



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

**Propuesta metodológica para cuantificar el
impacto de las instituciones educativas en
el rendimiento académico**

TESIS

Que para obtener el título de
Ingeniero Industrial

P R E S E N T A

Gabriel Gaspar Galindo

DIRECTORA DE TESIS

Dra. Esther Segura Pérez



Ciudad Universitaria, Cd. Mx., 2018



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

“[...] Nuestros estudiantes reciben todo de las universidades y cuando salen no sienten la obligación de devolver nada de lo que se les ha dado. ¡No parecen darse cuenta de que su educación la paga el pueblo! [...] Los latinoamericanos no estamos acostumbrados a darle a la patria”

- Miguel Ángel Asturias, Premio Nobel de Literatura 1967, en entrevista con Elena Poniatowska (1967).

Con este trabajo doy inicio a una perpetua retribución a mi alma máter y a mi país.

A mis padres

Agradecimientos

A mi mamá Delia Guadalupe Galindo Reyes y a mi papá Sergio Gaspar Maya, a quienes debo el poder culminar este trabajo y mis estudios universitarios. El logro es de ustedes.

A la Facultad de Ingeniería y a la Universidad Nacional Autónoma de México por la formación académica brindada. Ahora es mi turno de retribuir en algo a la institución y al país.

A mi asesora de tesis Dra. Esther Segura, y a todos los profesores de los cuales he adquirido valiosos conocimientos. Gracias.

Resumen

El rendimiento académico es influido por un sin número de factores, que dependen tanto de los niveles de motivación del estudiante, del contexto socioeconómico y sociodemográfico al que pertenece, así como de factores debidos a la docencia, a los programas educativos y a los programas de apoyo implementados por las instituciones educativas.

En este sentido, las instituciones educativas pueden analizar sus programas de apoyo desde un enfoque de mejora continua, en el que se busque maximizar su impacto en el desempeño académico de los estudiantes. Para ello, se vuelve necesario conocer en primera instancia, en qué medida estos programas influyen en el rendimiento académico de los estudiantes, en función de establecer metas medibles y realistas.

Con base en lo anterior, este trabajo presenta una propuesta metodológica de diez pasos que hace uso de modelos de regresión lineal, a través de su estimación con el método de mínimos cuadrados ordinarios, para aislar el efecto de los factores que influyen en el rendimiento académico y con ello, conocer el impacto de los programas de apoyo instrumentados por las instituciones educativas.

Abstract

Academic performance is influenced by a number of factors, which depend both on the levels of motivation of the student, socio-economic and socio-demographic context to which it belongs, as well as factors due to teaching or support programs implemented by educational institutions.

On the other hand, educational institutions can analyze their support programs from a continuous improvement approach, in which they seek to maximize their impact on the academic performance of students. For this, it becomes necessary to know in the first instance, to what extent these programs influence the academic performance of students, in terms of establishing measurable and realistic institutional goals.

Based on the above, this paper presents a methodological proposal of ten steps that makes use of linear regression models, through its estimation with the ordinary least squares method, to isolate the effect of the factors that influence academic performance. and with it, to know the impact of the support programs implemented by the educational institutions.

Índice

Introducción.....	i
1. El rendimiento académico	1
1.1. Estado del arte	1
1.1.1. Concepto del rendimiento académico.....	1
1.1.2. Variables explicativas del rendimiento académico.....	6
1.1.3. Estudios previos sobre factores que influyen en el rendimiento académico en estudiantes de la UNAM.....	12
1.2. Rendimiento académico en la Facultad de Ingeniería	13
1.3. Resumen del capítulo.....	20
2. Análisis econométricos y el modelo de regresión lineal	22
2.1. Econometría.....	22
2.2. Regresión lineal	28
2.2.1. Estimación de los parámetros por MCO.....	30
2.2.2. Coeficiente de determinación R^2	33
2.2.3. R^2 ajustada.....	35
2.2.4. Multicolinealidad	37
2.2.5. Forma funcional	41
2.2.6. Prueba de hipótesis	48
2.2.7. Valor p	54
2.2.8. Significancia práctica vs Significancia estadística	56
2.2.9. Intervalos de confianza	57
2.2.10. Prueba del estadístico F	58
2.2.11. Variables binarias (<i>Dummy variables</i>).....	63
2.2.12. Heterocedasticidad.....	66
<i>Prueba Breusch-Pagan</i>	67
<i>Prueba de White</i>	68
<i>Medidas correctivas a la heterocedasticidad</i>	70

2.3. Resumen del capítulo.....	73
3. Creación del modelo.....	75
3.1. Propuesta metodológica.....	75
3.1.1. Paso 1. Identificación de variables candidatas.....	75
3.1.2. Paso 2. Especificación primaria de la forma funcional del modelo.....	80
3.1.3. Paso 3. Determinación del tamaño de muestra	84
3.1.4. Paso 4. Identificación de posibles observaciones influyentes	87
3.1.5. Paso 5. Validación de la forma funcional	92
3.1.6. Paso 6. Validación del supuesto de homocedasticidad	96
3.1.7. Paso 7. Validación del supuesto de normalidad.....	99
3.1.8. Paso 8. Comprobación de observaciones influyentes	100
3.1.9. Paso 9. Diagnóstico de multicolinealidad	103
3.1.10. Paso 10. Validación de la significancia práctica y estadística	108
3.2. Diagrama de flujo	112
3.3. Resumen del capítulo.....	114
4. Técnicas de validación	118
4.1. Análisis de coeficientes y valores ajustados	119
4.2. Recolección de nuevas observaciones	121
4.3. Separación de los datos existentes.....	122
4.4. Resumen del capítulo.....	128
Conclusiones	131
Ventajas de la propuesta metodológica	134
Limitaciones de la propuesta metodológica	134
Investigaciones futuras	136
Jerarquización de las instituciones educativas con base en el valor añadido a sus estudiantes ..	136
Diseño de metas óptimas a través de la programación lineal	136

Referencias	139
Anexo A. Supuestos del modelo de regresión lineal	145
Anexo B. Estimación de la varianza poblacional	149
Anexo C. Forma matricial de la estimación por MCO	151
Anexo D. Funciones del lenguaje de programación estadístico R	154

Índice de Tablas y Figuras

Tabla 2.1. Resumen de las formas funcionales logarítmicas.....	42
Tabla 2.2. Resumen de los cálculos necesarios para determinar la variación exacta en Δy	44
Tabla 2.3. Resumen de la naturaleza de la curva debida a la combinación de signos entre β_1 y β_2	47
Tabla 2.4. Resumen de los tipos de errores en una prueba de hipótesis	52
Tabla 3.1. Variables propuestas para controlar factores debidos al estudiante.....	76
Tabla 3.2. Variables propuestas para controlar factores debidos al contexto	77
Tabla 3.3. Variables propuestas para controlar factores debidos al docente.....	79
Tabla 3.4. Variables de interés para la institución educativa.....	79
Tabla 3.5. Variables propuestas para medir el rendimiento académico	80
Figura 1.1. Expectativas del rendimiento académico centrado en la institución	3
Figura 1.2. Elementos medidos por el rendimiento académico	5
Figura 1.3. Factores vinculados al estudiante, a la institución o al contexto.....	7
Figura 1.4. Variables explicativas del rendimiento académico en función del estudiante.....	8
Figura 1.5. Variables explicativas del rendimiento académico en función de la institución	10
Figura 1.6. Variables explicativas del rendimiento académico en función del contexto	11
Figura 1.7. Estrategias de apoyo académico	17
Figura 1.8. Titulados por año.....	18
Figura 2.1. Representación gráfica de la variación explicada y no explicada por la recta de regresión de MCO en el caso especial de la regresión lineal simple.....	34
Figura 2.2. Prueba de hipótesis de dos colas	50
Figura 2.3. Distribución del estadístico F	60
Figura 3.1. Superposición de dos distribuciones.....	85
Figura 3.2. Tipos de observaciones	88
Figura 3.3. Patrones de gráficos de residuales.....	94
Figura 3.4. Gráfica de t_i contra x_j indicativa de heterocedasticidad existente.....	96

Figura 3.5. Gráfica de probabilidad normal de los residuales studentizados externos	99
Figura 3.6. Matriz de la proporción de varianza descompuesta π	106
Figura 3.7. Diagrama de flujo para la creación del modelo.....	112
Figura 3.8. Diagrama de flujo para la creación del modelo (continuación)	113

Nomenclatura

β_0	Intercepto
β	Parámetro poblacional del modelo de regresión lineal
x	Regresor, variable independiente y/o variable explicativa
μ	Término de error o perturbación
Y	Variable explicada o dependiente
$\hat{\beta}_0$	Estimación del intercepto
$\hat{\beta}$	Estimación del parámetro poblacional
$\hat{\mu}$	Residual
\hat{Y}	Estimación de la variable dependiente o explicada
n	Total de observaciones
k	Total de variables explicativas involucradas en el modelo

Subíndices

i	i-ésima observación
j	j-ésima variable independiente

Superíndices

T	Transpuesta
-----	-------------

Introducción

La educación puede ser analizada desde un enfoque de mejora continua, en el aspecto de que es posible medir y, en consecuencia, mejorar muchos de los procesos y programas que la integran. De esta premisa nace el proyecto de investigación presentado en esta tesis.

Considerando que la educación, y específicamente, el rendimiento académico, es influido por un sin número de factores, relativos al estudiante, al contexto sociodemográfico y socioeconómico del mismo, así como relativos a la propia institución educativa, ¿cómo es posible establecer metas de cumplimiento o indicadores de desempeño para las instituciones educativas, si éstas sólo controlan un número limitado del total de factores que influyen en el rendimiento académico? ¿De qué forma se puede medir el efecto de las instituciones educativas en el rendimiento académico, de tal manera que el efecto de otros factores se mantenga aislado?

Este trabajo propone el uso de modelos de regresión lineal, a través de su estimación con el método de mínimos cuadrados ordinarios, para aislar el efecto de cada factor y poder cuantificar su impacto en el rendimiento académico de los estudiantes. Con esto, será posible identificar el efecto que tienen las instituciones educativas, medido a través del impacto de programas de apoyo, tales como los programas de tutoría, de asesorías académicas, etc., en el rendimiento académico.

El primer capítulo de este trabajo, se integra por una investigación sobre el estado del arte que guarda el uso de modelos de regresión lineal, en la ponderación de factores que influyen en el rendimiento académico. Se analizan, además, las variables utilizadas para explicar el rendimiento académico en diversos estudios previos.

En el segundo capítulo se define el término *econometría* y se analizan los tipos de datos con los que es posible estimar modelos de regresión lineal. Así mismo, en este capítulo se presenta la estructura de los modelos de regresión lineal y la

matemática detrás del funcionamiento del método de mínimos cuadrados ordinarios para su estimación. Se explican a detalle los estadísticos y herramientas comúnmente usados en el análisis de este tipo de modelos.

En el tercer capítulo se presenta una propuesta metodológica de diez pasos que, haciendo uso de modelos de regresión lineal, permitirá cuantificar el impacto de acciones emprendidas por las instituciones educativas en el rendimiento académico; tales como programas de asesorías, de tutoría, talleres de ejercicios, etc. Se presenta además, un diagrama de flujo detallado para la aplicación de esta metodología. Cabe decir que, para cada operación matemática o análisis estadístico, se presenta la función o la combinación de funciones en el lenguaje de programación estadístico R, que permitirán realizar dichas operaciones de manera automática.

En el último capítulo de este trabajo, se presentan tres técnicas de validación de modelos de regresión lineal, que buscan conocer tanto la capacidad de ajuste a las observaciones, como la capacidad de predecir nuevas observaciones de manera precisa, de los modelos construidos y estimados.

Problemática abordada.

¿En qué medida las escuelas, colegios o universidades, pueden mejorar sus procesos educativos?

Para poder dar respuesta a esta pregunta, se debe conocer cuál es el estado actual de dichos procesos. Y para ello, a su vez, se debe saber cómo medir el impacto de estos; como bien lo señaló Peter Drucker: “Lo que no se mide, no se puede mejorar” (Aiteco Consultores, 2016).

Si el nivel de aprendizaje de los estudiantes (ya sea que se encuentre dentro de los estándares establecidos o no) es el producto obtenido de cualquier proceso educativo, el *rendimiento académico* puede ser entendido como un indicador de este nivel de aprendizaje e inclusive, del proceso educativo.

Considerando que el rendimiento académico es influido por una variedad de factores, algunos dependientes del estudiante, otros dependientes de la institución educativa, y algunos otros fuera del control de ambos (como el nivel socioeconómico y aspectos sociodemográficos); y conociendo además, que para poder mejorar los procesos educativos, se debe medir el impacto actual de los programas de evaluación, de apoyo al estudiante, etc.; la problemática abordada en esta tesis, consiste en establecer una manera de medir el impacto que guardan las políticas de la institución con respecto al nivel de aprendizaje de los estudiantes.

Justificación.

La utilidad en identificar el impacto, de manera cuantitativa, de las instituciones educativas en el rendimiento académico, deriva en la posibilidad de establecer **metas** e **indicadores** tanto para los procesos como para sus programas de enseñanza. El contar con metas e indicadores permitirá dar seguimiento a los programas de tutoría, asesorías, exámenes extraordinarios, etc.; así mismo, será posible evaluar el cumplimiento de las políticas y objetivos de enseñanza. Todo esto en función de aportar valor al proceso de mejora continua.

Recordando que una meta y, en consecuencia, su indicador de cumplimiento, deben ser medibles; contar con una manera de cuantificar el impacto de, por

ejemplo, las asesorías académicas en el rendimiento académico, permitirá establecer una línea base de la que partirá la definición de una meta alcanzable, asociada a los objetivos del programa educativo.

Sin una medida asociada al impacto del programa (por ejemplo, las asesorías académicas) en el rendimiento académico, no es posible establecer una meta medible y alcanzable, con base en los recursos disponibles de la institución, dado que no se cuenta con la referencia necesaria del impacto actual de dicho programa.

Por otra parte, cabe destacar que las T.I.C. (Tecnologías de la Información y Comunicación) empiezan a jugar ya un papel importante en el sector educativo. Aunado al aprendizaje en línea, caracterizado por eliminar la necesidad del estudiante y el profesor de compartir el mismo espacio físico para poder comunicarse, se han agregado en los últimos años tecnologías disruptivas tales como la realidad virtual y la realidad aumentada, las cuales prometen dirigir los alcances de la educación hacia nuevos horizontes.

Con un abanico tan amplio y en continua expansión como lo son las T.I.C., cabe preguntarse: ¿en qué medida estas tecnologías impactarán en el rendimiento académico?, ¿el beneficio obtenido justifica el costo asociado a su uso?, ¿existe un incremento en el valor agregado a los estudiantes con el uso de T.I.C.'s? De ser así, ¿en qué medida?

La cuantificación del impacto de las instituciones educativas en el rendimiento académico de los estudiantes, no se limita en conocer la influencia actual de sus programas y procesos educativos. Si bien esta información es relevante para la toma de decisiones en el corto plazo, su importancia a largo plazo radica en la posibilidad de establecer una estrategia de mejora continua que involucre desde el análisis de programas de apoyo, como las asesorías académicas, hasta el análisis del efecto en el uso de nuevas tecnologías, como la realidad aumentada, en los procesos de enseñanza.

Objetivo general.

Diseñar una metodología que permita medir el impacto de acciones emprendidas por las instituciones educativas en el rendimiento académico.

Objetivos específicos.

- Identificar las variables o factores comúnmente asociados con la variación del rendimiento académico.
- Diferenciar el impacto de variables dependientes del estudiante, de la institución educativa, y de aquellas que escapan del control de ambos, en el rendimiento académico.
- Conocer los supuestos en los que se basa la estimación de un modelo de regresión lineal mediante el método de mínimos cuadrados ordinarios, así como el efecto del incumplimiento de estos en la estimación de los parámetros.
- Comprender las principales técnicas de validación de un modelo de regresión lineal.

Hipótesis.

Es posible contar con un procedimiento o metodología, que permita medir el impacto individual de las variables que afectan el rendimiento académico de los estudiantes, diferenciando entre aquellas que son dependientes del estudiante y de la institución educativa, así como de aquellas que escapan del control de ambos.

Alcances y limitaciones.

Este trabajo es el producto de una investigación que pretende establecer una base metodológica para la estimación de modelos de regresión lineal, a través del método de Mínimos Cuadrados Ordinarios MCO. Esto, considerando las consecuencias del incumplimiento de los supuestos del Modelo de Regresión Lineal Clásico MRLC, la presencia de altas dependencias lineales entre las variables independientes, así como la existencia de observaciones influyentes.

Así mismo, la investigación realizada también incluye técnicas de validación para Modelos de Regresión Lineal MRL.

El alcance de este trabajo se limita al análisis de datos obtenidos en un punto dado en el tiempo.

Capítulo I.

El rendimiento académico



Vista panorámica del edificio de rectoría, UNAM
Foto: www.fundacionunam.org.mx

1.1 Estado del arte

1.1.1 Concepto del rendimiento académico.

En la búsqueda por saber qué estado guardan las investigaciones y estudios referentes al rendimiento académico, se vuelve indispensable conocer en primera instancia: ¿Qué entendemos en la actualidad por *rendimiento académico*? ¿El concepto ha diferido respecto al tiempo? ¿Respecto a diversos autores?

Ciertamente dicho concepto no resulta extraño o ajeno en el bagaje lingüístico de la sociedad, es un término usado de manera común e indistintamente por la mayoría de sus integrantes: estudiantes, docentes, directivos, padres de familia, líderes políticos, etc. Este uso, me atrevo a sintetizar, se ha referido mayoritariamente al nivel de aprendizaje encontrado en los estudiantes (ya sea que se encuentre dentro de los estándares establecidos o no), ponderado por las calificaciones obtenidas en sus cursos.

Sin embargo, con base en la revisión histórica de estrictas investigaciones sobre educación, así como en libros escritos por profesionales sumamente experimentados en la materia, cabe decir que dicha conceptualización del imaginario colectivo debe ser enriquecida por las aportaciones de estos grandes autores.

Es importante conocer que existen diversas concepciones sobre su significado, por lo que no se ha llegado a un acuerdo común sobre la definición estricta de

rendimiento académico. Encuentro que las diferencias se han debido principalmente, a las diversas aproximaciones con las que este fenómeno ha sido estudiado.

Una de las definiciones más sencillas y recientes es la presentada por Juan L. Castejón (2014) en la que establece que “La definición operativa y medida de los resultados cognitivos del aprendizaje es a lo que se denomina rendimiento académico”. En esta definición el autor atribuye al rendimiento académico como el producto del aprendizaje, como la manera de medir el aprendizaje, siendo éste un “constructo no observable ni medible de forma directa” (Castejón, 2014). En este esquema, el rendimiento académico se encuentra en función del estudiante, aunque asegura también que es influido por una multitud de variables.

Esta definición sostiene sin lugar a dudas lo que es entendido *per se* en el imaginario colectivo, sin embargo, otros autores abordan el tema de una manera más integradora. Tal es el caso de Dugan y Herson (2002) que en palabras de Noel Rodríguez-Ayán (2007) “señalan la importancia de distinguir los indicadores centrados en las instituciones [educativas]...de los indicadores de aprendizaje de los estudiantes” (Rodríguez-Ayán, 2007). Con este entendimiento, el enfoque para definir el rendimiento académico ya no se limita sólo a las capacidades o habilidades del estudiante, sino que ahora abarca también las habilidades y capacidades de la institución educativa.

Esta visión más integradora da pauta a una diferenciación en la definición del concepto, pudiéndose centrar ahora también en la institución educativa “como entidad responsable de satisfacer las expectativas de la sociedad respecto a la Educación Superior” (Rodríguez-Ayán, 2007).

Este enfoque ha sido estudiado por Reynolds (1990) definiendo el concepto en función de diversos sectores de la sociedad. La figura 1.1 muestra la alineación de los sectores en función de sus expectativas, con respecto al rendimiento académico (centrado en la institución).



Figura 1.1 Expectativas del rendimiento académico centrado en la institución.
Elaboración propia con base en la síntesis de Rodríguez-Ayán (2007).

De la definición con enfoque al estudiante, así como de las diversas definiciones con enfoque a la institución, se evidencia un patrón compartido: todas estas definiciones de *rendimiento académico* lo evalúan a partir de los resultados (o productos) obtenidos, ya sea por el estudiante o por la institución. Es decir, se basan en un sistema entradas – salidas.

Este sistema ha sido estudiado ampliamente por Murillo (2003) quien identifica esta concepción del entendimiento de lo que es el rendimiento académico como “Productividad Escolar”, entendido como todos aquellos estudios cuyas raíces y desarrollo son estrictamente economicistas, ya que buscan optimizar los insumos para conseguir los productos (entendiéndose como eficiencia). En otras palabras, los estudios que se han originado a partir de lo entendido por *rendimiento académico* explicado en párrafos anteriores, han tenido la finalidad de maximizar el producto (p.e. el aprendizaje de los estudiantes o la capacidad adaptativa de egresados) manipulando variables de entrada tales como los recursos financieros destinados a la institución o la cantidad de material didáctico disponible, entre otros.

Un claro ejemplo de este tipo de estudios es el realizado por Virreira (1979) quien se enfocó en establecer la manera de disminuir los costos de funcionamiento del sistema escolar manteniendo su rendimiento constante, o viceversa, al incrementar el rendimiento manteniendo constantes los costos. El resultado, un tanto previsible, fue que al realizar una mayor inversión en recursos didácticos se obtenía un mayor rendimiento (Murillo, 2007).

