Facultad de Economía,
Empresa y Turismo
Universidad de La Laguna

# MEMORIA DEL TRABAJO FIN DE GRADO

## ANÁLISIS DE LA DESIGUALDAD DE OPORTUNIDADES EN ESPAÑA A TRAVÉS DEL MODELO DE SELECCIÓN DE VARIABLES LASSO

Analysis of inequality of opportunity in Spain through the LASSO variable selection model

Autoría: D. Roberto Adrián Martín García

Tutorizado por: D. Gustavo Alberto Marrero Díaz

Grado en ECONOMÍA

FACULTAD DE ECONOMÍA, EMPRESA Y TURISMO

Curso Académico 2021 / 2022

La Laguna, 14 de Junio de 2022

**RESUMEN**

La desigualdad de oportunidades (DO) es aquella desigualdad que proviene de circunstancias que escapan del control de los individuos. Los modelos habituales para estudiar la DO presentan sesgos importantes en su estimación al alza y a la baja, por ello, en este trabajo se intentan mitigar estos sesgos realizando una selección de circunstancias a través de LASSO, con los cuales se construyen modelos de DO para el caso español. Se encuentra que circunstancias como el sexo, la edad, los estudios del padre y el tipo de escuela en donde estudió el individuo influyen considerablemente sobre la renta, y a través de un modelo cruzado se constataron distintas relaciones entre circunstancias que amplifican los efectos sobre los ingresos. Igualmente observamos la importancia de los canales educativos y laborales para mitigar el impacto de las circunstancias y se calculó que la ratio desigualdad de oportunidades sobre desigualdad total (DO/DT) ronda el 49%.

Inequality of opportunity (IO) is the inequality that comes from circumstances that are not in the control of individuals. Current models to study IO suffer from considerable biases in its estimations, so in this study we try to limit these biases by selecting circumstances with LASSO, and with those results some models are constructed for the case of IO in Spain. We find that some circumstances like sex, age, father's education and the type of school that the individual attended have considerable impacts in the individual's income, and with a cross-variable model we found a number of relationships between circumstances that amplified the effects on income. We observed the importance of the educational and occupational channels to limit the impact of circumstances and we calculated that the inequality of opportunity over total inequality ratio (IO/TI) is approximately 49%.

Palabras clave: Desigualdad, Desigualdad de oportunidades, LASSO.

## ÍNDICE DE CONTENIDOS

## ÍNDICE DE TABLAS



## ÍNDICE DE FIGURAS

## 1. INTRODUCCIÓN

En las últimas décadas la desigualdad se ha posicionado como uno de los temas prioritarios de debate a nivel académico, social y político a lo largo del mundo. Economistas como Stiglitz (2015) indican que "La creciente desigualdad presente en la mayoría de los países es uno de los problemas más críticos a los cuales se enfrenta el mundo actualmente". (p.379). En el artículo, Stiglitz no sólo argumenta que hay razones morales para oponerse a la creciente desigualdad presente en el *status quo*, sino que el problema se extiende al ámbito económico: "En contraste a aquellos que creen que la desigualdad es necesaria para un buen desempeño económico, investigaciones recientes muestran que la desigualdad es perjudicial para el desarrollo, la estabilidad y la eficiencia económica". (p.388). Con respecto al caso español, autores como Ayala (2016) indican que "El retrato que ofrecen las estadísticas comparadas en la actualidad no es muy distinto del que había hace cuatro décadas y en él domina la caracterización de España como un país con niveles muy altos de desigualdad en el contexto europeo". (p. 52).

Sin embargo, cabe destacar en este punto que dentro de la literatura económica se distinguen entre diferentes tipos de desigualdad (Roemer, 1993; Marrero y Rodríguez, 2012). Estos autores destacan dos categorías, una proveniente únicamente de diferenciales de esfuerzo entre individuos que denominan desigualdad de esfuerzo y otra proveniente de diferencias en circunstancias, que denominan desigualdad de oportunidades. Esta distinción es sumamente relevante, ya que distintos tipos de desigualdad afectan de manera distinta al crecimiento económico. Marrero y Rodríguez (2013) encuentran que la desigualdad de esfuerzos tiene una relación positiva con el crecimiento, mientras que la desigualdad de oportunidades tiene un impacto negativo sobre el mismo.

En este contexto, tomando en cuenta que España es un país con relativamente altos niveles de desigualdad con respecto a sus homólogos europeos, resulta fundamental el estudio de qué parte de esa desigualdad es de oportunidades, que además de ser perjudicial para el crecimiento es considerada socialmente como injusta. Este tipo de desigualdad es particularmente preocupante, ya que como indican instituciones como el Banco Mundial (2006) "las circunstancias en el momento del nacimiento no deben tener ningún peso en las oportunidades que una persona tenga en la vida". (p.13). En este sentido, el estudio de la desigualdad de oportunidades y en particular de cuáles circunstancias son relevantes resulta sumamente necesario para comprender la naturaleza y el alcance de este problema y para orientar la elaboración de políticas públicas para que realmente contribuyan a equilibrar el terreno de juego en España.

En este trabajo entenderemos como "circunstancias" a todas aquellas características que escapan del control del individuo y que pueden ejercer un impacto sobre su renta, como el género, educación y ocupación de los padres, tipo de escuela, características del entorno en donde creció, entre otros. Por otra parte, para entender bien los mecanismos a través de los cuales las circunstancias terminan afectando a la renta, es menester definir correctamente los canales, que son aquellas características mediante las cuales las circunstancias pueden ejercer su influencia sobre la renta del individuo. En este caso, estamos hablando, por ejemplo, del nivel educativo del individuo y de su ocupación, que han sido en parte fruto de las decisiones del

individuo, van a tener un impacto en su renta, pero que probablemente han sido afectadas o condicionadas por las circunstancias.

De esta forma, el objetivo principal de este trabajo es analizar el problema de la desigualdad de oportunidades en España, con un enfoque especial en la selección de las circunstancias más relevantes que influyen sobre la renta de los individuos en el país. Como objetivos complementarios se propone analizar la relevancia de los canales de la educación y ocupación dentro del contexto de la desigualdad de oportunidades y estimar la ratio desigualdad de oportunidades sobre desigualdad total (ratio DO/DT).

Para realizar este estudio, utilizaremos los datos provenientes del módulo de desigualdad y movilidad social del CIS (Estudio Nº 3178), a través de los cuales determinaremos las circunstancias más relevantes que influyen en la desigualdad de oportunidades en España, además estimaremos el valor de estos coeficientes para conocer el impacto particular de cada uno de ellos sobre la renta de los individuos. Esta encuesta se realizó con carácter nacional a población mayor de edad residente en España, y cuenta con 2482 observaciones con más de 50 variables, por lo que nos aporta información amplia y relevante con la cual podremos estimar la desigualdad de oportunidades en España.

La metodología convencional para analizar la desigualdad de oportunidades es el uso de modelos de regresión lineal (Ferreira y Ginoux, 2011; Marrero y Rodríguez, 2012; Palomino, et al, 2019), en donde la variable dependiente es la renta del individuo o del hogar (usualmente en logaritmo neperiano), y como variables independientes se tienen el conjunto de circunstancias. Esta metodología, aun cuando está muy arraigada en la literatura y en el análisis econométrico, presenta ciertos problemas que pueden afectar la fiabilidad de los resultados. Autores como Brunori et al. (2021) exponen que:

> Las aproximaciones actuales para estimar la desigualdad de oportunidades sufren de sesgos que son consecuencia de elecciones críticas en la etapa de selección del modelo. En primer lugar, los investigadores deben decidir las circunstancias a considerar para la estimación. […] Por una parte, descartar circunstancias relevantes de los modelos limita el alcance predictivo de las circunstancias y lleva a sesgos a la baja de las estimaciones. Por otra parte, incluir demasiadas circunstancias ocasiona un sobreajuste que lleva a sesgos al alza de las estimaciones. (p.2).

Es por esto que en los últimos años han surgido propuestas de incorporar herramientas provenientes del Machine Learning para el estudio de la desigualdad de oportunidades. En el mismo artículo, Brunori et al. (2021) indican que "Los métodos de Machine Learning permiten la creación de modelos flexibles que explican la desigualdad de oportunidades mientras imponen disciplina estadística a través del criterio de la replicabilidad fuera de la muestra. Estas características sirven para obtener estimaciones de la desigualdad de oportunidades que son menos propensas a sesgos al alza o a la baja". (p.1)

Siguiendo el artículo de Cabrera et al. (2019), en este trabajo se realiza un estudio de la desigualdad de oportunidades en España, seleccionando y estimando aquellas circunstancias que poseen el mayor impacto sobre la renta del individuo. Sin embargo, debido al gran número de variables, y para evitar los sesgos mencionados anteriormente, la selección de circunstancias relevantes se realizará a partir del uso del método LASSO, una metodología creada por Robert Tibshirani en 1996 que, mediante un problema de optimización, selecciona un subconjunto de

variables que juntas forman un modelo que minimiza el error cuadrático medio. Esta metodología se ha utilizado con cada vez mayor frecuencia en el ámbito académico e investigador, ya que corrige los errores típicos que surgen al aplicar únicamente métodos de regresión lineales, permitiendo obtener modelos más explicativos, menos sesgados y con un conjunto de variables seleccionadas bajo criterios puramente estadísticos, mitigando así la influencia del investigador sobre los resultados.

A partir del conjunto de variables seleccionadas por LASSO, se realizarán diversos modelos MCO para estudiar el problema de la desigualdad de oportunidades, incluyendo un modelo seleccionando las circunstancias más relevantes que explican la DO en España, un modelo con circunstancias y canales, donde podemos observar la forma en la que la educación y la ocupación canalizan los efectos de las circunstancias sobre la renta de los individuos, y por último, realizaremos un modelo con circunstancias cruzadas. Este último modelo resulta un aporte interesante de este trabajo al estudio de la desigualdad de oportunidades en España, ya que la inclusión de efectos cruzados de las variables es una manera de mejorar los modelos lineales paramétricos convencionales, ya que permiten considerar potenciales no-linealidades en los modelos. Sin embargo, la existencia de muchos potenciales efectos cruzados provoca que el procedimiento habitual paramétrico tienda a generar fácilmente problemas de sobre estimación y de sesgos al alza de la estimación de la desigualdad de oportunidades. En este contexto de múltiples potenciales variables es donde un método de selección de variables como LASSO puede cobrar más relevancia, que es lo que se empleará en este estudio.

Por último, a través de la estimación paramétrica de la renta explicada y la posterior aplicación de índices de Gini sobre estas variables, se procederá a estimar la ratio desigualdad de oportunidades sobre desigualdad total, para observar qué parte de la desigualdad de ingresos en España puede explicarse a través del conjunto de circunstancias seleccionado.

De esta manera, el aporte más importante de este trabajo proviene del uso de herramientas de Machine Learning para la selección de circunstancias relevantes, mejorando los resultados provenientes del uso exclusivo de métodos de regresión lineal al disminuir los posibles sesgos al alza y a la baja que puedan afectar los resultados debido a problemas como el sobreajuste, la eliminación indebida de variables relevantes, entre otras.

Este trabajo se divide en siete secciones. En la sección 2 se describe la metodología a emplear, en la sección 3 se describe la base de datos utilizada y se observan algunos estadísticos descriptivos preliminares como aproximación al estudio, en la sección 4 se exponen y discuten los resultados de los modelos MCO con LASSO, en la sección 5 se hace un estudio de la desigualdad de oportunidades a partir de los resultados de cada modelo, en la sección 6 se hace un estudio con LASSO de las circunstancias junto con los canales y por último en la sección 7 se encuentran las conclusiones de este trabajo.

## 2. METODOLOGÍA

Esta sección describe la metodología empleada para seleccionar las circunstancias que afectan a la renta de los individuos y estimar la desigualdad de oportunidades. Una manera convencional de estudiar las circunstancias relevantes que permiten explicar la desigualdad de oportunidades proviene del uso de métodos de regresión lineales paramétricos ex-ante (Ferreira y Gignoux, 2011; Marrero y Rodríguez, 2012; Palomino et al, 2019). En este caso nos

encontramos con un una variable objetivo *y,* que por lo general suele ser renta personal o del hogar en logaritmo neperiano para *i* número de individuos. Además, tendríamos que las variables explicativas son el conjunto de circunstancias de cada individuo, que representaremos en forma de vector C, con un vector de coeficientes β, de forma que,

$$\ln(y_i) = \beta C_i + \varepsilon_i,$$

donde ε es la parte correspondiente a la renta personal del individuo no explicada por el conjunto considerado de circunstancias.

