UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO INGENIERÍA CIVIL INDUSTRIAL

IMPACTO DE LA COMPOSICIÓN DE GÉNERO EN EL DESEMPEÑO DE EMPRESAS CHILENAS: USO DE MÉTODOS CORRECTIVOS DE ATRICIÓN

POR

Cristian Eduardo Bustos Bello

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniero Civil Industrial

Profesor Guía

Ph.D. Marcela Parada Contzen

Julio 2022

Concepción, Chile

Dedicatoria

Dedico este trabajo en primer lugar a mis padres, quienes fueron los que me acompañaron en todo momento durante este proceso, nunca dudaron y apoyaron toda decisión que tomé.

A los familiares que siempre se preocuparon por mí, ayudándome de muchas formas durante estos años.

A mis amigos, que fueron un apoyo durante mi estadía en la universidad, dentro y fuera de ella.

Y por último, a los que no están. A mis abuelos que forjaron en mí parte de la persona de la que soy el día de hoy. Y a Kaiser, que estuviste a mi lado desde que estaba en la básica. Ahora, desde donde estás, me diste una ayuda para seguir.

Cristian Eduardo Bustos Bello.

Resumen

La atrición en los datos se define como la pérdida de registros a lo largo del tiempo, presentando diferencias sistemáticas en la forma y en la cantidad según la naturaleza de los datos, siendo éste, uno de los principales obstáculos a la hora de aplicar métodos estadísticos en estudios longitudinales. Carrasco (2022), presenta una primera aproximación sobre el desempeño de empresas chilenas según la composición de género, basándose en los resultados de la Encuesta Longitudinal de Empresas realizada entre 2007 y 2017. Sin embargo, esta base de datos presenta altos grados de atrición. En la presente memoria de título se plantea una extensión de este trabajo, usando métodos correctivos para la presencia de atrición en los datos.

En la literatura, se presentan diversos tipos de modelos para lograr solucionar esta problemática, los cuales varían desde el uso del método de máxima verosimilitud, la inyección de registros artificiales, el uso de pesos en los coeficientes de estimación, la inclusión de variables dicotómicas, la aplicación de modelos de machine learning, entre otros.

De esta forma, para el presente trabajo se plantearon los métodos de imputación simple e imputación múltiple para lograr solucionar el problema de la atrición en los datos. El primero, se realizó en base a los criterios de moda, media, regresión lineal y por arrastre de la última observación. Mientras que el segundo, se implementó con el algoritmo de ecuaciones concatenadas (multiple imputation by chained ecuations, en inglés) junto a estimaciones con Random Forest.

Los resultados presentan que ambas imputaciones no afectaron la vista general de la base de datos, debido a que existe una estabilidad tanto en media, como en moda y composición. Además, se muestra que la presencia de mujeres en la gerencia de las empresas tiene un impacto estadísticamente significativo y negativo sobre las ventas y la productividad. Mientras que, para el número de trabajadores presentes, no hay resultados estadísticamente significativos.

Tabla de contenidos

Lista de Tablas

Lista de Figuras

1 Introducción

1.1 Antecedentes

La presente memoria de título se basa en el trabajo realizado por Carrasco (2022), dónde se plantea una primera aproximación del impacto de la composición de género en el rendimiento de empresas chilenas según los datos de la Encuesta Longitudinal de Empresas entre los años 2007 y 2017.[1] La motivación de este estudio se basa en extender la literatura que analiza diferencias en productividad en las empresas, según su composición de sexo en la gerencia general. De esta forma, se toma el caso de Chile que ha sido escasamente estudiado en la literatura de género y desempeño empresarial.[2] La presente memoria de título sigue manteniendo esta motivación general y particularmente se centra en mejorar el análisis estadístico de manera de generalizar los resultados a medida que se levantan ciertos supuestos de modelamiento.

En Carrasco (2022) se realiza un análisis estadístico basado en el uso de 4 regresiones lineales múltiples, las cuales buscan explicar el comportamiento de 4 variables dependientes de interés: productividad, utilidad de la empresa, número de empleados y cantidad de ventas. Dicha memoria concluye que el análisis obtenido tiene margen de mejora, principalmente por la atrición no modelada presente en los datos (registros no existentes) junto a la posibilidad de incluir otras variables de interés (por ejemplo, políticas impartidas) que podrían explicar también el rendimiento de la empresa. Además, tal como detalla Carrasco (2022), si bien existe una tendencia al alza dentro de la participación de mujeres en la gerencia general de las empresas, sigue habiendo una diferencia substancial entre la participación de hombres y mujeres (78% hombres 22% mujeres, según la ELE 5).

De esta forma, el problema de la atrición presente en los datos será tratado en la siguiente investigación, con el fin de conseguir un análisis con menores sesgos de estimación. Se utilizarán las

[1] Siguiendo la misma caracterización de Carrasco (2022), en esta memoria de título se utiliza el concepto género para indicar sexo (hombre o mujeres) pues esa es la única desagregación en información de género que permite la base de datos.

[2] Carrasco (2022) hace una revisión detallada de los resultados encontrados en países como Estados Unidos (Adams & Ferreira, 2009), Alemania (Gottschalk & Niefert, 2013), Dinamarca (Parrotta & Smith, 2013), España (Campbell & Mínguez-Vera, 2008) y en un conjunto de países de Europa (Faccio, Marchica, & Mura, 2016).

5 versiones de la Encuesta Longitudinal de Empresas (ELE 1 - 2007, ELE 2 - 2009, ELE 3 - 2013, ELE 4 - 2015 y ELE 5 - 2017), incluyendo también las empresas que no hayan participado en alguna por diversos motivos (voluntariamente no quisieron responder, quiebra, fallo a la hora de contactar, etc.) con su correspondiente ELE. Se considerará también el estudio de las variables dependientes ocupadas por Carrasco (2022) y la posibilidad de realizar cambios en las variables de control.

1.2 Justificación del tema

Si bien los resultados propuestos por Carrasco (2022) son correctos estadísticamente, están basados en un escenario ideal, donde sólo se utilizaron los datos de empresas que respondieron las 5 versiones de la ELE, generando atrición en los datos. De esta forma, el realizar una extensión que incluya también las empresas que presenten irregularidades dentro de los datos, junto a un análisis de los datos que nos permita disminuir el sesgo, entregará una visión más realista y estandarizada del análisis planteado inicialmente.

1.3 Objetivos

1.3.1 Objetivo general

El objetivo general del tema es analizar el impacto de la composición de género en la gerencia general en el desempeño de empresas chilenas utilizando métodos que corrijan atrición y sesgos por selección en los datos.

1.3.2 Objetivos específicos

- Revisar la literatura que presente la formalización de la atrición y métodos empíricos para abordar la atrición y/o sesgo de selección.
- Analizar la magnitud y características de la atrición en la base de datos conformada por las cinco rondas de la ELE.
- Formular el(los) modelo(s) empíricos apropiados y estimarlos con metodologías que permitan controlar por la atrición, utilizando las cinco rondas de la ELE.

- Analizar los resultados obtenidos y compararlos con los resultados obtenidos por Carrasco (2022).

1.4 Estructura del informe

La memoria de título consta de 6 capítulos. En el presente, se describen los antecedentes generales, la justificación del tema, los objetivos generales y específicos, y, por último, la estructura de la memoria.

El siguiente capítulo presenta el marco teórico, en el cual se hace una revisión bibliográfica de problemáticas y metodologías aplicadas en presencia de atrición y/o datos ausentes.

El capítulo 3 presenta la descripción de los datos a utilizar por medio de la definición de las variables, un análisis general y un análisis sobre datos ausentes.

En el capítulo 4, son detalladas las metodologías a ocupar para subsanar el problema de la pérdida de datos y la atrición, junto con presentar los modelos de estimación de parámetros.

El capítulo 5 presenta los resultados principales de los métodos correctivos anteriormente descritos junto a la comparación de los resultados de los modelos de estimación con los obtenidos por Carrasco (2022).

Finalmente, el capítulo 6 expone las conclusiones principales de la presente memoria de título y su margen de mejora.

2 Marco Teórico

La atrición en los datos se define como la pérdida de registros (también llamados individuos u observaciones) a lo largo del tiempo, presentando diferencias sistemáticas en la forma y en la cantidad según la naturaleza de los datos (Nunan, 2018).

En esta memoria de título, la atrición se observa en las empresas que, luego de contestar por primera vez alguna versión de la ELE, no continuaron su participación en las siguientes versiones de manera íntegra. Puede haber diferentes razones que haga que las empresas no continúen su participación: por ejemplo, 1) la empresa ya no existe porque cerró, ya sea por motivos financieros u otros, o porque se fusionó con otra, 2) la empresa no quiso contestar, 3) el administrador de la encuesta perdió el contacto con la empresa (por cambio de lugar, número de teléfono, etc), 4) la empresa no pudo ser contactada en el momento de tener que responder la encuesta.

Si bien la pérdida de datos es una gran amenaza para la integridad de este tipo de análisis, tiene poca atención econométrica formal (Fitzgerald, Gottschalk, & Moffitt, 1999), siendo muchas veces ignorada y/o usando sólo el caso completo (base de datos sin pérdida de valores/observaciones), que dentro de un estudio longitudinal serían los individuos que tienen tasa completa de respuesta desde su primera aparición.

Sin embargo, según lo estudiado por Fitzgerald, Gottschalk, & Moffitt (1998), bajo el contexto de una regresión lineal, la presencia de atrición en los datos afecta mayoritariamente en los interceptos (constantes), y, en algunos casos, la pendiente. De esta forma, es de suma importancia aplicar algún método estadístico que corrija este problema.

Los primeros acercamientos a la pérdida de datos y su impacto son atribuidos a Szabo (1969), Hoon Jr. (1969) y Finley (1972). Éstos, si bien no presentaron modelos para corregir la atrición, la describieron como un problema a considerar a futuro, principalmente por el impacto que tenían en el peso del total de la base de datos con la que trabajaron. Posteriormente, Rubin (1976) profundizó sobre este tema, planteando que los problemas de datos faltantes se pueden clasificar en 3 categorías según la probabilidad que tienen las propias observaciones de perderse. Si la probabilidad de que falten es la misma para todas las observaciones, se dice que los datos faltan completamente al azar (*missing completely at random*, MCAR). En otras palabras, es el caso cuando no existe relación entre los datos y la probabilidad de pérdida. Si la probabilidad de que falten los datos sí se puede explicar

por variables presentes en los datos, entonces, éstos faltan al azar (*missing at random*, MAR), siendo el supuesto que se presenta con mayor frecuencia en este tipo de problemas. En el caso de que ni MCAR ni MAR se cumplan, entonces se habla pérdida de datos que no es al azar (*missing not at random*, MNAR), el cual plantea que la probabilidad de que falte varía por razones que desconocemos.

Con respecto a los métodos para la corrección de la atrición, Lehnen & Koch (1974) usaron del método de máxima verosimilitud para estimar los parámetros de un modelo de regresión lineal, incluyendo también la estimación de un parámetro que indica la presencia o ausencia de sesgo por atrición. Comprobaron, a su vez, que el uso de mínimos cuadrados ordinarios para estimar los parámetros conduciría a un sesgo en los estimadores. De igual manera, Hausman & Wise (1979), propusieron un modelo de regresión lineal para variables categóricas. Sin embargo, éstos presentaron un enfoque basado en la comprobación de hipótesis sobre la similitud de los diversos subconjuntos de datos con atrición y el cambio neto (marginal) a través de los períodos de tiempo. Por su parte, Little & Rubin (1991), proponen un modelo basado en probabilidades y el uso de funciones de verosimilitud, en el cual abarcan la mayoría de casos "especiales", incluyendo tópicos como el análisis de componentes de la varianza y el modelado de variables latentes. Abowd, Crepon, & Kramarz (1997) propusieron el uso del método de los momentos generalizado para lidiar con este problema, mientras que Fitzgerald, Gottschalk, & Moffitt (1999), plantearon un modelo de sesgo de selección en el que solo se observa una parte de la población, para explorar la pérdida de los datos desde una perspectiva diferente, la de la selección en observables, concluyendo que la integridad del modelo va directamente de la mano con la selección de los individuos, y que el método de los mínimos cuadrados ponderados es efectivo, siempre y cuando se usen en paralelo ecuaciones relacionadas con la atrición.

Hasta este punto, los métodos expuestos son presentados bajo la lógica del perfeccionamiento de la estimación de parámetros en una regresión lineal para tratar con el problema de la pérdida de datos. Sin embargo, Hirano, Imbens, Ridder, & Rubin (2001), presentan dos modelos, en los cuales usan muestras de "refresco" (artificiales) para llenar el vacío que dejan los registros que presentan atrición. El primer modelo, se basa en el supuesto de falta al azar (MAR), el cual, tal como se detalló al inicio, se caracteriza por poseer patrones visibles y que los datos sí lo pueden explicar, en este caso, por las variables rezagadas y no por las contemporáneas que tienen valores perdidos. Mientras que, el segundo modelo, presenta un caso opuesto, debido a que permite que la probabilidad de atrición dependa de las variables contemporáneas, pero no de las antiguas. Ambos modelos poseen inferencias

muy diferentes, dependiendo principalmente de la naturaleza de los datos. Por otra parte, Baulch & Quisumbing (2011) plantearon el uso del ponderador de probabilidad inversa para el aumento de datos en los tramos que presente atrición. Este método, si es correctamente aplicado, tiene tendencia a aumentar la eficacia del modelo. Sin embargo, debe ser comprobable su integridad para los datos en específico, dado a que, un mal uso, puede empeorar el sesgo por atrición (Vandecasteele & Debels, 2007).

En los últimos años, destaca el trabajo de Zhu (2014), el cual presenta cuatro modelos de simulación para tratar con el problema de la atrición: caso completo (CC), sustitución media (MS), última observación realizada (LOCF) e imputación múltiple (MI), y el trabajo de Hill, Biemer, & Buskirk (2020), el cual propone el uso de machine learning para predecir la atrición.

3 Descripción de datos

La base de datos utilizada para esta memoria consta en la unión de las primeras 5 versiones de la Encuesta Longitudinal de Empresas (ELE) de propiedad del Ministerio de Economía y el Instituto Nacional de Estadísticas (Chile). Esta base de datos tiene rondas publicadas para los años 2007, 2009, 2013, 2015 y 2017. La ELE tiene como objetivo describir a las empresas de todo Chile, teniendo en consideración el tamaño de esta, su propio sector industrial, los años de antigüedad que tenga, etc., generando así, análisis generales de las características de las empresas o a nivel de detalle.

En el presente capítulo se definen las variables a utilizar dentro de la memoria de título y expone en detalle la base de datos en dos enfoques: análisis general y análisis de datos ausentes.

3.1 Definición de variables utilizadas

La definición de las variables utilizadas se presenta en las Tablas 1, 2 y 3:

Tabla 1: Descripción de las variables utilizadas

Nombre de la variable	Definición
ID	ID único de empresa (se mantiene a través de las versiones de la ELE)
Año de realización	Año en el cual la correspondiente versión de la ELE fue realizada
Tamaño	Asignación de número según tamaño de la empresa (1: Grande, 2: Mediana, 3: P1, 4: P2, 5: Micro)
Tasa de interés por deuda	Tasa anual de interés del crédito (en caso de que la empresa tenga deuda)
Ventas	Ventas anuales reportadas en dólares de diciembre de 2017
Edad	Edad de la empresa
Holding	1: si la empresa es parte de un grupo económico, 0: en otros casos
SRI	1: si la empresa es una sociedad de responsabilidad ilimitada, 0: en otros casos
Extranjeros	1: si la empresa tiene dueños extranjeros, 0: en otros casos
Porcentaje de participación de extranjeros	Porcentaje de participación de los dueños extranjeros (en caso de que los dueños sean extranjeros)
Familiar	1: si la empresa es un negocio familiar, 0: en otros casos
Dueño	1: si el CEO de la empresa es el dueño de esta, 0: en otros casos

Tabla 2: Descripción de las variables utilizadas (continuación)

Nombre de la variable	Definición
Propiedad	1: si el CEO es dueño de algún porcentaje de la empresa, 0: en otros casos
Género	1: si el CEO de la empresa es hombre, 0: en otros casos
Género (mujer)	1: si el CEO de la empresa es mujer, 0: en otros casos
Exporta	1: si la empresa realiza exportaciones directas, 0: en otros casos
Porcentaje de exportaciones	Porcentaje de las ventas atribuidas a las exportaciones directas (en caso de que la empresa realice exportaciones directas)
Número de trabajadores	Número de empleados de la empresa
Salarios	Salarios anuales reportados en dólares de diciembre de 2017
Externo	1: si la empresa tiene trabajadores subcontratados, 0: en otros casos
Cantidad de externos	Cantidad de trabajadores subcontratados de la empresa (en caso de que la empresa tenga trabajadores subcontratados)
Deuda	1: si la empresa tiene deuda, 0: en otros casos
IyD	1: si la empresa hace I+D, 0: en otros casos
Inventario	Inventario anual reportado en dólares de diciembre de 2017
Impuestos	Impuestos anuales reportados en dólares de diciembre de 2017
Utilidades	Utilidades anuales reportadas en dólares de diciembre de 2017
Región	Número de la región en la cual se ubica la empresa
Ventas (en log)	Ventas anuales reportadas en dólares de diciembre de 2017, en logaritmo
Productividad laboral de los empleados (en log)	Ventas anuales reportadas en dólares de diciembre de 2017 sobre número de empleados, en logaritmo
Inventario (en log)	Inventario anual reportado en dólares de diciembre de 2017, en logaritmo
Impuestos (en log)	Impuestos anuales reportados en dólares de diciembre de 2017, en logaritmo
Salarios (en log)	Salarios anuales reportados en dólares de diciembre de 2017, en logaritmo
Utilidades (en log)	Utilidades anuales reportadas en dólares de diciembre de 2017, en logaritmo
Número de trabajadores (en log)	Número de empleados de la empresa, en logaritmo

Tabla 3: Descripción de las variables utilizadas (continuación)

Nombre de la variable	Definición
ELE 2	1: si la ELE corresponde a la versión 2, 0: en otros casos
ELE 4	1: si la ELE corresponde a la versión 4, 0: en otros casos
Empresa grande	1: si la empresa es de tamaño grande, 0: en otros casos
Margen de utilidad	Relación entre las utilidades y las ventas de manera porcentual
Tasa de crecimiento de la utilidad	Tasa a la cual creció (o decreció) la utilidad de la empresa entre dos instantes de tiempo consecutivos

3.2 Análisis general

Dado a que las empresas viven en un entorno sumamente cambiante, tanto en factores internos (quiebra, crecimiento, cambio en el personal, etc.) como factores externos (influencia del entorno en el Q de ingresos, demanda de productos, etc.), se han publicado, hasta el año 2022, 5 versiones de la encuesta: ELE-1 (2007), ELE-2 (2009), ELE-3 (2013), ELE-4 (2015) y ELE-5 (2017), con el fin de que la información sobre las empresas participantes sea la más actualizada posible. Dentro de este rango, se han realizado 39.106 encuestas, las cuales, han tenido un comportamiento irregular en torno a la participación, tal como se aprecia en la Tabla 4.

