ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

ESTUDIO DE REGRESIÓN ENTRE EL RENDIMIENTO POSTCOSECHA DEL CACAO ECUATORIANO, LAS CARACTERÍSTICAS DE LOS PRODUCTORES Y SUS MÉTODOS DE CULTIVO

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

JHON ERICK BARRERA PÉREZ
johnbarrerac2@yahoo.es

DIRECTORA: DRA. ADRIANA UQUILLAS ANDRADE
adriana.uquillas@epn.edu.ec

QUITO, OCTUBRE 2018

DECLARACIÓN

Yo, JHON ERICK BARRERA PÉREZ, declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y la normativa institucional vigente.

Jhon Erick Barrera Pérez

CERTIFICACIÓN

Certifico que el siguiente trabajo fue desarrollado por JHON ERICK BARRERA PÉREZ, bajo mi supervisión.

Dra. Adriana Uquillas Andrade
Directora del Proyecto

AGRADECIMIENTOS

A mi familia, y especialmente a mis padres, pues con su apoyo constante y palabras de aliento me han fortalecido y acompañado a lo largo de mi carrera.
A la Dra. Adriana Uquillas, por su paciencia y sus útiles consejos que fueron de invaluable ayuda durante el desarrollo de este trabajo.
Al Ing. Marco Güilcapi y al Ing. Víctor Lema, del Ministerio de Agricultura del Ecuador, quienes resolvieron mis dudas e inquietudes acerca del cacao, su producción y su importancia en la economía del país; tema en el cual mis conocimientos eran escasos, por no decir inexistentes.
Finalmente, a todas las personas que contribuyeron de alguna forma a que llegara hasta este punto de mi carrera, sean profesores, amigos o compañeros. Gracias de todo corazón.

DEDICATORIA

A mis alumnos,
por enseñarme más
de lo que yo les enseñé a ellos.

Índice general

Índice de figuras

Índice de tablas

Resumen

El cacao es uno de los principales rubros de exportación del Ecuador a nivel mundial tanto en cantidad como en calidad. No obstante, en el país no se han llevado a cabo las suficientes iniciativas de estudio de su producción ni de la influencia de diversos factores en su productividad, a pesar de que los organismos gubernamentales han recolectado información precisa que podría ser usada para conocer más sobre este producto y mejorar su productividad. En este trabajo se proponen dos modelos de regresión, a partir de datos tomados por el Ministerio de Agricultura, Ganadería, Acuacultura y Pesca del Ecuador (MAGAP) en el año 2016, para estimar la transformación de masa tras los procesos postcosecha del cacao (fermentación y secado) dependiendo de las condiciones socioeconómicas del agricultor, su ubicación geográfica y sus métodos agrícolas. Los modelos difieren en la presencia de fincas de rendimiento atípicamente alto y encajan con la experiencia en el campo de los técnicos del MAGAP. El análisis de los coeficientes del modelo muestra que algunas provincias y cantones presenten rendimientos muy superiores para el tratamiento postcosecha del cacao, así como demuestra, *ceteris paribus,* la menor pérdida de masa en fermentación y secado de la variedad de cacao Nacional Fino de Aroma. El modelo prueba también el efecto diferenciado de la región geográfica de cultivo en el rendimiento postcosecha de la variedad de cacao CCN-51 y la influencia positiva de la capacitación cultural y la mecanización de los procesos de fermentación y secado. La información proporcionada por este trabajo permite afinar las políticas gubernamentales de apoyo al sector cacaotero e incrementar la comprensión del cultivo de cacao y su desarrollo en el país.

Palabras clave: rendimiento postcosecha del cacao, modelos de regresión, fermentación de cacao, secado de cacao, clusterización geográfica.

Abstract

Cocoa is one of the main export items of Ecuador worldwide both in quantity and quality. However, the country has not carried out sufficient initiatives to study its production or the influence of various factors on its productivity, despite the fact that government agencies have collected accurate information that could be used to learn more about this product and improve its productivity. In this document, two regression models are proposed, based on data taken by the Agriculture's Ministry of Ecuador (MAGAP) in 2016, to estimate mass transformation after cocoa post-harvest processes (fermentation and drying) depending on the farmers' socio-economic conditions, geographical location and agricultural methods. The models differ in the presence of uncharacteristically high yield farms and fit with the field experience of MAGAP technicians. The analysis of the coefficients of the models suggests that some provinces and cantons have much higher yields for the post-harvest treatment of cocoa, as well as demonstrate, *ceteris paribus,* the lowest mass loss in fermentation and drying of the Nacional Fino de Aroma variety of cocoa. The model also demonstrates the differentiated effect of the geographic region of cultivation on the postharvest yield of the CCN-51 cocoa variety and the positive influence of the cultural training and the mechanization of the fermentation and drying processes. The information provided by this work allows to fine-tune government policies to support the cocoa production sector and increase understanding of cocoa cultivation and its development in the country.
Keywords: cocoa postharvest yield, regression models, cocoa fermentation, cocoa drying, geographic clustering.

Capítulo 1

El cacao, su producción e importancia en el Ecuador

1.1. Generalidades

El cacao (*Theobroma cacao* L.) es un árbol originario de la cuenca amazónica, en particular de las zonas medias y bajas de la selva tropical húmeda. Puede llegar a alcanzar de 4 a 8 metros de altura. A partir del descubrimiento de América se extiende su cultivo a través de las zonas de clima tropical en el mundo para la producción de su principal derivado, el chocolate.
De acuerdo a [14], mediante la adaptación a cada medio y la selección artificial realizada por los agricultores, se diferenciaron tres tipos fundamentales de cacao en el mundo:

- **Cacao criollo:** Este tipo de cacao se extiende entre las selvas de México hasta el norte de Ecuador. Es particularmente apreciado por el fino aroma y sabor del chocolate que se produce a partir de sus semillas, a pesar de que es muy susceptible a enfermedades.

- **Cacao forastero amazónico:** Es la variedad de cacao más cultivado en el mundo, ocupando alrededor del 80 % de la superficie cultivada mundial. A partir de la cuenca del río Amazonas (donde aún se puede encontrar de forma silvestre) se extendió por África, el Sudeste de Asia y Oceanía. Este tipo de cacao produce un chocolate cuyo sabor no destaca particularmente.

- **Cacao trinitario:** Originario de la isla de Trinidad (de ahí su nombre) como una mezcla de las variedades criolla y forastera amazónica. Se encuentra distribuido en Venezuela, Colombia y algunos países de África. La variabilidad entre los fenotipos de cacao trinitario es muy amplia, considerándose el cacao de Trinidad como de una calidad superior al cacao trinitario cultivado en los países africanos, que se considera cacao corriente o común.

A continuación se expone una tabla con los principales productores de cacao a nivel mundial.

Ranking	País	Producción (en Tm)
1	Costa de Marfil	1,488,992
2	Ghana	835,466
3	Indonesia	777,500
4	Nigeria	367,000
5	Camerún	275,000
6	Brasil	256,186
7	Ecuador	128,446
8	México	82,000
9	Perú	71,175
10	República Dominicana	68,021

Tabla 1.1: Principales productores mundiales de cacao en 2013. Fuente: World Resources Institute.

Según [14], en el Ecuador se cultivan dos tipos principales de cacao: el **Nacional Fino de Aroma** que se considera cacao forastero, pero tiene características especiales de calidad que lo relacionan con el tipo criollo; y el **CCN-51 (Colección Castro Naranjal 51)**, un clon desarrollado por el agrónomo ecuatoriano Homero Castro en 1965, que presenta resistencia a plagas y una productividad más alta que el cacao Nacional; a pesar de que el chocolate que se obtiene a partir de este tipo no posee los sabores especiales del tipo Nacional, por lo que es considerado cacao ordinario o común. De acuerdo a [15], la relación entre la participación de estas dos variedades en la exportación ecuatoriana es de 75 % de cacao Nacional y 25 % de CCN-51.
El fruto del cacao, conocido como **mazorca**, varía de forma, color, tamaño, rugosidad y espesor dependiendo de su especie. En su interior almacena las semillas, que se encuentran recubiertas por una pulpa dulce. La cantidad de semillas en una mazorca depende de la efectividad de la fecundación de los ovarios, pudiendo llegar hasta a las 60 semillas por mazorca. La forma de las semillas es muy variable dependiendo de la especie, pudiendo ser entre redondeadas y alargadas. En el cacao de tipo nacional y criollo, el color interno de las semillas va del blanco al lila pálido; mientras que en los tipos trinitarios y forasteros amazónicos es de un color morado muy oscuro.

1.2. Cosecha y postcosecha del cacao

La cosecha de cacao se puede realizar a lo largo de todo el año. En Ecuador, existen dos picos de cosecha muy diferenciados: desde fines de febrero a mediados de mayo (ciclo 1), y de fines de octubre a inicios de enero (ciclo 2). Este último ciclo de cosechas es de menor intensidad que el anterior.
Para determinar el momento óptimo para la cosecha de las mazorcas la clave es el cambio de coloración en su cáscara. Las mazorcas de cacao Nacional, que son verdes en su proceso de maduración, se vuelven de un color amarillo; mientras que aquellas

que tenían un color rojizo se tornan rojas amarillentas o anaranjadas cuando están aptas para la cosecha. Estas mazorcas en general provienen de cacao trinitario. Otra señal es el desprendimiento de las semillas dentro de las mazorcas, produciendo un característico sonido a maraca cuando es sacudida.
A partir de la cosecha de las mazorcas se debe seguir un proceso hasta llegar a la comercialización o exportación del cacao. Este proceso se conoce como **beneficio de las almendras** y se puede observar en el siguiente diagrama de flujo.

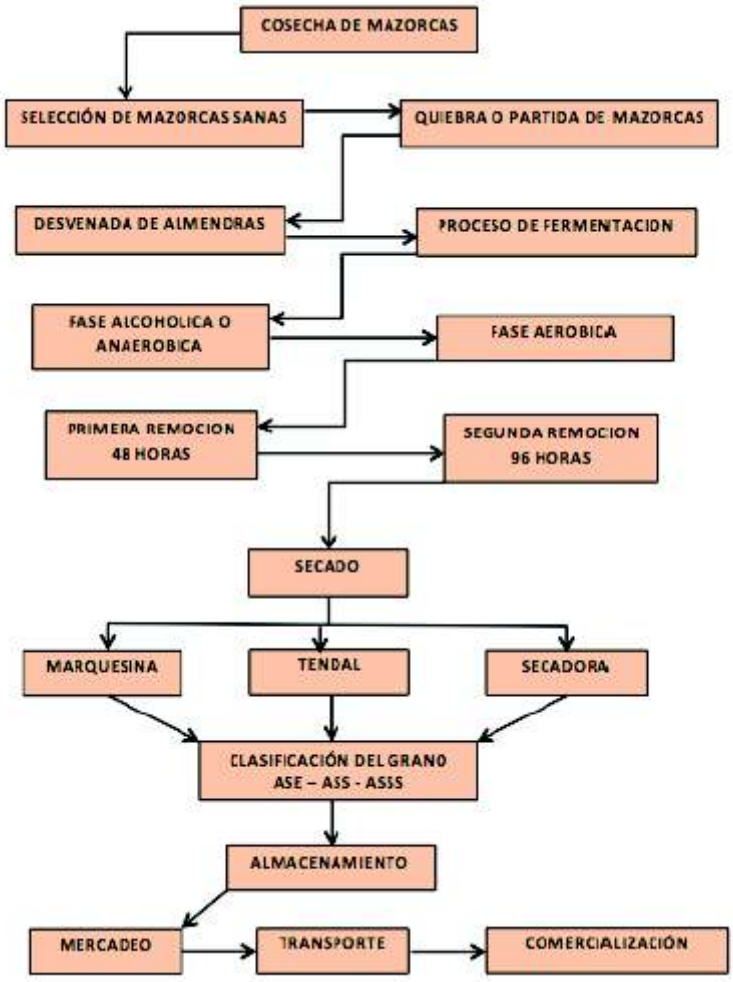


Figura 1.1: Flujo del manejo postcosecha del cacao. Fuente: AGROCALIDAD Cia. Ltda.

Podemos resumir el beneficio en tres etapas claramente diferenciadas: la preparación, la fermentación y el secado. Después de pasar por estas tres etapas, el cacao está listo para su comercialización.

1.2.1. Preparación

El primer paso involucra la **selección de las mazorcas**, mediante la cual se eliminan aquellas que podrían alterar la calidad del producto final. Se deben desechar

las mazorcas que no han completado su proceso de maduración, las que han sido dañadas ya sea por animales o plagas y aquellas que están sobremaduras o podridas. Una vez terminada la selección, se procede a la **quiebra de las mazorcas**, que permite extraer las almendras de su interior. Se debe quebrar las mazorcas mediante un golpe seco contra una superficie dura y no se las debe partir con un machete o cualquier objeto afilado, pues podría dañar las semillas o la baba que los cubre, lo cual causaría problemas más adelante en el proceso de fermentación.
Finalmente, se realiza la **desvenada de las almendras**, que separa a las semillas de la placenta, vena central o maguey a la cual se hallan adheridas dentro de la mazorca.
Las almendras de cacao sueltas tras estos procesos se conocen como **cacao en baba.**

1.2.2. Fermentación

La fermentación de las almendras es el proceso que mata al embrión de las semillas, quitándoles su facultad para germinar, y les concede el sabor a chocolate. Es importante que el cacao listo para ser fermentado no tenga almendras en mal estado, o restos de cáscaras u otras impurezas que podrían alterar el libre desarrollo del tratamiento y por ende, se obtendría un producto de baja calidad.
Este proceso comprende dos etapas: la etapa **alcohólica o anaeróbica** en la cual levaduras actúan sobre la pulpa de las semillas, transformando el azúcar en alcohol; y la etapa **aeróbica** en la cual bacterias transforman el alcohol en ácido acético. En estos procesos se registra un aumento de temperatura (hasta 50°C) de la masa fermentante; esto, unido a la penetración del alcohol y el ácido acético en la semilla, provoca la muerte del embrión y genera una cadena de reacciones químicas en el interior de la almendra que forma los precursores del sabor y aroma del chocolate.
Los métodos más usados para realizar la fermentación del cacao son, de acuerdo a [14], los siguientes:

- **Fermentación por montones:** Se apila el cacao en baba en un tendal sobre hojas de plátano o bijao y se cubre con las mismas hojas para mantener el calor de la reacción. Después de cada remoción se recomienda cambiar las hojas usadas por otras nuevas. La cantidad mínima para generar el suficiente calor para realizar la fermentación es de 80 kilos de cacao en baba.

- **Fermentación en saquillos:** Como su nombre lo indica, se almacena el cacao en baba en saquillos de plástico o yute. Dicho método no es recomendado, por cuanto la fermentación no se realiza en toda la masa pues es casi imposible realizar buenas remociones del cacao si éste está almacenado en saquillos.

- **Fermentación en cajones:** Este es el método más recomendado para la fermentación del cacao, por la facilidad para realizar las remociones y mejorar, por tanto, la calidad del producto fermentado. Se realiza en cajones de madera

de tamaño apropiado con paredes removibles para facilitar el volteo de la masa fermentante. Dichos cajones están construidos en escalera, para facilitar el desplazamiento del cacao a través de sus etapas de fermentación; además, deben tener perforaciones en su piso para facilitar el drenaje de los subproductos del proceso de fermentado y se cubrirán con hojas de plátano o bijao para evitar la fuga de calor. El cajón además deberá estar separado del piso mediante patas de madera, y será construido en maderas blancas (laurel, tillo, pechiche), puesto que las maderas oscuras o resinosas pueden afectar el sabor y la calidad del cacao fermentado.

La fermentación puede tomar diferentes lapsos de tiempo. En el caso del cacao Nacional, el tiempo recomendado es de 4 días, con remociones cada 48 horas. Estas remociones ayudan al fermentado parejo de los granos, puesto que las reacciones químicas generalmente suceden en las capas superiores. Para el cacao trinitario, el tiempo de fermentación recomendado es de 6 días, con remociones cada 2 días. Por estas razones no se recomienda mezclar diferentes tipos de cacao, pues cada uno tiene diferentes tiempos para alcanzar su fermentado óptimo.

Para el clon CCN-51, al efectuar este proceso se obtiene almendras con un sabor muy ácido. Por ello, se han desarrollado tratamientos alternativos para el fermentado de este tipo. Se suele lavar previamente la pulpa de las semillas o proceder a un secado antes de pasar al proceso de fermentación por el mismo tiempo recomendado de 4 días. Con ello se reduce la acidez de las semillas; sin embargo, no se desarrolla completamente el sabor básico del producto final.

La manera de determinar si el proceso de fermentado ha terminado es mediante el descenso de temperatura que se produce cuando las reacciones en la masa fermentante se han detenido, o tomando algunos granos de cacao y examinando su interior. El siguiente cuadro resume las características que diferencian las almendras fermentadas de las no fermentadas.

Características	**Almendras fermentadas**	**Almendras sin fermentación**
Aroma	Agradable	Desagradable, ácido
Sabor	Medianamente amargo	Astringente
Forma	Hinchada	Aplanada
Color interno	Café obscuro	Café violáceo
Textura	Quebradiza	Compacta, dura
Separación de la testa	Fácil	Difícil

Tabla 1.2: Diferencias entre las almendras fermentadas y no fermentadas. Fuente: Manual del Cultivo del Cacao, INIAP, 1994.

1.2.3. Secado

Al terminar la fermentación, el cacao queda con un alto índice de humedad (alrededor del 60 %). Esto puede favorecer la aparición de mohos que descomponen las semillas. Para evitar esto se procede al secado de los granos, lo cual reduce su humedad hasta el valor recomendado de 7 % y además favorece procesos de oxidación que disminuyen la acidez de las semillas de cacao.
El método más usado es el de **secado natural al sol**, por su facilidad y bajo coste. Además, en el secado artificial ciertas reacciones se detienen abruptamente y el cacao queda muy ácido. Se recomienda que los dos primeros días el cacao sea expuesto por máximo 4 horas en capas progresivamente más finas. A partir del tercer día el cacao puede ser expuesto por más tiempo. Las semillas deben ser removidas de tanto en tanto con el fin de conseguir un secado uniforme. Dependiendo del clima, este proceso puede tomar de 4 a 6 días.
El **secado artificial** en máquinas a base de combustible se usa cuando el clima no favorece el secado al sol. En estas secadoras se debe vigilar la temperatura muy cuidadosamente, pues si el secado se realiza muy rápido no se elimina la acidez del cacao, mientras que al tomar demasiado tiempo el grano queda muy frágil y se pierde demasiada masa en el producto resultante. La ventaja más importante del secado artificial es su velocidad; generalmente el proceso se termina entre 8 a 12 horas. Sin embargo, las desventajas son la interrupción de las reacciones de oxidación por las altas temperaturas a las que se exponen las almendras; la acidez del producto final; la falta de desarrollo del aroma y sabor a chocolate; y el coste alto para el agricultor en comparación con el secado natural.
Una vez terminadas las tres etapas del beneficio de las almendras el producto, conocido como **cacao seco**, está listo para ser comercializado.

1.2.4. Comercialización

La calidad del cacao seco influye mucho en los productos derivados que se pueden obtener y en el precio que alcanza en el mercado. Por ejemplo, el cacao Nacional Fino de Aroma es exportado para producir chocolates de alta calidad, mientras que el clon CCN-51 es usado en general para producir chocolate en polvo, manteca de cacao o licor de cacao.
Según criterios propuestos por el INEN [16], el cacao Nacional Fino de Aroma se clasifica en:

- Arriba[1] Superior Summer Selecto (ASSS)
- Arriba Superior Selecto (ASS)

[1]El nombre *cacao Arriba* es una denominación de origen que otorga el Instituto Ecuatoriano de Propiedad Intelectual (IEPI) a los productores de cacao Nacional Fino de Aroma que cumplen estándares que garantizan su calidad.

- Arriba Superior Época (ASE)

Y el cacao CCN-51 se clasifica en:

- Cacao Superior Selecto (CSS)
- Cacao Superior Corriente (CSC)

En promedio, según [29], el cacao ASSS es comercializado a un precio superior entre $1800 y $2000 al precio internacional, el cacao ASS recibe entre $800 y $1200 extra por tonelada y el cacao ASE recibe entre $200 y $300 sobre el precio promedio por tonelada.

1.3. Importancia del cacao en el Ecuador

Desde la época de la Colonia, el cacao ha sido un producto de suprema importancia para la economía del Ecuador. Exceptuando un breve lapso entre 1925 y 1950, la producción de cacao ecuatoriano ha ido siempre en aumento, concentrándose principalmente en las provincias de la Costa ecuatoriana. Según datos del Ministerio de Agricultura, Ganadería, Acuacultura y Pesca (MAGAP) [17], en el 2016 en Ecuador existían 559617 hectáreas de cacao sembrado, ya sea solo o asociado con otros cultivos. Las provincias con mayor superficie cultivada son Los Ríos y Manabí, con 126186 y 125839 hectáreas, respectivamente; mientras que la provincia con mejor rendimiento es Guayas, que en sus 101724 hectáreas sembradas logró un rendimiento de 0,80 Tm/ha.
En los últimos años se ha reactivado la producción de cacao fino de aroma en huertas tradicionales gracias a la creación por parte del MAGAP del Proyecto de Reactivación de Café y Cacao Fino de Aroma. Gracias a estos y otros esfuerzos, se ha incrementado la productividad promedio en el país de 0,18 Tm/ha a 0,49 Tm/ha, y el cacao representa el 5,8 % de las exportaciones agropecuarias; es decir, unos ingresos que en el año 2016 representaron 621 millones de dólares [30]. Aunque anteriormente se mencionó que el país es el séptimo productor mundial de cacao, en la exportación de cacao fino de aroma ocupa el primer lugar, con el 60 % del total de las exportaciones mundiales.
Se puede apreciar en la Figura 1.2 que las exportaciones de cacao ecuatoriano han aumentado significativamente en la última década. Sin embargo, y a pesar del renovado interés que existe en su cultivo y producción tanto por los entes estatales como por los agricultores, y de que iniciativas del Estado como el Proyecto mencionado anteriormente, la Gran Minga del Cacao Nacional, y las Encuestas de Superficie y Producción Agropecuaria Continua (ESPAC) realizadas anualmente por el Instituto Nacional de Estadística y Censos (INEC) han permitido levantar mucha información acerca del estado actual del cultivo de cacao en nuestro país, no existen estudios que hayan usado dicha información para mejorar u optimizar el proceso de tratamiento

Figura 1.2: Exportaciones netas de cacao ecuatoriano de 2005 a 2017. Fuente: Banco Central del Ecuador.

postcosecha del cacao. Los estudios que se han realizado tanto en Ecuador como en otros países se han enfocado solo en la parte de la fermentación [3, 37] o la del secado [1, 21, 23], mientras que aquellos que analizan ambas etapas [11, 28] se concentran en particularidades cualitativas del cacao (aroma, sabor) que provienen de características muy puntuales (genotipo, lugar de origen, altura). Se han realizado algunas investigaciones que intentan estudiar de forma cuantitativa ciertos cultivos, y puntualmente el rendimiento de estos, como la caña de azúcar [26], el trigo [33], la soja y el maíz [32], el plátano [35] y la vid [22]. En todos los estudios citados anteriormente, las técnicas usadas han sido las de regresión múltiple en sus diferentes formas, probando que esta herramienta es ampliamente utilizada para las estimaciones de rendimiento agrícola de diferentes especies de plantas.

En este proyecto se utilizarán los datos obtenidos por el MAGAP a través del Proyecto de Reactivación de Café y Cacao Fino de Aroma en colaboración con la Coordinación General del Sistema de Información Nacional (CGSIN) para realizar un análisis cuantitativo mediante un modelo de regresión lineal del rendimiento del cacao a través de las etapas de beneficio (fermentación y secado); en particular, la disminución de masa que el cacao en baba presenta al transformarse en cacao seco, y las variables socioeconómicas y de métodos de cultivo del agricultor que influyen en esta disminución.

En el siguiente capítulo hablaremos de las herramientas matemáticas que se van a utilizar en la modelación del rendimiento del cacao en el presente trabajo. El capítulo 3 tratará sobre el desarrollo del modelo de regresión para los datos entregados por el MAGAP, y en el capítulo 4 se entregarán algunas conclusiones y recomendaciones deducidas de los resultados del modelo.

Capítulo 2

Modelos de regresión lineal múltiple y algoritmos de agrupación

Hablaremos ahora de los fundamentos teóricos y matemáticos detrás de la modelación planeada para los datos del MAGAP. Se describirán los procesos seguidos para la creación del **modelo de regresión lineal múltiple**, y las técnicas empleadas para la agrupación de datos; concretamente, el **algoritmo** `k-means`.