Como expone Cabrera et al. (2019), si existe igualdad de oportunidades, el conjunto de circunstancias personales no debería afectar la distribución de los ingresos, y por consecuencia, la variabilidad del vector $\exp(\beta C_i)$ (una distribución suavizada ajustada por las circunstancias existentes) debería ser cero. Si aplicamos el índice de Gini a este vector, obtendríamos una estimación paramétrica de la desigualdad de oportunidades explicada por el conjunto seleccionado de circunstancias. Además, si aplicamos el índice de Gini a la renta personal de los individuos, tendríamos la estimación de la desigualdad total, lo cual nos permitiría calcular directamente el ratio de desigualdad de oportunidades sobre desigualdad total.

Sin embargo, el uso exclusivo de métodos de regresión lineales lleva consigo un conjunto de consecuencias que pueden sesgar considerablemente los resultados obtenidos. Como exponen Brunori et al. (2019) "Al medir la desigualdad de oportunidades con datos de encuesta, los investigadores se enfrentan a dos tipos de sesgos. Un sesgo a la baja, debido a la observabilidad parcial de las circunstancias que afectan a los resultados del individuo, y un sesgo al alza, que es consecuencia directa de la varianza en la muestra". (p.635).

Cabe destacar que estos sesgos a la baja y al alza no son los únicos inconvenientes que provienen del uso exclusivo de modelos de regresión lineales. Autores como Ranstam y Cook (2018) indican que "aplicar métodos de regresión estándar a un conjunto de variables posibles para generar un modelo tiende a llevar a un sobreajuste en términos del número de variables finalmente incluidas en el modelo, y también a una sobreestimación de qué tan bien funciona el modelo en términos del uso de las variables incluidas para explicar la variabilidad observada". (p.1348).

Tomando en cuenta estos inconvenientes, y para realizar una mejor estimación de la desigualdad de oportunidades, surge como una posible solución el uso del método LASSO de selección de variables. El método LASSO, o *least absolute shrinkage and selection operator* por sus siglas, fue publicado por Robert Tibshirani en 1996 como un método para mejorar las estimaciones por mínimo cuadrados ordinarios. En su artículo, expone que:

> Existen dos razones por las cuales los analistas de datos a menudo no están satisfechos con las estimaciones MCO. La primera es la precisión predictiva: Las estimaciones MCO a menudo tienen bajo sesgo pero gran varianza, y la precisión predictiva puede ser mejorada disminuyendo o dando un valor 0 a algunos coeficientes. Al hacer esto sacrificamos algo de sesgo para reducir la varianza, mejorando en términos generales la precisión predictiva. La segunda razón es la interpretación, ya que con un gran número de predictores, nos gustaría determinar un subconjunto más pequeño que exhiben los mayores efectos. (Tibshirani, 1996, p.267).

Con esto en mente, Tibshirani propone el método LASSO, que siguiendo la explicación brindada por Buhlmann y van de Geer (2011), podemos entenderlo como un problema de optimización de la forma:

$$\beta(\lambda) = \arg\min\left(\sum_{i=1}^{n} \frac{(Y_i - (X\beta)_i)^2}{n} + \lambda \sum_{j=1}^{p} |\beta_j|\right),$$

donde $\lambda \geq 0$ es un parámetro de penalización.

Si el parámetro de penalización toma valor cero, el problema a resolver es un mínimo cuadrado ordinario convencional. Sin embargo, si toma valores superiores a cero, el método penaliza aquellas variables que no aportan suficiente poder explicativo a la variable objetivo, pudiendo llegar a ser óptimo eliminarlas y asignarle un coeficiente igual a cero, eliminando de esta forma las variables más irrelevantes del modelo. Autores como Fonti y Belitser (2017) exponen que "Existen muchas ventajas derivadas de usar el método LASSO, en primer lugar, puede proveer una muy buena precisión predictiva, ya que disminuir y remover los coeficientes pueden reducir la varianza sin un cambio sustancial en el sesgo. [...] Además, LASSO ayuda a incrementar la interpretabilidad del modelo al eliminar variables irrelevantes que no están asociadas a la variable objetivo, resolviendo así el problema de sobreajuste". (p.5)

Sin embargo, el uso de LASSO requiere el cumplimiento de una serie de requisitos para que los resultados obtenidos tengan validez en la práctica, relacionados con el parámetro de penalización λ y con la propia estructura de los datos.

Con respecto a los datos, el proceso de selección de variables es sensible a la magnitud de las observaciones, debido a que la penalización toma la forma del parámetro de penalización λ multiplicada por el coeficiente β de cada predictor. Si tenemos variables explicativas (circunstancias) de distinta escala, puede darse el caso de que LASSO rechace de manera incorrecta aquellas con valores más elevados en términos absolutos, ya que se sobrepenalizaría de manera artificial el coeficiente asociado a esa variable. Para corregirlo, autores como Tibshirani et al. (2011) exponen que "Estandarizar los predictores es una práctica común antes de aplicar LASSO, de esta forma el término de penalización tiene sentido". (p. 248). En este trabajo, solo en el proceso de selección de variables, usamos este proceso de estandarización.

Por otra parte, se puede observar que el problema de optimización, y por tanto el proceso de selección de variables, depende del valor que tome el parámetro de penalización λ, lo cual tendrá repercusiones directas sobre las estimaciones finales. El número de coeficientes distintos de cero se hace máximo cuando λ se aproxima a cero (se incluyen todas las variables), y se hace cero cuando λ tiende a infinito (no se incluye ninguna variable en el modelo), por lo tanto, encontrar un valor óptimo de λ es fundamental para obtener estimaciones fiables.

Uno de los métodos más habituales en la práctica para encontrar el λ óptimo es el uso de la validación cruzada. En palabras de Brunori et al. (2019), estos métodos se basan en lo siguiente:

> La muestra original se divide en un conjunto de entrenamiento y un conjunto de pruebas. La relación entre predictores y variable objetivo se estima primero en el conjunto de entrenamiento, bajo un gran número de especificaciones de modelos posibles. Luego, los coeficientes derivados son usados para predecir los resultados en el conjunto de pruebas. La especificación seleccionada es aquel modelo que minimiza los errores de predicción en el conjunto de prueba. (p.638).

El parámetro lambda óptimo es aquel que minimiza el error cuadrático medio de los errores fuera de la muestra. De esta forma, este parámetro de penalización, y el estimador LASSO en sí mismo, constituye una manera relativamente sencilla y con una sólida base estadística para realizar el trade-off más apropiado entre minimizar el sesgo a la baja por ausencia de variables explicativas y el sesgo al alza por sobreajuste. Es por esto que para este trabajo se usa el método de validación cruzada para obtener el parámetro de penalización óptimo.

Por último, es necesario establecer que el método LASSO no está exento de sesgos, autores como Hastie et al. (2009) exponen que las disminuciones de LASSO provocan sesgos hacia cero de los coeficientes. Sin embargo, también indican que "Una aproximación para reducir este sesgo es correr LASSO para identificar el subconjunto de coeficientes distintos de cero, y después correr un modelo lineal no restringido para el conjunto seleccionado de predictores". (p.91). Este método es el que se empleará en este trabajo, por lo que se realizarán dos etapas para cada estimación. En primer lugar se utilizará el método LASSO para seleccionar las variables relevantes de cada modelo (usando variables estandarizadas), y luego se correrá un MCO con el conjunto de predictores seleccionados.

## 3. DESCRIPCIÓN DE LOS DATOS

Para el desarrollo de este trabajo se utilizarán los datos provenientes del módulo de desigualdad y movilidad social del Centro de Investigaciones Sociológicas de España (Estudio Nº 3178). Este cuestionario fue realizado del 13 al 24 de noviembre de 2017 en ámbito nacional, la población objetivo eran residentes mayores de 18 años de ambos sexos y se consiguieron 2482 encuestas definitivas. El tipo de muestreo realizado fue multietapas, estratificado por conglomerados, y los estratos se formaron por el cruce de las 17 comunidades autónomas, con el tamaño de entorno dividido en 7 categorías.

En total, la base de datos consta de 59 variables, provenientes de preguntas de muy diversa naturaleza, tratando desde la nacionalidad del encuestado hasta el peso y estatura corporal. Por lo tanto, hubo una preselección de variables a utilizar en este estudio, que como justificamos en anteriores apartados se dividió en tres categorías: Circunstancias, Canales y Variable objetivo.

La variable objetivo de este estudio será el ingreso personal del encuestado tras aplicarle el logaritmo neperiano. Las circunstancias son aquel conjunto de variables que escapan del control del individuo y que repercuten de manera directa o indirecta en su desenvolvimiento personal o profesional, en el caso de este estudio en particular, serán aquellas variables fuera de su control que últimamente determinen la renta del encuestado, expuestas próximamente en la Tabla 1. Por último entendemos por canales aquellas variables por las cuales las circunstancias pueden ejercer su influencia sobre la renta del individuo, en este caso será el nivel educativo y la ocupación del individuo que percibe rentas. La ocupación se divide en cuatro categorías según el sistema de Clasificación Internacional Uniforme de Ocupaciones (CIUO), que son no cualificados, cualificados, técnicos y directivos, mientras que la educación se ha agrupado en cuatro categorías, sin estudios primarios completos, estudios primarios o ESO, estudios de bachillerato o FP y universitarios.

Todas las variables se pasan a categóricas (i.e., la educación del padre tiene sin educación, educación primaria o ESO, secundaria o FP y terciaria), lo que permite modelizar la posible no linealidad del efecto de la propia variable sobre la renta (esto no permite considerar efectos cruzados, que será tratado con posterioridad). Tanto MCO como LASSO no permiten tener observaciones perdidas, por lo que la muestra final se restringe a 1185 observaciones. Aun así, es un tamaño muestral elevado en comparación con otras muestras existentes para España. En la Tabla 1 se recogen y explican las variables empleadas en este estudio, junto a su media y desviación típica:

**Tabla 1. Variables preseleccionadas para el estudio**

| Código | Descripción | μ | s |
|---|---|---|---|
| ***Circunstancias básicas*** | | | |
| Tamano_<50 | (1) 'Nació en pueblo de menos de 50 mil habitantes'; (0) 'Otro' | 0.51 | 0.5 |
| Tamano_50_400 | (1) 'Nació en ciudad de entre 50 y 400 mil habitantes'; (0) 'Otro' | 0.32 | 0.47 |
| Tamano_mas400 | (1) 'Nació en ciudad de más de 400 mil habitantes'; (0) 'Otro' | 0.17 | 0.38 |
| Nacion | (1) 'Nacionalidad española'; (0) 'Otra nacionalidad' | 0.91 | 0.29 |
| Sexo | (1) 'Mujer'; (0) 'Hombre' | 0.44 | 0.5 |
| Edad_hasta30 | (1) 'Hasta 30 años de edad'; (0) 'Otro | 0.15 | 0.35 |
| Edad_31_45 | (1) 'Entre 31 y 45 años de edad'; (0) 'Otro' | 0.32 | 0.47 |
| Edad_46_60 | (1) 'Entre 46 y 60 años de edad'; (0) 'Otro' | 0.25 | 0.44 |
| Edad_61_75 | (1) 'Entre 61 y 75 años de edad'; (0) 'Otro' | 0.2 | 0.4 |
| Edad_mas75 | (1) 'Más de 75 años de edad; (0) 'Otro' | 0.08 | 0.27 |
| ***Características de los padres*** | | | |
| Padest_sin | (1) 'Padre sin estudios o primarios incompletos'; (0) 'Otro' | 0.34 | 0.48 |
| Pades_primsecun | (1) 'Padre con estudios primarios o secundarios 1º etapa (bachillerato elemental, FP oficialía, FP1, FP Inicial, ESO)'; (0) 'Otro' | 0.46 | 0.5 |
| Padest_secunfp | (1) 'Padre con estudios de Secundaria 2ª etapa (bachillerato, BUP, COU) o FP (ciclos formativos medios y superiores, FP2 y maestrías)'; (0) 'Otro' | 0.1 | 0.3 |
| Padest_univ | (1) 'Padre con estudios universitarios'; (0) 'Otro' | 0.1 | 0.3 |
| Madest_sin | (1) 'Madre sin estudios o primarios incompletos'; (0) 'Otro' | 0.35 | 0.48 |
| Madest_primsecun | (1) 'Madre con estudios primarios o secundarios 1º etapa (bachillerato elemental, FP oficialía, FP1, FP Inicial, ESO)'; (0) 'Otro' | 0.49 | 0.5 |
| Madest_secunfp | (1) 'Madre con estudios de Secundaria 2ª etapa (bachillerato, BUP, COU) o FP (ciclos formativos medios y superiores, FP2 y maestrías)'; (0) 'Otro' | 0.09 | 0.28 |
| Madest_univ | (1) 'Madre con estudios universitarios'; (0) 'Otro' | 0.07 | 0.26 |
| Clasepad_alt | (1) 'Padre de clase alta (directivos y profesionales de alto | 0.11 | 0.31 |

