3

Fuentes estadísticas y metodología del análisis empírico

3.1. Introducción

El presente capítulo contempla la consecución de tres objetivos principales. De un lado, se efectúa una descripción minuciosa de las fuentes estadísticas utilizadas en la parte empírica de esta tesis doctoral. En segundo lugar, se presenta la metodología empleada para, a continuación, pasar a exponer los modelos econométricos desarrollados en los capítulos 4, 5 y 6 de esta investigación.

Por consiguiente, el capítulo se estructura en tres partes diferenciadas. En la primera parte se presenta la base de datos que permitirá la realización del estudio empírico en los capítulos siguientes, aludiendo a sus principales ventajas y a sus limitaciones con respecto a otras fuentes estadísticas disponibles. En la segunda parte, se plantean las principales hipótesis a contrastar, así como las variables dependientes e independientes incluidas en el análisis empírico. Por último, la tercera parte está dedicada a la exposición de las principales características de los modelos econométricos empleados con el fin de contrastar las distintas hipótesis seleccionadas.

3.2. Descripción del Módulo de Transición de la Educación al Mercado Laboral (EPA, 2000)

Desde el año 1999, el cuestionario general de los segundos trimestres de la Encuesta de Población Activa[66] incluye un módulo específico destinado a profundizar en determinados aspectos del mercado de trabajo que resultan de especial interés[67]. Los datos utilizados para llevar a cabo el análisis empírico desarrollado en los tres siguientes capítulos corresponden al módulo *ad hoc*

---

66 La Encuesta de Población Activa es una encuesta continua realizada con periodicidad trimestral por el INE y cuya finalidad principal consiste en proporcionar datos de las principales categorías poblacionales en relación con el mercado de trabajo. Se trata de una encuesta por muestreo, que abarca todo el territorio nacional y está dirigida a la población que reside en las viviendas familiares, es decir, las utilizadas durante todo o la mayor parte del año como vivienda habitual y permanente. La muestra final se eleva a 65.000 viviendas al trimestre y se entrevista a todas las personas que residen en el hogar.

67 El diseño muestral de estos módulos coincide con el de la EPA y el contenido de los mismos se realiza en coordinación con las Encuestas de Fuerza de Trabajo de otros países de la Unión Europea. En cuanto a la temática de estos módulos, el correspondiente al año 1999 estuvo dedicado al análisis de los accidentes y enfermedades laborales, el módulo de 2001 trata de las relaciones laborales especiales y de condiciones de horarios de trabajo y el módulo de 2002 analiza la situación de las personas discapacitadas con respecto al empleo.

de la Encuesta de Población Activa del segundo trimestre de 2000, que se denomina Módulo de Transición de la Educación al Mercado Laboral.

El Módulo de Transición de la Educación al Mercado Laboral es una ampliación del cuestionario habitual de la Encuesta de Población Activa, al que únicamente responden aquellos individuos que, teniendo una edad comprendida entre los 16 y los 35 años en el momento de realizarse la encuesta, hayan finalizado, abandonado o interrumpido durante más de un año sus estudios o formación iniciales, en la etapa comprendida entre el año 1991 y el segundo trimestre del año 2000[68].

Este módulo específico introduce una serie de cuestiones relacionadas con la incorporación de los jóvenes al mercado de trabajo tras haber finalizado su formación, con objeto de proporcionar información detallada acerca de diversos aspectos vinculados al proceso de transición desde la educación al mercado laboral.

En primer lugar, la encuesta facilita información minuciosa sobre el nivel de estudios alcanzado por los individuos en el momento en que salen del sistema educativo, así como del sector de estudios al que pertenece su titulación. En este sentido, es importante resaltar el hecho de que la información correspondiente al nivel educativo de los individuos contenida en el módulo se ajusta a la configuración actual de los niveles de enseñanza del sistema educativo español[69].

Por otra parte, también es posible obtener datos acerca de si el individuo ha estado buscando empleo una vez ha abandonado el sistema educativo y, en caso afirmativo, cuál fue la duración del proceso de búsqueda. Esta última cuestión resulta de particular relevancia en nuestro trabajo ya que permite estudiar el efecto de la longitud del periodo de búsqueda de desempleo en el proceso de inserción laboral de los individuos.

En tercer lugar, el cuestionario incluye preguntas referentes a si el joven ha obtenido un empleo significativo, esto es, aquel empleo con una duración mayor o igual a seis meses y con una jornada laboral de al menos veinte horas semanales[70]. Si la respuesta es afirmativa, se indaga acerca de algunos de los rasgos más relevantes de dicha ocupación. Estos datos se convierten en las cuestiones centrales de nuestro trabajo, en el momento en que se aborda, tanto la inserción laboral de los jóvenes en términos cuantitativos, como cuando se analizan los aspectos cualitativos relacionados con el primer empleo significativo.

Además, se cuenta con toda la información disponible referente a las características personales y familiares de los individuos que recoge el cuestionario general de la Encuesta de Población Activa del segundo trimestre de 2000. A partir de este cuestionario habitual pueden conocerse variables como la edad del individuo, la Comunidad Autónoma en que éste reside, su nacionalidad, la composición familiar, el nivel educativo y la situación laboral de los padres, entre otras.

La principal ventaja que puede atribuirse a la utilización del módulo específico de la Encuesta de Población Activa de 2000 con respecto a otras fuentes estadísticas en el estudio de la transición de los jóvenes desde el sistema educativo al mercado de trabajo es que dicho módulo permite la obtención de información amplia y actualizada acerca del proceso de inserción laboral de los jóvenes a lo largo de la década de los noventa.

---

68 En el Módulo de Transición de la Educación al Mercado Laboral (EPA, 2000), se considera educación inicial todo estudio o formación que se haya realizado desde el nivel primario y sin interrupciones de más de un año de duración, sin tener en cuenta aquellas interrupciones que se hayan producido como consecuencia de maternidad o paternidad, enfermedad grave, realización del servicio militar o prestación social sustitutoria o el haber estado a la espera de un diploma para acceder a estudios de mayor nivel. Por otra parte, en el concepto de formación se incluyen tanto los estudios generales como los profesionales, sean o no reglados y a tiempo completo y parcial. También se considera como formación inicial la preparación de oposiciones.

69 La adecuación de la clasificación de los niveles de estudio contemplados en la EPA con los distintos niveles de enseñanza actualmente impartidos en nuestro país se llevo a cabo en el año 2000. Anteriormente, la información disponible en la EPA no permitía la posibilidad de distinguir el sector de estudio cursado por el individuo, ni diferenciar los estudios de formación profesional medios y superiores.