Pero esta línea de investigación, enfocada en la “Productividad Escolar” no ha sido la única utilizada para estudiar los fenómenos educativos. Existe otra línea de investigación que en términos de Murillo (2003) debe ser claramente diferenciada de la primera, denominada “Eficacia Escolar”. Entendiendo a ésta última como la conformada por “estudios empíricos que buscan, por un lado, conocer qué capacidad tienen las escuelas para incidir en el desarrollo de los alumnos y, por otro, conocer qué hace que una escuela sea eficaz” (Murillo F. J., 2003) entendiendo a la eficacia como “[aquellos] procesos que hacen que se consigan mejor los objetivos” (Murillo F. J., 2003).

Esta línea de investigación, denominada “Eficacia Escolar”, ha sido liderada por el Movimiento de Investigación sobre Eficacia Escolar que en palabras de Gala-Emma Peñalba (2003): “analiza los factores que influyen en que los alumnos y alumnas de unos centros obtengan mejores resultados que los de otros situados en contextos socio-económicos análogos, así como los elementos que contribuyen a que los centros de enseñanza alcancen sus objetivos educativos” (Gala-Emma Peñalba, 2003).

Estas aportaciones de “Eficacia Escolar” dejan en claro que para un correcto entendimiento de lo que es el *rendimiento académico* no basta con considerar un sistema entradas – salidas (evaluado por sus resultados o productos) sino más bien, es necesario considerar un sistema entradas – proceso – salidas (evaluado tanto por el proceso como por los resultados obtenidos).

Entrando en mayor detalle sobre la conceptualización de “Eficacia Escolar”, Murillo (2003) hace un análisis de sus tres características más importantes:

- *Valor añadido*. Entendido como el valor que le aporta la institución al alumno.
- *Equidad*. Entendida como la capacidad de la institución para agregar dicho valor a cualquier alumno, no importando sus diferencias socio-económicas, lingüísticas, de género, o de cualquier otra índole.
- *Desarrollo Integral*. Entendido como la flexibilidad de la institución para agregar valor no sólo en las cuestiones ortodoxas (como pudieran

ser Matemáticas o Lengua) sino también en cuestiones afectivas, tales como el autoconcepto o la actitud creativa y crítica.

Con esto en mente, se ha llegado a una definición precisa de lo que una escuela eficaz significa, que en palabras del mismo autor se entiende como “[el] desarrollo integral de todos y cada uno de sus alumnos mayor de lo que sería esperable teniendo en cuenta su rendimiento previo y la situación social, económica y cultural de las familias” (Murillo F. J., 2003).

Como se mencionó al inicio de este apartado, los enfoques con los que se ha buscado definir qué se entiende por *rendimiento académico* así como su forma de evaluarlo han sido muy diversos, algunos tomando en cuenta sólo a un elemento para su definición (el estudiante) u otros considerando también a la institución, integrando con ello nuevas perspectivas. Así mismo, enfoques más integradores establecen la necesidad de medir no sólo los resultados obtenidos sino también los procesos involucrados para poder concebir una definición más holística del concepto.

Para fines de este trabajo de tesis, el rendimiento académico será entendido como un *producto* a evaluar (tasa de deserción, tasa de titulación, tasa de reprobación, avance de créditos esperado, etc.) obtenido del sistema Contexto – Institución – Estudiante.

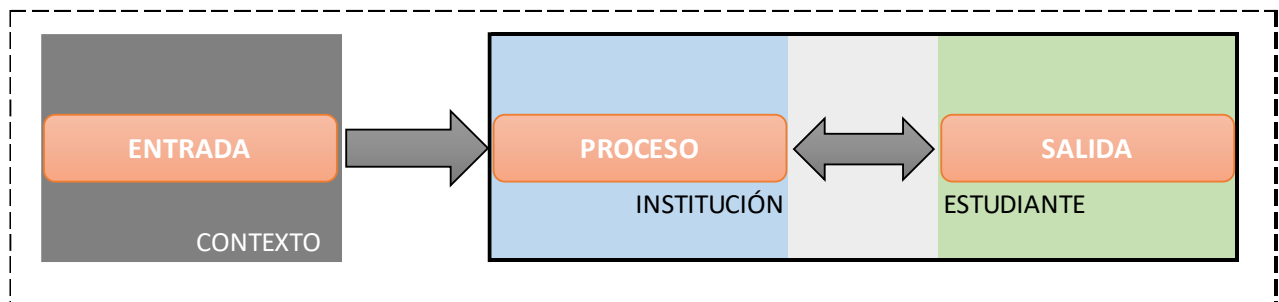


Figura 1.2 Elementos medidos por el rendimiento académico. Elaboración propia.

1.1.2 Variables explicativas del rendimiento académico.

Existe una miríada de trabajos de investigación referentes al análisis de las causas que influyen en el rendimiento académico de los estudiantes. Además, el abanico de los estudios realizados es muy amplio debido a los diferentes niveles de educación que han sido considerados en los análisis.

Al realizar una búsqueda en la base de datos ERIC (*Education Resources Information Center*) la mayor en lo que respecta a información especializada en educación, se encontraron 175,055 resultados que respondieron a las palabras clave “*factors of student achievement*”. La cantidad de publicaciones en otras bases de datos tales como *Taylor & Francis Online* (que a su vez contiene bases de datos del *Educational Research Abstract Online* o del *Research into Higher Education*) supera los 45,000.

El número de estudios que han analizado las causas del rendimiento académico responde a la variedad de aproximaciones y metodologías empleadas, así como al grupo o población en la que se han enfocado dichos estudios.

De los resultados obtenidos de estos análisis cabe destacar que, algunas de las causas o factores encontrados como significativos en niveles de educación básica no lo son más en niveles de educación superior, ya que el contexto tanto del estudiante como de la institución son radicalmente distintos. Sin embargo, es de considerar que algunos de estos factores pueden ser contextualizados al nivel de educación superior, por lo que los análisis realizados a nivel de educación básica pueden servir de sustento a la elección de dichos factores en niveles superiores.

Asimismo, ha sido muy importante para efectos de este trabajo de tesis diferenciar entre los distintos enfoques utilizados para el análisis de las variables explicativas, encontrando que en términos generales se pueden dividir en dos: aquellos orientados a la psicología (que consideran en mayor medida aspectos motivacionales, emocionales o cognitivos) y aquellos orientados a las ciencias de la educación (que consideran en mayor medida aspectos pedagógicos).

Finalmente se puede realizar una última categorización con base en el elemento del sistema a estudiar; es decir, con enfoque en el estudiante (variables dependientes de la persona) con enfoque en la escuela (variables dependientes de la institución) o

con enfoque en la sociedad (variables de contexto tanto del estudiante como de la institución).

Como resultado de lo anterior, y para tener una comprensión más adecuada de los factores que afectan al rendimiento académico, se presenta en la siguiente figura una síntesis de aquellos vinculados al estudiante, a la institución o al contexto.

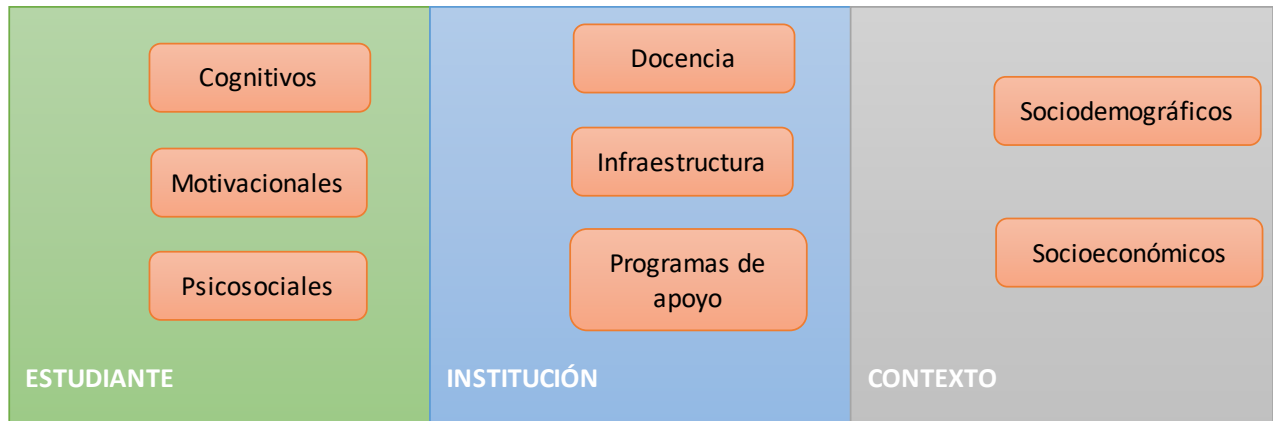


Figura 1.3 Factores vinculados al estudiante, a la institución o al contexto. Elaboración propia.

Cada factor engloba variables que se categorizaron a consideración propia, tomando en cuenta el modelo sobre factores explicativos de Castejón (2014), las variables consideradas en estudios como los de Rodríguez-Ayán (2007) o la Investigación Iberoamericana sobre Eficacia Escolar (2007) así como los diversos análisis encontrados en la literatura (Tomás-Miquel, Expósito-Langa, & Sempere-Castelló, 2014; Barahona, 2014; Garbanzo & Guiselle, 2007; Bailey, Taasobshirazi, & Carr, 2014; Reyes Carreto, Godinez Jaimes, Ariza Hernández, Sánchez Rosas, & Torreblanca Ignacio, 2014; García Jiménez, Alvarado Izquierdo, & Jiménez Blanco, 2000).

Es importante mencionar que no se excluye la posible relación entre dichos factores, por tanto, variables que se clasifican en un factor pudieran ser clasificadas en otro por alguien más. La finalidad de la clasificación es obtener un panorama general de las variables consideradas en los diversos estudios previos.

La figura 1.4 presenta las variables contenidas en los factores referentes al estudiante.

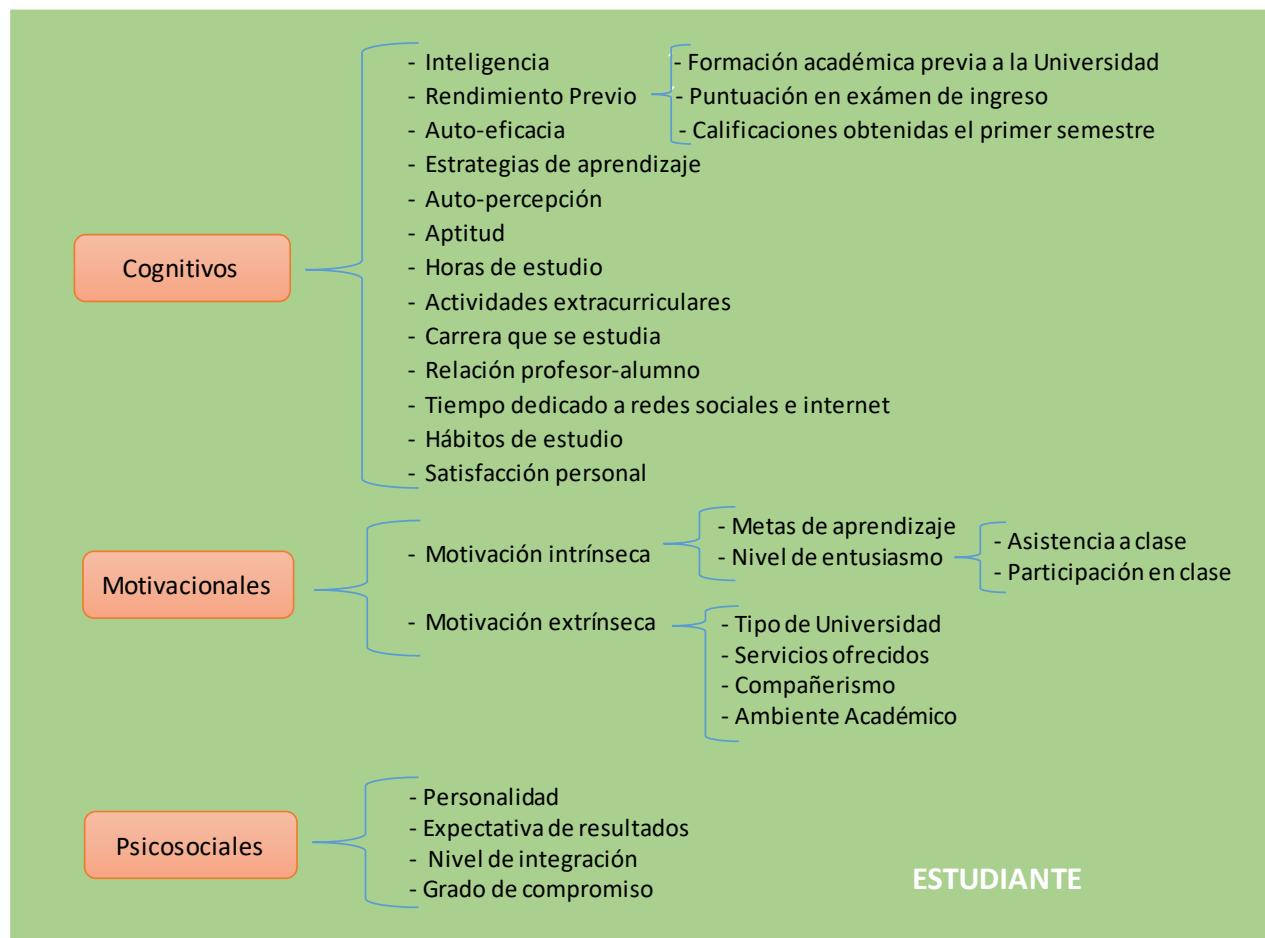


Figura 1.4 Variables explicativas del rendimiento académico en función del estudiante. Elaboración propia.

Las variables clasificadas dentro del factor cognitivo responden a la capacidad del estudiante para cumplir una determinada tarea, la percepción sobre su capacidad y las habilidades intelectuales que posea; así como al nivel de conocimientos adquirido.

Las variables integradas dentro del factor motivacional responden al “estado psicológico relacionado con los estudios que es positivo y significativo” (Salonava *et al.*, 2005). Este estado es influenciado por factores tanto internos como externos al estudiante, y la combinación de sus habilidades cognitivas como de los estímulos

motivacionales ha sido encontrado sumamente importante en la explicación del rendimiento académico.

Por otra parte, las variables integradas en el factor psicosocial son las relacionadas a la conducta humana; identificando como las principales la *personalidad*, el *nivel de integración*, el *grado de compromiso* y la *expectativa de resultados* que tenga el estudiante.

En la figura 1.5 se presentan las variables referentes a la institución. Las variables incluidas dentro del factor *docente*, corresponden a todas aquellas actitudes, aptitudes, conocimientos y técnicas de enseñanza utilizadas por la planta docente.

Por otro lado, el rendimiento académico de los estudiantes puede estar relacionado con aspectos infraestructurales de la misma universidad (Salonava *et al.*, 2005; Crampton & Thomson, 2011; Garbanzo & Guiselle, 2007) por lo que variables como la condición de las aulas y de las bibliotecas, así como la disponibilidad del equipo audiovisual son incluidas en este factor.

En dicha figura también se encuentra como factor relevante los *programas de apoyo*, en los que se consideran variables tales como becas económicas y de especie, asistencia médica o psicológica y cursos extracurriculares, por mencionar algunos.

Dentro de las variables encontradas en la literatura como explicativas del rendimiento académico, se identificaron algunas que no han podido ser categorizadas en alguno de los factores antes mencionados, dada su naturaleza intrínseca al sistema más que a la propia institución. Tales variables son la *vía de ingreso a la carrera* (primera elección o cambio de carrera) los *horarios de clase*, la *complejidad de las materias*, la *duración del plan de estudios* o el *número de grupos por materia*.

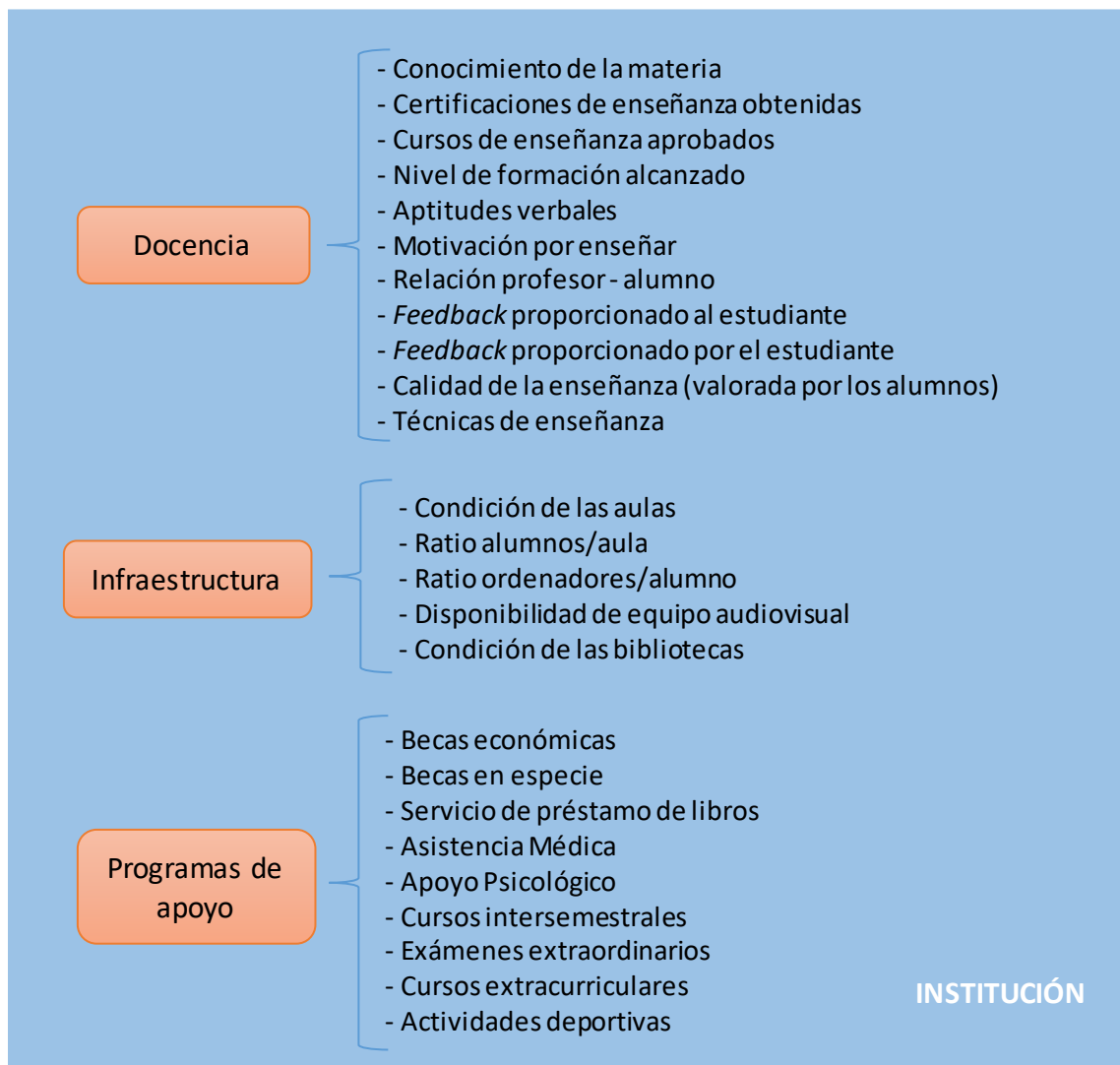


Figura 1.5 Variables explicativas del rendimiento académico en función de la institución.

Elaboración propia.

Finalmente, también se han identificado aquellas variables referentes al *contexto*; es decir, que no dependen del individuo en sí, ni de la institución. Su importancia radica en las interrelaciones que pueden producir entre sí y entre las variables que se encuentran en función del estudiante y de la institución (Garbanzo & Guiselle, 2007).

El cuadro resumen de dichas variables se presenta en la siguiente figura.