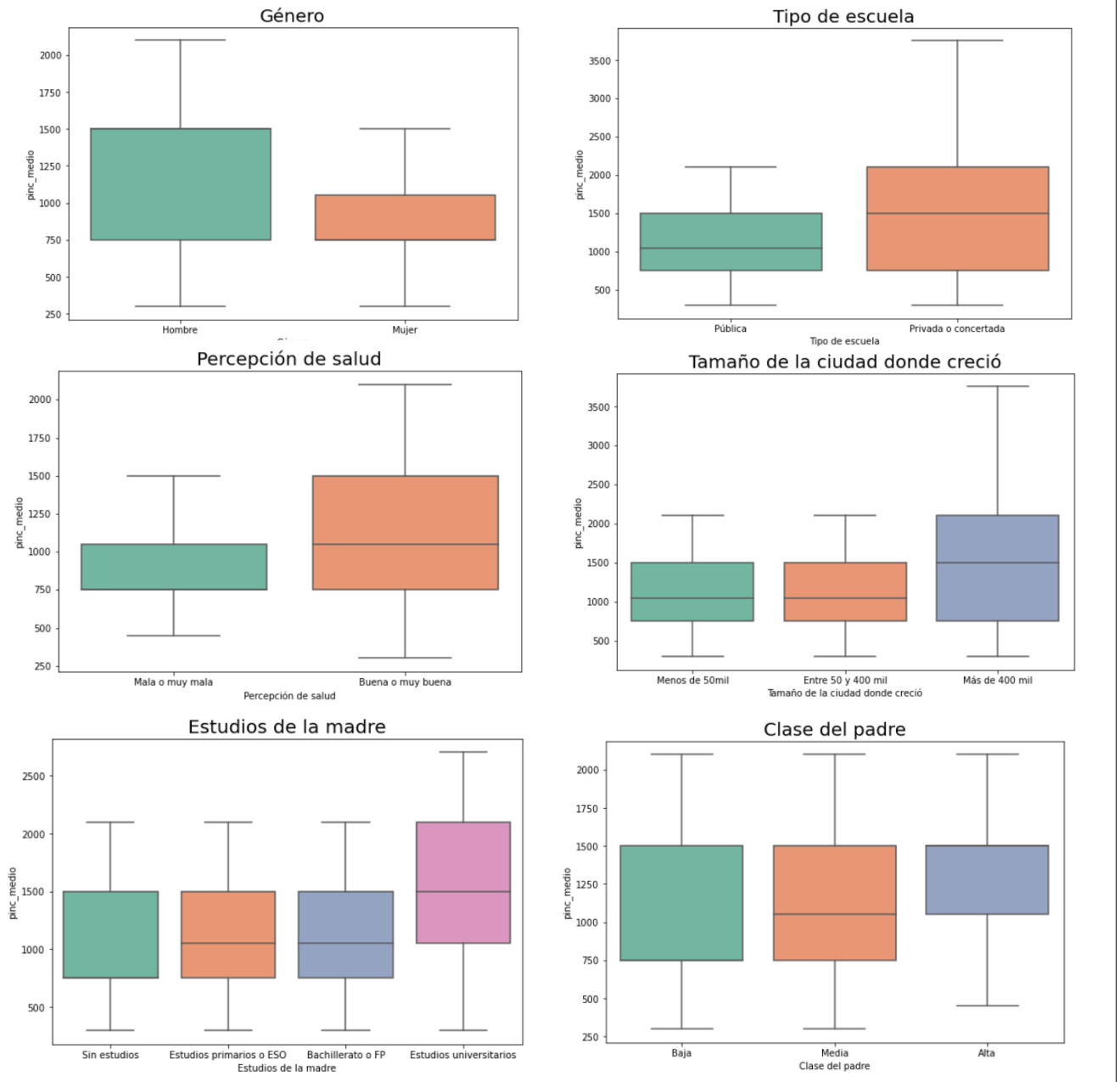
| | | | |
|---|---|---|---|
| | nivel); (0) Otro | | |
| Clasepad_med | (1) 'Padre de clase media (trabajadores de rutina manuales y no manuales); (0) Otro | 0.82 | 0.38 |
| Clasepad_baj | (1) 'Padre de clase baja (trabajadores no cualificados); (0) Otro | 0.07 | 0.26 |
| ***Circunstancias al crecer*** | | | |
| Escuela | (1) 'Estudió en escuela pública'; (0) 'Estudió en escuela privada o concertada' | 0.76 | 0.42 |
| Salud | (1) 'Percepción "Normal", "Buena" o "Muy buena" de su salud al crecer'; (0) 'Percepción "Mala", o "Muy mala" de su salud al crecer' | 0.98 | 0.13 |
| Hermanos | (1) 'Creció con 3 o más hermanos; (0) 'Creció con 2 o menos hermanos' | 0.61 | 0.49 |
| Ambiente | (1)'Creció en un ambiente con infraestructura cultural, deportiva, sanitaria y de transporte adecuada'; (0) 'No creció con alguna de estas' | 0.46 | 0.49 |
| Seguimiento | (1) 'Su familia dedicaba recursos y/o atención a su proceso de formación'; (0) 'Su familia no dedicaba atención al seguimiento de su formación' | 0.66 | 0.47 |
| ***Canales*** | | | |
| est_sin | (1) 'Sin estudios o primarios incompletos'; (0) 'Otro' | 0.06 | 0.24 |
| Est_primsecu | (1) 'Con estudios primarios o secundarios 1º etapa (bachillerato elemental, FP oficialía, FP1, FP Inicial, ESO)'; (0) 'Otro' | 0.44 | 0.5 |
| Est_secufp | (1) 'Con estudios de Secundaria 2ª etapa (bachillerato, BUP, COU) o FP (ciclos formativos medios y superiores, FP2 y maestrías)'; (0) 'Otro' | 0.26 | 0.44 |
| Est_univ | (1) Con estudios universitarios'; (0) 'Otro' | 0.24 | 0.43 |
| Ocu_directivos | (1) 'Ocupación CIUO-08=1-2 (directivos-as y profesionales)'; (0) 'Otro' | 0.19 | 0.4 |
| Ocu_tecnicos | (1) 'Ocupación CIUO-08=3 (técnicos: profesionales de apoyo)'; (0) 'Otro' | 0.13 | 0.33 |
| Ocu_cuali | (1) 'Ocupación CIUO-08=4-5-6-7-8 (cualificados-as)'; (0) 'Otro' | 0.56 | 0.5 |
| Ocu_no_cuali | (1) 'Ocupación CIUO-08=9 (no cualificados-as)'; (0) 'Otro' | 0.12 | 0.33 |
| ***Variables objetivo*** | | | |
| Pinc_medio | Ingresos personales | 1191.4 | 752.9 |
| Lnpinc_medio | Logaritmo neperiano de los ingresos personales | 6.92 | 0.57 |

Datos provenientes del CIS (2017). Elaboración propia.

Como primera aproximación a la distribución de los ingresos personales de los individuos con respecto a las circunstancias tenemos la Figura 1, en donde se exponen gráficos tipo box-plot en donde se toman en cuenta las categorías dentro de cada circunstancia y la renta asociada a cada una de ellas. Diferencias en la altura o posición de la caja indican distintos niveles de renta para las categorías en términos generales, lo cual constituiría un indicativo inicial

del impacto que tiene sobre los ingresos de los individuos características que escapan de su control, es decir, veríamos una aproximación básica pero intuitiva de la existencia de desigualdad de oportunidades en España.

**Figura 1. Distribución de los ingresos personales con respecto a las circunstancias**




Datos provenientes del CIS (2017). Elaboración propia.

Este primer vistazo a la distribución de los ingresos con respecto a este subconjunto de circunstancias nos brinda una evidencia inicial de que algunas características que escapan del control del individuo tienen impacto directo en su renta personal. Por ejemplo, destacamos el resultado del género. Al comparar las distribuciones de rentas de hombres y mujeres, observamos que los hombres gozan de mayor renta que las mujeres encuestadas, lo cual podría indicarnos que estamos ante otra evidencia empírica de brecha de género en España, en

concreto, observamos que los hombres encuestados ganan en media unos 1150 euros, mientras que las mujeres ganan unos 850 euros, preliminarmente una diferencia del 26%.

Igualmente destaca el tipo de escuela, con una mayor renta en términos generales para individuos provenientes de escuelas privadas o concertadas con respecto a la contraparte pública, con una diferencia del 30% aproximadamente. Observamos también que el tamaño de la ciudad en la que nació el individuo es relevante, con una renta considerablemente superior para los encuestados que crecieron en una ciudad grande de más de cuatrocientos mil habitantes. La salud del individuo al crecer también parece tener un impacto en la renta futura del individuo. Y, por último, observamos que mayores niveles de educación de la madre y la clase socioeconómica del padre ejercen una influencia sobre la renta de los hijos, con rentas más altas para niveles educativos superiores y clase alta respectivamente.

## 4. RESULTADOS: EFECTO DE LAS CIRCUNSTANCIAS SOBRE LA RENTA DE LOS INDIVIDUOS

En esta sección se exponen los resultados de los modelos estimados para explicar el problema de la desigualdad de oportunidades en España, lo que constituye el objetivo principal de este trabajo. En el primer subapartado tenemos un modelo MCO de referencia en el cual se incluyen todas las circunstancias, eliminando una categoría de cada variable para evitar problemas de multicolinealidad estricta. En el segundo estimamos un modelo MCO con un subconjunto de circunstancias seleccionadas mediante la técnica LASSO, directamente comparable con el primer modelo, con lo cual veremos las ventajas que nos ofrece LASSO a la hora del estudio de la DO. Finalmente, en el tercer apartado estimamos un modelo MCO con efectos cruzados también seleccionando las variables mediante LASSO, un aporte relevante de este trabajo ya que a partir de la selección por LASSO creamos un modelo con efectos cruzados pero sin sobreajuste, que suele ser un problema habitual en la práctica econométrica al tener un gran número de variables cruzadas.

### 4.1. MCO CON TODAS LAS CIRCUNSTANCIAS PRESELECCIONADAS

Estimamos en primer lugar un modelo log-lineal por MCO que contenga todas las circunstancias preseleccionadas. Este modelo, similar al estimado en Cabrera et al. (2019), servirá de referencia para compararlo con los modelos que estimaremos a continuación, en donde la selección de circunstancias se realizará a través de la herramienta LASSO. Como todas las variables explicativas son dummies, se ha de omitir una categoría de cada variable para evitar problemas de multicolinealidad estricta, lo cual tendría repercusiones serias sobre las estimaciones del modelo. El criterio para realizar esta eliminación es la de dejar fuera del modelo la "peor" categoría de cada predictor: "padest_sin" (padre sin estudios primarios completos), "madest_sin" (madre sin estudios primarios completos), "tamano_<50" (tamaño de ciudad en la que creció de un tamaño inferior a cincuenta mil habitantes), "edad_hasta30" (grupo de edad del individuo entre 18 y 30 años) y "clasepad_baj" (clase del padre baja). La Tabla 2 presenta las estimaciones por MCO de este primer modelo.