Tabla 4: Número de encuestas realizadas para cada tanda de la ELE.

Versión de la encuesta	Número de empresas participantes
ELE-1	10.213
ELE-2	7.062
ELE-3	7.267
ELE-4	8.084
ELE-5	6.480

En primera instancia la base de datos fue preprocesada para encontrar homología entre las versiones, debido a que la estructura de la encuesta en ninguna de estas fue idéntica, por lo que las preguntas podían cambiar de orden, enunciado o, incluso, ser eliminadas.

La composición del tipo de empresas que participaron en cada encuesta, según su tamaño, se puede ver en la Figura 1. Si bien, existe una clara baja en la participación, y en su peso con respecto al total, de parte de las microempresas, está la opción de que muchas de estas hayan mutado hacia otro tamaño de empresa, que hayan quebrado (por ser el escalón más pequeño y, por consiguiente, más inestable) o que no hayan contestado siguientes versiones de la ELE (presencia de atrición). Caso similar para las empresas de tipo P2, las cuales disminuyeron su participación en más de 60% entre la primera y la quinta versión. Por otra parte, las empresas grandes, son las que más han aumentado su peso en las versiones de la ELE, pasando de un 18% (con respecto al total) en la ELE-1, a un 54% en la ELE-4 y, finalmente, a un 39% en la ELE-5.

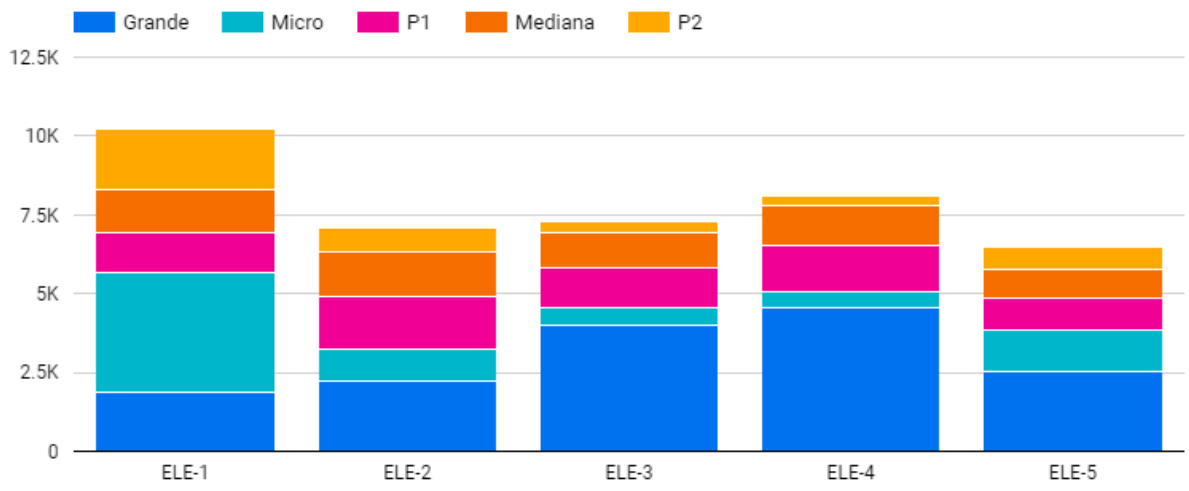


Figura 1: Cantidad de empresas participantes según la versión de la ELE y su tamaño.

Fuente: elaboración propia

Por otra parte, una de nuestras variables de interés a estudiar es el género, el cual toma valor igual a 1 en el caso de que el gerente general (CEO) de la empresa encuestada es hombre, y valor 0 si es mujer. Esta, fue la variable principal estudiada por Carrasco (2022), por lo que es de suma importancia tener

claridad con respecto a su comportamiento a través de las ELE. En la Figura 2 se presentan los pesos según el género del CEO para cada versión de la encuesta. En primera instancia, se aprecia un comportamiento irregular (aumenta y disminuye) en cuanto a la proporción de gerentes generales hombres, junto a la presencia de datos perdidos en la ELE-1.

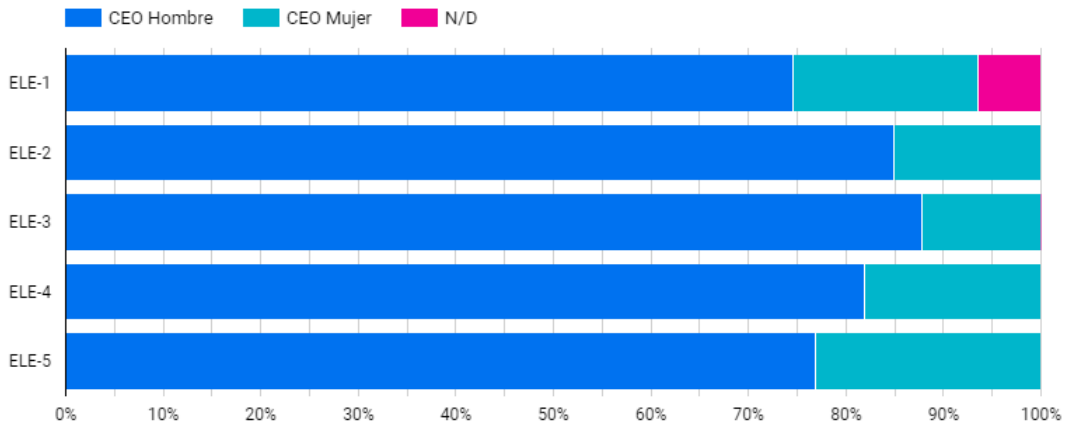


Figura 2: Pesos según el género del CEO de las empresas según la versión de la ELE.

Fuente: elaboración propia

Sin embargo, cuando se realiza este mismo análisis (excluyendo las observaciones con este valor perdido o *null*) según el tamaño de la empresa, se pueden observar fuertes tendencias. Esto, se ve demostrado en la Figura 3, la cual presenta los pesos para las empresas grandes y medianas, quienes muestran un alza en la proporción de las mujeres que son CEO de estas empresas, pasando de un 6% de la participación total a un 18% para estos dos tipos de empresas. Este crecimiento, recalca la importancia de estudiar el impacto que tiene esta variable sobre el rendimiento de la empresa en cuestión, especialmente para el caso de estas empresas, debido a que son las que más influencia tienen en el mercado.

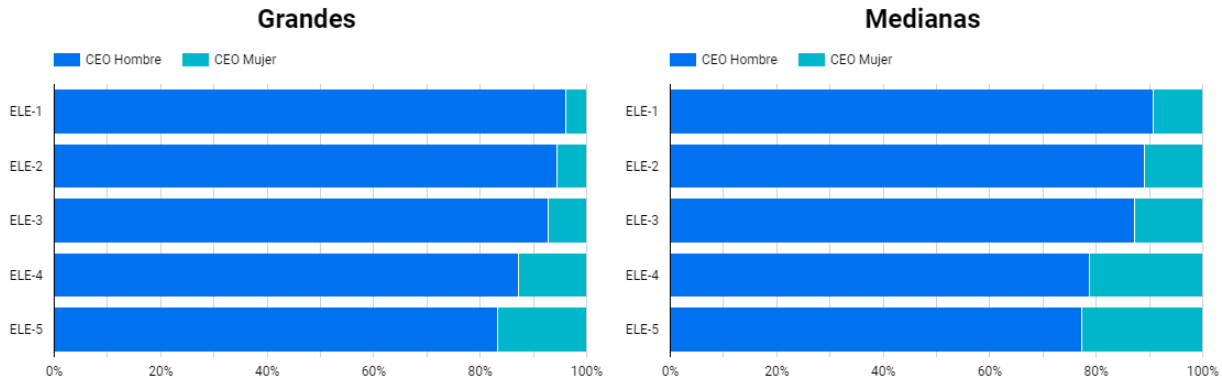


Figura 3: Pesos según el género del CEO de las empresas según la versión de la ELE para empresas grandes y medianas.

Fuente: elaboración propia

Bajo este concepto, en la Figura 4 y en la Figura 5 podemos observar el impacto del género sobre otras dos variables de interés para medir el rendimiento de una empresa: utilidades y deuda, utilizándose para la primera variable su versión logarítmica. En ambos casos, el CEO hombre tiene más influencia (tanto para la generación de utilidades, como para la creación de deuda).

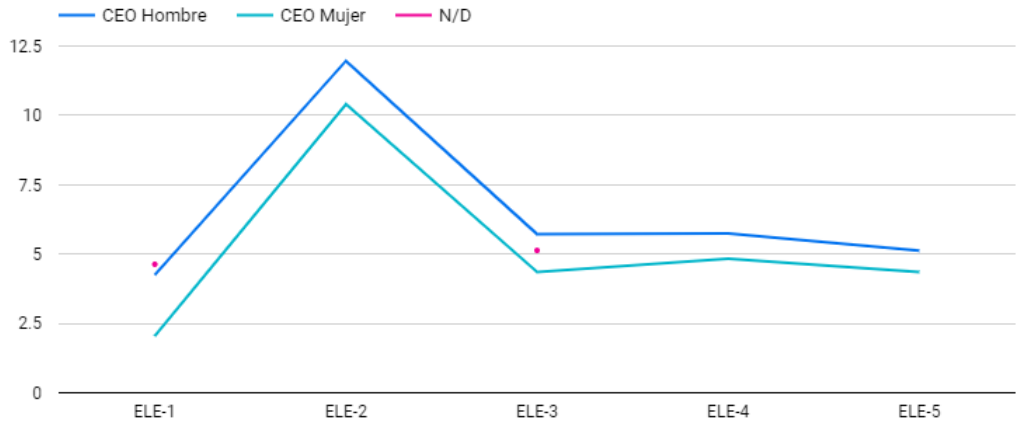


Figura 4: Utilidades promedio (en log) para cada versión de la encuesta según el género del CEO.

Fuente: elaboración propia

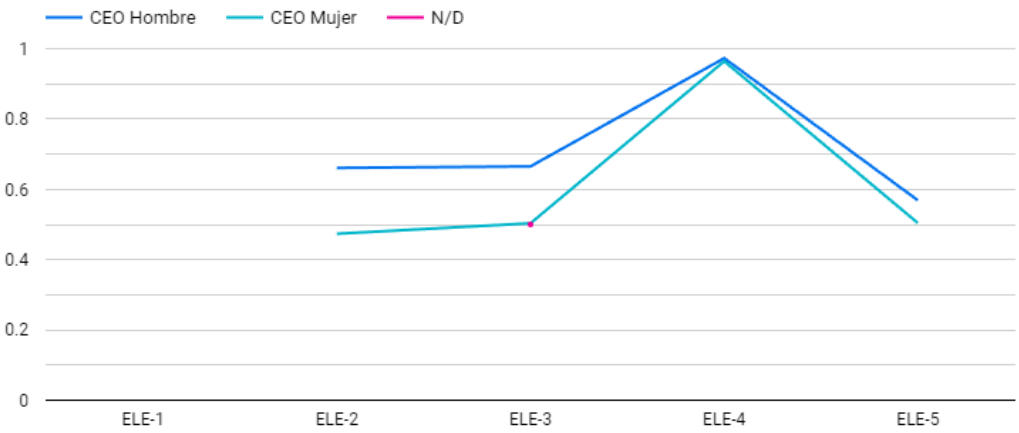


Figura 5: Deuda promedio para cada versión de la encuesta según el género del CEO.

Fuente: elaboración propia

3.3 Análisis de datos ausentes

La base de datos ya detallada será tema de estudio para nuestro problema central de sesgo por atrición. Sin embargo, existe otra fuente de ausencia de datos que no es propia de la tasa de deserción de las empresas y que está presente en nuestra base: la pérdida de datos. Esta, corresponde a los registros por variables que no están (también valores nulos o *null*) por algún motivo que no es detallado (como la no respuesta de la pregunta, la pérdida del dato a la hora de confeccionar la base original, la posibilidad de que ninguna alternativa sea la indicada para responder, información inconclusa, etc.), por lo que también se le conoce como ausencia implícita de datos. De esta manera, a continuación se presenta el análisis para la ausencia implícita (pérdida de datos) y para la ausencia explícita (atrición), con el fin de diferenciar el impacto de cada uno de éstos dentro de nuestra base.

3.3.1 Datos ausentes implícitos: pérdida de datos

Para cuantificar el impacto de la ausencia implícita, se presentan en las Tablas 5, 6 y 7 las variables que presentan algún grado de pérdida, junto a su frecuencia, porcentaje de pérdida con respecto al total y las correlaciones principales con respecto al resto de variables. Dado a que el establecimiento de correlaciones tiene como fin la posterior estimación de las variables para los datos perdidos, sólo se considerarán las variables que no presenten pérdida, haciendo la excepción para las utilidades, en

el cual se incluyen las ventas (en log) por su alta correlación (0,91) y bajo porcentaje de pérdida con respecto al total (1,31%).

Tabla 5: Variables con pérdida implícita en la base de datos

Variable	**Perdidos**	**Perdidos con respecto al total**	**Correlaciones principales**
Tasa de interés por deuda	28.842	73,75%	Año de realización (0,16) IyD (-0,10)
Ventas	514	1,31%	**con Ventas (en log)* ELE 2 (0,67) Tamaño (-0,63) Empresa grande (0,50) Número de trabajadores (en log) (0,42) SRI (-0,39) Dueño (-0,38) Holding (0,32)
Porcentaje de participación de extranjeros	1.115	2,85%	Exporta (0,22) Empresa grande (0,22) Holding (0,22)
Familiar	17.275	44,17%	Propiedad (0,24) Extranjeros (-0,19) Holding (-0,12)
Género	661	1,69%	Tamaño (-0,20) SRI (-0,14) Dueño (-0,13)
Exporta	2	0,01%	Empresa grande (0,29) Tamaño (-0,27) Extranjeros (0,26)

Tabla 6: Variables con pérdida implícita en la base de datos (continuación)

Variable	Perdidos	Perdidos con respecto al total	Correlaciones principales
Porcentaje de exportaciones	4	0,01%	Empresa grande (0,19) Tamaño (-0,18) Extranjeros (0,17)
Salarios	664	1,70%	**con Salarios (en log)* Número de trabajadores (en log) (0,88) Tamaño (-0,65) SRI (-0,50) Empresa grande (0,49) Dueño (-0,48)
Externo	3.200	8,18%	Empresa grande (0,32) Tamaño (-0,30) Número de trabajadores (en log) (0,27)
Cantidad de externos	3.200	8,18%	Número de trabajadores (0,16) Tamaño (-0,05)
Deuda	10.213	26,12%	Tamaño (-0,41) ELE 4 (0,36) SRI (0,32)
Inventario	3.037	7,77%	**con Inventario (en log)* Tamaño (-0,48) Empresa grande (0,42) Número de trabajadores (en log) (0,37) Dueño (-0,31) ELE 2 (0,25)

Tabla 7: Variables con pérdida implícita en la base de datos (continuación)

Variable	Perdidos	Perdidos con respecto al total	Correlaciones principales
Impuestos	2.521	6,45%	**con Impuestos (en log)* ELE 2 (0,69) Tamaño (-0,49) Empresa grande (0,35) Holding (0,30)
Utilidades	515	1,32%	**con Utilidades (en log)* $\widehat{\text{Ventas}}$ (en log) (0,91) ELE 2 (0,70) Tamaño (-0,51) Empresa grande (0,40) Holding (0,32) SRI (-0,31)
Región	2.308	5,9%	Tamaño (-0,20) Empresa grande (0,19)

Si bien existen distintos tipos de niveles de pérdida, para nuestras variables de interés como el género, ventas, utilidades y salarios, la caída no pasa el 2%, lo que es beneficioso para nuestro análisis posterior. Por otra parte, variables como la tasa de interés por deuda, si es una empresa familar o si la empresa presenta deuda, poseen pérdidas implícitas muy grandes, por lo que son fuertes candidatos a no ser considerados como variables explicativas dentro de los modelos a estudiar. En forma general, la mayoría de las variables con pérdida tienen correlaciones muy bajas con el resto de variables, lo que se puede deber a que estas variables se explican por otros con pérdida (como impuestos por ejemplo) o que tienen independencia del resto por su naturaleza (posible presencia de autocorrelación).

3.3.2 Datos ausentes explícitos: atrición

Para observar el impacto de la atrición en nuestra base de datos, se presenta en la Figura 6 la evolución en la participación en la ELE de las empresas a través del tiempo. Se puede apreciar, que de las 10.213 empresas que participaron en la primera versión, sólo 838 respondieron las 5 versiones de la ELE, siendo el resto, empresas que no respondieron todas las versiones o empresas que respondieron la encuesta por primera vez posterior a la ELE-1 (también llamadas muestras de refresco).

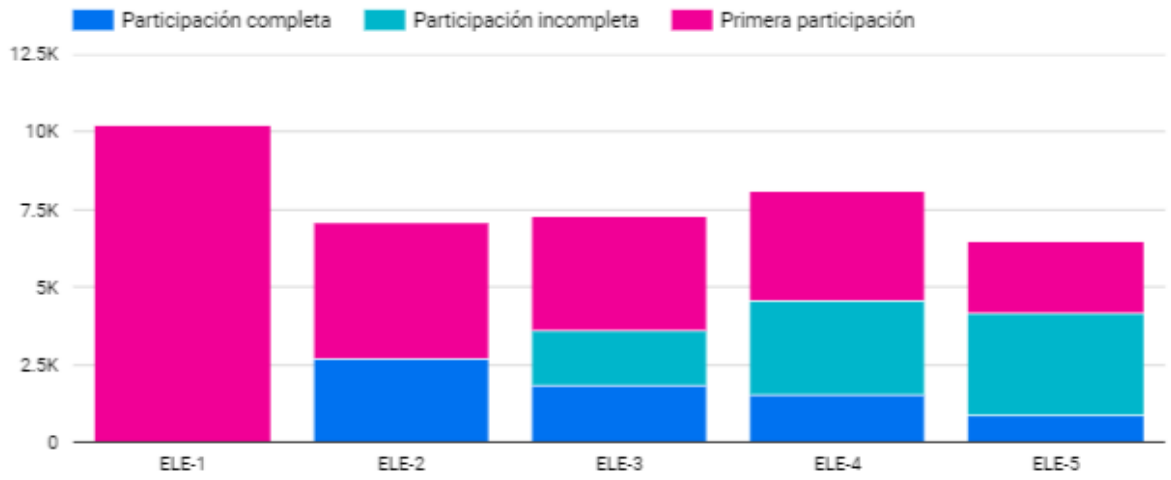


Figura 6: Cantidad de empresas según la versión de la ELE y estado de su participación.

