CAPÍTULO 3
DISEÑO DE LA INVESTIGACIÓN

En primer lugar tratamos los aspectos relacionados con la población, ya que este estudio pretende ser representativo de la totalidad de empresas manufactureras españolas. Para ello abordamos el tipo de fuente de información utilizada, dichos datos son elaborados por la Fundación SEPI (fundación tutelada por la Sociedad Estatal de Participaciones Industriales) a través de la Encuesta Sobre Estrategias Empresariales —ESEE—. Explicamos el muestreo utilizado para seleccionar a las empresas cuyo comportamiento es extrapolado a la población y se presenta la ficha técnica.

En un segundo lugar se especifica la forma a través de la cual hemos construido las variables utilizadas en el estudio empírico. Para ello diferenciamos dos apartados: el primero se centra en explicar la definición de la variable dependiente y el segundo en las variables explicativas del modelo. Al final de cada uno de estos apartados se presenta un cuadro resumen con los campos de la ESEE utilizados, así como la relación de hipótesis atribuidas a cada variable.

Finalmente, se explican teóricamente cada una de las técnicas estadísticas utilizadas para la contrastación de las hipótesis formuladas en el anterior capítulo. Tal es el caso de los árboles de decisión, la regresión logística binomial, la regresión logística multinomial y la regresión logística ordinal.

3.1. METODOLOGÍA Y DISEÑO DE LA INVESTIGACIÓN

3.1.1. Población y fuente de información

La población objeto de estudio son las empresas manufactureras españolas que realizan inversiones directas en el exterior. Para determinar si las empresas son o no susceptibles de su inclusión se siguió un doble criterio:

— Todas aquellas empresas pertenecientes a lo que se conoce como industria manufacturera que contasen con 10 o más trabajadores contratados.

— Todas aquellas empresas que cuentan con un porcentaje de participación entre el 5 por 100 y el 100 por 100 en, al menos, una empresa situada fuera del territorio nacional.

Con relación al primer aspecto, para determinar si la empresa pertenece al sector manufacturero español, utilizaremos el criterio aplicado por la Encuesta Sobre Estrategias Empresariales —ESEE— que se basa en los códigos de la Clasificación Nacional de Actividades Económicas —CNAE— de la empresa para hacer sus propias agrupaciones. La relación de tales subsectores se presenta en el cuadro 3.1.

Con relación al segundo aspecto, se ha utilizado el criterio de acotar el margen de participación de las empresas entre el 5 por 100 y el 100 por 100 ya que es el empleado por la mayoría de los autores, como se describe detalladamente en el apartado 3.2.1 y se muestra en el cuadro 3.7.

La fuente de información utilizada para la contrastación de las hipótesis se ha obtenido a partir de un estudio descriptivo longitudinal que abarca el período 2000-2002 (ambos inclusive), en concreto hemos utilizado una encuesta de panel [1]. Tanto el diseño del cuestionario como el trabajo de campo es realizado anualmente por la Fundación SEPI bajo el nombre de Encuesta Sobre Estrategias Empresariales —ESEE—, y toda la información está sometida a controles de validación y de consistencia lógica para certificar su calidad y consistencia temporal.

El período de investigación se ha fijado de acuerdo a la naturaleza de la encuesta que emplean, ya que es en el año 2000 cuando incorporan las preguntas relativas a las inversiones directas que realizan las

[1] El panel es una técnica cuantitativa caracterizada por la recogida de información externa primaria de manera periódica a partir de una muestra cuyo grueso es permanente (ESTEBAN *et al.,* 1997: 261).

CUADRO 3.1

CLASIFICACIÓN DE LA ESEE DE LAS EMPRESAS MANUFACTURERAS ESPAÑOLAS A PARTIR DEL CNAE-93

Sector ESEE	Concepto	CNAE-93
1	Industria cárnica	151
2	Productos alimenticios y tabaco	152 a 158 + 160
3	Bebidas	159
4	Textiles y vestido	171 a 177 y 181 a 183
5	Cuero y calzado	191 a 193
6	Industria de la madera	201 a 205
7	Industria del papel	211 + 212
8	Edición y artes gráficas	221 a 223
9	Productos químicos	241 a 247
10	Productos de caucho y plástico	251 a 252
11	Productos minerales no metálicos	261 a 268
12	Metales férreos y no férreos	271 a 275
13	Productos metálicos	281 a 287
14	Máquinas agrícolas e industriales	291 a 297
15	Máquinas de oficina, proceso de datos, etc.	300 + (331 a 335)
16	Maquinaria y material eléctrico	311 a 316 y 321 a 323
17	Vehículos de motor	341 a 343
18	Otro material de transporte	351 a 355
19	Industria del mueble	361
20	Otras industrias manufactureras	362 a 366, 371 a 372

Fuente: Fundación SEPI.

empresas manufactureras españolas en el extranjero, de manera tal que se pueda vincular la IDE con las características propias de las empresas (MERINO y MUÑOZ, 2002: 13).

3.1.2. Diseño de la muestra

La determinación de la muestra ha sido realizada por la Fundación SEPI con un carácter mixto, en función del tamaño y del sector de actividad al que pertenece la empresa.

Por un lado, el criterio del tamaño distingue entre empresas de 200 o menos trabajadores —en las que se aplica un muestreo aleatorio estratificado— y las de más de 200 trabajadores —donde el criterio para la selección es censal—. Por otro lado, el criterio del sector se aplica utilizando básicamente el CNAE-93 a dos dígitos (MORENO y RODRÍGUEZ, 1998: 26-27). Al tratarse de una encuesta de panel, los esfuerzos se han encaminado a reducir el deterioro de la muestra inicial —evitando el decaimiento de la colaboración de las empresas— así como a la incorporación cada año de todas las empresas de nueva creación mayores de 200 trabajadores y una muestra seleccionada aleatoriamente que representa aproximadamente el 5 por 100 de las empresas nuevas entre 10 y 200 trabajadores.

La delimitación de la muestra que utilizaremos para el período 2000-2002 se muestra en el cuadro 3.2 y recoge cómo de las 10.355 observaciones iniciales contempladas en la ESEE, la muestra se concreta hasta seleccionar aquellas que efectivamente responden a la encuesta y, al mismo tiempo, se catalogan como inversoras directas en el exterior.

CUADRO 3.2

DELIMITACIÓN DE LA MUESTRA

Número de observaciones para el período 2000-2002
10.355 observaciones[2] de empresas recogidas de manera global en la ESEE
5.302 observaciones de empresas en la muestra viva[3] de la ESEE
3.458 observaciones de empresas en la ESEE que realizan actividades internacionales
635 observaciones de empresas en la ESEE que realizan inversiones directas en el exterior

Fuente: Elaboración propia.

De manera más específica, la distribución de las observaciones queda representada a través de los cuadros 3.3, 3.4 y 3.5 donde se procede a la desagregación de los datos de la muestra viva.

[2] Téngase en cuenta que estamos hablando de *observaciones* y no de empresas, esto es debido a que estamos operando con un panel incompleto de datos.

[3] Precisamos que la muestra a la que nos estamos refiriendo es la muestra *viva* para explicitar que estamos considerando a las empresas que realmente contestan el cuestionario. Esto se debe a que en los paneles desaparecen un cierto número de empresas cada año.