**Tabla 2. Resultados de la estimación MCO con todas las circunstancias**

| Variable | Coeficientes | Error estándar | t | P>\|t\| |
|---|---|---|---|---|
| Tamano_<50 | - | | - | - |
| Tamano_50_400 | -0,0296 | 0,034 | -0,860 | 0,390 |
| Tamano_mas400 | 0,1034** | 0,044 | 2,344 | 0,019 |
| Nacion | 0,1382** | 0,055 | 2,510 | 0,012 |
| Sexo | -0,3001*** | 0,030 | -9,845 | 0,000 |
| Edad_hasta30 | - | - | - | - |
| Edad_31_45 | 0,2569*** | 0,048 | 5,335 | 0,000 |
| Edad_46_60 | 0,2731*** | 0,053 | 5,195 | 0,000 |
| Edad_61_75 | 0,3159*** | 0,057 | 5,509 | 0,000 |
| Edad_mas75 | 0,2487*** | 0,073 | 3,429 | 0,001 |
| Padest_sin | - | - | - | - |
| Pades_primsecun | 0,2116*** | 0,047 | 4,514 | 0,000 |
| Padest_secunfp | 0,1932*** | 0,068 | 2,846 | 0,005 |
| Padest_univ | 0,2066** | 0,081 | 2,553 | 0,011 |
| Madest_sin | - | - | - | - |
| Madest_primsecun | 0,0593 | 0,046 | 1,289 | 0,198 |
| Madest_secunfp | 0,0494 | 0,071 | 0,693 | 0,488 |
| Madest_univ | 0,1378* | 0,079 | 1,734 | 0,083 |
| Clasepad_alt | -0,0096 | 0,084 | -0,114 | 0,909 |
| Clasepad_med | -0,0235 | 0,059 | -0,400 | 0,689 |
| Clasepad_baj | - | - | - | - |
| Escuela | -0,2063*** | 0,039 | -5,343 | 0,000 |
| Salud | 0,1761 | 0,117 | 1,509 | 0,132 |
| Hermanos | 0,0053 | 0,033 | 0,161 | 0,872 |
| Ambiente | 0,0047 | 0,032 | 0,150 | 0,881 |
| Seguimiento | 0,0851** | 0,036 | 2,378 | 0,018 |
| Const | 6,4483*** | 0,157 | 40,958 | 0,000 |
| **$R^2$ ajustado** | 0,185 | | | |

Nota: Asteriscos denotan significatividad individual al 99% (***), 95% (**) y 90% (*). Se omiten del análisis las variables “padest_sin”, “madest_sin”, “edad_hasta30”, “tamano_<50” y “clasepad_baj” para evitar multicolinealidad estricta. Elaboración propia.

De las 12 circunstancias y 21 variables resultantes al categorizar cada circunstancia, 12 de estas variables son significativas al 95% de confianza. Dentro de este subconjunto de variables estadísticamente significativas, destaca el sexo del individuo, con un coeficiente muy significativo de -0,3001. Recordando que tenemos como variable dependiente un logaritmo neperiano, y las variables explicativas están expresadas como dummies (toman valor 1 y 0), los coeficientes de cada variable mide la ventaja en términos de renta (en tanto por uno) de pertenecer a la categoría asociada a la variable respecto a la categoría omitida. Por ejemplo, el coeficiente de -0,3001 asociado a la categoría mujer indica que la renta de las mujeres en la muestra es un 30% inferior a la renta de los hombres (la categoría omitida). Este resultado es una evidencia empírica de brecha de género en nuestro país. Veremos, además, que este resultado es robusto a diferentes especificaciones consideradas.

Con respecto a las variables de edad, vemos un impacto positivo para los subconjuntos de 31-45, 46-60 y 61-75, y más de 75, respecto a la categoría omitida, siguiendo la forma esperable del life-cycle, ya que empieza en un 25,7% para el primer rango de edad, 27,3% para el segundo, alcanza su máximo en el rango de 61-75 con un 31,6%, y finalmente desciende el impacto positivo para la renta de los individuos mayores de 75 años con un 24,9%.

El tipo de escuela es significativo y negativo, es decir, para aquellos individuos con una educación pública durante su crecimiento, se asocia un impacto negativo del 20,6% en sus niveles de renta con respecto a individuos con educación privada o concertada. Haber crecido en una ciudad de más de cuatrocientos mil habitantes posee un impacto significativo y positivo en la renta de los individuos de un 10,3% con respecto a individuos que hayan crecido en ciudades pequeñas. Tener la nacionalidad española en nuestro país corresponde a un incremento del 13,8% en la renta con respecto a los residentes no nacionales.

Ninguna de las tres categorías dummy de la educación de la madre salieron significativas al 95% de confianza, aunque la de educación universitaria si es significativa al 90%, teniendo esta categoría una ventaja en renta de un 13,8% respecto a aquellos individuos con madres con educación inferior a primaria (la categoría omitida). Por su parte, las categorías de la educación del padre salieron significativas al 95%, todas con signo positivo y con un impacto positivo bastante similar, rondando en las tres categorías entre un 19,3% y un 21,2% más de renta para el individuo. Una alta correlación entre estas variables puede estar haciendo que las variables (o algunas categorías) no sean significativas individualmente, pero si lo sean conjuntamente. Estos aspectos se tendrán en cuenta en el proceso de selección de la variales usando LASSO con validación cruzada, ya que este procedimiento seleccionará las variables no en base a su significación individual sino a su capacidad de reducir el error cuadrático de predicción fuera de la muestra (*out of sample*). Así, podría darse el caso que varias categorías fueran indivdualmente no significativas, pero que al incluirlas conjuntamente sí permitiera reducir el error de previsión fuera de la muestra.

Es destacable que la variable "Seguimiento", que toma valor 1 si los encuestados reconocen que sus padres dedicaron recursos y atención a su formación, es significativa, y está asociada con unos niveles de renta medio de 8,5% superiores en el futuro. Por último, hay que tomar en cuenta que en este tipo de análisis en donde las variables no son completamente independientes entre sí, aún cuando la variable no sea significativa estadísticamente a nivel individual, sigue gozando de la misma interpretación que el resto de predictores y posee capacidad predictiva en conjunto con otros regresores. En este sentido, destaca la variable "Salud", ya que haber crecido con buena salud reporta a los individuos un 17,6% más de renta a futuro con respecto a individuos que hayan crecido con mala salud.

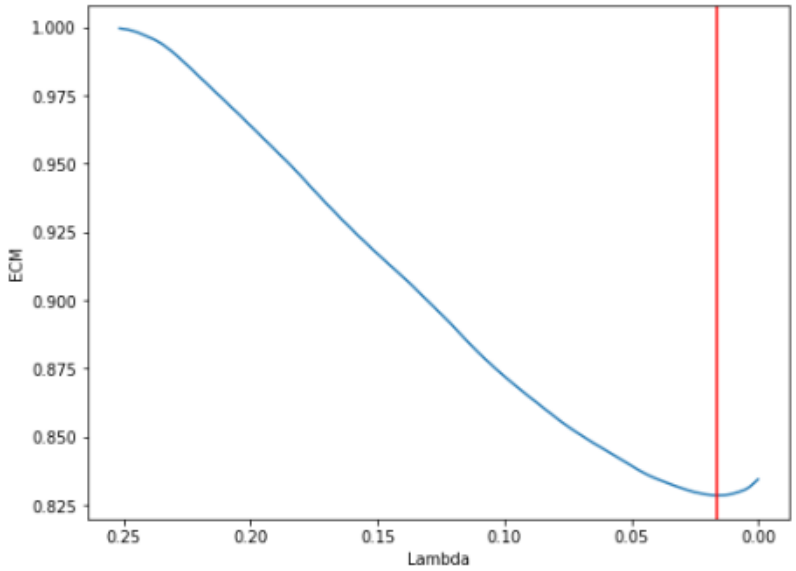
## 4.2. SELECCIÓN DE VARIABLES USANDO LASSO Y ESTIMACIÓN DEL MODELO RESULTANTE

En este apartado, realizaremos la selección de circunstancias relevantes para el estudio de desigualdad de oportunidades, para luego aplicar un MCO con ese conjunto de predictores. Por lo tanto, para el segundo modelo seleccionaremos las variables usando LASSO, incluyendo inicialmente todas las circunstancias preseleccionadas en la tabla 2. Cabe destacar en este punto que no se procederá a eliminar una categoría de cada variable dummy, ya que el propio

LASSO se encargará de seleccionar las categorías más representativas de cada predictor que en conjunto minimicen el error cuadrático medio.

Tras correr el método de validación cruzada, se seleccionó un valor del parámetro de penalización de 0.016219. Recordemos que el valor proveniente de la validación cruzada es considerado como óptimo en el sentido que es aquel que produce el modelo que minimiza el error cuadrático medio fuera de la muestra. En la Figura 2 podemos observar cómo evoluciona el error cuadrático medio a medida que se evalúan distintos valores de lambda.

**Figura 2. Evolución del error cuadrático medio con respecto al valor de lambda. Modelo MCO con circunstancias seleccionadas con LASSO**




Fuente: Elaboración propia.

Este valor de λ selecciona un total de 10 circunstancias y 13 variables, por lo que nos quedamos con 8 predictores menos con respecto al modelo estimado en la sección anterior. Uno de los resultados relevantes que nos ofrece el método LASSO es que nos permite recuperar el orden mediante el cual se van incluyendo los predictores en el modelo. Partiendo de un valor de λ suficientemente alto, en nuestro caso 0,25143, la penalización es tan grande que ninguna variable entraría en el modelo para este nivel de penalización (regularización en terminología de ML). A medida que el procedimiento va reduciendo el valor de λ, se van introduciendo variables en el modelo. En principio, las variables van entrando debido a que su aportación en ganancia de ajuste compensa la penalización impuesta. Así, las primeras variables que entran serán las que más aporten a mejorar el ajuste del modelo. Por lo tanto, podemos interpretar que las primeras variables son las que más estarían aportando a explicar la renta de los individuos, condicionado al resto de variables. Las variables incluidas, en orden de entrada con su λ asociado, son las siguientes:

**Tabla 3. Incorporación de variables según parámetro lambda. Modelo con circunstancias seleccionadas con LASSO**

| Orden | lambda | Variable incorporada |
|---|---|---|
| 1 | 0,247973 | Sexo |
| 2 | 0,234616 | Padest_sin |
| 3 | 0,195975 | Escuela |
| 4 | 0,136738 | Edad_hasta30 |
| 5 | 0,096736 | Tamano_mas400 |
| 6 | 0,092801 | Seguimiento |
| 7 | 0,073341 | Madest_sin |
| 8 | 0,067495 | Nacion |
| 9 | 0,045178 | Madest_univ |
| 10 | 0,042157 | Salud |
| 11 | 0,030661 | Edad_61_75 |
| 12 | 0,029010 | Tamano_50_400 |
| 13 | 0,017624 | Clasepad_med |

Fuente: Elaboración propia

El proceso de selección de lambda empezó con un valor de 0,25143. Podemos observar que la variable "Sexo" entra primero en el modelo, con un parámetro de penalización bastante elevado (0,247973), por lo tanto, no solo hemos constatado en el modelo MCO anterior una evidencia empírica de brecha de género, sino que también con esta herramienta vemos que el sexo es una de las variables explicativas de la desigualdad de oportunidades más relevantes. Luego, de manera destacada, tenemos que la categoría "el padre no posea estudios primarios", seguido del hecho de que el individuo haya estudiado en una escuela pública, que tenga menos de 30 años, y así sucesivamente. Nótese que con este proceso de selección de variables/categorías, los coeficientes estimados serán distintos, ya que la categoría omitida es diferente. No obstante, esto no afectará a los resultados cualitativos y a la estimación de la desigualdad de oportunidades.

Como expusimos en el apartado de metodología, para evitar los sesgos hacia cero que brinda LASSO cuando se aplica para un determinado nivel de regularización (valor de λ), sólo utilizaremos esta herramienta para realizar la selección de variables. Además, solo usamos las variables estandarizadas para LASSO. Para una mejor interpretación de los coeficientes estimados a continuación, se presentan los resultados de la estimación MCO utilizando como variables explicativas las seleccionadas por este método sin estandarizar, añadiendo en el modelo una constante. De este modo, el modelo estimado es comparable con el obtenido en la Tabla 2, teniendo en cuenta que las categorías omitidas pueden ser distintas en ambos modelos, tal y como comentamos anteriormente.