70 Como afirman Albert *et al.* (2003b), esta definición de empleo significativo es muy relevante para el caso de España, donde la contratación temporal es un fenómeno frecuente tanto entre los trabajadores jóvenes como adultos. También hay que puntualizar que quedan excluidos los trabajos ocasionales, el servicio militar obligatorio, la prestación social sustitutoria de esta definición, así como los posibles empleos significativos que tuvieran lugar antes de finalizar, abandonar o interrumpir por primera vez los estudios o formación iniciales (INE, 2001).

En nuestro país ha existido una limitada disponibilidad de encuestas cuyo principal objetivo fuese analizar este fenómeno. Para realizar este tipo de aproximaciones, los investigadores han utilizado, bien encuestas de carácter general (Encuesta de Población Activa, Encuesta Sociodemográfica, Panel de Hogares de la Unión Europea, etc.), que no están exclusivamente centradas en el colectivo juvenil, o bien otras encuestas más específicas que sólo consideran a jóvenes con un determinado nivel educativo o que habitan en una región concreta[71].

La elaboración del módulo *ad hoc* de la Encuesta de Población Activa del segundo trimestre de 2000 supone el fin de las limitaciones señaladas anteriormente. Como afirman Albert et al. (2003a), la principal novedad de esta fuente estadística es que aporta información acerca de dos hechos transcendentales en el ciclo vital de los individuos: el abandono del sistema educativo y la obtención del primer empleo significativo. Asimismo, dicha base de datos combina las ventajas del gran tamaño muestral de la Encuesta de Población Activa con información detallada de la transición de los jóvenes al mundo laboral (Krogan y Müller, 2002).

Además, el Módulo de Transición de la Educación al Mercado Laboral cuenta con otras dos características dignas de tenerse en cuenta. Primeramente, la información proporcionada es relativamente comparable con otros veinte países europeos: los antiguos Estados miembros de la Unión Europea, con la excepción de Alemania, cinco de los países de reciente adscripción a la Unión Europea (Hungría, Eslovaquia, Eslovenia, Lituania y Letonia) y Rumania[72]. Así, diversos trabajos han realizado estudios comparativos de la transición desde la educación al mercado laboral entre países europeos, tomando como fuente estadística los datos de este cuestionario en los distintos países participantes (Gang, 2002; Ianelli, 2002b; Smyth, 2002).

Por otra parte, el módulo específico de la Encuesta de Población Activa del segundo trimestre de 2000 proporciona un cierto enfoque longitudinal y retrospectivo en cuanto a las trayectorias laborales de los individuos, lo que otorga la posibilidad de profundizar en ciertos rasgos del mercado de trabajo durante las primeras etapas laborales de los jóvenes.

No obstante, además de las ventajas señaladas, la fuente estadística aquí descrita adolece de ciertas limitaciones. En primer lugar, existe una falta de sincronización entre las variables recogidas en el módulo, que hacen referencia al momento en que el individuo salió del sistema educativo y obtuvo su primer empleo (en el caso de que lo haya obtenido), y las características personales y familiares contenidas en el cuestionario habitual de la Encuesta de Población Activa, que se refieren al momento en que se realizó la entrevista (segundo trimestre del año 2000).

Por otra parte, tanto la Encuesta de Población Activa como el Módulo de Transición de la Educación al Mercado Laboral carecen de información relativa a los ingresos o rentas procedentes del trabajo que obtienen los individuos, que, en este caso, serían de gran utilidad para abordar el estudio de determinadas características cualitativas del primer empleo de los jóvenes egresados del sistema educativo.

A pesar de estas limitaciones, la calidad de los resultados derivados de la implantación del módulo específico de la Encuesta de Población Activa de 2000 en España ha propiciado la decisión de utilizar en el desarrollo empírico de la tesis doctoral los datos procedentes de la base de datos mencionada[73].

---

71 Con respecto a los estudios centrados en un determinado nivel educativo, puede destacarse el elaborado por García-Montalvo (2001) con una muestra de titulados universitarios. Por otra parte, García-Montalvo y Peiró (2000) se ocupan del estudio de la transición de los jóvenes residentes en de la Comunidad Valenciana, mientras que García- Espejo (1998) desarrolla su análisis para la región asturiana. Finalmente, existe un tercer tipo de estudios que se centran en un determinado nivel educativo y en una región concreta (Salas, 1999; Sáez y Rey, 2000; González-Betancor, 2003).

72 Lamentablemente, los datos no son estrictamente comparables, ya que, durante el desarrollo de la encuesta, algunos países no siguieron de forma exacta las indicaciones proporcionadas por Euroestat. Los estudios dedicados a evaluar la implementación del módulo (Ianelli, 2002a; Müller *et al.*, 2002) ponen de manifiesto que algunas naciones no han aplicado las definiciones indicadas por Euroestat mientras que, en otros casos, se eliminaron del cuestionario las variables que hacen referencia al entorno familiar del individuo.

73 De acuerdo con Smyth (2002) y Ianelli (2002a), España resulta ser uno de los países que con mayor rigor han implementado el módulo de transición, destacando por la cantidad y calidad de sus datos. Por otro lado, el interés de este tipo de datos, así como su utilidad para desarrollar análisis comparativos acerca de la transición de los jóvenes al mundo laboral en los países europeos, ha motivado la decisión de que el módulo vuelva a ser replicado en el año 2006.

3.3. Selección de la muestra, hipótesis a contrastar y definición de las variables utilizadas en el análisis

De los 180.853 encuestados en la Encuesta de Población Activa del segundo trimestre del año 2000, aproximadamente un 30% (53.918 individuos) tenían una edad comprendida entre los dieciséis y los treinta y cinco años, ambos inclusive. Dentro de este último colectivo, 15.009 jóvenes habían interrumpido, abandonado o finalizado sus estudios durante al menos un año entre 1991 y 2000, y constituyen el conjunto de individuos que contestan al módulo específico de transición al mercado laboral. En este trabajo, se han excluido del análisis aquellos individuos que salieron del sistema educativo en el año 2000, ya que sólo se dispone de información hasta la fecha de realización de la encuesta (segundo trimestre del año 2000) por lo que los datos para este año son escasos (sólo 333 individuos) y podrían presentar sesgos[74]. Por tanto, la muestra general estaría compuesta por 14.676 individuos que representan el 97,8% del total de la población que contesta al módulo de transición.

A continuación, los siguientes apartados están destinados a presentar las distintas hipótesis que pretenden ser contrastadas en el análisis empírico de la presente investigación, así como a realizar una descripción del conjunto de variables, tanto dependientes, como independientes, que se utilizan para contrastarlas.