CUADRO 3.3

DISTRIBUCIÓN DE LA MUESTRA SEGÚN SU ACTIVIDAD INTERNACIONAL

Actividad internacional	N.º de observaciones	Porcentajes
Nula	1.844	34,8%
Activa	3.458	65,2%
Total	5.302	100%

Fuente: Elaboración propia.

CUADRO 3.4

DISTRIBUCIÓN DE LA MUESTRA SEGÚN EL TIPO DE MODALIDAD DE ENTRADA

Modalidad de entrada	N.º de observaciones	Porcentajes
Inversiones Indirectas en el Exterior[4]	2.823	81,7%
Inversiones Directas en el Exterior	635	18,3%
Total	3.458	100%

Fuente: Elaboración propia.

CUADRO 3.5

DISTRIBUCIÓN DE LA MUESTRA SEGÚN EL TIPO DE IDE EMPLEADA

Tipo de IDE	N.º de observaciones	Porcentajes
Empresa Conjunta Internacional	262	41,26%
Filial de Plena Propiedad	373	58,74%
Total	635	100%

Fuente: Elaboración propia.

Al analizar los cuadros anteriores podemos hacernos una idea general del tipo de implicación internacional que presentan las empresas manu-

[4] El cálculo de las empresas que realizan actividades de inversión indirecta en el exterior se ha realizado a través del cálculo agregado de empresas cuya inversión en el capital de una empresa extranjera no alcanza el 5 por 100, así como las empresas que, careciendo de inversión alguna en el capital de empresas extranjeras, realizan actividades de exportación.

factureras españolas. Así pues, podemos ver cómo aproximadamente el 65 por 100 de las empresas [5] presentan algún tipo de actividad internacional, mientras que sólo el 35 por 100 son totalmente inactivas en este campo.

Si nos centramos en ese 65 por 100 de empresas internacionales podemos observar el grado de compromiso que adquieren, de manera tal que tan sólo el 18 por 100 de las mismas emprenden una IDE. Por tanto, podemos deducir que, aunque el porcentaje de participación internacional es relativamente elevado, el grado de implicación que tienen es bajo, ya que aproximadamente el 82 por 100 de las empresas de las que disponemos no compromete sus recursos de manera significativa en los mercados internacionales.

Finalmente, dentro de ese 18 por 100 de empresas que sí realizan una IDE, obtenemos la distribución final de la muestra. Obtenemos que aproximadamente el 59 por 100 de las mismas se decantan por la plena propiedad de sus filiales y el 41 por 100 optan por la ECI. Estos porcentajes finales son consistentes con estudios anteriores, como por ejemplo el de HENNART (1991), de forma que el tipo de distribución que adoptan las empresas, según el tipo de IDE, suele oscilar en unos porcentajes que se aproximan al 60 por 100 para las filiales de plena propiedad y en un 40 por 100 para la ECI.

Finalmente, presentamos la ficha técnica de los datos en el cuadro 3.6.

3.2. MEDICIÓN DE LAS VARIABLES

3.2.1. Variable dependiente

En primer lugar, debemos recordar que este estudio trata de identificar los factores que explican la elección de la ECI como modalidad de entrada en los mercados internacionales. Por ello, en este apartado vamos a exponer la forma en la que ha sido definida la variable dependiente del modelo, es decir, la modalidad de entrada —MO_ENT y MO_ENT_95.

En la literatura se han realizado una gran cantidad de trabajos sobre la IDE, por lo que se ha procedido a la revisión de las diferentes formas de medida, excluyendo aquellas mediciones que son incompatibles [6] con nuestro estudio. Por ello, nos hemos centrado en aquellos trabajos que tratan de definir la variable dependiente a través de una *proxy,* el porcentaje de participación accionarial en la «filial» extranjera.

---

[5] Aunque el término correcto sea «observaciones de empresa» por tratarse de un *pooled,* a partir de aquí se prescindirá de tal precisión para facilitar la lectura y compresión del trabajo.

[6] Por «incompatibles» nos estamos refiriendo a aquellos estudios que utilizan una encuesta en la que se pregunta abiertamente por la modalidad de entrada que ha seguido la empresa encuestada, ya que la ESEE carece de tal información.

CUADRO 3.6

FICHA TÉCNICA DEL ESTUDIO

Universo:	Empresas manufactureras localizadas en España
Ámbito geográfico:	Todo el territorio nacional
Diseño del cuestionario:	Fundación SEPI
Unidad de análisis:	Empresa
Tamaño muestral:	635 observaciones de empresas manufactureras españolas que desarrollan una IDE
Error muestral:	± 0,02 (p=q=50)
Nivel de confianza:	95% (K = 2 sigma)
Diseño muestral:	Aleatorio estratificado y censal según tamaño de la empresa y sector de pertenencia
Trabajo de campo:	Fundación SEPI
Período de análisis:	Años 2000, 2001 y 2002
Tratamiento de la información:	Answer Tree 3.0 SPSS 12.0 para Windows
Técnicas de análisis:	Árboles de Decisión Regresión Logística Binomial Regresión Logística Multinomial Regresión Logística Ordinal

Fuente: Elaboración propia.

De hecho, en el gráfico 3.1 podemos observar cómo la participación accionarial en empresas extranjeras es la magnitud principal del monto total de la IDE, de forma que es la forma más representativa de aproximar esta variable.

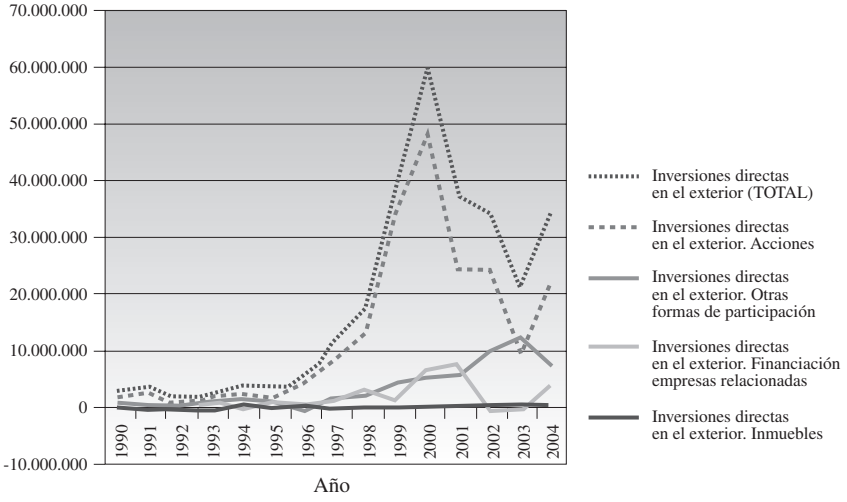
De esta forma podemos observar en el cuadro 3.7 cómo la gran mayoría de los autores optan por unos márgenes comprendidos entre el 5 por 100 y el 95 por 100 para identificar la ECI, y un margen del 95 por 100 al 100 por 100 para las filiales de plena propiedad.

Por ello, nosotros vamos a optar por seguir esta corriente mayoritaria para definir la variable MO_ENT [7].

[7] No obstante, en el estudio empírico y análisis del modelo se ha procedido a su comparación con la corriente minoritaria que disminuía el margen de identificación de la ECI al 10 por 100-95 por 100, para comprobar si este cambio era relevante a la hora de obtener y analizar los resultados. Al hacerlo no se obtuvo ningún cambio significativo, por lo que se puede afirmar que el cambio de intervalo no supone ninguna alteración en nuestros resultados ni conclusiones.

