significativo, así como el coeficiente del modelo final (p=.326), y que estos resultados son los mismos que los obtenidos en la Tabla 6.3 (en conjunto y para cada variable por separado). Por lo tanto, si eliminamos la interacción ‘Educ*D_gndr_fem’, nos quedaremos únicamente con tres bloques: las variables simples, la interacción entre la educación con la categoría laboral (‘Educ*D_jc_Custodial’ y ‘Educ*D_jc_Manager’) y la interacción del género con la categoría laboral (‘D_gndr_fem*D_jc_Custodial’ más ‘D_gndr_fem*D_jc_Manager’), lo cual correspondería a la Ecuación 6.5:

$$\begin{aligned}\text{Salary} = {} & b_0 + b_1 \cdot \text{‘Educ’} + \left[b_2 \cdot \text{‘}D_{\text{gndr}_{\text{fem}}}\text{’}\right] + \left[b_3 \cdot \text{‘}D_{\text{jc}_{\text{Custodial}}}\text{’} + b_4 \cdot \text{‘}D_{\text{jc}_{\text{Manager}}}\text{’}\right] \\ & + \left[b_6 \cdot \text{‘Educ} * D_{\text{jc}_{\text{Custodial}}}\text{’} + b_7 \cdot \text{‘Educ} * D_{\text{jc}_{\text{Manager}}}\text{’}\right] \\ & + \left[b_8 \cdot \text{‘}D_{\text{gndr}_{\text{fem}}} * D_{\text{jc}_{\text{Custodial}}}\text{’} + b_9 \cdot \text{‘}D_{\text{gndr}_{\text{fem}}} * D_{\text{jc}_{\text{Manager}}}\text{’}\right] + e\end{aligned} \quad \textbf{(6. 5)}$$

Tabla 6.4.

*Resultados de la regresión por bloques con 'Educ*D_gndr' en último lugar*

(a)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,851[a]	,724	,721	$9,012.577	,724	307,231	4	469	,000
2	,859[b]	,738	,735	$8,796.663	,014	12,653	2	467	,000
3	,862[c]	,743	,739	$8,727.378	,005	8,444	1	466	,004
4	,862[d]	,743	,739	$8,727.704	,001	,965	1	465	,326

a. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years)

b. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ* D_jc_Custodial , Educ* D_jc_Manager

c. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ* D_jc_Custodial , Educ* D_jc_Manager , 'D_gndr_fem * D_jc_Manager'

d. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ* D_jc_Custodial , Educ* D_jc_Manager , 'D_gndr_fem * D_jc_Manager', Educ * D_gndr_fem

(b)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
4	(Constant)	9476,710	4480,532		2,115	,035
	Educational Level (years)	1605,737	321,858	,271	4,989	,000
	D_gndr_fem	1006,269	5618,617	,029	,179	,858
	D_jc_Custodial	22503,524	9197,564	,306	2,447	,015
	D_jc_Manager	-7539,151	11739,896	-,169	-,642	,521
	Educ* D_jc_Custodial	-1707,978	835,672	-,242	-2,044	,042
	Educ* D_jc_Manager	2085,978	698,769	,810	2,985	,003
	'D_gndr_fem * D_jc_Manager'	-8158,625	3447,222	-,069	-2,367	,018
	Educ * D_gndr_fem	-414,946	422,376	-,155	-,982	,326

a. Dependent Variable: Current Salary

Nótese que en la Ecuación 6.5 se han utilizado los subíndices de los coeficientes correspondientes a los de la Ecuación primitiva 6.4 (no se han utilizado en la Ecuación 6.5 los coeficientes b5, b10 ni b11, que han resultado ser los coeficientes de las variables no significativas estadísticamente).

Siendo la sintaxis para obtener este modelo la que sigue:

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA CHANGE
      /CRITERIA=PIN(.05) POUT(.10)
      /NOORIGIN
      /DEPENDENT salary
```