**Tabla 4. Resultados MCO tras LASSO con circunstancias seleccionadas**

| Variable | Coeficientes | Error estándar | t | P>\|t\| |
|---|---|---|---|---|
| Tamano_<50 | - | | - | - |
| Tamano_50_400 | -0,0303 | 0,034 | -0,894 | 0,372 |
| Tamano_mas400 | 0,1034** | 0,043 | 2,387 | 0,017 |
| Nacion | 0,1404*** | 0,054 | 2,610 | 0,009 |
| Sexo | -0,3019*** | 0,030 | -10,001 | 0,000 |
| Edad_hasta30 | -0,2648*** | 0,044 | -5,982 | 0,000 |
| Edad_31_45 | - | - | - | - |
| Edad_46_60 | - | - | - | - |
| Edad_61_75 | 0,0539 | 0,040 | 1,360 | 0,174 |
| Edad_mas75 | - | - | - | - |
| Padest_sin | -0,2075*** | 0,046 | -4,549 | 0,000 |
| Pades_primsecun | - | - | - | - |
| Padest_secunfp | - | - | - | - |
| Padest_univ | - | - | - | - |
| Madest_sin | -0,0580 | 0,045 | -1,282 | 0,200 |
| Madest_primsecun | - | - | - | - |
| Madest_secunfp | - | - | - | - |
| Madest_univ | 0,0760 | 0,062 | 1,228 | 0,220 |
| Clasepad_alt | - | - | - | - |
| Clasepad_med | -0,0181 | 0,040 | -0,453 | 0,651 |
| Clasepad_baj | - | - | - | - |
| Escuela | -0,2042*** | 0,038 | -5,437 | 0,000 |
| Salud | 0,1774 | 0,116 | 1,531 | 0,126 |
| Hermanos | - | - | - | - |
| Ambiente | - | - | - | - |
| Seguimiento | 0,0835** | 0,035 | 2,395 | 0,017 |
| Const | 6,9743*** | 0,137 | 50,819 | 0,000 |
| **$R^2$ ajustado** | 0,19 | | | |

Nota: Asteriscos denotan significatividad individual al 99% (***), 95% (**) y 90% (*). Aquellas categorías sin valores estimados en la tabla fueron omitidas tras el proceso de selección de LASSO. Elaboración propia.

Podemos observar que con 8 variables menos mejoramos muy ligeramente el $R^2$ ajustado del modelo anterior, pasando de 0,185 a 0,19. De 13 variables seleccionadas, 7 son estadísticamente significativas al 95% de confianza. Igualmente cabe acotar que para todas las variables en común con el modelo anterior, los coeficientes son bastante similares, y gozan de la misma interpretación.

La diferencia fundamental de este modelo con respecto al anterior, además del número de variables seleccionadas, es lo que ocurre con la edad y los estudios del padre. LASSO ha eliminado todas las categorías de los estudios del padre menos la de padre sin estudios, la cual es significativamente negativa, ya que individuos que cumplan con esta condición poseen en término medio un 20,7% menos de renta con respecto a individuos con padre con estudios, igualmente, vemos que en edad la única variable seleccionada estadísticamente significativa es la del grupo de edad de 18 a 30 años, que es considerablemente negativa, con un 26,5% menos de renta para individuos de este grupo con respecto a los demás. Igualmente, vemos que se

seleccionan dos categorías de los estudios de la madre: Si la madre no terminó la primaria, los individuos en promedio ganarían 5,8% menos de renta en el futuro con respecto a individuos con madres con estudios, mientras que si son universitarias, los encuestados ganarían en promedio 7,6% más con respecto a madres sin educación terciaria.

Podemos observar las múltiples ventajas que nos ofrece LASSO, disminuyendo 8 variables ganando ligeramente en $R^2$ ajustado, y nos permitió observar el orden en el cual entran las variables explicativas, indicativo de su relevancia para nuestro modelo. Además, LASSO seleccionó las categorías más importantes para el estudio, indicándonos que la categorías que generan mejores ajustes son, por ejemplo, las de "edad_hasta30" y "padest_sin", es decir, la categoría de individuos con una edad entre 18 y 30 años e individuos con padres sin estudios primarios completos; la reducción de categorías en el modelo reduce problemas de colinealidad entre categorías que al final puede terminar implicando un aumento de la incertidumbre en las estimaciones y un exceso de error cuando hacemos predicciones fuera de la muestra.

## 4.3. ANÁLISIS DE EFECTOS CRUZADOS: SELECCIÓN MEDIANTE LASSO Y ESTIMACIÓN POR MCO POSTERIOR

En este apartado realizaremos el análisis de efectos cruzados, donde seleccionaremos mediante LASSO los efectos cruzados más relevantes para posteriormente incorporarlas a un MCO que permita observar las relaciones entre circunstancias que tienen impacto significativo sobre la renta.

La motivación de este apartado proviene de una crítica habitual que se realiza a la práctica econométrica convencional que es el uso constante de métodos lineales, ya que se puede perder el análisis de los impactos cruzados. Una alternativa es considerar polinomios de efectos cruzados entre todas las variables incluidas en el modelo, pero el mayor inconveniente que posee el introducir todos los efectos cruzados es la gran cantidad de variables que se terminan creando, lo cual llevará inevitablemente a un problema de sobreajuste en el modelo. Para ello, y explotando al máximo las capacidades selectivas de LASSO, se crearon variables cruzadas entre todas las circunstancias preseleccionadas. En total se crearon 353 variables, entre los predictores por sí solos y las variables cruzadas. Aplicando la validación cruzada nos encontramos con un parámetro de penalización óptimo de 0.00692, con el cual nos quedamos finalmente con 28 variables de las 353 iniciales.

Este ejercicio del modelo MCO con variables cruzadas nos permitirá analizar el problema de la desigualdad de oportunidades de una manera más completa, ya que estudiando los efectos cruzados podemos encontrarnos con relaciones importantes entre circunstancias que pueden escaparse del análisis si se emplean métodos de regresión lineales. Estas circunstancias cruzadas luego pueden considerarse a la hora de realizar políticas públicas redistributivas o cualquiera que pretenda atacar de manera directa el problema de la desigualdad de oportunidades.

En la siguiente tabla se recogen los resultados del MCO de las variables cruzadas:

**Tabla 5. Resultados MCO con variables cruzadas seleccionadas con LASSO**

| Variable | Coeficientes | Error estándar | t | P>\|t\| |
|---|---|---|---|---|
| Sexo | -0,2259*** | 0,064 | -3,520 | 0,000 |
| Tamano_50_400 Sexo | -0,0328 | 0,056 | -0,584 | 0,559 |
| Sexo Escuela | -0,0785 | 0,070 | -1,121 | 0,262 |
| Escuela | -0,0352 | 0,068 | -0,515 | 0,607 |
| Tamano_<50 Escuela | -0,0666 | 0,052 | -1,282 | 0,200 |
| Padest_sin Escuela | -0,0811 | 0,092 | -0,879 | 0,380 |
| Escuela Hermanos | -0,1409** | 0,067 | -2,101 | 0,036 |
| Edad_hasta30 | 0,1741 | 0,528 | 0,330 | 0,742 |
| Nacion edad_hasta30 | -0,1952 | 0,126 | -1,544 | 0,123 |
| Edad_hasta30 Salud | -0,1434 | 0,520 | -0,276 | 0,783 |
| Edad_hasta30 Seguimiento | -0,1294 | 0,133 | -0,975 | 0,330 |
| Padest_sin | 0,1819 | 0,171 | 1,063 | 0,288 |
| Tamano_<50 padest_sin | -0,0466 | 0,065 | -0,716 | 0,474 |
| Padest_sin clasepad_med | -0,0871 | 0,067 | -1,294 | 0,196 |
| Padest_sin Seguimiento | -0,1173* | 0,067 | -1,739 | 0,082 |
| Madest_sin | -0,0007 | 0,065 | -0,010 | 0,992 |
| Tamano_<50 Ambiente | 0,0744* | 0,041 | 1,796 | 0,073 |
| Tamano_50_400 Ambiente | -0,0748 | 0,048 | -1,558 | 0,119 |
| Tamano_mas400 edad_46_60 | 0,1804** | 0,075 | 2,405 | 0,016 |
| Tamano_mas400 madest_univ | 0,3210*** | 0,107 | 3,013 | 0,003 |
| Tamano_mas400 Seguimiento | 0,0432 | 0,059 | 0,726 | 0,468 |
| Nacion padest_sin | -0,1487 | 0,133 | -1,121 | 0,262 |
| Nacion Salud | 0,1068 | 0,074 | 1,444 | 0,149 |
| Nacion Hermanos | 0,0727 | 0,067 | 1,088 | 0,277 |
| Nacion Seguimiento | 0,1236** | 0,049 | 2,519 | 0,012 |
| Edad_61_75 Seguimiento | 0,1108** | 0,054 | 2,036 | 0,042 |
| Madest_sin Hermanos | -0,0348 | 0,087 | -0,401 | 0,688 |
| Madest_primsecun Hermanos | 0,0565 | 0,062 | 0,906 | 0,365 |
| Const | 7,0226*** | 0,079 | 89,199 | 0,000 |
| **$R^2$ ajustado** | 0,211 | | | |

Nota: Asteriscos denotan significatividad individual al 99% (***), 95% (**) y 90% (*). Elaboración propia.

En este punto, la interpretación de los coeficientes es un poco más compleja. El impacto de un predictor sobre nuestra variable independiente seguirá siendo una semielasticidad en los n niveles de renta del individuo, pero ahora muchos predictores tendrán efectos cruzados, por lo que el efecto global de una variable independiente sobre la renta puede que dependa también de otras variables correlacionadas, que pueden amplificar o disminuir el impacto de la circunstancia original.

Un ejemplo ilustrativo de lo encontrado lo tenemos con el género de los individuos. Según estas estimaciones de este último modelo, el punto de desventaja por ser mujer (respecto a ser hombre) es de un 22,6% menos de renta. Sin embargo, si una mujer, además, estudió en una escuela pública (respecto a privada y concertada), el impacto negativo sobre la renta se amplifica en unos 7,9 puntos porcentuales más (alcanza el 30% de diferencial), y si además creció en una ciudad de entre cincuenta mil y cuatrocientos mil habitantes, la diferencia aumenta aún más en otros 3,3 puntos porcentuales adicionales. La combinación de ambos aspectos,

haber estudiado en colegio público y haber crecido en ciudades de entre cincuenta mil y cuatrocientos mil habitantes, supone un diferencial de casi 35% (superior al 30% que obtuvimos en media en la Tabla 2). Por lo mismo, los diferenciales por ser mujer, disminuyen si estudiaron en centros concertados-privados, etc. Podemos ver en este punto algo consistente con la teoría de desigualdad de oportunidades, y es que si se cumplen numerosas circunstancias con impactos negativos sobre la renta del individuo, el impacto global será mayor.

La variable "Escuela", es decir, si el individuo estudió en una escuela pública, posee un impacto negativo del 3,5% frente a estudiar en privado-concertado. Esta disminución de la renta de los individuos se agrava si creció en una ciudad de menos de cincuenta mil habitantes, con un 6,7% menos de renta con respecto a individuos que crecieron en ciudades más grandes, si el padre no completó los estudios primarios, con un 8,1% con respecto a padres con estudios, y si el individuo posee 3 o más hermanos, con un 14,1% menos de ingresos con respecto a su contrapartida.

El impacto positivo de haber crecido en una ciudad con más de cuatrocientos mil habitantes depende de si el individuo pertenece al grupo de edad entre 46 y 60 años, con un 18% con respecto a otros grupos de edad, si su madre completó estudios universitarios, con un 32,1% respecto a madre sin estudios terciarios, y si sus padres prestaron atención o dirigieron recursos a su formación, con un 4,3% más de renta con respecto a su contrapartida.

El impacto de la nacionalidad española sobre la renta es algo difusa y muy dependiente de otros predictores, ya que por sí sola no aparece como variable independiente de la regresión. El impacto de la nacionalidad depende positivamente de haber crecido con buena salud, con un 10,7% más de renta con respecto a individuos que hayan crecido con mala salud, si posee tres o más hermanos, con un 7,3%, y que sus padres hayan dedicado recursos y atención a su educación, con un 12,4%. Por otra parte, efectos contrarios ligados a la nacionalidad son haber tenido un padre sin estudios primarios, con un 14,9% menos de renta, o tener menos de 30 años, con un 19,5% de impacto negativo.