GRÁFICO 3.1

EVOLUCIÓN DE LA IDE EN ESPAÑA
(1990-2004)



Fuente: Elaboración propia a partir de los datos del Banco de España.

CUADRO 3.7

RELACIÓN DE AUTORES Y LA FORMA DE MEDICIÓN DE LA IDE

Autores	**Construcción de la modalidad de IDE**	
	Empresa conjunta internacional	*Filial extranjera de plena propiedad*
Franko (1971)	(5%-95%]	(95%-100%]
Stopford y Wells (1972)	(5%-95%]	(95%-100%]
Gomes-Casseres (1989: 11)	(5%-95%]	(95%-100%]
Gomes-Casseres (1990: 8)	(5%-95%]	(95%-100%]
Hennart (1991: 487-488)	(5%-95%]	(95%-100%]
Hennart y Reddy (1997)	(5%-95%]	(95%-100%]
Makino y Newpert (2000: 708-709)	(5%-95%)	[95%-100%]
Chen y Hennart (2002: 8)	(5%-95%)	[95%-100%]
Pak y Park (2004: 8-9)	[5%-95%)	[95%-100%]
Chowdhury (1992: 120)	(10%-95%)	[95%-100%]
Chan (1995: 40)	(10%-95%)	[95%-100%]
Hennart y Larimo (1998: 525)	(10%-95%)	[95%-100%]
Chen y Hennart (2004: 1130)	(5%-80%]	(80%-100%]

Fuente: Elaboración propia.

Por otra parte, tenemos el estudio de STOPFORD y WELLS (1972) donde van más allá de esta distinción dicotómica, ya que discriminan entre las dos categorías básicas de la ECI —mayoritaria *vs.* minoritaria— y las filiales de plena propiedad —MO_ENT_95—. De tal forma que el umbral entre los diferentes tipos de ECI lo sitúan en el 50 por 100.

Por tanto, a partir de los datos de la ESEE, la creación de las dos variables dependientes alternativas queda de la siguiente manera:

CUADRO 3.8

PROPIEDADES DE LA VARIABLE DEPENDIENTE

Variable dependiente	Campos	Definición
Empresa conjunta internacional	**A27:** *«Indique, para la principal empresa participada, las siguientes características o rasgos»* **A27_1:** «Porcentaje de participación»	**MO_ENT:** Variable dicotómica. **Y=1,** si (A27_1 > 5 & A27_1 ≤ 95) → *Empresa conjunta internacional.* **Y=0,** si (A27_1 > 95) → *Filial de plena propiedad.*
		MO_ENT_95: Variable politómica. **Y=1,** si (A27_1 > 5 & A27_1 ≤ 50) → *ECI Minoritaria.* **Y=2,** si (A27_1 > 50 & A27_1 ≤ 95) → *ECI Mayoritaria.* **Y=3,** si (A27_1 > 95) → *Filial de plena propiedad.*

Fuente: Elaboración propia.

3.2.2. Variables independientes

En este apartado, vamos a identificar los factores explicativos de la ECI para definir las variables correspondientes. Por ello, en un primer momento se procederá a la explicación de la creación de cada una de las variables independientes y, al finalizar, se presentará un cuadro resumen.

a) *Tamaño de la empresa inversora*

A lo largo de la literatura se han utilizado diversas formas de medición para aproximar este factor, por ejemplo el volumen de los activos

de la empresa (DUBIN, 1975; KOGUT y SINGH, 1988; YU e ITO, 1988; CHAN, 1995; PADMANABHAN y CHO, 1996; TAN y VERTINSKY, 1996), las ventas anuales en el mercado doméstico (KIMURA, 1989; DELIOS y HENISZ, 2000; DOMKE-DAMONTE, 2000; GUILLÉN, 2003: 191) y el número de empleados (NORBURN y BIRLEY, 1986; GATIGNON y ANDERSON, 1988; DELIOS y BEAMISH, 1999; PAK y PARK, 2004).

Los dos primeros indicadores han sido descartados en nuestro estudio ya que tanto los métodos contables como el tipo de moneda varían entre los países, de forma que podríamos estar introduciendo sesgos difíciles de aislar (BROUTHERS y BROUTHERS, 2001: 182), por lo que se recomienda el uso del número de empleados para aproximar dicha variable (GATIGNON y ANDERSON, 1988: 318; ERRAMILLI, 1991: 487; ERRAMILLI y RAO, 1993: 36; CONTRACTOR y KUNDU, 1998: 335; LUO, 2002: 13).

Para ello vamos a utilizar una variable dicotómica que distinga entre las dos modalidades que nos interesan: empresas pequeñas *versus* empresas grandes (MERINO y SALAS, 1998: 22; CHEN *et al.,* 2004).

Para medir esta variable —TAMAÑO—, se ha utilizado el campo de la ESEE que recoge el personal de la empresa y el punto de corte que se ha elegido es el de 200 trabajadores. De tal manera que consideraremos empresas pequeñas las que tengan 200 o menos empleados, y empresas grandes las que tengan más de 200.

El motivo por el que se ha elegido este punto de corte y no otro es la forma según la cual ha sido diseñada la ESEE. A la hora de seleccionar las empresas a encuestar, la ESEE optó por hacerlo, entre otros criterios, según su tamaño y eligió el valor de 200 trabajadores como frontera. Como nos encontramos ante una encuesta de panel, cada año se sustituyen las empresas que se dan de baja por otras que pertenezcan, entre otros criterios, a su mismo tramo de tamaño. Así se perpetúan en el tiempo los criterios de exhaustividad y de muestreo aleatorio.

Por ello, en esta investigación hemos optado por mantener los valores según los cuales está creada la ESEE, para mantener su representatividad muestral.

b) *Tamaño del proyecto en el extranjero*

Con respecto a la medición del tamaño del proyecto, en la literatura se han utilizado diferentes indicadores para aproximar este dato. Así pues encontramos mediciones absolutas como el volumen de activos totales de la filial (FAGRE y WELLS, 1982; GOMES-CASSERES, 1990), la cantidad invertida en la misma (WEI *et al.,* 2004), el número de empleados (GATIG-

NON y ANDERSON, 1988); así como mediciones relativas como el número de empleados de la filial entre el número de empleados de la empresa matriz (HARZING, 2002) o la inversión realizada entre los activos totales (PADMANABHAN y CHO, 1996).

Al igual que en el caso anterior, vamos a seguir el trabajo de GATIGNON y ANDERSON (1988: 308), donde recomiendan la utilización del número de empleados de la filial extranjera (el tamaño de la filial —T_FILIAL—) como aproximación de este factor.

La lógica que se ha aplicado para crear esta variable es la misma que en el caso anterior. El tamaño de la filial se va a calcular a través del número de empleados que estén trabajando en dicha filial y, para ser consistentes, el punto de corte vuelve a ser el de 200 trabajadores.