3.3.1. Hipótesis a contrastar y variables dependientes

Como ya se adelantó en el capítulo introductorio, la tesis doctoral intenta ofrecer evidencia empírica acerca de tres fenómenos concretos. En primer lugar, en el capítulo 4 se pretenderá demostrar que el hecho de que un joven realice estudios universitarios se ve determinado por factores familiares, socioeconómicos y culturales. Para contrastar esta hipótesis se ha creado la variable dependiente "*completar estudios universitarios*" como una variable dicotómica que toma valor 1 para individuos que salgan del sistema educativo una vez alcanzados los estudios universitarios, y valor 0 para aquéllos cuyo nivel de estudios sea inferior[75].

En segundo lugar, el capítulo 5 trata de contrastar si la posesión de un título universitario sigue proporcionando a los individuos ciertas ventajas en su inserción laboral, concretamente en la obtención del primer empleo significativo. Con objeto de contrastar esta hipótesis se ha generado la variable "*logro de empleo significativo*" que toma valor 1 en el caso que el encuestado haya conseguido un empleo significativo, y valor 0 en caso contrario[76].

Por último, en este trabajo se pretende indagar no sólo en las oportunidades de los universitarios de obtener un empleo con respecto al resto de los jóvenes, sino también en la adecuación del empleo obtenido al nivel educativo que ostentan (ver el capítulo 6). Para este fin, se ha creado la variable ficticia "*adecuación al puesto de trabajo*" que toma tres posibles valores: valor 0, cuando el empleo obtenido se ajusta al nivel de estudios alcanzado; valor 1, en el caso que el individuo esté infraeducado en su primer puesto de trabajo y valor 2, cuando el individuo está sobreeducado en su primer empleo[77].

---

74 Se considera que pueden existir dos tipos de sesgos. En primer lugar, al realizarse las entrevistas entre los meses de abril y junio de 2000, éstas pueden no recoger adecuadamente el porcentaje de individuos que abandonan el sistema educativo en este año, ya que la mayoría de salidas de la educación se producen en los meses de junio y septiembre. Por otra parte, en el estudio de la inserción laboral de los individuos, debe tenerse en cuenta que los jóvenes que han salido del sistema educativo en el año 2000 apenas han tenido tiempo para encontrar un empleo, lo que puede comprometer la validez de los resultados. Por ese motivo, se ha considerado conveniente excluir del análisis a los jóvenes que salen del sistema educativo en 2000, decisión metodológica que también aparece en los trabajos de Albert *et al.* (2003a, 2003b), Corrales y Rodríguez (2003) y Rodríguez y Corrales (2003).

75 Esta variable dependiente se generó a partir del la variable NFORM de la Encuesta de Población Activa, que indica cuál es el nivel de estudios alcanzado por el entrevistado. Se considera que el individuo ha completado estudios universitarios si NFORM toma los valores 54 (enseñanzas universitarias de primer ciclo y equivalentes o personas que han aprobado tres cursos completos de una licenciatura o créditos equivalente), 55 (Enseñanzas universitarias de primer y segundo ciclo, de sólo segundo ciclo y equivalentes); 56 (Programas especialización profesional) ó 61 (Doctorado universitario).

76 La variable "empleo significativo" se creó a partir de pregunta M7 del módulo *ad hoc* de la EPA de 2000, en la que se pregunta al encuestado si tras interrumpir, abandonar o finalizar sus estudios ha encontrado un empleo significativo. La variable toma valor 1 si la respuesta del individuo a esta cuestión es afirmativa, y valor 0 en el caso de que sea negativa.

77 La medición del desajuste educativo puede hacerse a través de diferentes fórmulas que aparecen detalladas en el capítulo 6 de la presente tesis doctoral.

En el cuadro 3.1 se resumen las principales hipótesis planteadas en este trabajo, así como el capítulo en que serán abordadas y el modelo econométrico utilizado para su contrastación.

Cuadro 3.1. Resumen de las hipótesis a contrastar en el análisis empírico

Hipótesis a contrastar	Capítulo	Modelo econométrico
H1: El hecho de haber alcanzado estudios universitarios depende de factores económicos, sociales y culturales	Capítulo 4	Modelo *logit*
H2: Poseer un título universitario confiere a los individuos ventajas en términos de obtención del primer empleo significativo	Capítulo 5	Modelo *logit*
H3: El fenómeno de la sobreeducación en el primer empleo incide de forma más acusada entre los jóvenes universitarios	Capítulo 6	Modelo *logit multinomial* con sesgo de selección

Fuente: Elaboración propia

3.3.2. Variables independientes

En cuanto a las variables independientes que serán utilizadas para contrastar las tres hipótesis anteriormente planteadas, éstas se pueden ordenar en función de cinco categorías fundamentales: las características personales, las características familiares, los aspectos relacionados con el proceso de búsqueda de empleo, los principales rasgos del empleo conseguido y, finalmente, los factores de entorno. La utilización de unas u otras en los diferentes análisis empíricos realizados en capítulos posteriores dependerá de su oportunidad, teniendo en cuenta el fenómeno que se pretende explicar y la aproximación econométrica seleccionada.

3.3.2.1. Características personales[78]

Dentro de las características personales que serán incluidas como variables en los distintos modelos empíricos pueden destacarse las siguientes:

- Género masculino

  Se trata de una variable dicotómica que toma valor 1 en el caso de que el individuo sea varón y valor 0 cuando es mujer.

- Edad al salir del sistema educativo

  Es una variable discreta que recoge la edad del individuo en el momento de abandonar el sistema educativo.

- Edad en el momento de obtener un empleo significativo

  Es una variable discreta que recoge la edad que tenía el individuo en el momento en que consiguió su primer empleo significativo.

- Nacionalidad española

  Se construye una variable ficticia que toma valor 1 en el caso que el encuestado sea español y 0 en caso contrario.

- Nivel de estudios alcanzado

  La Encuesta de Población Activa desagrega el nivel educativo de los individuos en dieciocho posibles categorías. En aras de adoptar dicha clasificación a la configuración actual de nues-

78 La mayoría de las variables recogidas en este apartado han sido obtenidas directamente del cuestionario habitual de la Encuesta de Población Activa del segundo trimestre de 2000, excepto la edad en el momento de salida del sistema educativo y la edad en el momento de obtener el primer empleo, que resultan de restar de la edad actual, los años que han pasado desde aquéllos momentos hasta la fecha de la entrevista.

tro sistema educativo se ha procedido a la creación de las siguientes categorías: estudios primarios, primera etapa de estudios secundarios, bachillerato, formación profesional (en la que se distingue la formación profesional de grado medio y la de grado superior) y estudios universitarios (distinguiendo los que completan titulaciones de ciclo corto o de ciclo largo)[79]. En todos los casos, se trata de variables dicotómicas que toman valor 1 cuando el nivel de estudios del individuo coincida con el de la categoría y valor 0, en caso contrario.