Fuente: elaboración propia

En la Tabla 8 se enseña, de manera más detallada, los patrones formados por las empresas y su participación global en la ELE, cuantificando los distintos niveles de impacto de la atrición. Para nuestro análisis, más del 60% de las empresas totales participaron en sólo una versión de la encuesta, generando altos grados de pérdida en los datos para estas observaciones, desbalanceando así la base de datos para un estudio longitudinal. De igual manera, se puede apreciar que, aproximadamente, un 25% de las empresas presentes en el estudio tuvieron un comportamiento ideal para nuestro análisis de atrición (patrones en negrita). En otras palabras, son las empresas en las cuales, después de responder por primera vez la encuesta en cualquiera de las versiones de la ELE, no se retiró de esta posteriormente.

Tabla 8: Cantidad de empresas según su participación global en la ELE.

ELE-1	ELE-2	ELE-3	ELE-4	ELE-5	Frecuencia	Porcentaje
■					6.961	29%
	■				2.830	12%
				■	**2.307**	**10%**
		■			2.020	8%
			■		1.780	7%
			■	■	**1.763**	**7%**
		■	■	■	**839**	**3%**
■	■	■	■	■	**838**	**3%**
		■	■		813	3%
■	■				789	3%
■	■	■	■		667	3%
	■	■			561	2%
Otros patrones					1.962	< 10%
Total					**24.130**	**100%**

Por otra parte, con el objetivo de ver la relación entre las variables y el nivel de atrición de los datos, se presenta un análisis de cohorte en la Figura 7 y Figura 8. Este análisis presenta la tasa de retención de los individuos de cada encuesta con respecto a sus versiones posteriores, estudiando así la influencia de las variables sobre la tasa de pérdida. Para el primer caso, se presentan los análisis según el tamaño de la empresa, donde se observa que, para empresas más grandes, existe una tasa de atrición menor que empresas pequeñas. En la segunda, por su parte, se puede ver esta misma diferencia entre los géneros de los CEO, dado a que las empresas con CEO mujer, tienen mayor tasa de no respuesta en las siguientes encuestas con respecto a las empresas con CEO hombre. Si bien en ambos casos los porcentajes son concluyentes, existen otros factores que pueden influir en la no respuesta, especialmente según el tamaño de la empresa (como por ejemplo, la quiebra o cambio de propietario mencionados anteriormente). Sin embargo, empíricamente, se puede establecer el supuesto que la

pérdida de datos está sujeta a propiedades de las mismas observaciones más que el azar, describiendo este problema como uno de tipo MAR (*missing at random*).

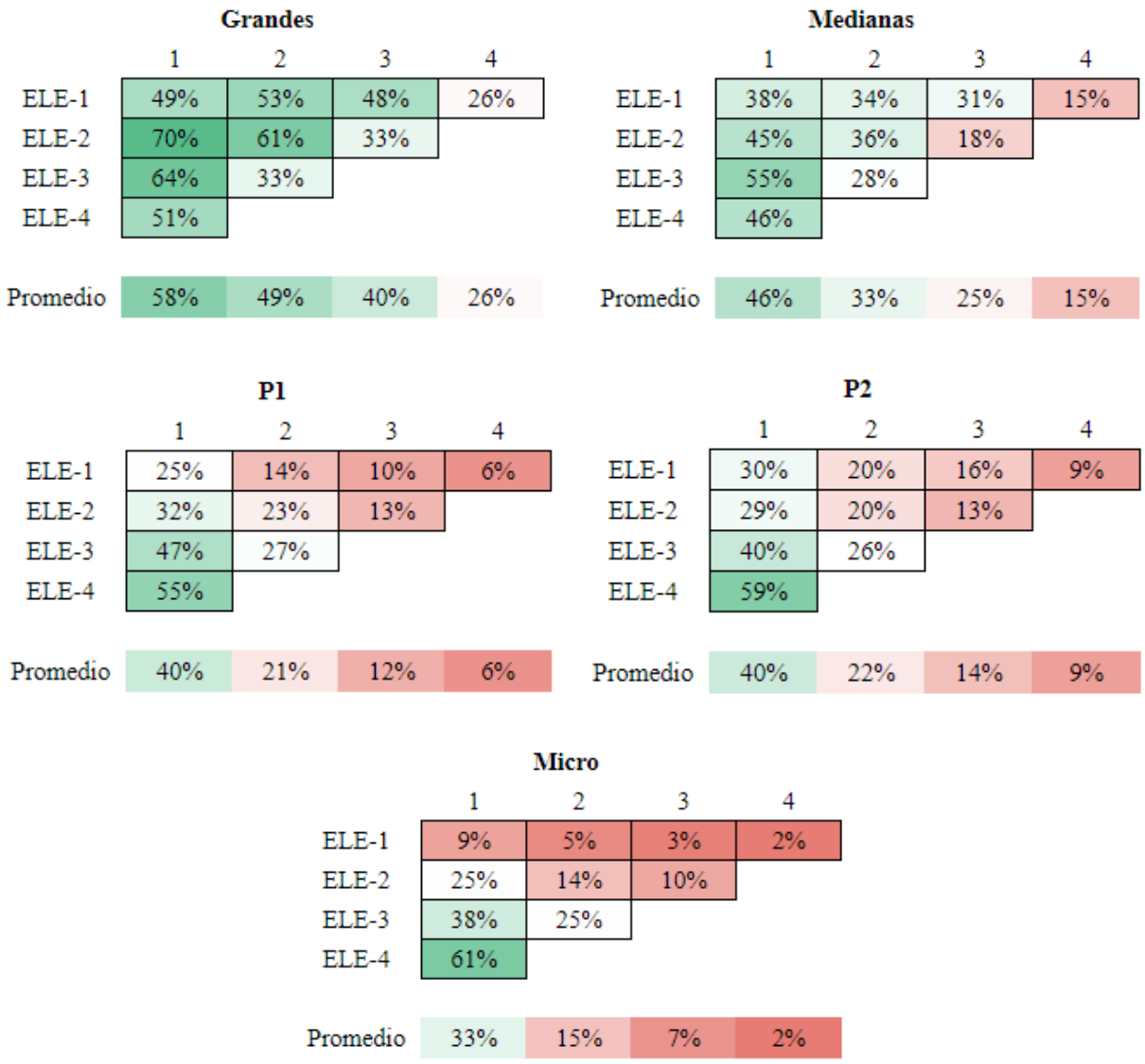


Figura 7: Análisis de cohorte de los participantes en cada versión de la ELE según el tamaño de la empresa.

Fuente: elaboración propia

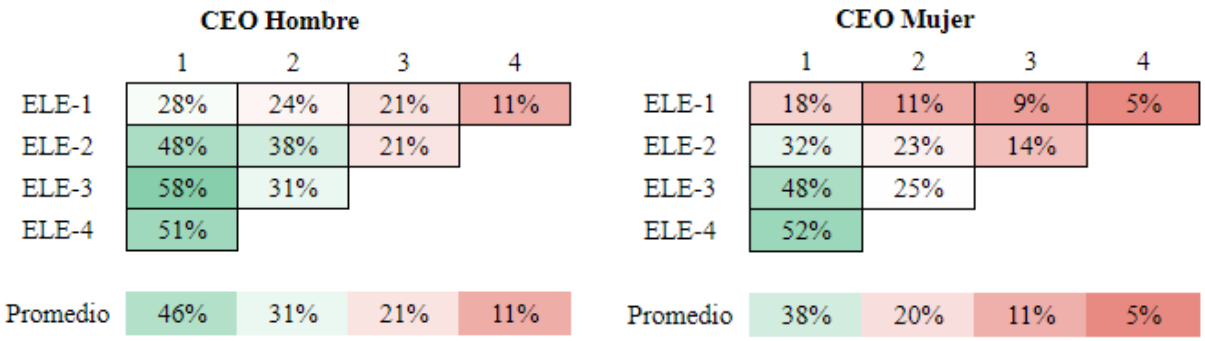


Figura 8: Análisis de cohorte de los participantes en cada versión de la ELE según el género del CEO.

Fuente: elaboración propia

4 Metodología

Tal como se presentó en el capítulo 2, existen muchos métodos para la corrección de los problemas relacionados con la atrición presente en los datos. Dentro de ellos, se encuentran los métodos de estimación de parámetros, la inyección de registros artificiales, el uso de pesos, machine learning, simulación, etc. Para el presente trabajo, en primera instancia se implementarán dos métodos que corregirán la ausencia de datos: imputación simple para la pérdida de datos e imputación múltiple para la atrición. Posterior, se realizará la elección de las variables dependientes y de control, y, por último, se ajustarán 4 modelos de estimación de parámetros para comprobar la significancia de estas variables.

4.1 Imputación simple

La imputación simple es uno de los métodos más recurrentes dentro del campo de la minería de datos para el caso de valores perdidos, y esto se debe a que su aplicación es simple y de pocos recursos computacionales. Si bien, para los casos *missing not at random* (MNAR) no es recomendado usar este método (Fielding, Fayers, McDonald, McPherson, & Campbell, 2008), es efectivo para nuestro caso de estudio (*missing at random*, MAR).

Si bien hay múltiples métodos de reemplazo de datos, para el presente trabajo se recurrirán a los siguientes:

- Por media: se imputa la media de los valores disponibles de la variable. Este método si bien mantiene estabilidad en el modelo, tiende a subestimar la varianza, disminuyendo así la aleatoriedad de la muestra.
- Por moda: se imputa la moda de los valores disponibles de la variable.
- Por regresión lineal: se imputa el valor perdido como la predicción en base a variables que no poseen pérdida. Se ocupa el método de mínimos cuadrados ordinarios para estimar los parámetros. Este método tiende a sobreestimar las correlaciones entre las variables.
- Por el arrastre de la última observación (LOCF en inglés): se imputa el valor a través de la copia de la última observación del mismo individuo. Este método se ocupa principalmente en modelos longitudinales y asume que la observación no tiene cambios a través del tiempo.

La imputación simple será ejecutada a través del entorno Google Colab y el lenguaje de programación Python.

4.2 Imputación múltiple

Si bien el método planteado de imputación simple de fácil implementación y efectivo para casos *missing at random* (MAR), tiene algunos problemas estadísticos asociados, tales como, la creación de sesgo para variables poco correlacionadas o la reducción en la eficiencia del modelo (Barnard & Meng, 1999). Este error crece según el peso de los datos faltantes con respecto al total de la muestra, debido a que el dato faltante es reemplazado directamente por otro (estimación única), para ser tratado como un dato real, sin considerar la variabilidad que puede tener este para distintos escenarios, subestimando así la varianza. Por lo que, según lo visto en el capítulo 3, el uso de este método no es el más recomendado para enfrentar el problema de la atrición dado a su alto porcentaje de pérdida.

El método de imputación múltiple (Rubin, Multiple Imputation for Nonresponse in Surveys, 1987), soluciona el problema de la poca variabilidad de las variables, dado a que, en lugar de hacer una sola estimación, se consideran múltiples valores "posibles". De este modo, el propósito central de este método no es asumir estimaciones como verdaderas, sino modelar los datos faltantes para llegar a una inferencia estadística válida en el global (Schafer, 1997).

4.2.1 Algoritmo MICE

La principal adaptación del método de imputación múltiple es el de ecuaciones encadenadas (también conocido como MICE: *multiple imputation by chained ecuations* en inglés). Este, plantea como supuesto principal que los datos ausentes correspondan al tipo *missing at random* (MAR) y que las variables tienen algún tipo de dependencia con las demás. De esta manera, las *n* estimaciones propias del método de imputación múltiple, serán una serie de estimaciones donde cada variable se turna para establecer una relación sobre el resto de variables. Este procedimiento proporciona una gran flexibilidad ya que a cada variable se le puede asignar un modelo de predicción único según sus características. El procedimiento del algoritmo MICE es el siguiente:

1) En primer lugar, se realizan *n* estimaciones puntuales para los valores faltantes, derivando en *n* sets de datos distintos. Al inicio, la primera estimación normalmente es muy alejada de la realidad, debido a que puede tomar valores aleatorios o algún estadístico de tendencia central. Sin embargo, como se

trata de un método iterativo (una iteración para cada set de datos), progresivamente las estimaciones se van acercando al valor buscado sin perder la variabilidad deseada. El número de set de datos (*n*) se define previamente. Si bien no hay un consenso con respecto a este número (depende de la estructura de los datos, porcentaje de pérdida, etc.), se plantea *n = 10* como un valor aceptable para aplicar este algoritmo.

2) Los *n* conjuntos de datos resultantes son analizados (en media, varianza, etc.) para realizar estimaciones en cada variable con datos ausentes.

3) Las estimaciones de parámetros de cada conjunto de datos imputados se combinan para obtener un solo conjunto final de estimaciones de parámetros.

Como se mencionó anteriormente, el método más usado para la predicción de los parámetros es el de regresión lineal múltiple. Sin embargo, no es el único existente, y esto se debe a que la estimación por MCO tiende a sobreestimar las correlaciones entre variables, lo que no es beneficioso si se tiene una base de datos con relaciones complejas o mucha atrición (como nuestro caso). Debido a esto y a que nuestra base de datos posee una alta dimensionalidad, se presenta Random Forest como uno de los métodos más recomendados (Wulff & Ejlskov, 2017), siendo este el método que será usado para estimar los parámetros del modelo.

4.2.2 Random Forest

Random Forest (o bosques aleatorios) es un método basado en machine learning que consta en un conjunto de árboles de decisión independientes, los cuales, si bien van entrenando con distintas muestras, lo realizan con igual distribución. En un final, estos árboles de decisión comparan sus predicciones para cada variable, compensando así los errores y generalizando los valores finales. Si bien los árboles predictores son ampliamente usados para variables de decisión, también son usados para variables continuas.

Para el caso del algoritmo MICE y Random Forest, está demostrado que las imputaciones que se intentan producir poseen un componente aleatorio por sobre la elección de la "mejor" predicción (Doove, van Buuren, & Dusseldorp, 2014), lo que ayuda en un modelo que presenta altos niveles de atrición.

Este método será ejecutado a través del entorno Google Colab y el lenguaje de programación Python, en el cual, para la implementación del algoritmo iterativo descrito se hará uso de la biblioteca importada *miceforest*.

4.3 Elección de variables

Posterior a la ejecución de la imputación simple y la imputación múltiple, dispondremos de una base de datos que no presenta pérdida, por lo que, previo a la elección de las variables (dependientes y de control), dispondremos de estadísticos más estables para todas las variables. En primera instancia, se usarán de base las variables propuestas por Carrasco (2022), para posteriormente evaluar posibles candidatos que tengan consistencia en su origen (bajo nivel de pérdida previo a la imputación simple) y alta correlación con las variables dependientes.

La correlación entre dos variables estadísticas (1) indica la fuerza y la dirección (creciente o decreciente) de una relación lineal (Kenney & Keeping, 1951). De esta forma, nos usaremos de este estadístico para medir la dependencia entre dos variables y, por consiguiente, la probabilidad de que las variables dependientes sean explicadas por los candidatos.

$$r_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} * \sqrt{n\sum y^2 - (\sum y)^2}} \quad (1)$$

Donde:

r_{xy}: Índice de correlación entre variables *x* e *y*

n: Número de observaciones

4.4 Modelos de estimación de parámetros

4.4.1 Mínimos Cuadrados Ordinarios

El método de mínimos cuadrados ordinarios (MCO) es uno de los métodos más populares para estimar los parámetros de un modelo de regresión lineal (Gujarati, 1978). MCO realiza esta la estimación de un conjunto de variables explicativas por el principio de mínimos cuadrados: se minimiza la suma de los cuadrados de las diferencias entre la variable dependiente observada en el conjunto de datos dado y los predichos por la función lineal de la variable de control (2).

$$S = \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2 \quad (2)$$

Donde:

- S: Suma de los errores cuadráticos
- Y_i: Variable dependiente observada
- $\widehat{Y}_i$: Variable dependiente predicha
- N: Número de observaciones

Generalizando a través de la ecuación de predicción del modelo de regresión lineal simple (3):

$$\widehat{Y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k X_{ki} \qquad \forall i, i \in \{1, \dots, N\} \quad (3)$$

Donde:

- $\hat{\beta}_0$: Estimador de la constante de la regresión
- $\hat{\beta}_k$: Estimadores de los parámetros k-ésimos que estima la pendiente entre la variable X_{ki} e Y_i para todo *i*
- X_{ki}: k-ésima variable de control para todo i
- K: Número de variables de control

Así, combinando (2) y (3), se obtiene la función a minimizar con el fin de encontrar los mejores estimadores (4).

$$S = \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \sum_{k=1}^{K} \hat{\beta}_k X_{ki})^2 \quad (4)$$

4.4.2 Mínimos Cuadrados Ordinarios Agrupados

El modelo de mínimos cuadrados ordinarios agrupados (MCA) presenta la misma premisa que el modelo de MCO, con la diferencia de que en este se incluye la presencia de la temporalidad (5). De esta forma, se nutre el modelo anterior con un enfoque más adecuado a un estudio longitudinal.

$$\hat{Y}_{it} = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k X_{kit} \qquad \begin{matrix} \forall i, i \in \{1, \dots, N\} \\ \forall t, t \in \{1, \dots, T\} \end{matrix} \qquad (5)$$

Donde:

$\hat{Y}_{it}$: Variable dependiente predicha para todo i y todo t

X_{kit}: k-ésima variable de control para todo i y todo t que incluye además efectos fijos de tiempo

T: Número de períodos

Así, adaptando (2) y combinándolo con (5), se obtiene la función a minimizar para este modelo (6).

$$S = \sum_{i=1}^{N} (Y_{it} - \hat{\beta}_0 - \sum_{k=1}^{K} \hat{\beta}_k X_{kit})^2 \qquad (6)$$

4.4.3 Efectos Fijos

El modelo de efectos fijos (FE) es un modelo estadístico que representa las cantidades observadas en las variables explicativas que son tratadas como si las cantidades fueran no-aleatorias, teniendo como fin eliminar el sesgo que puede causar una variable que sea constante dentro de un mismo grupo. De forma generalizada con respecto a una base de datos de panel, se tiene que (7):

$$\hat{Y}_{it} = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k X_{kit} + Y_i \sum_{k=1}^{K-1} Dummy_i \qquad (7)$$

Donde:

$Dummy_i$: Variable dummy para todo *i* hasta *n-1*

Por lo que, combinando (2) con (7), se obtiene la función para minimizar este modelo (8).

$$S = \sum_{i=1}^{N} (Y_{it} - \hat{\beta}_0 - \sum_{k=1}^{K} \hat{\beta}_k X_{kit} - Y_i \sum_{k=1}^{K-1} Dummy_i)^2 \quad (8)$$

4.4.4 Matching DID

El método de diferencias en diferencias, tiene como premisa inicial la creación de dos grupos: tratamiento y control. Luego, se realizan dos diferencias a las variaciones promedio, una para cada grupo, con el fin de poder eliminar las características no observables, constantes y no constantes, en el tiempo, respectivamente (Angrist & Pischke, 2009). Esto busca aislar todos los efectos exógenos que se pueden presentar en el modelo, dejando sólo variables netamente explicativas. Por su parte, el método de matching, crea un grupo de control que busca ser lo más similar al grupo de tratamiento, algo que se puede dificultar según la cantidad de variables que tenga (Rubin, 1973). Para este caso, la variante del *propensity score matching*, es la más efectiva para la presencia de muchas variables a estudiar (Rubin, 1997). Este método, se basa en el análisis de las variables disponibles y no en el trazado de reglas previas de correlación, por lo que podría haber relaciones entre variables que, por premisa, no existen (Rosenbaum & Rubin, 1983).