Un resultado interesante es lo que ocurre a los individuos que crecieron en una ciudad con menos de cincuenta mil habitantes. Por sí sola esta característica no aparece como regresor, sino que aparece como variable cruzada con otros predictores. Si el individuo además de haber crecido en este tipo de ciudades tuvo un padre sin estudios primarios, tiene un impacto negativo en la renta del 4,7%, si estudió en una escuela pública se amplifica el impacto negativo a un 6,7%, sin embargo, el efecto se mitiga enormemente si la ciudad posee infraestructuras culturales, sanitarias, deportivas y de transporte adecuadas, ya que si se cumplen estas condiciones, el individuo esperaría una mejora en su renta del 7,4%.

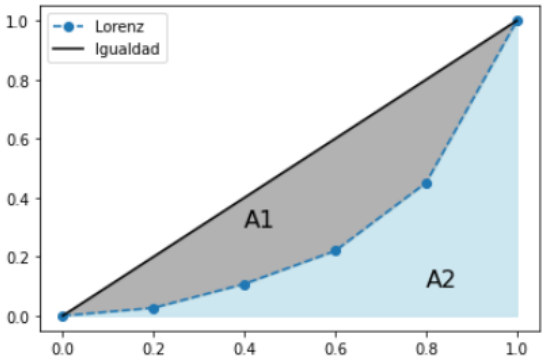
En este punto podemos observar la gran utilidad de realizar un análisis de efectos cruzados, ya que existen relaciones entre los distintos predictores que, si se cumplen ese conjunto de circunstancias, se pueden ampliar o mitigar los efectos de las circunstancias sobre la renta de los individuos, un análisis que se nos escapa al realizar un MCO lineal convencional. Igualmente observamos la potencia y utilidad del método de selección LASSO con la validación cruzada en este caso, ya que hemos sido capaces de pasar de estimar un MCO con 353 variables debido al enorme número de efectos cruzados, lo que claramente generaría un problema de sobreajuste, a pasar a tener un modelo bastante parsimonioso de tan sólo 28 variables. Esto permite afinar mucho en la comprensión de qué variables y qué efectos cruzados son los más relevantes y en la estimación de la desigualdad de oportunidades.

## 5. ESTIMACIÓN DE LA DESIGUALDAD DE OPORTUNIDADES

En este apartado se realiza la estimación paramétrica de la desigualdad de oportunidades, lo cual constituye uno de los objetivos de este estudio.

Como se expuso en el apartado de metodología, si aplicamos un índice de desigualdad a la variable de ingreso personal de los encuestados analizados en el estudio, obtendríamos una medida de la desigualdad total de ingresos presente en el conjunto de individuos. De forma que se procedió a construir el respectivo índice de Gini derivada de la siguiente curva de Lorenz, que proviene directamente de los datos reales de ingresos de las observaciones utilizadas en este trabajo, y no de las estimaciones por MCO realizadas a lo largo del estudio:

**Figura 3. Curva de Lorenz del ingreso personal de los encuestados**




Construido a partir de datos provenientes del CIS (2017).

Se computó que el índice de Gini de la variable de ingreso personal de los individuos es de 0,3058 para los datos de ingresos reales sacados directamente de la encuesta.

Por otra parte, si construimos el índice de Gini de la variable de ingreso personal explicada, es decir, de la $\exp(\beta C_i)$ estimada de cada modelo, tendríamos una medida de la desigualdad de ingresos derivada del conjunto de circunstancias seleccionadas. Por lo tanto, se procedió a calcular el índice de Gini de todas las variables explicadas de los tres modelos, para tener un cálculo de la desigualdad de oportunidades y del ratio desigualdad de oportunidades sobre desigualdad total en cada uno de ellos. En la siguiente tabla se exponen los resultados.

**Tabla 6. Estimaciones de desigualdad de oportunidades y ratio DO/DT**

| Índice de Gini | MCO inicial con todas las circunstancias | MCO con circunstancias seleccionadas con Lasso | MCO con variables cruzadas seleccionadas con LASSO | MCO con todas las variables cruzadas |
|---|---|---|---|---|
| Desigualdad total (DT) | 0,3058 | 0,3058 | 0,3058 | 0,3058 |
| Desigualdad de oportunidades (DO) | 0,14047 | 0,1407 | 0,1528 | 0,1806 |
| Ratio DO/DT | 0,4594 | 0,4601 | 0,4993 | 0,5908 |

Elaboración propia a partir de datos del CIS (2017).

Podemos observar que en los tres primeros modelos los resultados son bastante similares, y corresponden a que entre un 46% y un 49% de la desigualdad total en los ingresos se debe a la desigualdad de oportunidades, es decir, viene explicada por el conjunto de circunstancias seleccionadas en cada modelo. Si bien es cierto que las ganancias en porcentaje del ratio son francamente marginales, lo que expone esta tabla es que la proporción de la desigualdad total que se debe a la desigualdad de oportunidades es ligeramente inferior al 50%, un resultado consistente con los obtenidos por autores como Cabrera et al. (2019).

Por otra parte, en la última columna tenemos el cálculo de la ratio DO/DT para un modelo MCO con todas las variables cruzadas. En total se usaron 252 variables cruzadas más una constante en ese modelo, ya que se combinaron todas menos una categoría de cada variable para evitar problemas de multicolinealidad estricta. Vemos que el resultado de la desigualdad de oportunidades sobre desigualdad total varía considerablemente, siendo 10% superior al resto de ratios, lo cual apunta a un problema de sobreajuste claro.

Es en este punto donde constatamos la utilidad de la herramienta LASSO, ya que sin ella hubiésemos creado un modelo MCO con variables cruzadas que sesgaría al alza los resultados de las estimaciones de desigualdad de oportunidad y de la ratio DO/DT, debido a la enorme cantidad de predictores que se incluyeron en el modelo. Sin embargo, tras seleccionar las variables con LASSO, obtuvimos un modelo explicativo con circunstancias cruzadas válido, lo cual enriquece enormemente el estudio de la desigualdad de oportunidades al brindarnos una visión de las relaciones que existen entre distintas circunstancias, mientras que seguimos teniendo una estimación de la DO consistente con el resto de modelos, ya que no se produce el sesgo al alza por sobreajuste.

## 6. LOS CANALES DE LA EDUCACIÓN Y LA OCUPACIÓN: UNA PRIMERA APROXIMACIÓN AL PROBLEMA

En este apartado analizamos los canales de educación y ocupación de los individuos, que en el contexto del estudio de desigualdad de oportunidades corresponden a un conjunto de variables mediante las cuales las circunstancias pueden influir sobre la renta de los individuos. En este sentido, el estudio de los canales cobra suma relevancia, ya que constituyen variables objetivo sobre las cuales se pueden crear políticas públicas de diversa índole para atacar el problema de desigualdad de oportunidades. Influir directamente sobre las circunstancias constituye un objetivo muy difícil a corto plazo, ya que por definición son situaciones o características que escapan del control de los individuos. Sin embargo, se pueden crear leyes, reglamentos o normativas a nivel educativo o laboral, para evitar que las circunstancias de los individuos deriven irremediablemente en un impacto en su renta. En este sentido, cabe acotar que el estudio de los canales en este trabajo es relativamente superficial, siendo más que todo un análisis preliminar de los mismos en el caso español.

La metodología consiste en realizar una selección de variables entre circunstancias y canales con LASSO para obtener aquellas variables que, juntas, minimicen el error cuadrático medio fuera de la muestra, a partir de aquí, tendremos un punto de partida para comparar con los modelos realizados anteriormente. Si las variables educativas y ocupacionales son sumamente significativas y otras circunstancias dejan de ser estadísticamente significativas o

varían considerablemente en cuanto a magnitud, puede interpretarse este hecho como que parte del efecto de las circunstancias sobre la renta de los individuos se canaliza a través de estas nuevas variables incorporadas, por lo que veremos de manera preliminar cuáles circunstancias canalizan su efecto sobre la educación, la ocupación o ambas.

Al ser un ejercicio preliminar, nos centramos en el modelo sin efectos cruzados. Por lo tanto, incorporamos en LASSO todas las circunstancias y los canales (educación y ocupación del individuo) de manera lineal, y tras aplicar la validación cruzada se obtiene un lambda óptimo de 0.0139016. Con este parámetro pasamos de 27 variables/categorías a 20. En la siguiente tabla se muestran las variables finalmente seleccionadas organizadas de acuerdo a su orden de incorporación en el modelo:

**Tabla 7. Incorporación de variables según parámetro lambda. Modelo con circunstancias y canales seleccionadas con LASSO**

| Orden | lambda | Variable incorporada |
|---|---|---|
| 1 | 0,364671 | Ocu_directivos |
| 2 | 0,354714 | Est_univ |
| 3 | 0,250944 | Sexo |
| 4 | 0,195596 | Ocu_no_cuali |
| 5 | 0,180007 | Padest_sin |
| 6 | 0,158921 | Est_primsecu |
| 7 | 0,140305 | Escuela |
| 8 | 0,127347 | Edad_hasta30 |
| 9 | 0,112429 | Ocu_tecnicos |
| 10 | 0,084067 | Est_sin |
| 11 | 0,082912 | Padest_primsecun |
| 12 | 0,064625 | Madest_primsecun |
| 13 | 0,057055 | tamano_mas400 |
| 14 | 0,046356 | Tamano_50_400 |
| 15 | 0,034662 | Salud |
| 16 | 0,034186 | Nacion |
| 17 | 0,025918 | Hermanos |
| 18 | 0,022257 | Clasepad_baj |
| 19 | 0,020202 | Edad_61_75 |
| 20 | 0,019650 | Seguimiento |

Fuente: Elaboración propia

Podemos observar a priori que de las primeras cuatro variables seleccionadas, 3 constituyen canales: “ocu_directivos” (ocupación directivos), “est_univ” (estudios universitarios), “ocu_no_cuali” (ocupación no cualificados), y la otra nuevamente es el sexo. Esto da una idea de lo relevante que son estos dos canales en la determinación de la renta, como era esperable, y también de lo importante que es la circunstancia género, la cual puede que tenga efecto a través de estos dos canales, pero aún ejerce efectos significativos sobre la renta que va más allá de los mismos. Igualmente, cabe destacar que variables como el tamaño de la ciudad mayor a cuatrocientos mil habitantes o el seguimiento pierden importancia en cuanto a incorporación en el modelo, lo cual era de esperar de acuerdo a la teoría expuesta anteriormente.

En la siguiente tabla podemos encontrar los resultados del MCO con las variables seleccionadas:

**Tabla 8. Resultados MCO del modelo con circunstancias y canales seleccionadas con LASSO**

| Variable | Coeficientes | Error estándar | t | P>\|t\| |
|---|---|---|---|---|
| ***Circunstancias*** | | | | |
| Tamano_<50 | - | - | - | - |
| Tamano_50_400 | -0,0499 | 0,031 | -1,598 | 0,110 |
| Tamano_mas400 | 0,0416 | 0,040 | 1,033 | 0,302 |
| Nacion | 0,0571 | 0,051 | 1,125 | 0,261 |
| Sexo | -0,2908*** | 0,028 | -10,383 | 0,000 |
| Edad_hasta30 | -0,2154*** | 0,041 | -5,233 | 0,000 |
| Edad_31_45 | - | - | - | - |
| Edad_46_60 | - | - | - | - |
| Edad_61_75 | 0,0343 | 0,037 | 0,929 | 0,353 |
| Edad_mas75 | - | - | - | - |
| Padest_sin | -0,0462 | 0,048 | -0,961 | 0,337 |
| Pades_primsecun | 0,0819** | 0,041 | 1,990 | 0,047 |
| Padest_secunfp | - | - | - | - |
| Padest_univ | - | - | - | - |
| Madest_sin | - | - | - | - |
| Madest_primsecun | 0,0383 | 0,033 | 1,162 | 0,246 |
| Madest_secunfp | - | - | - | - |
| Madest_univ | - | - | - | - |
| Clasepad_alt | - | - | - | - |
| Clasepad_med | - | - | - | - |
| Clasepad_baj | 0,0606 | 0,054 | 1,121 | 0,262 |
| Escuela | -0,1208*** | 0,035 | -3,426 | 0,001 |
| Salud | 0,1407 | 0,106 | 1,323 | 0,186 |
| Hermanos | 0,0481 | 0,030 | 1,611 | 0,107 |
| Ambiente | - | - | - | - |
| Seguimiento | 0,0179 | 0,033 | 0,548 | 0,584 |
| Const | 6,8747*** | 0,129 | 53,106 | 0,000 |
| ***Canales*** | | | | |
| est_sin | -0,2257*** | 0,067 | -3,362 | 0,001 |
| Est_primsecu | -0,1315*** | 0,037 | -3,547 | 0,000 |
| Est_secufp | - | - | - | - |
| Est_univ | 0,1971*** | 0,045 | 4,385 | 0,000 |
| Ocu_directivos | 0,2750*** | 0,046 | 5,963 | 0,000 |
| Ocu_tecnicos | 0,1270*** | 0,046 | 2,786 | 0,005 |
| Ocu_cuali | - | - | - | - |
| Ocu_no_cuali | -0,1802*** | 0,045 | -4,031 | 0,000 |
| **$R^2$ ajustado** | 0,320 | | | |

Nota: Asteriscos denotan significatividad individual al 99% (***), 95% (**) y 90% (*). Las variables que no poseen estimaciones fueron excluidas del modelo tras la selección de variables. Elaboración propia.