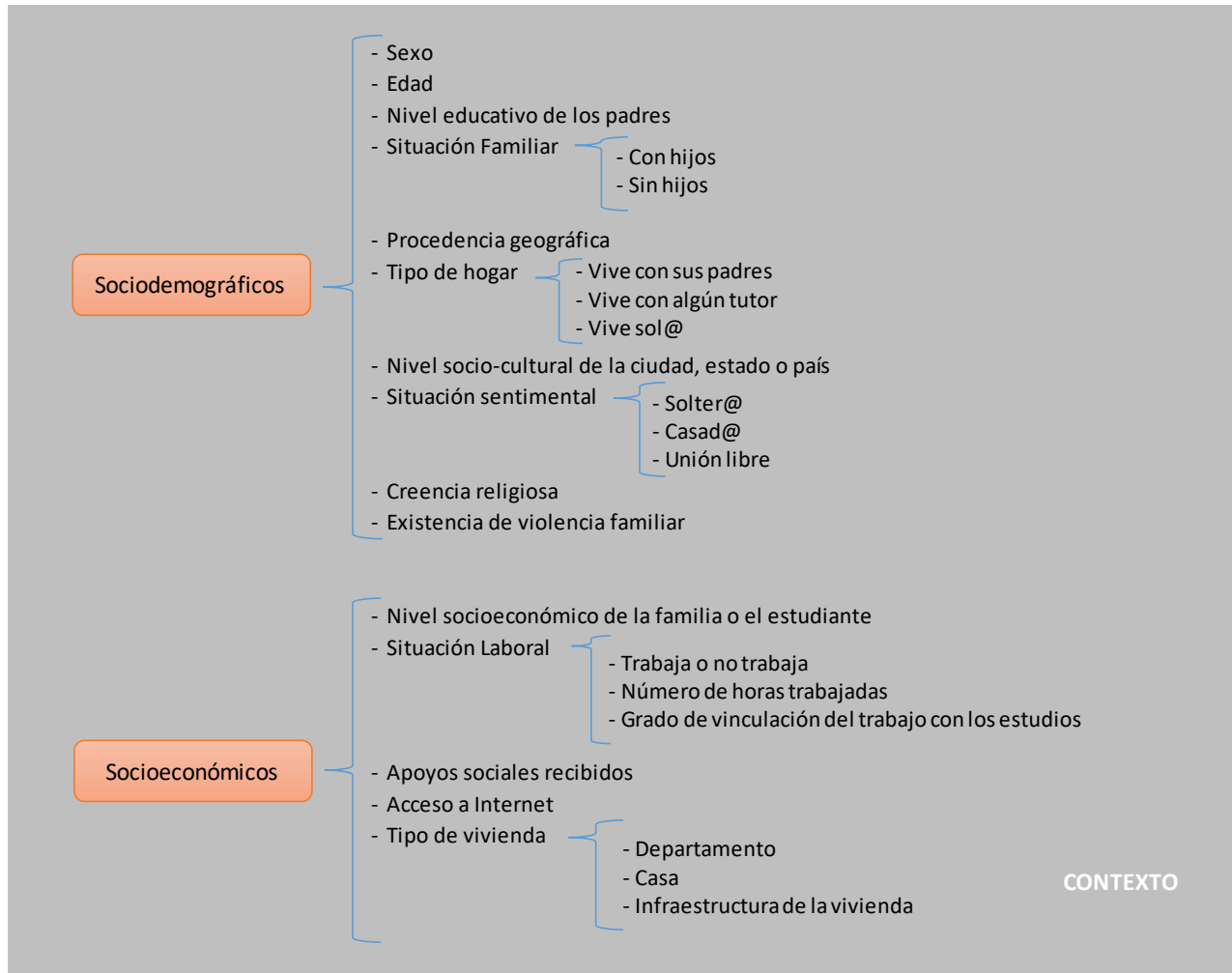


Figura 1.6 Variables explicativas del rendimiento académico en función del contexto. Elaboración propia.

Las variables explicativas del rendimiento académico referentes al contexto del estudiante, clasificadas en un factor sociodemográfico y en un factor socioeconómico, responden a las condiciones de entrada a la universidad que el individuo presenta, independientemente de sus capacidades cognitivas o de los servicios institucionales; por ello, su inclusión en cualquier análisis causal del desempeño estudiantil es imperante.

1.1.3 Estudios previos sobre factores que influyen en el rendimiento académico en estudiantes de la UNAM.

En la UNAM, también se han realizado diversos estudios referentes al rendimiento académico y los factores que lo influyen. Las Facultades de Estudios Superiores (FES) Iztacala y Aragón han sido las más activas en este ámbito.

Alvarado, Cepeda y Del Bosque (2014) encontraron en su estudio *Comparación de estrategias de estudio y autorregulación en universitarios* que los factores con mayor influencia en el rendimiento académico en estudiantes de Psicología de la FES Iztacala son las estrategias de estudio y la capacidad de autorregulación del estudiante, identificando la necesidad de que las universidades “diseñen e implementen programas para acrecentar la motivación de los estudiantes y la autorregulación en el proceso de aprendizaje, lo que puede mejorar la toma de conciencia y control sobre lo que se va a aprender” (Alvarado *et al.*, 2014), mejorando con ello el rendimiento académico.

Por otra parte, en la misma Facultad (FES-I) pero con enfoque en estudiantes de medicina de primer año, Osornio, Valadez, Cuellar y Monje (2008) encontraron que el tiempo de traslado y los ingresos económicos son los dos factores más importantes para explicar el rendimiento de los estudiantes. El estudio estuvo enfocado en analizar factores sociodemográficos, por lo que la capacidad de intervención de la institución quedó excluida; sin embargo, la importancia de su estudio radica en la creación de un perfil de entrada de sus estudiantes, en función de variables de contexto, con la finalidad de determinar un rendimiento académico esperado.

De los estudios encontrados, el realizado por Daniel Velázquez (1999) en estudiantes de Ingeniería Civil de la actual Facultad de Estudios Superiores (FES) Aragón, ha sido el más completo, además, cabe destacar el área de estudio de los estudiantes, que no pertenece a las ciencias sociales.

En su investigación de campo *Determinación de las Variables Cualitativas y Cuantitativas que influyen en el Rendimiento Académico de los alumnos de la Carrera de Ingeniería Civil del Campus Aragón – UNAM*, Velázquez (1999) encontró que los ambientes familiar y social son los que más influencia tienen en el

rendimiento de los estudiantes, siendo “el nivel educativo de sus familiares, los valores recibidos en su hogar y en su contexto social [...] la mayor influencia para el aprovechamiento escolar” (Velázquez, 1999).

Un aspecto interesante a considerar de sus resultados es que “el ambiente universitario no ha sido factor de cambio en sus actitudes. Desde el primer año hasta su egreso se perciben a sí mismos sin cambios significativos” (Velázquez, 1999).

Otras investigaciones conducidas por el mismo autor han aportado valiosa información referente al desempeño del personal docente y su influencia en los procesos de aprendizaje en estudiantes de ingeniería. Tal es el caso del *Estudio Longitudinal de la Evaluación del Desempeño del Personal Docente en los Procesos de Aula: el caso de Ingeniería Civil, Ingeniería en Computación, e Ingeniería Mecánica Eléctrica de la ENEP Aragón¹*, UNAM (2004), en el que descubrió que la evaluación del personal docente por su desempeño en los procesos de aula es independiente de la calificación otorgada a los alumnos, pero dependiente principalmente de los métodos de enseñanza y de los conocimientos que tengan sobre la materia, entre otros.

1.2 Rendimiento académico en la Facultad de Ingeniería

Una vez establecido el estado del arte que guardan las investigaciones referentes al tema de esta tesis, es necesario conocer ahora el estado que guardan las estrategias y apoyos brindados por parte de la Facultad de Ingeniería para cumplir con las metas académicas de los estudiantes en los tiempos establecidos, así como los resultados obtenidos.

Es importante conocer también los instrumentos utilizados por la Facultad para obtener información de los alumnos, así como para medir el avance en la consecución de las metas institucionales, analizando los productos del rendimiento académico (tasa de deserción, tasa de titulación, tasa de reprobación, avance en créditos, etc.) que han sido objeto de evaluación.

¹ Actualmente FES Aragón

Uno de los principales actores en la evaluación educativa dentro de la Facultad ha sido la Coordinación de Evaluación Educativa (CEE) cuya misión es:

Proponer y desarrollar lineamientos, métodos, instrumentos, estudios y programas de evaluación y desarrollo educativo con el propósito de fortalecer la formación integral de los alumnos y la mejora continua de los docentes de la Facultad de Ingeniería (CEE, 2016).

Las principales funciones que cumple dicha Coordinación son (CEE, 2016):

- Proponer proyectos, métodos, estudios de evaluación y desarrollo educativo que permitan a la Facultad de Ingeniería cumplir con sus fines sustantivos.
- Apoyar la coordinación, desarrollo y evaluación de procedimientos destinados a asegurar la acreditación de las carreras de la Facultad de Ingeniería ante organismos certificadores.
- Proponer instrumentos de evaluación y desarrollo educativo que contribuyan a beneficiar la calidad del desempeño de los docentes.
- Difundir en la comunidad los resultados de los estudios de evaluación, realizados en la CEE.
- Elaborar la evaluación del perfil de los alumnos de nuevo ingreso a la Facultad de Ingeniería.

Como resultado de su trabajo, la Coordinación ha podido crear perfiles de los alumnos de nuevo ingreso desde hace 20 años. Esto, con la aplicación de instrumentos como el *Cuestionario Sociodemográfico y de Antecedentes Escolares*, así como del *Sistema de Valoración de Conductas Orientadas hacia el Estudio (SIVACORE)*.

La aplicación del Cuestionario Sociodemográfico y de Antecedentes Escolares busca obtener información de la población estudiantil de recién ingreso en 4 áreas específicas (CEE, 2017):

- I. **Datos generales.** Comprende las características básicas del alumnado.

- II. **Perfil escolar.** Se explora la trayectoria escolar de los alumnos, en términos de sus estudios de bachillerato, el lugar e institución en que los realizaron, así como promedio y duración de los mismos.
- III. **Indicadores socioeconómicos.** En esta área se muestran resultados de algunas características socioeconómicas del alumnado tales como ingreso mensual en el hogar, situación laboral, bienes y servicios con que cuenta, etc.
- IV. **Hábitos y distribución del tiempo.** Recaba información relacionada con los hábitos de lectura, las actividades que se realizan en la biblioteca y el tiempo que emplean en diferentes actividades.

Mientras que la función del SIVACORE es evaluar las conductas de estudio que presentan los estudiantes (CEE, 2017), particularmente:

- i. La *Iniciativa* relacionada con la capacidad del estudiante de tomar decisiones académicas para llevar actividades por su cuenta.
- ii. La *Organización del trabajo escolar* relacionada con la organización y distribución de actividades escolares que permitan aprovechar al máximo los estudios.
- iii. La *Motivación escolar* que refiere a la disposición y voluntad del estudiante en el cumplimiento de sus actividades escolares.
- iv. La *Integración* que refiere a la capacidad del estudiante de adaptarse a la carrera y al ambiente escolar.
- v. La *Autorregulación* referida a la capacidad que tienen los estudiantes de valorar y corregir el desarrollo de sus actividades escolares.
- vi. La *Concentración* en términos de la capacidad de los estudiantes para mantener fija su atención en un objeto en profundidad y durante largo tiempo.
- vii. La *Administración del tiempo* que describe como planifican, organizan y controlan los estudiantes sus actividades escolares.

Finalmente, como complemento al trabajo de la CEE, la División de Ciencias Básicas realiza un *Diagnóstico de conocimientos previos o de antecedentes académicos* mediante una prueba estandarizada y en línea a todos los alumnos de nuevo ingreso.

Todos estos instrumentos han permitido la identificación de patrones, facilitando la creación y evaluación de perfiles de los alumnos.

Tal es el caso del estudio conducido por Ibarra García y Medina Mora (2013) en el cual, a través de una técnica de minería de datos conocida como *regla de inducción* encontraron que ciertas preguntas del SIVACORE definen con mayor exactitud a los estudiantes que aprueban todas sus asignaturas en el primer semestre.

Entre los resultados presentados por los autores se encuentra que, por ejemplo, la mayoría de los estudiantes (11.26%) que contestaron *nunca* a la pregunta “En los exámenes sucede que me preguntan temas que no revisé” y *siempre* a la pregunta “Me presento al primer día de clases puntualmente” aprobaron todas las materias del primer semestre.

De igual manera, la mayoría (30.32%) de quienes contestaron *siempre* a la pregunta “Mis compañeros tienen una buena opinión de mí como estudiante” y *siempre* a la pregunta “Tengo un lugar organizado para guardar útiles escolares” fueron quienes aprobaron las 5 materias del primer semestre.

La metodología también sirvió para detectar patrones de bajo rendimiento académico (aquellos estudiantes que sólo aprobaron 2 o menos materias en el primer semestre).

Las conclusiones generales encontradas en el estudio indican que los hábitos que mejor caracterizan a un estudiante que aprueba todas sus materias son: sus compañeros tienen una buena opinión de él o ella como estudiante, en general sabe organizarse para el estudio, puede o no tener un lugar de estudio, ya estudió todos los temas correspondientes al momento de presentar un examen, no se siente frustrado como estudiante y siempre participa en clase (Ibarra y Medina, 2013)

Por otra parte, el seguimiento a los alumnos no se limita a los de nuevo ingreso. La Facultad ha establecido estrategias y programas de apoyo para estudiantes de semestres avanzados, con la finalidad de reducir las tasas de reprobación, sobre todo en asignaturas con altos índices de rezago.

Los programas de apoyo impulsados consideran talleres de ejercicios, cursos extraordinarios, talleres de preparación para exámenes extraordinarios, conferencias clase, tutorías, entre otros. Así mismo, se han creado recientemente otros instrumentos de evaluación, tal es el caso de un nuevo examen estandarizado en línea, de aplicación semestral, desarrollado por la División de Ciencias Básicas, cuyo objetivo es evaluar el aprendizaje alcanzado por cohorte generacional al finalizar un semestre.



Figura 1.7 Estrategias de apoyo académico. Tomado del Segundo Informe de Actividades 2016, Facultad de Ingeniería.

Además de evaluar los índices de aprobación (o reprobación) y el avance en créditos, en la FI también se han impulsado programas de apoyo a la titulación, dirigidos a alcanzar una mayor tasa de graduados con título profesional.

Entre las acciones de apoyo a la titulación se encuentran la ampliación de las opciones de titulación (a 10), invitaciones a egresados con el 100% de créditos cubiertos a conocer la gama de opciones de titulación, fortalecimiento de la oferta de diplomados, aplicación de exámenes grupales de comprensión de lectura de inglés, entre otras.

Los resultados obtenidos con estas medidas han sido sobresalientes, sobre todo en el año 2016, alcanzando la mayor cantidad de titulados en más de 20 años.

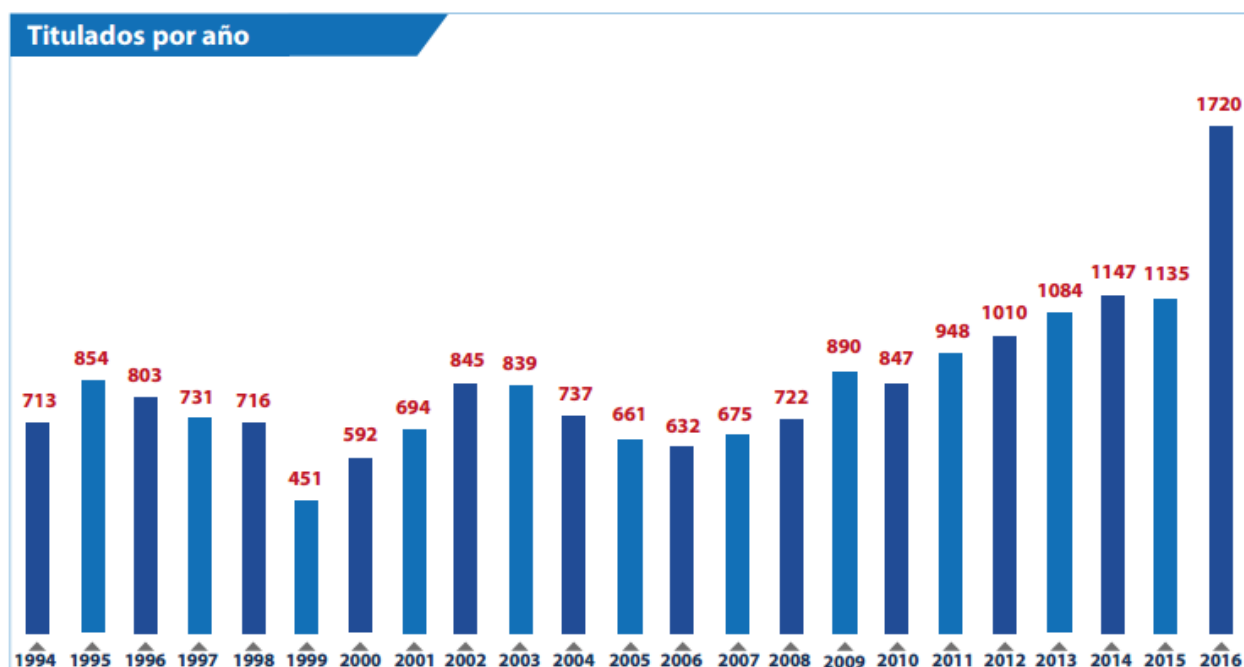


Figura 1.8 Titulados por año. Tomado del Segundo Informe de Actividades 2016, Facultad de Ingeniería.

Si bien es importante que el conjunto de las acciones impulsadas en la FI cumpla con su objetivo (incrementar la tasa de titulación, por ejemplo) también es importante determinar en qué medida cada una de las acciones han contribuido a dicho cumplimiento. De esta manera se podrán establecer metas para cada una de las acciones, con base en su contribución al cumplimiento del objetivo final, y mantener una mejora constante.

Los programas de apoyo de la FI también incluyen becas para estudiantes con escasos recursos, actividades culturales en sus instalaciones tales como exposiciones, conciertos, visitas guiadas, así como la famosa Feria Internacional del Libro en el Palacio de Minería; actividades deportivas tales como la posibilidad de participar en equipos representativos o en la SEFI Olimpiada, actividades de vinculación tales como la Semana SEFI y las ferias de empleo, etc.

1.3 Resumen del capítulo

El estudio del *rendimiento académico* ha ocupado a diversos autores a lo largo de la última mitad del siglo XX y principios del siglo XXI. La variedad de estudios ha desembocado en diversas concepciones del término. En este capítulo se analizaron las concepciones más reconocidas y aceptadas en función de encontrar una acorde a los objetivos de esta tesis.

En términos generales, el concepto se ha enfocado en los resultados obtenidos por el estudiante o por la institución, sin demeritar la influencia recíproca entre uno y otro. Así mismo, el estudio del rendimiento académico ha derivado en dos enfoques principales: el primero, denominado “Productividad Escolar”, refiere a todos aquellos estudios que buscan maximizar el producto (logro académico) manipulando variables de entrada (recursos financieros, material didáctico, etc.)

El segundo enfoque, denominado “Eficacia Escolar”, busca conocer la capacidad que tienen las escuelas para influir en el desarrollo del estudiante, además de conocer los procesos que hacen que se consigan mejor los objetivos. Ambos enfoques sirvieron para definir el término *rendimiento académico* que para propósitos de esta tesis será entendido como el producto a evaluar, obtenido del sistema Contexto – Institución – Estudiante.

También se analizaron las variables utilizadas para explicar el rendimiento académico en diversos estudios previos. Con esta información, se elaboró un mapa general que incluyó todas las variables encontradas en la literatura, clasificadas con base a su pertinencia en el elemento contextual, institucional o personal del estudiante. A su vez, cada elemento se integró por factores generales; de esta manera, el elemento *Estudiante* quedó integrado por los factores Cognitivos, Motivacionales y Psicosociales; el elemento *Institución* quedó integrado por la Docencia, la Infraestructura y los Programas de Apoyo, mientras que el elemento de *Contexto* quedó integrado por el factor Sociodemográfico y por el Socioeconómico.

En el estado del arte se analizaron también estudios previos realizados en la UNAM, específicamente en las Facultades de Estudios Superiores Iztacala y Aragón, debido a la similitud de las metodologías empleadas con la que se pretende trabajar en esta tesis. Los resultados obtenidos indican que las estrategias de estudio,

la capacidad de autorregulación del estudiante, así como el tiempo de traslado y los ingresos económicos son las variables que más repercuten en el rendimiento académico. Así mismo, hablando específicamente de estudiantes de ingeniería, se encontró que el ambiente familiar y social son los factores que más influencia tienen en el rendimiento de los estudiantes. Es de notarse además que el ambiente universitario no haya sido encontrado como un factor influyente en las actitudes de los estudiantes.

Finalmente, se analizaron las estrategias y programas de apoyo conducidos por la Facultad de Ingeniería que buscan ayudar a cumplir las metas académicas de los estudiantes, en los tiempos establecidos en los programas de estudios. Los talleres de ejercicios, cursos extraordinarios, ayuda en la preparación de exámenes, asesorías y tutorías figuraron como las acciones más comunes para impulsar el avance en tiempo y forma de los alumnos. En este apartado también se analizaron los instrumentos utilizados para la recopilación de información que han servido para la creación de perfiles de estudiantes por cohorte generacional, principalmente en alumnos de nuevo ingreso. Los principales a mencionar: el Cuestionario Sociodemográfico y de Antecedentes Escolares, el SIVACORE y el Diagnóstico de Conocimientos Previos o de Antecedentes Académicos.



Las Islas, Ciudad Universitaria
Foto: Gabriel Gaspar Galindo

2.1 Econometría

Hoy en día la cantidad de datos que se generan alrededor de casi cualquier interacción humana es inimaginable (pagos, transacciones, registros, llamadas, mensajes, ventas, likes, clicks, etc.) Así mismo, la creciente capacidad para el almacenamiento de datos, así como la considerable reducción de los costos asociados al mismo, están impulsando entre diversos sectores industriales y de servicios el uso de técnicas estadísticas avanzadas que permitan convertir dichos datos en *información* y ésta a su vez en *conocimiento* que oriente la toma de decisiones.

La teoría económica ha poseído una de las herramientas más importantes para el análisis de datos desde hace ya varias décadas; sin embargo, su aplicación se había enfocado casi exclusivamente a la comprobación (o refutación) de dichas teorías.

En la actualidad, la capacidad de esta herramienta para estimar relaciones causales ha atraído la atención de las empresas interesadas en conocer más a detalle el *¿por qué?* de la variación en su demanda. Este conocimiento les ha permitido

además descubrir el *¿cómo?* realizar previsiones más precisas e inclusive *influir* en la variación de su demanda de manera provechosa (Chase Jr., 2013).

De entre las diversas definiciones que existen de **econometría**, la de Wooldridge (2015) incluye este nuevo enfoque derivado del análisis empresarial:

La econometría se basa en el desarrollo de métodos estadísticos que se utilizan para estimar relaciones económicas, probar teorías económicas y evaluar e implementar políticas públicas y de negocios (p. 1).