De esta manera, el método conjunto de matching con diferencias en diferencias, plantea la creación de estos grupos de control y tratamiento a través de la técnica de matching, para la posterior aplicación de las dos diferencias del método DID (Herreros, 2021). Uno de los principales motivos para la aplicación de este método, es debido a que la creación de los grupos en el método DID está directamente ligada a que deben existir observaciones que tengan efecto con y sin tratamiento. Caso contrario para el matching, el cual no plantea necesario ese requisito, haciendo que el modelo sea mucho más flexible y nutrido. Además, el método de matching se preocupa netamente de las variables observables, por lo que también el método de DID se ocupa como un corrector del modelo, eliminando la influencia no observable en este.

El estimador principal del modelo presenta que:

$$\hat{\alpha}_{MDID} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left\{ \left(y^1_{ist_1} - y^0_{ist_0}\right) - \sum_{j \in I_0 \cap S_P} W(i,j)\left(y^0_{jst_1} - y^0_{jst_0}\right) \right\} \quad (9)$$

Los 4 métodos anteriormente descritos se aplicaron para cada una de las variables dependientes a escoger e implementados a través del software Stata.

5 Resultados

5.1 Imputación simple

En primer lugar, se ejecutó la imputación simple para solucionar el problema de la ausencia implícita de datos. En las Tablas 9, 10 y 11 se detallan los métodos ocupados para cada variable que presentaba pérdida.

Tabla 9: Variables con pérdida implícita en la base de datos y sus métodos de corrección

Variable	Método
Tasa de interés por deuda	Si ya existe con el mismo ID: Media de la tasa de interés por deuda del ID Si no existe con el mismo ID: Media de la tasa de interés por deuda según su año de realización
Ventas	Regresión Lineal Múltiple Variable Dependiente: Ventas (en log) Variables de Control: ELE 2, Tamaño, Empresa grande, Número de trabajadores (en log), SRI, Dueño, Holding Ver Anexo 2 para más detalles.
Porcentaje de participación de extranjeros	Si Extranjeros = 0: Porcentaje de participación de extranjeros = 0 Si Extranjeros = 1: - Si ya existe con el mismo ID: LOCF del porcentaje de participación de extranjeros del ID - Si no existe con el mismo ID: Media del porcentaje de participación de extranjeros según variable Exporta
Familiar	Si ya existe con el mismo ID: LOCF de Familar del ID Si no existe con el mismo ID: Moda de Familar según variable Propiedad
Género	Si ya existe con el mismo ID: LOCF del género del ID Si no existe con el mismo ID: Moda del género según tamaño de empresa
Exporta	Si ya existe con el mismo ID: LOCF de Exporta del ID Si no existe con el mismo ID: Moda de Exporta según variable Empresa grande

Tabla 10: Variables con pérdida implícita en la base de datos y sus métodos de corrección (continuación)

Variable	Método
Porcentaje de exportaciones	Si $\widehat{\text{Exporta}} = 0$: Porcentaje de exportaciones = 0 Si $\widehat{\text{Exporta}} = 1$: - Si ya existe con el mismo ID: LOCF del porcentaje de exportaciones del ID - Si no existe con el mismo ID: Media del porcentaje de participación de extranjeros según variable Empresa grande
Salarios	Regresión Lineal Múltiple Variable Dependiente: Salarios (en log) Variables de Control: Número de trabajadores (en log), Tamaño, SRI, Empresa grande, Dueño Ver Anexo 2 para más detalles.
Externo	Si ya existe con el mismo ID: LOCF de Externo del ID Si no existe con el mismo ID: Moda de Externo por variable Empresa grande
Cantidad de externos	Si $\widehat{\text{Externo}} = 0$: Cantidad de externos = 0 Si $\widehat{\text{Externo}} = 1$: - Si ya existe con el mismo ID: LOCF de la cantidad de externos del ID - Si no existe con el mismo ID: Media de la cantidad de externos según tamaño de empresa
Deuda	Si ya existe con el mismo ID: LOCF de deuda del ID Si no existe con el mismo ID: Moda de la deuda según tamaño de empresa
Inventario	Regresión Lineal Múltiple Variable Dependiente: Inventario (en log) Variables de Control: Tamaño, Empresa grande, Número de trabajadores (en log), Dueño, ELE 2 Ver Anexo 2 para más detalles.

Tabla 11: Variables con pérdida implícita en la base de datos y sus métodos de corrección (continuación)

Variable	Método
Impuestos	Regresión Lineal Múltiple Variable Dependiente: Impuestos (en log)* Variable de Control: ELE 2, Tamaño, Empresa grande, Holding Ver Anexo 2 para más detalles. *Si bien los impuestos pueden tomar valores negativos, éstos son aproximadamente un 0,5% de la base total, por lo que se toma el riesgo de estimar con su versión logarítmica y la posterior aplicación de la función exponencial.
Utilidades	Si ya existe con el mismo *ID*: Media de las utilidades del ID Si no existe con el mismo *ID*: Regresión Lineal Múltiple Variable Dependiente: Utilidades (en log)* Variables de Control: $\widehat{Ventas}$ $(en$ log$)$**, ELE 2, Tamaño, Empresa grande, Holding, SRI Ver Anexo 2 para más detalles. *Si bien las utilidades pueden tomar valores negativos, siendo representados por aproximadamente un 22% de los totales, se elige utilizar, en primera instancia, la media de las utilidades para un mismo *ID*, debido a que, una estimación positiva (con regresión lineal) para una empresa que tenga sólo utilidades < 0 no haría correspondencia. Así, si es que el *ID* no posee datos de utilidades en toda la base, se procederá a realizar la estimación en base al logaritmo de éste, asumiendo que este valor será positivo. A modo de comparación, en Anexo 3 se presenta una metodología distinta a la hora de estimar esta variable. Esta, se basa en la transformación previa de las utilidades para que sean todas positivas con el fin de no haber pérdida por negativos y poder estimar todos en base a utilidades (en log). Si bien el ajuste es mejor en este nuevo modelo de regresión, se presentan resultados muy similares tanto en media como en varianza para la muestra total. **Todas las estimaciones anteriores se han realizado en base a variables sin pérdida. Sin embargo, para el caso de las utilidades (en log), existe una fuerte correlación con el estimado de las ventas (en log), y no así con otras variables. Si bien habrá un error sobre estimado, este será menor al que se incurriría en el caso de estimar las utilidades (en log) sólo con variables sin pérdida (como ELE 2, Tamaño, etc.)
Región	Si ya existe con el mismo ID: Moda de la región del ID Si no existe con el mismo ID: Moda de la región según tamaño de empresa

En la Tabla 12 se presenta el impacto del método anteriormente descrito para la media y la varianza de estas variables.

Tabla 12: Variables con pérdida implícita junto a sus estadísticos de media-varianza antes y después de la aplicación de la imputación simple

Variables	**Media antes**	**Media después**	**Dif %**	**Varianza antes**	**Varianza después**	**Dif %**
Tasa de interés por deuda	9,51	8,79	-7,6%	105,43	47,74	-54,7%
Ventas	$6{,}7 \times 10^6$	$6{,}6 \times 10^6$	-1,2%	$3{,}1 \times 10^{16}$	$3{,}1 \times 10^6$	-1,3%
Porcentaje de participación de extranjeros	4,58	6,84	+49,2%	411,61	578,39	+40,5%
Familiar	0,39	0,36	-8,1%	0,24	0,23	-3,2%
Género	0,82	0,82	+0,3%	0,15	0,14	-1,3%
Exporta	0,11	0,11	0%	0,10	0,10	0%
Porcentaje de exportaciones	4,27	4,27	+0,1%	316,71	316,79	0%
Salarios	3574,06	3521,51	-1,5%	$8{,}7 \times 10^8$	$8{,}6 \times 10^8$	-1,6%
Externo	0,13	0,15	+14,9%	0,11	0,13	+12,3%
Cantidad de externos	208,26	242,64	+16,5%	$1{,}7 \times 10^7$	$1{,}6 \times 10^7$	-4,7%
Deuda	0,71	0,78	+10,%	0,20	0,17	-18,3%
Inventario	$1{,}3 \times 10^7$	$1{,}2 \times 10^7$	-7,8%	$4{,}5 \times 10^{18}$	$4{,}2 \times 10^{18}$	-7,8%
Impuestos	$2{,}1 \times 10^5$	$2{,}0 \times 10^5$	-6,4%	$2{,}7 \times 10^{14}$	$2{,}5 \times 10^{14}$	-6,4%
Utilidades	$1{,}9 \times 10^6$	$1{,}9 \times 10^6$	-1,3%	$5{,}4 \times 10^{16}$	$5{,}3 \times 10^{16}$	-1,3%
Región	10,4	10,1	-2,5%	13,52	14,39	+6,4%

Tal como se tenía de premisa, el método de imputación simple presenta estabilidad en torno a la media para las variables con datos ausentes (exceptuando el porcentaje de participación de extranjeros, quien,

a través de LOCF es muy probable que haya arrastrado datos *outliers*, sesgando la media total para la variable). Por otra parte, también se planteó que una de las desventajas principales que posee este método es la subestimación de la variabilidad. Sin embargo, para la mayoría de las variables (exceptuando la tasa de interés por deuda que presentaba una pérdida muy grande) la varianza se mantiene muy estable en comparación a su estado previo a la imputación. De esta forma, la aplicación del método de imputación simple no afectó de manera sustancial la base de datos, manteniendo a los estadísticos de tendencia central dentro de un rango aceptable para la aplicación del posterior método.

5.2 Imputación múltiple

Para solucionar el problema de la atrición se ejecutó el método de imputación múltiple con 10 iteraciones (sets de datos) bajo el algoritmo MICE y Random Forest para la estimación.

Como resultado, se imputaron 49.948 registros, los cuales se distribuyen de la siguiente manera:

Tabla 13: Comparación entre la cantidad de registros originales e imputados para cada tanda de la ELE.

Versión de la encuesta	Registros originales	Registros imputados
ELE-1	10.213	0
ELE-2	7.062	7.546
ELE-3	7.267	11.013
ELE-4	8.084	13.739
ELE-5	6.480	17.650

Con respecto al detalle, en la Figura 9, se presentan las empresas participantes según su tamaño, posterior a la realización de la imputación múltiple. Lo cual, comparando con la Figura 1, demuestra que los pesos de cada una de las clasificaciones se mantienen muy estables en torno a la base original, sólo cambiando la cantidad de encuestas respondidas.

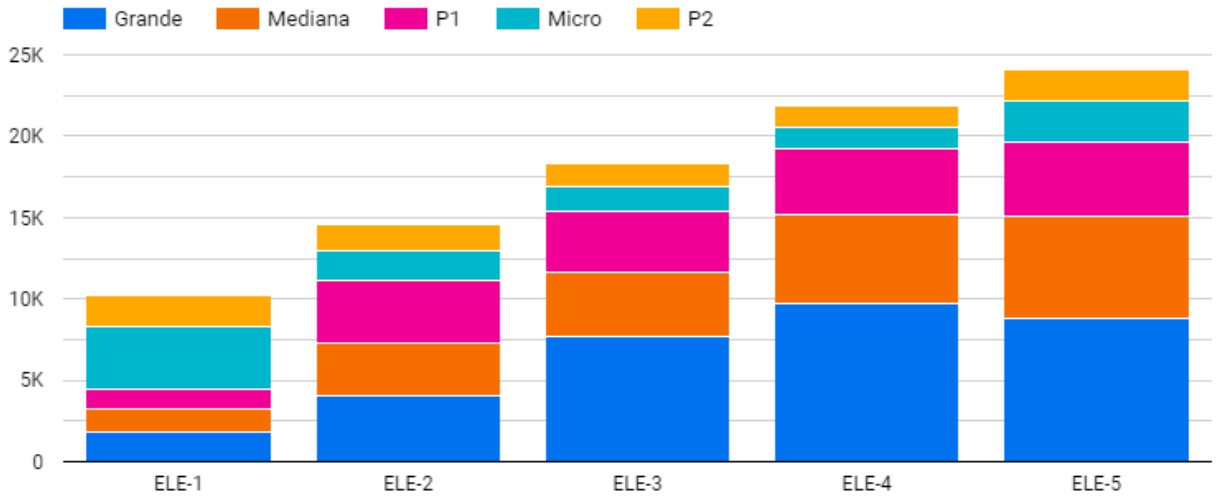


Figura 9: Cantidad de empresas participantes según la versión de la ELE y su tamaño, posterior a la imputación múltiple

Fuente: elaboración propia

5.3 Elección de variables

5.3.1 Variables dependientes

Carrasco (2022), planteó el uso de 4 variables dependientes que miden el performance de una empresa: productividad laboral de los empleados (ventas divido sobre número de empleados, en logaritmo), número de empleados o trabajadores (en logaritmo), ventas anuales (en logaritmo) y utilidades anuales. El uso de las utilidades tiene el problema que, al estar expresado en términos monetarios, hay alta dispersión de datos. Además, es posible que una empresa tenga pérdidas, y en este caso, la utilidad es negativa, imposibilitando la aplicación de logaritmo.

Para la sustitución de esta variable, se propone la inclusión del margen de utilidad de una empresa, llamado para nuestra base de datos como margen de utilidad (10):

$$\text{Margen de utilidad} = \frac{\text{Utilidad}}{\text{Ventas}} * 100 \qquad (10)$$

Dado a que existen outliers muy influyentes en esta variable creada, se realizará la estimación para los registros que se exclusivamente se encuentren dentro del rango [-4.000%, 4.000%], excluyendo de este modelo 804 registros (0,9% del total).

Además, para evaluar la robustez de la base, se propone la inclusión de la tasa de crecimiento de las utilidades (11):

$$\text{Tasa de crecimiento de la utilidad} = \frac{\text{Utilidad}_t}{\text{Utilidad}_{t-1}} - 1 \qquad (11)$$

Dado que para $t = 1$ no existen predecesores para realizar la comparación, se prescindirán de esos registros para este modelo, y, al igual que en la variable pasada, también se excluirán los registros que no se encuentren dentro del rango [-40, 40].

Si bien la elección del rango para excluir *outliers* se planteó bajo la sensibilidad de los límites con respecto a la media (con el fin de no perder muchos registros), en el Anexo 4, se plantea el método 1,5 IQR para la eliminación de estos datos atípicos. Este análisis muestra que la media se mantiene relativamente cercana a la actual, pero la desviación estándar es mucho menor.

De esta forma, en la Tabla 14 se presentan las variables dependientes a estudiar.

Tabla 14: Variables dependientes de los modelos de estimación

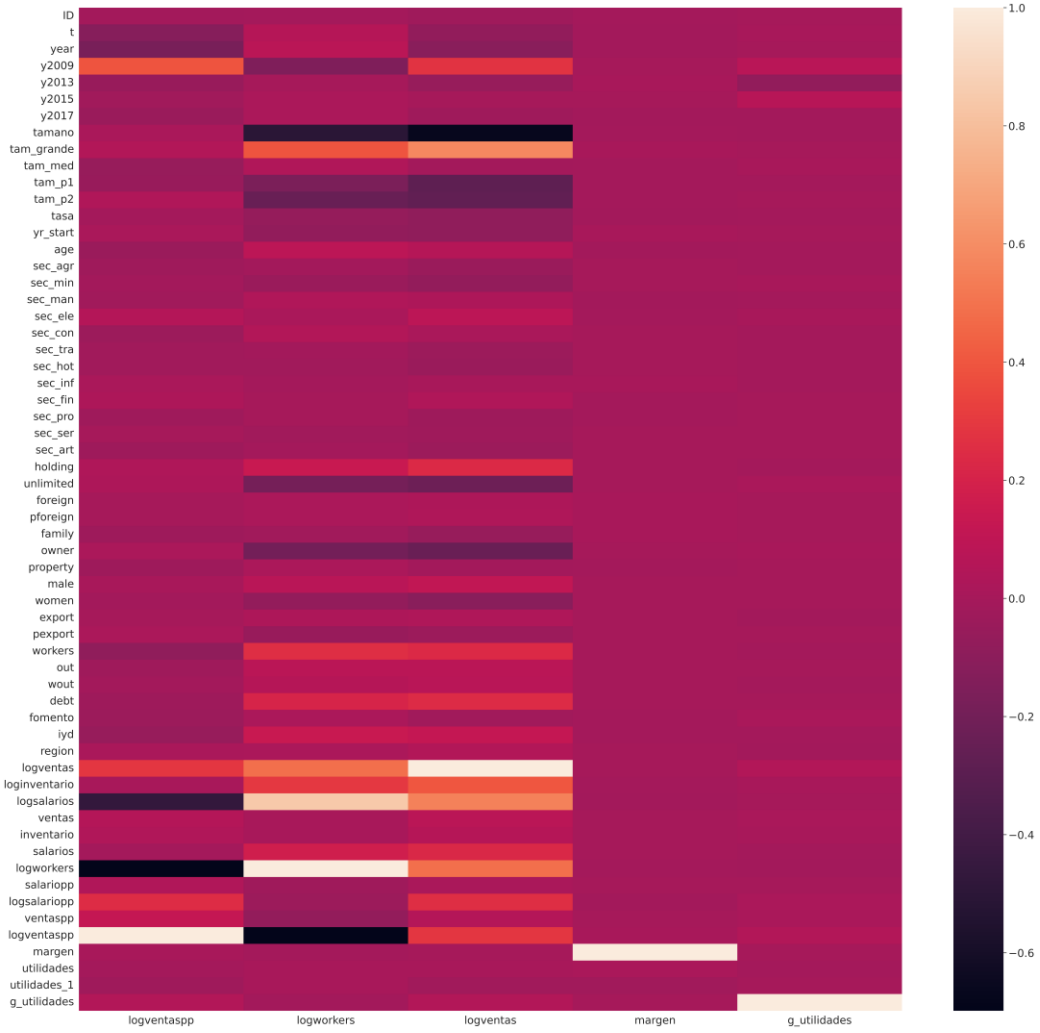
Variable	**Frecuencia**	**Media**	**Desv. Std.**	**Mín.**	**Max**
Productividad laboral de los empleados (en log)	89.054	4,331	4,248	-16,466	27,836
Número de trabajadores (en log)	89.054	4,051	3,597	-4,605	13,461
Ventas (en log)	89.054	8,383	3,466	-6,812	23,737
Margen de utilidad	88.520	4,005	127,126	-3.569,963	3.889,286
Tasa de crecimiento de la utilidad	53.328	-0,153	5,873	-39,769	39,982

5.3.2 Variables de control

Carrasco (2022), planteó el uso de las siguientes variables de control: Género (mujer), Edad, Extranjera, IyD, Exporta y Deuda, las cuales, debido a que todas mostraron algún nivel de significancia según la variable dependiente, se mantendrán para este trabajo.