```
/METHOD=ENTER      educ      D_gndr_fem      D_JC_custodial
      D_JC_manager
/METHOD=ENTER Educ_X_D_jc_Custodial Educ_X_D_jc_Manager
/METHOD=ENTER                  D_gndr_fem_X_D_jc_Custodial
      D_gndr_fem_X_D_jc_Manager
/SAVE PRED.
```

IMPORTANTE
Guarda el fichero de datos como: 'Comp_dat_Dms_U5.sav', ya que hemos pedido que nos guarde los valores predichos de dicha ecuación como una nueva variable, que llamaremos 'Y_by_EducXjobcat_GndrXjobcat', y que también contiene las variables principales anidadas bajo estas interacciones: 'Educ', 'Jobcat' y 'Gender'.

Los resultados de conjunto de dicha ecuación de regresión aparecen en la Tabla 6.6.

Tabla 6.6.
Resultados de la regresión por bloques

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,851[a]	,724	,721	$9,012.577	,724	307,231	4	469	,000
2	,859[b]	,738	,735	$8,796.663	,014	12,653	2	467	,000
3	,862[c]	,743	,739	$8,727.378	,005	8,444	1	466	,004

a. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years)

b. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ* D_jc_Custodial , Educ* D_jc_Manager

c. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ* D_jc_Custodial , Educ* D_jc_Manager , 'D_gndr_fem * D_jc_Manager'

d. Dependent Variable: Current Salary

Podemos observar en la tabla 6.6 que cada bloque añade capacidad explicativa al modelo anterior, aunque en el tercer modelo ('Sexo*Categoría laboral') no hayan mujeres que trabajen de 'Custodial'. Para determinar mejor la significación de este tercer modelo podemos observar la Tabla 6.7a, donde nos indica que toda la ecuación final es significativa.

Tabla 6.7.

Resultados de la regresión por bloques

(a)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	99821247460,197	4	24955311865,049	307,231	,000[b]
	Residual	38095247976,143	469	81226541,527		
	Total	137916495436,34	473			
2	Regression	101779436368,71	6	16963239394,785	219,216	,000[c]
	Residual	36137059067,629	467	77381282,800		
	Total	137916495436,34	473			
3	Regression	102422617317,22	7	14631802473,889	192,101	,000[d]
	Residual	35493878119,117	466	76167120,427		
	Total	137916495436,34	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years)

c. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ* D_jc_Custodial , Educ* D_jc_Manager

d. Predictors: (Constant), D_jc_Manager, D_jc_Custodial, D_gndr_fem, Educational Level (years), Educ* D_jc_Custodial , Educ* D_jc_Manager , 'D_gndr_fem * D_jc_Manager'

(b)

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
3	(Constant)	12790,121	2949,437		4,336	,000
	Educational Level (years)	1364,790	208,413	,231	6,548	,000
	D_gndr_fem	-4428,916	979,919	-,129	-4,520	,000
	D_jc_Custodial	19190,114	8556,539	,261	2,243	,025
	D_jc_Manager	-10852,562	11244,570	-,243	-,965	,335
	Educ* D_jc_Custodial	-1467,031	798,841	-,208	-1,836	,067
	Educ* D_jc_Manager	2326,925	654,288	,903	3,556	,000
	'D_gndr_fem * D_jc_Manager'	-9362,583	3221,906	-,079	-2,906	,004

a. Dependent Variable: Current Salary

En la Tabla 6.7.b, podemos comprobar que la interacción 'Educ*D_jc_Custodial' es no significativa (p = .067), pero no se quita, porque forma un bloque con 'Educ* D_jc_Manager', y todo el bloque es significativo. Del mismo modo, no se suprime 'D_jc_Manager' porque toda la variable a la que pertenece ('Jobcat'), es significativa en interacción con 'Educ' y con 'Gender'.

Las interacciones entre Educación y Categoría laboral, y Sexo y Categoría Laboral son significativas, como se comprueba en la Tabla 6.7.b, lo que nos indicaría que:

Para ‘Educ*Cat_lab’: que la pendiente de las rectas de Educación con la variable dependiente cambian según la Categoría Laboral (o que al menos una de ellas difiere de las otras dos).

Para ‘Sexo*Cat_lab’: las ordenadas en el origen de las rectas de Sexo con la variable dependiente son diferentes según la Categoría Laboral, o que al menos una de ellas se diferencia de las otras.

Así pues, la ecuación general quedaría como sigue:

$$\begin{aligned}'\text{Salary}' = {} & 12790{,}121 + 1364{,}79 \cdot \text{‘Educ’} + \left[-4428{,}916 \cdot \text{‘D}_{\text{gndr}_{\text{fem}}}\text{’}\right] \\ & + \left[19190{,}114 \cdot \text{‘D}_{\text{JC}_{\text{Custorial}}}\text{’} - 10852{,}562 \cdot \text{‘D}_{\text{JC}_{\text{Manager}}}\text{’}\right] \\ & + \left[-1467{,}031 \cdot \text{‘Educ} * \text{D}_{\text{JC}_{\text{Custodial}}}\text{’} + 2326{,}925 \cdot \text{‘Educ} * \text{D}_{\text{JC}_{\text{Manager}}}\text{’}\right] \\ & + \left[-9362{,}583 \cdot \text{‘D}_{\text{gndr}_{\text{fem}}} * \text{D}_{\text{JC}_{\text{Manager}}}\text{’}\right] + \text{e}\end{aligned} \tag{6. 6}$$

De la que podríamos extraer seis ecuaciones, una para cada grupo:

Para Administrativo (D_jc_Custodial=0, D_jc_Manager=0) Hombre (D_gndr_fem=0):

$$\begin{aligned}\text{Salary}_{\text{Clerical,Male}'} & \\ & = 12790{,}121 + 1364{,}79 \cdot \text{‘Educ’} + [-4428{,}916 \cdot 0] \\ & + [19190{,}114 \cdot 0 - 10852{,}562 \cdot 0] \\ & + [-1467{,}031 \cdot \text{‘Educ} * 0\text{’} + 2326{,}925 \cdot \text{‘Educ} * 0\text{’}] + [-9362{,}583 \cdot \text{‘0} * 0\text{’}] \\ & = 12790{,}121 + 1364{,}79 \cdot \text{‘Educ’}\end{aligned} \tag{6. 7}$$

Para Administrativo (D_jc_Custodial=0, D_jc_Manager=0) Mujer (D_gndr_fem=1):

$$\begin{aligned}\text{Salary}_{\text{Clerical,Female}'} & \\ & = 12790{,}121 + 1364{,}79 \cdot \text{‘Educ’} + [-4428{,}916 \cdot 1] \\ & + [19190{,}114 \cdot 0 - 10852{,}562 \cdot 0] \\ & + [-1467{,}031 \cdot \text{‘Educ} * 0\text{’} + 2326{,}925 \cdot \text{‘Educ} * 0\text{’}] + [-362{,}583 \cdot \text{‘1} * 0\text{’}] \\ & = 8361{,}205 + 1364{,}79 \cdot \text{‘Educ’}\end{aligned} \tag{6. 8}$$

Para Custodial (D_jc_Custodial=1, D_jc_Manager=0) Male (D_gndr_fem=0): (6. 9)

$$\text{Salary}_{\text{Custodial,Male'}} = 12790{,}121 + 1364{,}79 \cdot \text{'Educ'} + [-4428{,}916 \cdot 0] + [19190{,}114 \cdot 1 - 10852{,}562 \cdot 0] + [-1467{,}031 \cdot \text{'Educ} * 1' + 2326{,}925 \cdot \text{'Educ} * 0'] + [-9362{,}583 \cdot \text{'0} * 0'] = 31980{,}559 - 102{,}241 \cdot \text{'Educ'}$$

Para Custodial (D_jc_Custodial=1, D_jc_Manager=0) Female (D_gndr_fem = 1):

$$\text{Salary}_{\text{Custodial,Female'}} = 12790{,}121 + 1364{,}79 \cdot \text{'Educ'} + [-4428{,}916 \cdot 1] + [19190{,}114 \cdot 1 - 10852{,}562 \cdot 0] + [-1467{,}031 \cdot \text{'Educ} * 1' + 2326{,}925 \cdot \text{'Educ} * 0'] + [-9362{,}583 \cdot \text{'1} * 0'] = 27551{,}319 - 102{,}241 \cdot \text{'Educ'} \quad \textbf{(6. 10)}$$

Para Manager (D_jc_Custodial=0, D_jc_Manager=1) Male (D_gndr_fem=0)

$$\text{Salary}_{\text{Manager,Male'}} = 12790{,}121 + 1364{,}79 \cdot \text{'Educ'} + [-4428{,}916 \cdot 0] + [19190{,}114 \cdot 0 - 10852{,}562 \cdot 1] + [-1467{,}031 \cdot \text{'Educ} * 0' + 2326{,}925 \cdot \text{'Educ} * 1'] + [-9362{,}583 \cdot \text{'0} * 1'] = 1937{,}559 + 3691{,}715 \cdot \text{'Educ'} \quad \textbf{(6. 11)}$$

Para Manager (D_jc_Custodial=0, D_jc_Manager=1) Female (D_gndr_fem=0)

$$\text{Salary}_{\text{Manager,Female'}} = 12790{,}121 + 1364{,}79 \cdot \text{'Educ'} + [-4428{,}916 \cdot 1] + [19190{,}114 \cdot 0 - 10852{,}562 \cdot 1] + [-1467{,}031 \cdot \text{'Educ} * 0' + 2326{,}925 \cdot \text{'Educ} * 1'] + [-9362{,}583 \cdot \text{'1} * 1' = -11853{,}94 + 3691{,}715 \cdot \text{'Educ'} \quad \textbf{(6. 12)}$$

En resumen, tenemos las siguientes ecuaciones:

$$\text{Salary}_{\text{Clerical,Male'}} = 12790{,}121 + 1364{,}79 \cdot \text{'Educ'}$$

$$\text{Salary}_{\text{Clerical,Female'}} = 8361{,}205 + 1364{,}79 \cdot \text{'Educ'}$$

$$\text{Salary}_{\text{Custodial,Male'}} = 31980{,}559 - 102{,}241 \cdot \text{'Educ'}$$

$$\text{Salary}_{\text{Custodial,Female'}} = 27551{,}319 - 102{,}241 \cdot \text{'Educ'}$$

$$\text{Salary}_{\text{Manager,Male'}} = 1937{,}559 + 3691{,}715 \cdot \text{'Educ'}$$

$$\text{Salary}_{\text{Manager,Female'}} = -11853{,}94 + 3691{,}715 \cdot \text{'Educ'} \quad \textbf{(6. 13)}$$

Donde podemos destacar a simple vista varias cuestiones. En primer lugar, podemos observar que las pendientes en cada ecuación, según las diferentes categorías laborales, son iguales para hombres y mujeres, por lo que se comprueba la no existencia de interacción entre nivel educativo y género (‘Educ*Gender’); sí que observamos la existencia de pendientes diferentes para cada categoría laboral, lo que confirma la existencia de interacción entre el nivel educativo y la categoría laboral (‘Educ*Jobcat’); y por último, observamos que no existe equidistancia entre las diferentes ordenadas en el origen, lo que nos permite corroborar la interacción entre el género y la categoría laboral (‘Gender*Jobcat’). Esto último podemos observarlo mejor si realizamos un gráfico de dichas ordenadas en el origen para cada categoría laboral diferenciando las líneas por sexo, tal y como podemos ver en la Figura 6.2.

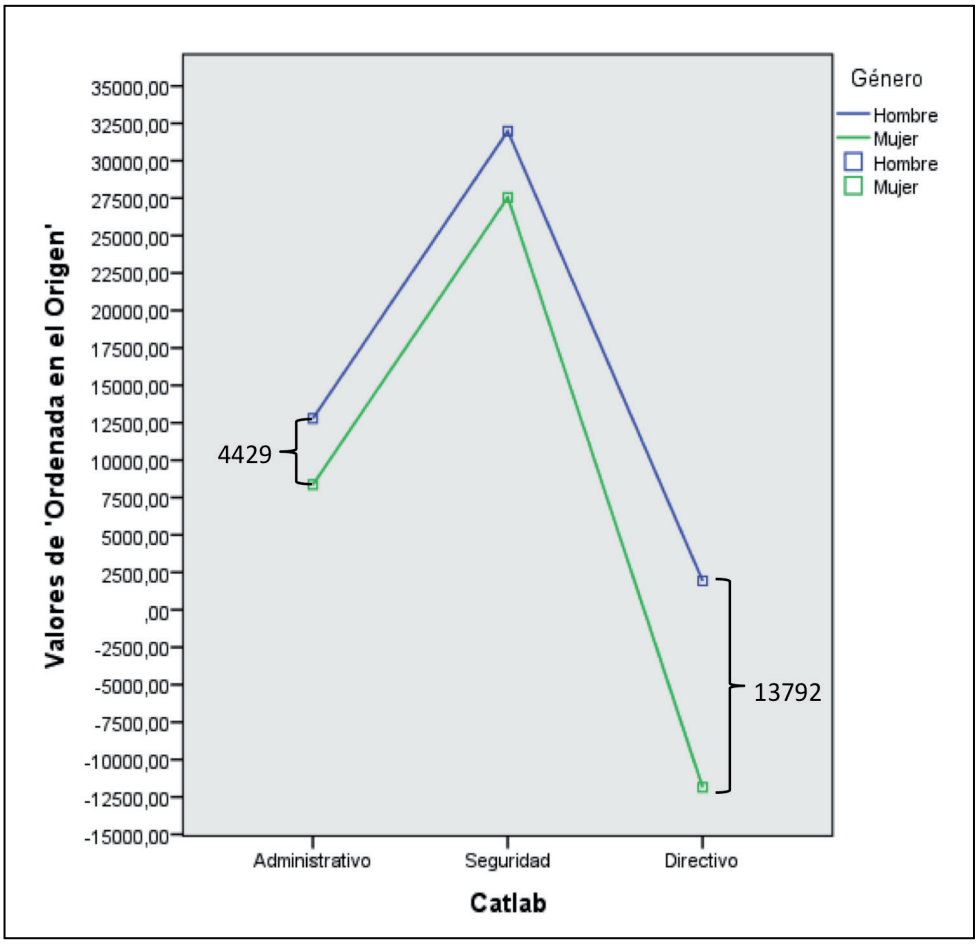


Figura 6.2. Ordenadas en el Origen para cada categoría laboral

Si existe interacción entre dos variables categóricas dentro de una ecuación de regresión, el significado es que son significativamente distintas las diferencias entre las categorías homólogas de las variables. Así, la interacción entre ‘Gender*Jobcat’ puede observarse en las diferencias de las ordenadas en el origen por sexo, que si bien en los ‘Clerical’ es de

4428.9 dólares (12790.1 – 8361.2), manteniéndose para los 'Custodial' (31980.5 – 27551.3 = 4429.2), para los 'Manager' es de 13791.5 dólares (1937.6 – (– 11853.9)); o dicho de otro modo, las mujeres Clerical y las mujeres Custodial cobran «de base» 4429 dólares menos que los hombres (para el valor 'cero' de 'Educ'), y las mujeres Manager cobran «de base» 13792 dólares menos que los hombres 'Manager' en la ordenada en el origen. Así pues, lo que cobran las diferentes categorías profesionales interacciona con el sexo de los sujetos en los respectivos valores de las intersecciones.

Si quisiéramos representar los valores pronosticados en la ecuación resultante para cada uno de los seis grupos existentes al evaluar la interacción 'Gender*Jobcat', esto no sería posible con la información que disponemos, ya que o bien representamos los valores por género (dos líneas) o por jobcat (tres líneas). Para poder realizar la representación conjunta de los seis grupos y representar sus diferentes líneas de pronóstico, debemos generar una nueva variable con las seis categorías de 'Gender*Jobcat', para lo cual podemos introducir en la ventana de sintaxis las siguientes instrucciones:

```
IF ( jobcat = 1 and  gender ='m') JOBCAT_GDR = 1.
IF ( jobcat = 2 and  gender ='m') JOBCAT_GDR = 2.
IF ( jobcat = 3 and  gender ='m') JOBCAT_GDR = 3.
IF ( jobcat = 1 and  gender ='f') JOBCAT_GDR = 4.
IF ( jobcat = 2 and  gender ='f') JOBCAT_GDR = 5.
IF ( jobcat = 3 and  gender ='f') JOBCAT_GDR = 6.
     EXECUTE.
```

Esta nueva variable llamada JOBCAT_GDR (de Jobcat y Gender) tendrá las seis categorías necesarias para representar las seis líneas de pronóstico para los diferentes grupos. No olvides incluir las etiquetas de valor para cada una de las categorías, para ello puedes utilizar el comando de sintaxis 'Value Label' según el siguiente esquema:

```
VALUE LABELS variable_name value 'label' value 'label'
     EXECUTE.
```

Puesto que tenemos guardados los valores pronosticados, y ahora tenemos una nueva variable que nos diferencia a los sujetos del estudio en los seis posibles grupos resultantes de combinar Jobcat y Gender, podemos realizar la nube de puntos de dichos valores pronosticados ('Y_by_EducXjobcat_gndrXjobcat') estableciendo marcas por esta nueva variable con la sintaxis:

```
GRAPH
     /SCATTERPLOT(BIVAR)=educ                                    WITH
     Y_by_EducXjobcat_gndrXjobcat BY JOBCAT_GDR
     /MISSING=LISTWISE.
     EXECUTE.
```

Obteniendo la Figura 6.3, a la que hemos añadido las líneas de ajuste en subgrupos. Cabe destacar el punto de las diez mujeres directivas, con 16 años de 'Educ', de cuyo grupo obtenemos una línea paralela al eje de abscisas (o pendiente igual a 'cero'). Además, también podemos observar que no aparece ninguna línea para el grupo Custodial-Female, ya que no existe ningún caso en esta combinación (como ya hemos visto anteriormente). Por último, los Clerical-Male y las Clerical-Female tienen líneas paralelas (las dos líneas inferiores del gráfico).

Para hacer correctamente el gráfico podemos, en primer lugar, eliminar la línea del grupo 'Manager-Female', paralela al eje de abscisas, ya que esa línea generada por el SPSS no aporta ninguna información del grupo (recordar que, según la Tabla 6.1, las diez 'Female' tienen 16 años de 'Educ', por lo que se genera un solo punto de pronóstico, y el generador de gráficos del SPSS estima una línea horizontal, con valor de pendiente igual a 'cero', pero la ecuación estimada para este grupo 'Manager-Female' es de ($Salary_{Manager,Female}$' = –11853,94 + 3691,715·'Educ'), y en segundo lugar añadir las dos líneas de regresión que nos faltarán ('Custodial-Female' y 'Manager-Female'). Para ello, desde el editor de gráficos del SPSS, seleccionamos el menú 'Options', y elegimos la opción 'Reference Line from Equation', tal y como podemos ver en la Figura 6.4(a), añadiendo las ecuaciones correspondientes a las líneas que faltan ('Custodial-Female', Ecuación 6.10, y 'Manager-Female', Ecuación 6.12). Lo que nos generaría el gráfico definitivo de interacciones entre 'Jobcat' y 'Educ' a la hora de predecir el 'Salary' (Figura 6.4(b)).

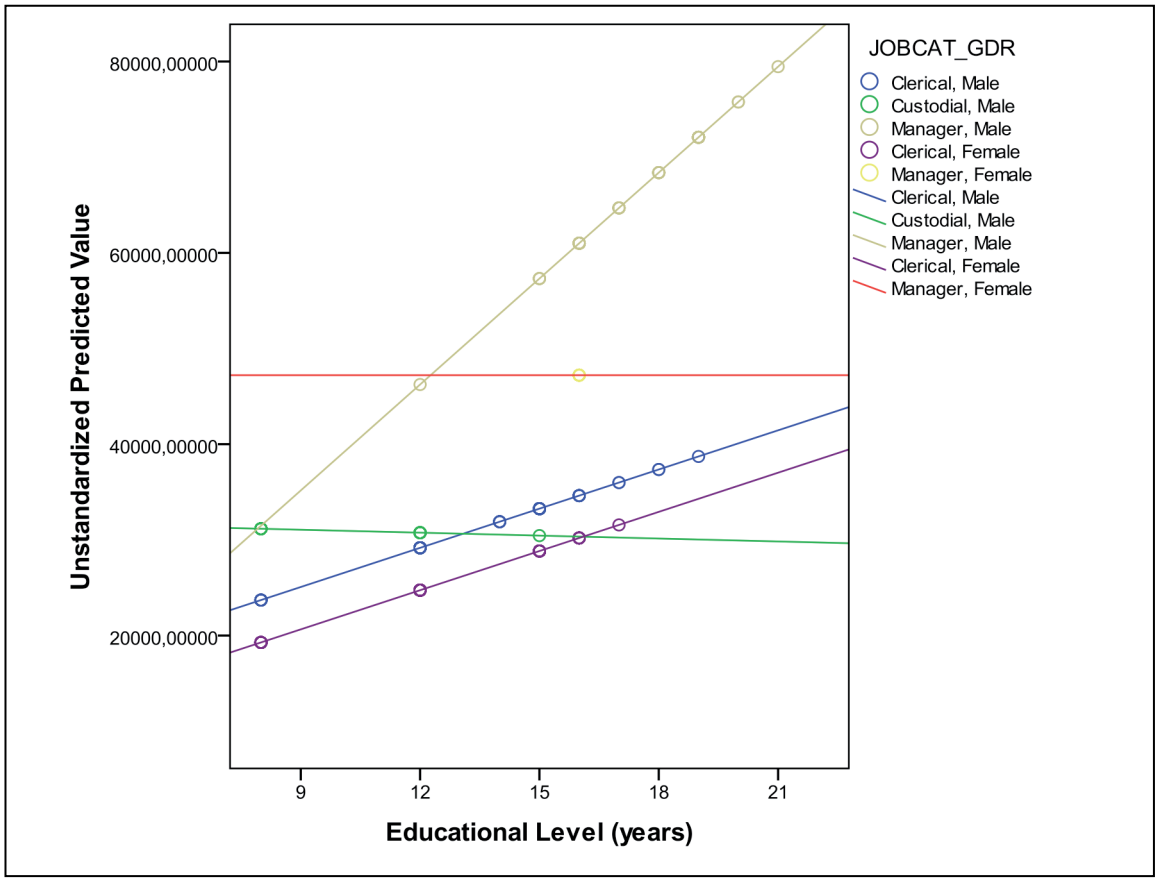
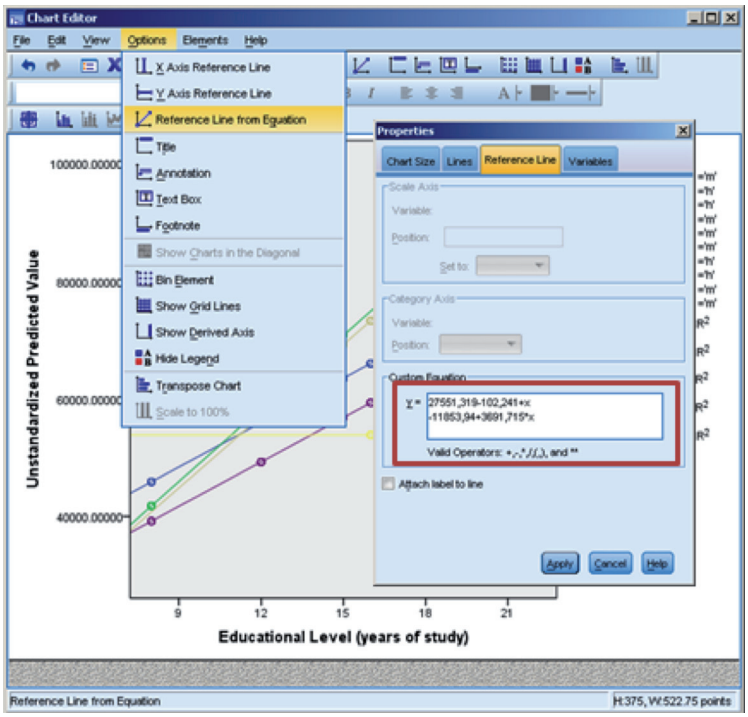


Figura 6.3. Valores esperados para la variable 'Salary' en función de 'Educational level' (years) y 'Jobcat_gdr', conforme a la Ecuación 6.6

Por lo tanto, podemos resumir que hemos obtenido significativo el modelo 'Salario' = f('Edu*Catlab', 'Sexo*Catlab'), que se corresponde con la Figura 6.4. En ésta podemos apreciar que no hay interacción global entre 'Educ*Gender*Jobcat' porque observamos que las líneas son paralelas según el género. Además, podemos decir que no hay interacción 'Educ*Gender' ya que las pendientes de 'Gender' no cambian dentro de la/s otra/s categoría/s de 'Jobcat'. Eso no pasa en la interacción 'Educ*Jobcat', porque como podemos observar, las pendientes cambian según 'Jobcat'. En cuanto a la interacción 'Gender*Jobcat', tal y como podemos ver en la Figura 6.2, las ordenadas en el origen no son equidistantes para el 'Gender' según la 'Jobcat', por lo que existiría interacción.

O dicho de otro modo: teniendo representados cada uno de los seis posibles grupos, una persona que entrara a trabajar en esta empresa tendría su propio pronóstico en 'Salary' en función del grupo al que pertenezca y a su correspondiente ecuación de regresión, estimada en la Ecuación 6.5, que a su vez se desglosa en las Ecuaciones 6.6 a 6.12.

(a)



(b)

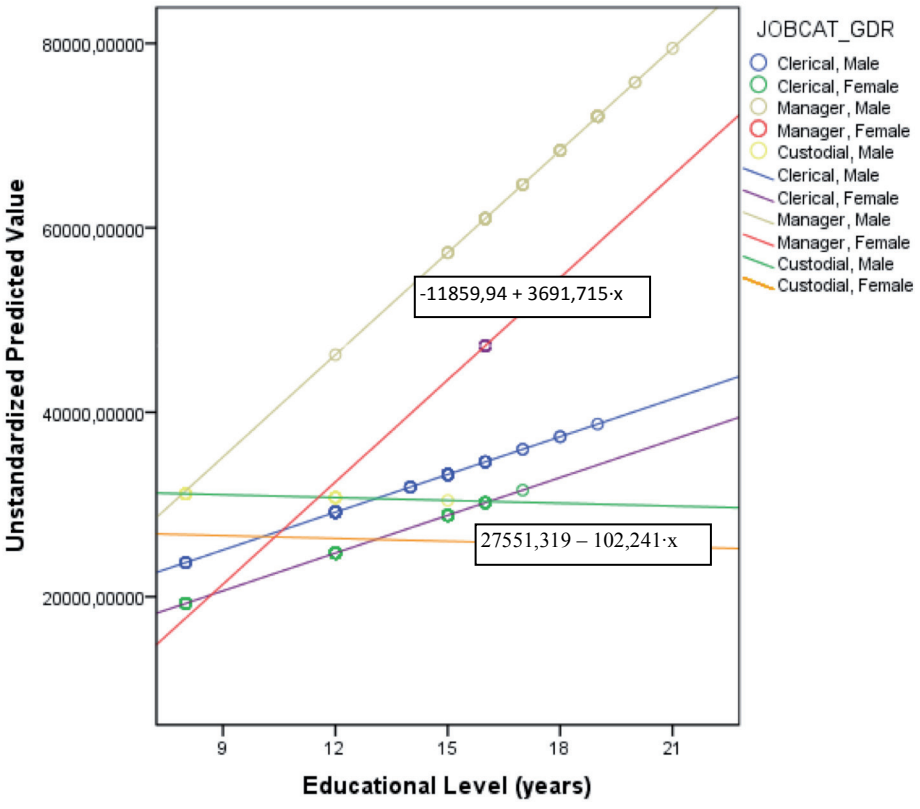


Figura 6.4. Inserción de rectas en el gráfico de dispersión a partir de sus respectivas ecuaciones en el editor de gráficos

6.4. Conclusiones

Cuando trabajes con la interacción de variable independientes categóricas, piensa en que seguramente hay diferencias en las respectivas ordenadas en el origen (p. ej.: 'Sexo*Catlab').

Cuando en la interacción aparezca una o alguna variable independiente continua con otra variable de grupos, piensa que las pendientes (de la/s recta/s o del plano o planos) están afectadas según los niveles de la otra variable, o de las otras variables (p. ej.: 'edu*catlab'), es decir, hay al menos una pendiente que significativamente diferente de las demás.

Siempre que sea posible haz una figura que apoye las explicaciones del modelo, relacionando figura y texto. Por ejemplo:

a. No hay interacción 'Gender*Educ' (ver Figura 6.5), ya que la posición relativa de las pendientes en ambos grupos (hombres y mujeres) no cambia (o no lo hace apenas).
b. Hay interacción, ya que la posición relativa de las pendientes en ambos grupos (hombres y mujeres) cambia y/o las ordenadas en el origen también cambian (Figura 6.6).

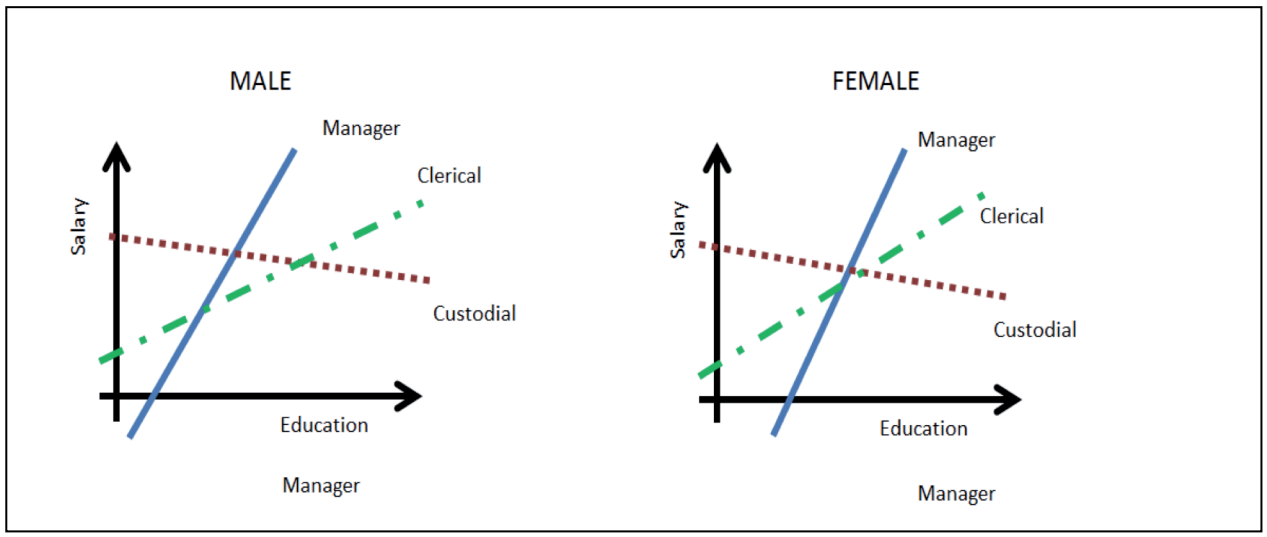


Figura 6.5. Ejemplo de **no** existencia de interacción

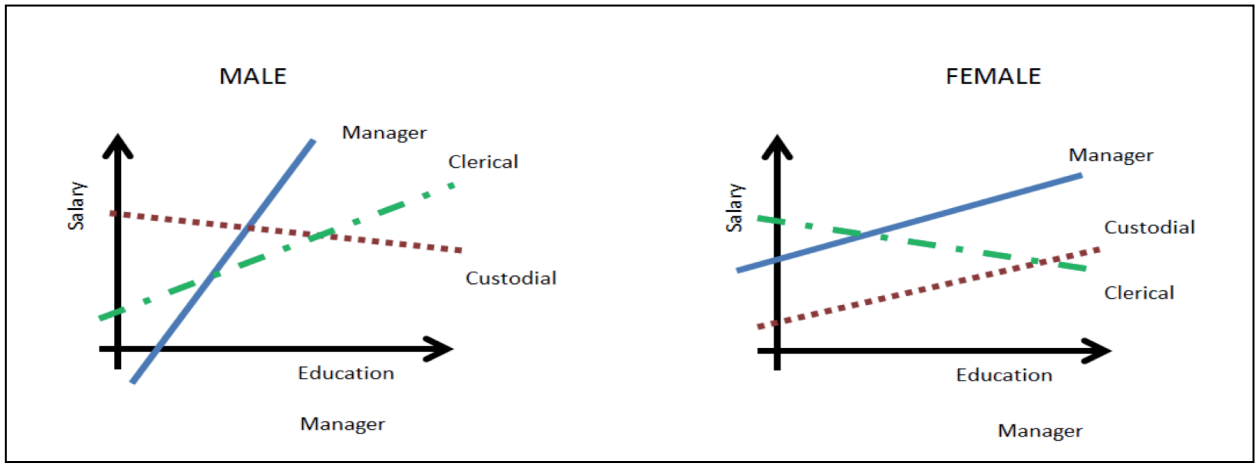


Figura 6.6. Ejemplo de existencia de interacción

En cualquier caso, cada conjunto de datos es diferente y se ha de explicar e interpretar adecuadamente cada interacción significativa, comparándola con el caso de que no hubiese interacción.

Lecturas recomendadas

Esta unidad es un desarrollo de las unidades anteriores, y se han de tener claros los conceptos que hemos desarrollado hasta ahora para poder seguir profundizando en la interacción de variables. De todos modos, para poder afianzar los conocimientos adquiridos, se recomienda la lectura del libro de Aiken y West (1991), en su capítulo 7 («Interactions between categorical and continuous variables») y Jaccard y Turrisi (2003), en sus capítulos 3 y 4. Pero ninguno de estos dos manuales, y ningún otro sobre interacción de variables, expone el caso que nos ocupa: el de la interacción entre una variable independiente continua con otras dos variables categóricas. Además, ningún libro especifica cómo llevar a cabo el ajuste del modelo más adecuado en función de la hipótesis global (interacción completa de las variables), ni especifica cómo representar cada correspondiente ecuación de pronóstico para cada grupo. Esperamos que con estos apuntes el lector comprenda y sepa cómo hacer el ajuste del modelo a los datos, su representación e interpretación.

Bibliografía

Aiken, L. S.; West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Jaccard, J.; Turrisi, R. (2003). *Interaction effects in multiple regression*. (2.ª ed.). Thousand Oaks, CA: Sage.

Actividades

Se desea investigar la relación de Salini = f('edu*sex*catlab').

Cuestiones:

a. Ajusta el mejor modelo explicativo (en el que todo sea estadísticamente significativo).

b. Escribe la ecuación de conjunto encontrada en la cuestión 'a)'.

c. Escribe la ecuación para cada grupo.

e. Haz una representación de los valores esperados para cada grupo.

f. Interpreta los resultados obtenidos, poniéndolos en relación con el gráfico que has hecho.

Unidad 7. Regresión lineal con interacción dos variables independientes continuas con otra de grupos

Introducción

El objetivo del estudio mediante regresión lineal, en esta unidad, será inicialmente la predicción de una variable dependiente en función de dos variables independientes continuas, de modo que el comportamiento de una variable independiente es distinto según el nivel de la otra variable independiente. En la segunda parte de esta unidad se incluirá la interacción de una variable categórica con otras dos continuas.

En ciencias de la salud, la interacción de variables se usa con más frecuencia que en ciencias sociales, y se suele poner como ejemplo de interacción de dos variables independientes continuas el siguiente: si una persona fuma una cantidad «x» de cigarrillos tendrá una determinada probabilidad de padecer cáncer, si otra persona bebe un determinado número «y» de centímetros cúbicos de alcohol, tendrá otra probabilidad, pero si una tercera persona fuma «x» cigarrillos y bebe «y» centímetros cúbicos de alcohol, tendrá una probabilidad de padecer cáncer mayor que la simple suma de las dos probabilidades. Incluyendo a este supuesto el género de los sujetos, tendremos también un ejemplo para la segunda parte de la unidad, donde realizaremos la regresión lineal de una variable dependiente sobre tres variables independientes, dos de las cuales serán variables continuas 'Educación' y 'Experiencia Previa en meses' y la otra será categórica 'Gender'.

En estos casos, una vez establecida la hipótesis de la interacción, se ha de llevar a cabo el cálculo de las variables de interacción; para ello, inicialmente realizamos el producto entre las variables independientes que interactúan. En esta unidad nos ocuparemos de cómo realizar las interacciones, cómo estimar los parámetros de la regresión, cómo formular la ecuación de regresión y cómo interpretar dicha interacción.

Objetivos

Cuando el alumno finalice la unidad sabrá:

- Una vez planteada la hipótesis sobre la interacción a partir de dos variables principales continuas, llevar a cabo la creación de las variables de interacción entre las variables principales.
- Incluir la variable independiente de interacción en la correspondiente ecuación de regresión.
- Estimar la ecuación de regresión con interacción entre dos variables independientes continuas.
- Representar gráficamente el diagrama de dispersión y las rectas de interacción entre dos variables independientes continuas.
- Interpretar la interacción entre las variables independientes continuas.
- Obtener la ecuación de pronóstico, y su valoración, de una variable dependiente y dos independientes continuas.
- Lo mismo para el caso de interacción entre dos variables continuas y otra de dos grupos.

7.1. Regresión con dos variables independientes continuas

En la unidad anterior tratamos la conveniencia de incluir en el modelo de regresión múltiple el término de interacción, como variable explicativa, desde la consideración de que a partir de tres variables independientes, una variable continua y otras dos politómicas, establecían grupos cuyos comportamientos en cuanto a la variable explicada configuraban sistemas diferentes, en ordenada en origen, en pendiente y/o en ambas.

Supongamos ahora que pretendemos ajustar una ecuación de regresión múltiple en la que queremos pronósticar la variable 'Salary' en función de la variable: 'Educ' con etiqueta 'Educational Level' (years of study), y de la variable 'Prevexp' con la etiqueta 'Previous Experience' (measured in months) y la interacción 'EducXPrevexp'; o lo que es lo mismo, en forma compacta: 'Salary' = f('Educ'*'Prevexp'); para ello partiremos desde el fichero de datos de la unidad anterior: 'Comp_dat_Dms_U6.sav'.

La Ecuación 7.1 contempla el objetivo que planteamos en esta unidad, de modo que el alumno comprenda el sentido de cada término de la expresión:

$$Salary = b_0 + b_1 \cdot Educ + b_2 \cdot Prevexp + b_3 \cdot Educ \cdot Prevexp + e \qquad (7.1)$$

Antes de realizar el ajuste de la ecuación hemos de obtener la nueva variable de interacción, simplemente desde la sintaxis:

```
COMPUTE EducXPrevex=educ * prevexp.
     EXECUTE.
```

IMPORTANTE
Guarda el fichero de datos como: 'Comp_dat_Dms_U7.sav', ya que hemos pedido que se genere la nueva variable EducXPrevex que utilizaremos a lo largo de esta unidad.

La Figura 7.1 contiene los valores obtenidos para la nueva variable que se ha generado tras correr la sintaxis COMPUTE.

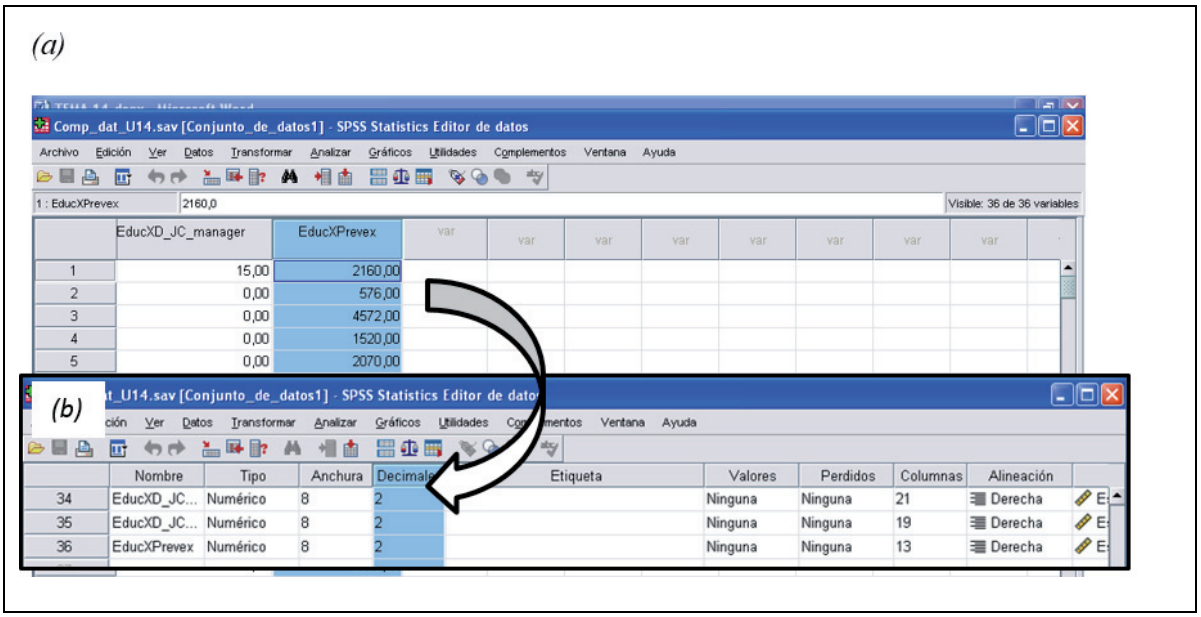


Figura 7.1. Variables 'EducXPrevexp', a) en modo 'Data view' y b) en modo 'Variable view', como aparecen en la matriz de datos del SPSS

Partiendo del modelo que se propone en la Ecuación 7.1, podemos estimar los parámetros de la ecuación mediante la sintaxis:

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA CHANGE
      /CRITERIA=PIN(.05) POUT(.10)
      /NOORIGIN
      /DEPENDENT salary
      /METHOD=ENTER educ prevexp EducXPrevex.
```