2.1. Regresión lineal múltiple

2.1.1. Especificación

Es un problema común en muchas ramas de la ciencia intentar predecir el valor Y que tomará una variable a partir de otras p variables $X_1, X_2, \ldots, X_p$ cuyos valores son conocidos. Es decir, encontrar una expresión

$$Y = f(X_1, X_2, \ldots, X_p)$$

para estimar o entender la relación matemática que tiene Y con las variables $X_i, 1 \leq i \leq p$, siendo $i \in \mathbf{N}$.
Cuando esta expresión es de la forma

$$\begin{aligned} Y &= \beta_0 + \sum_{i=1}^{p} \beta_i X_i + u \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + u \end{aligned} \tag{2.1}$$

hablamos de un **modelo de regresión lineal múltiple** para Y (conocida como variable **dependiente, endógena o regresada**), donde $\beta_0, \beta_1, \ldots, \beta_p$ son números reales, las variables X_i son denominadas variables **independientes, explicativas,**

exógenas o regresoras, y u es una perturbación aleatoria, conocida como **error estocástico**. Se puede ver que los valores de Y están formados por dos partes: una **determinista** que depende exactamente de las variables exógenas, y otra **estocástica**, representada por el término u. Este término puede ser útil para representar los errores de medición en la toma de datos, o también la aleatoriedad intrínseca en el comportamiento de muchos fenómenos.

Es importante destacar que la variable endógena del modelo debe ser una variable continua, pues en caso de ser una variable discreta o categórica, estaríamos hablando de otro tipo de modelos, conocidos como modelos de regresión multinomial. Sin embargo, las variables exógenas pueden ser continuas o discretas indistintamente.

El objetivo principal del modelo de regresión, contra lo que pueda parecer, no es estimar el valor de la variable dependiente a partir de las independientes, sino su **valor esperado o promedio** a partir de las variables independientes. Es decir, intentamos estimar $E(Y|X_1, \dots X_p)$. Por lo cual, otra manera de expresar el modelo de regresión lineal es

$$E(Y|X_1, \dots X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{2.2}$$

Se debe aclarar que la predicción de los valores exactos de Y es imposible debido que la relación entre Y y $X_1, \dots, X_p$ es estocástica. Ahora, si reemplazamos (2.2) en (2.1)

$$Y = E(Y|X_1, \dots X_p) + u$$

y tomamos esperanzas condicionales a cada lado

$$E(Y|X_1, \dots X_p) = E(E(Y|X_1, \dots X_p)) + E(u|X_1, \dots X_p).$$

Considerando que la esperanza de una constante es el mismo valor de la constante

$$E(Y|X_1, \dots X_p) = E(Y|X_1, \dots X_p) + E(u|X_1, \dots X_p)$$

de donde obtenemos que

$$E(u|X_1, \dots X_p) = 0.$$

Este resultado es clave para el modelo, como se verá más adelante. Dado que los únicos valores desconocidos para poder manejar Y en la ecuación (2.1) son los coeficientes $\beta_i, 0 \leq i \leq p$ en la población, el objetivo principal del análisis de regresión lineal múltiple es encontrar una estimación de dichos valores $\hat{\beta}_i, 0 \leq i \leq p$ a partir de una muestra aleatoria de n observaciones de las variables endógena y exógenas.

2.1.2. Estimación

Existen muchas maneras de estimar los coeficientes del modelo lineal múltiple, pero el más conocido y utilizado es el método de **mínimos cuadrados ordinarios**

(MCO). Dicho método es atribuido a Carl Friedrich Gauss[1] que lo utilizó para estimar la trayectoria del asteroide Ceres en 1801.

La **función de regresión poblacional (FRP)**, que se expresa de forma equivalente tanto por (2.1) como por (2.2), no es observable directamente en la mayoría de los casos. No obstante, a partir de una muestra de n observaciones elegidas aleatoriamente de la población podemos aproximar dicha función. Esta aproximación, conocida como **función de regresión muestral (FRM)** viene dada por la siguiente ecuación:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p \tag{2.3}$$

donde $\hat{Y}$ es el estimador muestral de $E(Y|X_1, \ldots X_p)$ y $\hat{\beta}_i, 0 \leq i \leq p$ son los estimadores muestrales de $\beta_i, 0 \leq i \leq p$ que son los coeficientes poblacionales del modelo. Ahora, agregando también los estimadores muestrales de u, el término de perturbación estocástica, se tiene una expresión equivalente para la FRM donde se toma en cuenta la relación no determinística entre Y y $X_1, \ldots, X_p$:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p + \hat{u} \tag{2.4}$$

Si reemplazamos (2.3) en (2.4) obtenemos

$$Y = \hat{Y} + \hat{u}$$

o lo que es lo mismo

$$\hat{u} = Y - \hat{Y}$$

de donde vemos que el término $\hat{u}$ no es más que la diferencia entre los valores observados y los esperados de Y en la muestra. Por ello, $\hat{u}$ también es denominado **residual** del modelo muestral.

Para la determinación de los coeficientes desconocidos de la FRM, el método de mínimos cuadrados ordinarios expresa que deben ser calculados de tal manera que la suma de los cuadrados de los residuales muestrales sea minimizada; es decir, que la expresión

$$\begin{aligned} \sum_{j=1}^{n} \hat{u_j}^2 &= \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 \\ &= \sum_{j=1}^{n} (Y_j - (\hat{\beta}_0 + \hat{\beta}_1 X_{j1} + \hat{\beta}_2 X_{j2} + \ldots + \hat{\beta}_p X_{jp}))^2 \end{aligned} \tag{2.5}$$

sea lo más pequeña posible. Empleando derivación multivariante llegamos a las siguientes $p + 1$ ecuaciones con $p + 1$ incógnitas, conocidas como **condiciones de**

[1]Matemático alemán (1777-1855), considerado uno de los más grandes de todos los tiempos. Despuntó tanto en matemáticas como en la física y astronomía. Hizo importantes aportes a la teoría de números, la geometría, el álgebra y la estadística. Es conocido aún como el *príncipe de las matemáticas.*

primer orden de MCO:

$$\begin{aligned}
\sum_{j=1}^{n}(Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_{j1} - \hat{\beta}_2 X_{j2} - \ldots - \hat{\beta}_p X_{jp}) &= 0 \\
\sum_{j=1}^{n} X_{j1}(Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_{j1} - \hat{\beta}_2 X_{j2} - \ldots - \hat{\beta}_p X_{jp}) &= 0 \\
&\vdots \\
\sum_{j=1}^{n} X_{jp}(Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_{j1} - \hat{\beta}_2 X_{j2} - \ldots - \hat{\beta}_p X_{jp}) &= 0
\end{aligned} \tag{2.6}$$

Resolviendo este sistema de ecuaciones encontramos los estimadores $\hat{\beta}_i, 0 \leq i \leq p$, conocidos como los **estimadores MCO** para (2.1).

Otra manera de entender el método de mínimos cuadrados ordinarios es a partir de la **formulación matricial** del modelo de regresión lineal. En efecto, la FRM puede ser expresada también como

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$$

donde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}, \ \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}, \ y \ \hat{\mathbf{u}} = \begin{bmatrix} \hat{u_1} \\ \hat{u_2} \\ \vdots \\ \hat{u_n} \end{bmatrix}.$$

En esta formulación la suma de cuadrados de residuales se expresa mediante

$$\begin{aligned}
\sum_{j=1}^{n} \hat{u_j}^2 = \mathbf{u}'\mathbf{u} &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}
\end{aligned}$$

Para encontrar el mínimo de esta suma procedemos a derivar esta expresión con respecto a $\hat{\boldsymbol{\beta}}$ e igualamos a cero:

$$\begin{aligned}
\frac{\partial}{\partial \hat{\boldsymbol{\beta}}}(\mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) &= 0 \\
-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= 0 \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{Y}
\end{aligned}$$

De donde el estimador MCO en forma matricial viene dado por la fórmula:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.7}$$

Sin importar la forma de expresar el modelo, los estimadores se pueden hallar fácilmente en cualquier paquete estadístico, pues su cálculo manual es bastante engorroso; sin embargo, algunas de sus propiedades más útiles no están garantizadas a menos que se cumplan ciertas condiciones previas. Estas condiciones, que según [24], se conocen como **supuestos del modelo lineal clásico** son:

1. **El modelo es lineal en los parámetros** $\beta_i, 0 \leq i \leq p$**:** el modelo sigue la ecuación (2.1). Al expresar que el modelo es lineal solo en los parámetros, no se especifica que lo sea necesariamente para las variables.

2. **La muestra es aleatoria:** es decir, las n observaciones de las variables exógenas y endógena provienen de una muestra representativa y aleatoria extraída de una población que sigue el modelo especificado por la ecuación (2.1).

3. **No existe colinealidad perfecta:** esta condición expresa que ninguna de las variables es una combinación lineal de las otras, ni tampoco toma un valor constante. En la formulación matricial, esta condición se puede entender como que la matriz $\mathbf{X}$ tiene rango completo de columnas igual a $p + 1$, por tanto ninguna columna es una combinación lineal de alguna de las otras.

4. **Esperanza condicional residual nula:** el valor esperado de u para cualquier valor fijo de las variables independientes debe ser cero. Se cumple entonces:

   $$E(u|X_1, X_2, \ldots, X_p) = 0$$

   En la formulación matricial, se expresa como

   $$E(\mathbf{u}|\mathbf{X}) = \mathbf{0_n}$$

   donde $\mathbf{0_n} = [0\ 0\ 0\ \ldots\ 0]'$, es decir, $\mathbf{0_n}$ es el vector cero de tamaño n.

5. **Homocedasticidad:** La varianza del residual u para cualquier valor fijo de las variables independientes debe ser constante. Es decir:

   $$V(u|X_1, X_2, \ldots, X_p) = \sigma^2$$

   En la formulacion matricial, decimos que

   $$V(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I_n}$$

   donde σ^2 es una constante y además $\mathbf{I_n}$ es la matriz identidad de tamaño n.

6. **Normalidad:** El error estocástico u es independiente de las variables explicativas, y sigue una distribución normal de media cero y varianza σ^2:

$$u \rightsquigarrow N(0, \sigma^2)$$

En la formulación matricial se dice que el vector $\mathbf{u}$ sigue una distribución normal multivariante con vector de medias $\mathbf{0_n}$ y matriz de varianzas y covarianzas $\sigma^2\mathbf{I_n}$:

$$\mathbf{u} \rightsquigarrow N(\mathbf{0_n}, \sigma^2\mathbf{I_n})$$

Dadas estas suposiciones, se tiene el siguiente teorema.

Teorema 2.1.1 (Teorema de Gauss-Markov)**.** *Bajo los supuestos del modelo lineal clásico 1 al 5, el estimador $\hat{\boldsymbol{\beta}}$ definido en (2.7) existe, es único y es el* ***estimador lineal insesgado de mínima varianza*** *de los valores reales $\boldsymbol{\beta}$ en la población, donde:*

- *Estimador insesgado se refiere a que el valor esperado del estimador muestral es el valor del parámetro poblacional:*

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

- *Estimador de mínima varianza quiere decir que si se calculara otro estimador lineal $\bar{\boldsymbol{\beta}}$, su varianza sería siempre superior a la del estimador calculado mediante (2.7). Es decir:*

$$V(\hat{\boldsymbol{\beta}}) \leq V(\bar{\boldsymbol{\beta}})$$

Demostración. Demostrar la existencia de $\hat{\boldsymbol{\beta}}$ es simple. De (2.7) observamos que el cálculo del estimador está sujeto a la existencia de la matriz inversa de $\mathbf{X'X}$. Dado que la matriz $\mathbf{X}$ es de rango completo por el supuesto 3, $\mathbf{X'}$ también lo es. Por tanto su producto $\mathbf{X'X}$ también será de rango completo y como esta es una matriz cuadrada, será invertible. De donde queda demostrada la existencia y unicidad de $\hat{\boldsymbol{\beta}}$. Ahora demostraremos el insesgamiento del estimador. Primero, lo expresaremos de la siguiente forma:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X'X})^{-1}\mathbf{X'Y} \\
&= (\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{X'X})^{-1}(\mathbf{X'X})\boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'u} \\
&= \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'u}
\end{aligned}$$

donde hemos usado el hecho de que $(\mathbf{X'X})^{-1}(\mathbf{X'X}) = \mathbf{I}_{p+1}$. Ahora calculamos la

esperanza condicional de $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= E(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0_n} \\ &= \boldsymbol{\beta} \end{aligned}$$

Aquí hemos usado el supuesto 4 que expresa que $E(\mathbf{u}|\mathbf{X}) = \mathbf{0_n}$. Por consiguiente, hemos demostrado que $\hat{\boldsymbol{\beta}}$ es insesgado.
Ahora demostraremos que $\hat{\boldsymbol{\beta}}$ es el estimador lineal de varianza mínima. Para ello, supongamos que existe otro estimador lineal $\bar{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}$, donde la matriz $\mathbf{A}$ es una matriz de dimensiones $n \times (p+1)$ formada por constantes y funciones no aleatorias de $\mathbf{X}$ para que este nuevo estimador también sea insesgado y lineal en los parámetros. Mediante una manipulación algebraica de $\bar{\boldsymbol{\beta}}$ obtenemos:

$$\bar{\boldsymbol{\beta}} = \mathbf{A}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}.$$

Calculando la esperanza de $\bar{\boldsymbol{\beta}}$:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + E(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'E(\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\mathbf{0_n} \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Para que este estimador sea insesgado se debe cumplir que $E(\bar{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, lo que solo sucedería si $\mathbf{A}'\mathbf{X} = \mathbf{I}_{p+1}$.
Primero procederemos a calcular la matriz de varianzas y covarianzas de $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= V(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}) \\ &= \cancel{V(\boldsymbol{\beta}|\mathbf{X})} + V((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(V(\mathbf{u}|\mathbf{X}))\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2\cancel{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Por tanto, tenemos que $V(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. A continuación calcularemos la matriz

de varianzas y covarianzas del estimador $\bar{\boldsymbol{\beta}}$:

$$\begin{aligned} V(\bar{\boldsymbol{\beta}}|\mathbf{X}) &= V((\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= V(\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \cancel{V(\boldsymbol{\beta}|\mathbf{X})} + V(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'V(\mathbf{u}|\mathbf{X})\mathbf{A} \\ &= \mathbf{A}'(\sigma^2\mathbf{I}_n)\mathbf{A} \\ &= \sigma^2\mathbf{A}'\mathbf{A} \end{aligned}$$

A partir de estas matrices calcularemos su diferencia:

$$\begin{aligned} V(\bar{\boldsymbol{\beta}}|\mathbf{X}) - V(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \\ &= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\ &= \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A} \end{aligned}$$

En esto proceso hemos usado el hecho de que $\mathbf{A}'\mathbf{X} = \mathbf{I}_n$ y definimos $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Se ve fácilmente que $\mathbf{M}$ es una matriz simétrica e idempotente, por lo cual sabemos que $\mathbf{A}'\mathbf{M}\mathbf{A}$ es semidefinida positiva para cualquier matriz $\mathbf{A}$ de dimensión adecuada. Consideremos entonces cualquier combinación lineal de los parámetros que viene dada por la fórmula:

$$\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + c_2\beta_2 + \ldots + c_p\beta_p$$

donde $\mathbf{c} = [c_0\ c_1\ c_2\ \ldots\ c_p]'$ un vector de tamaño $p+1$. Los estimadores que se pueden usar para $\mathbf{c}'\boldsymbol{\beta}$ son $\mathbf{c}'\bar{\boldsymbol{\beta}}$ y $\mathbf{c}'\hat{\boldsymbol{\beta}}$ y ambos son insesgados, pero tenemos que:

$$V(\mathbf{c}'\bar{\boldsymbol{\beta}}|\mathbf{X}) - V(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{c}'[V(\bar{\boldsymbol{\beta}}|\mathbf{X}) - V(\hat{\boldsymbol{\beta}}|\mathbf{X})]\mathbf{c} \geq 0$$

por definición de matriz semidefinida positiva. Por tanto tenemos que para cualquier vector $\mathbf{c}$ de dimensión adecuada se cumple que:

$$V(\mathbf{c}'\bar{\boldsymbol{\beta}}|\mathbf{X}) - V(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) \geq 0$$

o lo que es igual

$$V(\mathbf{c}'\bar{\boldsymbol{\beta}}|\mathbf{X}) \geq V(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}).$$

Es decir, el estimador MCO tiene menor varianza que cualquier otro estimador lineal de los parámetros poblacionales (o una combinación lineal de ellos), lo que completa la demostración del teorema. □

Procederemos ahora a definir también ciertos valores que serán de utilidad en los

procesos que realizaremos más adelante.

Definición 2.1.1. Se define la **suma de cuadrados total (SCT)** del modelo mediante la siguiente fórmula

$$SCT = \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \tag{2.8}$$

donde $\bar{Y}$ es el promedio de la variable endógena.

Podemos entender SCT como una medida de la variación total que existe entre los valores reales de Y alrededor de su media muestral. Sin embargo, mediante una descomposición de (2.8)

$$\begin{aligned} SCT &= \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \\ &= \sum_{j=1}^{n} (Y_j - \hat{Y}_j + \hat{Y}_j - \bar{Y})^2 \\ &= \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 + \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2 \\ &= \sum_{j=1}^{n} \hat{u_j}^2 + \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2 \end{aligned} \tag{2.9}$$

vemos que SCT puede dividirse en otras dos sumas de cuadrados, una formada por los residuos del modelo, y otra por la diferencia entre los valores ajustados de Y según el modelo y la media muestral de los valores observados. Definiremos a continuación estas sumas.

Definición 2.1.2. Se denomina **suma de cuadrados residuales (SCR)** a la primera parte de la SCT definida en (2.9); esto es:

$$SCR = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^{n} \hat{u_j}^2$$

También se define como **suma de cuadrados explicada por el modelo (SCE)** a la segunda parte de SCT:

$$SCE = \sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2$$

De estas definiciones es claro que $SCT = SCE + SCR$.

A partir de estas sumas de cuadrados se puede definir otros valores que ayudan a medir la adecuación del modelo. Definiremos dichos valores a continuación.

Definición 2.1.3. Se conoce como **coeficiente de determinación R^2 del modelo** al valor definido por la expresión:

$$R^2 = \frac{SCE}{SCT}$$

La raíz cuadrada del coeficiente R^2 se conoce como **coeficiente de correlación múltiple** R.

Este coeficiente mide la proporción de la variación total que es explicada por el modelo lineal. Su valor puede variar entre 0 y 1: así, mientras más cercano a 1 esté, mejor ajustan los valores dados por el modelo $\hat{Y}_t$ a los datos observados en la muestra Y_t, y por tanto, mejor es el modelo. El coeficiente de correlación múltiple en cambio calcula la correlación lineal existente entre los valores observados de Y y los valores ajustados de Y. Pero dado que estos valores ajustados son dependientes de las variables $X_i, 0 \leq i \leq p$, podemos considerarlo también como una medida de la correlación entre los valores observados de Y y las variables exógenas del modelo $X_i, 0 \leq i \leq p$.
Este coeficiente de determinación es una medida muy útil de la bondad de ajuste del modelo; no obstante, posee una debilidad: para hacer aumentar su valor solo basta con aumentar el número de variables X_i en el modelo, sin importar si son significativas o no. Para mejorar entonces este coeficiente debemos definir un valor que no tenga este problema.

Definición 2.1.4. Se denomina **coeficiente de determinación R^2 ajustado del modelo,** notado como $\bar{R^2}$, al valor calculado con la siguiente fórmula:

$$\bar{R^2} = 1 - \frac{n-1}{n-p}(1 - R^2)$$

donde R^2 es el coeficiente de determinación definido anteriormente.

Una vez calculados los estimadores, se debe proceder a verificar que el modelo cumpla con los supuestos definidos anteriormente. En particular, las condiciones 3, 5 y 6. Además se debe probar estadísticamente que las variables incluidas en el modelo sean significativas. Este proceso se conoce como **validación del modelo.**

2.1.3. Validación

Nos concentraremos en cuatro puntos principales:

- Probar estadísticamente la significatividad del modelo en general y las variables exógenas que lo forman en particular.
- Descartar la presencia de multicolinealidad entre las variables explicativas.
- Probar la homocedasticidad de los residuales.

- Probar que los residuales siguen una distribución normal.

Significatividad del modelo y sus variables

En la demostración del teorema 2.1.1 se tomó en cuenta solamente a los 5 primeros supuestos del modelo lineal clásico. El supuesto 6 no se utilizó, puesto que no es necesario para dicha demostración. Sin embargo la utilidad de este supuesto es la de definir las distribuciones de probabilidad de los estimadores $\hat{\beta}_i, 0 \leq i \leq p$, y conociendo estas, definir pruebas de hipótesis e intervalos de confianza para estos estimadores.
Primero trabajaremos sobre la significación total del modelo. Es decir, tendremos la hipótesis nula

$$H_o: \ \beta_1 = \ldots = \beta_p = 0$$

contra la hipótesis alternativa

$$H_a: \ \exists i, 1 \leq i \leq p, \ \beta_i \neq 0$$

Para realizar esta prueba introduciremos ahora un resultado conocido que aparece en [27].

Teorema 2.1.2 (Teorema de Cochran)**.** *Sean $X_1, X_2, \ldots, X_n$ variables independientes y normalmente distribuidas con media 0 y varianza 1. Sea $Q = \sum_{i=1}^{n} X_i^2$ su suma de cuadrados, que sigue una distribución χ_n^2. Sean también $Q_1, Q_2, \ldots, Q_m$ sumas de cuadrados con $f_1, f_2, \ldots, f_m$ grados de libertad respectivamente, que cumplen:*

$$\sum_{i=1}^{m} Q_i = Q$$

y

$$\sum_{i=1}^{m} f_i = n$$

Entonces cada Q_i es independiente de los demás y sigue una distribución $\chi_{f_i}^2$.

Dado este resultado podemos demostrar el siguiente teorema.

Teorema 2.1.3. *Dados los supuestos 1 al 6 del modelo lineal clásico, y suponiendo la hipótesis nula*

$$H_o: \ \beta_1 = \ldots = \beta_p = 0$$

tenemos que

$$F = \frac{SCE/p}{SCR/n - p - 1} \tag{2.10}$$

sigue una distribución F con p grados de libertad en el numerador y $n - p - 1$ en el denominador.

Demostración. A partir de

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_p X_{pj} + u_j$$

y sabiendo que $\beta_1 = \ldots = \beta_p = 0$, tenemos que

$$Y_j = \beta_0 + u_j$$

Se puede ver claramente que aplicando el supuesto 4 se tiene $\bar{Y} = \beta_0$. Por tanto, y suponiendo que $u \rightsquigarrow N(0, \sigma^2)$ tenemos que

$$\frac{Y_j - \bar{Y}}{\sigma} \rightsquigarrow N(0,1)$$

Podemos calcular la suma de cuadrados de esta variable, que será:

$$\sum_{j=1}^{n} \left(\frac{Y_j - \bar{Y}}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{j=1}^{n} (Y_j - \bar{Y})^2 = \frac{1}{\sigma^2} SCT$$

Y dado que conocemos que $SCT = SCE + SCR$ tenemos que, por el teorema 2.1.2,

$$\frac{1}{\sigma^2} SCE \rightsquigarrow \chi^2_p \quad y \quad \frac{1}{\sigma^2} SCR \rightsquigarrow \chi^2_{n-p-1}$$

pues SCT tiene $n-1$ grados de libertad dado que se pierde un grado de libertad al calcular $\bar{Y}$; SCE tiene p grados de libertad pues depende de los estimadores MCO $\hat{\beta}_i, 1 \leq i \leq p$ ($\hat{\beta}_0$ no entra en esta suma pues se elimina al realizar las diferencias); y SCR tiene $n-(p+1)$ grados de libertad pues a partir de la suma residual se calculan todos los estimadores MCO $\hat{\beta}_i, 0 \leq i \leq p$.
Ahora, por un resultado de teoría de probabilidades se tiene que si $U \rightsquigarrow \chi^2_n$ y $V \rightsquigarrow \chi^2_m$, entonces

$$\frac{U/n}{V/m} \rightsquigarrow F_{n,m}$$

Aplicando este resultado tenemos que

$$\frac{\frac{1}{\sigma^2} SCE/p}{\frac{1}{\sigma^2} SCR/n-p-1} = \frac{SCE/p}{SCR/n-p-1} \rightsquigarrow F_{p,n-p-1}$$

que es lo que queríamos demostrar. □

Este teorema nos muestra que para probar la hipótesis de que todos los coeficientes poblacionales sean cero se puede usar el estadístico F definido en (2.10). Si el F calculado supera el valor crítico para cierto nivel de confianza, se rechaza H_o y se acepta que el modelo de regresión lineal es significativo y explica la variabilidad de Y.
Ahora trabajaremos en la significatividad de variables aisladas. Decir que una varia-

ble X_i es significativa equivale a decir que el β_i asociado a ella es distinto de cero. Es decir, probaremos la hipótesis

$$H_o: \ \beta_i = 0$$

contra la hipótesis alternativa

$$H_a: \ \beta_i \neq 0$$

Para ello debemos conocer la distribución de probabilidad muestral de $\hat{\beta}_i$. Para ello tenemos el siguiente teorema.