Se considerará que un proyecto requerirá una elevada inversión cuando el número de empleados en la filial extranjera sea de 200 trabajadores o más y, en caso contrario, requerirá de una pequeña inversión.

c) *Intensidad investigadora de la empresa*

Para la construcción de una variable que recoja el efecto que causa la intensidad investigadora de la empresa en la elección de la ECI, tendremos en cuenta el grado de inversión en I+D, ya que esta *proxy* está ampliamente aceptada en la literatura (ANDERSON, 1988; KOGUT y SINGH, 1988; PADMANABHAN y CHO, 1996; LUO, 1998; MUTINELLI y PISCITELLO, 1998; DELIOS y BEAMISH, 1999; LU, 2002; PAK y PARK, 2004; TSAI y CHENG, 2004).

Por ello vamos a crear una variable dicotómica —C_TECNO_SEC— que distinga entre las empresas que destinan muchos recursos a la I+D, de las que no lo hacen. Para ello vamos a crear, en primer lugar, una variable continua que recoja la intensidad tecnológica de la empresa —gastos en I+D normalizado sobre el total de las ventas— y, en segundo lugar, procederemos a establecer las dos categorías a través de su comparación con la media del sector. Por ello si la comparación con la media del sector es positiva para la empresa, presupondremos que se caracteriza por una elevada intensidad investigadora. Si es negativa presupondremos lo contrario.

d) *Experiencia doméstica*

Para su medición vamos a fijarnos en el análisis de MERINO y RODRÍGUEZ (1997: 737), que miden esta variable a través de la *proxy* de la edad de la empresa.

Así pues, para determinar su efecto vamos a definir una variable dicotómica —EXP_DOMES— que distinga entre las empresas que tienen un elevado nivel de experiencia doméstica de las que tienen un nivel bajo. Para ello vamos a seguir dos pasos: En primer lugar construiremos una variable continua que recoja la edad por exceso de la empresa (calculada como el año en el que se responde el cuestionario más uno [8]) y, a esta suma, le restaremos su fecha de constitución. En segundo lugar diferenciaremos dos categorías comparando la edad de la empresa y la edad media de su sector. Si la empresa tiene más edad que la media del sector presupondremos una elevada experiencia local, en caso contrario se supondrá una baja experiencia local.

e) *Experiencia internacional*

Para la medición de este factor se encuentran varias propuestas en los estudios realizados hasta el momento. Así se propone utilizar como *proxy* el porcentaje de ingresos atribuidos a las operaciones extranjeras (AGARWAL y RAMASWAMI, 1992), más utilizado es el número de países donde opera la empresa o donde tiene filiales (CHO, 1985; HEDLUNG y KVERNELAND, 1985; CAVES y MEHRA, 1986; KOGUT y SINGH, 1988; TSAI y CHENG, 2004) y, la *proxy* más utilizada es el número de años que la empresa lleva operando en los mercados internacionales (ERRAMILLI, 1991; HENNART, 1991; CONTRACTOR y KUNDU, 1998; DOMKE-DAMONTE, 2000, BROUTHERS y BROUTHERS, 2001; ERRAMILLI *et al.,* 2002; HARZING, 2002; CHEN y HENNART, 2004, TSAI y CHENG, 2004).

No obstante, por limitaciones de la encuesta que utilizamos, no podemos utilizar ninguna de las variables anteriores, ya que no contamos con tales datos.

Así pues, para la definición de la variable dicotómica que recoja su efecto —EXP_INTER— y distinga entre niveles altos y bajos de experiencia internacional, utilizaremos las preguntas de la ESEE, concretamente las que recogen los mecanismos a través de los cuales se está presente en el extranjero. Si la empresa afirma haber utilizado previamente mecanismos que supongan su participación directa en las operaciones internacionales, identificaremos a tal empresa con la característica de tener una elevada experiencia internacional; en caso contrario supondremos que tiene una baja experiencia internacional.

[8] Procederemos de esta forma para evitar que empresas que se hayan constituido el mismo año en el que responden al cuestionario figuren con una edad igual a cero años. De manera que el cálculo de la edad será por exceso y, al construirse esta variable por comparación con la media del sector, no supondrá sesgo alguno esta operación.

f) *Estrategias de internacionalización*

Con respecto a este factor, el objetivo es crear una variable que recoja la tipología de las estrategias de internacionalización, es decir, crear una variable dicotómica —STRAT_INTER— cuyas categorías se relacionen con las estrategias multipaís y globales.

Por ello, estableceremos una correspondencia simple con una de las preguntas del cuestionario de la ESEE en la que se pregunta por el grado de estandarización de los productos que fabrica ya que, según HARZING (2002: 212-213), las empresas que adoptan una estrategia global se caracterizan por producir productos altamente estandarizados; mientras que las empresas con estrategias multipaís deben modificar sus productos para adecuarlos a las necesidades específicas del mercado de destino (HARZING, 2000: 110).

Por ello, como por estrategias multipaís entendemos aquéllas en las que se produce una adaptación del producto a los diferentes mercados y por estrategias globales a aquéllas cuyos productos son iguales para todos los destinos, podemos atribuir el valor de estrategia multipaís a las empresas que en el cuestionario respondan que sus productos se diseñan, en su mayoría, de manera específica para cada cliente; mientras que si responde que los productos son muy estandarizados y que, en su mayoría, son iguales para todos los clientes, presupondremos que se refiere a una estrategia global.

g) *Estrategias de crecimiento*

Para la creación de la variable que recoja el efecto que causan las estrategias de crecimiento en la elección de la ECI, hemos utilizado dos variables:

En primer lugar, para recoger la dirección del crecimiento de la empresa y distinguir entre las estrategias de expansión, de diversificación relacionada y de diversificación no relacionada, hemos utilizado una variable tricotómica —STRAT_CREC—, esta variable se corresponde íntegramente con el contenido proporcionado por la ESEE sobre el índice de diversificación de las empresas, creado explícitamente para medir este efecto.

En segundo lugar, hemos procedido a la transformación de la anterior variable en otra dicotómica[9] —STRAT_CREC_2— de manera que en una

[9] La utilidad de crear esta segunda variable para recoger el mismo efecto que su predecesora es para conseguir una mayor capacidad interpretativa del modelo de regresión logística, como se expondrá en el capítulo 4.

de las categorías tengamos a las empresas que realizan estrategias de expansión, mientras que en la otra categoría recojamos a aquellas que realicen actividades de diversificación (tanto relacionada como no relacionada).

h) *Estrategias de negocio*

El último tipo de estrategia que vamos a estudiar es el referido a las estrategias de negocio. En este caso, se utilizará como *proxy* para distinguir entre las estrategias de liderazgo en costes y la diferenciación de producto la intensidad publicitaria normalizada por las ventas totales del período —STRAT_NEGO—. Este ha sido el indicador seleccionado porque se ha querido resaltar el tipo de valor añadido que condiciona que las empresas prefieran un tipo de modalidad de entrada y no otra: la imagen de marca promovida por fuertes inversiones en publicidad (ANDERSON y GATIGNON, 1986) y por ser el más empleado en los estudios empíricos (KOGUT y SINGH, 1988; HENNART y PARK, 1993; MUTINELLI y PISCITELLO, 1998; DELIOS y BEAMISH, 1999; LU, 2002; TSAI y CHENG, 2004).