Nótese que en este párrafo de sintaxis no hemos hecho el análisis en dos bloques porque éstos solo se usan para grupos de variables que nos interesa comprobar su significación, como una sola variable, caso del bloque con las variables *dummy* que representan a una sola variable primitiva del sistema. En este caso es la misma variable y, por tanto, solo se utiliza el epígrafe /METHOD=ENTER. En la Tabla 7.1 se muestran los resultados obtenidos.

Tabla 7.1.
*Significación del modelo: Salary = f('Educ*Prevexp')*

(a)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,671[a]	,451	,447	$12,118.094	,451	128,592	3	470	,000

a. Predictors: (Constant), EducXPrevex, Educational Level (years of study), Previous Experience (months)

(b)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56650604686	3	18883534895	128,592	,000[b]
	Residual	69018660093	470	146848213,0		
	Total	1,257E+11	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), EducXPrevex, Educational Level (years of study), Previous Experience (months)

(c)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2147,057	3867,547		-,555	,579
	Educational Level (years of study)	4406,249	280,086	,780	15,732	,000
	Previous Experience (months)	73,137	21,985	,469	3,327	,001
	EducXPrevex	-5,200	1,776	-,400	-2,929	,004

a. Dependent Variable: Current Salary

En el apartado (a) podemos observar que el modelo lineal que incluye la interacción explica el 45.1% de la varianza de la variable 'Salary'. En (b) queda patente el valor significativo de ajuste global del modelo, el valor p-value del estadístico F, para los parámetros considerados, es significativo. En (c) mostramos que los coeficientes estimados b_1, b_2 y b_3, son significativos. Recuérdese que, tal y como hemos señalado ya en otras unidades, si únicamente fuera significativo el coeficiente de la interacción habría de dejarse igualmente los coeficientes de las variables simples implicadas. El coeficiente de 'Educ' es de 4406.249, (t = 15.732, p = 0.000), lo que indica que posee un efecto positivo sobre el 'Salary'; es decir, a mayor nivel de años de estudio, mayor salario cobra también actualmente, aproximadamente unos 4406 dólares por cada

incremento de un año en sus estudios. Asimismo, el coeficiente de 'Prevexp' es de 73.137 (t = 3.327, p = 0.001), que produce un efecto directo en el 'Salary', de tal modo que un nivel más elevado en 'Prevexp', hace también que se cobre mayor salario. Por último, el coeficiente 'Educ*Prevex' es de -5.2 (t = -2.929, p = 0.004), lo cual estaría indicando que la interacción actúa de forma inversa entre las variables que forman la interacción, en nuestro caso concreto, las personas con alta educación les favorece más en su salario el tener baja experiencia previa, mientras que para baja educación, cuanta más experiencia previa, mejor salario.

El valor de la ordenada en el origen es de -2147.057 (p = .579) que ha de dejarse en la ecuación, aunque es estadísticamente no significativo, con el fin de no cometer sesgos de pronóstico.

El conjunto de los apartados que acabamos de comentar, nos llevan a plantear, redondeando a las milésimas, la Ecuación de pronóstico 7.2:

$$\text{Salary}' = -2147.057 + 4406.249 \cdot \text{Educ} + 73.137 \cdot \text{Prevexp} + [-5.200 \cdot \text{Educ} \cdot \text{Prevex}] \quad \textbf{(7. 2)}$$

Teniendo en cuenta todo lo anterior, es claro que se plantea la conveniencia del modelo de regresión lineal donde se incluya el término de interacción entre las variables independientes como estrategia más adecuada para entender el efecto de dichas variables en el pronóstico del 'Salary'. A partir de la Ecuación 7.2 podemos generar diferentes ecuaciones de regresión entre los niveles cruzados de 'Educ' y 'Prevexp'. Supongamos que queremos pronosticar el 'Salary' en función de 'Educ' (dando a 'Prevexp' un valor fijo), donde la Ecuación 7.2 daría lugar a la Ecuación:

$$\text{Salary}' = (-2147.057 + 73.137 \cdot \text{Prevexp}) + (4406.249 - 5.200 \cdot \text{Prevexp}) \cdot \text{Educ} \quad \textbf{(7. 3)}$$

O lo que es lo mismo, al tener 'Prevexp' un valor fijo, el valor de la ordenada en el origen será (-2147.057 + 73.137·Prevexp), y el de la pendiente + (4406.249 – 5.200·Prevexp), lo que nos indica que la pendiente de 'Educ' en la Ecuación 7.3 depende del valor que tome 'Prevexp'.

Si se desea pronosticar 'Salary' en función de 'Prevexp' (dejando a 'Educ' con un valor fijo), daría lugar a la Ecuación (7.4).

$$\text{Salary}' = (-2147.057 + 4406.249 \cdot \text{Educ}) + (73.137 - 5.200 \cdot \text{Educ}) \cdot \text{Prevexp} \quad (7.4)$$

Desde las Ecuaciones 7.3 y 7.4 solo tendríamos que darle valores a 'Prevexp' o a 'Educ', y obtendremos valores pronosticados de Salary en función de una u otra variable. Si bien en estos casos lo que se suele hacer es tomar como variable independiente la más 'causante' de las dos variables independientes de la interacción (en nuestro caso 'Educ'), y de la otra, 'Prevexp', se tomarían tres valores que serían: (a) la media de 'Prevexp' más 1,96 desviaciones típicas de la misma variable, que nos daría un nivel alto de 'Prevexp', (b) la media que nos indicaría un nivel medio de 'Prevexp' y (c) la media -1,96 desviaciones típicas de 'Prevexp' nos indicaría un nivel bajo.

Desde esta perspectiva calculamos la media y la desviación típica de 'Prevexp' desde la siguiente sintaxis:

```
DESCRIPTIVES VARIABLES=prevexp
     /STATISTICS=MEAN STDDEV.
```

Desde donde obtenemos la Tabla 7.2.

Tabla 7.2.
Estadísticos descriptivos de la variable 'Prevexp'

Descriptive Statistics

	N	Mean	Std. Deviation
Previous Experience (months)	474	95,86	104,586
Valid N (listwise)	474		

Sabiendo que la media de 'Prevexp' es 95.9 y su desviación típica 104.6, ya podemos establecer los tres niveles de 'Salary' desde la Ecuación 7.3.