Teorema 2.1.4. *Bajo los supuestos 1 al 6 del modelo lineal clásico, se tiene que*

$$\hat{\beta}_i \rightsquigarrow N(\beta_i, V(\beta_i))$$

donde $V(\beta_i) = \sigma^2 a_{ii}$ *siendo* a_{ii} *el término ubicado en la i-ésima posición de la diagonal de la matriz* $(\mathbf{X}'\mathbf{X})^{-1}$.

Demostración. Recordemos, del teorema 2.1.1, que el estimador $\hat{\boldsymbol{\beta}}$ puede expresarse como

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

Es decir, cada estimador $\hat{\beta}_i, \ 0 \leq i \leq p$ puede ser visto como $\hat{\beta}_i = \beta_i + \sum_{j=1}^{n} w_{ij}u_j$ donde w_{ij} son valores que solo dependen de las variables independientes X_i. Por tanto, cada estimador puede ser visto como una combinación lineal de los u_j, que tienen una distribución normal por el supuesto 6. Y dado que la combinación de variables distribuidas normalmente también está distribuida normalmente, tenemos que $\hat{\beta}_i, \ 0 \leq i \leq p$ sigue una distribución normal. Su media y varianza ya fueron calculadas en el teorema 2.1.1. □

Entonces tenemos una distribución conocida para probar la hipótesis de significatividad de una variable. Sin embargo, el problema es que σ^2, la varianza del término de error no siempre es conocida, sino que se estima a partir de la información que el mismo modelo proporciona. El siguiente teorema proporciona un estimador para σ^2.

Teorema 2.1.5. *Dados lo supuestos 1 al 6 del modelo lineal clásico, tenemos que*

$$\hat{\sigma^2} = \frac{SCR}{n-p-1}$$

es un estimador insesgado para σ^2.

Demostración. Vimos en la demostración del teorema 2.10 que

$$\frac{1}{\sigma^2} SCR \rightsquigarrow \chi^2_{n-p-1}$$

Por tanto tenemos que

$$E\left(\frac{SCR}{\sigma^2}\right) = E(\chi^2_{n-p-1})$$

Dado que $E(\chi^2_{n-p-1}) = n-p-1$

$$\frac{1}{\sigma^2}E(SCR) = n-p-1$$

o lo que es lo mismo

$$\frac{1}{n-p-1}E(SCR) = E\left(\frac{SCR}{n-p-1}\right) = \sigma^2$$

□

Entonces, conociendo este nuevo estimador de σ^2, podemos reemplazarlo en la distribución conocida de $\hat{\beta}_i$. No obstante, su distribución no será la misma. El siguiente teorema nos da la distribución de probabilidad de los estimadores al reemplazar σ^2 por $\hat{\sigma^2}$.

Teorema 2.1.6. *Dados los supuestos 1 al 6 del modelo lineal clásico, se tiene que*

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{a_{ii}}} \tag{2.11}$$

sigue una distribución t con $n-p-1$ grados de libertad.

Demostración. La distribución t con m grados de libertad se define de la siguiente manera: si $X \rightsquigarrow N(0,1)$ y $Y \rightsquigarrow \chi^2_m$, tenemos que

$$\frac{X}{\sqrt{\frac{Y}{m}}} \rightsquigarrow t_m$$

Por propiedades de la distribución normal y partiendo del teorema 2.1.4 tenemos que

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{a_{ii}}} \rightsquigarrow N(0,1)$$

Sabemos que $\frac{SCR}{\sigma^2} \rightsquigarrow \chi^2_{n-p-1}$. Ahora, mediante una manipulación algebraica tenemos que

$$\frac{SCR}{\sigma^2} = \frac{n-p-1}{n-p-1}\frac{SCR}{\sigma^2} = \frac{n-p-1}{\sigma^2}\frac{SCR}{n-p-1} = (n-p-1)\frac{\hat{\sigma^2}}{\sigma^2}$$

que tendrá la misma distribución. Por tanto:

$$\frac{\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{a_{ii}}}}{\sqrt{\frac{\cancel{(n-p-1)}\frac{\hat{\sigma^2}}{\sigma^2}}{\cancel{(n-p-1)}}}} = \frac{\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{a_{ii}}}}{\sqrt{\frac{\hat{\sigma^2}}{\sigma^2}}} = \frac{\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{a_{ii}}}}{\frac{\hat{\sigma}}{\sigma}} = \frac{(\hat{\beta}_i-\beta_i)\cancel{(\sigma)}}{\hat{\sigma}\cancel{\sigma}\sqrt{a_{ii}}} = \frac{\hat{\beta}_i-\beta_i}{\hat{\sigma}\sqrt{a_{ii}}}$$

tiene una distribución t_{n-p-1} con lo cual finaliza la demostración. □

Para probar entonces la hipótesis nula

$$H_o : \ \beta_i = 0$$

contra la hipótesis alternativa

$$H_a : \ \beta_i \neq 0$$

calculamos el estadístico definido en (2.11) y reemplazando en él $\beta_i = 0$ tendríamos que este equivaldría a

$$\frac{\hat{\beta}_i}{\sqrt{V(\hat{\beta}_i)}} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{a_{ii}}} \tag{2.12}$$

Si este estadístico es mayor al valor crítico de la distribución t para un nivel de confianza dado, se rechaza la hipótesis nula y se acepta la alternativa de que la variable es significativa en el modelo.
Entonces ya tenemos dos estadísticos para realizar las pruebas de hipótesis necesarias en el modelo. Estos estadísticos son calculados generalmente de forma automática por los paquetes informáticos más utilizados, así que no es necesario hacerlo de forma manual.
Una manera equivalente de probar las hipótesis mencionadas es a partir del **p-valor** de los estadísticos. Este p-valor se define como la probabilidad de que bajo la distribución dada del estadístico y considerando la hipótesis nula verdadera, se supere el valor calculado. Si este p-valor es menor que un nivel de confianza acordado (en este trabajo usaremos 0.05 o 5 %) se rechaza la hipótesis nula y se acepta la alternativa. Este p-valor también es fácilmente calculado en paquetes informáticos.
Resumiendo entonces:

- **Significación total del modelo:** Calculamos el estadístico F del modelo definido en (2.10) y a partir de este su p-valor dado por la distribución F. Si el p-valor es menor que $0,05$, rechazamos la hipótesis nula de no significatividad del modelo y aceptamos que al menos una de las variables independientes explica la variabilidad de Y.

- **Significatividad de variables independientes aisladas:** Calculamos el estadístico definido en (2.12) y su p-valor en la distribución t. Si el p-valor es menor a $0,05$, rechazamos la hipótesis nula y aceptamos que la variable independiente es significativa en el modelo.

Además, existen pruebas para verificar si la forma funcional para el modelo es la correcta. Deseamos revisar si el modelo (2.1) mejoraría al incluir potencias de las variables explicativas $X_i, 0 \leq i \leq p$ como variables exógenas. Para ello existe la **prueba de error de especificación de la ecuación de regresión**, conocida

como **RESET** por sus siglas en inglés.
Para realizar dicho test se añaden dos variables más al modelo, que son los valores ajustados de Y elevados al cuadrado y al cubo. Exponentes más grandes son posibles mas no son muy comunes. Por tanto, el modelo queda expresado de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \gamma_1 \hat{Y}^2 + \gamma_2 \hat{Y}^3 + u$$

Y se prueba la hipótesis nula de

$$H_o : \ \gamma_1 = \gamma_2 = 0$$

contra la hipótesis alternativa

$$H_a : \ \gamma_1 \neq 0 \ o \ \gamma_2 \neq 0$$

Estas hipótesis se pueden probar con los estadísticos t respectivos para γ_1 y γ_2 o mediante el estadístico definido por

$$\frac{(R_n^2 - R_o^2)/2}{(R_n^2/n - k - 3)}$$

donde R_o^2 es el estadístico R cuadrado del modelo original y R_n^2 es el estadístico R cuadrado de la regresión con las nuevas variables. Este valor sigue una distribución $F_{2,n-k-3}$. En caso de aceptar la hipótesis nula, se acepta que el modelo está bien especificado y potencias superiores de las variables no deberían ser significativas. Caso contrario, y dependiendo de qué coeficiente resulte significativo, se debería agregar potencias más altas o los productos cruzados de las variables explicativas del modelo original.
Si bien esta prueba determina que algunas variables deberían agregarse al modelo, no especifica claramente cuáles son esas variables. Para determinarlas se puede proceder a probar su significatividad individualmente mediante el estadístico t o revisar otros métodos que desarrollaremos en secciones posteriores.

Multicolinealidad

La **multicolinealidad** se presenta cuando alguna de las variables X_i es una combinación lineal exacta de las otras, o bien la correlación lineal entre un grupo de variables es muy alta. El primer caso se conoce como **multicolinealidad perfecta** mientras que el segundo se define como **multicolinealidad aproximada.** Es de destacar que es imposible que no exista correlación entre variables explicativas y con ella, cierto grado de multicolinealidad en un modelo lineal. Sin embargo, hay que distinguir en qué punto la multicolinealidad se convierte en un problema para la formulación del modelo.
Antes de hablar en detalle sobre los efectos de la multicolinealidad, demostraremos

dos resultados que emplearemos en las siguientes secciones.

Teorema 2.1.7. *El estimador* $\hat{\beta}_i,\ 1 \leq i \leq p$ *puede ser expresado también como*

$$\hat{\beta}_i = \frac{\sum_{j=1}^{n} \hat{r_{ji}} Y_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \tag{2.13}$$

donde $\hat{r_i}$ *es el residuo obtenido al hacer una regresión de la variable* X_i *sobre las otras variables explicativas del modelo.*

Demostración. Sean $\hat{X}_i$ los valores ajustados obtenidos al efectuar la regresión de X_i sobre las otras variables explicativas del modelo. Por tanto, es claro que $X_{ji} = \hat{X}_{ji} + \hat{r_{ji}},\ 1 \leq j \leq n$. Si sustituimos esta ecuación en la correspondiente del sistema (2.6) tenemos:

$$\sum_{j=1}^{n} (\hat{X_{ji}} + \hat{r_{ji}})(Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_{j1} - \hat{\beta}_2 X_{j2} - \ldots - \hat{\beta}_p X_{jp}) = 0$$

Tomando en cuenta que el segundo paréntesis de la ecuación anterior no es más que la fórmula de los residuos $\hat{u}$ del modelo, y dado que por definición $\sum_{j=1}^{n} \hat{X}_{ji} \hat{u_{ji}} = 0$ puesto que $\hat{X}_i$ es función lineal de las variables $X_1, X_2, \ldots, X_p$, la ecuación previa se simplifica a:

$$\sum_{j=1}^{n} \hat{r_{ji}} (Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_{j1} - \hat{\beta}_2 X_{j2} - \ldots - \hat{\beta}_p X_{jp}) = 0$$

Dao que $\hat{r_i}$ son residuos pero de la regresión auxiliar de X_i sobre las otras variables exógenas, tenemos también que $\sum_{j=1}^{n} X_{jm} \hat{r_{ji}} = 0$ para $1 \leq m \leq p,\ m \neq i$. Por tanto, la ecuación se puede simplificar a:

$$\sum_{j=1}^{n} \hat{r_{ji}} (Y_j - \hat{\beta}_i X_{ji}) = 0$$

Reemplazando otra vez aquí $X_{ji} = \hat{X}_{ji} + \hat{r_{ji}}$:

$$\sum_{j=1}^{n} \hat{r_{ji}} (Y_j - \hat{\beta}_i (\hat{X_{ji}} + \hat{r_{ji}})) = 0$$

Separando la sumatoria:

$$\sum_{j=1}^{n} \hat{r_{ji}} Y_j - \hat{\beta}_i \sum_{j=1}^{n} \hat{r_{ji}} \hat{X_{ji}} - \hat{\beta}_i \sum_{j=1}^{n} (\hat{r_{ji}})^2 = 0$$

Por definición de los residuos tenemos que $\sum_{j=1}^{n} \hat{r_{ji}} \hat{X_{ji}} = 0$ de donde:

$$\sum_{j=1}^{n} \hat{r_{ji}} Y_j - \hat{\beta}_i \sum_{j=1}^{n} (\hat{r_{ji}})^2 = 0$$

Despejando $\hat{\beta}_i$ de esta ecuación obtenemos el resultado buscado. □

A partir del teorema 2.1.7 podemos encontrar una expresión para la varianza de los estimadores que también dependerá de una regresión auxiliar. Esta es la finalidad del siguiente teorema.

Teorema 2.1.8. *Bajo los supuestos 1 al 5 del modelo lineal clásico, tenemos que*

$$V(\hat{\beta}_i) = \frac{\sigma^2}{\sum_{j=1}^{n} (X_{ji} - \bar{X}_i)^2} \left(\frac{1}{1 - R_i^2} \right) \tag{2.14}$$

donde σ^2 representa la varianza de los residuales, $\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$ es la suma de las variaciones al cuadrado entre los valores de la variable X_i asociada al coeficiente $\hat{\beta}_i$ y su promedio muestral $\bar{X}_i$, y R_i^2 es el estadístico R cuadrado de la regresión auxiliar que tiene como variable endógena a X_i y como variables exógenas las restantes variables del modelo.

Demostración. Recordemos que según el teorema 2.1.7 los estimadores MCO se pueden expresar de acuerdo a (2.13). Calculemos entonces la varianza de $\hat{\beta}_i$ condicional a los valores de $X_i, 1 \leq i \leq p$:

$$V(\hat{\beta}_i | X_1, X_2, \ldots, X_p) = V \left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} Y_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \right)$$

Dado que $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + u$ por el supuesto 1, reemplazamos esta expresión en la ecuación anterior:

$$V \left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} Y_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \right) = V \left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} (\beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \ldots + \beta_p X_{jp} + u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \right)$$

Recordando de la demostración del teorema 2.1.7 que $\sum_{j=1}^{n} X_{jm} \hat{r_{ji}} = 0$ para $1 \leq m \leq p,\ m \neq i$ y que por la definición de residual $\sum_{j=1}^{n} \hat{r_{ji}} = 0$ la mayoría de los términos se eliminan, quedando solo la siguiente expresión:

$$V \left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} (\beta_0 + \beta_1 X_{j1} + \ldots + \beta_p X_{jp} + u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \right) = V \left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} (\beta_i X_{ji} + u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \right)$$

Conociendo que $X_{ji} = \hat{X_{ji}} + \hat{r_{ji}}$:

$$V \left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} (\beta_i X_{ji} + u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \right) = V \left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} (\beta_i (\hat{X_{ji}} + \hat{r_{ji}}) + u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \right)$$

Recurrimos otra vez a la demostración del teorema 2.1.7 donde usamos el hecho de que $\sum_{j=1}^{n} \hat{X_{ji}} \hat{r_{ji}} = 0$:

$$\begin{aligned} V\left(\frac{\sum_{j=1}^{n} \hat{r_{ji}}(\beta_i(\hat{X_{ji}} + \hat{r_{ji}}) + u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2}\right) &= V\left(\frac{\sum_{j=1}^{n} \hat{r_{ji}}(\beta_i \hat{r_{ji}} + u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2}\right) \\ &= V\left(\frac{\beta_i \sum_{j=1}^{n} (\hat{r_{ji}})^2 + \sum_{j=1}^{n} \hat{r_{ji}} u_j)}{\sum_{j=1}^{n} \hat{r_{ji}}^2}\right) \\ &= V\left(\frac{\beta_i \cancel{\sum_{j=1}^{n} (\hat{r_{ji}})^2}}{\cancel{\sum_{j=1}^{n} \hat{r_{ji}}^2}} + \frac{\sum_{j=1}^{n} \hat{r_{ji}} u_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2}\right) \\ &= V\left(\beta_i + \frac{\sum_{j=1}^{n} \hat{r_{ji}} u_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2}\right) \end{aligned}$$

Considerando que la varianza de una constante es cero, y que dado el supuesto 2 de muestreo aleatorio se puede considerar que los residuos u_j son independientes:

$$\begin{aligned} V\left(\beta_i + \frac{\sum_{j=1}^{n} \hat{r_{ji}} u_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2}\right) &= \cancelto{0}{V(\beta_i)} + V\left(\frac{\sum_{j=1}^{n} \hat{r_{ji}} u_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2}\right) \\ &= \frac{1}{\left(\sum_{j=1}^{n} \hat{r_{ji}}^2\right)^2} V\left(\sum_{j=1}^{n} \hat{r_{ji}} u_j\right) \\ &= \frac{1}{\left(\sum_{j=1}^{n} \hat{r_{ji}}^2\right)^2} \left(\sum_{j=1}^{n} (\hat{r_{ji}})^2 V(u_j | X_1, X_2, \ldots, X_p)\right) \end{aligned}$$

Usando el supuesto 5 de homocedasticidad tenemos que $V(u|X_1, X_2, \ldots, X_p) = \sigma^2$. Reemplazando en nuestra última ecuación:

$$\begin{aligned} \frac{1}{\left(\sum_{j=1}^{n} \hat{r_{ji}}^2\right)^2} \left(\sum_{j=1}^{n} (\hat{r_{ji}})^2 V(u_j | X_1, \ldots, X_p)\right) &= \frac{1}{\left(\sum_{j=1}^{n} \hat{r_{ji}}^2\right)^2} \left(\sum_{j=1}^{n} \sigma^2 (\hat{r_{ji}})^2\right) \\ &= \frac{\sigma^2}{\left(\sum_{j=1}^{n} \hat{r_{ji}}^2\right)^2} \left(\sum_{j=1}^{n} \hat{r_{ji}}^2\right) \\ &= \frac{\sigma^2}{\sum_{j=1}^{n} \hat{r_{ji}}^2} \end{aligned}$$

La expresión del denominador es la suma de cuadrados de residuales de la regresión auxiliar. Notaremos a esta suma de cuadrados como SCR_i, y a la suma de cuadrados total y explicada de la regresión auxiliar como SCT_i y SCE_i respectivamente. Aquí también se cumplirá que $SCT_i = SCR_i + SCE_i$ o lo que es igual $SCR_i = SCT_i -$

SCE_i. Por tanto, podemos escribir:

$$\begin{aligned}\frac{\sigma^2}{\sum_{j=1}^{n} \hat{r_{ji}}^2} &= \frac{\sigma^2}{SCT_i - SCE_i} \\ &= \frac{\sigma^2}{SCT_i(1 - \frac{SCE_i}{SCT_i})}\end{aligned}$$

Pero $\frac{SCE_i}{SCT_i}$ no es más que el estadístico R cuadrado de la regresión auxiliar. Expresándolo como R_i^2 y reemplazando la expresión de SCT dada por (2.9):

$$\begin{aligned}V(\hat{\beta}_i|X_1, \ldots, X_p) &= \frac{\sigma^2}{SCT_i(1 - R_i^2)} \\ &= \frac{\sigma^2}{\sum_{j=1}^{n}(X_{ji} - \bar{X}_i)^2} \left(\frac{1}{1 - R_i^2}\right)\end{aligned}$$

que es precisamente lo que queríamos demostrar. □

A partir de estos dos resultados podemos profundizar en los diferentes efectos que la multicolinealidad provoca en el modelo. Según [12], la multicolinealidad causa los siguientes problemas:

- **Los estimadores MCO siguen siendo insesgados, pero sus varianzas y covarianzas crecen y se dificulta su estimación precisa.**
  En el teorema 2.1.8 encontramos que la varianza de los estimadores MCO depende tanto de la varianza de los residuales como de los resultados de la regresión auxiliar de una variable sobre las otras del modelo. Vemos que si la correlación entre la variable X_i y las otras variables explicativas del modelo es muy alta, el valor R_i^2 aumentará y con ello la fracción $\frac{1}{1-R_i^2}$ también se hará más grande, por lo cual la varianza del estimador crecerá también.
  Hay que destacar que en el caso de multicolinealidad aproximada, los estimadores MCO siguen siendo insesgados. En efecto, si revisamos en la demostración del teorema 2.1.1, para la parte del insesgamiento no se usó el supuesto 3 de ausencia de multicolinealidad perfecta. Ahora, en el caso de existir multicolinealidad perfecta, el estimador MCO ni siquiera existe, dado que la matriz $\mathbf{X}$ tendrá una columna que es combinación lineal de las otras. En este caso, por álgebra de matrices su determinante valdrá cero, al igual que el de $\mathbf{X}'$. De donde $\mathbf{X}'\mathbf{X}$ no será invertible y el estimador $\hat{\boldsymbol{\beta}}$ no podrá ser calculado.
  Una consecuencia de este incremento de las varianzas de los estimadores es que algunas variables exógenas resultarán no significativas a pesar de que en realidad sí lo sean. Esto sucede porque para probar la significatividad de una variable se usó el estadístico definido en (2.12) que depende de la varianza de $\hat{\beta}_i$. Si ésta crece, el estimador disminuye su valor, y por tanto es más fácil aceptar la hipótesis nula de no significatividad de la variable.

- **A pesar de que el modelo incluya variables no significativas, la bondad de ajuste, dada por el valor R^2, es muy alta.**
  Al presentar multicolinealidad, se vio que aunque algunas variables sean significativas, la prueba de hipótesis usando el estadístico (2.12) suele mostrar no significatividad. Sin embargo, el valor de R^2 será alto y por tanto el estadístico (2.10) también, con lo cual la hipótesis nula de $\beta_1 = \beta_2 = \ldots = 0$ se rechaza. Es decir, tenemos un modelo significativo formado por variables no significativas.

- **Los estimadores MCO son muy sensibles a pequeños cambios en los datos.**
  En el caso de multicolinealidad perfecta vimos que el estimador $\hat{\boldsymbol{\beta}}$ no puede ser calculado pues la matriz $\mathbf{X}$ es no invertible. En el caso de existir multicolinealidad aproximada, la matriz $\mathbf{X}$ sí es invertible, sin embargo, al ser una columna de $\mathbf{X}$ *casi* una combinación lineal de las otras, la matriz tendrá un determinante muy cercano a cero. Y en ese caso, la matriz estaría mal condicionada, con lo cual pequeñas variaciones en los datos podrían causar grandes variaciones en el cálculo de los estimadores.

Para detectar la multicolinealidad se puede recurrir a varios métodos. En este trabajo usaremos tres:

1. **Matriz de correlaciones:** El análisis de la matriz de correlaciones entre las variables en el modelo es un primer paso para la detección de la multicolinealidad. Esta matriz cuadrada de p filas y columnas, que notaremos como $\mathbf{C}$ y sus elementos como c_{ij}, está formada mediante la siguiente fórmula:

$$c_{ij} = corr(X_i, X_j) = \frac{s_{ij}}{s_i s_j}, i, j \in \{1, \ldots, p\}$$

   donde s_{ij} es la covarianza muestral entre las variables X_i y X_j, y s_i y s_j representan las desviaciones estándar de las variables X_i y X_j respectivamente. Esencialmente la estructura de esta matriz es la siguiente:

$$\mathbf{C} = \begin{bmatrix} 1 & c_{12} & c_{13} & \ldots & c_{1p} \\ c_{21} & 1 & c_{23} & \ldots & c_{2p} \\ c_{31} & c_{32} & 1 & \ldots & c_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & c_{p3} & \ldots & 1 \end{bmatrix}$$

   La diagonal principal de la matriz de correlaciones está formada por unos, puesto que la correlación de una variable consigo misma tiene ese valor siempre. Además es simétrica, lo que quiere decir que la parte de la matriz bajo la diagonal principal es igual que la parte sobre esta diagonal. Dado que la correlación entre dos variables siempre está entre 0 y 1, es fácil al revisar la

matriz notar valores de correlación altos que podrían indicar la presencia de multicolinealidad.

Desgraciadamente, este análisis funciona para detectar relaciones entre pares de variables. En caso de existir multicolinealidad entre tres o más variables, el análisis de la matriz ya no es de mucha utilidad.

2. **Factor de inflación de la varianza:** Denominado como FIV (o VIF, por sus siglas en inglés), este valor mide el incremento de la varianza de los coeficientes del modelo $\hat{\beta}_i$ al añadir una nueva variable al modelo. Definiremos como FIV de la variable X_i a la segunda fracción de la ecuación (2.14); es decir

$$FIV_i = \frac{1}{1 - R_i^2}$$

Recordando que el coeficiente de determinación R^2 es el cuadrado del coeficiente de correlación múltiple R, es sencillo observar que mientras aumenta la correlación entre las variables explicativas del modelo, R_i^2 también lo hace y con él FIV_i, y éste crece de forma mucho más pronunciada. La siguiente tabla muestra estos incrementos.

Coef. de correlación R_i	R_i^2	FIV_i
0,00	0,00	1
0,50	0,25	1,33
0,70	0,49	1,96
0,80	0,64	2,78
0,90	0,81	5,76
0,95	0,9025	10,26
0,97	0,9409	16,92
0,99	0,9801	50,25
0,995	0,990025	100
0,999	0,998001	500

Tabla 2.1: Relaciones entre algunos valores de R_i, R_i^2 y FIV_i. Fuente: [12]

Anteriormente mencionamos que un efecto de la multicolinealidad es el incremento de la varianza de los estimadores MCO. EL FIV ayuda a medir dicho efecto. Pero dado que la multicolinealidad es una cuestión de grado y no de presencia o ausencia, ¿en qué punto la multicolinealidad se vuelve un problema? Según [31], si el FIV se encuentra entre 1 y 5, la multicolinealidad no afecta a la estimación del modelo. Si el FIV se encuentra entre 5 y 10, la multicolinealidad puede representar un problema para el cálculo de los coeficientes. Y si el FIV es superior a 10, las estimaciones dadas por (2.6) están completamente sesgadas y la multicolinealidad es un problema grave.