Así pues, el liderazgo en costes y la diferenciación de producto se han considerado como dos alternativas por las que puede optar la empresa, y el *continuum* que hay entre ellas se ve representado por el cociente anterior. Así, según aumenten las inversiones en campañas publicitarias se considerará que la empresa está tratando de diferenciar más su producto (CAVES, 1991: 210); mientras que según disminuye, se considerará una estrategia de liderazgo en costes.

i) *Distancia cultural*

Un factor controvertido a la hora de su medición es la distancia cultural. Hay varios autores que han tratado de crear medidas que recojan el efecto que este factor produce en las empresas que se dirigen al exterior, sin embargo, no están libres de controversia. Por ejemplo, podemos encontrar el índice de HOFSTEDE [10] (1980, 2001) como uno de los más utilizados, pero la falta de actualización de los datos, la falta de información sobre ciertos países, así como los sesgos propios de la investigación hacen cada vez más aconsejable utilizar nuevos indicadores.

Por ello nosotros vamos a utilizar un doble criterio —el económico y el lingüístico— para medir la variable distancia cultural —DIS_CUL—. Se considerarán como países con una baja distancia cultural a aquellos

[10] El razonamiento que justifica la exclusión de este indicador para la medición de la distancia cultural lo podemos encontrar en el anexo II.

que pertenezcan a la Unión Europea, ya que todos ellos han debido cumplir con unos criterios de convergencia fijados por Maastricht que, desde un punto de vista económico, hacen compatibles a todos los países miembros; y, por otra parte, consideraremos también países de baja distancia cultural a aquellos que compartan con España una distancia lingüística nula según el indicador de WEST y GRAHAM (2004: 249).

Por tanto, para aplicar este doble criterio, la ESEE nos proporciona tanto la lista desagregada de países donde la empresa tiene la filial participada más importante como el número exacto de filiales agrupadas según cuatro bloques de países —Unión Europea, Resto de la OCDE, Iberoamérica y Resto del Mundo—. De manera que nosotros utilizaremos esta última alternativa y procederemos de la siguiente manera: si la suma total de filiales extranjeras en la UE y en Iberoamérica es superior a la suma total de filiales extranjeras en el resto de la OCDE y el resto del mundo, consideraremos que, de manera general, la empresa se enfrenta a una baja distancia cultural; sin embargo, en caso contrario consideraremos que la empresa se enfrenta con una elevada distancia cultural.

j) *Riesgo País*

Para la medición del riesgo país de aquellos países a los que se dirige la empresa hemos utilizado la información facilitada por la empresa *Sovereign* sobre la clasificación de los países según su riesgo país.

Así, por una parte hemos utilizado una variable con cinco categorías —R_PAÍS— basada en el Índice del Riesgo País construido y facilitado por la empresa *Sovereign* en colaboración con *Euromoney.* De manera que cada uno de los valores de la variable se corresponden con las diferentes categorías de clasificación: riesgo país bajo, medio-bajo, medio, medio-alto y alto.

Por otra parte, hemos creado a partir de los datos anteriores una variable dicotómica[11] —R_PAÍS_2— para distinguir sólo dos tipos de países, los que consideraremos de riesgo bajo (aquellos que en la variable R_PAÍS obtienen una calificación de riesgo país bajo o medio-bajo) y los que consideraremos de alto riesgo (los que en la variable R_PAÍS tienen una calificación de riesgo medio, medio-alto o alto).

Así pues, en el cuadro 3.9 se presenta de manera esquemática la definición de cada variable, los campos de la ESEE utilizados para ello, así como la relación de hipótesis formuladas.

[11] Véase nota 9 de este capítulo.

CUADRO 3.9

RESUMEN DE LA CONSTRUCCIÓN DE LAS VARIABLES Y SUS HIPÓTESIS

Factor	Campos	Definición	Hipótesis
Tamaño de la empresa	**G1**: *«Personal ocupado en la empresa, al 31 de diciembre de 200X, según las modalidades que se indican»:* **G1_9:** «Total del personal de la empresa»	**TAMAÑO:** Variable dicotómica: **X=1,** si G1_9 ≤ 200 trabajadores → *Empresa pequeña.* **X=0,** si G1_9 > 200 trabajadores → E*mpresa grande*	**H1:** Existe mayor probabilidad de que las pequeñas empresas opten por la ECI como modalidad de entrada en el país de destino.
Tamaño del proyecto en el extranjero	**A27:** *«Indique, para la principal empresa participada, las siguientes características o rasgos»:* **A27_2:** «Número de trabajadores»	**T_FILIAL:** Variable dicotómica: **X=1,** si A27_2 > 200 trabajadores → *Elevado tamaño del proyecto en el extranjero.* **X=0,** si A27_2 ≤ 200 trabajadores → *Bajo tamaño del proyecto en el extranjero.*	**H2:** Existe mayor probabilidad de que las empresas que acometan una elevada inversión opten por la ECI como modalidad de entrada en el país de destino.
Intensidad investigadora de la empresa	**E2_1_2:** «Gastos externos (I+D)» **E2_2_2:** «Gastos Internos (I+D)» **HA1_9:** «Total de ventas» **A13_2_1:** Sector	**C_TECNO_SEC :** Variable dicotómica. **X = 1**, si $\left[\frac{(E2\_1\_2 + E2\_2\_2)}{HA1\_9} - Media\_del\_Sector\right] \leq 0$ → *Baja intensidad investigadora.* **X = 0**, si $\left[\frac{(E2\_1\_2 + E2\_2\_2)}{HA1\_9} - Media\_del\_Sector\right] > 0$ → *Alta intensidad investigadora.*	**H3:** Existe mayor probabilidad de que las empresas con una baja intensidad investigadora opten por la ECI como modalidad de entrada en el país de destino.

Experiencia doméstica	**A6:** «Año de constitución de la empresa»	**EXP_DOMES :** Variable dicotómica. **X = 1**, si $\left[[(200X+1)-A6]-Media\_del\_Sector\right] \leq 0$ → *Bajo nivel de experiencia local.* **X = 0**, si $\left[[(200X+1)-A6]-Media\_del\_Sector\right] > 0$ → *Alto nivel de experiencia local.*	**H4.1:** La experiencia doméstica de las empresas que han decidido introducirse en un mercado extranjero a través de una IDE condicionará la elección de la ECI. **H4.1a:** Existe mayor probabilidad de que las empresas con un *bajo nivel de experiencia doméstica* opten por la ECI como modalidad de entrada en el país de destino. **H4.1b:** Existe mayor probabilidad de que las empresas con un *alto nivel de experiencia doméstica* opten por la ECI como modalidad de entrada en el país de destino.
Experiencia internacional	**F3_1:** «Dispone de medios propios» **F3_2:** «Utiliza una empresa matriz instalada en el extranjero» **F3_4:** «Participa en alguna modalidad de acción colectiva hacia la exportación»	**EXP_INTER:** Variable dicotómica. **X=1,** si (F3_1 = NO & F3_2 = NO & F3_4 = NO) → *Baja experiencia internacional.* **X=0,** si (F3_1 = SÍ o F3_2 = SÍ o F3_4 = SÍ) → *Alta experiencia internacional.*	**H4.2:** Existe mayor probabilidad de que las empresas con un bajo nivel de experiencia internacional opten por la ECI como modalidad de entrada en el país de destino.
Estrategias de internacionalización	**A15:** «Indique, si en su mayoría, los productos que fabrica son o no muy estandarizados»	**STRAT_INTER:** Variable dicotómica. **X=1,** si A15= «Los productos en su mayoría se diseñan específicamente para cada cliente» → *Estrategia Multipaís.* **X=0,** si A15= «Los productos son muy estandarizados» → *Estrategia Global.*	**H5.1:** Existe mayor probabilidad de que las empresas que desarrollen una estrategia multipaís opten por la ECI como modalidad de entrada en el país de destino.