Para el nivel alto de 'Prevexp':

$$\text{Salary}' = (-2147.057 + 73.137 \cdot (95.9 + 1.96 \cdot 104.6)) + (4406.249 + (-5.2) \cdot (95.9 + 1.96 \cdot 104.6)) \cdot \text{Educ} = 19861.036 + 2841.486 \cdot \text{Educ} \quad (7.5)$$

Para el nivel medio de 'Prevexp':

$$\text{Salary}' = (-2147.057 + 73.137 \cdot (95.9)) + (4406.249 + (-5.2) \cdot (95.9)) \cdot \text{Educ} = 4866.781 + 3907.569 \cdot \text{Educ} \quad (7.6)$$

Para el nivel bajo de 'Prevexp':

$$\text{Salary}' = (-2147.057 + 73.137 \cdot (95.9 - 1.96 \cdot 104.6)) + (4406.249 + (-5.2) \cdot (95.9 - 1.96 \cdot 104.6)) \cdot \text{Educ} = -10127.474 + 4973.652 \cdot \text{Educ} \quad (7.7)$$

Desde las tres Ecuaciones de regresión 7.5 a 7.7 podemos calcular los tres pronósticos del salario a las que podemos llamar Yp_alto. Yp_medio e Yp_bajo mediante la sintaxis:

```
COMPUTE Yp_prevexp_high=(-2147.057 +73.137*(95.9 +
1.96*104.6)) + (4406.249 + (-5.2)*(95.9 +
1.96*104.6))*Educ.
     VARIABLE LABELS  Yp_prevexp_high 'COMPUTE
          Yp_prevexp_high=(-2147.057 + 73.137*(95.9 +
          1.96*104.6)) + (4406.249 + (-5.2)*(95.9 +
          1.96*104.6))*Educ'.
     EXECUTE.
COMPUTE Yp_prevexp_mean=(-2147.057 +73.137*(95.9)) +
(4406.249 + (-5.2)*(95.9))*Educ.
     VARIABLE LABELS  Yp_prevexp_mean 'COMPUTE
          Yp_prevexp_mean= (-2147.057 +73.137*(95.9)) +
          (4406.249 + (-5.2)*(95.9))*Educ'.
     EXECUTE.
COMPUTE Yp_prevexp_low=(-2147.057 +73.137*(95.9 -
1.96*104.6)) + (4406.249 + (-5.2)*(95.9 -
1.96*104.6))*Educ.
     VARIABLE LABELS  Yp_prevexp_low 'COMPUTE
          Yp_prevexp_low= (-2147.057 +73.137*(95.9 -
          1.96*104.6)) + (4406.249 + (-5.2)*(95.9 -
          1.96*104.6))*Educ'.
     EXECUTE.
```

Compruébese que las tres nuevas variables aparecen en la matriz de datos. Con los valores pronosticados realizaremos el diagrama de dispersión y las correspondientes rectas de pronóstico de la regresión mediante la sintaxis:

```
GRAPH
      /SCATTERPLOT(OVERLAY)=educ educ educ educ WITH
            salary Yp_prevexp_high Yp_prevexp_mean
            Yp_prevexp_low (PAIR)
      /MISSING=LISTWISE.
```

Por medio de la cual se obtiene la Figura 7.2, que es el diagrama de dispersión de conjunto y las rectas de regresión (de 'Salary' en función de 'EducXPrevex') para cada uno de los tres grupos.

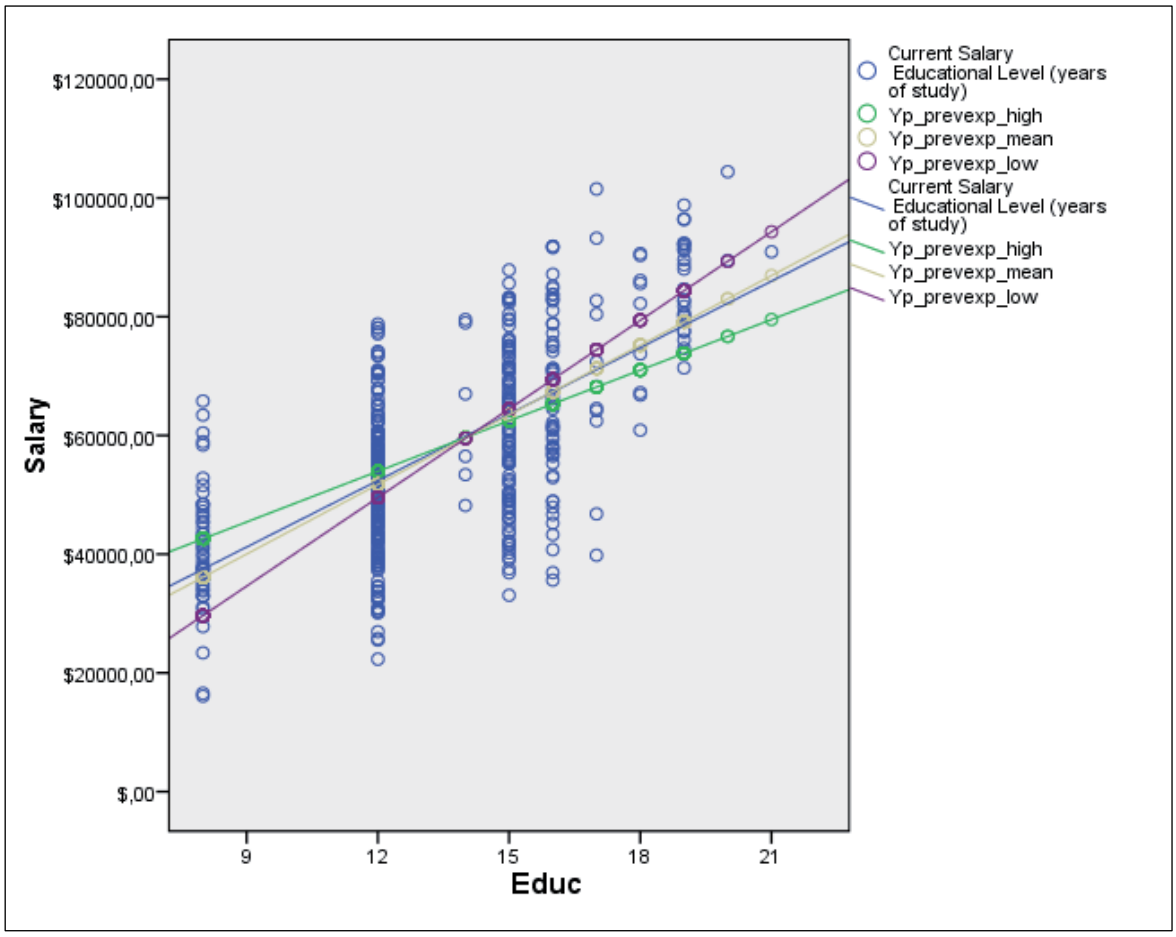


Figura 7.2. Diagrama de dispersión y rectas de regresión de 'Salary' = f('Educ') para cada uno de los tres niveles de 'Prevexp'

Lo más destacado en la Figura 7.2 son dos aspectos: (a) el valor pronosticado de la variable salario depende del nivel de 'Prevexp', y (b) se comprueba lo comentado anteriormente, que para un valor elevado de 'Educ', es mejor tener una 'Prevexp' baja con el fin de recibir un salario elevado, mientras que para valores bajos de 'Educ' se da lo contrario, es mejor alta 'Prevexp' con el fin de cobrar más salario.

7.2. Regresión con tres variables independientes, dos continuas y una categórica

Supongamos que en esta ocasión nos interesa comprobar cuál es la interacción entre las variable independientes 'Educ', 'Prevexp' y 'Gender' sobre la variable dependiente 'Salary', que en forma funcional compacta sería:

$$\text{Salary} = \text{f('Educ} * \text{Prevexp} * \text{Gender')} \quad (7.8)$$

Como podemos observar en la matriz de datos de trabajo, y ya sabemos de temas previos, la variable 'Gender' tiene dos categorías que hemos recodificado en una variable *dummy* ('D_gndr_Fem') donde el valor 0 corresponde a 'Male' y el 1 a 'Female'. En cuanto a las variables continuas, la variable 'Educ' representa los años de educación recibidos por los sujetos del estudio, con un recorrido entre 8 y 21 años de estudio; la variable 'Prevexp' representa la experiencia previa en meses de los mismos sujetos, y su recorrido va de 0 meses de experiencia previa a 476 meses. Asimismo, incluiremos la interacción entre 'Educ' y 'Prevexp' ya realizada en el primer apartado de esta unidad, pudiendo plantear la hipótesis completa de forma funcional:

$$\begin{aligned} &Salary = f('Educ', 'PrevExp', 'Gender'. && \text{Vars. Principales} \\ &['Educ * PrevExp'], ['Educ * Gender'], ['PrevExp * Gender'], && \text{Interacción de 2 Vis} \\ &['Educ * PrevExp * Gender']) && \text{Interacción de 3 Vis} \end{aligned} \quad (7.9)$$

Siendo la ecuación de regresión:

$$\begin{aligned} Salary = b_0 + b_1 \cdot Educ + b_2 \cdot Prevexp + b_3 \cdot Gender + [b_4 \cdot Educ * Prevexp] \\ + [b_5 \cdot Educ * D_{gndr_{Fem}}] + [b_6 \cdot Prevexp * D_{gndr_{Fem}}] \\ + [b_7\ Educ * Prevexp * D_{gndr_{Fem}}] + e \end{aligned} \quad (7.10)$$

De anteriores unidades, y del modelo estudiado anteriormente, ya disponemos de todas las variables implicadas en este modelo, a excepción de *Prevexp*D_gndr_Fem* y de *Educ*Prevexp*D_gndr_Fem*, por lo que tendremos que calcularlas mediante la sintaxis:

```
COMPUTE PrevexXD_gndr_Fem = Prevexp * D_gndr_Fem.
COMPUTE EducXPrevexXD_gndr_Fem = Educ*Prevexp*D_gndr_Fem.
EXECUTE.
```

Las variables obtenidas pueden observarse en la Figura 7.3.

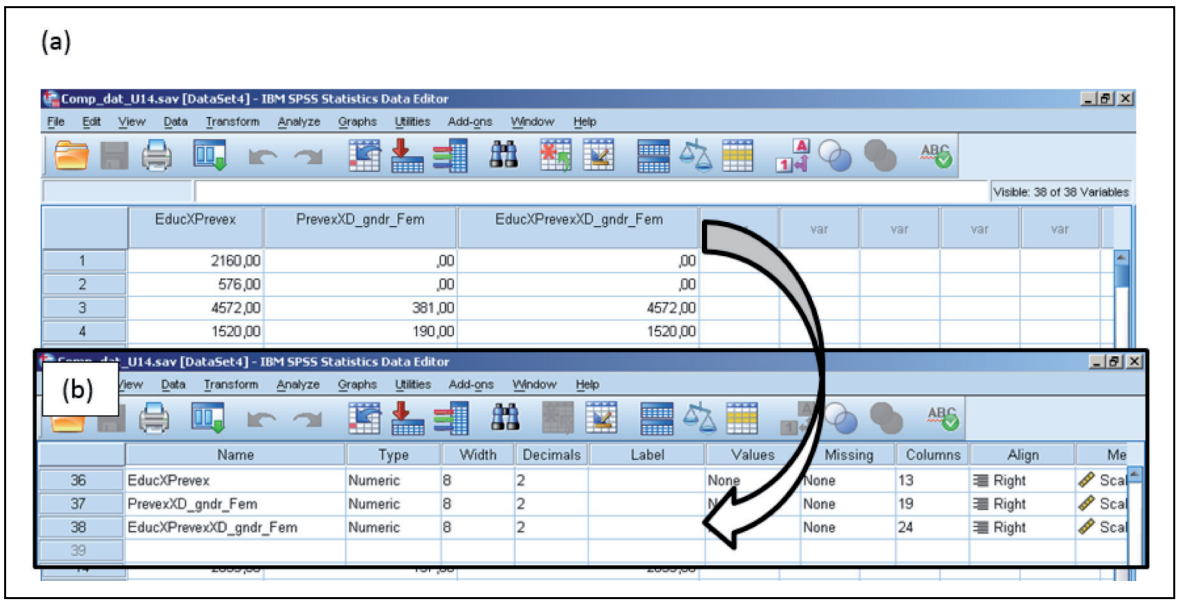


Figura 7.3. Variables *'EducXPrevex'*, '*Prevexp*D_gndr_Fem'* y de '*Educ*Prevexp *D_gndr_Fem'* en modo a) 'Data view' y b) 'Variable view', como aparecen en la matriz de datos