Además, teniendo la base de datos con la corrección de ambas problemáticas, se realiza un análisis a través de sus correlaciones para observar la fuerza de los candidatos con respecto a las variables dependientes, tal como se ve en la Figura 10.

Figura 10: Correlaciones entre las variables dependientes y el total de variables



Fuente: elaboración propia

Se puede apreciar relación medianamente considerable entre la presencia de holding y las ventas (en log), holding y el número de trabajadores (en log), y la presencia de externos y el número de trabajadores (en log), todas positivas. Por otra parte, el tamaño de la empresa (variable donde 1 es empresa grande y 5 es microempresa) tiene una relación muy fuerte con las ventas (en log) y el número de trabajadores (en log), pero de manera inversa. De esta manera, se incluirán estas 4 variables como variables explicativas, siendo presentadas, junto a las planteadas por Carrasco (2022), en la Tabla 15.

Tabla 15: Variables de control a usar en los modelos de estimación

Variable	**Para la base de datos completa (n = 89.054)**				
	Frecuencia	**Media**	**Desv. Std.**	**Mín.**	**Max**
Género (mujer)	89.054	0,241	0,428	0	1
Edad	89.054	16,963	11,340	0	190
Extranjera	89.054	0,155	0,362	0	1
IyD	89.054	8,383	3,466	0	1
Exporta	89.054	0,200	0,400	0	1
Deuda	89.054	0,729	0,444	0	1
Holding	89.054	0,269	0,443	0	1
Externo	89.054	0,241	0,428	0	1
Tamaño	89.054	2,390	1,373	1	5

5.4 Modelos de estimación de parámetros

Para cada una de las variables dependientes anteriormente explicadas, se implementarán los 4 modelos econométricos descritos en el capítulo 4 (MCO, MCA, FE y Matching DID). En el caso del modelo de Matching DID, los grupos de control y tratamiento se definen de la siguiente manera:

- Grupo de control: No hay cambio de Género (mujer) entre dos períodos de tiempo (0 a 0, 1 a 1).
- Grupo de tratamiento: Hay cambio de Género (mujer) desde 0 a 1 entre dos períodos de tiempo.

Además, con el acrónimo *DIFF*, se hará referencia a la diferencia de la variable correspondiente entre dos períodos de tiempo (DID).

5.4.1 Productividad laboral de los empleados

En primer lugar, se utilizó la productividad laboral de los empleados (en log) como variable dependiente. En las Tablas 16 y 17, se pueden ver los resultados de manera resumida. La principal desigualdad a la hora de comparar los resultados, es la abismal diferencia entre los R^2. Carrasco (2022) obtuvo 0,57 a la hora de ajustar el modelo, mientras que los efectuados en el presente trabajo no superan el 0,10 de ajuste. Dado esto, se puede plantear como hipótesis, que el algoritmo de imputación múltiple afectó negativamente en la estimación puntual de las variables relacionadas con las ventas y el número de trabajadores. A pesar de esto, los modelos de MCO, MCA y FE son significativos y poseen la mayoría de las variables como buenos predictores de la productividad laboral de los empleados (en log). Al igual como afirmó Carrasco (2022), la variable Género (mujer), tiene una influencia negativa dentro de la productividad de la empresa, pero, para nuestros modelos, con menos porcentaje.

Tabla 16: Comparación de modelos para la productividad laboral de los empleados (en log)

Variable	**Carrasco (2022), Regresión Lineal Múltiple**	**MCO**	**MCA**	**FE**	**Matching DID**
Género (mujer)	-0,631* [0,238]	-0,318** [0,033]	-0,318** [0,033]	-0,356** [0,049]	-
Edad	-0,048** [0,006]	-0,059** [0,001]	-0,059** [0,001]	-0,207** [0,002]	0,000 [0,001]

Nota: (*): significativo al 10%, (**): significativo al 5%

Tabla 17: Comparación de modelos para la productividad laboral de los empleados (en log) (continuación)

Variable	Carrasco (2022), Regresión Lineal Múltiple	MCO	MCA	FE	Matching DID
Extranjera	-3,586** [0,104]	-0,340** [0,039]	-0,340** [0,039]	-0,501** [0,047]	0,051** [0,017]
IyD	-5,744** [0,079]	-0,866** [0,033]	-0,866** [0,033]	-1,120** [0,041]	0,002 [0,016]
Exporta	0,320** [0,188]	-0,100** [0,035]	-0,100** [0,035]	-0,314** [0,042]	0,037** [0,016]
Deuda	-0,306 [0,112]	-0,499** [0,032]	-0,499** [0,032]	-0,776** [0,037]	0,050** [0,015]
Holding	-	0,236** [0,033]	0,236** [0,033]	-0,249** [0,046]	0,015 [0,015]
Externo	-	-0,388** [0,033]	-0,388** [0,033]	-0,725** [0,039]	0,016 [0,015]
Tamaño	-	0,071** [0,011]	0,071** [0,011]	0,006 [0,015]	0,029** [0,006]
R^2	0,572	0,046	0,046	0,040	0,001
n	16.425	89.054	89.054	89.054	60.586
Prueba F	257,59**	475,43**	475,43**	1303,95**	-
LR chi2	-	-	-	-	45,96**

Nota: (*): significativo al 10%, (**): significativo al 5%

Para el modelo de Matching DID, se aprecia que las variables de control son menos explicativas para la variable de tratamiento (en comparación de los modelos anteriores). Sin embargo, tal como se aprecia en la Tabla 18, el ajuste Matching a la diferencia en la productividad laboral de los empleados (en log) es significativo al 5%, lo que quiere decir, que los grupos tratamiento-control tienen diferencias lo suficientemente significativas entre sí (con respecto a esta variable) como para ser influyentes en él. En este caso, la diferencia de los coeficientes para los grupos es de 1,198 (ATE,

efecto promedio del tratamiento), lo que se interpreta como una mayor influencia del grupo de tratamiento por sobre el de control (> 0). En este caso, como nuestro grupo de tratamiento es el cambio de gerencia de hombre a mujer entre dos períodos, se puede afirmar que existe una relación positiva entre este suceso y la productividad laboral de los empleados (en log), contrastando con lo obtenido por Carrasco (2022) y los modelos ya estimados (MCO, MCA y FE).

Tabla 18: Modelo de Matching DID para la productividad laboral de los empleados (en log)

Variable	**Sample**	**Grupo de Tratamiento**	**Grupo de Control**	**Diferencia entre grupos**	**Prueba T [Error estándar]**
DIFF de productividad laboral de los empleados	Unmatched	0,071	-1,179	1,250	16,02** [0,078]
	ATT	0,071	-0,560	0,630	3,42** [0,184]
	ATU	-1,179	0,088	1,268	
	ATE			1,198	

Nota: (*): significativo al 10%, (**): significativo al 5%

5.4.2 Número de trabajadores

Posteriormente, se utilizó el número de trabajadores (en log) como variable dependiente. En la Tabla 19, se pueden ver los resultados de manera resumida. Nuevamente existe una desigualdad marcada dentro de los R^2, sin embargo, no es alarmante. De igual manera, los modelos de MCO, MCA y FE son significativos y tienen todas sus variables como buenos predictores del número de trabajadores (en log). Para la variable Género (mujer) existe una diferencia con respecto a lo visto por Carrasco (2022), debido a que, tanto MCO como MCA estiman su influencia como negativa (-3,9%). Sin embargo, el modelo de FE, concuerda muy cercano a lo que estimó Carrasco (2022), afirmando que la influencia de la mujer dentro de una gerencia es positiva en torno a la cantidad de trabajadores (14,5%).

Tabla 19: Comparación de modelos para el número de trabajadores (en log)

Variable	Carrasco (2022), Regresión Lineal Múltiple	MCO	MCA	FE	Matching DID
Género (mujer)	0,188* [0,087]	-0,039* [0,023]	-0,039* [0,023]	0,145** [0,030]	-
Edad	0,012** [0,002]	0,035** [0,001]	0,035** [0,001]	0,085** [0,001]	0,000 [0,001]
Extranjera	-0,297** [0,038]	0,247** [0,027]	0,247** [0,027]	0,369** [0,030]	0,051** [0,017]
IyD	1,984** [0,029]	0,566** [0,023]	0,566** [0,023]	0,403** [0,026]	0,002 [0,016]
Exporta	-0,008 [0,072]	0,246** [0,024]	0,246** [0,024]	0,367** [0,026]	0,037** [0,016]
Deuda	-0,010 [0,040]	0,105** [0,022]	0,105** [0,022]	-0,091** [0,023]	0,050** [0,015]
Holding	-	0,134** [0,023]	0,134** [0,023]	0,323** [0,029]	0,015 [0,015]
Externo	-	0,513** [0,023]	0,513** [0,023]	0,601** [0,024]	0,016 [0,015]
Tamaño	-	-1,409** [0,008]	-1,409** [0,008]	-1,257** [0,009]	0,029** [0,006]
R^2	0,652	0,3707	0,3707	0,3441	0,001
n	16.425	89.054	89.054	89.054	60.586
Prueba F	380,60**	5.827,68**	5.827,68**	3620,92**	-
LR chi2 (modelo)	-	-	-	-	45,96**

Nota: (*): significativo al 10%, (**): significativo al 5%

Para el modelo de Matching DID, se aprecia que las variables de control son menos explicativas para la variable de tratamiento (en comparación de los modelos anteriores). Sin embargo, tal como se aprecia en la Tabla 20, el ajuste Matching a la diferencia en el número de trabajadores (en log) es significativo al 5%, lo que quiere decir, que los grupos tratamiento-control tienen diferencias lo suficientemente significativas entre sí (con respecto a esta variable) como para ser influyentes en él. En este caso, la diferencia de los coeficientes es de -0,241 (ATE), lo que se interpreta como una mayor influencia del grupo de control por sobre el de tratamiento (< 0). En este caso, como nuestro grupo de control es el status-quo en el género de la gerencia entre dos períodos, se puede afirmar que existe una relación positiva entre este suceso y el número de trabajadores (en log). En otras palabras, este modelo concluye que el cambio de gerencia desde CEO hombre a CEO mujer no es más beneficioso que una continuidad del género en la gerencia, alineándose a lo comprobado con MCO y MCA.

Tabla 20: Modelo de Matching DID para el número de trabajadores (en log)

Variable	**Sample**	**Grupo de Tratamiento**	**Grupo de Control**	**Diferencia entre grupos**	**Prueba T [Error estándar]**
DIFF de número de trabajadores (en log)	Unmatched	0,548	0,943	-0,396	-8,57** [0,046]
	ATT	0,548	0,928	-0,380	-3,32** [0,114]
	ATU	0,943	0,719	-0,224	
	ATE			-0,241	

Nota: (*): significativo al 10%, (**): significativo al 5%

5.4.3 Ventas

Para analizar las ventas por separado de la productividad, se utilizaron las ventas (en log) como variable dependiente. En las Tablas 21 y 22, se pueden ver los resultados de manera resumida. Nuevamente existe una desigualdad marcada dentro de los R^2, sin embargo, no es alarmante. De igual manera, los modelos de MCO, MCA y FE son significativos y tienen todas sus variables como buenos

predictores de las ventas (en log). Al igual como afirmó Carrasco (2022), la variable Género (mujer) tiene una influencia negativa dentro de las ventas de la empresa, acercándose a lo estimado posteriormente (-44% vs. -35%).

Tabla 21: Comparación de modelos para las ventas (en log)

Variable	**Carrasco (2022), Regresión Lineal Múltiple**	**MCO**	**MCA**	**FE**	**Matching DID**
Género (mujer)	-0,442* [0,194]	-0,357** [0,023]	-0,357** [0,023]	-0,211** [0,035]	-
Edad	-0,035** [0,004]	-0,024** [0,001]	-0,024** [0,001]	-0,122** [0,002]	0,000 [0,001]
Extranjera	-3,884** [0,091]	-0,093** [0,028]	-0,093** [0,028]	-0,132** [0,034]	0,051** [0,017]
IyD	-3,759 [0,068]	-0,300** [0,023]	-0,300** [0,023]	-0,796** [0,030]	0,002 [0,016]
Exporta	0,311 [0,166]	0,147** [0,025]	0,147** [0,025]	0,052* [0,030]	0,037** [0,016]
Deuda	-0,317* [0,166]	-0,394** [0,023]	-0,394** [0,023]	-0,866** [0,027]	0,050** [0,015]
Holding	-	0,370** [0,023]	0,370** [0,023]	0,074** [0,034]	0,015 [0,015]
Externo	-	0,125** [0,023]	0,125** [0,023]	-0,123** [0,028]	0,016 [0,015]
Tamaño	-	-1,339** [0,037]	-1,339** [0,037]	-1,251** [0,052]	0,029** [0,006]
R^2	0,473	0,2904	0,2904	0,185	0,001
n	16.425	89.054	89.054	89.054	60.586
Prueba F	151,04**	4.048,39**	4.048,39**	1950,65**	-

Nota: (*): significativo al 10%, (**): significativo al 5%

Tabla 22: Comparación de modelos para las ventas (en log) (continuación)

Variable	Carrasco (2022), Regresión Lineal Múltiple	MCO	MCA	FE	Matching DID
LR chi2 (modelo)	-	-	-	-	45,96**

Nota: (*): significativo al 10%, (**): significativo al 5%

Para el modelo de Matching DID, se aprecia que las variables de control son menos explicativas para la variable de tratamiento (en comparación de los modelos anteriores). Además, tal como se ve en la Tabla 23, el ajuste Matching a las ventas (en log), es no significativo al 5% (en los tratados por matching, ATT), lo que quiere decir, que los grupos tratamiento-control no tienen diferencias lo suficientemente significativas entre sí (con respecto a esta variable) como para ser influyentes en él. De esta forma, este modelo no es lo suficientemente explicativo.

Tabla 23: Modelo de Matching DID para las ventas (en log)

Variable	Sample	Grupo de Tratamiento	Grupo de Control	Diferencia entre grupos	Prueba T [Error estándar]
DIFF de ventas (en log)	Unmatched	0,618	-0,236	0,854	13,32** [0,064]
	ATT	0,618	0,368	0,250	1,59 [0,157]
	ATU	-0,236	0,807	1,043	
	ATE			0,956	

Nota: (*): significativo al 10%, (**): significativo al 5%

5.4.4 Margen de utilidad

Como primera variable propuesta, se utilizó el margen de utilidad como variable dependiente. En las Tablas 24 y 25, se pueden ver los resultados de manera resumida. Si bien todos los modelos presentan un R^2 muy bajo, tienen una prueba F significativa. Tanto MCO, MCA y FE, poseen muy pocas variables explicativas para esta variable, entre los cuales no se encuentra Género (mujer), siendo no significativo para todos los casos al 5% y 10%.

Tabla 24: Comparación de modelos para la variable margen de utilidad

Variable	**Carrasco (2022), Regresión Lineal Múltiple**	**MCO**	**MCA**	**FE**	**Matching DID**
Género (mujer)	-	-0,940 [1,007]	-0,940 [1,007]	-1,298 [1,469]	-
Edad	-	0,038 [0,038]	0,038 [0,038]	-0,220** [0,070]	0,000 [0,001]
Extranjera	-	-3,176** [1,207]	-3,176** [1,207]	-1,301 [1,417]	0,051** [0,017]
IyD	-	1,101 [1,020]	1,101 [1,020]	1,131 [1,244]	-0,002 [0,016]
Exporta	-	-2,120* [1,088]	-2,120* [1,088]	-0,387 [1,266]	0,038** [0,016]
Deuda	-	-1,826* [0,988]	-1,826* [0,988]	-1,698 [1,122]	0,053 [0,016]
Holding	-	-0,437 [1,018]	-0,437 [1,018]	-2,886** [1,390]	0,011 [0,015]
Externo	-	4,355** [1,023]	4,355** [1,023]	-4,138** [1,170]	0,014 [0,015]
Tamaño	-	-3,059** [0,343]	-3,059** [0,343]	-2,327** [0,455]	0,027** [0,006]

Nota: (*): significativo al 10%, (**): significativo al 5%

Tabla 25: Comparación de modelos para la variable margen de utilidad (continuación)

Variable	Carrasco (2022), Regresión Lineal Múltiple	MCO	MCA	FE	Matching DID
R^2	-	0,0012	0,0012	0,0060	0,0010
n	-	88.520	88.520	88.520	59.973
Prueba F	-	12,15**	12,15**	5,41**	-
LR chi2 (modelo)	-	-	-	-	42,47**

Nota: (*): significativo al 10%, (**): significativo al 5%

Para el modelo de Matching DID, se aprecia que las variables de control son menos explicativas para la variable de tratamiento (en comparación de los modelos anteriores), debido a que son muy pocos los que intentan predecir dentro de un modelo con muy poco ajuste. Además, tal como se ve en las Tablas 26 y 27, el ajuste Matching al margen de utilidad, es no significativo al 5% (en los tratados y no matcheados), lo que quiere decir, que los grupos tratamiento-control no tienen diferencias lo suficientemente significativas entre sí (con respecto a esta variable) como para ser influyentes en él. De esta forma, este modelo no es lo suficientemente explicativo.