En la actualidad, la utilidad del análisis econométrico estriba en la capacidad de comprobar relaciones causales establecidas por la mera intuición, sin la necesidad de respaldarse en modelos económicos formales. Esta capacidad aumenta con la disponibilidad de bastas bases de datos accesibles para cualquier analista.

Existen dos objetivos fundamentales que guían cualquier análisis econométrico: el primero de ellos es encontrar (en caso de existir) la magnitud de la relación entre una variable dependiente y una variable independiente que se crea es *explicativa* de la primera.

Este análisis se realiza siempre bajo el supuesto *ceteris paribus*, el cual se interpreta como: “si todos los demás factores considerados en el estudio permanecen constantes” (Wooldridge, 2015). Esto quiere decir que, al realizar un análisis econométrico se supone que, a excepción de la variable independiente bajo estudio, todas las demás variables independientes consideradas no sufren cambio alguno.

Este supuesto es esencial para inferir relaciones causales, ya que sin él no sería posible establecer el efecto de la variable independiente sobre la variable dependiente.

El segundo objetivo fundamental del análisis econométrico es obtener pronósticos. La ventaja de utilizar técnicas econométricas sobre otras técnicas de pronósticos es que con las primeras es posible cuantificar el impacto de cambios en precio, publicidad, promociones, eventos de marketing, etc. y con ello moldear la

demanda futura (Chase Jr., 2013). Esta aproximación a los pronósticos es entendida como *proactiva*, mientras que el uso de técnicas más convencionales (suavizado exponencial, método Holt de dos parámetros, método Holt – Winters, etc.) es entendida como *reactiva* al considerar un número limitado de factores (nivel, tendencia, estacionalidad y ciclicidad) que no permiten cuantificar la variación en la demanda (y, por ende, moldearla) debido a otros factores causales tales como las promociones o la publicidad.

Makridakis, Wheelwright y Hyndman (1983) presentan en su clásico *Forecasting: Methods and applications* una serie de pasos necesarios para establecer un modelo econométrico:

1. Determinar las variables a incluir en la ecuación.
2. Determinar la forma funcional (p.e. lineal, exponencial, logarítmica, etc.) de la ecuación.
3. Estimar los parámetros de la ecuación.
4. Probar la significancia estadística de los resultados.
5. Probar la validez de las suposiciones involucradas

En este capítulo se presentarán los criterios de selección de la forma funcional del modelo (punto 2), la principal técnica de estimación de parámetros (punto 3), así como los estadísticos necesarios para los pasos 4 y 5. La determinación de las variables (paso 1) se presenta en el siguiente capítulo.

Finalmente, una cuestión importante a considerar en la realización de un análisis econométrico, es conocer el *tipo de datos* con los que se trabajará. De esto dependerán las técnicas o métodos de estimación empleados.

Existen cuatro tipos de datos con los cuales se puede realizar un análisis econométrico, el primero de ellos concierne a los **datos de corte transversal**.

Este tipo de datos consiste en “una muestra de individuos, hogares, empresas, ciudades, estados, países u otras unidades, tomada en algún punto dado en el tiempo” (Wooldridge, 2015).

Al hablar de *muestras* se entiende que, para que los resultados obtenidos puedan ser estimaciones confiables del comportamiento poblacional, éstas deben ser obtenidas de manera aleatoria.

La forma más usual de obtener este tipo de datos es a través de *entrevistas*. Las entrevistas se realizan en un solo momento a una cantidad determinada de personas, hogares, empresas, etc. Los datos de corte transversal nos permiten obtener un panorama del estado que guarda el fenómeno social de nuestro interés (nivel educativo de las personas, nivel de cobertura de un seguro de vida, etc.) en un momento dado. Esto se puede entender como la *fotografía* tomada en una fecha establecida, del fenómeno social que nos interesa analizar.

Todos los trabajos y análisis presentados en el estado del arte del primer capítulo utilizaron éste tipo de datos. Por consiguiente, es fácil deducir que, si no todos, la gran mayoría de trabajos relacionados a la evaluación del rendimiento académico han utilizado datos de corte transversal.

Por otra parte, otro tipo de datos son los **datos de series de tiempo**, los cuales consisten de “observaciones de una o varias variables a lo largo del tiempo” (Wooldridge, 2015).

La principal diferencia con los datos de corte transversal, es que con los datos de series de tiempo se cuenta con la capacidad de identificar patrones dependientes del tiempo, tales como la tendencia o la estacionalidad. Bajo la suposición de que los patrones identificados hasta el momento seguirán presentándose en el futuro, es posible hacer deducciones sobre el valor futuro de alguna variable de nuestro interés.

Los datos de series de tiempo se definen por su *periodicidad*; es decir, por la frecuencia con que los datos son recolectados. Esta frecuencia depende en gran parte de la naturaleza de la variable o del costo asociado a la recolección de observaciones. Por ejemplo, es muy común que, en cuestiones de planeación de la demanda, los datos sean recabados de manera mensual. Esto permite a las empresas desagregar los recursos disponibles (financieros, humanos, materiales, etc.) en

semanas o incluso días, sin incurrir en el costo que conllevaría recabar la información de manera semanal o diariamente.

Por otra parte, la periodicidad de la serie de tiempo del Censo Nacional de Población y Vivienda es decenal, debido en gran medida a la incapacidad de períodos inferiores de identificar cambios sustanciales a nivel población, así como al costo en el que se incurre en recabar estas observaciones.

Los datos de series de tiempo pueden entenderse como una *película* de la variable en cuestión, en la cual podemos vislumbrar escenas acontecidas en momentos anteriores y estimar (con cierto margen de error) lo que acontecerá después.

De los datos de corte transversal y de series de tiempo se derivan los siguientes dos tipos. El primero de ellos es la **combinación de cortes transversales** (Wooldridge, 2015).

En este tipo de datos se combinan *dos* muestras de datos de corte transversal tomadas de manera independiente; es decir, que fueron tomadas en momentos distintos a individuos, empresas, hogares, etc., distintos.

Como ejemplo, se puede suponer que en el año 2010 se realizó una entrevista de 5 preguntas a 100 personas en la Delegación Coyoacán, para conocer su opinión sobre una nueva política pública que restringe la conducción de automóviles con mascotas en las piernas del conductor. Para el año 2015, se volvió a entrevistar a 100 personas (diferentes de las primeras 100) preguntándoles las mismas 5 preguntas.

En este ejemplo, la recolección de datos se obtuvo en dos años distintos, entrevistando en cada uno de ellos a cien personas distintas; sin embargo, las preguntas fueron las mismas. Por ende, el tipo de datos obtenido al combinar las dos bases de datos, es considerado como una combinación de cortes transversales.

Tal y como se vislumbra en el ejemplo anterior, éste tipo de datos son muy útiles para evaluar el impacto de políticas públicas. Wooldridge (2015) enfatiza además en

la importancia que tienen al incrementar el tamaño de la muestra, así como en la capacidad de observar cómo ha cambiado con el tiempo una relación clave.

Finalmente, el último tipo de datos existente es el de los **datos de panel o longitudinales**. Estos consisten en una serie de tiempo por *cada* unidad de corte transversal (Wooldridge, 2015).

Continuando con el ejemplo anterior, si en lugar de tomar las muestras en dos momentos distintos (2010 y 2015) se tomaran en cada uno de los años de ese intervalo (2010, 2011, 2012, 2013, 2014 y 2015) y además las personas entrevistadas fueran las *mismas* en cada año, entonces estaríamos hablando de datos tipo panel o longitudinales.

La característica fundamental de los datos de panel, que los distingue de las combinaciones de cortes transversales, es que durante un intervalo de tiempo se vigilan las mismas unidades (personas, empresas o condados, etc.) de un corte transversal (Wooldridge, 2015).

Es importante evaluar cada situación en particular para determinar qué tipo de datos es el idóneo. La decisión de elegir qué tipo de datos utilizar estará siempre en función de la naturaleza de la variable de interés, así como de la disponibilidad de la información.

2.2 Regresión lineal

La regresión lineal es la técnica por excelencia para determinar la existencia de alguna relación entre varias variables (independientes) y una variable de interés (dependiente).

El modelo general de la regresión lineal es:

$$Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik} + \mu_i; i = 1, 2, \dots, n \quad (2.1)$$

donde:

β_0 es el intercepto.

β_k es el parámetro asociado a x_{ik} .

x_{ik} es una variable explicativa de Y asociada a la observación i

μ es el término de error o perturbación.

Y es la variable dependiente o explicada por todas las x_{ik} .

Las constantes $\beta_1, \beta_2, \dots, \beta_k$ describen la **dirección** y **fuerza** de la relación entre sus x asociadas y la variable dependiente Y .

Así mismo, el parámetro β_j representa el cambio esperado en la variable dependiente por un cambio *unitario* en x_j cuando se mantienen constantes las demás variables explicativas.

Por otra parte, no importando el número de factores o variables explicativas que agreguemos al modelo para explicar la variación de Y , siempre habrá factores que no se podrán incluir (ya sea por desconocimiento de su existencia o por falta de información). El modelo de regresión lineal nos permite incluir de manera implícita todos estos **factores no observables**. Estos se encuentran contenidos en el término de error μ .

También es muy importante aclarar el aspecto de linealidad. El modelo de regresión lineal debe su nombre a que éste es lineal en los **parámetros** β . Por tanto, es posible que un modelo contenga una relación no lineal entre la variable dependiente y las variables explicativas y, aun así, considerarse como modelo de regresión lineal (siempre y cuando la relación de los parámetros β con Y sea lineal).

La función del modelo de regresión lineal es obtener los valores de los parámetros β para conocer la dirección y fuerza de la relación entre las variables explicativas y la variable dependiente. Sin embargo, en primera instancia esto resulta casi imposible. Los parámetros β de la ecuación (2.1) describen el comportamiento **poblacional** entre las variables involucradas.

Como es casi imposible conocer los valores de las β poblacionales, éstas se deben *estimar* a través de **muestras**. Para obtener estimadores confiables de los parámetros poblacionales las muestras deben ser obtenidas de manera *aleatoria*.

Por otra parte, existe una suposición fundamental que permite inferir *causalidad* entre las variables explicativas x y la explicada Y , y ésta es que el término de error μ no está relacionado con los regresores (o variables independientes).

El propósito del modelo de regresión lineal (en este contexto) es *aislar el impacto* de alguna variable explicativa sobre la variable dependiente; por ello, dicha suposición es muy importante, ya que sin ella se deja abierta la posibilidad de encontrar una variación en Y debida no sólo a las x_k sino a una combinación entre éstas y el término de error μ .

Utilizando conceptos de probabilidad, lo anterior se puede expresar matemáticamente como:

$$E(\mu|x_1, x_2, \dots, x_k) = 0 \quad (2.2)$$

La ecuación (2.2) establece que la esperanza de μ dado cualquier valor de x_1, x_2, \dots, x_k es igual a cero. En otras palabras, el valor promedio de μ no depende de las variables explicativas: no se encuentran correlacionados. A la ecuación (2.2) se le conoce como **supuesto de media condicional cero** (Wooldridge, 2015).

El supuesto de media condicional cero permite establecer explícitamente un **efecto causal** entre alguna variable explicativa x_j y la variable explicada Y ; sin embargo, para analizar el efecto causal de *una* sola variable explicativa se debe considerar que no hay cambio en todos los demás factores relevantes.²

2.2.1 Estimación de los parámetros por MCO.

Dadas n observaciones de las variables explicativas ($x_{i1}, x_{i2}, \dots, x_{ik}; i = 1, 2, \dots, n$ y $n > k$) y de la variable dependiente ($Y_i; i = 1, 2, \dots, n$) el modelo de regresión lineal queda expresado como:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \mu_i \\ &= \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \mu_i; i = 1, 2, \dots, n \end{aligned} \quad (2.3)$$

La siguiente ecuación establece de manera explícita la estimación de la variable dependiente para la muestra i :

$$\hat{y}_i = \hat{\beta}_0 + \sum_{i=1}^k \hat{\beta}_i x_{ik} \quad (2.4)$$

donde

² Consultar el Anexo A para mayor referencia sobre los supuestos en los que se basa la construcción y estimación de modelos de regresión lineal.

\hat{y}_i estimación de la variable dependiente Y_i .

$\hat{\beta}_0$ estimación del intercepto β_0 .

$\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ estimación de los parámetros poblacionales $\beta_1, \beta_2, \dots, \beta_k$.

La ecuación (2.4) representa la *recta de regresión* de la muestra i . Por otra parte, \hat{y}_i también es conocido como el **valor ajustado**, dado que éste valor se *ajusta* a su correspondiente recta de regresión lineal.

La diferencia entre el valor real Y_i y su estimado \hat{y}_i se conoce como el **residual** de la observación i :

$$\hat{\mu}_i = Y_i - \hat{y}_i = Y_i - \hat{\beta}_0 - \sum_{j=1}^n \hat{\beta}_j x_{ij} \quad (2.5)$$

Cabe destacar que el residual $\hat{\mu}_i$ *no* es lo mismo que el término de error μ_i de la ecuación (2.3). La diferencia radica en que el primero se puede calcular a partir de los datos, mientras que el segundo jamás será observable (Wooldridge, 2015).

El método de Mínimos Cuadrados Ordinarios (MCO) minimiza la suma de los residuales al cuadrado³, por lo que para las n observaciones se tiene:

$$\sum_{i=1}^n \hat{\mu}_i^2 = \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right)^2 \quad (2.6)$$

³ La elevación al cuadrado se debe a las propiedades básicas de la sumatoria. Para mayor referencia consultar *Introducción a la Econometría*, Wooldridge, 2015, pp. 27-28.

La ecuación (2.6) también es conocida como la **función de mínimos cuadrados** (Montgomery y Runger, 2013).

En cuanto a los valores de las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, éstos se obtienen resolviendo la derivada parcial para cada parámetro en la ecuación (2.6). El sistema de ecuaciones para resolver una regresión lineal con k variables explicativas queda como el siguiente (tomado de Montgomery y Runger, 2013):

$$\begin{aligned}
 n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{ik} &= \sum_{i=1}^n Y_i \\
 \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{i1} x_{ik} &= \sum_{i=1}^n x_{i1} Y_i \\
 \vdots & \\
 \hat{\beta}_0 \sum_{i=1}^n x_{ik} + \hat{\beta}_1 \sum_{i=1}^n x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{ik} x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{ik}^2 &= \sum_{i=1}^n x_{ik} Y_i
 \end{aligned} \tag{2.7}$$

Hoy en día se puede utilizar cualquier paquete estadístico o econométrico para encontrar la solución a este sistema de ecuaciones lineales.

La interpretación que se le da al valor de cada estimación $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ es que cada uno de ellos mide el cambio en \hat{y}_i por un aumento unitario en su respectiva x , manteniendo todas las demás variables independientes constantes (Wooldridge, 2015).

Suponiendo que se desea analizar el impacto de x_1 sobre \hat{y}_i , lo anterior queda expresado como:

$$\Delta \hat{y}_i = \hat{\beta}_1 \Delta x_1, \quad (2.8)$$

manteniendo constantes x_2, x_3, \dots, x_k . Por tanto, las variables x_2, x_3, \dots, x_k han sido *controladas* al estimar el efecto de x_1 sobre \hat{y}_i (Wooldridge, 2015).

En cuanto al intercepto $\hat{\beta}_0$, su valor no provee información adicional, a menos que dada la naturaleza del análisis, se encuentre interesante saber el valor de \hat{y}_i cuando el valor de todas las variables explicativas es igual a cero.

Más que nada, la importancia de la estimación del intercepto radica en poder pronosticar valores para Y_i .

2.2.2 Coeficiente de determinación R^2 .

Para cualquier análisis de regresión lineal es muy importante conocer qué tan adecuadas son las variables independientes seleccionadas. Es decir, qué tan bien *explican* a la variable dependiente. Para ello, es necesario contar con un indicador que nos permita conocer la pertinencia de las variables independientes involucradas.

Partiendo del sentido común, es fácil deducir que un indicador de éste tipo debe tomar en cuenta la *variación* existente entre la estimación obtenida \hat{y}_i , y el valor real Y_i . Conocer esta diferencia nos permitirá evaluar la pertinencia de las variables explicativas seleccionadas.

El *coeficiente de determinación* R^2 nos indica la *proporción* de la variación muestral en Y_i que es explicada por la recta de regresión de MCO (Wooldridge, 2015). Su valor está dado por:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (2.9)$$

En donde el numerador, también conocido como la *suma explicada de cuadrados* indica la variación existente entre la estimación \hat{y}_i de la variable dependiente y la media de los valores reales u observaciones \bar{Y} . Así mismo, el denominador, también conocido como la *suma total de cuadrados*, indica la variación existente entre las observaciones puntuales Y_i y la media de las observaciones \bar{Y} .

En otras palabras, R^2 es la razón entre la variación explicada por la recta de regresión de MCO y la variación real o total entre las n observaciones.

Lo anterior se puede observar en la siguiente figura:

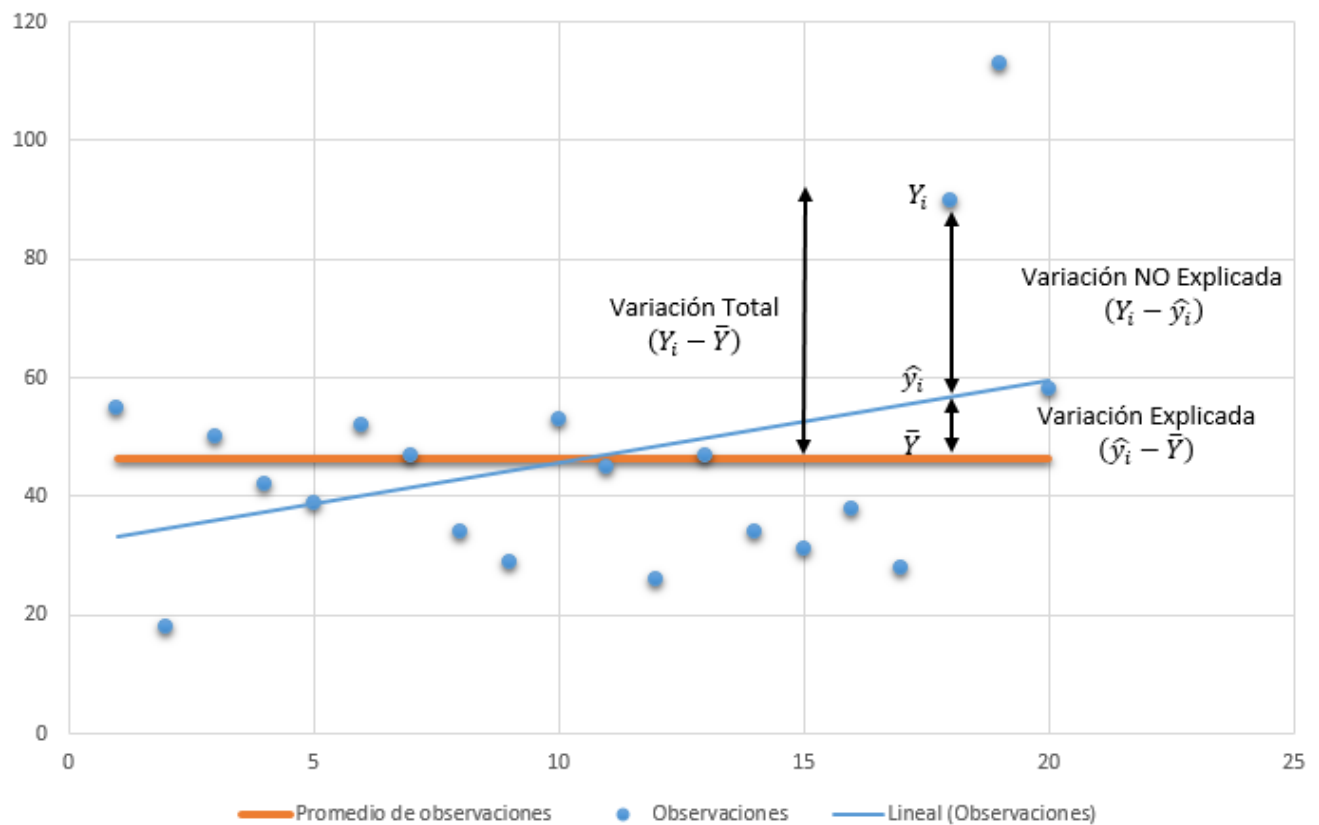


Figura 2.1 Representación gráfica de la variación explicada y no explicada por la recta de regresión de MCO en el caso especial de la regresión lineal simple. Elaboración propia con base en Makridakis *et al.* (1983).

La distancia existente entre el *punto de estimación* \hat{y}_i y la observación Y_i (como se vio anteriormente, también conocido como el *residual de la observación i*) es la variación *no captada* o *no explicada* por la recta de MCO. Mientras que la distancia existente entre el *promedio de observaciones* \bar{Y} y el *punto de estimación* \hat{y}_i es la variación *captada* o *explicada* por el método de MCO. La recta que une todos los puntos de estimación \hat{y}_i es conocida como la *recta de regresión lineal*.

En la figura 2.1 se aprecia una recta de regresión lineal para el caso especial de una sola variable explicativa, denominada comúnmente como *regresión lineal simple*. El análisis del coeficiente de determinación R^2 se realiza sobre una recta de regresión lineal simple sencillamente porque es más fácil entenderlo e interpretarlo. Las conclusiones que se derivan de su análisis aplican también para hiperplanos de regresiones lineales múltiples (con más de una variable explicativa).