- Sector de estudios

Desde el año 2000, la Encuesta de Población Activa recoge información acerca del sector de estudios al que pertenece la titulación del individuo, estableciendo veintiséis categorías diferentes[80]. En este trabajo se ha realizado una agregación mediante la cual el análisis se realiza considerando siete posibles sectores de estudio, intentado que la clasificación se asemejase lo más posible a las ramas de enseñanza que existen en nuestro país y que el número de individuos en cada categoría resultase significativo[81]. En concreto, la clasificación adoptada se resume en las siguientes categorías: estudios básicos y personales, Humanidades, Ciencias Sociales, Enseñanzas Técnicas, Ciencias de la Salud y Ciencias y otros sectores de estudio.

3.3.2.2. Características familiares[82]

Entre las características familiares recogidas en el análisis empírico pueden destacarse las siguientes:

- Estudios de los padres

El Módulo de Transición de la Educación al Mercado Laboral establece cuatro categorías en el nivel de estudios de los padres: estudios primarios o inferiores, educación secundaria obligatoria o equivalente, bachillerato o equivalente y estudios superiores o equivalentes[83]. Estas variables se incorporarán en el análisis empírico mediante la construcción de cuatro variables ficticias, teniendo en cuenta, tanto el nivel de estudios del padre, como el de la madre.

- Situación laboral de los padres

A través del cuestionario habitual de la Encuesta de Población Activa puede conocerse información de la situación laboral de los padres del individuo (ocupado, desempleado, inactivo). A partir de esa información, se han generado tres variables ficticias que recogen estos tres posibles estados de los padres en cuanto a su situación laboral. Como en el caso anterior, se considera esta información, tanto para el padre, como para la madre.

- Situación socioeconómica de los padres

En el caso de que el padre o la madre del individuo se encuentre ocupado, a través del cuestionario de la Encuesta de Población Activa es posible conocer qué tipo de actividad realiza. La Encuesta de Población Activa estructura las ocupaciones de los individuos según los códi-

---

79 La tabla A.1.1. del Anexo 1 recoge la equivalencia entre las categorías de niveles de estudio que aparecen en la Encuesta de Población Activa y las empleadas en esta investigación.

80 Estas categorías responden a la clasificación de estudios realizada por Euroestat, en virtud de poder armonizar los diferentes sectores de estudios existentes en los distintos países de la Unión Europea.

81 La equivalencia entre los sectores de estudio definidos en la Encuesta de Población Activa y los considerados en este trabajo se presenta en la tabla A.1.2. del Anexo 1.

82 Todas las variables que aparecen en este apartado han sido tomadas del cuestionario habitual de la Encuesta de Población Activa del segundo trimestre de 2000, al carecer el módulo de información a este respecto. Si bien estas variables corresponden al año 2000 y no al momento en el que el joven sale del sistema educativo, se estima conveniente incorporarlas en el análisis por su alto poder explicativo de los fenómenos objeto de estudio. Además, consideramos que no resulta ilógico suponer que, para padres de individuos entre 16 y 35 años, dichas características permanecerán razonablemente estables. De hecho, la mayoría de investigaciones basadas en esta fuente estadística para España (Corrales y Rodríguez, 2003; Albert et al. 2003a; Albert et al. 2003b; Congregado y García, 2002) así como otras realizadas para el ámbito internacional (Ianelli, 2002b; Gang, 2002) hacen uso de la información del cuestionario habitual en su análisis.

83 El módulo específico de la EPA de 2000 sólo aporta información acerca de la educación de los padres en los casos en que éstos no hayan sido encuestados en la vivienda (preguntas M13 y M14 del módulo). En el resto de los casos, la información sobre el nivel educativo de los padres se obtiene a partir del cuestionario habitual de la Encuesta de Población Activa del segundo trimestre de 2000. La correspondencia entre los niveles educativos de los padres que aparecen en la Encuesta de Población Activa y los que se utilizan en este trabajo está recogida en la tabla A.1.3. del Anexo 1.

gos a dos dígitos de la Clasificación Nacional de Ocupaciones de 1994 (CNO-94). Así, existen sesenta y seis categorías ocupacionales diferentes que se han agrupado en cinco epígrafes distintos: directivos y gerentes de empresas; profesionales científicos e intelectuales, técnicos y profesionales de apoyo; empleados administrativos y trabajadores de los servicios; trabajadores cualificados y, por último, trabajadores no cualificados[84].

- Número de hermanos menores de dieciséis años en el hogar

Para introducir esta información en el análisis empírico se han construido tres variables ficticias; una para cuando el encuestado no tenga hermanos menores de dieciséis años, otra cuando tiene un hermano menor y la última recoge la posibilidad de que el número de hermanos menores de dieciséis años sea mayor o igual a dos. En estos casos, cada una de las variables toma valor 1, cuando coincida con la categoría correspondiente, y valor 0, en caso contrario.

3.3.2.3. Características de la búsqueda de empleo

Dentro de este bloque de características se consideran aquellas variables que indican si el individuo ha buscado empleo desde que salió del sistema educativo y, en caso afirmativo, cuál ha sido la duración del periodo de búsqueda.

- Búsqueda continua de empleo[85]

La variable ficticia toma valor 1 en el caso de que el individuo haya realizado una búsqueda continua de empleo desde que salió del sistema educativo y valor 0 en caso contrario.

- Duración de la búsqueda de empleo[86]

En el caso en que el individuo haya buscado empleo, existen tres diferentes categorías: búsquedas con una duración inferior a seis meses, búsquedas que se prolongan entre seis y once meses y búsquedas superiores a un año. Cada variable toma valor 1, en el caso de que coincida con la categoría correspondiente, y valor 0, en caso contrario.

3.3.2.4. Características del empleo obtenido[87]

En este apartado se recogen las principales características del primer puesto de trabajo, en el caso de que el individuo haya obtenido un empleo significativo.

- Empleo indefinido

En el caso de que el individuo haya encontrado su primer empleo significativo, la variable ficticia toma valor 1, en el caso de que se trate de un empleo indefinido y valor 0 cuando sea un empleo temporal.

- Jornada completa

Si el joven ha obtenido un empleo significativo, se crea una variable ficticia que toma valor 1, si el individuo trabaja a jornada completa y valor 0 cuando trabaja a tiempo parcial.

- Sector público

En el caso de que el joven consiga un empleo significativo, se genera una variable ficticia que toma valor 1 en el caso de que el individuo trabaje en el Sector Público y 0 cuando lo hace en el sector privado.

---

84 La equivalencia entre esta ordenación y la Clasificación Nacional de Ocupaciones a dos dígitos se encuentra en la tabla A.1.4 del Anexo 1.