3. **FIV generalizado:** El FIV definido en la sección anterior ayuda a medir el aumento de las varianzas de los estimadores por el ingreso de una nueva variable

al modelo. Un generalización, propuesta en [25], mide el aumento de la varianza por el ingreso de un grupo de variables al modelo lineal. Este factor se conoce como **factor de inflación de la varianza generalizado** (o GVIF, por sus siglas en inglés) y se usa comúnmente con las variables categóricas que tienen más de dos valores posibles. Para dichas variables se usan n variables binarias o *dummy* para representar las $n+1$ categorías de la variable en cuestión.
El GVIF se calcula de la siguiente forma:

$$GVIF = \frac{det(\mathbf{C}_{11}) \times det(\mathbf{C}_{22})}{det(\mathbf{C})} \tag{2.15}$$

donde $\mathbf{C}_{11}$ representa la matriz de correlaciones entre las variables que ingresan al modelo, $\mathbf{C}_{22}$ es la matriz de correlaciones de las variables que ya están dentro del modelo, y $\mathbf{C}$ es la matriz de correlaciones de todas las variables explicativas. Dado que se comparan matrices de diferentes dimensiones, se sugiere que el valor usado sea $GVIF^{1/p_1}$, donde p_1 es el número de variables que ingresan al modelo.

Hemos mencionado algunos métodos para detectar la multicolinealidad en los modelos de regresión lineal. Pero aún no hemos mencionado de qué manera podemos corregir el modelo para eliminarla. Algunas de las formas de corrección son:

- **Eliminar las variables correlacionadas,** pues aportan información similar al modelo, y su eliminación no representará un cambio muy grande en los estadísticos de bondad de ajuste como el R^2 y el $\bar{R}^2$. En este trabajo, en caso de presentarse multicolinealidad, se eliminarán variables con un FIV superior a 5 que indican correlaciones superiores a 0,90 como podemos ver en la tabla 2.1.

- **El análisis de componentes principales (ACP),** una técnica del análisis multivariante que resume la información de varias variables en factores o componentes principales no correlacionados entre sí. Para más información sobre el ACP, se puede consultar [13].

- **Regresión contraída o *ridge*,** una técnica sugerida en [36] para eliminar la multicolinealidad. Consiste en ponderar ciertas variables a partir de técnicas matriciales que reducen la varianza de los estimadores. Para los detalles más técnicos sobre este método, ver [5].

Heterocedasticidad

El supuesto 5 del modelo lineal clásico menciona que los residuos deben tener varianza constante e independiente de las variables explicativas del modelo. Cuando esta condición no se cumple se dice que el modelo presenta **heterocedasticidad.**

Matemáticamente, decimos que

$$V(u_j|X_1, \ldots, X_p) = {\sigma_j}^2, 1 \leq j \leq n.$$

Es decir, la varianza del término residual varía a lo largo de la muestra.
Un ejemplo típico de estimación heterocedástica que se cita en [12] es la predicción de los errores cometidos en cierta disciplina (en este ejemplo, mecanografía) como función del tiempo de práctica. La forma precisa del modelo sería

$$Errores = \beta_1 + \beta_2 * Tiempo\_practica + u$$

Es claro que los errores se reducen conforme el tiempo de práctica aumenta (por tanto $\beta_2 < 0$) pero lo más importante que también su varianza se reduce, pues los errores de personas con un tiempo de práctica menor son más dispersos que los de aquellos con un tiempo de práctica superior, que no presentan tanta variación. Se pueden ver estas ideas en el siguiente gráfico.

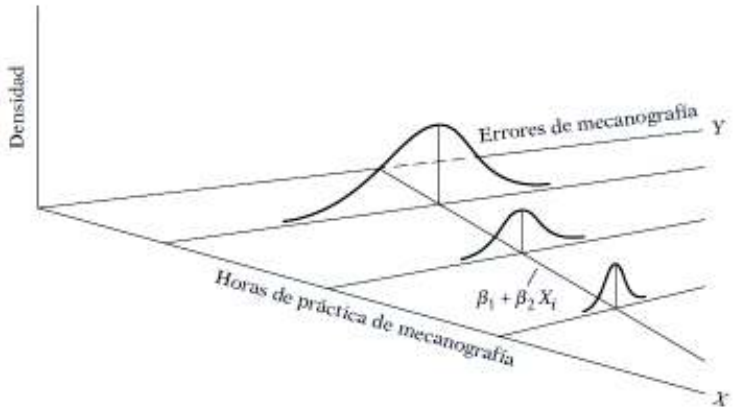


Figura 2.1: Heterocedasticidad del modelo de aprendizaje de mecanografía. Fuente: [12].

La heterocedasticidad en el modelo puede producirse por muchas razones. Según [2], algunas de estas causas son:

- La naturaleza misma del fenómeno a modelar, como en el ejemplo anterior.

- Caundo los datos muestrales no provienen de observaciones individuales sino de promedios de grupos de diferentes tamaños. En estos casos sucede que la varianza dependerá del tamaño del grupo promediado; de forma precisa, la varianza disminuye si el tamaño del grupo aumenta.

- Cuando el modelo deja fuera una variable explicativa importante. Supongamos

que el modelo correctamente especificado es de la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

y se intenta estimar el modelo incorrecto

$$Y = \beta_0 + \beta_1 X_1 + v$$

tendremos que $v = \beta_2 X_2 + u$ con lo cual el residuo de este modelo no cumplirá con los supuestos del modelo lineal clásico. En efecto, y suponiendo que el modelo original sí los cumple, $E(v) = \beta_2 E(X_2)$ y $V(v) = \sigma^2 + (\beta_2 X_2)^2$ con lo cual se violan los supuestos 4 y 5 del modelo lineal clásico.

Si expresamos el modelo de regresión en la forma matricial, la condición de heterocedasticidad es equivalente a expresar que el término de error estocástico tiene una matriz de varianzas y covarianzas igual a $\sigma^2\mathbf{\Sigma}$, donde $\mathbf{\Sigma}$ es una matriz diagonal cuyos términos no son todos iguales a uno, a diferencia de $\mathbf{I_n}$.

Al igual que con la multicolinealidad, la presencia de heterocedasticidad en el modelo no afecta al insesgamiento de los estimadores MCO. En efecto, si se revisa la demostración del teorema 2.1.1, no se usa el supuesto 5 de homocedasticidad en la prueba de insesgamiento del estimador. Sin embargo, la heterocedasticidad afecta a otras propiedades que hemos revisado del estimador MCO.

El primer efecto de la heterocedasticidad es que el estimador dado por (2.7) ya no es el estimador de varianza mínima. En concreto, si suponemos que $V(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{\Sigma}$, al calcular la matriz de covarianzas del estimador $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= V(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}) \\ &= \cancel{V(\boldsymbol{\beta}|\mathbf{X})} + V((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(V(\mathbf{u}|\mathbf{X}))\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{\Sigma})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

que es una expresión más compleja que la hallada al suponer homocedasticidad de los residuales, y que no podemos garantizar que necesariamente sea la menor de todos los estimadores insesgados y lineales.

Otra consecuencia de la heterocedasticidad es que las deducciones sobre las distribuciones de los estadísticos de prueba para la significatividad tanto parcial como completa del modelo ya no son válidas. El teorema 2.10 no se cumple puesto que al ser las varianzas residuales diferentes en cada observación, no se garantiza que la suma de cuadrados siga una distribución χ^2_{n-1} y por tanto sus sumas parciales tampoco lo seguirán al fallar las condiciones previas para aplicar el teorema de Cochran. Esto tambien provoca que el estimador dado para la varianza del error en el teorema

2.1.5 ya no sea insesgado, y por tanto cualquier estimador que dependa de este valor lo será también. En particular, si vemos las varianzas $V(\hat{\beta}_i)$ de los estimadores MCO que fueron dadas tanto en el teorema 2.1.4 como en el teorema 2.1.8, estas dependen directamente del valor de σ^2. Por lo mencionado anteriormente, sustituir σ^2 por $\hat{\sigma^2}$ en estas fórmulas provocaría sesgo en los estimadores de las varianzas. Y como vimos en el teorema 2.1.6, estas varianzas son un factor clave para probar la significatividad de las variables explicativas.
Antes de pasar a las medidas correctivas para este problema daremos una definición.

Definición 2.1.5. Se define como **desviación estándar del estimador** $\hat{\beta}_i$ $(de(\hat{\beta}_i))$ a la raíz cuadrada de $V(\hat{\beta}_i)$, expresada de cualquiera de sus formas. Es decir:

$$\begin{aligned} de(\hat{\beta}_i) &= \sqrt{V(\hat{\beta}_i)} \\ &= \sigma\sqrt{a_{ii}} \\ &= \frac{\sigma}{\sqrt{\sum_{j=1}^{n}(X_{ji} - \bar{X}_i)^2 \left(\frac{1}{1-R_i^2}\right)}} \end{aligned}$$

Si se reemplaza σ^2 por $\hat{\sigma^2}$ en las fórmulas anteriores, obtenemos el **error estándar del estimador** $\hat{\beta}_i$ $(ee(\hat{\beta}_i))$.

Con esta definición, el estadístico (2.11) puede ser formulado como

$$\frac{\hat{\beta}_i - \beta_i}{ee(\hat{\beta}_i)}$$

Pero como vimos anteriormente, este error estándar no se puede usar en caso de heterocedasticidad. Para corregir este problema, en [24] se define el **error estándar robusto a heterocedasticidad** o simplemente **error estándar robusto** como

$$ee(\hat{\beta}_i) = \sqrt{\frac{\sum_{j=1}^{n} \hat{r_{ji}} u_j}{\sum_{j=1}^{n} \hat{r_{ji}}^2}}$$

el cual puede ser usado cuando se tiene evidencia de heterocedasticidad de forma desconocida en el modelo para realizar las pruebas de significancia de variables.
En este trabajo usaremos dos técnicas exploratorias y dos pruebas estadísticas para la detección de la heterocedasticidad:

1. **Método gráfico:** Esta técnica analiza los gráficos de dispersión de los residuos al cuadrado del modelo contra los valores de las variables, tanto la endógena como las exógenas, en busca de patrones sistemáticos que podrían revelar heterocedasticidad. En caso de que los residuos sean homocedásticos, el gráfico de dispersión tomaría la forma de una banda horizontal casi constante. Si se observa patrones en los gráficos, se tendría evidencia de dependencia entre los residuos al cuadrado y las variables; y más aún, se sabría cuál variable es la

causante de dicha dependencia. Para aclarar estas ideas veamos el siguiente gráfico.

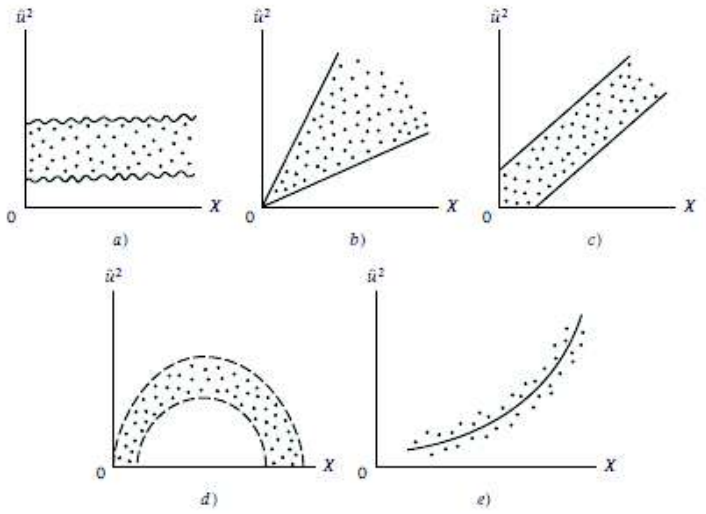


Figura 2.2: Algunas formas posibles para los gráficos de dispersión. Fuente:[12].

Si los residuos son homocedásticos, se esperaría que todos los gráficos de dispersión fueran semejantes a la parte a) del gráfico. Los otros gráficos muestran tendencias claras. En particular, b) y c) muestran tendencias lineales, mientras que d) y e) muestran tendencias cuadráticas.
La evidencia dada por este método sirve tanto como apoyo para posteriores pruebas como para el desarrollo de medidas correctivas para la heterocedasticidad, lo que veremos más adelante.

2. **Prueba de Park:** La prueba de Park es una formalización del método gráfico que en parte usa sus conclusiones y las prueba matemáticamente. Esta prueba sugiere que σ^2 es una función de cierta variable del modelo; exactamente, supone que tiene una forma

$$\sigma_j^2 = \sigma^2 (X_{ji})^m e^{v_j}$$

Tomando logaritmos a ambos lados tenemos

$$ln\ \sigma_j^2 = ln\ \sigma^2 + m\ ln(X_{ji}) + v_j$$

donde v_j es un término de error estocástico. Dado que σ_j^2 es desconocido, se usa como aproximación $\hat{u_j}^2$, quedando la anterior ecuación como

$$ln\ \hat{u_j}^2 = ln\ \sigma^2 + m\ ln(X_{ji}) + v_j$$

el cual es un modelo lineal, y como tal, puede ser estimado mediante MCO.

En caso de que el modelo original resulte ser heterocedástico, se espera que m resulte significativo; si esto no sucede, podemos aceptar que el modelo cumple con el supuesto de homocedasticidad. Para probar la significatividad, se usa el procedimiento mencionado en secciones anteriores.
La prueba de Park tiene un carácter estrictamente exploratorio, pues en [12] se menciona que la estimación puede fallar ya que el término v_j puede ser también heterocedástico. Por tanto, se necesita pruebas o contrastes estadísticos más fuertes, como las que veremos a continuación.

3. **Contraste de Breusch-Pagan:** Como hemos mencionado, la heterocedasticidad supone que la varianza de los residuales posee cierta dependencia con las variables del modelo; esto es, existe una función f tal que

$$V(u|X_1, \ldots, X_p) = f(X_1, \ldots, X_p) \tag{2.16}$$

El contraste de Breusch-Pagan supone que dicha dependencia es un modelo lineal, por tanto la función f mencionada anteriormente sería:

$$f(X_1, \ldots, X_p) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p + v$$

Si el modelo fuera homocedástico, $\theta_1 = \theta_2 = \ldots = \theta_p = 0$ con lo que la significatividad de la regresión sería baja. Mas si el modelo es satisfactorio para explicar la varianza del modelo original, se puede hablar de que existen evidencias de heterocedasticidad. El proceso entonces para realizar este contraste según [39] es:

a) Se estima el modelo (2.1) mediante MCO y se obtienen sus residuales $\hat{u}$.

b) Se calcula una variable nueva $\hat{e}$ de la siguiente manera:

$$\hat{e_j} = \frac{\hat{u_j}^2}{\tilde{\sigma^2}}, 1 \leq j \leq n.$$

donde $\tilde{\sigma^2}$ es un estimador de la varianza residual dado por:

$$\tilde{\sigma^2} = \frac{SCR}{n}$$

c) Se estima el modelo

$$\hat{e} = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p + v$$

mediante MCO.

d) El estadístico de prueba para la hipótesis nula $H_o = \theta_1 = \theta_2 = \ldots = \theta_p = 0$ es $LM = \frac{SCE}{2}$, donde SCE es la suma de cuadrados explicada en la regresión de $\hat{e}$. Este sigue una distribución chi cuadrado con p grados de

libertad:

$$LM \rightsquigarrow \chi^2_p$$

Otros autores [24] mencionan que se puede probar la significatividad de la regresión de $\hat{e}$ usando otros estadísticos. En concreto se menciona al estadístico F de la regresión, que como su nombre indica, sigue una distribución $F_{p,n-p-1}$ y al estadístico $n * R^2_{\hat{e}}$ donde n es el tamaño de la muestra y $R^2_{\hat{e}}$ es el coeficiente de determinación de la regresión en $\hat{e}$. Este estadístico también sigue una distribución chi cuadrado con p grados de libertad.
Dado que se conoce la distribución de los estadísticos (cualquiera de ellos) se puede desarrollar las pruebas de hipótesis adecuadas al nivel de confianza elegido, con lo cual se comprueba si existe o no heterocedasticidad a dicho nivel.

4. **Contraste de White:** La idea detrás de este contraste es similar al anterior; sin embargo, este contraste añade más términos al modelo. De acuerdo a [20] el procedimiento es el siguiente:

   a) Se estima el modelo (2.1) por MCO y se obtienen los residuales $\hat{u}$.

   b) Se efectúa la regresión auxiliar de $\hat{u}^2$ sobre las variables del modelo, sus cuadrados, y sus productos cruzados de segundo orden.

   c) Se calcula el estadístico $n*R^2_{\hat{u}^2}$ que sigue una distribución chi cuadrado con m grados de libertad, donde m es el número de variables en la regresión auxiliar.

   Este contraste es más general que el anterior, pero posee la debilidad de que al añadir muchas variables a la regresión de $\hat{u}^2$ los grados de libertad aumentan demasiado. En [24] se muestra un procedimiento modificado, que efectúa la regresión auxiliar sobre los valores predichos por el modelo original $\hat{Y}$ y sus cuadrados $\hat{Y}^2$:

   $$\hat{u}^2 = \theta_0 + \theta_1\hat{Y} + \theta_2\hat{Y}^2 + v$$

   Este proceso disminuye los grados de libertad y puede ser usado como complemento al anterior. Al igual que en el contraste de Breusch-Pagan, la prueba de hipótesis se puede realizar con el estadístico sugerido o con el estadístico F de la regresión.

Sea cual sea el método usado para detectar la heterocedasticidad, los contrastes pueden proporcionar pistas sobre cuál podría ser la variable que causa este fenómeno. Una vez que se tienen las pruebas de presencia de heterocedasticidad se debe proceder a corregir el modelo y su estimación. Existen dos métodos muy comunes para eliminar la heterocedasticidad:

1. **Transformación de variables.** Consiste en tomar o bien la variable dependiente Y o algunas de las independientes $X_i, 1 \leq i \leq p$ y aplicarles una transformación algebraica que solucione el problema de la varianza. Para conocer

qué transformación sería más adecuada para cada modelo se puede recurrir a la información proporcionada por las pruebas mencionadas anteriormente, en especial del método gráfico o de la prueba de Park. No obstante, sea cual sea la variable a ser transformada, este procedimiento en general se realiza bajo prueba y error; realizando una transformación y revisando mediante los contrastes anteriores si se ha resuelto el problema sobre las varianzas de los residuos. Simbolicemos con Y' la variable transformada que reemplaza a la variable original Y. Las transformaciones más usadas suelen ser:

- la transformacion logarítmica ($Y' = ln(Y)$),
- la transformación cuadrática ($Y' = Y^2$),
- la transformación inversa ($Y' = \frac{1}{Y}$).

Una familia de transformaciones usada de forma frecuente en modelos econométricos es la familia de **transformaciones Box-Cox,** de las que las transformaciones anteriores son casos particulares. Estas transformaciones parten del criterio de minimizar la suma de cuadrados de los residuales del modelo. Para más información sobre esta familia de transformaciones, se puede revisar [18].

2. **Estimación mediante mínimos cuadrados ponderados (MCP).** En ciertas ocasiones se puede estimar la forma de la función de la varianza residual expresada en (2.16). En estos casos se puede ponderar las diferentes observaciones de tal manera que los residuos sean homocedásticos.
Para ejemplificar la situación supongamos que el modelo (2.16) se puede expresar como

$$V(u_j|X_1, \ldots, X_p) = \sigma^2 h_j,\ 1 \leq j \leq n. \tag{2.17}$$

donde h_j es una función de las variables exógenas y endógena que se puede calcular para cada observación de la muestra. Supongamos también que el modelo original cumplía con todos los supuestos del modelo lineal clásico exceptuando el supuesto 5 de homocedasticidad. Entonces si ponderamos cada observación del modelo original dividiendo para $\sqrt{h_j}$ (que conocemos que existe, puesto que las varianzas siempre son positivas y por tanto h_j también lo es):

$$\begin{aligned} \frac{Y_j}{\sqrt{h_j}} &= \hat{\beta}_0 \frac{1}{\sqrt{h_j}} + \hat{\beta}_1 \frac{X_{j1}}{\sqrt{h_j}} + \hat{\beta}_2 \frac{X_{j2}}{\sqrt{h_j}} + \ldots + \hat{\beta}_p \frac{X_{jp}}{\sqrt{h_j}} + \frac{\hat{u_j}}{\sqrt{h_j}} \\ \tilde{Y_j} &= \hat{\beta}_0 H_j + \hat{\beta}_1 \tilde{X_{j1}} + \hat{\beta}_2 \tilde{X_{j2}} + \ldots + \hat{\beta}_p \tilde{X_{jp}} + \hat{v_j} \end{aligned} \tag{2.18}$$

obtenemos un nuevo modelo de regresión lineal para la variable transformada $\tilde{Y}$ definida mediante

$$\tilde{Y_j} = \frac{Y_j}{\sqrt{h_j}}, 1 \leq j \leq n$$

en función de las variables $\tilde{X_i}, 1 \leq i \leq p$ que se obtienen transformando las

variables originales del modelo siguiendo la fórmula

$$\tilde{X_{ji}} = \frac{X_{ji}}{\sqrt{h_j}}, 1 \leq i \leq p, 1 \leq j \leq n$$

y la nueva variable H definida por

$$H_j = \frac{1}{\sqrt{h_j}}, 1 \leq j \leq n$$

El nuevo modelo no posee término independiente a diferencia del original.
Este modelo es lineal en los parámetros con lo cual se cumple el supuesto 1 del modelo lineal clásico. La aleatoriedad de la muestra no ha cambiado, con lo que también se cumpliría el supuesto 2. Dado que el modelo original no tenía multicolinealidad, el nuevo modelo tampoco la tendrá por lo cual cumple el supuesto 3.
Para corroborar el supuesto 4 calcularemos la esperanza de los residuales $\hat{v}$ para valores fijos de las variables explicativas originales $X_i,\ 1 \leq i \leq p$. Es claro que si dichas variables tienen valores fijos, las variables transformadas $H, \tilde{X_1}, \tilde{X_2}, \ldots, \tilde{X_p}$ también tendrán valores fijos. Recordando que $\hat{v_j} = \frac{\hat{u_j}}{\sqrt{h_j}}$:

$$E(\hat{v_j}) = E\left(\frac{\hat{u_j}}{\sqrt{h_j}}\right) = \frac{1}{\sqrt{h_j}} E(\hat{u_j})$$

donde $\sqrt{h_j}$ sale como constante pues también tiene un valor fijo. Y dado que el modelo original cumplía con el supuesto 4, $E(\hat{u_j}) = 0$. Por tanto, reemplazando en la ecuación anterior tenemos que:

$$E(\hat{v_j}) = 0$$

lo que demuestra que el modelo transformado cumple el supuesto 4.
Ahora verificaremos si el modelo nuevo presenta homocedasticidad a diferencia del original que era heterocedástico. Calcularemos entonces la varianza de los residuales $\hat{v}$ para valores fijos de las variables $X_i,\ 1 \leq i \leq p$. Al igual que en la comprobación del supuesto anterior usaremos el hecho de que $\hat{v_j} = \frac{\hat{u_j}}{\sqrt{h_j}}$:

$$V(\hat{v_j}) = V\left(\frac{\hat{u_j}}{\sqrt{h_j}}\right) = \left(\frac{1}{\sqrt{h_j}}\right)^2 V(\hat{u_j})$$

Por propiedades de la varianza $\sqrt{h_j}$ sale pero elevada al cuadrado. Usando (2.17):

$$\left(\frac{1}{\sqrt{h_j}}\right)^2 V(\hat{u_j}) = \frac{1}{h_j}\sigma^2 h_j = \sigma^2$$

de donde la varianza residual es constante, por tanto el modelo transformado es homocedástico.

Al cumplir el modelo (2.18) los supuestos del modelo lineal clásico se puede utilizar las ecuaciones (2.6) para estimar sus parámetros $\hat{\beta}_i, 0 \leq i \leq p$ mediante MCO. Esta estimación para las variables transformadas se conoce como **estimación por mínimos cuadrados ponderados (MCP)** para el modelo original (2.1).

Es importante que para que esta estimación sea realizada con exactitud y elimine la heterocedasticidad la expresión h_j de la ecuación (2.17) debe estar acorde a la estructura real de la varianza. Para estimar esta estructura se puede recurrir a estudios previos sobre el fenómeno a modelar o a la información que se obtiene al realizar los contrastes de Park, Breusch-Pagan o White como guía para idear la estructura más plausible para la varianza de los residuales y realizar la ponderación de las observaciones. Una vez realizada la estimación MCP se puede volver al modelo original multiplicando el modelo ponderado por $\sqrt{h_j}$.