CUADRO 3.9 *(Cont.)*

Factor	Campos	Definición	Hipótesis
Estrategias de crecimiento	**ÍNDICES DIVER/EXP:** Elaboración a partir de los dígitos del CNAE.	**STRAT_CREC:** Variable tricotómica. **X=0,** si ÍNDICE = 0 → *Expansión.* **X=1,** si ÍNDICE = 1 → *Diversificación Relacionada.* **X=2,** si ÍNDICE = 2 → *Diver. NO Relacionada.* **STRAT_CREC_2:** Variable dicotómica. **X=1,** si ÍNDICE = (1 o 2) → *Diversificación.* **X=0,** si ÍNDICE = 0 → *Expansión.*	**H5.2:** Existe mayor probabilidad de que las empresas que desarrollen una estrategia de diversificación opten por la ECI como modalidad de entrada en el país de destino.
Estrategias de negocio	**HA7_2_2:** «Gastos de publicidad, propaganda y relaciones públicas» **HA1_9:** «Total de ventas»	**STRAT_NEGO**: Variable continua $\frac{HA7\_2\_2}{HA1\_9}$	**H5.3:** Existe mayor probabilidad de que las empresas que desarrollen una estrategia de liderazgo en costes opten por la ECI como modalidad de entrada en el país de destino.

Distancia cultural	**A26:** *«Indique la localización geográfica de las empresas participadas»:* **A26_1_2:** «**UE**/número de empresas» **A26_2_2:** «**Resto de la OCDE**/ número de empresas» **A26_3_2:** «**Iberoamérica**/número de empresas» **A26_4_2:** «**Resto del mundo**/ número de empresas»	**DIS_CUL**: Variable dicotómica. **X=1**, si $(A26\_2\_2+A26\_4\_2) \geq (A26\_1\_2+A26\_3\_2)$ $\rightarrow$ *Alta distancia cultural*. **X=0**, si $(A26\_2\_2+A26\_4\_2) < (A26\_1\_2+A26\_3\_2)$ $\rightarrow$ *Baja distancia cultural*.	**H6a:** Existe mayor probabilidad de que las empresas que se dirijan a países caracterizados por una alta distancia cultural opten por la ECI como modalidad de entrada en el país de destino. **H6b:** Existe mayor probabilidad de que las empresas que se dirijan a países caracterizados por una baja distancia cultural opten por la ECI como modalidad de entrada en el país de destino.
Riesgo país	**EUROMONEY: Índice del Riesgo País** **1:** «Bajo riesgo» **2:** «Medio-Bajo riesgo» **3:** «Medio riesgo» **4:** «Medio-Alto riesgo» **5:** «Alto riesgo»	**R_PAÍS:** Variable ordinal con las cinco categorías. **R_PAÍS_2:** Variable dicotómica. **X=1,** si *R_PAÍS* $\geq 3 \rightarrow$ *Elevado riesgo país.* **X=0,** *R_PAÍS* $< 3 \rightarrow$ *Bajo riesgo país.*	**H6c:** Existe mayor probabilidad de que las empresas que se dirijan a países caracterizados por una *alta distancia cultural* y un *elevado riesgo país* opten por la ECI como modalidad de entrada en el país de destino.

Fuente: Elaboración propia.

3.3. TÉCNICAS ESTADÍSTICAS APLICADAS AL ESTUDIO

En este apartado vamos a tratar de explicar brevemente el funcionamiento de cada una de las técnicas estadísticas utilizadas para la contrastación de las hipótesis. En primer lugar trataremos el funcionamiento del árbol de decisiones, en segundo lugar explicaremos la regresión logística binomial, a continuación expondremos los principales argumentos de la regresión logística multinomial para, finalmente, concluir con la regresión logística ordinal.

3.3.1. Árboles de decisión

Como se explica en el trabajo de DÍAZ MARTÍNEZ *et al.* (2005: 6-7), los árboles de decisión son un modo de representación de la regularidad subyacente en los datos. Se presenta en forma de un conjunto de condiciones excluyentes y exhaustivas organizadas en una estructura jerárquica arborescente compuesta por nodos internos y externos conectados por ramas. Un nodo interno contiene una pregunta que es una unidad que evalúa una función de decisión para determinar cuál es el próximo nodo hijo a visitar. En contraste, un nodo externo, también llamado nodo hoja o nodo terminal, no tiene nodos hijos y se asocia con una etiqueta o valor que caracteriza a los datos que llegan al mismo. La estructura de condición y ramificación de un árbol de decisión es idónea para el problema que nos ocupa, el de clasificación. Debido al hecho de que la clasificación trata con clases o etiquetas disjuntas, es decir, una instancia es de una clase o de otra, pero no de varias clases a la vez, un árbol de decisión conducirá un ejemplo hasta una y sólo una hoja, asignándole, por tanto, una única clase.

En general, un árbol de decisión se emplea de la siguiente manera: en primer lugar, se presenta una instancia, un vector compuesto por varios atributos —en nuestro caso, una empresa caracterizada por un conjunto de ratios financieros—, al nodo inicial (o nodo raíz) del árbol de decisión. Dependiendo del resultado de la función de decisión usada por el nodo interno, el árbol nos conducirá hacia uno de los nodos hijos. Esto se repite hasta que se alcanza un nodo terminal y se asigna una etiqueta o valor a los datos de entrada. En cuanto al mecanismo de generación del árbol, existe una gran diversidad de ellos pero todos se basan en utilizar un conjunto de casos de entrenamiento sobre el que se van haciendo particiones recursivas (el conjunto se divide sucesivamente dotándole de una estructura ramificada) de acuerdo con ciertas reglas que se seleccionan de manera que se minimice una «función de impureza» que mida el grado en que los distintos subconjuntos generados son más o menos

puros, es decir, sus elementos son más o menos homogéneos (entendida la homogeneidad en el sentido de pertenencia a la misma clase).

3.3.2. Regresión logística binomial

3.3.2.1. *Delimitación de la técnica estadística*

La regresión logística binomial múltiple es una técnica de análisis multivariante en la que la variable explicada es dicotómica, toma los valores [12] 0 y 1, y las explicativas pueden ser tanto categóricas como continuas. De manera tal que se intenta predecir la probabilidad de pertenencia a un grupo a partir de las variables explicativas. La expresión de la probabilidad condicional de la ocurrencia del suceso se denota de la siguiente manera:

$$\boxed{P(Y=1\,|\,x)=\Lambda(\beta' X_i)=\pi(x)}$$

La forma específica de la ecuación del modelo presenta la siguiente forma:

$$\boxed{\pi(x)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n}}=\frac{1}{1+e^{-\beta_0-\beta_1x_1-\beta_2x_2-\ldots-\beta_nx_n}}}$$

De esta manera el predictor lineal $g = \beta_0 + \beta_1x_1 + \beta_2x_2 +...+ \beta_nx_n$ no proporciona el grupo al que pertenece una determinada observación sino, de forma indirecta, la probabilidad de pertenencia a uno de los grupos [denominados aquí $\pi(x)$].