85 Esta variable ha sido creada a partir de la pregunta M11 del módulo especifico de la Encuesta de Población Activa del segundo trimestre de 2000. Este módulo define la búsqueda continua de empleo como aquélla que los entrevistados realizaron, una vez habían salido del sistema educativo, de forma activa e ininterrumpida durante al menos un mes.

86 La información a este respecto esta tomada de la pregunta M12 del módulo.

87 La información acerca del tipo de contrato, la jornada laboral, si se trabaja en el sector privado o en el Sector Público, y sobre el sector de la economía al que pertenece su puesto de trabajo se han obtenido del cuestionario habitual de la Encuesta de Población Activa del segundo trimestre de 2000. También se debe puntualizar que esta información sólo es relevante en el caso que el individuo obtenga un empleo significativo, por lo que únicamente será utilizada en el capítulo 6, en el que se analiza la inserción laboral de los jóvenes desde un punto de vista cualitativo.

■ Sector económico

Atendiendo a la Clasificación Nacional de Actividades Económicas de 1993 (CNAE-93), se ha procedido a agrupar las actividades económicas en cuatro categorías, que coinciden con la tradicional división de actividades por sectores de la economía española. Las categorías generadas son, por tanto, el sector de la agricultura, el sector industrial, el sector de la construcción y, por último, el sector servicios[88]. Para cada una de estas cuatro posibles categorías se ha generado una variable ficticia que toma valor 1, en el caso de que el sector de ocupación del individuo coincida con la categoría, y valor 0 en caso contrario.

3.3.2.5. Características de entorno

Las características de entorno a las que aludiremos en el análisis empírico son:

■ Área geográfica de residencia

Otra de las variables que puede resultar relevante en el estudio analítico es la región de residencia de los individuos. A este respecto se han creado diecisiete variables ficticias que recogen las Comunidades Autónomas españolas en las que el individuo puede residir[89].

■ Año en que el individuo sale del sistema educativo[90]

Se han creado nueve variables ficticias que recogen el año en que el individuo abandona el sistema educativo.

■ Tasa de crecimiento del empleo

Variable continua que muestra la tasa de crecimiento del empleo en el año que el individuo sale del sistema educativo.

■ Tasa de desempleo

Variable continua que muestra la tasa desempleo en el año que el individuo sale del sistema educativo.

Los cuadros 3.2 y 3.3. recogen, respectivamente, las variables dependientes y explicativas utilizadas en los diversos análisis empíricos desarrollados en este trabajo.

Cuadro 3.2. Descripción de las variables dependientes utilizadas en el análisis

Variables	Descripción
Finalización de estudios universitarios	Valor 1: el individuo finaliza los estudios universitarios antes de salir del sistema educativo. Valor 0: el individuo abandona el sistema educativo antes de completar estudios universitarios.
Obtención de un empleo significativo	Valor 1: el individuo ha obtenido un empleo significativo tras abandonar, interrumpir o finalizar sus estudios. Valor 0: el individuo no ha obtenido un empleo significativo.
Adecuación educación-empleo	Valor 0: el individuo está adecuado en su primer empleo significativo. Valor 1: el individuo está infraeducado en su primer empleo significativo. Valor 2: el individuo está sobreeducado en su primer empleo significativo.

Fuente: Elaboración propia

88 La equivalencia entre la CNAE-93 y la agregación empleada en nuestro estudio se presenta en la tabla A.1.5. del Anexo 1.

89 Esta variable está tomada del cuestionario habitual de la Encuesta de Población Activa del segundo trimestre de 2000, que recoge la Comunidad Autónoma en que reside el individuo en el momento de efectuarse la encuesta. Sería deseable conocer cuál era el lugar en que vivía el encuestado cuando en la fecha en que interrumpió, abandonó o finalizó sus estudios pero, lamentablemente, el módulo no proporciona información a este respecto. A pesar de este inconveniente, se han incorporado al análisis las variables representativas de las distintas Comunidades Autónomas, ya que pensamos que no debe existir gran discrepancia entre ambas al darse en España poca movilidad y al tratarse de una definición de área geográfica suficientemente amplia.

90 Dicha información ha sido obtenida de la pregunta M3 del módulo de transición en la que se le pregunta al encuestado el año en que abandono, interrumpió o finalizó sus estudios.

Cuadro 3.3. Descripción de las variables dependientes

Variables	Descripción
Características Personales	
Género	Valor 1: varón. Valor 0: mujer.
Edad al salir del sistema educativo	Edad actual menos el tiempo transcurrido desde que el individuo salió del sistema educativo.
Edad al obtener el primer empleo significativo	Edad actual menos el tiempo transcurrido desde que el individuo logró su primer empleo significativo.
Nacionalidad española	Valor 1: español. Valor 0: no español.
Nivel de estudios alcanzado	Variables ficticias que recogen las siguientes categorías: Estudios primarios. Primera etapa de estudios secundarios. Bachillerato. Formación profesional: • FP de grado medio. • FP de grado superior. Estudios universitarios: • Universitarios ciclo corto. • Universitarios ciclo largo.
Sector de estudios	Variables ficticias que recogen las siguientes categorías: Básicos y personales. Humanidades. Ciencias Sociales. Enseñanzas Técnicas. Ciencias de la Salud. Ciencias y otros sectores de estudio.
Características Familiares	
Nivel de estudios de los padres	Variables ficticias que recogen las siguientes categorías: Estudios primarios o inferiores. Estudios secundarios de primera etapa. Bachillerato o equivalente. Estudios superiores.
Situación laboral de los padres	Variables ficticias que recogen los siguientes epígrafes: Ocupado. Parado. Inactivo.
Ocupación de los padres	Variables ficticias que recogen los siguientes epígrafes: Directivos o gerentes. Profesionales y técnicos. Administrativos y trabajador de los servicios. Trabajadores cualificados. Trabajadores no cualificados.
Número de hermanos menores de 16 años	Variable ficticia dividida en: No tiene hermanos menores de 16 años. Tiene un hermano menor. Tiene dos o más hermanos menores

(continúa)

Cuadro 3.3. Continuación

Variables	Descripción
Características de la búsqueda de empleo	
Búsqueda de empleo	Valor 1: el individuo buscó empleo después de salir del sistema educativo. Valor 0: el individuo no buscó empleo después de salir del sistema educativo.
Duración de la búsqueda de empleo	Variables ficticias con las siguientes categorías: Duración de la búsqueda es menor de 6 meses. Duración de la búsqueda entre 6 y 11 meses. Duración de la búsqueda igual o superior a un año.
Características del empleo obtenido	
Contrato indefinido	Valor 1: el individuo tiene contrato indefinido. Valor 0: el individuo tiene contrato temporal
Jornada parcial	Valor 1: el individuo tiene jornada parcial. Valor 0: el individuo tiene jornada completa.
Sector Público	Valor 1: el individuo trabaja en el sector público. Valor 0: el individuo trabaja en el sector privado.
Sector económico	Variables ficticias que recogen las siguientes categorías: Agricultura. Industria. Construcción. Servicios.
Características de entorno	
Comunidad Autónoma residencia	17 variables ficticias (una para cada Comunidad Autónoma).
Año de salida del sistema educativo	Variables ficticias que recogen el año en que el individuo abandonó, interrumpió o finalizó sus estudios.
Tasa de crecimiento del empleo	Variable continua que recoge la tasa de crecimiento del empleo en el año en que el individuo abandona el sistema educativo.
Tasa de desempleo	Variable continua que recoge la tasa desempleo en el año en que el individuo abandona el sistema educativo.