La estimación MCP es un caso particular de la **estimación por mínimos cuadrados generalizados (MCG),** que como su nombre lo indica, es una generalización del método MCO de estimación que resuelve ciertos problemas de especificación de modelos y relaja ciertas condiciones para los cálculos de los estimadores. Para más información sobre este tema se puede acudir a [2].

Normalidad de residuos

Se mencionó anteriormente que el supuesto de normalidad de residuos no es necesario para la demostración del teorema de Gauss-Markov. Sin embargo, se probó también que para conocer las distribuciones muestrales de los estimadores $\hat{\beta}_i, 0 \leq i \leq p$ del modelo para realizar pruebas de hipótesis o construir intervalos de confianza este supuesto es imprescindible.

Existen casos en los que se puede relajar el supuesto de normalidad de residuos y las pruebas de hipótesis se pueden seguir realizando. Únicamente se exige que la muestra sea grande. Para ello se tiene el siguiente teorema, mencionado en [24] cuya demostración va más allá del alcance de este trabajo.

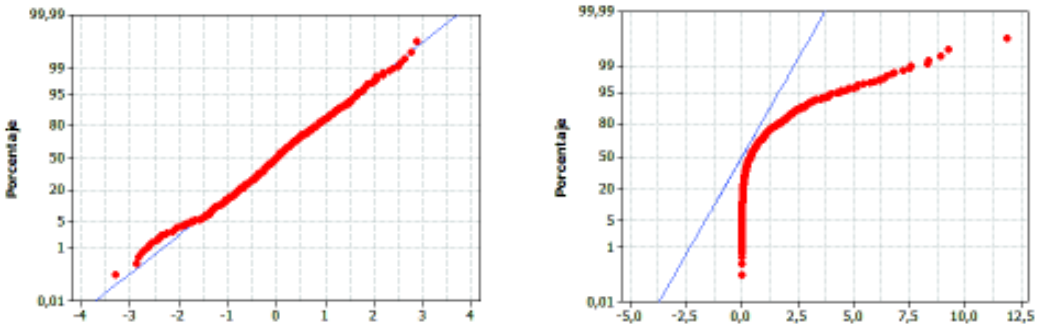
Teorema 2.1.9 (Normalidad asintótica)**.** *Dados los supuestos 1 a 5 del modelo lineal clásico se cumple que:*

1. $\sqrt{n}(\hat{\beta}_i - \beta_i)$ *está distribuida de forma asintóticamente normal con media cero y varianza* σ^2/a_j^2*, donde* $a_j^2 = plim\left(\frac{1}{n}\sum_{j=1}^{n} \hat{r_{ji}}\right)$ *donde plim significa el límite en probabilidad y* $\hat{r_{ji}}$ *es el residuo de la regresión auxiliar de la variable* X_i *asociada al estimador* β_i *contra todas las demás variables exógenas.*

2. $\hat{\sigma^2}$ *es un estimador consistente de* σ^2*, donde* $\hat{\sigma^2}$ *viene dado por el teorema 2.1.5.*

3. $(\hat{\beta}_i - \beta_i)/ee(\hat{\beta}_i)$ *está distribuida de forma asintóticamente normal con media cero y varianza 1 para todo* i *entre* 1 *y* p.

La forma más sencilla de revisar la normalidad de los residuos del modelo es a través de los **gráficos de probabilidad normal**. Se muestran ejemplos de estos gráficos en la siguiente figura.



(a) Datos normales (b) Datos no normales

Figura 2.3: Gráfico de probabilidad normal para datos normales y datos que no lo son. Fuente: elaboración propia.

Los gráficos de probabilidad normal presentan una línea central que representa la distribución normal ideal. Mientras más ajustados estén los datos observados a dicha línea, es más seguro que provengan de una distribución normal. Si bien este gráfico proporciona evidencias para afirmar o descartar la normalidad de los residuos del modelo, existen pruebas estadísticas diseñadas para probar la normalidad de forma más categórica. En este trabajo usaremos dos:

1. La **prueba de Anderson-Darling**, cuyo fundamento teórico se puede ver en [40]. La hipótesis nula de esta prueba se define como que los datos siguen una ley de probabilidad teórica con función de distribución $F(x)$. Para realizar esta prueba se deben ordenar los residuales $\hat{u_j}$ de tal manera que $\hat{u_1} \leq \hat{u_2} \leq \hat{u_3} \leq \ldots \leq \hat{u_n}$; es decir, de menor a mayor. Luego se calcula el estadístico A^2 mediante la fórmula

$$A^2 = -N - \frac{1}{N}\sum_{j=1}^{N}(2j-1)\left[ln\ F(u_j) + ln\ (1 - F(u_{N+1-j}))\right]$$

donde $F(u_j)$ es la función de distribución a la que se supone pertenecen los datos. Para la distribución normal estándar se tiene que

$$F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{w^2}{2}}dw$$

Una vez calculado el estadístico se lo compara con los valores críticos para el nivel de confianza deseado. En la siguiente tabla se dan estos valores para la

distribución normal.

Nivel de confianza	Valor crítico
0,10	0,632
0,05	0,751
0,01	0,870

Tabla 2.2: Valores críticos de A^2 para la distribución normal. Fuente: [6].

Si el estadístico calculado supera al valor crítico deseado se rechaza la hipótesis nula.

2. La **prueba de Jarque-Bera**, propuesta en [8], es una prueba muy sencilla de aplicar para detectar si los residuales siguen una distribución normal. Para esta prueba se debe calcular el **coeficiente de asimetría** S y la **curtosis** K de los datos a evaluar. El coeficiente de asimetría de una variable X se calcula mediante la fórmula

$$S = \frac{1}{ns^3} \sum_{j=1}^{n} (X_j - \bar{X})^3$$

y la curtosis se calcula con

$$K = \frac{1}{ns^4} \sum_{j=1}^{n} (X_j - \bar{X})^4$$

donde n es el tamaño de la muestra, $\bar{X}$ es el promedio muestral y s es la desviación estándar muestral de la variable X. A partir de estos datos se calcula el estadístico de prueba JB de la siguiente forma:

$$JB = n \left[\frac{S^2}{6} + \frac{(K-3)^2}{24} \right]$$

Este estadístico asintóticamente sigue una distribución chi cuadrado de 2 grados de libertad.

La elección de la prueba de normalidad depende de la cantidad de observaciones. Si bien ambas pruebas son asintóticas, Anderson-Darling funciona bien independientemente del tamaño de la muestra, mientras sea superior a 20; mientras que Jarque-Bera va mejorando su desempeño para muestras más grandes. Además la prueba de Anderson-Darling detecta mejor desviaciones situadas hacia las colas de la distribución normal en comparación con Jarque-Bera. Para más información sobre estos resultados y comparaciones entre estas y otras pruebas de normalidad se puede consultar [7].

2.2. Algoritmo `k-means`

La **agrupación de datos** (o data clustering, en inglés) es la creación de conjuntos o **clústeres** de individuos por semejanza, cercanía u otras propiedades comunes mediante técnicas o algoritmos matemáticos. Según [4], el agrupamiento se utiliza para tres fines principales:

- **Encontrar estructuras subyacentes:** para mejorar la visualización de los datos, identificar anomalías y encontrar detalles no perceptibles a simple vista.
- **Clasificación natural:** identificar la similitud entre diferentes tipos de organismos a través de sus características, para construir relaciones filogenéticas.
- **Compresión:** para organizar datos y representarlos por medio de los puntos prototípicos de sus respectivos clústeres.

Existen numerosas técnicas destinadas a realizar las tareas de agrupamiento de datos. Uno de los algoritmos más utilizados es el **algoritmo `k-means`**. Este algoritmo es muy popular debido a la sencillez de la idea detrás de su uso y la simplicidad de su implementación computacional.
Dados $X = \{x_1, x_2, \ldots, x_N\}$ N puntos en $\mathbb{R}^d$, el objetivo de `k-means`, de acuerdo a [4], es encontrar un conjunto de K clústeres $C = \{c_j, j = 1, \ldots, K\}$ que minimice las distancias al cuadrado entre los puntos de un clúster y su punto medio o **centroide** μ_j. La función de distancias dentro de cada clúster se define como:

$$D(c_j) = \sum_{x_i \in c_j} [d(x_i, \mu_j)]^2$$

donde $d(x, y)$ es la distancia entre dos puntos en $\mathbb{R}^d$ que puede ser definida de varias maneras, aunque la más común sea la distancia euclidiana. Por tanto, lo que se espera lograr con el algoritmo es minimizar la función

$$D(C) = \sum_{j=1}^{K} \sum_{x_i \in c_j} [d(x_i, \mu_j)]^2$$

Para ello, el algoritmo sigue los siguientes pasos:

1. Se selecciona un grupo aleatorio de K puntos que serán los centroides iniciales de los clústeres.
2. Se asigna cada uno de los N puntos al clúster c_j si la distancia al respectivo centroide μ_j es la menor posible.
3. Se recalcula el centroide de cada clúster a partir de los puntos que fueron asignados a él en el paso 2.

4. Se repiten los pasos 2 y 3 hasta que se estabilicen las asignaciones de los puntos a los clústeres.

Es evidente que para el correcto desempeño del algoritmo se debe partir de una asignación precisa de los centroides iniciales y del valor de K. Para el primer punto, el enfoque tradicional es escoger aleatoriamente los K centroides originales de entre los N puntos que van a ser particionados. Sin embargo, en [9] se propone utilizar una modificación al algoritmo `k-means` conocida como `k-means++`, que mejora los resultados del algoritmo original al seleccionar los puntos originales por medio de una ponderación probabilística. Antes de explicar esta ponderación, definamos $I(x)$ de la siguiente manera:

$$I(x) = \min_{c_j} d(x, c_j)$$

Es decir, $I(x)$ es la mínima distancia de un punto $x \in X$ al conjunto de centroides. Con esta definición en mente, `k-means++` toma los siguientes pasos:

1. Se selecciona al azar un elemento de X para ser el primer centroide c_1.

2. A cada punto $x \in X$ se le asigna una probabilidad $P(x)$ definida por

$$P(x) = \frac{I(x)^2}{\sum_{x \in X} I(x)^2}$$

   Por tanto, a mayor distancia de un punto x a c_1, mayor probabilidad de elección como nuevo centroide. Además, $P(c_1) = 0$, con lo que no se puede repetir centroides.

3. Se selecciona un nuevo centroide c_2 de los x, con probabilidad $P(x)$.

4. Se repite los pasos 2 y 3 hasta seleccionar los K centroides iniciales.

5. Se ejecuta `k-means` de la forma usual.

Para optimizar la elección de K se pueden seguir muchas estrategias diferentes. Para una revisión a profundidad de estas estrategias, se puede consultar [38]. Las más usadas y sencillas de aplicar son:

- **Criterio sencillo:** La regla más sencilla a aplicar para definir K es la siguiente:

$$K = \sqrt{\frac{N}{2}}$$

  A pesar de su simplicidad, esta regla no toma en cuenta la posible distribución de los puntos y puede ser errónea, mas es una aproximación válida en la mayoría de casos.

- **Método del codo:** Este método analiza la función $D(C)$, que en este caso se suele conocer como **función de inercia,** para un conjunto de posibles valores

de K, y escoge el valor para el cual la función no tiene ya descensos muy radicales; en otras palabras, escoge el punto en el cual la función alcanza un *codo*, de ahí su nombre. Para dar una idea sobre el uso de este método se muestra a continuación un ejemplo de gráfico de la función $D(C)$ dependiendo de K.

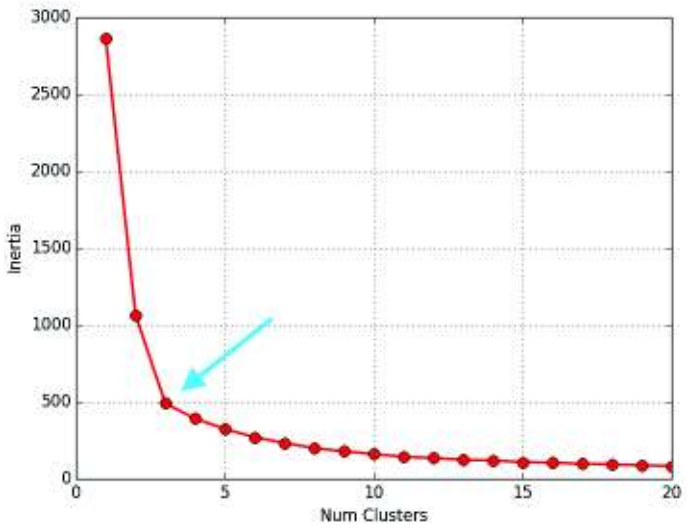


Figura 2.4: Gráfico de la función de inercia y su *codo*. Fuente: [34].

En el gráfico anterior el *codo* de la función de inercia se alcanza cuando $K = 3$ y por tanto ese sería el valor óptimo de K para ejecutar `k-means`.
Si bien este método es una mejora de la regla anterior, podría suceder que el codo de la función no esté bien definido o que existan múltiples valores posibles para él.

- **Algoritmos mejorados:** Se han propuesto algunas mejoras posibles a `k-means` que analizan el desempeño de los valores posibles de K de acuerdo a medidas estadísticas. Uno de ellos el algoritmo `X-means`, propuesto en [10], que analiza los resultados de `k-means` para diferentes valores de K mediante el criterio de información bayesiano (BIC, por sus siglas en inglés). Otra opción es el algoritmo `G-means`, que prueba que cada clúster siga una distribución gaussiana $d-$dimensional y escoge el valor de K en el que mejor se cumpla esta hipótesis. Para más información sobre `G-means`, se puede consultar [19].

Capítulo 3

Desarrollo del modelo de regresión para el rendimiento postcosecha del cacao ecuatoriano

Se mencionó en capítulos anteriores que existe un interés renovado en el cultivo y producción del cacao en el Ecuador. Gracias el trabajo de algunas entidades gubernamentales (MAGAP e INEC) se poseen actualmente grandes cantidades de datos acerca de muchas plantaciones en el país y las personas que las cultivan. Entre ellas, el cacao.

Entre los muchos procesos relacionados con el cultivo y producción de cacao, en este trabajo se desea hacer énfasis en el rendimiento postcosecha de forma cuantitativa, es decir, dejando aparte ciertas variables como sabor, aroma, que si bien son importantes en la percepción del cacao en el mercado, ya han sido estudiadas en otros trabajos. Lo que nos interesa en particular es la reducción de masa que experimenta el cacao tras ser sometido a los procesos postcosecha (fermentación y secado). Es claro que para conveniencia de los productores se espera que dicha reducción no sea muy grande, pues sería una pérdida en el momento de la comercialización del cacao seco. Además como parte de la política agropecuaria ecuatoriana se espera aumentar los rendimientos del cacao ecuatoriano, como se puede ver en [30], con lo que conocer los factores que ayudarán a cumplir esta meta es también de suma importancia para el MAGAP. Actualmente, para medir el rendimiento postcosecha se utiliza el **índice de mazorca**, que se define como la cantidad de mazorcas necesarias para generar un kilo de cacao seco. Sin embargo, este índice no toma en cuenta la enorme variabilidad existente entre los tamaños de mazorcas en las variedades de cacao cultivadas en el país. Además, este índice se calcula a partir de 2 mediciones: la primera se realiza en el momento de la cosecha del cacao y la segunda después de la fermentación y el secado. La desventaja de esta metodología es que si no se acierta en la duración del proceso postcosecha, la segunda medición no necesariamente se realizará en el momento exacto, con lo cual se perdería información necesaria para conocer el ren-

dimiento de las diferentes granjas cacaoteras.
El objetivo del modelo es estimar el rendimiento postcosecha a partir de variables socioeconómicas y geográficas del agricultor y sus métodos de cultivo. Mediante este modelo hallaremos las posibles determinantes de las diferencias de rendimiento entre los productores, y con ellas se podrá proponer métodos para mejorar el rendimiento postcosecha de estos. Por otro lado, las instituciones gubernamentales conocerán las diferentes regiones en las cuales el rendimiento postcosecha del cacao es menor o mayor, y podrán ajustar sus políticas de acuerdo a dicha información.
Para construir este modelo, primero hablaremos de la obtención de datos muestrales realizada por el MAGAP. Luego definiremos las variables exógenas y endógena del modelo, calcularemos sus parámetros, validaremos los supuestos necesarios y discutiremos e interpretaremos sus resultados.

3.1. Obtención de los datos muestrales

El proceso de extracción de la muestra mencionado en esta sección fue calculado e implementado por los expertos del MAGAP antes de la realización de este trabajo. De acuerdo a cifras del MAGAP calculadas a partir de mapas satelitales, en 2016 existían 547866 hectáreas cultivadas con cacao en el país. Para estimar el número de productores de cacao en esa área, se conoce de estudios previos que para cualquier cultivo la extensión promedio de una finca agrícola es de 4 hectáreas. A esta extensión se la conoce como **unidad de producción agropecuaria (UPA),** y es independiente del producto cultivado.
Tomando esta estimación, podemos calcular la población total de agricultores de cacao en el país.

$$\frac{547866\ Ha}{4\ Ha/agric.} = 136966,5\ agric. \approx 136967\ agricultores$$

Por tanto se estimó que existen 136967 cultivadores de cacao en el Ecuador, lo que se consideró como la población de este estudio. Bajo la condición de que ningún cantón productor podía tener menos de 4 puntos muestrales, se escogió una muestra de 569 agricultores distribuidos aleatoriamente por el país. A estos agricultores, entre agosto y noviembre del 2016, se les aplicó un cuestionario de 45 preguntas sobre sus datos socioeconómicos, los datos de su hacienda y sus métodos de cultivo de cacao. Después de ello se escogió un número de mazorcas maduras (de 5 a 20), se midió el peso, longitud, diámetro y espesor de cada una y se extrajo y se contó sus semillas. Después se procedió a fermentar las semillas en baba por 15 días por el método de montones, luego se secó en tendales al sol y se pesó el cacao seco resultante. De estas 569 observaciones se eliminaron 4 por tener datos inconsistentes (el peso seco era superior al peso en baba), por lo que el tamaño de muestra definitivo del estudio es de 565 agricultores.

3.2. Definición de las variables del modelo

3.2.1. Variable dependiente

La variable dependiente Y en este estudio será el **coeficiente de transformación postcosecha (CTP)**, definido de la siguiente forma:

$$CTP = \frac{peso\ seco}{peso\ en\ baba}$$

Este coeficiente es un número adimensional que oscila entre 0 y 1 y podría entenderse como el porcentaje en masa de cacao en baba que queda despúes de los procesos de fermentación y secado. Mientras más alto sea su valor, mejor es el rendimiento postcosecha de la finca.

3.2.2. Variables independientes

A partir de las 45 preguntas del cuestionario y las 6 variables medidas por los técnicos se obtuvieron 72 variables (15 cuantitativas y 57 cualitativas) candidatas a usarse en el modelo. De estas, se descartaron las variables sobre fertilización puesto que los efectos de los diferentes elementos del fertilizante (N, P, K, Mg) sobre la plantación dependen del tipo de suelo y su contenido en dichos elementos, y esta información no es conocida.

Clusterización provincial

Se calculó el CTP promedio para cada una de las 18 provincias y 99 cantones muestreados. Con estos valores se procedió a agrupar en clústeres mediante el algoritmo `k-means++`, eligiendo el valor apropiado de k a partir del método gráfico o del codo, definido en la sección 2.2.

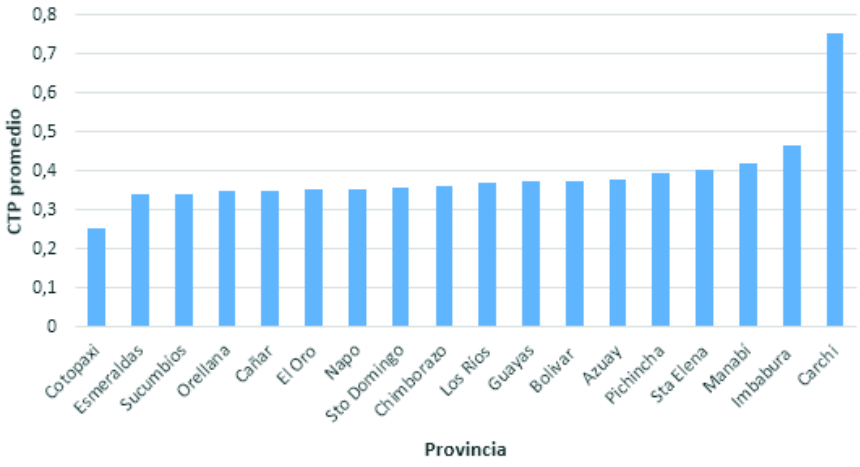


Figura 3.1: CTP promedio por provincia.

La figura 3.1 muestra los CTP promedio para cada provincia. A partir de estos datos, se tomó valores de k entre 2 y 12 y se ejecutó `k-means`, específicamente en su variante `k-means++` para la asignación óptima de centroides, para cada uno de los valores de k. También se calculó la función de inercia $D(C)$ para cada k. La siguiente figura muestra los valores de esta función para los k tomados.

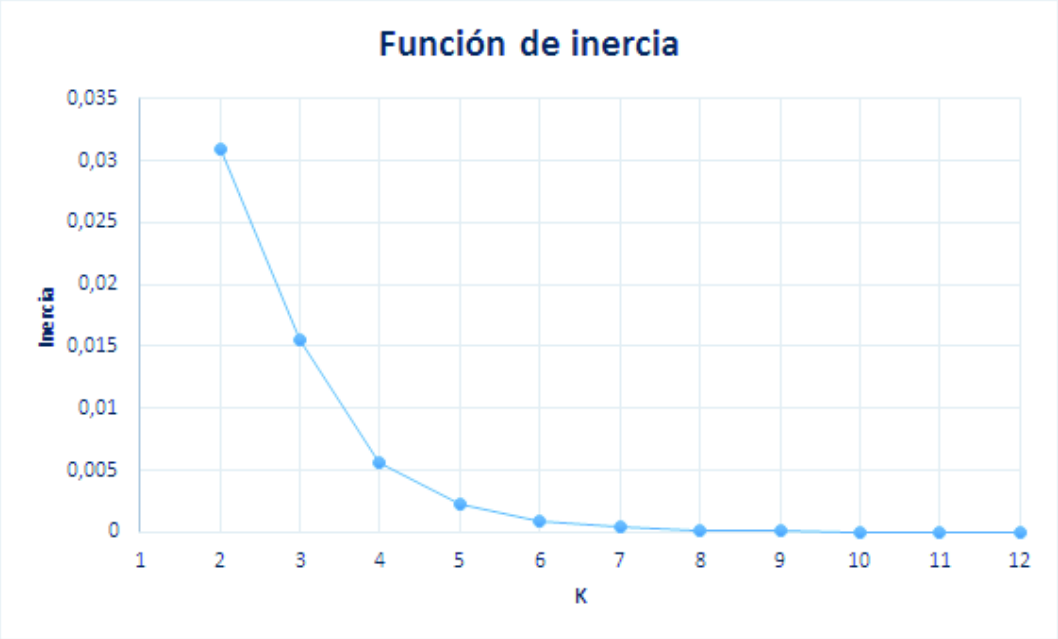


Figura 3.2: Función de inercia para la clusterización de provincias.

En la figura 3.2 vemos que el codo de la función de inercia se alcanza para $k = 4$ que será el valor que usaremos para el algoritmo. La clusterización elegida entonces se muestra en la tabla 3.1 y en la figura 3.3.

Clúster	**Provincias**
Clúster 1	Cotopaxi
Clúster 2	Esmeraldas, Sucumbíos, Orellana, Cañar, Napo, Sto. Domingo, Chimborazo, Los Ríos, Guayas, Bolívar, Azuay, Pichincha, Sta. Elena
Clúster 3	Manabí, Imbabura
Clúster 4	Carchi

Tabla 3.1: Clústeres de provincias por rendimiento.

Vemos que hay 4 provincias que destacan en su rendimiento, tanto Cotopaxi por su rendimiento bajo, como Manabí, Imbabura y Carchi, por su rendimiento alto. Las demás provincias quedan en el clúster de rendimiento intermedio.

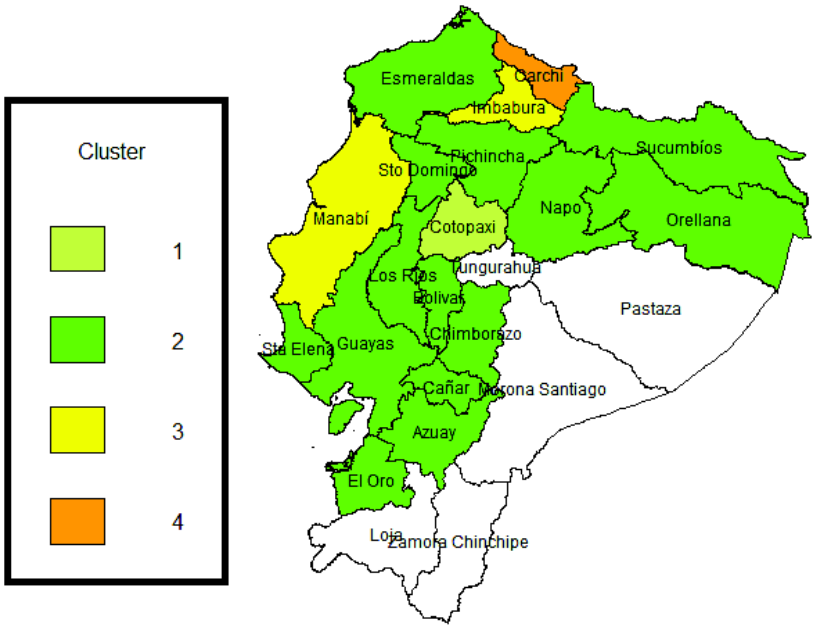


Figura 3.3: Mapa de la clusterización provincial.