Si bien el coeficiente de determinación nos permite conocer la proporción de la variación explicada respecto a la variación total o, en otras palabras, la proporción de la variación total en Y_i que puede ser explicada por las variables explicativas seleccionadas; confiar ciegamente en el valor de R^2 puede hacernos caer en una trampa.

La razón por la que hay que tener cuidado con el uso de R^2 es que éste coeficiente nunca disminuye y, en general, aumenta cuando se agrega otra variable explicativa a la regresión, independientemente de si se trata de una variable relevante o no. En palabras de Wooldridge (2015): “El hecho de que la R^2 nunca disminuya cuando se agrega cualquier variable a la regresión hace de R^2 un instrumento poco confiable para decidir si agregar una o varias variables al modelo”.

2.2.3 R^2 ajustada.

Al inicio del apartado anterior comentaba sobre la importancia de contar con un indicador que permitiera conocer la pertinencia de las variables independientes o explicativas involucradas. En términos estadísticos, este indicador debe ser capaz de

medir la *varianza* de la variable dependiente en función de las variables explicativas seleccionadas; en otras palabras, la *varianza explicada por el modelo de regresión*.

En este contexto cabe hacer énfasis que la *variación* y la *varianza* **no** son lo mismo. Anteriormente se ha establecido a R^2 como una medida de la proporción de la **variación** muestral en Y_i explicada por la recta de regresión; sin embargo, R^2 no puede ser considerada una medida de la proporción de la **varianza**, ya que la obtención de R^2 no involucra *grados de libertad*.

El término grados de libertad en este contexto, se refiere a un *ajuste* aplicado al valor original de R^2 que permite “sancionar” la adición de más variables independientes a un modelo de regresión.

El valor de R^2 ajustada está dado por:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{(n - 1)}{(n - k - 1)} \quad (2.10)$$

En donde n son las observaciones y k el número de variables explicativas involucradas en el modelo.

Como este valor sí toma en cuenta grados de libertad (al igual que la *varianza*)⁴ la R^2 ajustada puede ser interpretada como la proporción de la **varianza** en la variable dependiente, explicada por las variables independientes (Makridakis *et al.*, 1983) convirtiéndola en un mejor indicador de ajuste.

⁴ La definición matemática de la *varianza* para datos no agrupados está dada por:

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (\bar{x} - x_i)^2$$

en donde el factor $1/(n-1)$ representa un ajuste a los grados de libertad para volver a la *varianza* muestral una estimación insesgada de la *varianza* poblacional.

2.2.4 Multicolinealidad.

Existe una cuestión inherente en la elaboración de un modelo de regresión, y ésta es que al seleccionar las variables explicativas o independientes casi podemos asegurar que éstas se encuentran (en cierto grado) correlacionadas entre ellas mismas.

Esto tiene sentido ya que al pensar en fenómenos naturales o sociales éstos deben ser analizados desde una perspectiva holística. Ninguna variable se encuentra totalmente aislada del resto del sistema. A la correlación existente entre dos o más variables independientes se le denomina *multicolinealidad*.

Si bien la existencia de multicolinealidad es un evento común para la mayoría de los análisis de regresión lineal, su presencia en gran medida (es decir, una correlación alta entre dos o más variables independientes) se interpreta como síntoma de algún problema en el modelo de regresión.

Para Charles W. Chase Jr. (2013) la existencia de multicolinealidad en el modelo de regresión es un indicador importante para desestimar los valores de los coeficientes $\hat{\beta}_j$ calculados en la muestra, debido a su “inestabilidad” y a la posible falta de significancia estadística. Así mismo, el autor indica que la presencia de correlación alta entre dos o más variables vuelve poco confiable al coeficiente de determinación R^2 .

Para Makridakis *et al.* (1983) los resultados obtenidos en softwares estadísticos no pueden ser confiables si existe multicolinealidad, argumentando de igual manera la “inestabilidad” existente en los coeficientes de regresión obtenidos.

Sobre este tema cabe recordar que en los análisis econométricos uno de los principales objetivos es encontrar el efecto parcial de *cierta* variable independiente sobre la variable dependiente. La inclusión de *otras* variables independientes al modelo permite reducir el efecto del factor inobservable (el error μ).

Al respecto, Wooldridge (2015) hace hincapié en el punto anterior al argumentar que la multicolinealidad *per se* no es un problema. Más bien se vuelve un problema cuando la variable independiente asociada al parámetro de interés (supongamos β_1 de la ecuación 2.3) se encuentra altamente correlacionada con alguna otra variable

independiente. Esto es un problema ya que no es posible determinar de manera *puntual* el efecto parcial de la variable de interés sobre la variable dependiente.

Es importante mencionar que este autor no demerita el problema de la multicolinealidad, pero aclara que en el contexto de un análisis econométrico, si se encuentra una alta correlación entre dos o más variables independientes que *no* incluyen a la variable explicativa de interés, entonces no hay de qué preocuparse.

Por otra parte, cabe destacar el término *inestabilidad* utilizado en los argumentos de los dos primeros autores referidos en este apartado. Este término hace referencia a que una correlación alta entre dos variables independientes (por ejemplo x_1 y x_2) impactará a la estimación del parámetro de interés ($\hat{\beta}_1$) de tal manera que su varianza se incrementará. Mientras más fuerte sea la correlación existente entre ambas variables, más grande será la varianza de la estimación del parámetro de interés: $Var(\hat{\beta}_j) \rightarrow \infty$ a medida que $R_j^2 \rightarrow 1$.⁵

Lo anterior resulta ser un gran problema ya que es muy probable que, al utilizar una muestra distinta para la estimación del modelo, se obtengan estimaciones de los coeficientes β completamente diferentes, derivando en conclusiones contradictorias. Además, con varianzas que tienden a infinito, análisis de inferencia estadística tienen nula confiabilidad.

Montgomery, Peck y Vining (2012) establecen cuatro principales causas de multicolinealidad:

- 1.- El método empleado para recolectar los datos.
- 2.- Restricciones inherentes al modelo o a la población.
- 3.- Especificación del modelo (o forma funcional).
- 4.- Exceso de regresores (variables explicativas).

⁵ Consultar Anexo A.

La primera causa hace referencia a una técnica de muestreo *sesgada*. Es decir, se recaba información **solo** de un extracto (región o subespacio) de la población. Esto puede originar multicolinealidad entre dos o más variables de interés involucradas.

La segunda causa hace referencia a restricciones *físicas* presentes en el modelo o en la población. Para ejemplificar esta cuestión, Montgomery *et al.* (2012) hacen mención de la relación existente entre los ingresos de una familia y el tamaño de su casa. La restricción *física* subyace en que las familias con ingresos económicos altos no tendrán casas pequeñas y viceversa. Por tanto, independientemente de la técnica de muestreo empleada, siempre existirá una correlación (ya sea positiva o negativa) en el ingreso económico de las familias y el tamaño de su hogar.

Otra causa de multicolinealidad se deriva de la *forma funcional del modelo*. Este tema se trata a detalle en la siguiente sección; sin embargo, se puede adelantar que la forma (o especificación) del modelo de regresión puede incluir términos polinomiales y/o logarítmicos⁶. Al utilizar, por ejemplo, el término lineal y cuadrático de una *misma* variable en el modelo de regresión, si el rango de la variable es pequeño, es muy probable que exista multicolinealidad.

La última causa de multicolinealidad se debe al exceso de regresores en el modelo. Esto sucede cuando se incluyen más variables explicativas que observaciones existentes; es decir $j > i$.

Si bien la presencia de *alta* multicolinealidad entre variables de interés produce estimaciones poco confiables, los modelos con altas correlaciones entre regresores **no necesariamente son malos para realizar predicciones** de nuevas observaciones. En palabras de Montgomery *et al.* (2012): “Si bien el método de mínimos cuadrados generalmente producirá estimadores pobres de los parámetros individuales del modelo en presencia de multicolinealidad, esto no necesariamente implica que el modelo ajustado [la ecuación de regresión] sea un predictor pobre. Si las predicciones se encuentran confinadas a regiones en el espacio en x donde la

⁶ Ver apartado *Forma funcional* del presente capítulo.

multicolinealidad se mantiene, el modelo ajustado [la ecuación de regresión] producirá predicciones satisfactorias”.

La *matriz de correlación* es una herramienta que ha sido muy popular para identificar variables correlacionadas. Ésta, hace uso de la magnitud de los coeficientes de correlación⁷ de las variables explicativas involucradas en el modelo para evaluar si la multicolinealidad puede ser o no un problema.

La mayor desventaja de utilizar la matriz de correlación en el contexto de la regresión lineal, es que sólo puede identificar relaciones lineales (variables correlacionadas) entre **pares de regresores**. Es decir, la matriz de correlación es incapaz de identificar relaciones lineales que dependan de tres o más variables. En consecuencia, en aplicaciones de estimación de modelos de regresión lineal esta desventaja es contundente en la decisión de utilizar otras técnicas de detección de multicolinealidad. En el Capítulo III se presenta una técnica más eficiente que permite identificar relaciones lineales dependientes entre más de dos variables.

Finalmente, ¿qué hacer en caso de encontrarse con un nivel de multicolinealidad que no permita medir el efecto parcial de una o varias variables de interés? Todos los autores mencionados en esta sección concuerdan en incrementar el tamaño de la muestra (es decir, recolectar más datos) o combinar las variables altamente correlacionadas en una sola (considerando el cambio de unidades en caso de ser necesario). Existen, además, otras técnicas para lidiar con el problema de multicolinealidad que involucran métodos de estimación diferentes al MCO. Sin embargo, dichas técnicas quedan fuera del alcance de esta tesis.⁸

⁷ El coeficiente de correlación (denominado comúnmente como R) mide el nivel de asociación lineal entre dos variables. Su rango de valores va desde -1 (correlación negativa perfecta) pasando por 0 (inexistencia de correlación) hasta +1 (correlación positiva perfecta). Al elevar R al cuadrado obtenemos el coeficiente de determinación R^2 .

⁸ Consultar *Introduction to Linear Regression Analysis*, 5a Edición, de Douglas C. Montgomery, Elizabeth A. Peck y G. Geoffrey Vining, pp. 304-321, para mayor referencia en cuanto a métodos de estimación diferentes del MCO que proporcionan mejores resultados en presencia de multicolinealidad.

2.2.5 Forma funcional.

Si bien el modelo de regresión lineal requiere que los *parámetros* sean lineales, esto no limita de manera alguna la relación entre las variables explicativas y la variable explicada.

Para una gran cantidad de trabajos econométricos, más que conocer el impacto por el cambio en una *unidad*, resulta mucho más útil conocer el impacto dado un cambio *porcentual*.

Para poder evaluar el impacto de una variable dado un cambio porcentual, se requiere hacer uso del concepto *elasticidad* perteneciente al campo de la economía. En este contexto, la elasticidad mide la variación porcentual que experimenta una variable dependiente dada la variación de un 1% en una variable independiente.

La *forma funcional* del modelo de regresión lineal que permite calcular la elasticidad de dos variables, es:

$$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + \mu \quad (2.11)$$

La ecuación (2.11) representa a un **modelo de elasticidad constante**, en donde β_1 es la elasticidad entre y y x_1 . En otras palabras, el coeficiente β_1 indica la *variación porcentual* experimentada por y dada una variación de 1% en el valor de x_1 . El término $\ln()$ indica el logaritmo natural de la variable entre paréntesis.

Considerando la variación del error igual a cero, es decir $\Delta\mu = 0$, entonces:

$$\% \Delta y \approx \beta_1 \% \Delta x_1 \quad (2.12)$$

Cabe destacar que el cambio porcentual en y encontrado mediante la forma de la ecuación (2.12) es sólo una aproximación, y ésta sólo es válida para incrementos en y pequeños.

La forma del modelo (2.11) es sólo una de cuatro posibles formas funcionales en las que se emplean logaritmos; dependiendo del tipo de análisis que se esté realizando, convendrá utilizar una u otra forma.

En la siguiente figura se presenta un resumen de las formas funcionales logarítmicas:

Modelo	Variable dependiente	Variable independiente	Forma Funcional	Interpretación de β_1
Nivel - Nivel	y	x	$y = \beta_0 + \beta_1 x_1 + \mu$	$\Delta y = \beta_1 \Delta x$
Nivel - ln	y	$\ln(x)$	$y = \beta_0 + \beta_1 \ln(x_1) + \mu$	$\Delta y = (\beta_1/100)\% \Delta x$
ln - Nivel	$\ln(y)$	x	$\ln(y) = \beta_0 + \beta_1 x_1 + \mu$	$\% \Delta y = (100\beta_1) \Delta x$
ln - ln	$\ln(y)$	$\ln(x)$	$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + \mu$	$\% \Delta y = \beta_1 \% \Delta x$

Tabla 2.1 Resumen de las formas funcionales logarítmicas. Elaboración propia con base en Wooldridge (2015).

La primera forma funcional, referente al modelo Nivel – Nivel, es la forma más común encontrada en modelos de regresión lineal y su interpretación es la que se ha discutido a lo largo de este capítulo (véanse ecuaciones (2.3) y (2.8)).

La segunda forma funcional, referente al modelo Nivel – ln, es utilizada cuando se desea conocer la variación de y en sus unidades de medición original dado un incremento porcentual de 1% en x .

Por su parte, la tercer forma funcional ln – Nivel, es utilizada cuando se desea conocer la variación porcentual en y dada una variación de x en sus unidades de medición original.

El modelo ln – ln ha sido explicado de manera detallada en párrafos anteriores (véanse ecuaciones (2.11) y (2.12)).

Anteriormente se hizo mención de que la expresión (2.12) sólo es válida para incrementos pequeños en y . Esto se debe a que el coeficiente β_1 está relacionado con $\ln(y)$ y no con y (la variable de interés).

Para poder conocer la *variación exacta* de Δy (independientemente del tamaño de su incremento) basta con despejar el logaritmo natural de éste término. Para la ecuación (2.11) la estimación exacta del cambio porcentual en y está dada por:

$$\% \Delta y = 100 * [e^{\beta_1} \Delta x] \quad (2.13)$$

donde la multiplicación por 100 convierte la *variación proporcional* de Δy en una *variación porcentual*. Cuando $\Delta x = 1$:

$$\% \Delta y = 100 * [e^{\beta_1}] \quad (2.14)$$

Por otra parte, cuando $\Delta x = -1$ se tiene:

$$\% \Delta y = 100 * [(e^{\beta_1})(-1)] \quad (2.15)$$

Las ecuaciones (2.14) y (2.15) calculan la *variación exacta* en Δy , dado un incremento o una disminución respectivamente, de la variable independiente. Esto resulta de vital importancia cuando el interés del análisis radica en conocer el comportamiento de la variable dependiente dado un incremento o una reducción específica, de la variable independiente.

Cabe destacar que, una vez conocido el cálculo exacto de la *variación* en y para la forma funcional de la ecuación (2.11), la expresión (2.12) no deja de tener relevancia. Mientras que para Δy “pequeñas” el cálculo de la *variación porcentual* con esta expresión resulta sumamente exacta, para Δy “grandes” el uso simple del coeficiente β_1 da una estimación situada siempre entre los valores absolutos de las

estimaciones correspondientes a un aumento y a una disminución (Wooldridge, 2015).

En la siguiente figura se presenta un resumen de los cálculos que permiten estimar la variación porcentual exacta en Δy , tanto para el modelo *ln – Nivel* como para el modelo *ln – ln* (el otro par de modelos no necesita ningún despeje ya que Δy se encuentra libre):

Modelo	Cálculo para Δy		
	exacto	Con $\Delta x = 1$	Con $\Delta x = -1$
ln - Nivel	$\% \Delta y = 100 * [e^{\beta_1 \Delta x} - 1]$	$\% \Delta y = 100 * [e^{\beta_1} - 1]$	$\% \Delta y = 100 * [e^{-\beta_1} - 1]$
ln - ln	$\% \Delta y = 100 * [e^{\beta_1 \Delta x}]$	$\% \Delta y = 100 * [e^{\beta_1}]$	$\% \Delta y = 100 * [e^{\beta_1}(-1)]$

Tabla 2.2 Resumen de los cálculos necesarios para determinar la variación exacta en Δy . Elaboración propia

Existen otro tipo de modelos cuya forma funcional se encuentra en términos de **funciones cuadráticas**. Es decir:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \mu \quad (2.16)$$

En este caso, la variable dependiente y depende de sólo una variable explicativa x , pero lo hace de forma cuadrática.

Es común el uso de modelos con funciones cuadráticas en aplicaciones donde se sospecha una relación *no lineal* entre la variable explicativa y la explicada. Un ejemplo de esto es el impacto de los años de educación en el salario percibido. La interpretación del coeficiente beta en un modelo *lineal en las variables* (véanse ecuaciones (2.3) y (2.8)) indicaría un *rendimiento constante* de la educación. Es decir, se le daría el mismo valor a un año de estudios de bachillerato que a un año de estudios universitarios. En la práctica, esto parece poco realista. Tiene mayor sentido pensar que un año de estudios universitarios es más valioso que un año de estudios de bachillerato.

Las funciones cuadráticas en los modelos de regresión permiten captar *rendimientos no constantes* entre dos variables (Wooldridge, 2015).

Al hacer uso de funciones cuadráticas es muy importante saber interpretar el efecto de x sobre y con base en los coeficientes beta. La interpretación de β_1 **no** es la misma para un modelo de la forma (2.16) que para uno de la forma (2.3).

Considerando la estimación de los parámetros poblaciones mediante un modelo de regresión con función cuadrática:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \quad (2.17)$$

El efecto aproximado de x sobre y está dado por:

$$\Delta \hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x) \Delta x \quad (2.18)$$

$$\frac{\Delta \hat{y}}{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$$

En palabras de Wooldridge (2015): “Esto indica que la pendiente de la relación entre x y y dependen del valor de x ”. Cuando $x = 0$ el parámetro $\hat{\beta}_1$ se interpreta como la variación aproximada de y al pasar de $x = 0$ a $x = 1$.

Las gráficas de los modelos con función cuadrática suelen tener formas parabólicas o en U; por consiguiente, para un rango de valores de x la relación con y será de *rendimiento creciente no constante*, mientras que para otro rango la relación será de *rendimiento decreciente no constante*. La división entre los dos rangos estará

marcada por el *punto de inflexión*, el cual indicará el *máximo* o el *mínimo* de la función.

Debido a lo explicado en el párrafo anterior, al interpretar la gráfica del efecto de x sobre y es muy importante identificar el rango de valores para los cuales la relación entre x y y **no tiene sentido alguno** (p.e. que a una mayor emisión de gases de efecto invernadero el impacto negativo sobre el planeta disminuya).

Si la forma funcional utilizada es correcta y se han controlado factores o variables independientes suficientes para explicar una gran parte de la variación en y , entonces éste rango de valores que explica una relación sin sentido pertenece a valores de x aislados, obtenidos en la muestra; es decir, valores poco usuales. Por consiguiente, la parte de la curva que pertenece a este rango de valores puede ignorarse.

Por otro lado, si los valores que pertenecen al rango de la curva sin sentido son frecuentes (digamos, más del 20% del total de los datos obtenidos de la muestra) entonces es posible que la forma funcional utilizada no sea la más adecuada o que el número de variables independientes no sea suficiente para explicar a y (Wooldridge, 2015).

La naturaleza de la curva en formas funcionales cuadráticas está determinada por los signos de los coeficientes $\hat{\beta}_1$ y $\hat{\beta}_2$. En la siguiente figura se presenta una tabla resumen de las posibles combinaciones de signos entre estos dos coeficientes y el tipo de gráfica obtenido:





Coeficiente β_1	Coeficiente β_2	Forma de la gráfica	Punto de Inflexión PI	Efecto de x sobre y	
				$0 < x < PI$	$PI < x < \infty$
Positivo	Negativo		Es un Máximo	Creciente	Decreciente
Negativo	Positivo		Es un Mínimo	Decreciente	Creciente
				$x > 0$	
Positivo	Positivo		No hay	Creciente	
Negativo	Negativo		No hay	Decreciente	

Tabla 2.3 Tabla resumen de la naturaleza de la curva debida a la combinación de signos entre β_1 y β_2 .
Elaboración propia.

Para el caso en el que los coeficientes beta tengan el mismo signo, no existirá un punto de inflexión. El *mínimo* valor esperado de la variable dependiente será cuando $x = 0$ para el caso en el que ambos coeficientes sean positivos. Por el contrario, cuando ambos coeficientes sean negativos, el *máximo* valor de y se encontrará cuando $x = 0$.

La expresión (2.19) presenta el cálculo para obtener el punto de inflexión para cualquier forma funcional cuadrática:

$$x^* = \frac{-\widehat{\beta}_1}{2\widehat{\beta}_2} \tag{2.19}$$

Cabe destacar que es posible hacer uso de formas funcionales cuadráticas junto con logaritmos. Para ello, sería necesario combinar los argumentos de los efectos parciales de la forma funcional cuadrática (ver ecuación 2.18) y la forma funcional logarítmica (ver ecuación 2.12). Sin embargo, al hacer esto se dificultaría la interpretación de los coeficientes.

2.2.6 Prueba de hipótesis.

La prueba de hipótesis es una herramienta sumamente útil para inferir aspectos del comportamiento poblacional con base en una muestra representativa.

Esta herramienta, nacida de la estadística aplicada, ha provisto de una solución *ad hoc* a los estudios econométricos, permitiendo evaluar la pertinencia de las variables explicativas incluidas en los modelos de regresión.