Fuente: Elaboración propia.

En suma, con la inclusión todas estas variables se pretende dotar del máximo poder explicativo a los modelos econométricos desarrollados en los capítulos siguientes, cuyas principales características se procede a exponer a continuación.

3.4. Los modelos econométricos

Tras presentar las características más destacadas de la fuente de datos, así como enumerar las distintas variables incluidas en el análisis empírico, se procede a realizar una exposición de los principales aspectos de los modelos econométricos desarrollados en los siguientes capítulos.

3.4.1. El modelo *logit*

3.4.1.1. Características y especificación del modelo *logit*

El modelo *logit* pertenece al grupo de modelos de elección discreta o cualitativa. Este tipo de modelos da lugar al desarrollo de análisis empíricos para situaciones en los que la variable

dependiente no es una variable continua (como ocurre en el caso de los modelos de regresión lineal), sino que se trata de una variable dicotómica, es decir, que presenta dos valores que, por convención, son 0 y 1.

Siendo así, el objetivo del modelo *logit* reside en explicar el comportamiento de una variable dicotómica dependiente (*Y*) en función de un grupo de variables explicativas (el vector *X*). De esta forma, el planteamiento de un modelo *logit* permite obtener una función lineal de las variables independientes, que permita clasificar a los individuos en una de las dos subpoblaciones o grupos establecidos por los dos valores de la variable dependiente. Desde el punto de vista analítico, el modelo *logit* se puede representar como sigue.

Sea la variable dicotómica *Y*, que sólo puede tomar los valores 1 (ocurre el suceso) y 0 (no ocurre el suceso). Se puede expresar la probabilidad P_i de que se produzca el evento (*Y*=1), a partir de un conjunto $[x_1,.....,x_K]$ de variables independientes, de la siguiente forma:

$$P_i = E\left(Y = 1/(x_1,.....x_K\right) = \left[\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_K x_K)}}\right] \qquad (3.1.)$$

Si se denota por $\left[z_i = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K\right]$, podemos reescribir la expresión (3.1.) por la (3.2.)

$$P_i = \left[\frac{1}{1+e^{-z_i}}\right] \qquad (3.2.)$$

La expresión anterior proporciona valores continuos de P_i entre 0 y 1 para cualquier x_i, aunque el exponente de la función *e* puede tomar valores entre $\left[-\infty,+\infty\right]$, permitiendo calcular la probabilidad de que un individuo pertenezca a una de las dos poblaciones (Damodar, 1997).

La probabilidad de que no ocurra el suceso (*Y*=0) se puede expresar como:

$$P_i = \left[\frac{1}{1+e^{z_i}}\right] \qquad (3.3.)$$

Dividiendo la expresión la expresión (3.2.) por (3.3.) se obtendrá:

$$\left[\frac{P_i}{1-Pi}\right] = \left[e^{-\beta_0+\beta_1 x_{1i}+...+\beta_K x_K}\right] = e^{z_i} \qquad (3.4.)$$

Operando sobre la ecuación anterior y tomando logaritmos neperianos se obtendrá la expresión:

$$\left[L = Ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_P x_p\right] = e^{z_i} \qquad (3.5.)$$

que permite calcular el logaritmo neperiano de la probabilidad de que ocurra cierto suceso, una vez estimados los coeficientes de regresión logística. Los estimadores obtenidos con la utilización de este modelo son consistentes, esto es, insesgados y asintóticamente eficientes.

En este tipo de modelos, debe de tenerse en cuenta que la interpretación de los parámetros obtenidos en la estimación no es directa, ya que los coeficientes estimados no indican el incremento en la probabilidad, dado el aumento en una unidad en la variable independiente (Cabrer, 2000). La magnitud de la variación en la probabilidad depende del valor concreto que tome la función de distribución, que estará condicionado por la pendiente de la función en cada punto. De esta forma, cuanto más elevada sea dicha pendiente, mayor será el impacto del cambio en el valor de una variable explicativa sobre la probabilidad.

No obstante, el signo de los coeficientes sí que indica perfectamente la dirección del cambio. Así, los valores positivos de los parámetros indican que la probabilidad de que se produzca el suceso aumenta en el modelo previamente especificado. Por el contrario, los valores negativos de los coeficientes indican que la probabilidad de que ocurra el suceso disminuye con las variables planteadas.

La interpretación del modelo *logit* es la siguiente: la tasa de cambio en la probabilidad

$$\left[\frac{\partial P_i}{\partial x_j}\right] \text{ viene determinada por } \beta_j P_i(1-P_i) \qquad (3.6.)$$

donde B_j es el coeficiente del vector *j-ésimo*. Esto significa que, en los modelos *logit*, todos los regresores condicionan el cálculo de los cambios en la probabilidad. La constante es el valor del logaritmo de las probabilidades si el resto de las variables es cero.

Una mención especial merece el caso en que una de las variables explicativas no sea continua, sino que sea una variable ficticia que tan solo tome dos valores, 0 y 1. En este caso el efecto de una variación de la variables X_j sobre la variable dependiente se calcula a través de la diferencia entre los valores obtenidos por:

$$E(Y/Xi) = 1 \text{ y } E(Y/Xi) = 0 \qquad (3.7.)$$

3.4.1.2. Bondad del ajuste del modelo

Si bien en los modelos de regresión lineal el coeficiente de determinación (R^2) es el estadístico más utilizado para valorar la capacidad explicativa del modelo, en el caso de los modelos *logit* han surgido diversas alternativas para medir el grado de ajuste.

En primer lugar, una de las medidas más utilizadas es la que se basa en el logaritmo de la función de máxima verosimilitud, denominada *test de la razón de verosimilitud* o *contraste de la ratio de verosimilitud*. El contraste de la ratio de verosimilitud queda definido como:

$$2\left[LogL(\beta) - Log(\beta_0)\right] \qquad (3.8.)$$

donde β corresponde al modelo formulado y β_0 al modelo restringido, en el que únicamente se considera el término constante, estimándose los modelos por el procedimiento de máxima verosimilitud.