Clusterización cantonal

Se siguió un procedimiento similar al anterior para la agrupación de cantones por su CTP promedio. Las figuras 3.4 y 3.5 muestran el valor del CTP para cada uno de los 99 cantones de la muestra usada.

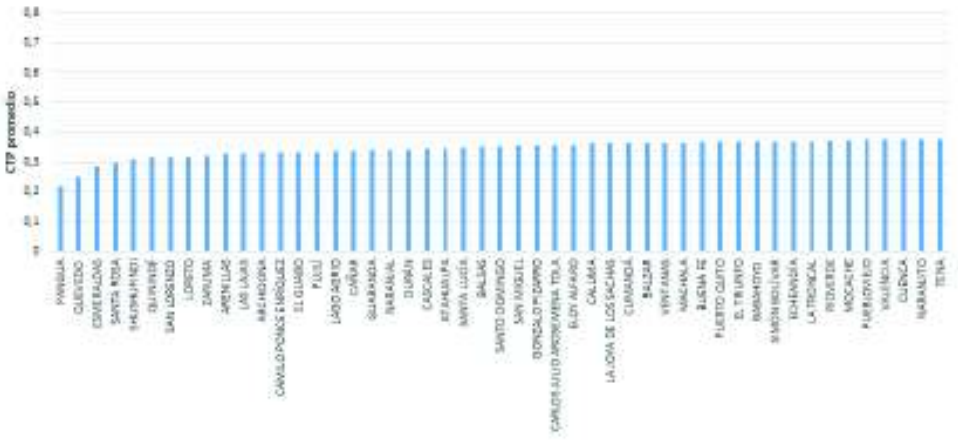


Figura 3.4: CTP promedio por cantón, primera parte.

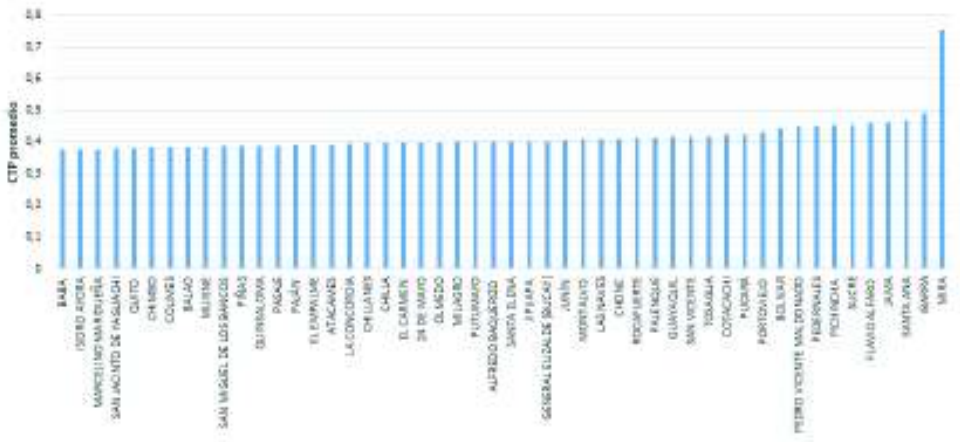


Figura 3.5: CTP promedio por cantón, segunda parte.

Las observaciones de la provincia de Carchi provienen de un solo cantón (Mira) por lo cual no se lo toma en cuenta al realizar la clusterización cantonal puesto que redundaría con la clusterización de provincias. Tenemos entonces 98 cantones en total. Tomando valores de k desde 2 hasta 25, se aplicó el algoritmo `k-means++` y se calculó la función de inercia para cada valor de k. Para obtener el valor óptimo de k se muestra la función de inercia en la figura 3.6.

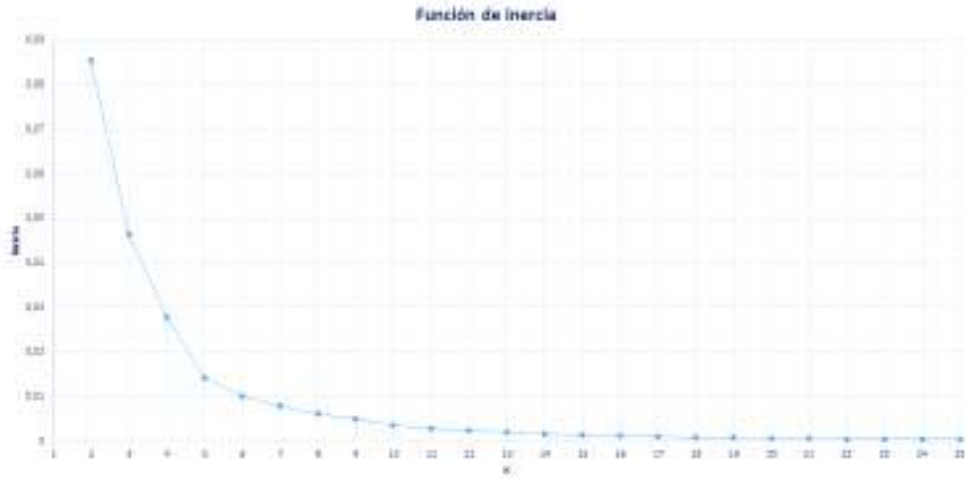


Figura 3.6: Función de inercia para la clusterización de cantones.

El codo se alcanza en $k = 5$, por lo cual se usarán 5 clústeres para la agrupación cantonal. Las asignaciones se muestran en la tabla 3.2 y en la figura 3.7.
Una vez obtenidas estas agrupaciones geográficas, se definen 9 variables cualitativas más que representarán la pertenencia de cada granja a su respectivo clúster provincial y cantonal. Con lo cual se tiene un total de 15 variables cuantitativas y 66 cualitativas para el modelo.

Clúster	Cantón
Clúster 1	Pangua, Quevedo, Esmeraldas.
Clúster 2	Santa Rosa, Shushufindi, Quinindé, San Lorenzo, Loreto, Zaruma, Arenillas, Las Lajas, Archidona, Camilo Ponce Enríquez, El Guabo, Pujilí, Lago Agrio, Cañar, Guaranda, Naranjal, Durán, Cascales, Atahualpa, Santa Lucía.
Clúster 3	Balsas, Santo Domingo, San Miguel, Gonzalo Pizarro, Carlos Julio Arosemena Tola, Eloy Alfaro, Caluma, La Joya de los Sachas, Cumandá, Balzar, Ventanas, Machala, Buena Fe, Puerto Quito, El Triunfo, Babahoyo, Simón Bolívar, Echeandía, La Troncal, Rioverde, Mocache, Puebloviejo, Valencia, Cuenca, Naranjito, Tena, Baba, Isidro Ayora, Marcelino Maridueña, San Jacinto de Yaguachi, Quito, Chimbo, Colimes, Balao, Muisne, San Miguel de los Bancos, Piñas, Quinsaloma.
Clúster 4	Pasaje, Paján, El Empalme, Atacames, La Concordia, Chillanes, Chilla, El Carmen, 24 de Mayo, Olmedo, Milagro, Putumayo, Alfredo Baquerizo, Santa Elena, Jipijapa, General Elizalde (Bucay), Junín, Montalvo, Las Naves, Chone, Rocafuerte, Palenque, Guayaquil, San Vicente, Tosagua, Cotacachi, Pucará, Portoviejo.
Clúster 5	Bolívar, Pedro Vicente Maldonado, Pedernales, Pichincha, Sucre, Flavio Alfaro, Jama, Santa Ana, Ibarra.

Tabla 3.2: Clústeres de cantones por rendimiento.

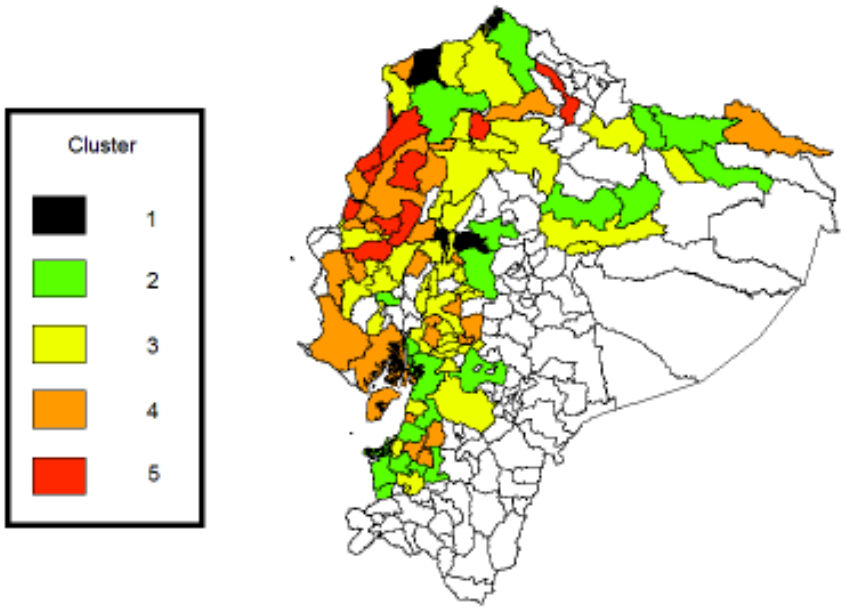


Figura 3.7: Mapa de la clusterización cantonal.

3.3. Estimación de los parámetros

3.3.1. Modelo 1

A partir de las pruebas de significatividad mencionadas en la sección 2.1.3 se obtuvo 8 variables para el modelo de regresión con un p-valor menor a 0,05. En la siguiente tabla se especifican dichas variables (además del intercepto) con sus respectivos coeficientes, error estándar, estadístico t y el p-valor de este estadístico.

Variable	**Coeficiente**	**Err. estándar**	**Estad. t**	**P-valor**
Intercepto	0,458576	0,006754	67,89	0,000
ClustProv4	0,32887	0,02721	12,09	0,000
ClustCant1	-0,19725	0,01256	15,71	0,000
ClustCant2	-0,107105	0,008493	-12,61	0,000
ClustCant3	-0,064620	0,007815	-8,27	0,000
ClustCant4	-0,037958	0,007718	-4,92	0,000
ImbIbarLit	0,26281	0,02696	9,75	0,000
ImbIbarCar	-0,11590	0,02373	-4,88	0,000
Variedad	-0,033819	0,004159	-8,13	0,000

Tabla 3.3: Variables regresoras para el CTP.

Describiremos ahora cada una de estas variables.

1. **ClustProv4:** Variable dummy, creada mediante la clusterización provincial.

$$ClustProv4 = \begin{cases} 1 & \text{si la provincia del productor pertenece al Clúster 4.} \\ 0 & \text{caso contrario.} \end{cases}$$

2. **ClustCant1:** Variable dummy, creada mediante la clusterización cantonal.

$$ClustCant1 = \begin{cases} 1 & \text{si el cantón del productor pertenece al Clúster 1.} \\ 0 & \text{caso contrario.} \end{cases}$$

3. **ClustCant2:** Variable dummy, creada mediante la clusterización cantonal.

$$ClustCant2 = \begin{cases} 1 & \text{si el cantón del productor pertenece al Clúster 2.} \\ 0 & \text{caso contrario.} \end{cases}$$

4. **ClustCant3:** Variable dummy, creada mediante la clusterización cantonal.

$$ClustCant3 = \begin{cases} 1 & \text{si el cantón del productor pertenece al Clúster 3.} \\ 0 & \text{caso contrario.} \end{cases}$$

5. **ClustCant4:** Variable dummy, creada mediante la clusterización cantonal.

$$ClustCant4 = \begin{cases} 1 & \text{si el cantón del productor pertenece al Clúster 4.} \\ 0 & \text{caso contrario.} \end{cases}$$

6. **ImbIbarLit:** Variable dummy, creada porque en el cantón Ibarra hay una marcada diferencia entre el rendimiento en sus parroquias.

$$ImbIbarLit = \begin{cases} 1 & \text{si el productor está en la parroquia Lita} \\ 0 & \text{caso contrario.} \end{cases}$$

7. **ImbIbarCar:** Variable dummy, creada porque en el cantón Ibarra hay una marcada diferencia entre el rendimiento en sus parroquias.

$$ImbIbarCar = \begin{cases} 1 & \text{si el productor está en la parroquia Carolina} \\ 0 & \text{caso contrario.} \end{cases}$$

8. **Variedad:** Variable dummy, representa la variedad de cacao que cultiva cada productor.

$$Variedad = \begin{cases} 1 & \text{si la variedad cultivada es CCN51.} \\ 0 & \text{si la variedad cultivada es Nacional Fino de Aroma.} \end{cases}$$

Para probar la significatividad global mediante el estadístico F se muestra a continuación la tabla ANOVA del modelo.

Fuente	**Grados de libertad (GL)**	**Suma de cuadrados (SC)**	**SC/GL**	**Estad. F**	**P-valor**
Explicada	8	1,93456	0,24182	118,34	0,000
Residual	556	1,13615	0,00204		
Total	564	3,07071			

Tabla 3.4: Análisis de varianza del modelo.

Al ser el p-valor del estadístico F menor que 0,05 se acepta la significatividad total del modelo. A partir de esta tabla también se puede calcular los valores del estadístico R cuadrado y R cuadrado ajustado. Siendo precisos se tiene que

$$R^2 = 63,0\,\% \;\; y \;\; \bar{R^2} = 62,5\,\%$$

Al hacer un análisis detallado de los residuales del modelo se observa la presencia de 5 observaciones que tienen un residuo muy elevado y al no ajustarse a la distribución normal esperada podían ser datos atípicos, como se puede apreciar en la siguiente figura.

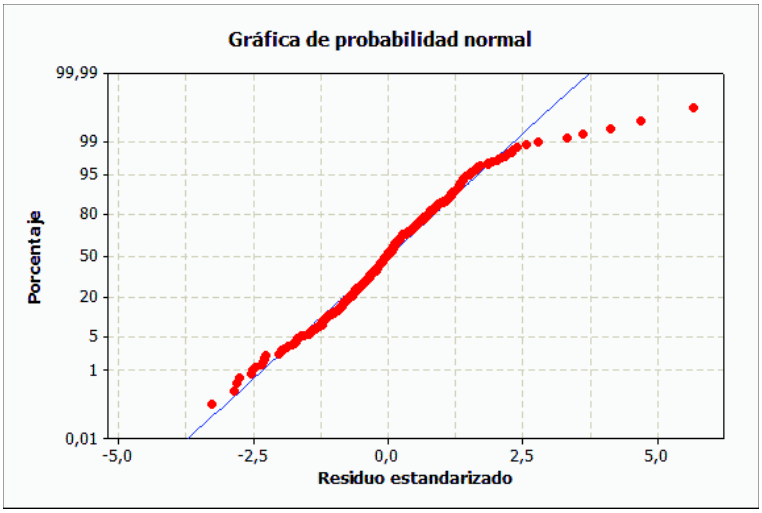


Figura 3.8: Gráfico de probabilidad normal de residuos. Se aprecian los 5 puntos que tienen un residuo alto en la parte superior derecha.

Al revisar más de cerca las observaciones problemáticas y compararlas con aquellas que provienen del mismo cantón, se confirma que se tratan de datos atípicos con un rendimiento anormalmente alto. Esta comparación se puede ver en la tabla 3.5.

Cantón	**N°**	**CTP**		**Cantón**	**N°**	**CTP**
Machala	94	0,288432		Guayaquil	197	0,357143
	101	0,330979			198	0,330033
	102	*0,543326*			199	0,404624
	103	0,313107		Cotacachi	257	0,436667
El Empalme	179	0,356500			258	0,320337
	180	0,390952			259	0,296996
	181	0,346791			*260*	*0,641822*
	182	*0,583632*			494	0,388462
	183	0,344784		La Joya de los Sachas	495	0,342034
	184	0,358960			496	0,245543
	185	0,386241			497	0,275845
	186	0,457518			498	0,401928
	187	0,335385			499	0,265531
	188	0,342535			500	0,374228
Guayaquil	*196*	*0,572687*			*501*	*0,605354*

Tabla 3.5: Comparación de los datos atípicos con los que provienen de sus respectivos cantones. Las observaciones atípicas son la 102, 182, 196, 260 y 501 (en cursiva en la tabla).

Por tanto, se procede a calcular otro modelo de regresión para el CTP en el cual estas observaciones atípicas son eliminadas. Podemos descartarlas sin problemas por varias razones:

- Este trabajo busca variables determinantes para el CTP esperado de las fincas cacaoteras y no se enfoca en analizar los factores detrás de plantaciones con rendimiento atípico.

- Analizando estos puntos conjuntamente con los expertos en producción del cacao del MAGAP se llegó a la conclusión de que estas observaciones atípicas pueden ser debidas a errores en el pesaje del cacao, ya sea en baba o en seco, o a un proceso de fermentación o secado incompleto, lo que entra en la categoría de errores de medición. Podrían deberse también a factores no tomados en cuenta en la toma de datos, lo cual queda fuera del alcance de este trabajo.

- Dado que los datos eliminados son apenas 5, representan menos del 1 % (concretamente, el 0,89 %) de la muestra original. Por tanto, su eliminación no supondrá una reducción radical del tamaño de la muestra con lo que las propiedades asintóticas de los estimadores dadas en el teorema 2.1.9 seguirían siendo válidas.

Desde este punto, **Modelo 1** se referirá al modelo calculado con la muestra completa, y denominaremos **Modelo 2** al que se calculará eliminando los datos atípicos.

3.3.2. Modelo 2

Una vez eliminadas las observaciones mencionadas se probaron varios modelos con los cuales se obtuvieron ciertas conclusiones que contradecían la experiencia previa de los expertos del MAGAP. Por tanto, se probó ajustar el modelo usando interacciones entre variables. De estas interacciones algunas resultaron significativas al nivel del 0,05 y encajaban con lo esperado por los técnicos en cultivo de cacao. La tabla 3.6 muestra la información de las variables explicativas del Modelo 2.

Variable	**Coeficiente**	**Err. estándar**	**Estad. t**	**P-valor**
Intercepto	0,458206	0,006157	74,20	0,000
ClustProv4	0,31168	0,02463	12,66	0,000
ClustCant1	-0,20677	0,01144	-18,07	0,000
ClustCant2	-0,096181	0,009571	-10,05	0,000
ClustCant3	-0,077210	0,007435	-10,39	0,000
ClustCant4	-0,032996	0,007446	-4,43	0,000
ImbIbarLit	0,26318	0,02416	10,89	0,000
ImbIbarCar	-0,12585	0,02142	-5,88	0,000
Variedad	-0,016259	0,005598	-2,90	0,004
BPCCPlant	-0,011365	0,005352	-2,12	0,034
PostcosechaM-CapacLabCult	0,03117	0,01570	1,99	0,048
Variedad-ClustCant2	-0,03239	0,01047	-3,10	0,002
Variedad-ClustCant4	-0,032648	0,008376	-3,90	0,000

Tabla 3.6: Variables regresoras para el CTP, Modelo 2.

Describiremos ahora las variables que han sido introducidas al Modelo 2.

1. **PostcosechaM:** Variable dummy, informa sobre la mecanización de las tareas agrícolas.

$$PostcosechaM = \begin{cases} 1 & \text{si la postcosecha se realiza de forma mecanizada.} \\ 0 & \text{caso contrario.} \end{cases}$$

2. **CapacLabCult:** Variable dummy, indica qué tema de capacitación consideró el agricultor de mayor provecho para su producción, en el caso de haber recibido capacitación.

$$CapacLabCult = \begin{cases} 1 & \text{si el tema es Realización de Labores Culturales.} \\ 0 & \text{caso contrario.} \end{cases}$$

3. **BPCCPlant:** Variable dummy, informa sobre los beneficios recibidos por el agricultor a través del Proyecto Café y Cacao del MAGAP.

$$BPCCPlant = \begin{cases} 1 & \text{si el beneficio recibido fue de Plantas.} \\ 0 & \text{caso contrario.} \end{cases}$$

En la tabla 3.7 vemos las sumas de cuadrados y el estadístico F calculado para el Modelo 2.

Fuente	**Grados de libertad (GL)**	**Suma de cuadrados (SC)**	**SC/GL**	**Estad. F**	**P-valor**
Explicada	12	1,94680	0,16223	99,12	0,000
Residual	547	0,89529	0,00164		
Total	559	2,84209			

Tabla 3.7: Análisis de varianza del Modelo 2.

De acuerdo a este estadístico, afirmamos que el Modelo 2 también es significativo. Además tenemos que

$$R^2 = 68,5\,\% \quad y \quad \bar{R}^2 = 67,8\,\%$$

por lo cual este modelo explica más variabilidad del CTP que el Modelo 1; algo esperable, dada la eliminación de los 5 puntos extraños y la entrada de más variables al modelo.
Procederemos ahora a validar los modelos calculados.

3.4. Validación de los modelos

3.4.1. Modelo 1

Prueba RESET de forma funcional

Como mencionamos en el capítulo 2, la prueba RESET se hace con las potencias superiores de los valores ajustados de la ecuación. La siguiente tabla resume la información de los estadísticos para la prueba y su p-valor respectivo.

Fuente	Estadístico	GL	P-valor
$\hat{Y}^2$	$t = -0,415721$	10	0,6778
$\hat{Y}^3$	$t = 0,398098$	10	0,6907
Total	$F = 0,100478$	(2,554)	0,9044

Tabla 3.8: Resultados de la prueba RESET, Modelo 1.

Al no ser ninguno de los p-valores menores a 0,05 rechazamos la hipótesis de que potencias superiores o productos de las variables regresoras sean significativas en el Modelo 1.

Multicolinealidad

La figura 3.9 muestra la matriz de correlaciones para las variables explicativas del Modelo 1. Vemos que ningún valor supera el 0,90 por lo cual no parece haber multicolinealidad alta entre pares de variables.

	CLUSTPROV4	CLUSTCANT1	CLUSTCANT2	CLUSTCANT3	CLUSTCANT4	IMBIBARLIT	IMBIBARCAR	VARIEDAD
CLUSTPROV4	1.000000							
CLUSTCANT1	-0.013629	1.000000						
CLUSTCANT2	-0.033469	-0.085456	1.000000					
CLUSTCANT3	-0.058344	-0.148964	-0.365811	1.000000				
CLUSTCANT4	-0.048537	-0.123925	-0.304323	-0.530496	1.000000			
IMBIBARLIT	-0.005338	-0.013629	-0.033469	-0.058344	-0.048537	1.000000		
IMBIBARCAR	-0.006169	-0.015752	-0.038681	-0.067430	-0.056096	-0.006169	1.000000	
VARIEDAD	0.051207	0.016662	0.099890	0.211812	-0.145536	-0.087213	0.027855	1.000000

Figura 3.9: Matriz de correlaciones para el Modelo 1.

Revisaremos ahora los valores del FIV y FIV generalizado. Tenemos dos paquetes de variables que provienen de una misma fuente: la clusterización cantonal (ClustCant1, ClustCant2, ClustCant3 y ClustCant4) y la división para el cantón Ibarra (ImbIbarLit y ImbIbarCar). El FIV para cada variable se muestra en la tabla 3.9. Ningún valor excede el límite establecido de 5, por lo que se concluye que la multicolinealidad no representa un problema para la estimación del modelo.
Idéntica conclusión obtenemos al analizar los FIV generalizados de las dos fuentes mencionadas en la tabla 3.10. Por tanto, la multicolinealidad no representa un problema para el Modelo 1.

Variable	FIV
ClustProv4	1,081
ClustCant1	1,416
ClustCant2	2,859
ClustCant3	4,015
ClustCant4	3,499
ImbIbarLit	1,061
ImbIbarCar	1,094
Variedad	1,159

Tabla 3.9: FIV de las variables del Modelo 1.

Fuente	GVIF	GVIF^{1/p_1}
Clust. cantonal	1,3727	1,0824
Ibarra	1,1553	1,0748

Tabla 3.10: FIV generalizado del Modelo 1.

Heterocedasticidad

Comenzaremos con el análisis de los diagramas de dispresión de los residuos al cuadrado contra cada variable regresora. Dichos diagramas se pueden ver en la figura 3.10.