Una **hipótesis estadística** es un enunciado acerca de los parámetros de una o más poblaciones (Montgomery y Runger, 2013). En el contexto de un modelo de regresión, una hipótesis estadística es una conjetura sobre el valor de alguno de los parámetros β_j de la población bajo estudio.

En una prueba de hipótesis se busca encontrar “evidencia suficiente” para aceptar o rechazar un argumento que *infiera* algún aspecto relevante sobre la población. En el caso concreto del MRL, uno de los principales argumentos a evaluar es que el valor del parámetro β_j es igual a cero (es decir, que a nivel población la variable explicativa asociada a β_j no tiene ningún efecto sobre la variable dependiente).

Para poder realizar la prueba de alguna hipótesis estadística en un modelo de regresión, se necesita considerar un supuesto adicional a los supuestos de Gauss-Markov (Ver Anexo A). En palabras de Wooldridge (2015), este supuesto adicional, conocido como el **supuesto de normalidad** establece que:

El error poblacional μ es independiente de las variables explicativas x_1, x_2, \dots, x_k y está distribuido normalmente, con media cero y varianza σ^2 : $\mu \sim Normal(0, \sigma^2)$.

Esta serie de supuestos (Gauss-Markov más el de normalidad) son mejor conocidos como los **supuestos del modelo lineal clásico (MLC)** y su importancia radica en que bajo ellos los estimadores de MCO son los **estimadores insesgados de menor varianza** (Wooldridge, 2015).

Retomando la conjetura del principal argumento a evaluar en análisis econométricos, éste se puede expresar formalmente como:

$$\begin{aligned} H_0: \beta_j &= 0 \\ H_1: \beta_j &\neq 0 \end{aligned} \tag{2.20}$$

Al enunciado $H_0: \beta_j = 0$ de la ecuación (2.20) se le llama la **hipótesis nula**, mientras que al enunciado $H_1: \beta_j \neq 0$ se le conoce como la **hipótesis alternativa**.

Montgomery y Runger (2013) establecen un procedimiento para llevar a cabo pruebas de hipótesis. El primero de ellos es *identificar el parámetro de interés*: como se mencionó anteriormente, para análisis econométricos el parámetro de interés son las β_j del MRL.

Como segundo paso, se debe *establecer la hipótesis nula*. Para la mayoría de las aplicaciones econométricas, el enunciado H_0 de la ecuación (2.20) será la hipótesis nula a evaluar. Sin embargo, es posible que en ocasiones se desee evaluar otro valor para el parámetro. Para ilustrar lo anterior, Wooldridge (2015) utiliza como ejemplo la elasticidad entre dos variables (una dependiente y la otra independiente) en la cual resulta interesante conocer si un aumento del 1% en la variable explicativa conduce a un aumento promedio de 1% en la variable explicada; es decir, que la elasticidad entre ambas sea uno. Esta hipótesis nula se establece como: $H_0: \beta_j = 1$.

El tercer paso consiste en *especificar una hipótesis alternativa apropiada*. El enunciado H_1 de la ecuación (2.20) es la hipótesis alternativa apropiada para evaluar la pertinencia de las variables explicativas en el modelo poblacional. Sin embargo, al igual que la hipótesis nula, resulta interesante poder evaluar otro tipo de argumentos (p.e. si el impacto de la variable explicativa de interés es positivo o negativo a nivel poblacional). Para ello, se hace uso de las desigualdades “mayor que” o “menor que”; es decir:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j > 0$$

(2.21)

ó

$$H_0: \beta_j = 0$$

$$H_1: \beta_j < 0$$

(2.22)

A la ecuación (2.20) se le conoce como prueba **de dos colas**, mientras que a las ecuaciones (2.21) y (2.22) se les conoce como pruebas **de una cola**. Esto se debe a la región de interés en la distribución del error poblacional; es decir: $y|x_1, \dots, x_k \sim Normal(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \sigma^2)$. La siguiente figura ejemplifica la región de interés para una prueba de dos colas:

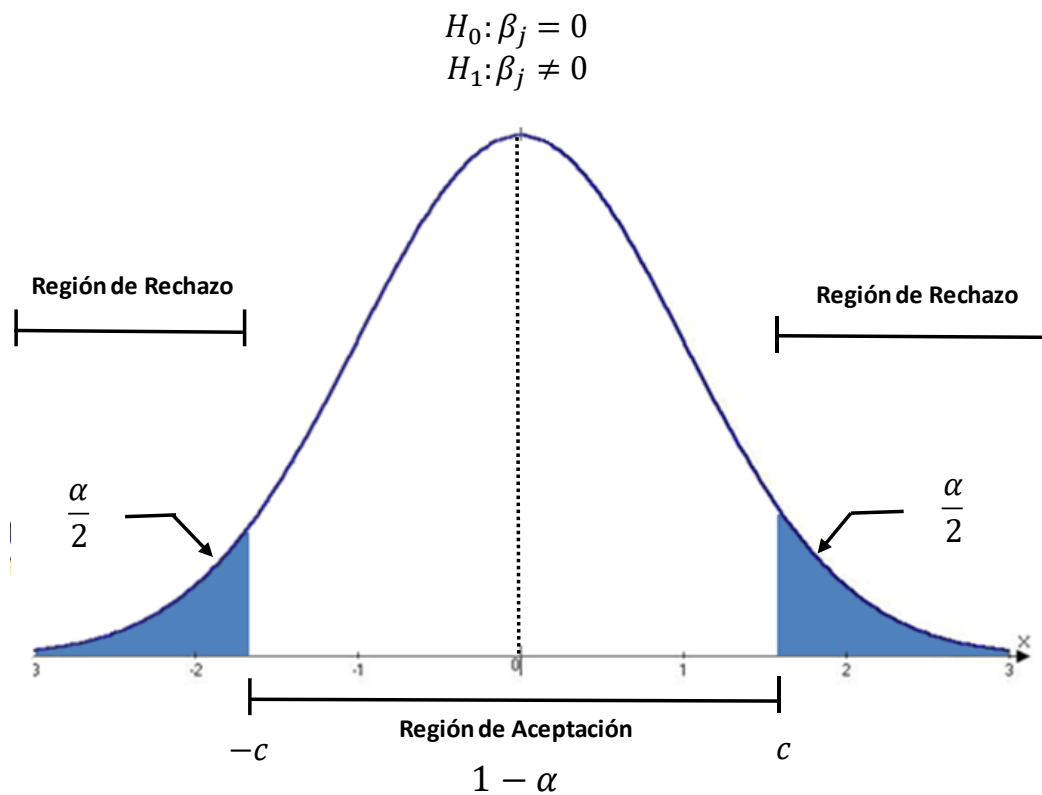


Figura 2.2 Prueba de hipótesis de dos colas. Elaboración propia.

En la figura 2.2 se aprecia que para una prueba de dos colas existen dos regiones de interés. Éstas se encuentran determinadas por el *valor crítico* c que establece el inicio de las regiones sombreadas (conocidas como regiones de rechazo).

Para el caso de pruebas de una cola, sólo existe una región crítica. Ésta se encuentra de lado derecho de la gráfica si la desigualdad ocupada en el enunciado de la hipótesis nula es “mayor que” ($>$), por otra parte si la desigualdad utilizada es “menor que” ($<$) la región crítica se encontrará de lado izquierdo de la gráfica.

Siguiendo con el procedimiento establecido por Montgomery y Runger (2013), el cuarto paso en la elaboración de pruebas de hipótesis es *elegir el nivel de significancia* α .

En este punto cabe recordar que en principio, la prueba de hipótesis supone que H_0 es verdadera, y lo que se hace en la prueba es buscar evidencia suficiente que demuestre lo contrario. Una cómica analogía con juicios legales podría establecer que: “La hipótesis nula es verdadera hasta que no se demuestre lo contrario”.

En estadística nunca se tendrá la certeza de estar en lo correcto, pero al trabajar con variables aleatorias es posible asociar probabilidades de éxito o de fracaso. El **nivel de significancia** es la probabilidad de rechazar H_0 cuando ésta es verdadera; es decir, es la probabilidad de cometer un error. A éste error, comúnmente se le denomina **error tipo I**.

Sin embargo, es posible que nuestra primera suposición (que la hipótesis nula es verdadera) sea falsa en realidad. Debido a esto, existe una probabilidad asociada a aceptar H_0 cuando debería de rechazarse. A este tipo de error se le conoce como **error tipo II**.

El hecho de cometer alguno de estos dos errores se debe a que en una prueba de hipótesis se trabaja con los datos de una muestra en particular, y el analista no tiene forma de saber si esa muestra pertenece al “común” de los datos (p.e. al 95% de ellos) o si fue extraída de un conjunto de datos poco usual.

A continuación, se presenta una tabla resumen de los tipos de errores:

Decisión	En realidad:	
	H_0 es verdadera	H_0 es falsa
No puede rechazarse H_0	no hay error	error tipo II
Se rechaza H_0	error tipo I	no hay error

Tabla 2.4 Resumen de los tipos de errores en una prueba de hipótesis. Elaboración propia con base en Montgomery y Runger (2013).

El nivel de significancia, denotado comúnmente como α , es una probabilidad que el analista controla. Es decir, el analista establece una probabilidad suficientemente aceptable para él de cometer un error tipo I. En palabras de Montgomery y Runger (2013): “Puesto que el analista puede controlar directamente la probabilidad de rechazar incorrectamente H_0 , al rechazo de la hipótesis nula H_0 siempre se le considera una **conclusión robusta**”.

Por otra parte, la probabilidad de cometer un error tipo II, denotada comúnmente como β , depende del verdadero valor del parámetro así como del tamaño de la muestra seleccionada. Para Montgomery y Runger (2013) la decisión de aceptar H_0 se considera una **conclusión débil**, a menos que se sepa que β es aceptablemente pequeña.

Finalmente, es importante definir a los complementos de α y β . El primero de ellos, obtenido de $1-\alpha$, es la *probabilidad de no rechazar H_0 cuando ésta es verdadera*; mientras que $1-\beta$ puede interpretarse como la *probabilidad de rechazar correctamente una H_0 falsa*. A la probabilidad $1-\beta$ se le conoce también como la **potencia de una prueba estadística** (Montgomery y Runger, 2013).

El siguiente paso en el procedimiento consiste en *establecer un estadístico de la prueba apropiado*. En el caso de MRL el estadístico más usado es el **estadístico t**, definido como:

$$t = \frac{(\hat{\beta}_j - a_j)}{ee(\hat{\beta}_j)} \quad (2.23)$$

En donde a_j es el valor hipotético de β_j y $ee(\hat{\beta}_j)$ es el error estándar de $\hat{\beta}_j$.

La mayoría de los paquetes estadísticos proveen un valor del estadístico t en su salida. Este valor hace referencia a la prueba de hipótesis de la ecuación (2.20), y el valor t es calculado como $t = \hat{\beta}_j / ee(\hat{\beta}_j)$ en el cual a_j es igual a cero (Wooldridge, 2015).

En este punto cabe preguntarse, ¿por qué usar el estadístico t como estadístico de prueba? La respuesta es sencilla: para la mayoría de las aplicaciones se desconoce el valor de la varianza σ^2 , por lo que es necesario reemplazarlo por su estimador $\hat{\sigma}^2$ (el procedimiento para obtener al estimador se explica en el Anexo B). La raíz cuadrada de $\hat{\sigma}^2$, conocida como el error estándar, es a su vez el estimador de la desviación estándar. Por tanto, el estadístico t permite medir a cuántas desviaciones estándar estimadas se encuentra $\hat{\beta}_j$ de cero (Wooldridge, 2015).

Cabe destacar un aspecto importante respecto al uso del estadístico t , y este es que para tamaños de muestra “pequeños” (digamos $n < 30$) la distribución a utilizar para encontrar el *valor crítico* (en donde empieza la región de rechazo) será una distribución t con $n - k - 1$ grados de libertad; en donde n es el tamaño de muestra y k el número de variables independientes incluidas en el modelo.

La apariencia de la distribución t es similar a la de la distribución normal estándar, la diferencia radica en que la primera tiene colas de mayor peso que la segunda; es decir, la probabilidad es más grande en las colas de la distribución t que en las de la distribución normal (Montgomery y Runger, 2013). Sin embargo, el peso de las colas se empieza a parecer a medida que el tamaño de la muestra o los grados de libertad aumentan. Por tanto, para un valor $n - k - 1$ “grande” ($n > 30$) es posible utilizar la distribución normal estándar.

El último paso para realizar una prueba de hipótesis consiste en *establecer la región de rechazo del estadístico*. En la figura 2.5 se aprecia un ejemplo de las regiones de rechazo para una prueba de dos colas.

La región de rechazo está determinada por el *valor crítico* c que no es más que el límite entre la(s) región(es) de rechazo y la región de aceptación (Montgomery y Runger, 2013). Obtener c es muy sencillo, sólo se necesita establecer un nivel de significancia (digamos, del 5%) y calcular los grados de libertad mediante la fórmula $n - k - 1$. Con estos dos datos se consulta el valor de c en tablas de distribución t .

Cabe destacar que para pruebas de dos colas, c es el percentil 97.5 en la distribución t con $n - k - 1$ grados de libertad, mientras que para pruebas de una cola c corresponde al percentil 95 (Wooldridge, 2015).

La *regla de rechazo* para una prueba de dos colas está determinada por:

$$|t| > c \quad (2.24)$$

En donde $|t|$ es el valor absoluto del estadístico t y c el valor crítico obtenido.

La interpretación de la regla de rechazo es que, si el valor absoluto del estadístico t obtenido es mayor que el valor crítico c , para un nivel de significancia dado, hay evidencia suficiente para rechazar la hipótesis nula a favor de la hipótesis alternativa. Por otra parte, si la regla de rechazo no se cumple, entonces, para un nivel de significancia dado, no hay evidencia suficiente que justifique el rechazo de la hipótesis nula.

2.2.7 Valor p

En el procedimiento para la realización de pruebas de hipótesis hay un paso que depende en gran medida de quién esté realizando el análisis, y éste es la elección del nivel de significancia.

Si un analista presenta sus resultados sin mayor información que la decisión de la aceptación o rechazo de la hipótesis nula a un nivel de significancia dado (escogido por él o ella), restringe a otros usuarios de la información el poder obtener conclusiones propias respecto de los datos. Esto se vuelve un problema importante cuando, mientras que para una persona el nivel de significancia del 5% es aceptable, para otra no lo sea.

A fin de evitar estas dificultades, es común presentar junto con el valor del estadístico de prueba el *valor-p* para cada variable explicativa.

La definición formal del *valor-p*, tomada de Montgomery y Runger (2013), es:

El valor-p es el nivel de significación más bajo que llevaría al rechazo de la hipótesis nula H_0 con los datos dados (p. 511).

Explicado por los mismos autores en otras palabras: “El valor-p es la probabilidad de que el estadístico de la prueba tome un valor que es *al menos tan extremo* como el valor observado del estadístico *cuando la hipótesis nula es verdadera*”.

Con el uso del *valor-p* se elimina la restricción de poder obtener conclusiones propias dado un nivel de significancia pre-seleccionado, permitiendo a los usuarios de la información tomar decisiones con *cualquier* nivel de significancia deseado.

Al ser el *valor-p* una probabilidad, sólo puede tomar valores de 0 a 1. Un *valor-p* “alto” (digamos > 0.10) proporciona *poca evidencia* contra la hipótesis nula, mientras que un valor “pequeño” proporciona *evidencia contra* la hipótesis nula.

Si bien tampoco existe una ley que rijan a partir de qué valor de p se puede aceptar o rechazar la hipótesis nula, su uso aporta la ventaja de que cualquier persona pueda definir, con su criterio, *qué tan significativos* son los datos.

Con el uso de programas estadísticos o econométricos la obtención del *valor-p* es bastante sencilla. Para una distribución t y prueba de dos colas, este valor está determinado por:

$$P(|T| > |t|) \quad (2.25)$$

En donde T denota una variable aleatoria con distribución t y $n - k - 1$ grados de libertad y t denota el valor numérico del estadístico de prueba (Wooldridge, 2015).

Para la prueba de una cola (ya sea derecha o izquierda) basta con dividir el resultado de la expresión (2.25) entre dos. Esto se debe a que la distribución t es simétrica respecto a cero (Ver figura 2.5).

2.2.8 Significancia práctica vs Significancia estadística

Una vez establecida la *significancia estadística* de una variable, es importante analizar si dicha variable cuenta con *significancia práctica* para justificar su inclusión en el modelo. La primera se encuentra determinada por el valor del estadístico de la prueba (el estadístico t en este caso) mientras que la segunda se encuentra determinada por la magnitud y el signo del estimador $\hat{\beta}_j$.

Al analizar si una variable explicativa pertenece al modelo o no, es muy importante no dejarse llevar sólo por el resultado de la significancia estadística. Además de ésta, se debe observar la magnitud del coeficiente para poder determinar si una variable es importante o no.

Lo anterior cobra relevancia al trabajar con muestras de tamaño grande, debido a la precisión de los estimadores obtenida. Con muestras de tamaño grande es posible encontrar significancia estadística en variables con *pequeñas* desviaciones de cero. En otras palabras, es probable obtener un valor del estadístico t lo suficientemente grande para rechazar la hipótesis nula $H_0: \beta_j = 0$ aun cuando el verdadero valor del parámetro es *muy cercano* a cero. En casos como éste, se debe ponderar si la significancia práctica encontrada justifica la inclusión de dicha variable en el modelo, a pesar de haber obtenido significancia estadística.

Un caso similar ocurre con muestras de tamaño pequeño. Es posible encontrar variables sin significancia estadística cuyo impacto práctico, sin embargo, sea grande. En casos como éste, es necesario calcular el *valor-p* del estadístico *t* para poder determinar si efectivamente la evidencia es suficiente para rechazar la hipótesis nula o si, por el contrario, la no significancia estadística pueda deberse al error de muestreo.

2.2.9 Intervalos de confianza.

En la estimación de un modelo de regresión lineal se cuentan con pistas que permiten analizar la precisión de las estimaciones, tales como el error estándar. Sin embargo, en ocasiones se desea contar con más información que la proporcionada por una estimación puntual.

Los *intervalos de confianza* proporcionan información complementaria a las estimaciones puntuales, permitiendo a los analistas conocer un rango (en lugar de un valor puntual) sobre el cual sea muy probable que se encuentre el verdadero valor del parámetro.

Para Montgomery y Runger (2013):

La interpretación de un intervalo de confianza es que, si se toma un número infinito de muestras aleatorias y se calcula un intervalo de confianza del $100(1-\alpha)$ por ciento para $[\beta_j]$ en cada muestra, entonces $100(1-\alpha)$ por ciento de estos intervalos incluirán el valor real de $[\beta_j]$ (p. 515).

Para estos autores, $1-\alpha$ denota el **coeficiente de confianza**, que multiplicado por 100 no es más que el porcentaje de las muestras cuyos intervalos contendrán a β_j . En la obtención de la muestra a analizar, se espera que ésta sea parte de ese $100(1-\alpha)$ por ciento.

Considerando la distribución *t* con $n - k - 1$ grados de libertad, el intervalo de confianza para β_j está determinado por:

$$\hat{\beta}_j - t_{\frac{\alpha}{2}, n-k-1} * ee(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2}, n-k-1} * ee(\hat{\beta}_j) \quad (2.25)$$

2.2.10 Prueba del estadístico F .

En el apartado de pruebas de hipótesis se estableció una manera de evaluar la significancia estadística de forma *individual* para cada variable explicativa en el modelo. Ahora, se analizará otra prueba capaz de analizar la significancia estadística *conjunta* o *general* de la regresión: la *prueba del estadístico F* o simplemente *Prueba F* .

La aplicación más común de esta prueba consiste en evaluar si los parámetros de las variables explicativas involucradas en el modelo son igual a cero. Se utiliza la misma metodología que para las pruebas de hipótesis antes vistas, sólo que, en lugar de un parámetro, la hipótesis nula contiene a todos los parámetros asociados a una variable explicativa:

$$\begin{aligned} H_0 &= \beta_1 = \beta_2 = \dots = \beta_k = 0 \\ H_1 &= \beta_j \neq 0 \text{ para al menos una } j \end{aligned} \quad (2.26)$$

Por consiguiente, el estadístico de prueba t ya no es de utilidad (debido a que éste se obtiene de un coeficiente en particular). El estadístico F es el indicado para ser utilizado como estadístico de prueba. Cabe destacar que del estadístico F no se pueden obtener conclusiones sobre una variable en particular, éste mide el efecto *en conjunto* de todas las variables involucradas (Wooldridge, 2015).

El estadístico F para la prueba de significancia general de una regresión está determinado por:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \quad (2.27)$$

En donde R^2 es el coeficiente de determinación, k el número de variables explicativas en el modelo y n el número de observaciones.

Si bien la prueba de significancia general es la aplicación más común del estadístico F , ésta es un caso especial de *pruebas de exclusión*.

Las pruebas de exclusión permiten evaluar si un *grupo* de variables ayudan o no en explicar a y . En este caso se compara la *suma de residuales al cuadrado* (SRC) del modelo que tiene a *todas* las variables explicativas contra la SRC del modelo *sin* el grupo de variables a evaluar. Siempre que se eliminan variables explicativas de un modelo, la SRC aumenta. La pregunta es si este aumento es suficientemente grande, *en relación* con la SRC del modelo que tiene todas las variables, como para rechazar la hipótesis de que los parámetros de las variables a evaluar son igual a cero (Wooldridge, 2015).