Partiendo del modelo lineal que se propone en la Ecuación 7.9, podemos estimar los parámetros de la ecuación mediante la sintaxis:

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA
      /CRITERIA=PIN(.05) POUT(.10)
      /NOORIGIN
      /DEPENDENT salary
      /METHOD=ENTER educ prevexp D_gndr_fem
      /METHOD=ENTER EducXPrevex
      /METHOD=ENTER EducXD_gndr_Fem
      /METHOD=ENTER PrevexXD_gndr_Fem
      /METHOD=ENTER EducXPrevexXD_gndr_Fem.
```

Obsérvese que, a la hora de proceder con el ajuste del modelo y tal y como vimos en la unidad anterior, con el fin de poner a prueba las interacciones, en primer lugar debemos empezar realizando la regresión lineal por bloques con el método de análisis 'Introducir' (/METHOD=ENTER), incluyendo el modelo completo, donde en el último bloque pondremos a prueba la interacción más alta y en los bloques previos habremos introducido todas sus correspondientes interacciones de menor orden anidadas, hasta llegar hasta las variables principales, comprobando al final la

significación de la interacción de mayor orden; si es significativa la interacción más alta y también lo es el ajuste en conjunto del modelo, hemos finalizado el ajuste del modelo ya que todo ajusta, por lo que deberíamos realizar la ecuación global, la ecuación para cada grupo y tratar de realizar el gráfico correspondiente, así como su correcta interpretación. Si esta interacción más alta no ajustara, deberíamos probar el modelo con las interacciones anidadas de nivel más alto dentro de la ecuación. Si alguna de estas no fuera significativa, deberíamos de ir eliminándolas una a una, suprimiendo en primer lugar aquéllas que tengan el valor de *p* más grande, ya que alguna interacción puede convertirse en significativa si eliminamos su colinealidad con otras.

En la Tabla 7.3, de resumen del modelo resultante, el SPSS nos da la significación de mejora para el modelo de bloque en bloque para los cinco bloques analizados, observándose cómo se va modificando el estadístico R^2 conforme se van introduciendo bloques en el modelo.

Tabla 7.3.
*Resultados de la regresión por bloques 'Salary' = f('Educ*PrevExp*Gender')*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,700[a]	,490	,487	$11,677.373
2	,704[b]	,495	,491	$11,633.504
3	,732[c]	,536	,531	$11,165.669
4	,742[d]	,551	,545	$10,989.662
5	,742[e]	,551	,545	$11,000.212

a. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study)

b. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex

c. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex, EducXD_gndr_fem

d. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex, EducXD_gndr_fem, PrevexXD_gndr_Fem

e. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex, EducXD_gndr_fem, PrevexXD_gndr_Fem, EducXPrevexXD_gndr_Fem

Si observamos además la Tabla 7.4 de coeficientes, podemos determinar que para los diferentes modelos existen interacciones que no resultan significativas, concretamente, si atendemos al último bloque analizado donde se incluía la interacción de mayor nivel, podemos observar cómo la interacción entre EduxXprevexXD_gndr_Fem no es significativa, mientras que en el bloque anterior donde no se incluía esta interacción, sí que son todos los coeficientes significativos.

Tabla 7.4.

*Coeficientes de la regresión por bloques 'Salary' = f('Educ*PrevExp*Gender')*

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	61579572734,208	3	20526524244,736	150,531	,000[b]
	Residual	64089692044,906	470	136361046,904		
	Total	125669264779,114	473			
2	Regression	62195548612,205	4	15548887153,051	114,889	,000[c]
	Residual	63473716166,909	469	135338414,002		
	Total	125669264779,114	473			
3	Regression	67322695631,390	5	13464539126,278	108,000	,000[d]
	Residual	58346569147,724	468	124672156,299		
	Total	125669264779,114	473			
4	Regression	69268426300,119	6	11544737716,687	95,591	,000[e]
	Residual	56400838478,995	467	120772673,403		
	Total	125669264779,114	473			
5	Regression	69281094429,004	7	9897299204,143	81,793	,000[f]
	Residual	56388170350,110	466	121004657,404		
	Total	125669264779,114	473			

a. Dependent Variable: Current Salary

b. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study)

c. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex

d. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex, EducXD_gndr_fem

e. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex, EducXD_gndr_fem, PrevexXD_gndr_Fem

f. Predictors: (Constant), D_gndr_fem, Previous Experience (months), Educational Level (years of study), EducXPrevex, EducXD_gndr_fem, PrevexXD_gndr_Fem, EducXPrevexXD_gndr_Fem

Si el alumno lo considera oportuno, puede volver a correr la sintaxis anterior eliminando el bloque de la interacción EducXPrevexXD_gndr_Fem, que hemos visto que no resulta significativa. Por lo que, en definitiva, la ecuación general quedará conforme a la Ecuación 7.11.

$$Salary = -10432.238 + 5085.358 \cdot Educ + 87.846 \cdot Prevexp + 37342.278 \cdot Gender + [-5.456 \cdot Educ \cdot Prevexp] + [-3102.582\ Educ \cdot D_{gndr_{Fem}}] + [-42.802\ Prevexp \cdot D_{gndr_{Fem}}] + e \quad (7.11)$$

De la que podemos extraer dos ecuaciones diferentes, según el género:

1. Para Hombres (D_gndr_fem=0):

$$\text{Salary}'_{Male} = -10432.238 + 5085.358 \cdot' Educ' + 87.846 \cdot' Prevex' + 37342.278 \cdot 0 + [-5.456 \cdot 'Educ * Prevex'] + [-3102.582\ Educ * 0] + [-42.802\ Prevex * 0] \quad (7.12)$$

2. Para Mujeres (D_gndr_fem=1):

$$\begin{aligned}\text{Salary}'_{Female} = & -10432{,}238 + 5085{,}358 \cdot' Educ' + 87{,}846 \cdot' Prevexp' \\ & + 37342{,}278 \cdot 1 + [-5{,}456 \cdot' Educ * Prevex'] \\ & + [-3102{,}582\ Educ * 1] + [-42{,}802\ Prevex\ * 1]\end{aligned} \quad \textbf{(7. 13)}$$

En resumen, el salario según el género vendría explicado por las Ecuaciones presentadas en 7.14:

$$\begin{aligned}\text{Salary}_{Male} = & -10432{,}238 + 5085{,}358 \cdot' Educ' + 87{,}846 \cdot' Prevexp' \\ & + [-5{,}456 \cdot Educ * Prevex]\end{aligned}$$

$$\begin{aligned}\text{Salary}_{Female} = & \ 26910{,}04 + 1982{,}776 \cdot' Educ' + 45{,}044 \cdot' Prevexp' \\ & + [-5{,}456 \cdot Educ * Prevex]\end{aligned} \quad \textbf{(7. 14)}$$

Y tal y como se ha realizado anteriormente, desde las Ecuaciones 7.14 podríamos generar diferentes ecuaciones de regresión entre los niveles cruzados de 'Educ' y 'Prevexp', lo que nos permitirá pronosticar el 'Salary' en función de 'Prevexp' (Ecuaciones 7.15) o en función de 'Educ' (Ecuaciones 7.16):

$$\begin{aligned}\text{Salary}'_{Male_{Educ}} &= (-10432{,}238 + 87{,}846 \cdot' Prevexp') \\ &+ \big(5085{,}358 - (5{,}456 \cdot Prevexp)\big) *' Educ'\end{aligned}$$

$$\begin{aligned}\text{Salary}'_{Female_{Educ}} &= (26910{,}04 + 45{,}044 \cdot' Prevexp') \\ &+ (1982{,}776 - 5{,}456 \cdot Prevexp) * 'Educ'\end{aligned} \quad \textbf{(7. 15)}$$

$$\begin{aligned}\text{Salary}'_{Male_{Prevexp}} &= (-10432{,}238 + 5085{,}358 \cdot' Educ') \\ &+ (87{,}846 - 5{,}456 \cdot Educ) * 'Prevexp'\end{aligned}$$

$$\begin{aligned}\text{Salary}'_{Female_{Prevexp}} &= (26910{,}04 + 1982{,}776 \cdot' Educ') \\ &+ (45{,}044 - 5{,}456 \cdot Educ) * 'Prevexp'\end{aligned} \quad \textbf{(7. 16)}$$

Donde, de nuevo, tomaremos 'Educ' como variable independiente y de 'Prevexp' tomaremos los tres niveles señalados anteriormente (la media + 1SD, la media y la media – 1SD) que nos dividirían dicha variable en Experiencia Previa Alta, Media y Baja, con el fin de poder obtener los valores pronosticados de Salary. Recordemos que la media y la desviación típica para 'Prevexp' eran 95.86 y 104.586, respectivamente; por lo tanto, para los hombres en los diferentes niveles obtendríamos las Ecuaciones de 7.17 para el nivel alto, de 7.18 para el nivel medio y de 7.19 para el nivel bajo, ocurriendo lo mismo para las mujeres:

High level, male:

$$\text{Salary}'_{\text{Male,HighLevel}_{\text{Educ}}} = -10432.238 + 87.846 * \left(95 + (1.96 * 104.586)\right)$$
$$+(5085.358 - 5.456 * (95 + 1.96 * 104.586)) * Educ$$
$$\text{Salary}'_{\text{Male,HighLevel}_{\text{Educ}}} = 15920.55 + 3448.558 * Educ \quad \text{(7. 17)}$$

Mean level, male:

$$\text{Salary}'_{\text{Male,MeanLevel}_{\text{Educ}}} = (-10432.238 + 87.846 * 95)$$
$$+(5085.358 - 5.456 * 95) * Educ$$
$$\text{Salary}'_{\text{Male,MeanLevel}_{\text{Educ}}} = -2086{,}868 + 4567{,}228 * Educ \quad \text{(7. 18)}$$

Low level, male:

$$\text{Salary}'_{\text{Male,LowLevel}_{\text{Educ}}} = -10432{,}238 + 87{,}846$$
$$\cdot [95 - (1{,}96 * 104{,}586)] + [5085{,}358 - 5{,}456 \cdot [95 - (1{,}96 * 104{,}586)]] * \text{Educ}$$
$$\text{Salary}'_{\text{Male,LowLevel}_{\text{Educ}}} = -20094{,}29 + 5685{,}45 * Educ \quad \text{(7. 19)}$$

High level, female:

$$\text{Salary}'_{\text{Female,HighLevel}_{\text{Educ}}} = 26910{,}04 + 45{,}044$$
$$\cdot [95 + (1{,}96 * 104{,}586)] + \left[1982{,}776 - 5{,}456 \cdot [95 + (1{,}96 * 104{,}586)]\right] * Educ$$
$$\text{Salary}'_{\text{Female,HighLevel}_{\text{Educ}}} = 40423{,}24 + 345{,}976 * Educ \quad \text{(7. 20)}$$

Mean level, female:

$$\text{Salary}'_{\text{Female,MeanLevel}_{\text{Educ}}} = 26910{,}04 + 45{,}044 \cdot (95)$$
$$+ [1982{,}776 - 5{,}456 \cdot 95] * Educ$$
$$\text{Salary}'_{\text{Female,MeanLevel}_{\text{Educ}}} = 31189{,}22 + 1464{,}456 * Educ \quad \text{(7. 21)}$$

Low level, female:

$$\text{Salary}'_{\text{Female,LowLevel}_{\text{Educ}}} = 26910{,}04 + 45{,}044$$
$$\cdot [95 - (1{,}96 * 104{,}586)] + \left[1982{,}776 - 5{,}456 \cdot [95 - (1{,}96 * 104{,}586)]\right] * Educ$$
$$\text{Salary}'_{\text{Female,LowLevel}_{\text{Educ}}} = 21955{,}71 + 2577{,}16 * Educ \quad \text{(7. 22)}$$

Desde las Ecuaciones de regresión 7.17 hasta la 7.22 podemos calcular los tres pronósticos para la variable 'Salary' según el sexo y los niveles de 'Prevexp', llamándolos Yp_M_High, Yp_M_Med, Yp_M_Low, Yp_F_High, Yp_F_Med e Yp_F_Low, a través de la sintaxis:

```
COMPUTE Yp_M_High = 15920.55+ 3448.558*Educ.
     VARIABLE LABELS  Yp_M_High 'Male High Prevexp'.
```

```
      EXECUTE.
COMPUTE Yp_M_Med =-2086.868+ 4567.228*Educ.
      VARIABLE LABELS  Yp_M_Med 'Male Medium Prevexp'.
      EXECUTE.
COMPUTE Yp_M_Low = -20094.29+ 5685.45*Educ.
      VARIABLE LABELS  Yp_M_Low 'Male Low Prevexp'.
      EXECUTE.
COMPUTE Yp_F_High = 40423.24 +345.976*Educ.
      VARIABLE LABELS  Yp_F_High 'Female High Prevexp'.
      EXECUTE.
COMPUTE Yp_F_Med =31189.22 + 1464.456*Educ.
      VARIABLE LABELS Yp_F_Med 'Female Medium Prevexp'.
      EXECUTE.
COMPUTE Yp_F_Low = 21955.71 + 2577.16*Educ.
      VARIABLE LABELS  Yp_F_Low 'Female Low Prevexp'.
      EXECUTE.
```

Para con estas nuevas seis variables realizar el diagrama de dispersión correspondiente, que aparece en la Figura 7.4, utilizando para ello la siguiente sintaxis:

```
GRAPH
      /SCATTERPLOT(OVERLAY)=educ educ educ educ educ educ
      WITH Yp_M_High Yp_M_Med Yp_M_Low Yp_F_High Yp_F_Med
      Yp_F_Low (PAIR)
      /MISSING=LISTWISE.
```

Donde podemos observar, además del comportamiento diferencial de las variables por sexo (en general, a más años de educación los hombres crecen más en salario, aunque en niveles bajos de educación los hombres cobran menos que las mujeres), también las diferencias observadas en la interacción de educación con experiencia previa, donde vemos quienes tienen una experiencia previa mayor y menor nivel educativo empiezan cobrando más, aunque conforme va aumentando el nivel educativo se cobra menos.

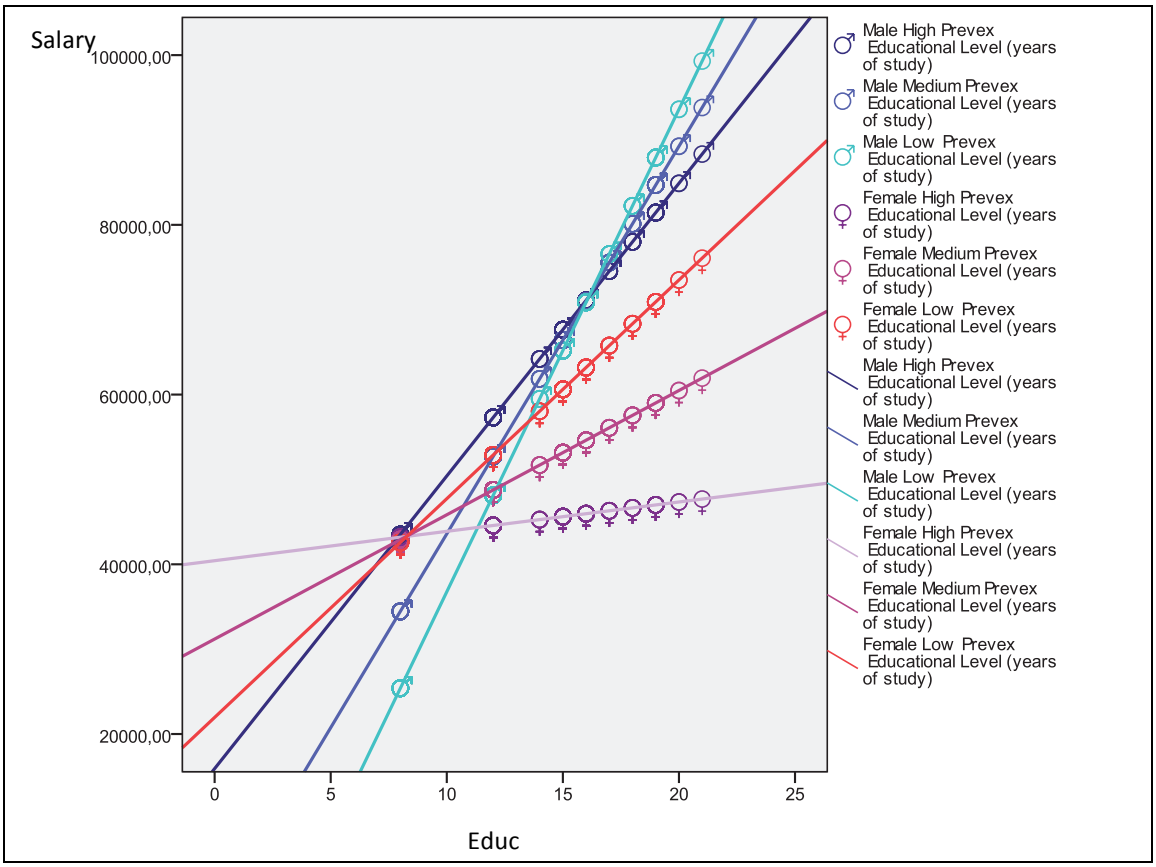


Figura 7.4. Diagrama de dispersión entre 'Salary' y 'Educ' (eje X: 'Educ', eje Y: 'Salary') para cada uno de los tres niveles de 'Prevexp' y del género

7.3. Conclusiones

En esta unidad se ha comprobado cómo cuando hay interacción entre dos variables continuas (por ejemplo: educación y experiencia previa):

- la pendiente de cada una de ellas cambia en función del valor que se tome en la otra variable, y que a su vez,
- cuando hay interacción de las variables continuas con una variable de grupos (por ejemplo, el género), la pendiente también cambia según el grupo al que se pertenezca.

La interpretación de las interacciones es diferente en cada caso, pero con los ejemplos expuestos, queda claro que el analista de datos puede llevar a cabo un pronóstico detallado en función del grupo al que pertenece cada persona (Ecuaciones 7.11 a 7.13) y del nivel de interés para cada valor de las variables independientes continuas.

Hay publicaciones (Aguinis, 2004; Mossholder, Kemery y Bedeian, 1990) en las que se recomienda llevar a cabo la interpretación de la interacción en función de los signos de cada coeficiente (de las variables principales y de la interacción), pero este sistema no es del todo recomendable, pues aparte de no ser fidedigno, se ha de tener en cuenta el sistema de codificación de cada grupo; lo ideal es hacer la ecuación de regresión para cada grupo (cuando la interacción contiene una variable de grupos), y en el caso de variables continuas, lo recomendable es llevar a cabo la representación de distintos valores «fijos» de una de las variables de interacción, tal como hemos hecho en los anteriores ejemplos (con tres valores de corte: la media más una desviación típica, la media y la media menos una desviación típica).

Lecturas recomendadas

No hay en los manuales ejemplos de interacción de dos variables independientes continuas con otra categórica, pero como repaso puede servir el manual de Aiken y West (1991), en el que se expone el ejemplo de la interacción entre tres variables independientes.

Bibliografía

Aguinis, H. (2004). *Regression analysis for categorical moderators*. NY: Guilford Press.

Aiken, L. S.; West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Mossholder, K. W.; Kemery, E. R.; Bedeian, A. G. (1990). «On using regression coefficients to interpret moderator effects». *Educational and Psychological Measurement, 50*, 255-263.

Actividades

A) Se quiere hacer la regresión lineal con interacción para la regresión cuya función será: Salary= f('edu* jobtime').

1. Para hacer este ejercicio previamente tendrás obtener la variable 'eduX jobtime'', haciendo el producto de 'Educ' por 'Jobtime'.
2. Haz la ecuación de regresión: Salary = f('eduXjobtime').

   Interpreta la significación de los estadísticos de conjunto y comenta el valor de R^2.
3. Escribe la ecuación de regresión. Interpreta la significación de los coeficientes obtenidos.
4. Escribe la ecuación de regresión para la media +1.96 DT, la media y la media –1.96 DT.
5. Haz una figura con los pronósticos obtenidos en la regresión.
6. Interpreta los resultados obtenidos.
7. En el caso de que la interacción del modelo propuesto sea estadísticamente no significativa, estima el modelo que mejor ajuste a los datos.

B) Se quiere hacer la regresión lineal con interacción para la regresión cuya función será: Salbegin= f('edu*prevex*center'), es decir, 'Beginning Salary' es función de la interacción entre 'Educational Level (years of study)', 'Previous Experience (months)' y 'Work center'.

Se pide:

1. Haz la ecuación que mejor ajuste a los datos a partir de la hipótesis planteada.
2. Interpreta la significación de los estadísticos de conjunto y comenta el valor de R^2.
3. Escribe la ecuación de regresión general. Interpreta la significación de los coeficientes obtenidos.
4. Escribe la ecuación de regresión para cada centro de trabajo.