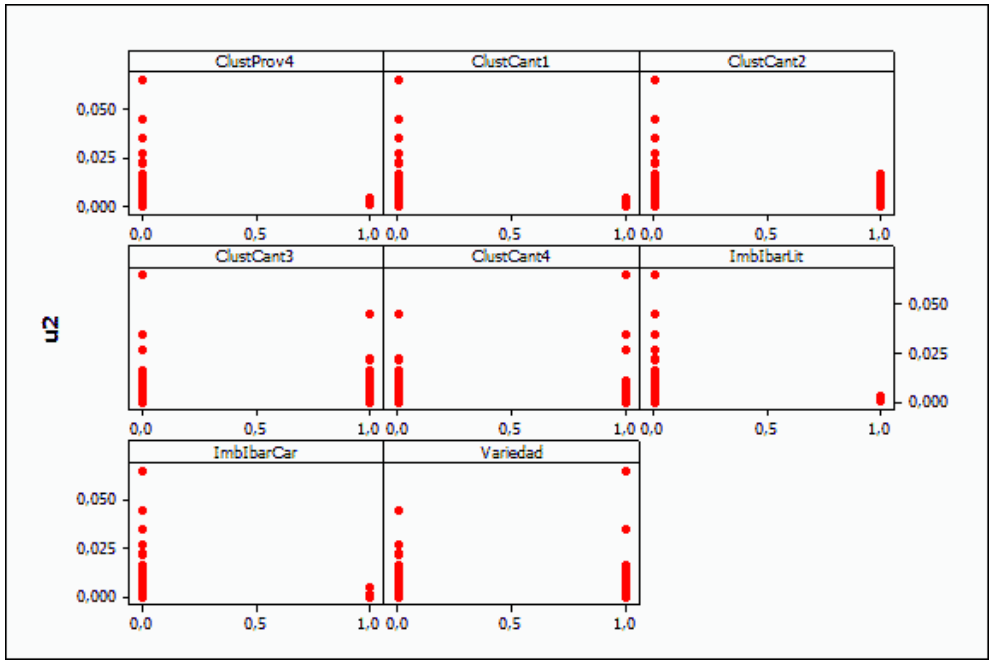


Figura 3.10: Diagrama de dispersión de los residuos al cuadrado contra las variables regresoras del Modelo 1.

Vemos algunas tendencias claras en los diagramas de dispersión, pero dichas tendencias se ven amplificadas por los puntos atípicos que conocemos. Para probar si estas tendencias en la varianza residual son significativas debemos realizar los contrastes estadísticos.

Hay que aclarar que en este caso no se puede usar la prueba de Park, pues esta depende de los logaritmos de las variables y en el caso de variables binarias el logaritmo no puede ser calculado.
Comenzaremos con la revisión del contraste de Breusch-Pagan. La tabla 3.11 muestra los tres estadísticos obtenidos para el Modelo 1 y su p-valor.

Estadístico	**GL**	**Valor**	**P-valor**
$LM = SCE/2$	8	14,802	0,0631
F de la regresión	(8,556)	0,693	0,6982
$n * R^2_{\hat{e}}$	8	5,575	0,6947

Tabla 3.11: Resultados del contraste de Breusch-Pagan, Modelo 1.

Dado que ningún estadístico de prueba tiene un p-valor menor a 0,05 podemos asumir que no hay evidencia fuerte de heterocedasticidad según el contraste de Breusch-Pagan.
Ahora presentaremos los resultados del contraste de White en la tabla 3.12.

Estadístico	**GL**	**Valor**	**P-valor**
F de la regresión	(13,551)	1,568	0,0901
$n * R^2_{\hat{u}^2}$	13	20,153	0,0915

Tabla 3.12: Resultados del contraste de White, Modelo 1.

También con un 95 % de confianza se puede afirmar que el Modelo 1 no presenta heterocedasticidad según el contraste de White.

Normalidad de residuos

En secciones anteriores presentamos la figura 3.8 como evidencia de la existencia de observaciones atípicas en el Modelo 1. De la misma forma, podemos presentar esa figura para aseverar que los residuos no se ajustan a una distribución normal.
La tabla 3.13 muestra los resultados de las pruebas de Anderson-Darling y Jarque-Bera para los residuales del Modelo 1. Es claro que no se puede aceptar que estos residuos sigan una distribución normal. Recordemos que en estas pruebas se espera que los p-valores sean mayores que cierto nivel de significancia, que tradicionalmente se define como 0,05.

Prueba	**Estadístico**	**P-valor**
Anderson-Darling	3,072	<0,005
Jarque-Bera	288,214	0,001

Tabla 3.13: Resultados de las pruebas de normalidad de residuos, Modelo 1.

Sin embargo, dado que el tamaño de muestra es grande, gracias al teorema 2.1.9 entendemos que el cumplimiento estricto del supuesto 6 de normalidad no es indispensable para las pruebas de hipótesis realizadas.

3.4.2. Modelo 2

Prueba RESET de forma funcional

La tabla 3.14 resume la información obtenida de la prueba de forma funcional para el Modelo 2.

Fuente	Estadístico	GL	P-valor
$\hat{Y}^2$	$t = 0,140935$	14	0,8880
$\hat{Y}^3$	$t = -0,219186$	14	0,8266
Total	$F = 0,247916$	(2,545)	0,7891

Tabla 3.14: Resultados de la prueba RESET, Modelo 2.

Podemos entonces afirmar, dado que ningún p-valor es menor a 0,05, que la forma funcional estimada es correcta y no se deben añadir productos cruzados ni potencias superiores al Modelo 2.

Multicolinealidad

Al igual que con el Modelo 1, comenzaremos analizando la matriz de correlaciones de las variables independientes. Hemos añadido a este modelo interacciones de variables que ya estaban presentes en el modelo, por lo cual es esperable encontrar valores de correlación altos en la matriz. Sin embargo, al revisarla con detenimiento en la figura 3.11, se observa que ninguna correlación es superior a 0,90 por lo que consideramos que la colinealidad entre pares de variables no es un problema para este modelo.

	CLUSTPROV4	CLUSTCANT1	CLUSTCANT2	CLUSTCANT3	CLUSTCANT4	VARIEDAD	IMBIBARLIT	IMBIBARCAR
CLUSTPROV4	1.000000							
CLUSTCANT1	-0.013753	1.000000						
CLUSTCANT2	-0.033801	-0.086312	1.000000					
CLUSTCANT3	-0.058593	-0.149621	-0.367712	1.000000				
CLUSTCANT4	-0.048454	-0.123729	-0.304077	-0.527117	1.000000			
VARIEDAD	0.061269	0.016111	0.097917	0.219874	-0.151400	1.000000		
IMBIBARLIT	-0.005386	-0.013753	-0.033801	-0.058593	-0.048454	-0.087908	1.000000	
IMBIBARCAR	-0.006225	-0.015895	-0.039065	-0.067719	-0.056000	0.027709	-0.006225	1.000000
POSTCOSECHAM*CAPACLABCULT	-0.008257	-0.021085	-0.009518	-0.089825	0.100498	-0.102095	-0.008257	-0.009543
BPCCPLANT	-0.027739	-0.011184	-0.103026	-0.024916	0.078266	-0.222252	-0.027739	0.032059
VARIEDAD*CLUSTCANT2	-0.027284	-0.069671	0.807198	-0.296816	-0.245451	0.310369	-0.027284	-0.031533
VARIEDAD*CLUSTCANT4	-0.030179	-0.077064	-0.189394	-0.328314	0.522849	0.343306	-0.030179	-0.034879

(a)

	POSTCOSECHAM*CAPACLABCULT	BPCCPLANT	VARIEDAD*CLUSTCANT2	VARIEDAD*CLUSTCANT4
POSTCOSECHAM*CAPACLABCULT	1.000000			
BPCCPLANT	0.103273	1.000000		
VARIEDAD*CLUSTCANT2	0.007381	-0.123084	1.000000	
VARIEDAD*CLUSTCANT4	-0.046266	-0.063322	-0.152879	1.000000

(b)

Figura 3.11: Matriz de correlaciones del Modelo 2. Por su gran tamaño se la presenta dividida en dos partes.

El análisis del FIV para las variables del Modelo 2 también apoya la afirmación de falta de multicolinealidad grave entre las variables al no ser ninguno de ellos superior a 5, aun cuando algunos se acercan bastante a esta cantidad. Vemos estos valores en la tabla 3.15.

Variable	FIV
ClustProv4	1,106
ClustCant1	1,469
ClustCant2	4,525
ClustCant3	4,522
ClustCant4	4,010
ImbIbarLit	1,064
ImbIbarCar	1,113
Variedad	2,595
BPCCPlant	1,072
PostcosechaM-CapacLabCult	1,041
Variedad-ClustCant2	3,998
Variedad-ClustCant4	2,970

Tabla 3.15: FIV de las variables del Modelo 2.

El analisís del FIV generalizado aquí aumenta en una fuente. Además de la clusterización cantonal y la del cantón Ibarra que también fueron analizadas en el Modelo 1, estudiaremos también el ingreso de las interacciones entre la variedad y el clúster cantonal (Variedad-ClustCant2 y Variedad-ClustCant4). Los resultados se muestran en la tabla 3.16.

Fuente	GVIF	GVIF^{1/p_1}
Clust. cantonal	10,4841	1,7994
Ibarra	1,1797	1,0861
Inter. variedad-cantón	10,1826	1,7863

Tabla 3.16: FIV generalizado del Modelo 2.

Ninguna de las pruebas muestra que el Modelo 2 presente problemas fuertes de multicolinealidad entre sus variables independientes.

Heterocedasticidad

Analizaremos la gráfica de dispersión de residuos al cuadrado de Modelo 2 (mostrada en la figura 3.12) en busca de patrones para encontrar evidencias de heterocedasticidad.
Mostramos también los resultados del contraste de Breusch-Pagan, vistos con más claridad en la tabla 3.17. Apenas un estadístico de prueba posee un p-valor inferior al 0,05 por lo cual este contraste no afirma categóricamente la presencia de heterocedasticidad.

Estadístico	GL	Valor	P-valor
$LM = SCE/2$	12	22,994	0,0278
F de la regresión	(12,547)	1,453	0,1378
$n * R^2_{\tilde{e}}$	12	17,303	0,1386

Tabla 3.17: Resultados del contraste de Breusch-Pagan, Modelo 2.

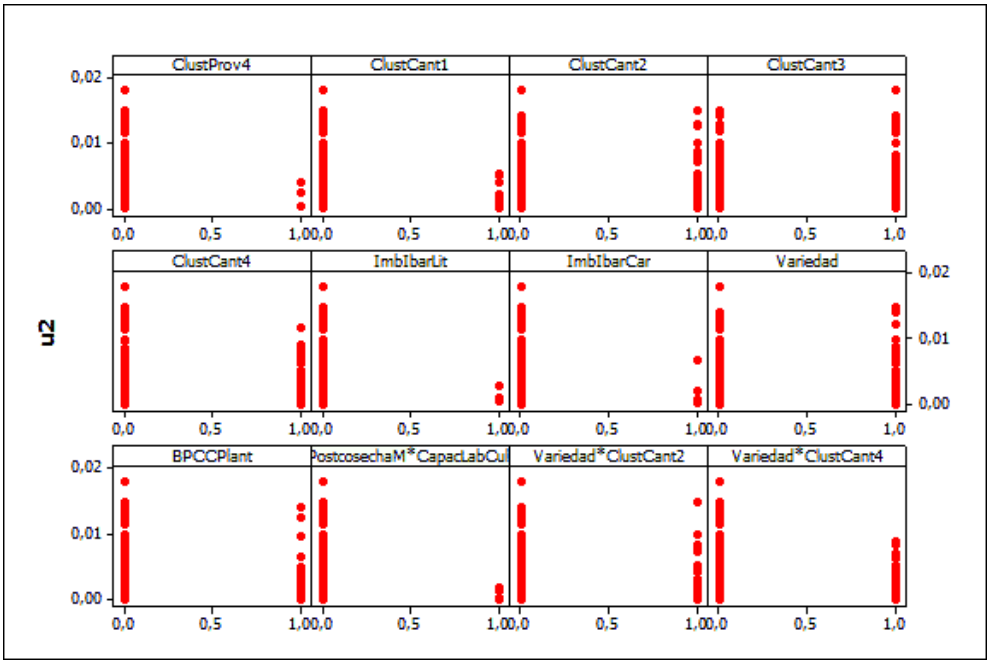


Figura 3.12: Diagrama de dispersión de los residuos al cuadrado contra las variables regresoras del Modelo 2.

El contraste de White también rechaza la presencia de heterocedasticidad en el Modelo 2. Los resultados de este contraste se ven con detalle en la tabla 3.18. Al no ser ningún estadístico significativo al nivel de 95 % de confianza podemos descartar que la heterocedasticidad sea un problema relevante para el Modelo 2.

Estadístico	GL	Valor	P-valor
F de la regresión	(26,533)	1,386	0,0984
$n * R^2_{\hat{u}^2}$	266	35,463	0,1020

Tabla 3.18: Resultados del contraste de White, Modelo 2.

Normalidad de residuos

Cabría esperar que por la eliminación de los puntos atípicos los residuos del Modelo 2 se ajusten mucho mejor a una distribución normal. Y al analizar el gráfico de probabilidad normal de este modelo que se presenta en la figura 3.13 vemos que en realidad el ajuste es mucho más exacto. Aun cuando este modelo también presenta puntos que se alejan del eje del gráfico, no están tan separados como en el Modelo 1. Sin embargo, para probar la normalidad más fuertemente revisaremos los resultados de las pruebas estadísticas de normalidad. Dichos resultados se presentan en la tabla 3.19.

Aun cuando ninguno de los resultados sea estadísticamente significativo, se aprecia claramente la disminución en los valores de los estadísticos de prueba tanto para la de Anderson-Darling como para la de Jarque-Bera en comparación con los del Modelo

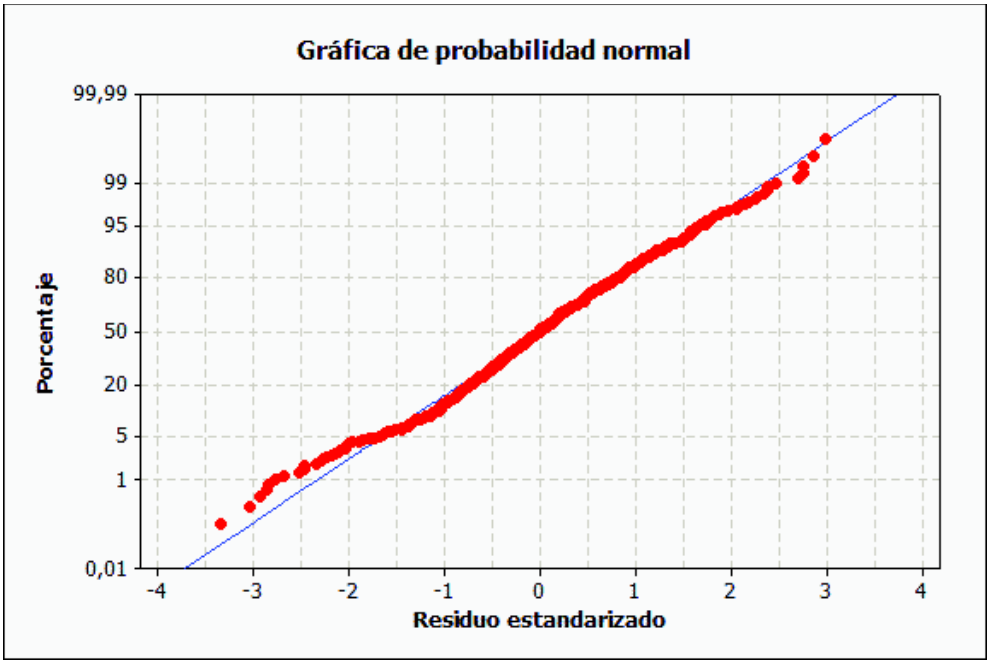


Figura 3.13: Gráfico de probabilidad normal para los residuales del Modelo 2.

1. Esta disminución se debe al mejor ajuste de los datos, aun cuando no logren pasar el nivel de significancia requerido. También recordemos que el teorema 2.1.9 permite relajar la condición de normalidad de residuos y asumir que el Modelo 2 tiene residuos asintóticamente normales al ser su tamaño de muestra bastante grande.

Prueba	Estadístico	P-valor
Anderson-Darling	1,355	<0,005
Jarque-Bera	10,992	0,004

Tabla 3.19: Resultados de las pruebas de normalidad de residuos, Modelo 2.

Una vez validados los supuestos del modelo lineal clásico para el Modelo 1 y 2, podemos discutir sus resultados.

3.5. Discusión e interpretación de parámetros

3.5.1. Modelo 1

Dado que todas las variables independientes del modelo son variables dummy binarias, primero definiremos la **población base** del estudio; es decir, aquella para cual todas las variables toman el valor de cero.
La descripción de esta población base es la siguiente:

- Son agricultores que residen en los cantones del clúster 5 (el de mejor rendimiento postcosecha).
- Cultivan la variedad de cacao Nacional Fino de Aroma.

Para estos agricultores, su CTP esperado es de 45,86 %. Conocido esto, damos ahora la influencia de cada variable en el CTP del cacao.

1. Los agricultores que residen en la provincia de Carchi (exactamente en el cantón Mira) tienen un CTP esperado 32,89 puntos porcentuales superior al de la población base. Estos agricultores tienen un CTP esperado de 78,75 %.

2. Los agricultores que residen en el clúster cantonal 1 tienen un CTP esperado 19,73 puntos porcentuales menor al de la población base. Estos agricultores tienen un CTP esperado de 26,13 %.

3. Los agricultores que residen en el clúster cantonal 2 tienen un CTP esperado 10,71 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 35,15 %.

4. Los agricultores que residen en el clúster cantonal 3 tienen un CTP esperado 6,46 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 39,40 %.

5. Los agricultores que residen en el clúster cantonal 4 tienen un CTP esperado 3,79 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 42,07 %.

6. Los agricultores que residen en la parroquia Lita del cantón Ibarra tienen un CTP esperado 26,28 puntos porcentuales mayor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 72,14 %.

7. Los agricultores que residen en la parroquia Carolina del cantón Ibarra tienen un CTP esperado 11,59 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 34,27 %.

8. Los agricultores que cultivan la variedad CCN-51 tienen un CTP esperado 3,38 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 42,48 %.

Este modelo posee pocas variables que expliquen el CTP a partir de métodos de cultivo del agricultor. Las variables que más influencia ejercen en el CTP son las variables geográficas. La única variable de cultivo (Variedad) apenas cambia el CTP en 3 puntos porcentuales, lo cual no es un cambio fuerte. En el caso de fermentar y secar 10 Kg de cacao en baba de diferentes variedades, se obtendrían apenas 338 gramos menos de cacao seco CCN-51 que su contraparte Nacional.

3.5.2. Modelo 2

La población base de este modelo es la siguiente:

- Son agricultores que residen en los cantones del clúster 5, el de mejor rendimiento.
- Cultivan la variedad de cacao Nacional Fino de Aroma.
- No han mecanizado sus procesos de postcosecha (fermentación y secado).
- No han recibido capacitación en la realización de labores culturales.
- No recibieron el beneficio de plantas por el Proyecto Café y Cacao del MAGAP.

Para estos agricultores, el CTP esperado es de 45,82 %. Definida la población base, el Modelo 2 muestra las siguientes influencias.

1. Los agricultores que residen en la provincia de Carchi (exactamente en el cantón Mira) tienen un CTP esperado 31,17 puntos porcentuales superior al de la población base. Estos agricultores tienen un CTP esperado de 76,99 %.
2. Los agricultores que residen en el clúster cantonal 1 tienen un CTP esperado 20,68 puntos porcentuales menor al de la población base. Estos agricultores tienen un CTP esperado de 25,14 %.
3. Los agricultores que residen en el clúster cantonal 2 y cultivan la variedad de cacao Nacional Fino de Aroma tienen un CTP esperado 9,62 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 36,20 %.
4. Los agricultores que residen en el clúster cantonal 3 tienen un CTP esperado 7,72 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 38,10 %.
5. Los agricultores que residen en el clúster cantonal 4 y cultivan la variedad de cacao nacional Fino de Aroma tienen un CTP esperado 3,30 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 42,52 %.
6. Los agricultores que residen en la parroquia Lita del cantón Ibarra tienen un CTP esperado 26,32 puntos porcentuales mayor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 72,14 %.
7. Los agricultores que residen en la parroquia Carolina del cantón Ibarra tienen un CTP esperado 12,56 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 33,26 %.
8. Los agricultores que cultivan la variedad CCN-51 tienen un CTP esperado 1,63 puntos porcentuales menor al de los agricultores de la población base. Estos agricultores tienen un CTP esperado de 44,19 %. Si además de cultivar

esta variedad residen en el clúster cantonal 2, su CTP esperado es 4,87 puntos porcentuales menor que los agricultores de la población base y por tanto el CTP esperado toma un valor de 40,95 %. En caso de cultivar la variedad CCN-51 y residir en el clúster cantonal 4 su CTP esperado es 4,89 puntos porcentuales menor que los agricultores de la población base. Por lo cual su CTP esperado toma un valor de 40,93 %.

9. Los agricultores que recibieron como beneficio del Proyecto Café y Cacao plantas tienen un CTP esperado 1,14 puntos porcentuales menor al de los agricultores del grupo base. Estos agricultores tienen un CTP esperado de 44,68 %.

10. Los agricultores que mecanizaron sus procesos de postcosecha y recibieron capacitación en la realización de labores culturales tienen un CTP esperado 3,12 puntos percentuales mayor que los agricultores del grupo base. Estos agricultores tienen un CTP esperado de 48,95 %.

Vemos que los efectos más fuertes sobre el CTP vienen dados por las clusterizaciones geográficas. Este modelo tiene más variables de cultivo que el Modelo 1, aun cuando el impacto de estas no es comparable al de las variables geográficas. También los signos de las variables están acorde a lo esperado tanto por el sentido común como por la experiencia en campo de los técnicos del cacao del MAGAP. Estos indicaron que la única variable que no sigue lo esperado (BPCCPlant) podría ser consecuencia de que a pesar de que se conoce que las plantas del Proyecto Café y Cacao son superiores en rendimiento de cosecha, postcosecha, resistencia a plagas y adaptación al medio, estas propiedades no se presentan inmediatamente sino a partir de un lapso de entre 18 a 24 meses a partir de su siembra que es el tiempo que toma la maduración y aclimatación. Durante este lapso, se puede afectar el rendimiento de las plantas antiguas puesto que se altera el ciclo de polinización en la granja y con este la normal producción de mazorcas; además de que para la siembra de estas nuevas plantas se precisa de un tratamiento de renovación parcial o completo de la finca productora en la región de siembra, lo cual suele traer consigo la necesidad de eliminar parte de las plantas antiguas y someter a las restantes a podas lo que influiría en el rendimiento esperado de la plantación.
Por tener mejores características de ajuste expresadas en un valor de R cuadrado más elevado, por no presentar puntos atípicos de residuos muy altos, y por presentar interacciones útiles para la estimación del CTP decidimos entonces que el Modelo 2 será el modelo escogido que presentaremos como resultado de este trabajo.

Capítulo 4

Conclusiones y recomendaciones

- El rendimiento postcosecha del cacao ecuatoriano se puede modelar mediante el uso de herramientas estadísticas como la regresión lineal múltiple. El modelo hallado cumple con los supuestos necesarios y por tanto sus parámetros y conclusiones son válidos tanto de forma teórica como de forma práctica.

- De todas las variables incluidas en el modelo, aquellas que inciden con más intensidad en la estimación del CTP son las variables geográficas o de ubicación del agricultor, seguidas por las agrícolas o de métodos de cultivo.

- Una limitación de este estudio es que muchas de las variables tomadas en cuenta como candidatas para explicar el CTP eran *percepciones* subjetivas de los agricultores más que variables medidas con total precisión, en particular para las variables socioeconómicas y agrícolas, como se puede revisar en las boletas de toma de datos (figuras A.1 y A.2). Posteriores trabajos deberían validar con precisión la información obtenida de los encuestados a fin de garantizar que las conclusiones sean aplicables y precisas.

- Al analizar los resultados del primer modelo de regresión (tabla 3.3) obtenido en este trabajo se halló evidencia de fincas con un rendimiento muy superior al pronosticado por dicho modelo. El cálculo de un segundo modelo (tabla 3.6) sin estas fincas reveló algunas interacciones interesantes que influyeron sobre la estimación y añadieron nuevas variables exógenas significativas para el CTP, como la interacción hallada entre la variedad de cacao sembrada y la región geográfica en la cual se siembra; sin embargo, las estimaciones de los coeficientes de las variables comunes a los dos modelos no registraron cambios muy drásticos.

- Las variables determinantes agrícolas de los modelos calculados en este trabajo encajan en su mayoría con la información consultada a los expertos del MAGAP acerca del proceso de beneficio del cacao en Ecuador. Además de determinar que la mecanización de los procesos postcosecha y la capacitación recibida influyen

de manera positiva en los rendimientos de los agricultores, se presentó una estimación numérica precisa de este impacto en el rendimiento mencionado.