Tabla 26: Modelo de Matching DID para el margen de utilidad

Variable	Sample	Grupo de Tratamiento	Grupo de Control	Diferencia entre grupos	Prueba T [Error estándar]
DIFF de margen de utilidad	Unmatched	-1,010	-1,054	0,044	0,02 [2,206]
	ATT	-1,010	-7,933	6,923	1,12 [6,155]

Nota: (*): significativo al 10%, (**): significativo al 5%

Tabla 27: Modelo de Matching DID para el margen de utilidad (continuación)

Variable	Sample	Grupo de Tratamiento	Grupo de Control	Diferencia entre grupos	Prueba T [Error estándar]
DIFF de margen de utilidad	ATU	-1,054	8,719	9,773	
	ATE			9,462	

Nota: (*): significativo al 10%, (**): significativo al 5%

5.4.5 Tasa de crecimiento de la utilidad

Por último, como segunda variable propuesta, se utilizó la tasa de crecimiento de la utilidad como variable dependiente. En las Tablas 28 y 29, se pueden ver los resultados de manera resumida. Si bien todos los modelos presentan un R^2 muy bajo, tienen una prueba F significativa (al 10% para MCO y MCA, y al 5% para FE). Tanto MCO, MCA y FE, poseen muy pocas variables explicativas para esta variable, entre los cuales, nuevamente, no se encuentra Género (mujer), siendo no significativo para todos los casos al 5% y 10%.

Tabla 28: Comparación de modelos para la variable tasa de crecimiento de la utilidad

Variable	Carrasco (2022), Regresión Lineal Múltiple	MCO	MCA	FE	Matching DID
Género (mujer)	-	0,028 [0,058]	0,028 [0,058]	0,005 [0,104]	-
Edad	-	-0,002 [0,002]	-0,002 [0,002]	0,019** [0,006]	0,001 [0,001]
Extranjera	-	0,025 [0,065]	0,025 [0,065]	-0,025 [0,088]	0,050** [0,024]

Nota: (*): significativo al 10%, (**): significativo al 5%

Tabla 29: Comparación de modelos para la variable tasa de crecimiento de la utilidad (continuación)

Variable	**Carrasco (2022), Regresión Lineal Múltiple**	**MCO**	**MCA**	**FE**	**Matching DID**
Género (mujer)	-	0,028 [0,058]	0,028 [0,058]	0,005 [0,104]	-
Edad	-	-0,002 [0,002]	-0,002 [0,002]	0,019** [0,006]	0,001 [0,001]
Extranjera	-	0,025 [0,065]	0,025 [0,065]	-0,025 [0,088]	0,050** [0,024]
IyD	-	-0,033 [0,059]	-0,033 [0,059]	-0,099 [0,092]	-0,046** [0,022]
Exporta	-	-0,117 [0,060]	-0,117 [0,060]	-0,212** [0,081]	0,059** [0,023]
Deuda	-	-0,042 [0,058]	-0,042 [0,058]	0,226** [0,078]	0,056** [0,023]
Holding	-	-0,107 [0,057]	-0,107 [0,057]	0,063 [0,097]	0,025 [0,021]
Externo	-	0,045 [0,057]	0,045 [0,057]	-0,005 [0,076]	0,025 [0,022]
Tamaño	-	-0,069 [0,022]	-0,069 [0,022]	-0,195** [0,039]	-0,002 [0,009]
R^2	-	0,0003	0,0003	0,0001	0,0013
n	-	53.328	53.328	53.328	27.852
Prueba F	-	1,87*	1,87*	6,80**	-
LR chi2 (modelo)	-	-	-	-	26,60**

Nota: (*): significativo al 10%, (**): significativo al 5%

Para el modelo de Matching DID, se aprecia que las variables de control son menos explicativas para la variable de tratamiento (en comparación de los modelos anteriores), debido a que son muy pocos

los que intentan predecir dentro de un modelo con muy poco ajuste. Además, tal como se ve en la Tabla 30, el ajuste Matching a la tasa de crecimiento de la utilidad, es no significativo al 5% (en los tratados y no matcheados), lo que quiere decir, que los grupos tratamiento-control no tienen diferencias lo suficientemente significativas entre sí (con respecto a esta variable) como para ser influyentes en él. De esta forma, este modelo no es lo suficientemente explicativo.

Tabla 30: Modelo de Matching DID para la tasa de crecimiento de la utilidad

Variable	**Sample**	**Grupo de Tratamiento**	**Grupo de Control**	**Diferencia entre grupos**	**Prueba T [Error estándar]**
DIFF de tasa de crecimiento de la utilidad	Unmatched	0,059	0,098	-0,039	-0,24 [0,165]
	ATT	0,059	-0,261	0,320	0,87 [0,367]
	ATU	0,098	-0,083	-0,181	
	ATE			-0,121	

Nota: (*): significativo al 10%, (**): significativo al 5%

6 Conclusiones

En este trabajo, se emplearon métodos para la corrección de la atrición de la base de datos de las 5 versiones de la Encuesta Longitudinal de Empresas, basado en el trabajo realizado por Carrasco (2022).

Para el caso de ambos métodos correctivos (imputación simple e imputación múltiple), se obtuvieron resultados que no cambiaron de mayor forma los pesos y la composición de la base de datos. Esto se debe, a que, para el caso de la imputación simple, mantuvieron en la mayoría de las variables la media y la varianza. Mientras que para la imputación múltiple, hubo, de igual forma, estabilidad en torno a la composición de la muestra (actividades, tamaño de empresa, etc.). Así, se comprobó empíricamente que el algoritmo basado en árboles de decisión con un componente aleatorio evita el sesgo excesivo al imputar datos.

En torno a los resultados obtenidos al adaptar los 4 modelos de estimación de parámetros, se puede afirmar que la imputación múltiple provocó bajas en las correlaciones y mayor dispersión en la base de datos (caída generalizada en los R^2), especialmente para la productividad laboral de los empleados (en log). A pesar de esto, los modelos de estimación para la productividad laboral de los empleados (en log), al número de trabajadores (en log) y ventas (en log) fueron significativos. De esta forma, si bien la imputación múltiple tuvo un correcto desempeño, existe la hipótesis que puede haber afectado a variables como ventas (en log) o el número de trabajadores (en log) en demasía, tema en el cual existe una opción de mejora a futuro.

En comparación con lo trabajado por Carrasco (2022), hay bastante similitud en torno a afirmar la influencia negativa (a excepción de Matching DID) de Género (mujer) dentro de la productividad laboral de los empleados (en log) y ventas (en log). Mientras que para el número de trabajadores (en log), las estimaciones fueron irregulares en torno al signo, no lográndose una afirmación con respecto a la influencia de Género (mujer) sobre éste.

Por otra parte, tal como lo había descrito Carrasco (2022), las utilidades son uno de los principales reflectores del status de una empresa, pero que, para términos de ajuste de regresión, no tenían buen rendimiento. De esta manera, en el presente trabajo se planteó el uso de la variable margen de utilidad y tasa de crecimiento de la utilidad, quienes otorgaban una visión proporcional de las utilidades, más que neta. Sin embargo, al igual que lo comprobado por Carrasco (2022) con utilidades, estas variables

presentaron niveles de ajuste muy bajos, pudiéndose deber a que el rango de utilidades crece de manera exponencial para las empresas con mayores ingresos (incluso habiendo eliminado *outliers*). Por lo que, se puede afirmar, que el margen de utilidad, la tasa de crecimiento de la utilidad y las utilidades poseen la suficiente dispersión y poca linealidad que provocan que, al seleccionarlos como variables dependientes dentro de un modelo lineal, torne éstos a no significativos, siendo no recomendado su uso para efectos de predicción de performance de una empresa.

Por último, otras oportunidades de mejora radican en la experimentación de uso de nuevos métodos de corrección con respecto a los usados en la presente memoria. Para imputación simple, por ejemplo, evaluar el uso de regresión estocástica, hot-deck o cold-deck para variables contínuas, análisis discriminante o regresión logística para variables categóricas, etc. Mientras que para los casos de imputación múltiple, se propone la posibilidad de incurrir en mayor cantidad de iteraciones para el algoritmo MICE u ocupar métodos de machine learning distintos, como lo son el k vecino más cercano (k-nearest), simulación bayesiana o redes neuronales (datawig para python), entre otros. También, para el caso del tratamiento de *outliers*, se propone el uso de métodos más eficientes para la detección de datos atípicos, con el fin de profundizar el estudio sobre la significancia que pueden tener las utilidades para este tipo de modelos.

7 Referencias

Abowd, J. M., Crepon, B., & Kramarz, F. (1997). *Moment Estimation With Attrition: An Application To Economic Models.* Journal of the American Statistical Association.

Adams, R. B., & Ferreira, D. (2009). *Women in the boardroom and their impact on governance and performance.* Journal of Financial Economics, 94 (2).

Angrist, J. D., & Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Barnard, J., & Meng, X.-L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research.*

Baulch, B., & Quisumbing, A. (2011). *Testing and adjusting for attrition in household panel data.* Chronic Poverty Research Centre.

Campbell, K., & Mínguez-Vera, A. (2008). Gender Diversity in the Boardroom and Firm Financial Performance. *Journal of Business Ethics, 83(3).*

Carrasco, F. (2022). *Análisis del impacto de la composición de género en la gestión de las empresas en Chile.* Departamento Ingeniería Civil Industrial, Universidad de Concepción.

Chabé-Ferret, S. (2015). *Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes.* Journal of Econometrics 185.1.

Doove, L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis.*

Dorsett, R. (2010). Adjusting for Nonignorable Sample Attrition Using Survey Substitutes Identified by Propensity Score Matching: An Empirical Investigation Using Labour Market Data. *Journal of Official Statistics.*

Faccio, M., Marchica, M.-T., & Mura, R. (2016). CEO gender, corporate risk-taking, and the efficiency of. *Journal of Corporate Finance, 39.*

Fielding, S., Fayers, P. M., McDonald, A., McPherson, G., & Campbell, M. K. (2008). Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes.*

Finley, E. H. (1972). A review of "longitudinal study" in developmental psychology. *Dissertations and Theses.* Portland State University. Department of Psychology.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. A. (1998). *An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics.* Journal of Human Resources, Vol. 33.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. A. (1999). *Sample Attrition in Panel Data: The Role of Selection on Observables.* Annals of Economics and Statistics, GENES.

Gottschalk, S., & Niefert, M. (2013). Gender differences in business success of German start-up firms. *International Journal of Entrepreneurship and Small Business, 18(1).*

Gujarati, D. N. (1978). *Basic Econometrics.*

Hausman, J. A., & Wise, D. A. (1979). *Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment.* Econometrica, Vol. 47.

Heckman, J. J., Ichimura, H., Smith, J. A., & Todd, P. E. (1998). *Characterizing selection bias using experimental data.*

Herreros, P. (2021). *Políticas ambientales voluntarias y sus efectos en la intensidad de uso de energía.* Departamento Ingeniería Civil Industrial, Universidad de Concepción.

Hill, C. A., Biemer, P. P., & Buskirk, T. D. (2020). *Using Machine Learning Models to Predict Attrition in a Survey Panel.* Big Data Meets Survey Science (A Collection of Innovative Methods).

Hirano, K., Imbens, G. W., Ridder, G., & Rubin, D. B. (2001). *Combining Panel Data Sets with Attrition and Refreshment Samples.* Combining Panel Data Sets with Attrition and Refreshment Samples. , 69(6).

Hoon Jr., P. W. (1969). Academic performance as a function of observable study behaviors. *ProQuest Dissertations Publishing.* The University of Nebraska, Lincoln.

Kenney, J. F., & Keeping, E. S. (1951). *Mathematics of statistics. Part 2.*

Lehnen, R. G., & Koch, G. G. (1974). *Analyzing Panel Data With Uncontrolled Attrition.* Public Opinion Quarterly, vol. 38.

Little, R. J., & Rubin, D. B. (1991). *Statistical Analysis with Missing Data.* Journal of Educational Statistics, 16(2).

Nunan, D. (2018). Catalogue of bias: attrition bias. *BMJ Evidence-Based Medicine.*

Orchid, B. (2015). *The Use of Propensity Score Methods for Addressing Attrition in Longitudinal Studies: Practical Guidance and Applications for Evaluating Early Childhood Interventions.* Hyatt Regency Miami.

Parrotta, P., & Smith, N. (2013). *Female-Led Firms: Performance and Risk Attitudes.*

Rosenbaum, P. R., & Rubin, D. B. (1983). *The central role of the propensity score in observational studies for causal effects.*

Rubin, D. B. (1973). *Matching to Remove Bias in Observational Studies.*

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, Volume 63, Issue 3.*

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. *Wiley Series in Probability and Statistics.*

Rubin, D. B. (1997). *Estimating causal effects from large data sets using propensity scores.*

Schafer, J. L. (1997). The Analysis of Incomplete Multivariate Data. *Chapman & Hall.* London.

Szabo, M. (1969). The relationship of intellective, personality and biographical variables to success and it's prediction in an independent study science course at the college level. *ProQuest Dissertations Publishing.* Purdue University.

Tucker, J. (2010). *Selection bias and econometric remedies in accounting and finance research.* Journal of Accounting Literature 29.

Vandecasteele, L., & Debels, A. (2007). *Attrition in Panel Data: The Effectiveness of Weighting.* European Sociological Review, 23(1).

Wulff, J. N., & Ejlskov, L. (2017). Multiple Imputation by Chained Equations in Praxis: Guidelines and Review. *Electronic Journal on Business Research Methods, 15(1).*

Young, R., & Johnson, D. R. (2015). *Handling Missing Values in Longitudinal Panel Data With Multiple Imputation.* J Marriage Fam. 2015 Feb; 77(1).

Zhu, X. (2014). *Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study.* Biostatistics & Data Management, Regeneron Pharmaceuticals, Inc.

Anexo 1: Variables de la base de datos

Variable	Tipo	Descripción
ID	Entero	ID único de la empresa N.I. (se mantiene a través de las versiones de la ELE)
Year	Entero	Año en el cual la correspondiente versión de la ELE fue realizada
Tamaño	Entero	Asignación de número según tamaño de la empresa (1: Grande, 2: Mediana, 3: P1, 4: P2, 5: Micro)
Actividad	Categórico	Asignación de letra según actividad de la empresa
Tasa	Flotante (%)	Tasa anual de interés del crédito
Ventas	Flotante	Ventas en dólares nominales (dic. 2017)
Año_inicio	Entero	Año de fundación de la empresa
Edad	Entero	Edad de la empresa
Holding	Entero	1: si es parte de un grupo económico, 0: en otros casos
SRI	Entero	1: si es una sociedad de responsabilidad ilimitada, 0: en otros casos
Extranjeros	Entero	1: si la empresa tiene dueños extranjeros, 0: en otros casos
P_extranjeros	Flotante (%)	Porcentaje de participación de los dueños extranjeros (en caso de que Extranjeros = 1)
Familiar	Entero	1: si la empresa es un negocio familiar, 0: en otros casos

Dueño	Entero	1: si el CEO de la empresa es el dueño de esta, 0: en otros casos
Propiedad	Entero	1: si el CEO es dueño de algún porcentaje de la empresa, 0: en otros casos
Género	Entero	1: si el CEO de la empresa es hombre, 0: en otros casos
Género_mujer	Entero	1: si el CEO de la empresa es mujer, 0: en otros casos
Exporta	Entero	1: si la empresa realiza exportaciones directas, 0: en otros casos
P_exporta	Flotante (%)	Porcentaje de las ventas atribuidas a las exportaciones directas (en caso de que Exporta = 1)
Trabajadores	Entero	Número de trabajadores
Salarios	Flotante	Salario en dólares nominales (dic. 2017) total pagado a los trabajadores
Externo	Entero	1: si la empresa tiene trabajadores subcontratados, 0: en otros casos
W_externo	Entero	Cantidad de trabajadores subcontratados de la empresa (en caso de que Out = 1)
Deuda	Entero	1: si la empresa tiene deuda, 0: en otros casos
Fomento	Entero	1: si la empresa recibe financiación pública, 0: en otros casos
IyD	Entero	1: si la empresa hace I+D, 0: en otros casos
Inventario	Flotante	Inventario en dólares nominales (dic. 2017)
Impuestos	Flotante	Impuestos pagados en dólares nominales (dic. 2017)

Utilidades	Flotante	Utilidades en dólares nominales (dic. 2017)
Región	Entero	Región en la cual se ubica la empresa
Log_ventas	Flotante	Logaritmo de Ventas
Ventas_pp		Ventas nominales por trabajador (dic. 2017)
Log_ventas_pp	Flotante	Logaritmo de Ventas_pp
Log_inventario	Flotante	Logaritmo de Inventario
Log_impuestos	Flotante	Logaritmo de Impuestos
Log_salarios	Flotante	Logaritmo de Salarios
Log_utilidades	Flotante	Logaritmo de Utilidades
Salario_pp	Flotante	Salario nominal por trabajador (dic. 2017)
Log_salario_pp	Flotante	Logaritmo de Salario_pp
Log_trabajadores	Flotante	Logaritmo de Trabajadores
Y_2009	Entero	1: si la ELE corresponde a la versión 2, 0: en otros casos
Y_2013	Entero	1: si la ELE corresponde a la versión 3, 0: en otros casos
Y_2015	Entero	1: si la ELE corresponde a la versión 4, 0: en otros casos
Y_2017	Entero	1: si la ELE corresponde a la versión 5, 0: en otros casos
Reg_met	Entero	1: si la región en la cual se ubica la empresa es la Metropolitana, 0: en otros casos
Sec_agr	Entero	1: si la actividad de la empresa es la agricultura, ganadería y/o pesca, 0: en otros casos
Sec_min	Entero	1: si la actividad de la empresa es la minería, 0: en otros casos

Sec_man	Entero	1: si la actividad de la empresa es la manufactura, 0: en otros casos
Sec_ele	Entero	1: si la actividad de la empresa es sobre la electricidad, gas y/o agua, 0: en otros casos
Sec_con	Entero	1: si la actividad de la empresa es la construcción, 0: en otros casos
Sec_tra	Entero	1: si la actividad de la empresa es el transporte, 0: en otros casos
Sec_hot	Entero	1: si la actividad de la empresa es el alojamiento y comidas, 0: en otros casos
Sec_inf	Entero	1: si la actividad de la empresa es sobre la información y comunicación, 0: en otros casos
Sec_fin	Entero	1: si la actividad de la empresa es en el sector financiero, 0: en otros casos
Sec_pro	Entero	1: si la actividad de la empresa es de profesionales, científicas y técnicas, 0: en otros casos
Sec_ser	Entero	1: si la actividad de la empresa es de servicio, 0: en otros casos
Sec_art	Entero	1: si la actividad de la empresa es artística, 0: en otros casos
Tam_grande	Entero	1: si la empresa es grande, 0: en otros casos
Tam_med	Entero	1: si la empresa es mediana, 0: en otros casos
Tam_p1	Entero	1: si la empresa es pequeña 1, 0: en otros casos