En cuanto a los resultados, tenemos un $R^2$ ajustado de 0,32, lo cual es un incremento considerable con respecto a modelos anteriores. De las 20 variables seleccionadas, 10 son estadísticamente significativas, y dentro de este subconjunto solo 4 son circunstancias. Las circunstancias que han seguido siendo significativas son el sexo, con un impacto negativo del 29,1% sobre la renta de las mujeres encuestadas con respecto a su contrapartida masculina, la edad hasta 30 años, con una influencia negativa del 21,5% para los individuos de este grupo de edad con respecto a otros, los estudios primarios o de ESO del padre, con un impacto positivo del 8,2% con respecto a padres con otros estudios, y el tipo de escuela, con un impacto negativo del 12,1% para individuos procedentes de escuelas públicas.

Con respecto a los canales, vemos que todas las categorías son estadísticamente significativas, salvo una que LASSO no incluye para evitar la multicolinealidad exacta, que en el caso de la ocupación es la ocupación cualificada y en el caso de los estudios son los de bachillerato o FP. Observamos que la renta del individuo es proporcional al nivel de ocupación que posee de acuerdo a la clasificación CIUO con respecto a la categoría omitida, con un impacto negativo del 18,02% si el individuo posee un trabajo no cualificado, otro positivo del 12,70% para trabajadores técnicos, y del 27,50% para directivos. Por otra parte, la relación con los estudios también es prácticamente lineal con respecto a la categoría omitida, con un impacto negativo del 22,57% para individuos sin estudios primarios, del 13,15% para individuos con primaria o ESO, y como era de esperar, un impacto positivo del 19,71% para los encuestados con estudios universitarios.

Observamos en este punto cómo muchas circunstancias dejaron de ser estadísticamente significativas a nivel individual. La explicación es que los impactos de ciertas circunstancias se ven absorbidos por los canales, los cuales reciben la influencia negativa o positiva de las circunstancias y lo termina reflejando en menor o mayor renta para los individuos. Como se comentó anteriormente, la importancia de este análisis es su impacto sobre la creación de políticas públicas, ya que es sumamente complejo actuar sobre las circunstancias y su papel sobre la renta, ya que las circunstancias por definición escapan del control de los individuos. Sin embargo, sí se puede actuar con mayor facilidad sobre los canales, que en este caso es el canal educativo y ocupacional. A través de leyes laborales o reformas educativas se pueden atacar los principales problemas de desigualdad de oportunidades para que las circunstancias no lleguen a tener un impacto sobre la renta de los individuos, de esta teoría surgen propuestas como las cuotas de género, el *affirmative action* en países como Estados Unidos, entre otras políticas.

Aunque sea usando este procedimiento sencillo, es relevante distinguir a través de qué canales se focalizan las distintas circunstancias, por ello, en la siguiente tabla se exponen los resultados de todos los modelos MCO estimados, incluyendo de uno a uno y conjuntamente los distintos canales. De esta manera podemos analizar qué circunstancias apuntan a ser canalizadas por cada canal por separado, y de manera conjunta. No obstante, tenemos en cuenta que, al existir correlación entre los distintos canales, puede que salgan resultados similares.

**Tabla 9. Comparación de resultados MCO. Modelos con y sin canales**

| Variable | MCO inicial con todas las circunstancias | MCO con circunstancias seleccionadas con LASSO | MCO con educación y circunstancias seleccionadas con LASSO | MCO con ocupación y circunstancias seleccionadas con LASSO | MCO con circunstancias y canales seleccionadas con LASSO |
|---|---|---|---|---|---|
| ***Circunstancias*** | | | | | |
| Tamano_<50 | - | - | - | - | - |
| Tamano_50_400 | -0,0296 | -0,0303 | -0,0513* | -0,0396 | -0,0499* |
| Tamano_mas400 | 0,1034** | 0,1034** | 0,0554 | 0,0535 | 0,0416 |
| Nacion | 0,1382** | 0,1404*** | 0,1217** | 0,0373 | 0,0571 |
| Sexo | -0,3001*** | -0,3019*** | -0,3203*** | -0,2686*** | -0,2908*** |
| Edad_hasta30 | - | -0,2648*** | -0,2411*** | -0,2245*** | -0,2154*** |
| Edad_31_45 | 0,2569*** | - | -0,0389 | - | - |
| Edad_46_60 | 0,2731*** | - | - | 0,0108 | - |
| Edad_61_75 | 0,3159*** | 0,0539 | 0,0605 | - | 0,0343 |
| Edad_mas75 | 0,2487*** | - | - | -0,0496 | - |
| Padest_sin | - | -0,2075*** | -0,0690 | -0,1395** | -0,0462 |
| Pades_primsecun | 0,2116*** | - | 0,0853** | 0,0155 | 0,0819** |
| Padest_secunfp | 0,1932*** | - | - | -0,0416 | - |
| Padest_univ | 0,2066** | - | - | - | - |
| Madest_sin | - | -0,058 | - | -0,0099 | - |
| Madest_primsecun | 0,0593 | - | 0,0318 | 0,0310 | 0,0383 |
| Madest_secunfp | 0,0494 | - | - | - | - |
| Madest_univ | 0,1378* | 0,0760 | - | - | - |
| Clasepad_alt | -0,0096 | - | - | - | - |
| Clasepad_med | -0.0235 | -0,0181 | - | - | - |
| Clasepad_baj | - | - | - | 0,0587 | 0,0606 |
| Escuela | -0,2063*** | -0,2042*** | -0,1404*** | -0,1412*** | -0,1208*** |
| Salud | 0,1761 | 0,1774 | 0,1585 | 0,1411 | 0,1407 |
| Hermanos | 0,0053 | - | 0,0341 | 0,0335 | 0,0481* |
| Ambiente | 0,0047 | - | - | - | - |
| Seguimiento | 0,0851** | 0,0835** | 0,0196 | 0,0531 | 0,0179 |
| Const | 6,4483*** | 6,9743*** | 6,8953*** | 6,8956*** | 6,8747*** |
| ***Canales*** | | | | | |
| est_sin | - | - | -0,2974*** | - | -0,2257*** |
| Est_primsecu | - | - | -0,1754*** | - | -0,1315*** |
| Est_secufp | - | - | - | - | - |
| Est_univ | - | - | 0,3309*** | - | 0,1971 |
| Ocu_directivos | - | - | - | 0,4489*** | 0,2750*** |
| Ocu_tecnicos | - | - | - | 0,2091*** | 0,1270*** |
| Ocu_cuali | - | - | - | - | - |
| Ocu_no_cuali | - | - | - | -0,1920*** | -0,1802*** |
| $R^2$ | 0,199 | 0,199 | 0,297 | 0,302 | 0,331 |

Nota: Asteriscos denotan significatividad individual al 99% (***), 95% (**) y 90% (*). Las variables que no poseen estimaciones fueron excluidas del modelo tras la selección de variables. Elaboración propia.

Comentamos en las siguientes subsecciones los resultados más relevantes derivados del análisis de esta tabla: ¿cuáles son las circunstancias más relevantes canalizadas por la educación y cuáles por la ocupación?

## 6.1. CIRCUNSTANCIAS CANALIZADAS POR LA EDUCACIÓN

Para analizar el papel de la educación del individuo como potencial canalizador, comparamos los resultados de la columna 2 y la 4 de la tabla 8. En esta comparativa, podemos observar que haber crecido en una ciudad de más de cuatrocientos mil habitantes es significativa al 95% en el modelo MCO inicial, con un 10,3% más de renta para los individuos, sin embargo, al añadir la educación esta variable deja de ser significativa y cae hasta el 5,5%, por lo tanto, la educación es uno de los medios por los cuales se puede mitigar las diferencias en renta derivadas de haber crecido o no en una gran ciudad.

Por otra parte, el hecho de haber tenido un padre sin estudios en el modelo MCO inicial era sumamente significativa al 99%, disminuyendo la renta en un 20,8%, sin embargo, al incluir variables educativas del individuo, el impacto deja de ser estadísticamente significativo a nivel individual y cae al -6,9% de renta para el individuo. La traducción de esto a políticas redistributivas es que la mejor forma de aliviar las consecuencias negativas sobre la renta de tener un padre sin estudios de primaria completados, es a través de la consecución de mayores niveles educativos por parte del individuo, especialmente si tomamos en cuenta que en el modelo MCO con ocupación no ocurre esta disminución tan abrupta del coeficiente de padest_sin, por lo que ese no es el medio canalizador de esta circunstancia, sino la educación. Evitar la trasmisión intergeneracional educativa es clave en este aspecto.

Igualmente se ven mitigados impactos como haber estudiado en una escuela pública, pasando de una disminución del 20,4% en la renta a un 14%, y de manera considerable podemos ver que la variables seguimiento, en la cual toma valor 1 si los padres dedicaban recursos o atención a la formación del individuo era estadísticamente significativa en el MCO inicial con un impacto positivo del 8,4%, mientras que en el nuevo modelo deja de ser individualmente significativa y el impacto se reduce hasta el 2% aproximadamente.

## 6.2. CIRCUNSTANCIAS CANALIZADAS POR LA OCUPACIÓN

Para analizar el papel de la ocupación del individuo como potencial canalizador, comparamos los resultados de la columna 2 y la 5 de la tabla 8.

Podemos observar que la ocupación hace que el impacto positivo de haber crecido en una ciudad de más de cuatrocientos mil habitantes disminuya de un 10,34% a un 5,35%, por lo que tanto la educación como la ocupación son canales importantes para equilibrar el terreno de juego para individuos con diferencias en esta circunstancia.

Por otra parte, podemos observar que el hecho de ser nacional español, que tanto en el MCO inicial como en el MCO con educación era estadísticamente significativo y de una magnitud considerable (del 14,04% y 12,17% respectivamente), con la ocupación ahora no es estadísticamente significativa, y el impacto cae hasta el 3,73%. La interpretación de este hecho es que mayores niveles de ocupación son los responsables realmente de eliminar las diferencias

existentes entre nacionales españoles e inmigrantes, un hecho directamente plasmable en políticas redistributivas.