Al modelo con todas las variables se le conoce como **modelo no restringido (MNR)**, mientras que al modelo sin el grupo de variables a evaluar se le conoce como **modelo restringido (MR)**.

Para una prueba de exclusión de un grupo de variables, el estadístico F se define como:

$$F = \frac{(SRC_r - SRC_{nr})/q}{SRC_{nr}/(n - k - 1)} \quad (2.28)$$

En donde SRC_r es la suma de residuales al cuadrado del modelo restringido, SRC_{nr} es la suma de residuales al cuadrado del modelo no restringido, q es el

número de variables *excluidas* en el MR y k el número de variables explicativas en el MNR (Wooldridge, 2015).

Para poder utilizar a F como estadístico de prueba, se debe conocer su distribución de muestreo bajo la hipótesis nula: $H_0 = \beta_1 = \beta_2 = \dots = \beta_q = 0$. F está distribuida como una variable aleatoria F con $(q, n - k - 1)$ grados de libertad (Wooldridge, 2015); esto es:

$$F \sim F_{q, n-k-1} \quad (2.29)$$

La siguiente figura muestra de manera gráfica la distribución del estadístico F . En ella se puede apreciar que el valor de F **no puede ser** menor a cero:

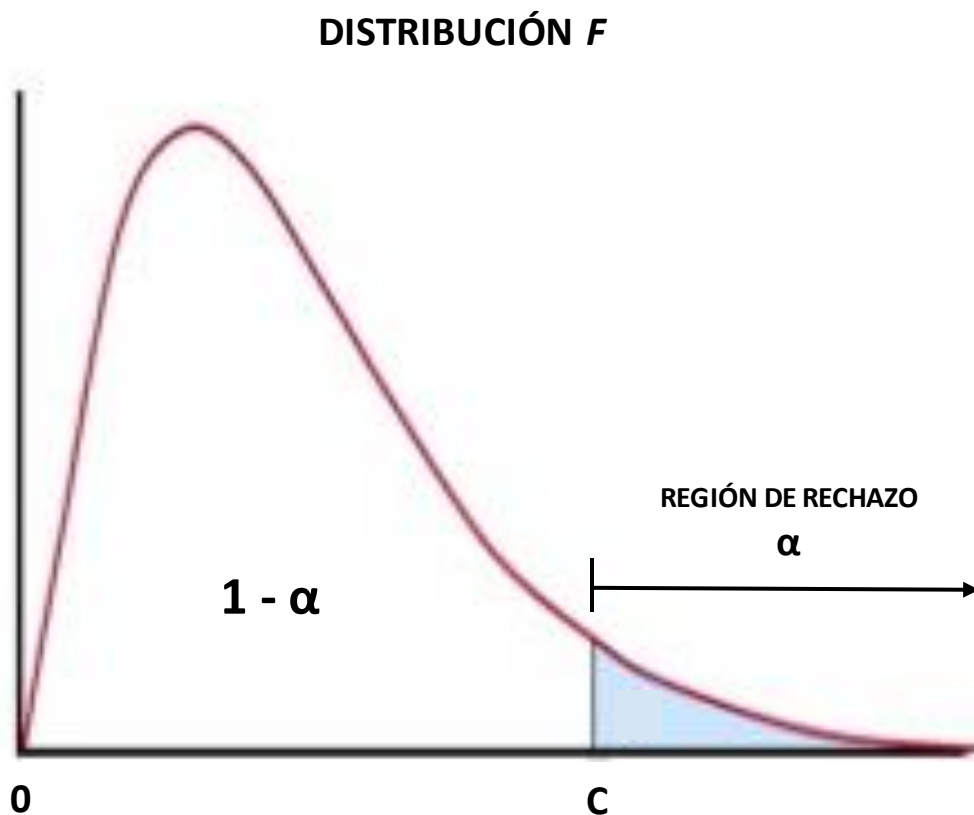


Figura 2.3 Distribución del estadístico F . Modificado de Bolívar Andrés *et al.* (2015).

La hipótesis nula se rechazará si para un nivel de significancia elegido, $F > c$.

Si el interés de un analista radica en el efecto de cierto grupo de variables, el estadístico F tiene dos grandes ventajas sobre la prueba del estadístico t . La primera de ellas es que con sólo un cálculo se puede determinar si la hipótesis nula se rechaza o no, mientras que, si se elige usar el estadístico t , se tendría que calcular este valor para cada variable del grupo de interés (además, en presencia de multicolinealidad las conclusiones basadas en pruebas t serían erróneas).

La segunda ventaja (y la más importante) es que al usar el estadístico F el problema de la multicolinealidad pierde importancia. Es de esperarse que las variables del grupo de interés se encuentren correlacionadas en cierta medida; al reportarse conclusiones basadas en el estadístico F éstas siempre se referirán al impacto en *conjunto*, considerando de antemano la correlación entre variables. No se puede decir lo mismo del estadístico t , en donde la multicolinealidad sí tiene un importante efecto negativo en los resultados obtenidos.

¿Y qué pasa en el caso de querer usar el estadístico F para una prueba de hipótesis con un solo parámetro (excluir sólo una variable)? La variable aleatoria F con $q = 1$ y $(n - k - 1)$ grados de libertad *es igual* al cuadrado de una variable aleatoria t con $(n - k - 1)$ grados de libertad; es decir:

$$F_{1,n-k-1} = t^2_{n-k-1} \quad (2.30)$$

Para Montgomery y Runger (2013) utilizar el estadístico F para medir la contribución de cada variable de manera individual “como si se tratara de la última variable agregada al modelo” resulta ser muy útil. En contraste, Wooldridge (2015) encuentra que: “Para la prueba de una sola hipótesis, el estadístico t es más flexible debido a que puede emplearse en pruebas contra alternativas de una cola. Dado que

los estadísticos t son más fáciles de obtener que los F , en realidad no hay razón para emplear un estadístico F en la prueba de hipótesis de un solo parámetro.”⁹

Tanto el estadístico t como el estadístico F aportan relevante información acerca de los parámetros de cualquier modelo de regresión, y de ambos se pueden extraer conclusiones interesantes. Por lo que a mi consideración ambas pruebas se deben complementar, más que sustituir, en función de encontrar suficiente evidencia que avale el rechazo o el no rechazo de la hipótesis nula.

Finalmente se presenta una forma alternativa de obtener el estadístico F , haciendo uso de los coeficientes de determinación R^2 . Las sumas de residuales al cuadrado y los coeficientes de determinación se encuentran relacionados linealmente, por lo que la fórmula (2.28) también se puede expresar como:

$$F = \frac{(R^2_{nr} - R^2_r)/q}{(1 - R^2_{nr})/(n - k - 1)} \quad (2.31)$$

En donde R^2_{nr} es el coeficiente de determinación del modelo no restringido (MNR) mientras que R^2_r es el coeficiente de determinación del modelo restringido (MR).

A (2.31) se le conoce como **forma R-cuadrada del estadístico F** (Wooldridge, 2015). Esta forma es utilizada cuando la salida del programa estadístico incluye los valores de R^2 .

⁹ En este punto, el argumento del autor es que a diferencia del estadístico t en el cual sólo hay que obtener el cociente entre el coeficiente y el error estándar de la estimación, para obtener F se debe calcular la suma de residuales al cuadrado (SRC).

2.2.11 Variables binarias (*Dummy variables*).

En la elaboración de análisis causales, existen por lo general características o variables que no pueden ser medidas de forma *cuantitativa*. La naturaleza de estas variables se define con un simple *sí* o *no*: ¿El individuo en cuestión es hombre o no lo es? ¿El estudiante en cuestión asistió a asesorías o no lo hizo? ¿La empresa en cuestión realizó eventos de marketing o no los hizo?

Estas características, medidas de forma *cualitativa*, se enfocan por lo general en eventos de *éxito* o de *fracaso*; por lo que un evento de *éxito* puede ser representado por un 1, mientras que un evento de *fracaso* puede ser representado por un 0. De ahí el término *binario* a éste tipo de variables.

Una definición más precisa, tomada de Chase Jr. (2013) establece que:

[Las variables binarias] son variables que pueden tomar el valor de cero o de 1 y son usadas para indicar la presencia o ausencia de una o más características cualitativas (p. 191).

La inclusión de una o más variables binarias al modelo de regresión no altera en absoluto la mecánica para la obtención de las estimaciones por MCO. La única diferencia, respecto a variables independientes cuantitativas, es la *interpretación de los coeficientes* (Wooldridge, 2015).

Una de las aplicaciones más interesantes de las variables binarias es poder encontrar *diferencias* entre dos grupos de interés. Para integrar variables binarias a un modelo de regresión se debe tomar en cuenta quién será considerado el **grupo de referencia**. El grupo de referencia es aquel contra el que se hacen las comparaciones (Wooldridge, 2015).

Como ejemplo de lo anterior, en una investigación sobre el efecto de las asesorías académicas, se podría considerar como grupo de referencia a aquellos estudiantes que *no* hayan asistido a asesorías; por lo que el modelo de regresión contendrá una variable binaria (denominada *aseso*) con el valor de 1 si el estudiante acudió a asesorías y con el valor de 0 si no fue así. La importancia de la definición de un grupo de referencia radica en la interpretación de los coeficientes.

En este caso, el coeficiente de la variable *aseso* indicará: *la diferencia encontrada en el rendimiento académico para estudiantes que sí asistieron a asesorías con relación a aquellos que no, dados los mismos niveles en rendimiento previo, motivación, calidad de la enseñanza, etc.*

En otras palabras, la variable binaria *aseso* permitirá identificar el impacto de las asesorías académicas entre estudiantes con características similares. Es de esperarse que la diferencia encontrada entre aquellos que sí asistieron a asesorías y aquellos que no, sea positiva ($\delta_0 > 0$).

Al considerar incluir *dummy variables* en un modelo de regresión se debe tener cuidado en no sobre especificar al modelo con variables binarias que representen *a un mismo grupo de interés*. Si se agrega más de una variable para un mismo grupo de interés, indudablemente se presentará multicolinealidad (además de violarse el tercer supuesto de los supuestos Gauss-Markov), y por ende los resultados obtenidos de la estimación serán poco confiables (Montgomery y Runger, 2013; Wooldridge, 2015).

La regla de oro indica que para g grupos de interés, se deben integrar $g-1$ variables binarias (Montgomery y Runger, 2013; Wooldridge, 2015). Esto se debe a que el intercepto (β_0) obtenido de la estimación (de un modelo con variables binarias) *es el intercepto* del grupo de referencia, mientras que los coeficientes de las variables binarias representan *la diferencia estimada entre el intercepto de ese grupo y el grupo de referencia* ($\beta_0 + \delta_0$). Por consiguiente integrar más variables que $g-1$ no tiene ninguna utilidad y sólo conllevará a problemas en la obtención de los resultados (Wooldridge, 2015).

Un modelo con variables binarias puede ser sujeto a la transformación *log – nivel* en su forma funcional y su interpretación será la misma que la presentada en la figura 2.2; es decir, el coeficiente de la variable binaria tendrá una interpretación porcentual. Así mismo, se pueden incluir variables explicativas cuantitativas al

modelo, con forma funcional cuadrática, y su interacción con variables binarias no afectará el procedimiento de estimación.

Si bien es sumamente común encontrar variables binarias dentro del grupo de variables explicativas de algún modelo, cabe la posibilidad de encontrar también *variables dependientes binarias*. Al igual que en el caso anterior, la diferencia respecto a otros modelos radica en la interpretación de los coeficientes, pero además en este caso, también cambia la interpretación de la variable dependiente.

A los modelos con variables dependientes binarias se les conoce como **modelos de probabilidad lineal (MPL)**. Esto se debe a que la variable y , ahora restringida a los valores de cero o uno, se interpreta como una probabilidad de *éxito* (la probabilidad de que y sea igual uno). El *valor esperado* de y , condicionado a los valores de las variables independientes, ahora se convierte en la *probabilidad* de que y sea igual a uno, condicionada a los valores de las variables independientes; es decir:

$$P(y = 1|x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (2.32)$$

En el MPL, β_j mide la variación de la probabilidad de éxito al variar x_j , permaneciendo los demás factores constantes [...] con esto en mente, el MPL permite estimar el efecto de diversas variables explicativas sobre un evento cualitativo (Wooldridge, 2015).

El mismo autor advierte sobre la importancia de utilizar valores cercanos al promedio de la muestra al sustituir en las variables independientes, ya que al utilizar valores extremos es posible que el analista encuentre predicciones para y menores a cero o mayores a uno (lo cual es imposible en términos de probabilidad).

Cabe mencionar que existen otros modelos para analizar el comportamiento de variables dependientes binarias que inclusive superan las limitaciones del MPL, tales como el modelo *logit* o *probit*; sin embargo, el análisis de estos modelos queda fuera del alcance de esta tesis.

Finalmente, con respecto a los MPL cabe preguntarse: ¿existe alguna métrica que permita evaluar su “efectividad” de predicción? Wooldridge (2015) nos comparte una, definida como el **porcentaje predicho correctamente**. Esta métrica se obtiene definiendo como 1 al valor predicho \tilde{y}_i cuando el valor ajustado $\hat{y}_i \geq 0.5$ y como cero ($\tilde{y}_i = 0$) cuando el valor ajustado $\hat{y}_i < 0.5$. Empleando los datos de los valores reales (y_i) y los valores predichos (\tilde{y}_i) se puede obtener la proporción de predicciones correctas con la cual es posible evaluar la “efectividad” del modelo.

2.2.12 Heterocedasticidad.

La obtención de estadísticos de prueba t y F analizados en secciones anteriores, se ha basado en el importante supuesto de *Homocedasticidad* que establece una varianza del término del error μ **constante** para cualesquiera sean los valores de las variables explicativas (ver Anexo A).

Dado que dichos estadísticos son sumamente relevantes para el análisis inferencial, es muy importante conocer formas de probar la existencia de heterocedasticidad y, en caso de existir, saber utilizar las metodologías estadísticas existentes para atenuar o eliminar su impacto sobre análisis inferenciales tales como pruebas de hipótesis e intervalos de confianza.

Antes que otra cosa, cabe preguntarse ¿por qué el problema de la heterocedasticidad es tan importante? Suponiendo un modelo de la forma (2.3) *con* heterocedasticidad, si se seleccionan a dos grupos de individuos (empresas, ciudades, etc.) y se obtienen sus valores correspondientes de las variables x_1, x_2, \dots, x_k , entonces se contará con *dos* valores de varianza para el error μ , uno por cada grupo de individuos seleccionado.

Con dos valores de varianza diferentes se tendrán también dos estimaciones de la varianza del error diferentes, en ese caso ¿cuál estimación utilizar para calcular $Var(\hat{\beta}_j)$? En este punto es muy fácil deducir que cualquiera sea la decisión (utilizar la estimación del grupo 1 o del grupo 2) se estará *sesgando* el procedimiento para obtener $Var(\hat{\beta}_j)$ y por consiguiente se obtendrá un error estándar $ee(\hat{\beta}_j)$ sesgado, que no servirá para realizar pruebas de hipótesis ni para obtener intervalos de

confianza. El problema empeora al considerar n grupos de individuos diferentes; he ahí la importancia de analizar a detalle el supuesto de homocedasticidad.

Existen varias maneras para detectar heterocedasticidad en los modelos de regresión; una de las más sencillas es la comparación gráfica entre los residuales y cada una de las variables explicativas. Esta comparación permite detectar como *variable más culpable de heterocedasticidad* aquella cuyo gráfico se separe más de la aleatoriedad (Pérez López, 2007).

Existen, además, pruebas más rigurosas, tales como la prueba Goldfeld-Quandt, la prueba Glesjer, la prueba de White y la prueba Breusch-Pagan. En este apartado se analizarán solamente las últimas dos.

Prueba Breusch-Pagan.

Esta prueba hace uso de la *prueba de hipótesis* para encontrar evidencia que señale la existencia de heterocedasticidad. Para ello, se establece un modelo como el siguiente:

$$\mu^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k + v \quad (2.33)$$

En donde μ^2 representa al cuadrado del error en la ecuación (2.3) y v es un término de error con media cero dadas las x_j . Esta ecuación establece de manera explícita una relación entre el cuadrado del error de la ecuación original y las variables explicativas. La importancia radica en el valor de los coeficientes δ que, en caso de ser cero, indicarán el cumplimiento del supuesto de homocedasticidad (Wooldridge, 2015).

Con base a los datos obtenidos, se debe encontrar evidencia suficiente que señale la presencia de heterocedasticidad, por lo que la hipótesis nula a evaluar es:

$$H_0: \delta_1 = \delta_2 = \dots = \delta_k = 0 \quad (2.34)$$

Wooldridge (2015) resume la prueba de Breusch-Pagan en tres sencillos pasos:

1.- Estimar el modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \mu$ por MCO, como de costumbre. Obtener los residuales cuadrados de MCO, $\hat{\mu}^2$ (uno para cada observación).

2.- Ejecutar la estimación del modelo (2.33) utilizando los residuales cuadrados del modelo original, es decir:

$$\hat{\mu}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k + error \quad (2.35)$$

Conservar la R-cuadrada de esta regresión, $R^2_{\hat{\mu}^2}$.

3.- Formar el estadístico F y calcular el valor-p (usando la distribución $F_{k, n-k-1}$). Si el valor-p es suficientemente pequeño, es decir, menor que el nivel de significancia elegido, se rechaza la hipótesis nula de homocedasticidad.

$$F = \frac{R^2_{\hat{\mu}^2}/k}{(1 - R^2_{\hat{\mu}^2})/(n - k - 1)} \quad (2.36)$$

Es posible que después de realizar una comparación visual de las gráficas entre los residuales y las variables explicativas, se tenga una idea de qué variables sean posibles causantes de heterocedasticidad; con esta información se podría reducir el número de variables en el modelo (2.33) a sólo aquellas encontradas como variables de interés.

Prueba de White.

Esta otra prueba consiste en la inclusión de los cuadrados y los productos cruzados de todas las variables explicativas en la ecuación (2.35). Con esta

modificación, White pretendía sustituir el supuesto de homocedasticidad por el supuesto más débil de que el error cuadrado no está correlacionado con ninguna de las variables independientes (x_j), ni con los cuadrados de las variables independientes (x_j^2), ni con ninguno de los productos cruzados ($x_j x_h$ para $j \neq h$) (Wooldridge, 2015).

El problema de este enfoque es que el número de variables independientes en la ecuación (2.35) aumenta de forma sustancial por cada término cuadrado y cada producto cruzado añadido; sin embargo, el mismo Wooldridge (2015) presenta una modificación a la prueba de White que elimina este inconveniente al utilizar como variables independientes en la ecuación (2.35) al valor ajustado \hat{y} y su cuadrado \hat{y}^2 obtenidos de la estimación del modelo original, *en lugar* de las variables explicativas, sus funciones cuadráticas y sus productos cruzados.

La justificación de esta sustitución es que los valores ajustados son funciones lineales de las variables independientes, por lo que si éstos se elevan al cuadrado se obtiene una función particular de todos los cuadrados y productos cruzados de las variables independientes (Wooldridge, 2015).

Con base a lo anterior, es posible estimar la ecuación:

$$\hat{\mu}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error \quad (2.37)$$

y establecer la hipótesis nula:

$$H_0: \delta_1 = \delta_2 = 0 \quad (2.38)$$

para poder probar la existencia de heterocedasticidad.

El procedimiento, obtenido de Wooldridge (2015) para este caso en particular de la prueba de White, se resume a continuación:

- 1.- Estimar el modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \mu$ por MCO, como de costumbre. Obtener los residuales $\hat{\mu}$ de MCO y los valores ajustados \hat{y} y calcular sus respectivos cuadrados, $\hat{\mu}^2$ y \hat{y}^2 .
- 2.- Ejecutar la regresión de la ecuación (2.37) y conservar su R-cuadrada, $R^2_{\hat{\mu}^2}$.
- 3.- Formar el estadístico F y calcular el valor-p, empleando la distribución $F_{2,n-3}$.

Sobre las pruebas para detectar heterocedasticidad cabe destacar que, para su buen funcionamiento, éstas requieren de una especificación correcta de la forma funcional. En palabras de Wooldridge (2015): “Es mejor utilizar pruebas explícitas para formas funcionales primero, puesto que la especificación errónea de las formas funcionales es más importante que la heterocedasticidad. Una vez satisfechos con la forma funcional, se puede probar la heterocedasticidad”.

Medidas correctivas a la heterocedasticidad.

Una vez verificada la presencia de heterocedasticidad, se debe elegir un curso de acción para mitigar sus efectos sobre la estimación de la varianza poblacional de algún parámetro. En general, existen dos alternativas para abordar este problema: la primera de ellas consiste en utilizar otro método de estimación que considere el problema de la heterocedasticidad, tal como el método de Mínimos Cuadrados Ponderados (MCP). La base de este método consiste en determinar una función de las variables explicativas $h(x_1, x_2, \dots, x_k)$ que permita identificar el *tipo* de heterocedasticidad, y con base en ello, corregirla.

La cuestión con el método de MCP es que se basa en suposiciones tales como la *forma* de heterocedasticidad (la naturaleza de la función $h(x_1, x_2, \dots, x_k)$), las cuales, en caso de ser incorrectas, conllevarán a resultados erróneos. Por ende, se deben analizar a detalle dichas suposiciones y realizar procedimientos adecuados para comprobar su veracidad. Este análisis queda fuera del alcance de esta tesis.