Así, se comparará el valor de la función de máxima verosimilitud en el máximo para ambos modelos, valor que se distribuye como una χ^2 con *k-1* grados de libertad (siendo *k* el número total de variables independientes, incluida la constante). Si el valor obtenido es mayor que el recogido en las tablas, se rechaza la hipótesis nula:

$$H_0 = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \ldots\ldots + \beta_k \qquad (3.9.)$$

y por tanto, las variables son conjuntamente explicativas (Martín *et al.*, 1997).

La segunda posibilidad de medir la bondad del ajuste del modelo es calcular el pseudo R^2 *de Mc Fadden*, definido como:

$$\text{R2 de McFadden} = \left[1 - \frac{LogL(\beta)}{Log(\beta_0)}\right] \qquad (3.10.)$$

donde β corresponde al modelo formulado y β_0 al modelo restringido en el que únicamente se considera el termino constante. Se trata de una medida similar al R^2 estimado en los modelos de

regresión, ya que su valor está comprendido entre 0 y 1, aunque en el caso de modelos de variable dependiente cualitativa su interpretación no es tan directa (Maddala, 1996).

Finalmente, al proporcionar este modelo probabilidades estimadas y al tener en la muestra de datos realizaciones de la variable dependiente, resulta posible comparar si la predicción de la probabilidad coincide con la realización muestral o, lo que es lo mismo, calcular un coeficiente de determinación en términos de la proporción de predicciones correctas.

Dado que la variable dependiente es 0 ó 1, después de calcular la $\hat{y}_i$, se considera que la observación *i-ésima* pertenece al primero de los grupos (*Y=1*) si $\hat{y}_i \geq 0{,}5$. En caso contrario, se clasifica a la observación dentro del segundo grupo (*Y=0*)[91].

De esta manera se puede contar el número de predicciones correctas y definir el valor predicho, a través de la expresión:

$$\hat{y}_i^* = \begin{cases} 1 \text{ si } \hat{y}_i \geq 0{,}5 \\ 0 \text{ si } \hat{y}_i < 0{,}5 \end{cases} \qquad (3.11.)$$

El número de aciertos (que estará formado por las observaciones que realmente toman valor 1 y son predichas como 1, más las observaciones que son realmente 0 y son predichas como 0) dividido por la muestra total, constituye esta medida de bondad del ajuste que se conoce como el porcentaje de respuestas correctas. Se obtendría, por lo tanto, un estadístico para conocer la proporción de predicciones estimadas correctamente por el modelo.

3.4.2. El modelo logit multinomial

Este modelo (Maddala, 1983) constituye una extensión del modelo *logit binomial* al caso en que la variable dependiente *Y* toma más de dos valores (*j = 0,.....,J*).

El modelo *logit multinomial* se basa en el supuesto de que el individuo compara las alternativas dos a dos, sin considerar el resto. De esta forma, el cálculo de la probabilidad de un determinado suceso *j* se realiza de la siguiente manera (Mc Fadden, 1974):

$$P(y_i = j \mid X) = Pj = \frac{e^{X'Bj}}{1 + \sum_{s=1}^{J-1} e^{X'Bs}} \qquad \forall j = 0,1,...,J \qquad (3.12.)$$

Donde $j = 0,1,...,J$ se refiere a los diferentes posibles valores que puede tomar la variable dependiente, X hace referencia al vector de características de los individuos y β_j es el vector de coeficientes ligado a las variables explicativas, es decir, los parámetros a estimar.

Sin embargo, como ocurre en el modelo binomial, no todos los parámetros pueden estimarse. Si en el caso *binomial* (*J=2*) únicamente se estima uno de ellos, en el modelo *multinomial* se estiman *J-1* conjuntos de parámetros.

La interpretación de los coeficientes en un modelo *logit* multinomial resulta complicada. En cualquier caso, debe tenerse en cuenta que los parámetros del modelo son relativos a la categoría que se toma de referencia, y que la interpretación de los resultados en términos de probabilidades o en *odd ratios* es más sencilla que usando los efectos marginales.

El efecto parcial en P (Yi=j) para una variable concreta Xk es el siguiente:

$$\frac{\partial P_j}{\partial X_k} = P_j \left[\beta_{jk} - \sum_{j=0}^{J} P_s \, \beta_{sk} \right] \qquad (3.13.)$$

---

91 Cuando la distribución de los posibles valores que toma la variable está lejos de ser el 50%, puede alterarse el punto de corte en el cual se clasifica a los individuos como pertenecientes al valor 1 ó al valor 0.

y, por tanto, la elasticidad aparece definida como:

$$\frac{\partial P_j}{\partial X_k}\frac{X_k}{P_j} = X_k\left[\beta_{jk} - \sum_{j=0}^{J} P_s\,\beta_{sk}\right] \qquad (3.14.)$$

De acuerdo con Amemiya (1981) y Maddala (1983), los efectos marginales y la elasticidad no resultan indicadores apropiados del efecto de una variable[92] sobre P_j.

Alternativamente, la interpretación de un modelo *logit multinomial* puede realizarse a través de los *odds ratios* (Green, 2003) tomados en logaritmos, donde el logaritmo del ratio de dos probabilidades es una función de las variables independientes y se define como:

$$Ln\,\frac{P_j}{P_s} = X(\beta_j - \beta_s) \qquad (3.15.)$$

Si se igualan los coeficientes de una categoría (en general, la categoría tomada como referencia, β_s) a cero, entonces se obtiene que:

$$Ln\,\frac{P_j}{P_s} = X(\beta_j) \qquad (3.16.)$$

Esta aproximación permite obtener una sencilla interpretación lineal del efecto de cada variable independiente. De esta forma, se puede obtener el cambio en el *odd ratio* para la categoría *j* asociado a una variable concreta X_k examinando $e^{\beta jk}$. En consecuencia, un cambio unitario en la variable X_k (si X_k es una variable ficticia, el cambio se mide con respecto al individuo de referencia), genera un cambio en el *odd ratio* de la categoría relevante con respecto a la categoría de referencia de $e^{\beta jk}$.

Finalmente, resta añadir que los indicadores de bondad del ajuste del modelo *logit multinomial* son equivalentes a los del *caso binomial*, que pueden consultarse en el epígrafe 3.4.1.2. del presente capítulo.

3.4.3. El modelo de selección muestral

En muchos casos, al realizar la especificación del modelo econométrico, debe considerarse la posibilidad de que se tenga que trabajar con una selección de la muestra no aleatoria, en la que se observe o no la variable dependiente Y en función del resultado de otra variable.

En nuestro trabajo, el problema de selección muestral aparece cuando intentamos determinar el desajuste educativo en el primer empleo significativo, ya que éste sólo puede ser observado en el caso de que el individuo haya obtenido un puesto de trabajo significativo antes del momento de ser entrevistado.