Es relevante hablar en este punto de la variable sexo, ya que en el modelo inicial tenía un impacto estadísticamente significativo del -30,19%, y con la ocupación cae hasta el -26,96%. Observamos por lo tanto que entre la educación y la ocupación, es ésta última la que parece el canal apropiado para atacar la brecha de género, pero el impacto limitado sobre la magnitud del coeficiente expone la gran complejidad que posee la variable sexo en términos de la desigualdad de oportunidades, donde la solución recae en una medida holística que vaya más allá de una simple reforma laboral o educativa. En este punto, es menester matizar que estos canales de educación y ocupación se están midiendo de manera imperfecta, ya que se categorizan únicamente en tipos de trabajo. Esto apunta a que, en igualdad de puestos de trabajo y de niveles educativos, sigue habiendo un efecto muy relevante de ser mujer sobre su renta. En este sentido puede haber múltiples explicaciones sobre este resultado, sin embargo, lo que queda claro es que al final estamos ante un caso de desigualdad de oportunidades, ya que encontramos que de manera sistemática y estadísticamente significativa el hecho de ser mujer impacta de manera relevante sobre sus ingresos, aun cuando abordamos en el estudio variables como niveles educativos o categorías ocupacionales. Sin embargo, determinar que existe desigualdad de oportunidades en España por esta circunstancia no es suficiente, lo que motiva a realizar posteriormente un estudio más exhaustivo de carácter intrasectorial para determinar las causas últimas de esta desigualdad.

Igualmente cabe destacar el impacto considerable sobre la variable edad hasta 30, con un impacto negativo considerable del 26,5% en el modelo inicial que cae al 22,5% al incorporar la ocupación. Sin embargo, en el modelo en el que se incluye tanto la educación como la ocupación, cae hasta el 21,5%, lo cual nos da indicios de que una forma de mejorar la situación económica de los jóvenes viene de una acción conjunta a nivel educativo y laboral, y no sólo enfocándonos en una de estas dos categorías, sin embargo, resulta un caso muy similar al del sexo.

Observamos que el hecho de que la madre no haya terminado estudios primarios en el modelo MCO inicial impactaba la renta negativamente en un 5,8%, mientras que con la ocupación esto cae hasta un -0,99%, lo cual es considerable a nivel de políticas.

En términos de la variable Escuela, vemos que el impacto es similar al MCO con educación, disminuyendo el impacto negativo sobre la renta en un 6% aproximadamente. Como es de esperar, en el modelo MCO con ambos canales la disminución es aún menor, pasando de un -20,4% inicialmente a un -12,1%, por lo que la mejor forma de amortiguar los impactos negativos de la renta provenientes de haber estudiado en un instituto público, es mediante medidas laborales y educativas en conjunto.

## 7. CONCLUSIONES

En este trabajo se realizó un estudio de la desigualdad de oportunidades en España. Para ello se trabajó sobre el módulo de desigualdad y movilidad social del CIS (2017), y a través de métodos LASSO de selección de variables se crearon un conjunto de modelos con las circunstancias más relevantes que ayudan a explicar parte de las diferencias de rentas en nuestro país. La estimación de estos modelos nos permitió analizar la desigualdad de

oportunidades en España, minimizando los sesgos habituales que provienen del uso exclusivo de modelos de regresión lineales a través de la selección con criterios estadísticos de las circunstancias más relevantes para explicar la DO, por lo que el objetivo principal de este trabajo fue completado. También se analizó el papel de la ocupación y de la educación de los individuos como potenciales aspectos canalizadores del efecto de las circunstancias sobre la renta de los individuos, lo cual constituía un objetivo complementario, al igual que el cálculo de la ratio desigualdad de oportunidades sobre desigualdad total, en donde obtuvimos unos resultados que rondaban el 49%. Es decir, aproximadamente el 49% de la desigualdad total se explica por el conjunto de circunstancias seleccionadas en este trabajo.

En total se estimaron cuatro modelos: Un MCO referencial con todas las circunstancias, un MCO con circunstancias, un MCO con variables cruzadas y un MCO con circunstancias y canales, en éstas tres últimas las variables fueron seleccionadas mediante la técnica LASSO con validación cruzada, para realizar el trade-off óptimo entre el sesgo al alza debido al sobreajuste, y el sesgo a la baja por eliminación excesiva de variables explicativas.

A lo largo del trabajo pudimos observar que independientemente del modelo elegido, hay predictores sumamente importantes a la hora de explicar la desigualdad de oportunidades, las más notables son el sexo, con un impacto negativo en la renta que rondaba consistentemente el 30% para las mujeres, lo cual constituye una evidencia empírica de la brecha de género que existe en nuestro país. Otras variables sumamente relevantes son el tamaño de la ciudad en la que creció el individuo, con un aumento de un 10% aproximadamente para las personas que vivan en grandes ciudades de más de cuatrocientos mil habitantes con respecto a habitantes de ciudades más pequeñas, tener la nacionalidad española está asociada con aumentos en la renta de un 14% con respecto a la población inmigrante residente. Por otra parte, tenemos que haber estudiado en una escuela pública está asociado con disminuciones en la renta de los individuos de un 20% con respecto a su contrapartida que estudió en privadas o concertadas. La educación del padre es relevante, especialmente si no terminó los estudios primarios, con un impacto negativo del 21%. Por otra parte, el estado de salud influye de manera considerable en la renta, ya que haber crecido con buena salud reporta un 17% más de renta en los individuos con respecto a su contrapartida.

En el modelo de variables cruzadas pudimos observar la gran complejidad existente en el problema de desigualdad de oportunidades, ya que las circunstancias en muchas ocasiones no influyen en la renta por sí mismas únicamente, sino que sus efectos se ven amplificados o mitigados de acuerdo al cumplimiento de otras circunstancias. Algunos resultados interesantes en este sentido corresponden a la variable sexo, donde los impactos negativos sobre la renta se ven amplificados si la mujer estudió en una escuela pública o si creció en una ciudad menor a cuatrocientos mil habitantes. Por otra parte, si el individuo estudió en una escuela pública, se amplifica el efecto negativo si el individuo creció en una ciudad pequeña, si es mujer, si el padre no completó los estudios primarios y si el individuo posee tres o más hermanos. Igualmente tenemos variables que solo se ven influenciadas en conjunto con otros predictores, es el caso de la nacionalidad española, en donde depende positivamente de haber crecido con buena salud, tener tres o más hermanos, y que sus padres hayan dedicado recursos y atención a su educación, mientras que depende negativamente de tener un padre sin estudios primarios o tener menos de treinta años.

De esta forma, podemos comprobar la magnitud y complejidad del problema de la desigualdad de oportunidades, ya que en muchos casos las circunstancias se relacionan entre sí, modificando la magnitud y el sentido del impacto sobre la renta de los individuos, por lo que tratar este problema tiene que hacerse de manera holística y prestando atención a más de un indicador, a más de una circunstancia y con un enfoque multidisciplinar que permita mitigar los impactos negativos sobre la renta por parte de las circunstancias, ya que en palabras del Banco Mundial (2006) "las circunstancias en el momento del nacimiento no deben tener ningún peso en las oportunidades que una persona tenga en la vida". (p.13).

Por otra parte, se realizó una estimación paramétrica de la desigualdad de oportunidades en España a partir de la variable dependiente explicada de los tres modelos iniciales, tras lo cual se estimó la ratio DO/DT, uno de los objetivos complementarios de este trabajo. Se llegó al resultado de que alrededor de un 49% de la desigualdad total de ingresos proviene del conjunto de circunstancias que constituye la desigualdad de oportunidades. Este es un resultado sumamente preocupante, ya que el hecho de que casi la mitad de la desigualdad total se deba a factores que escapan del control de los individuos va en contra de los principios de igualdad material y de los valores democráticos que deberían caracterizar a nuestra sociedad.

En este sentido, en este trabajo se desprenden algunas conclusiones interesantes que pueden dar pie a la aplicación de algunas políticas redistributivas o de otra naturaleza que mitiguen los impactos negativos de la desigualdad de oportunidades en nuestro país. Partiendo de lo más general a lo más específico, encontramos que existen circunstancias que escapan del control de los individuos que afectan de manera directa a su renta, lo cual en términos de justicia e igualdad no debería ocurrir. La solución a estos problemas debe darse a nivel institucional, como menciona Stiglitz (2015) "Cada aspecto de nuestro marco económico, legal y social influye en la desigualdad: Desde el sistema educativo y su financiación, al sistema de salud, a las leyes fiscales". (p.382).

Por otra parte, pudimos comprobar a partir del modelo MCO con variables cruzadas que las circunstancias en muchos casos no actúan únicamente por sí solas, sino que se relacionan con otras variables que mitigan o amplifican los impactos negativos sobre la renta, en este sentido, en muchos casos debe atacarse el problema de la desigualdad de oportunidades de manera holística, con un enfoque multidisciplinar que permita disipar los efectos de todas las circunstancias relevantes, especialmente tomando en cuenta que un porcentaje de la renta perdida por un individuo puede deberse a más de una combinación de factores que escapan de su control.

Igualmente, con respecto al último objetivo complementario, encontramos que algunos canales mediante los cuales se pueden mitigar estos efectos son la educación y la ocupación de los individuos, lo cual permite tener una herramienta concreta, las reformas educativas y laborales, que con un enfoque correcto pueden tener un impacto considerable sobre muchas de las circunstancias analizadas en este trabajo. Por ejemplo, la educación fue sumamente relevante para canalizar los efectos de circunstancias como el tamaño de la ciudad en la que se creció, tener padre sin estudios primarios completos, o haber estudiado en una escuela o instituto públicos. Por otra parte, la ocupación fue relevante para canalizar circunstancias como la nacionalidad, la edad hasta los 30 años, o la falta de estudios primarios de la madre. Cabe destacar igualmente que el impacto de muchas de estas variables, en concreto el tamaño de la ciudad, la nacionalidad y la escuela pública, se pueden corregir de mejor manera con un enfoque

dual que tome en cuenta el aspecto laboral junto al educativo para obtener resultados más óptimos.

Por último, cabe destacar algunas posibles extensiones a este trabajo. Tomando en cuenta como ejemplo la variable sexo, en el apartado 6 observamos que aun teniendo en cuenta variables educativas y laborales junto a las circunstancias en la regresión, vemos que el efecto negativo sobre la renta derivado de ser mujer disminuye de manera muy tenue e insuficiente. Esto apunta a que aún en igualdades educativas y de puestos de trabajo, el hecho de ser mujer posee un impacto negativo sobre la renta de una magnitud prácticamente igual a la obtenida en modelos donde incorporamos exclusivamente las circunstancias, por lo que un estudio de carácter intrasectorial para diseccionar realmente las causas últimas de esta desigualdad de oportunidades resulta relevante y necesario para apuntar realmente a una solución aplicable a este problema.

## 8. REFERENCIAS BIBLIOGRÁFICAS


Ayala, L. (2006). La Desigualdad en España: Fuentes, Tendencias y Comparaciones Internacionales. *Estudios sobre la Economía Española.* 2016/24.

Banco Mundial. (2006). *Informe sobre el desarrollo mundial 2006: Equidad y desarrollo*. World development report.

Brunori, P., Peragine, V. y Serlenga, L. (2019). Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare.* 52, 635–661.

Brunori, P., Hufe, P. y Mahler, D. (2021) The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests. *IZA Discussion Paper*. 14689.

Buhlmann, P. y van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cabrera, L., Marrero, G., Rodríguez, J. y Salas, P. (2019). Inequality of Opportunity in Spain: new Insights from New Data. *Hacienda Pública Española / Review of Public Economics*. 237, 153-185.

CIS (2017). Desigualdad Social y Movilidad Social en España. *CIS*, 3178.

Ferreira, F. y Gignoux, J. (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth*. 57, 622-657.

Fonti, V. y Belitser, E. (2017). Paper in Business Analytics Feature Selection using LASSO.

Hastie, T., Tibshirani, R. y Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Marrero, G. y Rodríguez, J. (2012). Inequality of opportunity in Europe. *Review of Income and Wealth*. 58, 597-621.

Marrero, G. y Rodríguez, J. (2013). Inequality of opportunity and growth. *Journal of Development Economics*. 104, 107-122.

Palomino, J., Marrero, G. y Rodríguez, J. (2019). Channels of inequality of opportunity: the role of education and occupation in Europe. *Social Indicators Research*. 143, 1045-1074.

Ranstam, J. y Cook, J.A. (2018). *LASSO regression*. *British Journal of Surgery*. 10, 1348.

Roemer, J. (1993) A pragmatic approach to responsibility for the egalitarian planner. *Philosophy Public Affairs*. 20, 146-166.

Stiglitz, J. (2015). The price of inequality: How today's divided society endangers our future. *Sustainable Humanity, Sustainable Nature: Our Responsibility*. 2-6: 379-399.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. 58, 267-288.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. y Tibshirani, R. (2011). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society*. 74, 245-266.