Si despejamos el predictor lineal en la ecuación anterior obtenemos:

$$\boxed{g(x)=Ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_o+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n} \text{ donde } \left[\frac{\pi(x)}{1-\pi(x)}\right]$$

recibe el nombre de *odds* ratio.

Si hubiéramos planteado una regresión usual de la forma $y = \beta_0 + \beta_1x_1 + \beta_2x_2 +...+ \beta_nx_n + \varepsilon$, con $y = \{0,1\}$, tendríamos una serie de problemas, como que el predictor lineal tomase valores entre $-\infty$ y $+\infty$; y que el término de error no se pueda distribuir de forma normal, como se admite de forma habitual (debido a que para unos valores dados de las variables

[12] Interprétense los valores 0 y 1 como la ausencia o presencia de la característica objeto de estudio, en nuestro caso la elección por parte de la empresa de la ECI como modalidad de entrada en el país de destino.

explicativas ε es dicotómica) y tampoco tendrá varianza constante [pues el error $\varepsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)$ aumentará en valor absoluto cuando el predictor lineal aproxime sus valores a $-\infty$ o $+\infty$].

Para paliar estos inconvenientes es por lo que se realiza una transformación del predictor lineal: $g \rightarrow \frac{1}{1+e^{-g}}$, utilizando una transformación logística, que mapea el intervalo $[-\infty, +\infty]$ —en el cual toma valores g—, en el intervalo [0,1], de modo que la variable transformada puede ser considerada una probabilidad $\pi(x) = \frac{1}{1+e^{-g}}$.

a) *Estimación de los coeficientes*

El método utilizado para la estimación de los coeficientes es el de máxima verosimilitud, ya que proporciona valores que maximizan la probabilidad de obtener el conjunto de datos observados.

Para ello se construye la función de verosimilitud, que representa la probabilidad de aparición de los datos observados a partir de los parámetros desconocidos.

$$\imath(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

No obstante, es más fácil desde un punto de vista matemático operar con el logaritmo de la función anterior:

$$L(\beta) = \ln\left[\imath(\beta)\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\}$$

Así pues, los estimadores de máxima verosimilitud serán aquellos que maximicen el valor de esta función.

b) *Significatividad de los coeficientes*

Para contrastar si los coeficientes individuales estimados en el modelo son o no significativos se utiliza el estadístico de Wald.

$$W_k = \frac{\hat{\beta}_k^2}{\hat{S}_{\hat{\beta}_k}^2} = \left(\frac{\hat{\beta}_k}{\hat{S}_{\beta_k}} \right)^2$$

Lo que se trata de contrastar es si la variable independiente en cuestión no tiene ningún efecto en la predicción de la probabilidad de *Y,* es decir, la hipótesis nula contrasta si el coeficiente es igual a cero (H_0:$\beta_k = 0$), con un $\alpha = 0{,}05$.

c) *Interpretación de los coeficientes*

Para la interpretación de los coeficientes debemos tener en cuenta la forma a partir de la cual se han estimado los $\beta_0,\beta_1...\beta_n$. De modo que $\exp(\beta_i)$ se puede interpretar como el factor por el que se multiplica el *odds* ratio cuando aumenta en una unidad la variable x_j, *ceteris paribus.*

3.3.2.2. *Adecuación del modelo*

Una vez estimados los coeficientes de las variables independientes y contrastada su significatividad individual, se deben realizar una serie de comprobaciones. Por una parte debe estudiarse la bondad de ajuste global del mismo y, por otra, su capacidad predictiva.

a) *Bondad de ajuste del modelo global*

El primer aspecto a tratar es la comprobación del ajuste del modelo de regresión logística. Para ello vamos a proceder a comprobar el ajuste global del mismo a través de diferentes estadísticos.

• Pruebas ómnibus sobre los coeficientes del modelo

El estadístico utilizado es el χ^2 del modelo, con él se contrasta la hipótesis nula (H_0) de que todos los coeficientes del modelo, excepto la constante, son iguales a cero, frente a la hipótesis alternativa (H_1) que afirma lo contrario.

• R^2 para la regresión logística

En la regresión logística, una aproximación equivalente al R^2 de la regresión lineal, es el *Pseudo–*R^2 y su forma de calcularlo es la siguiente:

$$Pseudo - R^2 = \frac{G}{G+N}$$

donde G es el estadístico χ^2 del modelo y N es el tamaño de la muestra. Su valor está comprendido entre [0,1]. Cuanto más se aproxime a 0 su valor significará que el ajuste del modelo es muy bajo; y cuando los valores se acerquen a 1 representará todo lo contrario.

Sin embargo, se suelen utilizar con mayor frecuencia dos estadísticos que no son más que una versión de esta idea básica: Pseudo-R^2 de Cox y Snell; y el Pseudo-R^2 de Nagelkerke —que es una modificación del anterior que resuelve su principal deficiencia, que no alcanza el valor máximo 1 ni siquiera ante un modelo perfecto—. Su cálculo se lleva a cabo según las siguientes expresiones:

$$Pseudo - R^2(Cox\_y\_Snell) = 1 - \left[\frac{-2LL0}{-2LL1}\right]^{2/n}$$

$$Pseudo - R^2(Nagel\,\mathrm{ker}\,ke) = \frac{1 - \left[\frac{-2LL0}{-2LL1}\right]^{2/n}}{1 - \left(-2LL0\right)^{2/n}}$$

• Prueba de HOSMER y LEMESHOW

HOSMER y LEMESHOW (1989: 140-145) proponen una variante del estadístico de Pearson para comprobar la bondad de ajuste del modelo. Lo que se plantea es subdividir la muestra en varios grupos de aproximadamente el mismo tamaño, normalmente 10 grupos. Posteriormente se procede a la imputación de los casos a los grupos en función de su probabilidad estimada de ocurrencia del evento que se analice. Finalmente se elabora una tabla de contingencia con las frecuencias observadas (*fo*) y esperadas (*fe*), donde el χ^2 se calcula a partir de la suma de estos valores (CEA D'ANCONA, 2002: 166):

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

Con *g*–2 grados de libertad, siendo *g* el número de grupos, por lo que habitualmente serán 8 grados de libertad.

La H_0 es que no existan diferencias entre los valores observados y los pronosticados a partir del modelo de regresión logística, frente a la hipótesis alternativa (H_1) que afirma lo contrario. Por tanto, lo que realmente nos interesa es *no* rechazar la H_0, por lo que su grado de significatividad nos interesaría que fuera superior al 5 por 100.

b) *Capacidad predictiva del modelo*

Otro aspecto relevante es el relativo a que la regresión logística es una técnica predictiva (CEA D'ANCONA, 2002: 171). Por ello, debemos evaluar la eficacia de las predicciones a través de la tabla de clasificación.

• Tabla de clasificación

En esta ocasión vamos a proceder a tratar de averiguar cuál es la capacidad predictiva del modelo. Para ello compararemos los valores reales de la variable dependiente con los valores estimados por el modelo. Así, al compararlos podremos ver cuál es el porcentaje de datos correctamente clasificados y cuál es el de incorrectos.