Tam_p2	Entero	1: si la empresa es pequeña 2, 0: en otros casos
Margen_utilidad	Flotante	Utilidad por ventas

Fuente: elaboración propia

Anexo 2: Regresión Lineal Múltiple dentro de imputación simple

Variable dependiente: Ventas (en log)

OLS Regression Results

Dep. Variable:	logventas	R-squared:	0.861
Model:	OLS	Adj. R-squared:	0.861
Method:	Least Squares	F-statistic:	3.416e+04
Date:	Fri, 01 Jul 2022	Prob (F-statistic):	0.00
Time:	19:42:22	Log-Likelihood:	-69336.
No. Observations:	38592	AIC:	1.387e+05
Df Residuals:	38584	BIC:	1.388e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.4296	0.038	222.153	0.000	8.355	8.504
y2009	7.2025	0.020	356.576	0.000	7.163	7.242
tamano	-0.9405	0.010	-94.001	0.000	-0.960	-0.921
tam_grande	0.8691	0.026	33.214	0.000	0.818	0.920
unlimited	-0.1268	0.037	-3.459	0.001	-0.199	-0.055
owner	-0.1564	0.036	-4.345	0.000	-0.227	-0.086
holding	0.6822	0.020	34.700	0.000	0.644	0.721
logworkers	0.1680	0.003	65.027	0.000	0.163	0.173

Omnibus:	34391.727	Durbin-Watson:	1.841
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3555103.662
Skew:	-3.863	Prob(JB):	0.00
Kurtosis:	49.381	Cond. No.	35.3

Fuente: Google Colab

Variable dependiente: Salarios (en log)

OLS Regression Results

Dep. Variable:	logsalarios	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.796
Method:	Least Squares	F-statistic:	3.002e+04
Date:	Fri, 01 Jul 2022	Prob (F-statistic):	0.00
Time:	19:42:27	Log-Likelihood:	-84161.
No. Observations:	38442	AIC:	1.683e+05
Df Residuals:	38436	BIC:	1.684e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2759	0.054	41.890	0.000	2.169	2.382
logworkers	1.0017	0.004	251.335	0.000	0.994	1.009
tamano	-0.4573	0.015	-31.189	0.000	-0.486	-0.429
tam_grande	-0.8174	0.038	-21.598	0.000	-0.892	-0.743
unlimited	-0.3978	0.054	-7.308	0.000	-0.504	-0.291
owner	-0.3340	0.053	-6.251	0.000	-0.439	-0.229

Omnibus:	16369.602	Durbin-Watson:	1.541
Prob(Omnibus):	0.000	Jarque-Bera (JB):	267220.960
Skew:	-1.623	Prob(JB):	0.00
Kurtosis:	15.502	Cond. No.	35.3

Fuente: Google Colab

Variable dependiente: Inventario (en log)

OLS Regression Results

Dep. Variable:	loginventario	R-squared:	0.313
Model:	OLS	Adj. R-squared:	0.313
Method:	Least Squares	F-statistic:	3293.
Date:	Fri, 01 Jul 2022	Prob (F-statistic):	0.00
Time:	19:42:30	Log-Likelihood:	-1.1671e+05
No. Observations:	36069	AIC:	2.334e+05
Df Residuals:	36063	BIC:	2.335e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.2645	0.165	7.678	0.000	0.942	1.587
tamano	-1.0992	0.043	-25.407	0.000	-1.184	-1.014
tam_grande	2.4666	0.113	21.913	0.000	2.246	2.687
logworkers	0.2628	0.011	23.100	0.000	0.241	0.285
owner	-0.7906	0.091	-8.710	0.000	-0.969	-0.613
y2009	5.1605	0.086	60.030	0.000	4.992	5.329

Omnibus:	4452.583	Durbin-Watson:	1.551
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2513.109
Skew:	-0.508	Prob(JB):	0.00
Kurtosis:	2.200	Cond. No.	31.8

Fuente: Google Colab

Variable dependiente: Impuestos (en log)

OLS Regression Results

Dep. Variable:	logimpuestos	R-squared:	0.252
Model:	OLS	Adj. R-squared:	0.252
Method:	Least Squares	F-statistic:	2455.
Date:	Fri, 01 Jul 2022	Prob (F-statistic):	0.00
Time:	19:42:33	Log-Likelihood:	-1.1184e+05
No. Observations:	36398	AIC:	2.237e+05
Df Residuals:	36392	BIC:	2.238e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5813	0.140	18.450	0.000	2.307	2.856
y2009	3.9362	0.074	53.422	0.000	3.792	4.081
tamano	-1.2040	0.035	-34.489	0.000	-1.272	-1.136
holding	0.3629	0.072	5.006	0.000	0.221	0.505
tam_grande	0.8255	0.096	8.571	0.000	0.637	1.014
logworkers	0.1483	0.009	15.682	0.000	0.130	0.167

Omnibus:	3374.977	Durbin-Watson:	1.889
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4301.369
Skew:	-0.832	Prob(JB):	0.00
Kurtosis:	2.740	Cond. No.	31.7

Fuente: Google Colab

Variable dependiente: Utilidades (en log)

OLS Regression Results

Dep. Variable:	logutilidad	R-squared (uncentered):	0.886
Model:	OLS	Adj. R-squared (uncentered):	0.886
Method:	Least Squares	F-statistic:	4.060e+04
Date:	Fri, 01 Jul 2022	Prob (F-statistic):	0.00
Time:	19:42:38	Log-Likelihood:	-72413.
No. Observations:	31476	AIC:	1.448e+05
Df Residuals:	31470	BIC:	1.449e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
logventas_reg	0.7574	0.006	130.387	0.000	0.746	0.769
y2009	2.0374	0.060	33.896	0.000	1.920	2.155
tamano	-0.3308	0.011	-31.033	0.000	-0.352	-0.310
tam_grande	-0.4686	0.051	-9.103	0.000	-0.569	-0.368
holding	0.7200	0.038	19.133	0.000	0.646	0.794
unlimited	-0.7906	0.039	-20.191	0.000	-0.867	-0.714

Omnibus:	9687.157	Durbin-Watson:	1.583
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39150.167
Skew:	-1.484	Prob(JB):	0.00
Kurtosis:	7.588	Cond. No.	49.7

Fuente: Google Colab

Anexo 3: Regresión Lineal Múltiple dentro de imputación simple (variante para utilidades)

Dado que para la base original, el menor valor de las utilidades es – 244.369.760, se propone:

$$\text{Utilidades'} = \text{Utilidades} + 244.369.761$$

De esta manera, todo valor de esta transformada de utilidades, será positivo, siendo posible la aplicación global de la función logarítmica y exponencial.

Variable dependiente: Utilidades' (en log)

OLS Regression Results

Dep. Variable:	logutilidadtrans	R-squared (uncentered):	0.983
Model:	OLS	Adj. R-squared (uncentered):	0.983
Method:	Least Squares	F-statistic:	3.716e+05
Date:	Mon, 15 Aug 2022	Prob (F-statistic):	0.00
Time:	22:49:51	Log-Likelihood:	-90418.
No. Observations:	38591	AIC:	1.808e+05
Df Residuals:	38585	BIC:	1.809e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
logventas	1.3606	0.005	292.455	0.000	1.351	1.370
y2009	-8.6590	0.051	-168.979	0.000	-8.759	-8.559
tamano	3.2136	0.009	355.023	0.000	3.196	3.231
tam_grande	2.9041	0.043	67.090	0.000	2.819	2.989
holding	-0.2521	0.034	-7.369	0.000	-0.319	-0.185
unlimited	0.0532	0.037	1.449	0.147	-0.019	0.125

Omnibus:	20635.859	Durbin-Watson:	1.764
Prob(Omnibus):	0.000	Jarque-Bera (JB):	709247.421
Skew:	1.963	Prob(JB):	0.00
Kurtosis:	23.632	Cond. No.	44.1

Fuente: Google Colab

Comparando esta nueva estimación con la base original, se tiene que:

Variables	Media antes	Media después	Dif %	Varianza antes	Varianza después	Dif %
Utilidades	$1{,}9 \times 10^{6}$	$1{,}9 \times 10^{6}$	-1,3%	$5{,}4 \times 10^{16}$	$5{,}3 \times 10^{16}$	-1,3%

Anexo 4: Detección de outliers bajo criterio 1,5 IQR

Para la detección de outliers del margen de utilidad y de la tasa de crecimiento de la utilidad, se utilizó el criterio 1,5 IQR, el cual establece límites de aceptación de los datos:

Rango intercuartílico (IQR): Q3 – Q1

Límite inferior (mínimo de la nueva muestra): (Q1 - 1.5 * IQR)

Límite superior (máximo de la nueva muestra): (Q3 + 1.5 * IQR)

Con los nuevos límites establecidos, y, en base a la base original, se tiene que:

Variable	Frecuencia	Media	Desv. Std.	Mín.	Max
Margen de utilidad	74.779	6,969	10,360	-21,436	36,868
Tasa de crecimiento de la utilidad	46.335	-0,536	1,176	-4,185	4,304

Anexo 5: MCO para variables dependientes

Variable dependiente: Productividad laboral de los empleados (en log)

Source	SS	df	MS		
Model	73696.1868	9	8188.4652		
Residual	1533622.16	89044	17.2231948		
Total	1607318.34	89053	18.0490084		

Number of obs	=	89054			
F(9, 89044)	=	475.43			
Prob > F	=	0.0000			
R-squared	=	0.0459			
Adj R-squared	=	0.0458			
Root MSE	=	4.1501			

logventaspp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.317997	.0327549	-9.71	0.000	-.3821964	-.2537976
age	-.0589168	.001244	-47.36	0.000	-.061355	-.0564785
foreign	-.3395722	.0392618	-8.65	0.000	-.4165251	-.2626194
iyd	-.866117	.0332198	-26.07	0.000	-.9312275	-.8010064
export	-.0998364	.0354485	-2.82	0.005	-.1693151	-.0303577
debt	-.4993299	.0321468	-15.53	0.000	-.5623374	-.4363224
holding	.2362612	.0331107	7.14	0.000	.1713645	.3011579
out	-.3880891	.0332953	-11.66	0.000	-.4533475	-.3228307
tamano	.0706567	.0111182	6.36	0.000	.0488651	.0924482
_cons	5.913222	.0518822	113.97	0.000	5.811534	6.014911

Fuente: stata

Variable dependiente: Número de trabajadores (en log)

Source	SS	df	MS		
Model	427157.896	9	47461.9884	Number of obs	= 89054
Residual	725194.596	89044	8.14422753	F(9, 89044)	= 5827.68
				Prob > F	= 0.0000
Total	1152352.49	89053	12.9400749	R-squared	= 0.3707
				Adj R-squared	= 0.3706
				Root MSE	= 2.8538

logworkers	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.0391981	.022524	-1.74	0.082	-.0833448	.0049487
age	.0354145	.0008554	41.40	0.000	.0337378	.0370912
foreign	.2466054	.0269984	9.13	0.000	.1936887	.2995221
iyd	.5657729	.0228437	24.77	0.000	.5209995	.6105462
export	.2464797	.0243762	10.11	0.000	.1987026	.2942568
debt	.1048451	.0221058	4.74	0.000	.0615179	.1481722
holding	.1335708	.0227686	5.87	0.000	.0889445	.1781971
out	.5126834	.0228955	22.39	0.000	.4678084	.5575584
tamano	-1.409306	.0076454	-184.33	0.000	-1.424291	-1.394321
_cons	6.369365	.0356769	178.53	0.000	6.299438	6.439291

Fuente: stata

Variable dependiente: Ventas (en log)

Source	SS	df	MS		
Model	310629.821	9	34514.4246	Number of obs	= 89054
Residual	759142.728	89044	8.52547873	F(9, 89044)	= 4048.39
				Prob > F	= 0.0000
Total	1069772.55	89053	12.0127626	R-squared	= 0.2904
				Adj R-squared	= 0.2903
				Root MSE	= 2.9198

logventas	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.3571951	.0230451	-15.50	0.000	-.4023633	-.3120268
age	-.0235023	.0008752	-26.85	0.000	-.0252177	-.0217868
foreign	-.0929668	.0276231	-3.37	0.001	-.1471079	-.0388257
iyd	-.3003441	.0233722	-12.85	0.000	-.3461534	-.2545347
export	.1466433	.0249402	5.88	0.000	.0977607	.1955259
debt	-.3944849	.0226173	-17.44	0.000	-.4388146	-.3501552
holding	.369832	.0232954	15.88	0.000	.3241731	.4154908
out	.1245943	.0234253	5.32	0.000	.078681	.1705077
tamano	-1.338649	.0078223	-171.13	0.000	-1.353981	-1.323318
_cons	12.28259	.0365024	336.49	0.000	12.21104	12.35413

Fuente: stata

Variable dependiente: Margen de utilidad

Source	SS	df	MS		
Model	1764950.29	9	196105.587	Number of obs	= 88520
Residual	1.4288e+09	88510	16142.6354	F(9, 88510)	= 12.15
				Prob > F	= 0.0000
				R-squared	= 0.0012
				Adj R-squared	= 0.0011
Total	1.4305e+09	88519	16160.9328	Root MSE	= 127.05

margen	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.9400378	1.006505	-0.93	0.350	-2.912778	1.032702
age	.038316	.0382677	1.00	0.317	-.0366884	.1133203
foreign	-3.17633	1.206891	-2.63	0.008	-5.541827	-.8108342
iyd	1.100533	1.019994	1.08	0.281	-.8986457	3.099713
export	-2.119626	1.088217	-1.95	0.051	-4.252521	.0132695
debt	-1.826123	.9883694	-1.85	0.065	-3.763318	.1110723
holding	-.4369439	1.017649	-0.43	0.668	-2.431526	1.557638
out	-4.355259	1.022615	-4.26	0.000	-6.359574	-2.350944
tamano	-3.059025	.3426028	-8.93	0.000	-3.730524	-2.387527
_cons	14.02335	1.595754	8.79	0.000	10.89569	17.15102

Fuente: stata

Variable dependiente: Tasa de crecimiento de la utilidad

Source	SS	df	MS		
Model	580.77192	9	64.5302134	Number of obs =	53328
Residual	1838712.6	53318	34.4857759	F(9, 53318) =	1.87
				Prob > F =	0.0513
Total	1839293.37	53327	34.4908465	R-squared =	0.0003
				Adj R-squared =	0.0001
				Root MSE =	5.8725

g_utilidades	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	.0284273	.0581946	0.49	0.625	-.0856345	.1424891
age	-.0017185	.0023372	-0.74	0.462	-.0062994	.0028623
foreign	.0247814	.0649271	0.38	0.703	-.1024763	.1520391
iyd	-.0329329	.059215	-0.56	0.578	-.1489947	.0831289
export	-.1171297	.0595754	-1.97	0.049	-.2338979	-.0003614
debt	-.0422158	.0582846	-0.72	0.469	-.1564541	.0720225
holding	-.1074698	.0567104	-1.90	0.058	-.2186225	.003683
out	.0453223	.056666	0.80	0.424	-.0657435	.1563881
tamano	-.0687688	.0220395	-3.12	0.002	-.1119663	-.0255712
_cons	.1042989	.0958682	1.09	0.277	-.0836037	.2922015

Fuente: stata

Anexo 6: MCA para variables dependientes

Variable dependiente: Productividad laboral de los empleados (en log)

Source	SS	df	MS		
Model	73696.1868	9	8188.4652		
Residual	1533622.16	89044	17.2231948		
Total	1607318.34	89053	18.0490084		

Number of obs	=	89054			
F(9, 89044)	=	475.43			
Prob > F	=	0.0000			
R-squared	=	0.0459			
Adj R-squared	=	0.0458			
Root MSE	=	4.1501			

logventaspp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.317997	.0327549	-9.71	0.000	-.3821964	-.2537976
age	-.0589168	.001244	-47.36	0.000	-.061355	-.0564785
foreign	-.3395722	.0392618	-8.65	0.000	-.4165251	-.2626194
iyd	-.866117	.0332198	-26.07	0.000	-.9312275	-.8010064
export	-.0998364	.0354485	-2.82	0.005	-.1693151	-.0303577
debt	-.4993299	.0321468	-15.53	0.000	-.5623374	-.4363224
holding	.2362612	.0331107	7.14	0.000	.1713645	.3011579
out	-.3880891	.0332953	-11.66	0.000	-.4533475	-.3228307
tamano	.0706567	.0111182	6.36	0.000	.0488651	.0924482
_cons	5.913222	.0518822	113.97	0.000	5.811534	6.014911

Fuente: stata

Variable dependiente: Número de trabajadores (en log)

Source	SS	df	MS	Number of obs =	89054
				F(9, 89044) =	5827.68
Model	427157.896	9	47461.9884	Prob > F =	0.0000
Residual	725194.596	89044	8.14422753	R-squared =	0.3707
				Adj R-squared =	0.3706
Total	1152352.49	89053	12.9400749	Root MSE =	2.8538

logworkers	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.0391981	.022524	-1.74	0.082	-.0833448	.0049487
age	.0354145	.0008554	41.40	0.000	.0337378	.0370912
foreign	.2466054	.0269984	9.13	0.000	.1936887	.2995221
iyd	.5657729	.0228437	24.77	0.000	.5209995	.6105462
export	.2464797	.0243762	10.11	0.000	.1987026	.2942568
debt	.1048451	.0221058	4.74	0.000	.0615179	.1481722
holding	.1335708	.0227686	5.87	0.000	.0889445	.1781971
out	.5126834	.0228955	22.39	0.000	.4678084	.5575584
tamano	-1.409306	.0076454	-184.33	0.000	-1.424291	-1.394321
_cons	6.369365	.0356769	178.53	0.000	6.299438	6.439291

Fuente: stata

Variable dependiente: Ventas (en log)

Source	SS	df	MS		
Model	310629.821	9	34514.4246	Number of obs	= 89054
Residual	759142.728	89044	8.52547873	F(9, 89044)	= 4048.39
				Prob > F	= 0.0000
Total	1069772.55	89053	12.0127626	R-squared	= 0.2904
				Adj R-squared	= 0.2903
				Root MSE	= 2.9198

logventas	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.3571951	.0230451	-15.50	0.000	-.4023633	-.3120268
age	-.0235023	.0008752	-26.85	0.000	-.0252177	-.0217868
foreign	-.0929668	.0276231	-3.37	0.001	-.1471079	-.0388257
iyd	-.3003441	.0233722	-12.85	0.000	-.3461534	-.2545347
export	.1466433	.0249402	5.88	0.000	.0977607	.1955259
debt	-.3944849	.0226173	-17.44	0.000	-.4388146	-.3501552
holding	.369832	.0232954	15.88	0.000	.3241731	.4154908
out	.1245943	.0234253	5.32	0.000	.078681	.1705077
tamano	-1.338649	.0078223	-171.13	0.000	-1.353981	-1.323318
_cons	12.28259	.0365024	336.49	0.000	12.21104	12.35413