Para solucionar este problema, suele recurrirse procedimiento de estimación bietápico (Heckman, 1979), en el que se incluyen dos ecuaciones:

$$z_i^* = \gamma' W_i + u_i$$
$$y_i^* = \beta' X_i + \varepsilon_i$$

La primera, denominada ecuación de selección muestral define, en nuestro caso concreto, qué variables influyen en la probabilidad de haber conseguido un empleo significativo. Se trata de un modelo probit *univariante*, en el que z^* es una variable latente que se aproxima mediante z,

92 Se puede observar que este efecto depende del valor que alcancen el resto de variables cuando se toma la derivada parcial, y que puede o no tener el mismo signo que el coeficiente, cuyo signo puede cambiar dependiendo del valor específico de la variable.

de manera que z_i=1 si $z_i^* > 0$ y z=0 si $z_i^* \leq 0$. La segunda ecuación sería, en nuestro análisis, la ecuación de desajuste educativo, que sólo se observa para los jóvenes que han encontrado su primer empleo significativo.

La existencia del sesgo de selección aludido implica que sólo se pueda observar Yi cuando el individuo haya conseguido un empleo significativo, por lo que, en un estado previo, debe tenerse en cuenta la obtención o no por parte del individuo de un empleo significativo. Dicha circunstancia se modeliza de la siguiente manera:

$$Z_i = 1 \text{ si } Z_i^* = \delta' W_i + u_i > 0 \qquad (3.17.)$$
$$Z_i = 0 \text{ si } Z_i^* = \delta' W_i + u_i \leq 0$$

siendo Z_i una variable que toma valor 1 cuando el individuo haya encontrado una empleo significativo, y valor 0, en caso contrario y Z_i^* una variable latente que viene explicada por el vector de características individuales W_i y por un término de perturbación aleatoria ui que se distribuye según una Normal $(0, \sigma^2)$. Entonces, la probabilidad de haber obtenido un empleo significativo viene definida por la siguiente expresión:

$$P(Z_i = 1) = P(\delta' W_i + u_i > 0) = \varphi \frac{\varphi' W_i}{\sigma} \qquad (3.18.)$$

Tomando la esperanza condicional de Y_i , se obtiene que:

$$E(Y_i / Z_i, \delta' W_i + u_i > 0) = \beta' X_i + E(\varepsilon_i / u_i > -\delta' W_i) \qquad (3.19.)$$

Asumiendo que u_i y ε_i siguen una distribución Normal bivariante $(0,0,1,\sigma_\varepsilon,\rho)$, siendo ρ el coeficiente de correlación entre ambas perturbaciones, se puede escribir:

$$E(\varepsilon_i / u_i > -\delta' W_i) = \rho' \sigma_\varepsilon \lambda_i \qquad (3.20.)$$

Si se adopta la normalización $\sigma_\varepsilon = 1$ resulta que:

$$E(\varepsilon_i / W_{1i}, Z_i > 0) = \rho' \lambda_i \qquad (3.21.)$$

El parámetro λ_i se interpreta como el término de corrección del sesgo de selección y tiene la forma:

$$\Lambda_i = \frac{\varphi(\delta'' W / \sigma_u)}{\varphi(\delta' W / \sigma_u)}, \text{ si } Zi > 0 \text{ y } \Lambda_i = -\frac{\varphi(\delta'' W / \sigma_u)}{\varphi(\delta' W / \sigma_u)}, \text{ si } Zi \leq 0 \qquad (3.22.)$$

donde φ representa la función de densidad de una Normal $(0, 1)$, y $\varnothing$ su correspondiente función de distribución.

La manera de contrastar si los términos de perturbación ε_i y u_i están correlacionados es introduciendo como variable independiente en la segunda ecuación la *inversa del ratio de Mills* (λ_i), obtenida en la primera etapa. Si el coeficiente asociado a esta variable resulta significativo, se entendería que existen características de los jóvenes que explican simultáneamente la probabilidad de obtener un empleo significativo y el grado de desajuste experimentado en dicho puesto de trabajo.

De esta forma, si existiera correlación entre ambas ecuaciones, no sería adecuado estimar el desajuste educativo ignorando la ecuación de selección muestral, ya que se obtendrían estimadores inconsistentes de los parámetros.

3.5. Recapitulación

En esta parte de la investigación se ha realizado un estudio de diversos aspectos metodológicos que condicionarán el análisis empírico efectuado en capítulos sucesivos.

En primer lugar, se han presentado las principales características del Módulo de Transición de la Educación al Mercado Laboral, del segundo trimestre de la Encuesta de Población Activa de 2000, que será la fuente estadística empleada en el estudio empírico posterior. Algunas de las razones que han motivado esta elección radican en el hecho de que dicha base de datos suministra cierta información acerca de dos hechos transcendentales en el ciclo vital de los individuos: la salida del sistema educativo y el logro del primer empleo significativo, que resultan de vital importancia en el análisis empírico desarrollado en los capítulos posteriores. Además, esta fuente estadística combina las ventajas del gran tamaño muestral de la Encuesta de Población Activa con información detallada de la transición de los jóvenes al mundo laboral.

A continuación, se procede a exponer las tres hipótesis principales que serán contrastadas en los siguientes capítulos, con el fin de presentar las variables que resultan pertinentes para su contrastación. De esta manera, en primer lugar, se contrastará si el entorno socioeconómico de los individuos ejerce influencia en la realización de estudios universitarios. En segundo lugar, se planteará si la obtención de un título universitario confiere al individuo ciertas ventajas a la hora de obtener su primer empleo significativo. Finalmente, se cuestionará si el fenómeno de la sobreeducación en el primer empleo significativo incide de forma más acusada entre los graduados universitarios. A nuestro juicio, dado el crecimiento en el número de titulados universitarios en las últimas décadas, la pertinencia de efectuar un análisis pormenorizado de las tres cuestiones enunciadas resulta un hecho indudable.

Las variables explicativas que se utilizarán en el contraste de las hipótesis mencionadas pueden agruparse en cinco diferentes categorías: las características personales, las características familiares, los aspectos relacionados con el proceso de búsqueda de empleo, los rasgos del puesto de trabajo obtenido y, finalmente, los factores de entorno.

Por último, el capítulo concluye con la presentación de los modelos mediante los cuales se tratarán de contrastar las hipótesis presentadas. En el estudio se repasan sus principales características, su formulación econométrica, así como los principales indicadores para medir la bondad del ajuste del modelo.

En definitiva, con la elaboración de este capítulo se ha pretendido describir las principales características de la fuente de datos utilizada, así como presentar la metodología del análisis empírico que será empleada en los capítulos sucesivos de esta tesis doctoral.