3.3.3. Regresión logística multinomial

3.3.3.1. *Delimitación de la técnica estadística*

La regresión logística multinomial múltiple es un tipo de técnica de análisis multivariante en la que la variable explicada tiene más de dos categorías, en nuestro caso tres —ECI Minoritaria, ECI Mayoritaria y Filial de Plena Propiedad—. Así pues, se trata de predecir la probabilidad de pertenencia a una de las tres categorías de la variable dependiente a partir de las variables explicativas, que pueden ser tanto categóricas como continuas.

La expresión de la probabilidad condicional de la ocurrencia del suceso se denota a través de dos funciones, ya que al ser una variable dependiente tricotómica una de sus categorías servirá de referencia ($Y=0$). Así obtendremos una función que compare el *logit* de $Y = 1$ *versus* $Y = 0$, y otra función que compare el *logit* de $Y = 2$ *versus* $Y = 0$. La comparación de los grupos $Y = 1$ *versus* $Y = 2$ se podrá obtener de la diferencia de las funciones anteriores.

$$g_1(x) = Ln\left[\frac{P(Y=1|x)}{P(Y=0|x)}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + ... + \beta_{1p}x_p$$

y

$$g_2(x) = Ln\left[\frac{P(Y=2|x)}{P(Y=0|x)}\right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + ... + \beta_{2p}x_p$$

La expresión de la probabilidad condicional de la ocurrencia del suceso se denotará de la siguiente manera:

$$P(Y=0|x) = \frac{1}{1+e^{g_1(x)}+e^{g_2(x)}} \qquad P(Y=1|x) = \frac{e^{g_1(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

$$P(Y=2|x) = \frac{e^{g_2(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

Por lo que la expresión general para la probabilidad condicional de las tres categorías del modelo será la siguiente:

$$P(Y=j|x) = \frac{e^{g_j(x)}}{\sum_{k=0}^{2} e^{g_k(x)}} \quad \text{donde} \quad g_0(x) = 0$$

a) Estimación de los coeficientes

El método utilizado para la estimación de los coeficientes es el mismo que en la regresión logística binomial. HOSMER y LEMESHOW (1989: 218) recomiendan generar tres variables binarias codificadas como 0 y 1 para indicar el grupo de pertenencia de la observación. Su codificación será la siguiente:

$$\text{Si } Y=0 \text{ entonces } \begin{cases} Y_{00}=1 \\ Y_{01}=0 \\ Y_{02}=0 \end{cases} \text{; si } Y=1 \begin{cases} Y_{10}=0 \\ Y_{11}=1 \\ Y_{12}=0 \end{cases} \text{; y si } Y=2 \begin{cases} Y_{20}=1 \\ Y_{21}=0 \\ Y_{22}=1 \end{cases}$$

Así, la función de verosimilitud quedaría como se muestra a continuación:

$$\iota(\beta) = \prod_{i=1}^{n} \left[\pi_0(x_i)^{y_{0i}} \pi_1(x_i)^{y_{1i}} \pi_2(x_i)^{y_{2i}} \right]$$

Y si tomamos logaritmos y tenemos en cuenta que $\sum_j Y_{ij} = 1$, la función de log-verosimilitud será:

$$L(\beta) = \ln\left[\iota(\beta) \right] = \sum_{i=1}^{n} y_1 g_1(x_i) + y_2 g_2(x_i) - \ln(1 + e^{g_1(x_i)} + e^{g2(x_i)})$$

Finalmente, y al igual que en la regresión logística binaria, los estimadores de máxima verosimilitud serán aquellos que maximicen el valor de esta función.

b) *Significatividad e interpretación de los coeficientes*

La forma de estimar si los coeficientes son significativos o no y la manera de interpretarlos es la misma que en el caso de la regresión logística binomial.

3.3.3.2. *Adecuación del modelo*

Al igual que se hizo en el modelo de regresión logística binomial, debe estudiarse la bondad de ajuste global del mismo y, por otra, su capacidad predictiva.

a) *Bondad de ajuste del modelo global*

El primer aspecto a tratar es la comprobación del ajuste de ambos modelos. Para ello vamos a proceder a comprobar el ajuste global del mismo a través de diferentes estadísticos.

• Razón de Verosimilitud: –2LL

Este estadístico ayuda a medir el efecto conjunto de las variables predictoras en la probabilidad de Y, y lo hace mediante la comparación de dos logaritmos –2LL para un modelo con ninguna variable indepen-

diente ($L0$) y –2LL para otro modelo con todas las variables explicativas ($L1$). De manera tal que:

$$-2LL = -2\log\left(\frac{L0}{L1}\right) = \left\{-2\log(0) - \left[-2\log(1)\right]\right\} = -2LL0 - (-2LL1)$$

El motivo por el cual se utiliza –2LL es porque la verosimilitud suele ser un valor inferior a 1, así conseguimos que se aproxime a la distribución χ^2_k (donde k es el número de variables independientes) lo que facilita la comprobación de su significatividad.

El estadístico χ^2_k contrasta la hipótesis nula (H_0) de que todos los coeficientes del modelo, excepto la constante, son iguales a cero, frente a la hipótesis alternativa (H_1) que afirma lo contrario.

- R_2 para la regresión logística

La explicación sobre estos estadísticos ya fue realizada en el modelo de regresión logística binaria.

b) *Capacidad predictiva del modelo*

Como ya señalamos anteriormente, debemos evaluar la eficacia de las predicciones a través de la tabla de clasificación o de confusión.

3.3.4. Regresión logística ordinal

3.3.4.1. *Delimitación de la técnica estadística*

La regresión logística ordinal se utiliza cuando la variable dependiente muestra más de dos categorías y, entre ellas, se presenta una relación jerárquica o de orden. Este análisis implica minimizar las diferencias de la suma de los cuadrados entre la dependiente y una combinación ponderada de las variables independientes. Los coeficientes estimados expresan cómo los cambios en las variables predictoras afectan a la variable explicada. Por tanto, su funcionamiento y la forma de ser interpretado son prácticamente iguales a la regresión logística multinomial, a excepción del tratamiento de la variable dependiente.

Así pues, el modelo de regresión latente es $y^* = \beta' x + \varepsilon$, donde y^* no se observa. La diferencia estriba en el comportamiento de lo que sí se observa donde: $y = 0$ si $y^* \leq 0$; $y = 1$ si $0 < y^* \leq \mu_1$; e $y = 2$ si $\mu_1 < y^* \leq \mu_2$.

Por tanto:

$$\begin{aligned} P(y = 0 \mid x) &= \Lambda(-\beta' X_i) \\ P(y = 1 \mid x) &= \Lambda(\mu_1 - \beta' X_i) - \Lambda(-\beta' X_i) \\ P(y = 2 \mid x) &= \Lambda(\mu_2 - \beta' X_i) - \Lambda(\mu_1 - \beta' X_i) \end{aligned}$$

Para que todas las probabilidades sean positivas se debe cumplir que $0 < \mu_1 < \mu_2$, donde μ_i es un parámetro que representa el valor del umbral y se estima a la vez que β, y $\Lambda(\beta' X_i)$ representa la función de distribución logística.

a) *Estimación, significatividad e interpretación de los coeficientes*

La forma de estimar si los coeficientes son significativos o no y la manera de interpretarlos es la misma que en los anteriores modelos.

3.3.4.2. *Adecuación del modelo*

Tanto la bondad de ajuste del modelo como su capacidad predictiva siguen las mismas pautas que en el modelo de regresión logística multinomial.