- El modelo de regresión determinó que, *ceteris paribus*, las provincias con un mejor rendimiento postcosecha en el país, medido a través del CTP, son las provincias de Carchi, Imbabura y Manabí, en ese orden; resultado en sí inesperado en el caso de las dos primeras provincias, pues al ser provincias de la Sierra no tienen las características consideradas ideales por los expertos para el cultivo de cacao, y además históricamente no poseen gran tradición cacaotera, a diferencia de algunas provincias de la Costa en las que se ha cultivado y procesado cacao desde hace mucho más tiempo.

- La producción de cacao en Ecuador es reconocida mundialmente más por su calidad que por su cantidad. La variedad de cacao Nacional Fino de Aroma cultivada en nuestro país generalmente es usada para la producción de chocolates finos. Este trabajo además revela que esta variedad presenta un rendimiento postcosecha ligeramente superior al de la otra variedad de cacao cultivada, la CCN-51, que a pesar de no poseer el reconocimiento cualitativo de sus características de aroma y sabor, es resistente a plagas y tiene una productividad superior al cacao Nacional.

- El Modelo 2 muestra que la variedad cultivada de cacao no solo influye en el rendimiento postcosecha de cacao de forma significativa, sino que su influencia es diferente de acuerdo al lugar en el cual se cultiva esta variedad. Se debe analizar a qué factor con precisión responde este cambio de rendimiento mediante estudios que se enfoquen en esta característica particular.

- El Modelo 2 encontró evidencia de que, *ceteris paribus,* las agricultores que recibieron plantas entregadas por el Proyecto Café y Cacao registran un CTP esperado ligeramente inferior al de los demás. Se conoce que estas plantas tienen características deseables en cuanto a resistencia a enfermedades y cantidad de mazorcas producidas, y su rendimiento postcosecha debería también ser superior a las plantaciones originales. Aun cuando la disminución de rendimiento es poco notoria, se deberían realizar posteriores investigaciones sobre este particular para tener más información, para determinar si esta falla en el rendimiento es debido a una característica intrínseca de las plantas, u obedece a posibles factores de cuidado del agricultor o a su ventana de preparación antes de alcanzar el pico de producción, de la que se habló en la sección 3.5.2.

- Al revisar la clusterización cantonal realizada en este estudio se han identificado cantones en los cuales se registra un muy alto rendimiento postcosecha de cacao. Es recomendable realizar un estudio más a profundidad en dichos cantones para determinar las posibles razones de estas anomalías, e incluir en estos

estudios la información que por falta de datos no se incluyó en este trabajo, como la fertilización, el tipo de suelo, la temperatura promedio, la humedad del ambiente, y otras.

- La iniciativa de toma de datos con la cual se realizó este trabajo se ha previsto realizarse anualmente. A partir de dicha información futura se podría refinar aún más la estimación de los determinantes para el CTP y reafirmar o mejorar las conclusiones de este estudio utilizando la información de diferentes ventanas de tiempo, para lo cual se puede aplicar modelos de datos de panel. Los resultados podrían ser aún mejores si la toma de datos se realiza dos veces al año de acuerdo a los picos de producción de los que se habló en el capítulo 1 de este trabajo.

- Mediante los resultados hallados en este trabajo se espera que las entidades encargadas tanto del monitoreo y levantamiento de información como del apoyo al agricultor apliquen políticas acordes a la ubicación geográfica de los productores, identifiquen los temas de capacitación más provechosos para ellos y incentiven las prácticas agrícolas que muestran resultados superiores; todo ello con el fin de mejorar la productividad del sector cacaotero, que redundará en beneficios tanto para las personas que viven de su cultivo en particular, como para la industria, el comercio exterior y el país en general.

Anexos

Anexo A

Boleta para la toma de datos del agricultor

Presentamos a continuación el ejemplo de la boleta física que se usó para el ingreso de datos en este estudio. Está dividida en cinco partes:

1. Datos básicos: fechas de toma de información, ubicación y responsables de la toma de datos.
2. Datos socioeconómicos del agricultor.
3. Datos básicos de la hacienda.
4. Datos productivos del cultivo de cacao.
5. Datos tomados por los investigadores en el campo.

Además se tomó información individual y física de cada una de las mazorcas de cacao muestreadas y usadas para el estudio; sin embargo en este trabajo solamente se utilizó la información total sumada.

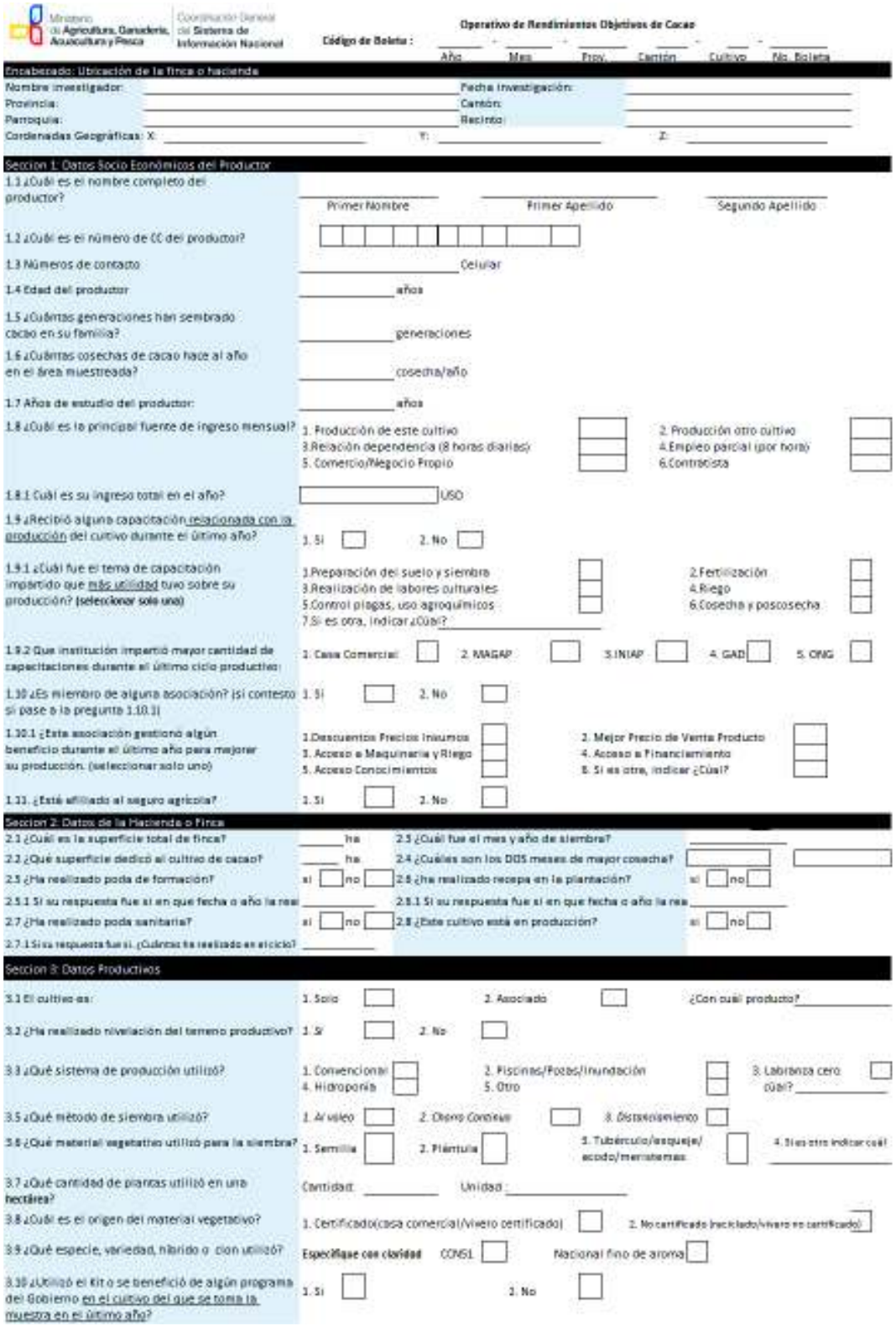


Figura A.1: Boleta de toma de datos, parte 1.

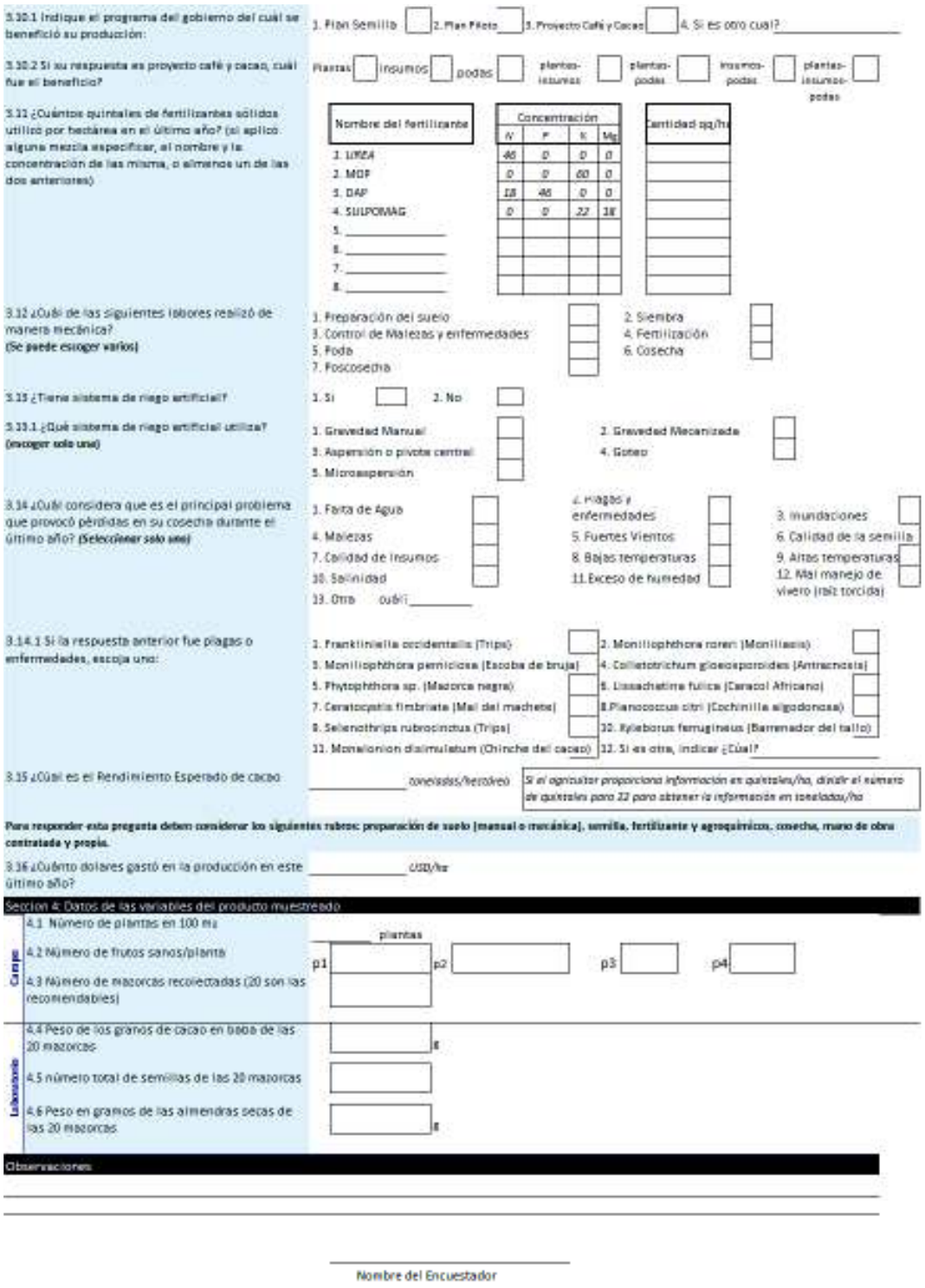


Figura A.2: Boleta de toma de datos, parte 2.

Anexo B

Promedios del CTP provinciales y cantonales

Provincia	CTP promedio
Azuay	0,3759
Bolívar	0,3743
Cañar	0,3502
Carchi	0,7536
Chimborazo	0,3629
Cotopaxi	0,2515
El Oro	0,3513
Esmeraldas	0,3403
Guayas	0,3737
Imbabura	0,4663
Los Ríos	0,3700
Manabí	0,4214
Napo	0,3536
Orellana	0,3466
Pichincha	0,3930
Sta. Elena	0,4019
Sto. Domingo	0,3579
Sucumbíos	0,3408

Tabla B.1: CTP promedio por provincias.

Cantón	CTP promedio
24 de Mayo	0,3979
Alfredo Baquerizo Moreno	0,4018
Archidona	0,3312
Arenillas	0,3258
Atacames	0,3910
Atahualpa	0,3440
Baba	0,3754
Babahoyo	0,3664
Balao	0,3833
Balsas	0,3500
Balzar	0,3629
Bolívar	0,4424
Buena Fe	0,3659
Caluma	0,3619
Camilo Ponce Enríquez	0,3316
Cañar	0,3368
Carlos Julio Arosemena Tola	0,3547
Cascales	0,3433
Chilla	0,3971
Chillanes	0,3963
Chimbo	0,3821
Chone	0,4100
Colimes	0,3826
Coronel Marcelino Maridueña	0,3769
Cotacachi	0,4240
Cuenca	0,3737
Cumandá	0,3629
Durán	0,3398
Echeandía	0,3674
El Carmen	0,3978
El Empalme	0,3903
El Guabo	0,3322
El Triunfo	0,3660
Eloy Alfaro	0,3571
Esmeraldas	0,2828
Flavio Alfaro	0,4601
General Antonio Elizalde	0,4032
Gonzalo Pizarro	0,3546
Guaranda	0,3384

Continúa en la siguiente página.

Viene de la página anterior.

Guayaquil	0,4161
Ibarra	0,4905
Isidro Ayora	0,3764
Jama	0,4611
Jipijapa	0,4030
Junín	0,4056
La Concordia	0,3950
La Joya de los Sachas	0,3624
La Troncal	0,3680
Lago Agrio	0,3358
Las Lajas	0,3280
Las Naves	0,4095
Loreto	0,3152
Machala	0,3644
Milagro	0,4006
Mira	0,7536
Mocache	0,3720
Montalvo	0,4083
Muisne	0,3837
Naranjal	0,3386
Naranjito	0,3742
Olmedo	0,3988
Paján	0,3898
Palenque	0,4126
Pangua	0,2192
Pasaje	0,3884
Pedernales	0,4498
Pedro Vicente Maldonado	0,4486
Pichincha	0,4541
Piñas	0,3869
Portoviejo	0,4310
Pucará	0,4240
Puebloviejo	0,3730
Puerto Quito	0,3659
Pujilí	0,3322
Putumayo	0,4009
Quevedo	0,2472
Quinindé	0,3140
Quinsaloma	0,3872

Continúa en la siguiente página.

Viene de la página anterior.

Quito	0,3805
Rioverde	0,3698
Rocafuerte	0,4113
San Jacinto de Yaguachi	0,3781
San Lorenzo	0,3145
San Miguel	0,3537
San Miguel de los Bancos	0,3848
San Vicente	0,4168
Santa Ana	0,4674
Santa Elena	0,4019
Santa Lucía	0,3464
Santa Rosa	0,2947
Santo Domingo	0,3509
Shushufindi	0,3090
Simón Bolívar	0,3668
Sucre	0,4545
Tena	0,3750
Tosagua	0,4172
Valencia	0,3732
Ventanas	0,3636
Zaruma	0,3212

Tabla B.2: CTP promedio por cantones.

Anexo C

Tablas de la función de inercia para las clusterizaciones

k	Función de inercia
2	0,030860
3	0,015561
4	0,005566
5	0,002284
6	0,000944
7	0,000402
8	0,000193
9	0,000153
10	0,000057
11	0,000041
12	0,000028

Tabla C.1: Función de inercia de la clusterización por provincias.

k	Función de inercia
2	0,085536
3	0,046308
4	0,027929
5	0,014264
6	0,010131
7	0,007908
8	0,006099
9	0,004945
10	0,003543
11	0,002759
12	0,002385
13	0,002049
14	0,001549
15	0,001406
16	0,001237
17	0,001105
18	0,000751
19	0,000895
20	0,000622
21	0,000553
22	0,000468
23	0,000436
24	0,000405
25	0,000382

Tabla C.2: Función de inercia de la clusterización por cantones.

Bibliografía

[1] Zambrano Alexis, Gómez Álvaro, Ramos Gladys, Romero Carlos, Lacruz Carlos, y Rivas Eliana. Caracterización de parámetros físicos de calidad en almendras de cacao criollo, trinitario y forastero durante el proceso de secado. *Agronomía Tropical*, 60(4):389–396, 2010.

[2] Novales Cinca Alfonso. *Econometría*. McGraw-Hill/Interamericana de España, Madrid, España, 2 ed$^{\underline{\text{ón}}}$., 2000.

[3] Hernández P. Alicia. Evaluación del proceso de fermentación de cacao en Costa Rica. En *Memoria Seminario Regional sobre Tecnología Poscosecha y Calidad Mejorada del Cacao*, págs. 129–140. Instituto Interamericano de Cooperación para la Agricultura (IICA), Turrialba, Costa Rica, 1989.

[4] Jain Anil K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, (31):651–666, 2010.

[5] Hoerl Arthur E. y Kennard Robert W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[6] Law Averill M. y Kelton David W. *Simulation Modeling and Analysis*. McGraw-Hill, New York, USA, 2 ed$^{\underline{\text{ón}}}$., 1991.

[7] Yazici Berna y Yolacan Senay. A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2):175–183, 2007.

[8] Jarque Carlos M. y Bera Anil K. A test of normality of observations and regressions residuals. *International Statistical Review*, 55(2):163–172, 1987.

[9] Arthur D. y Vassilvitskii S. k-means++: The advantages of careful seeding. En *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, págs. 1027–1035. Society for Industrial and Applied Mathematics (SIAM), New Orleans, USA., 2007.

[10] Pelleg D. y Moore A.W. X-means: Extending k-means with efficient estimation of the number of clusters. En Langley Pat, ed., *Proceedings of the 17th International Conference on Machine Learning*, págs. 727–734. International Machine Learning Society (IMLS), Stanford, USA., 2000.

[11] Sukha D.A., Butler D.R., Comissiong E.A., y Umaharan P. The impact of processing location and growing environment on flavor in cocoa (Theobroma cacao L.) - implications for "terroir" and certification - processing location study. En Mohammed M.A. y Francis J.A., eds., *Proceedings of the III International Conference on Postharvest and Quality Management of Horticultural Products of Interest for Tropical Regions*, págs. 255–262. International Society for Horticultural Science (ISHS), Port of Spain, Trinidad y Tobago, 2014.

[12] Gujarati Damodar N. y Porter Dawn C. *Econometría*. McGraw-Hill/Interamericana Editores, México D.F., México, 5 ed$^{\underline{\text{ón}}}$., 2010.

[13] Peña Daniel. *Análisis de datos multivariantes*. McGraw-Hill/Interamericana de España, Madrid, España, 1 ed$^{\underline{\text{ón}}}$., 2002.

[14] Asociación Nacional de Exportadores de Cacao (ANECACAO). *Manual del Cultivo de Cacao*, 2006.

[15] Asociación Nacional de Exportadores de Cacao (ANECACAO). Cacao CCN51. http://www.anecacao.com/index.php/es/quienes-somos/cacaoccn51.html, 2015. Consultado en 2018-03-25.

[16] Instituto Ecuatoriano de Normalización (INEN). *Norma Técnica Ecuatoriana NTE INEN 176. Cacao en grano. Requisitos*, 2018.

[17] Coordinación General del Sistema de Información Nacional (CGSIN). Boletín situacional cacao. Inf. téc., Ministerio de Agricultura, Ganadería, Acuacultura y Pesca (MAGAP), 2016. Recuperado de http://sipa.agricultura.gob.ec/index.php/situacionales-de-cultivo-2016/boletin-situacional-cacao.

[18] Box G.E.P. y Cox D.R. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[19] Hamerly Greg y Elkan Charles. Learning the k in k-means. En Thrun Sebastian, Saul Lawrence K., y Schölkopf Bernhard, eds., *Advances in Neural Information Processing Systems 16*, págs. 281–288. Neural Information Processing System Foundation (NIPS), Vancouver and Whistler, Canadá, 2004.

[20] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[21] Tinoco Héctor A. y Ospina Diana Y. Análisis del proceso de deshidratación de cacao para la disminución del tiempo de secado. *Revista EIA*, (13):53–63, 2010.

[22] Martínez-Casasnovas J.A. y Bordes X. Viticultura de precisión: Predicción de cosecha a partir de variables del cultivo e índices de vegetación. *Revista de Teledetección*, (24):67–71, 2005.

[23] Nogales Jairo, Graziani de Fariñas Lucía, y Ortiz de Bertorelli Ligia. Cambios fisicos y químicos durante el secado al sol del grano de cacao fementado en dos diseños de cajones de madera. *Agronomia Tropical*, 56(1):5–20, 2006.

[24] Wooldridge Jeffrey M. *Introducción a la econometría: un enfoque moderno*. Cengage Learning Editores, México D.F., México, 4 ed$^{\underline{\text{ón}}}$., 2010.

[25] Fox John y Monette Georges. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

[26] Figueroa Rodríguez Katya Angélica, García García Ana María Teresa, Mayett Moreno Yesica, Hernández Rosas Francisco, y Figueroa Sandoval Benjamín. Factores que explican el rendimiento de caña de azúcar a nivel municipal en México. *Revista Mexicana de Ciencias Agrícolas*, 6(6):1345–1358, 2015.

[27] Brownlee Kenneth Alexander. *Statistical Theory and Methodology in Science and Engineering*. John Wiley & Sons, Inc, New York, USA, 2 ed$^{\underline{\text{ón}}}$., 1965.

[28] Ortiz de Bertorelli Ligia, Graziani de Fariñas Lucía, y Robedas L. Gervaise. Influencia de varios factores sobre características del grano de cacao fermentado y secado al sol. *Agronomía Tropical*, 59(2):119–127, 2009.

[29] Carrera Almeida María Luisa. *Análisis sobre el desarrollo de la comercialización internacional del cacao nacional fino o de aroma del 2002 al 2012, su producción e impacto político, económico y social*. Disertación de grado, Pontificia Universidad Católica del Ecuador, 2014.

[30] Ministerio de Agricultura, Ganadería, Acuacultura y Pesca (MAGAP). La política agropecuaria ecuatoriana: hacia el desarrollo territorial rural sostenible: 2015-2025. I parte, 2016. Recuperado de http://www.agricultura.gob.ec/la-politica-agropecuaria-ecuatoriana-hacia-el-desarrollo-territorial-rural-sostenible-2015-2025/.

[31] Akinwande M.O., Dikko H.G., y Samson A. Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics*, 1(5):754–767, 2015.

[32] Bocco Mónica, Sayago Silvina, Violini Soraya, y Willington Enrique. Modelos simples para estimar rendimiento de cultivos agrícolas a partir de imágenes satelitales: una herramienta para la planificación. En Granitto Pablo M. y Milone Diego, eds., *Anales de las 44 JAIIO. Jornadas Argentinas de Informática e Investigación Operativa*, págs. 26–35. Sociedad Argentina de Informática e Investigación Operativa (SADIO), Rosario, Argentina, 2015.

[33] Rivas Raúl, Ocampo Dora, y Carmona Facundo. Modelo de predicción de rendimiento de trigo a partir de NDVI: aplicación en el contexto de la agricultura

de precisión. En *Anais do XV Simpósio Brasileiro de Sensoramiento Remoto*, págs. 584–590. Instituto Nacional de Pesquisas Espaciais, Curitiba, Brasil, 2011.

[34] Moya Ricardo. Selección del número óptimo de clústers. https://jarroba.com/seleccion-del-numero-optimo-clusters/, 2016. Consultado en 2018-05-30.

[35] Ávila G. de Hernández Rita M, Rodríguez Pérez Vianel, y Hernández Caraballo Edwin A. Predicción del rendimiento de un cultivo de plátano mediante redes neuronales artificiales de regresión generalizada. *Publicaciones en Ciencia y Tecnología*, 6(1):31–40, 2012.

[36] O'Brien Robert M. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 1(41):673–690, 2007.

[37] Rivera Fernández Ruben Darío, Mecías Gallo Freddy Wilberto, Guzmán Cedeño Ángel Monserrate, Peña Galeas Mayra Mercedes, Medina Quinteros Hugo Nolti, Casanova Ferrín Lola Margarita, Barrera Álvarez Alexandra Elizabeth, y Nicela Morante Pedro Eduardo. Efecto del tipo y tiempo de fermentación en la calidad física y química del cacao (Theobroma cacao l.) tipo nacional. *Ciencia y Tecnología*, 5(1):7–12, 2012.

[38] Kodinariya Trupti M. y Makwana Prashant R. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95, 2013.

[39] Breusch T.S. y Pagan A.R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294, 1979.

[40] Anderson T.W. y Darling D.A. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.