Fuente: stata

Variable dependiente: Margen de utilidad

Source	SS	df	MS		
Model	1764950.29	9	196105.587	Number of obs	= 88520
Residual	1.4288e+09	88510	16142.6354	F(9, 88510)	= 12.15
				Prob > F	= 0.0000
				R-squared	= 0.0012
				Adj R-squared	= 0.0011
Total	1.4305e+09	88519	16160.9328	Root MSE	= 127.05

margen	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.9400378	1.006505	-0.93	0.350	-2.912778	1.032702
age	.038316	.0382677	1.00	0.317	-.0366884	.1133203
foreign	-3.17633	1.206891	-2.63	0.008	-5.541827	-.8108342
iyd	1.100533	1.019994	1.08	0.281	-.8986457	3.099713
export	-2.119626	1.088217	-1.95	0.051	-4.252521	.0132695
debt	-1.826123	.9883694	-1.85	0.065	-3.763318	.1110723
holding	-.4369439	1.017649	-0.43	0.668	-2.431526	1.557638
out	-4.355259	1.022615	-4.26	0.000	-6.359574	-2.350944
tamano	-3.059025	.3426028	-8.93	0.000	-3.730524	-2.387527
_cons	14.02335	1.595754	8.79	0.000	10.89569	17.15102

Fuente: stata

Variable dependiente: Tasa de crecimiento de la utilidad

Source	SS	df	MS		
Model	580.77192	9	64.5302134	Number of obs =	53328
Residual	1838712.6	53318	34.4857759	F(9, 53318) =	1.87
				Prob > F =	0.0513
Total	1839293.37	53327	34.4908465	R-squared =	0.0003
				Adj R-squared =	0.0001
				Root MSE =	5.8725

g_utilidades	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	.0284273	.0581946	0.49	0.625	-.0856345	.1424891
age	-.0017185	.0023372	-0.74	0.462	-.0062994	.0028623
foreign	.0247814	.0649271	0.38	0.703	-.1024763	.1520391
iyd	-.0329329	.059215	-0.56	0.578	-.1489947	.0831289
export	-.1171297	.0595754	-1.97	0.049	-.2338979	-.0003614
debt	-.0422158	.0582846	-0.72	0.469	-.1564541	.0720225
holding	-.1074698	.0567104	-1.90	0.058	-.2186225	.003683
out	.0453223	.056666	0.80	0.424	-.0657435	.1563881
tamano	-.0687688	.0220395	-3.12	0.002	-.1119663	-.0255712
_cons	.1042989	.0958682	1.09	0.277	-.0836037	.2922015

Fuente: stata

Anexo 7: Efectos Fijos para variables dependientes

Variable dependiente: Productividad laboral de los empleados (en log)

```
Fixed-effects (within) regression               Number of obs      =     89054
Group variable: ID                              Number of groups   =     24130

R-sq:  within  = 0.1531                         Obs per group: min =         1
       between = 0.0000                                        avg =       3.7
       overall = 0.0402                                        max =         5

                                                F(9,64915)         =   1303.95
corr(u_i, Xb)  = -0.6010                        Prob > F           =    0.0000
```

logventaspp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.3558949	.0487718	-7.30	0.000	-.4514876	-.2603022
age	-.206889	.0023094	-89.59	0.000	-.2114155	-.2023626
foreign	-.5013089	.0470333	-10.66	0.000	-.5934942	-.4091237
iyd	-1.198617	.0413696	-28.97	0.000	-1.279702	-1.117533
export	-.3143026	.0420841	-7.47	0.000	-.3967874	-.2318178
debt	-.7755375	.0372763	-20.81	0.000	-.8485989	-.702476
holding	-.2489345	.0461804	-5.39	0.000	-.3394482	-.1584209
out	-.7249083	.0388988	-18.64	0.000	-.80115	-.6486666
tamano	.0064538	.0150875	0.43	0.669	-.0231178	.0360253
_cons	9.146431	.0711716	128.51	0.000	9.006935	9.285928
sigma_u	3.1586951					
sigma_e	4.0718697					
rho	.3756891	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(24129, 64915) =     1.14        Prob > F = 0.0000
```

Fuente: stata

Variable dependiente: Número de trabajadores (en log)

```
Fixed-effects (within) regression               Number of obs      =     89054
Group variable: ID                              Number of groups   =     24130

R-sq:  within  = 0.3342                         Obs per group: min =         1
       between = 0.3423                                        avg =       3.7
       overall = 0.3441                                        max =         5

                                                F(9,64915)         =   3620.92
corr(u_i, Xb)  = -0.0786                        Prob > F           =    0.0000
```

logworkers	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	.1447592	.0304564	4.75	0.000	.0850647	.2044537
age	.0851784	.0014422	59.06	0.000	.0823517	.088005
foreign	.3689824	.0293708	12.56	0.000	.3114157	.4265491
iyd	.4031072	.025834	15.60	0.000	.3524727	.4537418
export	.3672226	.0262801	13.97	0.000	.3157135	.4187316
debt	-.0906359	.0232778	-3.89	0.000	-.1362604	-.0450114
holding	.3229699	.0288382	11.20	0.000	.2664471	.3794928
out	.6014142	.024291	24.76	0.000	.5538038	.6490247
tamano	-1.257435	.0094217	-133.46	0.000	-1.275902	-1.238969
_cons	5.183999	.0444444	116.64	0.000	5.096888	5.27111
sigma_u	2.2567957					
sigma_e	2.5427505					
rho	.44063118	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(24129, 64915) =     1.96        Prob > F = 0.0000
```

Fuente: stata

Variable dependiente: Ventas (en log)

```
Fixed-effects (within) regression               Number of obs      =     89054
Group variable: ID                              Number of groups   =     24130

R-sq:  within  = 0.2129                         Obs per group: min =         1
       between = 0.1991                                        avg =       3.7
       overall = 0.1851                                        max =         5

                                                F(9,64915)         =   1950.65
corr(u_i, Xb)  = -0.2864                        Prob > F           =    0.0000
```

logventas	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-.2111357	.0354145	-5.96	0.000	-.2805482	-.1417233
age	-.1217107	.0016769	-72.58	0.000	-.1249974	-.1184239
foreign	-.1323266	.0341521	-3.87	0.000	-.1992648	-.0653883
iyd	-.7955099	.0300396	-26.48	0.000	-.8543875	-.7366323
export	.0529199	.0305584	1.73	0.083	-.0069745	.1128144
debt	-.8661733	.0270673	-32.00	0.000	-.9192253	-.8131214
holding	.0740354	.0335329	2.21	0.027	.008311	.1397598
out	-.1234941	.0282455	-4.37	0.000	-.1788552	-.0681329
tamano	-1.250981	.0109554	-114.19	0.000	-1.272454	-1.229509
_cons	14.33043	.0516796	277.29	0.000	14.22914	14.43172
sigma_u	2.0763844					
sigma_e	2.9566945					
rho	.33028693	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(24129, 64915) =     0.91        Prob > F = 1.0000
```

Fuente: stata

Variable dependiente: Margen de utilidad

```
Fixed-effects (within) regression               Number of obs      =     88520
Group variable: ID                              Number of groups   =     24104

R-sq:  within  = 0.0008                         Obs per group: min =         1
       between = 0.0005                                        avg =       3.7
       overall = 0.0006                                        max =         5

                                                F(9,64407)         =      5.41
corr(u_i, Xb)  = -0.0159                        Prob > F           =    0.0000
```

margen	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	-1.298938	1.469296	-0.88	0.377	-4.17876	1.580884
age	-.2204135	.0696191	-3.17	0.002	-.3568671	-.0839599
foreign	-1.301039	1.416542	-0.92	0.358	-4.077462	1.475384
iyd	1.130926	1.243633	0.91	0.363	-1.306596	3.568447
export	-.3866107	1.265589	-0.31	0.760	-2.867166	2.093944
debt	-1.698428	1.122033	-1.51	0.130	-3.897615	.5007583
holding	-2.885757	1.390051	-2.08	0.038	-5.610258	-.1612552
out	-4.138323	1.170072	-3.54	0.000	-6.431665	-1.844981
tamano	-2.32682	.4552656	-5.11	0.000	-3.219141	-1.434499
_cons	16.61835	2.144937	7.75	0.000	12.41427	20.82243
sigma_u	83.358638					
sigma_e	121.95299					
rho	.31843638	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(24103, 64407) =     1.31        Prob > F = 0.0000
```

Fuente: stata

Variable dependiente: Tasa de crecimiento de la utilidad

```
Fixed-effects (within) regression               Number of obs      =     53328
Group variable: ID                              Number of groups   =     21724

R-sq:  within  = 0.0019                         Obs per group: min =         1
       between = 0.0001                                        avg =       2.5
       overall = 0.0001                                        max =         4

                                                F(9,31595)         =      6.80
corr(u_i, Xb)  = -0.0907                        Prob > F           =    0.0000
```

g_utilidades	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
women	.0053742	.1035548	0.05	0.959	-.1975973	.2083456
age	.0191939	.0059131	3.25	0.001	.007604	.0307838
foreign	-.0248225	.088423	-0.28	0.779	-.1981351	.1484901
iyd	-.0986068	.092028	-1.07	0.284	-.2789852	.0817717
export	-.211749	.0807391	-2.62	0.009	-.3700007	-.0534972
debt	.225784	.07773	2.90	0.004	.0734302	.3781378
holding	.06282	.0972431	0.65	0.518	-.1277801	.2534202
out	-.0046195	.0763979	-0.06	0.952	-.1543623	.1451233
tamano	-.1947609	.0389549	-5.00	0.000	-.271114	-.1184078
_cons	-.1975355	.1688316	-1.17	0.242	-.5284521	.1333811
sigma_u	3.8973977					
sigma_e	5.9697657					
rho	.2988464	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(21723, 31595) =     0.92        Prob > F = 1.0000
```

Fuente: stata

Anexo 8: Matching DID para variables dependientes

Variable dependiente: DIFF de productividad laboral de los empleados (en log)

Probit regression — Number of obs = 60586; LR chi2(8) = 45.96; Prob > chi2 = 0.0000; Log likelihood = -20917.711; Pseudo R2 = 0.0011

TRATAMIENTO	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0002045	.0006192	0.33	0.741	-.001009	.001418
foreign	.0511276	.0172193	2.97	0.003	.0173784	.0848769
iyd	.001676	.0162035	0.10	0.918	-.0300823	.0334342
export	.0373764	.0158156	2.36	0.018	.0063783	.0683744
debt	.0504972	.015459	3.27	0.001	.0201981	.0807964
holding	.0150579	.0152697	0.99	0.324	-.0148702	.044986
out	.0158379	.0151319	1.05	0.295	-.01382	.0454958
tamano	.0288179	.0057531	5.01	0.000	.017542	.0400937
_cons	-1.36094	.0251068	-54.21	0.000	-1.410148	-1.311731

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
DIFF_logventaspp	Unmatched	.070556312	-1.17945247	1.25000878	.078023707	16.02
	ATT	.070556312	-.559842418	.630398731	.184185788	3.42
	ATU	-1.17945247	.08814527	1.26759774	.	.
	ATE			1.19777361	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	53,947	53,947
Treated	6,639	6,639
Total	60,586	60,586

Fuente: stata

Variable dependiente: DIFF de número de trabajadores (en log)

Probit regression — Number of obs = 60586
LR chi2(8) = 45.96
Prob > chi2 = 0.0000
Log likelihood = -20917.711 — Pseudo R2 = 0.0011

TRATAMIENTO	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0002045	.0006192	0.33	0.741	-.001009	.001418
foreign	.0511276	.0172193	2.97	0.003	.0173784	.0848769
iyd	.001676	.0162035	0.10	0.918	-.0300823	.0334342
export	.0373764	.0158156	2.36	0.018	.0063783	.0683744
debt	.0504972	.015459	3.27	0.001	.0201981	.0807964
holding	.0150579	.0152697	0.99	0.324	-.0148702	.044986
out	.0158379	.0151319	1.05	0.295	-.01382	.0454958
tamano	.0288179	.0057531	5.01	0.000	.017542	.0400937
_cons	-1.36094	.0251068	-54.21	0.000	-1.410148	-1.311731

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
DIFF_logworkers	Unmatched	.547828024	.943408269	-.395580245	.046178132	-8.57
	ATT	.547828024	.928166481	-.380338457	.114410858	-3.32
	ATU	.943408269	.718968452	-.224439816	.	.
	ATE			-.241523154	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	53,947	53,947
Treated	6,639	6,639
Total	60,586	60,586

Fuente: stata

Variable dependiente: DIFF de ventas (en log)

Probit regression — Number of obs = 60586
LR chi2(8) = 45.96
Prob > chi2 = 0.0000
Log likelihood = -20917.711 — Pseudo R2 = 0.0011

TRATAMIENTO	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0002045	.0006192	0.33	0.741	-.001009	.001418
foreign	.0511276	.0172193	2.97	0.003	.0173784	.0848769
iyd	.001676	.0162035	0.10	0.918	-.0300823	.0334342
export	.0373764	.0158156	2.36	0.018	.0063783	.0683744
debt	.0504972	.015459	3.27	0.001	.0201981	.0807964
holding	.0150579	.0152697	0.99	0.324	-.0148702	.044986
out	.0158379	.0151319	1.05	0.295	-.01382	.0454958
tamano	.0288179	.0057531	5.01	0.000	.017542	.0400937
_cons	-1.36094	.0251068	-54.21	0.000	-1.410148	-1.311731

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
DIFF_logventas	Unmatched	.618384336	-.236044197	.854428533	.064162331	13.32
	ATT	.618384336	.368324063	.250060274	.156816097	1.59
	ATU	-.236044197	.807113722	1.04315792	.	.
	ATE			.956250461	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	53,947	53,947
Treated	6,639	6,639
Total	60,586	60,586

Fuente: stata

Variable dependiente: DIFF de margen de utilidad

Probit regression — Number of obs = 59973
LR chi2(8) = 42.47
Prob > chi2 = 0.0000
Log likelihood = -20632.415 — Pseudo R2 = 0.0010

TRATAMIENTO	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0000915	.0006264	0.15	0.884	-.0011361	.0013192
foreign	.0510235	.0173541	2.94	0.003	.0170101	.0850369
iyd	-.0015042	.0163223	-0.09	0.927	-.0334954	.030487
export	.0382413	.0159146	2.40	0.016	.0070492	.0694334
debt	.0527872	.015588	3.39	0.001	.0222354	.0833391
holding	.0114809	.0153848	0.75	0.456	-.0186727	.0416346
out	.0143645	.0152372	0.94	0.346	-.0154999	.0442289
tamano	.0266342	.005839	4.56	0.000	.0151899	.0380785
_cons	-1.356402	.0253445	-53.52	0.000	-1.406076	-1.306727

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
DIFF_margen	Unmatched	-1.01003626	-1.05372163	.043685369	2.20580782	0.02
	ATT	-1.01003626	-7.93390235	6.92386609	6.15486073	1.12
	ATU	-1.05372163	8.71943514	9.77315678	.	.
	ATE			9.46263431	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	53,437	53,437
Treated	6,536	6,536
Total	59,973	59,973

Fuente: stata

Variable dependiente: DIFF de tasa de crecimiento de la utilidad

Probit regression — Number of obs = 27852
LR chi2(8) = 26.60
Prob > chi2 = 0.0008
Log likelihood = -10139.877 — Pseudo R2 = 0.0013

TRATAMIENTO	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0007833	.000876	0.89	0.371	-.0009337	.0025003
foreign	.0499631	.0241652	2.07	0.039	.0026003	.097326
iyd	-.0464605	.0222038	-2.09	0.036	-.0899791	-.0029419
export	.0587482	.0225802	2.60	0.009	.0144919	.1030045
debt	.0564488	.0234334	2.41	0.016	.0105201	.1023775
holding	.0253273	.0213462	1.19	0.235	-.0165106	.0671651
out	.0250863	.021894	1.15	0.252	-.0178251	.0679978
tamano	-.0017623	.0086371	-0.20	0.838	-.0186907	.0151661
_cons	-1.263867	.0380082	-33.25	0.000	-1.338361	-1.189372

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
DIFF_g_utilida~s	Unmatched	.058887528	.097979039	-.039091511	.165481728	-0.24
	ATT	.058887528	-.2614644	.320351928	.367449541	0.87
	ATU	.097979039	-.082723189	-.180702227	.	.
	ATE			-.121173712	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	24,543	24,543
Treated	3,309	3,309
Total	27,852	27,852

Fuente: stata

UNIVERSIDAD DE CONCEPCIÓN – FACULTAD DE INGENIERÍA

RESUMEN DE MEMORIA DE TÍTULO

Departamento	**:** Departamento de Ingeniería Industrial
Carrera	**:** Ingeniería Civil Industrial
Nombre del memorista	**:** Cristian Eduardo Bustos Bello
Título de la memoria	**:** Impacto de la composición de género en el desempeño de empresas chilenas: Uso de métodos correctivos de atrición
Fecha de la presentación oral	**:** 18 de agosto del 2022
Profesor Guía	**:** Marcela Verónica Parada Contzen
Profesor Revisor	**:** Hernaldo Del Carmen Reinoso Alarcón
Concepto	**:**
Calificación	**:**

Resumen

La atrición se define como la pérdida de registros a lo largo del tiempo en estudios longitudinales. Carrasco (2022), presenta una primera aproximación sobre el desempeño de empresas chilenas según la composición de género, basándose en los resultados de la Encuesta Longitudinal de Empresas. Sin embargo, esta base de datos presenta altos grados de atrición. En la presente memoria de título se plantea una extensión de este trabajo, usando métodos correctivos para la atrición.

En la literatura, se presentan diversas soluciones a esta problemática, las cuales varían desde el uso del método de máxima verosimilitud, la inyección de registros artificiales, la aplicación de modelos de machine learning, entre otros.

Para subsanar este problema, se utilizaron los métodos de imputación simple (moda, media, regresión lineal y por arrastre de la última observación) e imputación múltiple (ecuaciones concatenadas y Random Forest) para solucionar el problema de la atrición.

Los resultados presentan que ambas imputaciones no afectaron significativamente en la media, moda y composición de la base. Además, la presencia de mujeres en gerencia general tiene un impacto estadísticamente significativo y negativo sobre las ventas y la productividad. Mientras que, para el número de trabajadores presentes, no hay resultados estadísticamente significativos.