El otro curso de acción se basa en una modificación a la obtención de $Var(\hat{\beta}_j)$ utilizando el método usual de MCO. Con base en Wooldridge (2015), White (1980) demostró que en presencia de heterocedasticidad es posible utilizar los residuales cuadrados obtenidos de la regresión del modelo original para calcular una estimación robusta a la heterocedasticidad. Siendo:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \hat{\mu}_i^2}{SRC_j} \quad (2.39)$$

Donde \hat{r}_{ij} denota el i -ésimo residual de regresar x_j sobre el resto de las variables independientes, $\hat{\mu}_i^2$ el i -ésimo residual cuadrado de la regresión inicial de Y sobre x_1, x_2, \dots, x_k y SRC_j es la suma de residuales cuadrados de esta regresión; a la raíz cuadrada obtenida de (2.39) se le conoce como **error estándar de $\hat{\beta}_j$ robusto a la heterocedasticidad** (Wooldridge, 2015).

Este procedimiento realiza una estimación que, bajo justificación asintótica, permite seguir haciendo uso del estadístico t usual, con la única diferencia de utilizar el error estándar robusto a la heterocedasticidad para su obtención.

Al respecto, Wooldridge (2015) advierte sobre la tentación de utilizar el error estándar robusto a la heterocedasticidad para cualquier aplicación, independientemente de la existencia o no de heterocedasticidad:

Una razón por la que se utilizan los errores estándar usuales en el trabajo con cortes transversales es que, si el supuesto de homocedasticidad se satisface y los errores están distribuidos normalmente, los estadísticos t usuales tienen distribuciones t exactas, sin importar el tamaño de muestra. Los errores estándar robustos y los estadísticos t robustos sólo se justifican a medida que el tamaño de la muestra se vuelve grande [...] con tamaños de muestra pequeños, el estadístico t robusto puede tener distribuciones que no estén muy próximas a la distribución t y que podrían echar a perder la inferencia (p. 276).

También es posible obtener un estadístico F robusto a la heterocedasticidad, mejor conocido como *estadístico de Wald robusto a la heterocedasticidad*. Esta versión del estadístico no se puede calcular con los residuales cuadrados o R-cuadradas de los modelos restringido y no restringido, si no que se necesitan realizar análisis matriciales y asintóticos. Muchos programas estadísticos ya proveen de comandos para obtener este valor.

2.3 Resumen del capítulo

En este capítulo se presentó el término *econometría*, el cual hace referencia al uso de métodos estadísticos para estimar relaciones y probar teorías económicas. Se analizaron aplicaciones alternativas a estos métodos como en la planeación de la demanda futura para cualquier rama industrial, por ejemplo. Así mismo, se presentaron los tipos de datos existentes para análisis de esta naturaleza.

En el segundo apartado de este capítulo se presentó el modelo de regresión lineal y el método más común para su estimación: los Mínimos Cuadrados Ordinarios o MCO. Se analizaron los supuestos clave que permiten deducir *causalidad* entre dos variables, una dependiente y otra independiente. El supuesto más importante para este caso es el denominado **supuesto de media condicional cero**. Así mismo se presentaron indicadores que permiten analizar la proporción o el porcentaje *explicado* por la recta de regresión de MCO dada la dispersión de los datos en la muestra; siendo los más importantes el *coeficiente de determinación* R^2 y el *coeficiente de determinación ajustado* \bar{R}^2 .

Dado que la varianza de una estimación tiende a infinito en presencia de variables independientes altamente correlacionadas, se encontró que la *multicolinealidad* es un problema serio, al producir estimaciones sesgadas e inconsistentes. Las alternativas más aceptadas para resolver este problema son incrementar el tamaño de la muestra o combinar las variables altamente correlacionadas en una sola.

Por otra parte, se presentaron las herramientas básicas de la inferencia estadística, tales como los intervalos de confianza y la prueba de hipótesis. Se encontró que los estadísticos de prueba t y F fungen un importante papel en el análisis inferencial; sin embargo, al obtenerse con un nivel de significancia predeterminado y a juicio del analista, es muy importante presentar además los *valores-p* de estos estadísticos, en función de proveer mayor evidencia que respalde las conclusiones obtenidas. Al respecto, se destacó la importancia de diferenciar entre *significancia práctica* y *significancia estadística* para determinar las variables a incluir en el modelo.

Se analizaron también las diferentes *formas funcionales* en las que se puede estimar una ecuación de regresión, especialmente con términos logarítmicos y

cuadráticos. Se encontró que los primeros permiten medir el impacto dado un cambio porcentual en la variable de interés, mientras que los segundos permiten captar rendimientos no constantes entre dos variables. Se advirtió además, del aumento en la dificultad, al tratar de interpretar los coeficientes en ecuaciones con términos logarítmicos y cuadráticos combinados.

Debido a la necesidad en muchas aplicaciones de considerar información cualitativa, se presentaron a las *variables binarias* o *dummy variables* como una solución para la integración al análisis cuantitativo, de este tipo de información. Al tomar valores de cero o uno, se encontró que la inclusión de estas variables no afecta en absoluto al método de estimación MCO. La única diferencia radica en la interpretación de los coeficientes. Se analizó además al Modelo de Probabilidad Lineal (MPL) el cual deriva del uso de una variable dependiente binaria. La interpretación de la variable dependiente cambia ya que en este modelo, el valor esperado de y se vuelve una probabilidad de éxito condicionada al valor de las variables explicativas.

Finalmente, se designó un apartado para tratar el importante problema de la *heterocedasticidad*. Este problema se definió como el **sesgo** que se presenta al calcular $Var(\hat{\beta}_j)$ debido a una varianza no constante del término de error μ condicionado a los valores de las x_i . Se vio que esto repercute para el cálculo de estadísticos t y F y en la obtención de intervalos de confianza y pruebas de hipótesis. Se presentaron además dos pruebas estadísticas para detectar este problema: la *prueba Breusch-Pagan* y la *prueba de White*; así mismo, se analizaron medidas correctivas a este problema, siendo la más sencilla la obtención de estadísticos t y F robustos a la heterocedasticidad.



Cabús restaurado, Facultad de Ingeniería
Foto: www.twitter.com/fundacion_unam

3.1 Propuesta metodológica

En el desarrollo de la metodología se utilizarán como ejemplo, variables que corresponden a la Facultad de Ingeniería de la UNAM; sin embargo, con modificaciones pertinentes en función de la información disponible, esta metodología podrá ser aplicada por cualquier institución educativa.

A continuación, se presentan los pasos detallados para la creación de un modelo que permita cuantificar el impacto de las acciones emprendidas por las instituciones educativas para elevar el rendimiento académico.

3.1.1 Paso 1. Identificación de variables candidatas.

Con base en la investigación presentada en el primer capítulo, respecto a las variables que han sido relacionadas con el rendimiento académico en estudios previos, para este paso se cuenta con un marco de referencia que será utilizado como un primer *filtro*, que ayudará a delimitar variables candidatas que *pueden* ser incluidas en el (o los) modelo(s).

En el caso concreto de la Facultad de Ingeniería, las variables propuestas para controlar el factor cognitivo, motivacional y psicosocial del estudiante, son:

Descripción	Tipo	Medición	Identificador
<i>Factor Cognitivo</i>			
Rendimiento Previo	Cuantitativa	Promedio de bachillerato	<i>prombach</i>
Formación previa en matemáticas	Binaria	Mala o Regular = 0, Buena = 1	<i>nivelmate</i>
Capacidad evaluada al ingreso de la carrera	Cuantitativa	Puntuación en examen de ingreso	<i>examingr</i>
Capacidad autopercebida para el estudio	Binaria	Baja o Media = 0, Alta = 1	<i>capestudio</i>
Iniciativa	Binaria	Baja o Media = 0, Alta = 1	<i>iniciativa</i>
Administración del tiempo	Binaria	Mala o Regular = 0, Buena = 1	<i>admintempo</i>
<i>Factor Motivacional</i>			
Motivación escolar	Binaria	Baja o Media = 0, Alta = 1	<i>motiv</i>
<i>Factor Psicosocial</i>			
Nivel de Integración	Binaria	Baja o Media = 0, Alta = 1	<i>nivintegr</i>
Nivel de Autorregulación	Binaria	Baja o Media = 0, Alta = 1	<i>nivautorr</i>

Tabla 3.1 Variables propuestas para controlar factores debidos al estudiante. Elaboración propia.

La columna *Tipo*, hace referencia al tipo de variable del que se trata, para la tabla 3.1 sólo existen dos tipos de variables: cuantificable o binaria. La primera es utilizada para aquellas variables que pueden ser medidas con un valor numérico, como es el caso del promedio de bachillerato. La segunda se utiliza para registrar aspectos cualitativos, del tipo *cierto* o *falso*.¹⁰

En la columna *Medición* se ha establecido la forma de registrar los datos correspondientes a cada variable. Para el caso de las variables binarias, estudiantes que se clasifiquen en las categorías *Mala*, *Regular*, *Baja* o *Media* contarán con un valor de cero en las variables respectivas; éstos serán considerados el **grupo de referencia**. Estudiantes que se clasifiquen en las categorías *Buena* o *Alta* contarán con el valor de uno, y los coeficientes asociados a estas variables indicarán la *diferencia* existente entre este grupo y el grupo de referencia.

¹⁰ Las variables candidatas han sido obtenidas de los cuestionarios que la Facultad de Ingeniería realiza a sus estudiantes al momento de ingresar a la institución. Dichos cuestionarios son de opción múltiple; sin embargo, con la finalidad de disminuir el número de variables, se han agrupado las opciones *Mala* o *Regular* (y *Baja* o *Media*) en un solo grupo. Todos los grupos con valor igual a cero son considerados grupos de referencia.

Cabe aclarar que la mayoría de las variables para estos factores y para los subsecuentes, son de tipo binario u ordinal, debido al método de recolección de información que para el caso de la Facultad de Ingeniería se realiza con encuestas en línea de opción múltiple. Como se verá en secciones subsecuentes, esto representa una importante limitación en el análisis de los datos recolectados.

Por último, la columna *Identificador* no es más que un nombre abreviado de la variable, el cual pueda ser usado en las ecuaciones de regresión.

Por otra parte, las variables propuestas para controlar los factores sociodemográfico y socioeconómico, son:

Descripción	Tipo	Medición	Identificador
<i>Factor Sociodemográfico</i>			
Género	Binario	Hombre = 1, Mujer = 0	<i>genero</i>
Tipo de hogar	Ordinal	Vive con sus padres = 0, Vive con algún tutor = 1, Vive sol@ = 2	<i>hogar</i>
Nivel educativo del padre ¹	Ordinal	Nivel básico = 0, Nivel Intermedio = 1, Nivel Superior = 2	<i>edupadre</i>
Nivel educativo de la madre ¹	Ordinal	Nivel básico = 0, Nivel Intermedio = 1, Nivel Superior = 2	<i>edumadre</i>
Tiempo de traslado a la universidad (ida/vuelta)	Ordinal	< 1 hora = 0, Entre 1 y 2 horas = 1, > 2 horas = 2	<i>traslado</i>
<i>Factor Socioeconómico</i>			
Ingreso familiar promedio (mensual)	Ordinal	< \$5,000 = 0, Entre \$5,000 y \$9,000 = 1, > \$9,000 = 2	<i>ingreso</i>
Dependientes del ingreso familiar	Ordinal	1 o 2 personas = 0, De 2 a 4 personas = 1, Más de 4 personas = 2	<i>depingreso</i>
Acceso a computadora e internet en el hogar	Binario	Lo tiene = 1, No lo tiene = 0	<i>compeinter</i>
Tipo de vivienda	Binario	Propia = 1, Rentada = 0	<i>vivienda</i>
Situación laboral (Trabaja o no trabaja)	Binario	Trabaja = 1, No trabaja = 0	<i>trabaja</i>
Situación laboral (Hrs trabajadas por semana)	Ordinal	Menos de 25 = 0, Entre 25 y 40 = 1, Más de 40 = 2	<i>hrstrabajo</i>

¹ El nivel básico lo integran primaria y secundaria, el nivel intermedio bachillerato y el nivel superior estudios universitarios y de posgrado

Tabla 3.2 Variables propuestas para controlar factores debidos al contexto. Elaboración propia.

En la tabla 3.2 se agregó un nuevo tipo de variable, denominada **variable ordinal**, la cual integra aspectos cualitativos pertenecientes a diferentes categorías, como es el caso del *Nivel educativo de los padres*, cuyas categorías son *Nivel Básico*, *Nivel Intermedio* o *Nivel Superior*.

La interpretación del coeficiente asociado a una variable ordinal es la misma que la de una variable binaria, la diferencia radica en que para la primera se necesita establecer más de una variable explicativa, que permitan identificar diferencias entre cada categoría y el grupo de referencia.

Como ejemplo de lo anterior, se establece un modelo de regresión para evaluar el impacto del ingreso familiar mensual (*ingreso*) en el promedio de un estudiante (*promindiv*):

$$promindiv = \beta_0 + \beta_1 ingreso_1 + \beta_2 ingreso_2 + \mu \quad (3.1)$$

La variable *ingreso*₁ agrupa a aquellas familias con un ingreso promedio de entre \$5,000 y \$9,000 mensuales. Por otra parte, la variable *ingreso*₂ agrupa a aquellas familias con un ingreso promedio mayor a \$9,000 mensuales (ver figura 3.2). En la ecuación anterior, no se ha agregado una variable para las familias con ingresos inferiores a los \$5,000 mensuales dado que estas familias son el grupo de referencia.

La interpretación de β_1 es la siguiente: este parámetro, indica la diferencia existente en el promedio individual, para estudiantes con ingresos familiares mensuales de entre \$5,000 y \$9,000, con relación a aquellos estudiantes cuyos ingresos familiares son menores a los \$5,000 mensuales.

El parámetro asociado a *ingreso*₂ indica lo mismo, pero ahora tomando en cuenta a los estudiantes de la tercer categoría; es decir, el parámetro β_2 indica la diferencia existente en el promedio individual, para estudiantes con ingresos familiares mensuales mayores a \$9,000, con relación a estudiantes con ingresos familiares mensuales menores a \$5,000.

La importancia de las variables ordinales radica en que permiten identificar diferencias entre categorías diferentes de un mismo grupo.

A continuación se presentan las variables candidatas para controlar el factor docente:¹¹

¹¹ Las variables propuestas para medir el factor docente han sido obtenidas de las encuestas realizadas a profesores y estudiantes al finalizar cada semestre.

Descripción	Tipo	Medición	Identificador
<i>Factor Docente</i>			
Nivel de conocimiento en la materia ¹	Binario	Medio o Bajo = 0, Alto = 1	<i>conocmat</i>
Experiencia	Cuantitativa	Años impartiendo clase	<i>exper</i>
Certificaciones de enseñanza obtenidas	Cuantitativa	Número de certificaciones obtenidas	<i>certifi</i>
Calidad de la enseñanza ¹	Binario	Baja o Media = 0, Alta = 1	<i>calense</i>
Uso de recursos didácticos no tradicionales	Binario	Sí usa = 1, No usa = 0	<i>recdidac</i>
Participación en actividades académicas ²	Ordinal	Siempre = 0, Casi siempre = 1, Algunas veces = 2, Nunca = 3	<i>actiaca</i>

¹ Valorado por los alumnos

² Este factor se refiere a las actividades académicas convocadas por la coordinación a la que pertenece el docente

Tabla 3.3 Variables propuestas para controlar el factor debido al docente. Elaboración propia.

Por último, se definen las variables de interés para la institución educativa, dado que éstas permitirán evaluar la pertinencia de los programas de apoyo, además de que, sobre ellas, se tiene control en cuanto a los recursos financieros, materiales y/o humanos asignados:

Descripción	Tipo	Medición	Identificador
<i>Programas de Apoyo</i>			
Asesorías académicas	Binario	Acudió = 1, No acudió = 0	<i>asesoaca</i>
Tutorías	Binario	Acudió = 1, No acudió = 0	<i>tutoria</i>
Exámenes extraordinarios en tres etapas	Binario	Realizo examen de este tipo = 1, No realizo examen de este tipo = 0	<i>tresetapas</i>
Cursos extraordinarios	Binario	Acudió = 1, No acudió = 0	<i>cursoextra</i>
Exámenes extraordinarios c/taller preparación	Binario	Realizo examen de este tipo = 1, No realizo examen de este tipo = 0	<i>extrataller</i>
Talleres de ejercicios	Binario	Acudió = 1, No acudió = 0	<i>taller</i>
Asesorías psicopedagógicas	Binario	Acudió = 1, No acudió = 0	<i>asesopsico</i>

Tabla 3.4 Variables de interés para la institución educativa. Elaboración propia.

Todas las variables de la tabla 3.4 son de tipo binario debido a que, como se analizó en el capítulo II, este tipo de variables permiten identificar la **diferencia** entre dos grupos de interés.

También se presentan las variables pertinentes a la medición del rendimiento académico, éstas son:

Descripción	Tipo	Medición	Identificador
<i>Rendimiento Académico</i>			
Promedio individual	Cuantitativa	Nominal	<i>promindiv</i>
Promedio generacional	Cuantitativa	Nominal	<i>promgene</i>
Avance en créditos	Cuantitativa	Nominal o Porcentaje	<i>creditos</i>
Tasa de titulación	Cuantitativa	Porcentaje	<i>titulado</i>
Tasa de reprobación	Cuantitativa	Porcentaje	<i>reprobado</i>
Tasa de deserción	Cuantitativa	Porcentaje	<i>deserto</i>
Tasa de rezago	Cuantitativa	Porcentaje	<i>rezago</i>
Reprobación	Binaria	Reprobo al menos una materia de la DCB = 1, No reprobo = 0	<i>reprobo</i>

Tabla 3.5 Variables propuestas para medir el rendimiento académico. Elaboración propia.

Al identificar variables candidatas, lo más importante es contar con información disponible. Al respecto, la Facultad de Ingeniería cuenta con información de todas las variables presentadas en las figuras anteriores.

3.1.2 Paso 2. Especificación primaria de la forma funcional del modelo.

Una vez identificadas y clasificadas variables candidatas, es momento de seleccionar aquellas que vayan a formar parte del modelo, al menos en una primera instancia. Esta decisión gira en torno a las preguntas de investigación a las que se desee encontrarles respuesta.

Se presentan a continuación, tres ejemplos de modelos de regresión, propuestos con base en el tipo de información que la Facultad de Ingeniería posee. Las ecuaciones estimadas de estos modelos servirán para responder preguntas relacionadas con la naturaleza del rendimiento académico de sus estudiantes.

i) **¿Qué factores tienen mayor influencia en el rendimiento académico de los estudiantes?**

$$\begin{aligned} \ln(\text{promindiv}) &= \beta_0 + \beta_1 \text{prombach} + \beta_2 \text{examingr} + \beta_3 \text{nivelmate} + \beta_4 \text{motiv} + \beta_5 \text{nivautorr} \\ &+ \beta_6 \text{hogar}_1 + \beta_7 \text{hogar}_2 + \beta_8 \text{ingreso}_1 + \beta_9 \text{ingreso}_2 + \beta_{10} \text{traslado}_1 \\ &+ \beta_{11} \text{traslado}_2 + \beta_{12} \text{trabaja} + \beta_{13} \text{calense} + \beta_{14} \text{exper} + \mu \end{aligned} \quad (3.2)$$

Esta ecuación integra elementos propios del **estudiante**, del **contexto** y de la **institución**. Las estimaciones de los parámetros β_1 , β_2 , y β_{14} indicarán la *variación porcentual que resulta en la variable promindiv* (promedio individual) *al aumentar en una unidad las variables prombach, examingr y exper*, respectivamente.

La interpretación de las estimaciones de los demás parámetros será un poco distinta, dado que se trata de variables de tipo ordinal o binario. En el caso de la variable *nivelmate*, la estimación de su parámetro asociado β_3 indicará la *diferencia porcentual que se predice en el promedio individual, entre estudiantes con buena formación previa en matemáticas y aquellos con una formación previa deficiente, dados los mismos niveles en rendimiento previo (prombach, examingr), motivación (motiv), tipo de hogar (hogar₁, hogar₂), etc.*

Para las estimaciones restantes, su interpretación será muy similar a esta última. Por ejemplo, para el caso de *motiv*, la estimación de β_4 indicará la *diferencia porcentual que se predice para promindiv entre estudiantes con una motivación alta y aquellos con una motivación baja o media*, manteniendo constantes los demás elementos controlados (propios del estudiante, del contexto y de la institución).

La magnitud de las diferencias estimadas (es decir, de los $\hat{\beta}_j$) proveerán de información a las autoridades académicas para determinar qué factores resultan tener una mayor influencia en el rendimiento académico de los estudiantes.

ii) **¿Cuál es el efecto del programa de asesorías académicas en el rendimiento académico de los estudiantes?**

$$\begin{aligned} \ln(\text{promindiv}) &= \beta_0 + \beta_1 \text{asesoaca} + \beta_2 \text{prombach} + \beta_3 \text{examingr} + \beta_4 \text{nivelmate} \\ &+ \beta_5 \text{motiv} + \beta_6 \text{nivautorr} + \beta_7 \text{hogar}_1 + \beta_8 \text{hogar}_2 + \beta_9 \text{ingreso}_1 \\ &+ \beta_{10} \text{ingreso}_2 + \beta_{11} \text{traslado}_1 + \beta_{12} \text{traslado}_2 + \beta_{13} \text{trabaja} + \beta_{14} \text{calense} \\ &+ \beta_{15} \text{exper} + \mu \end{aligned} \quad (3.3)$$

Para evaluar el efecto de un programa de apoyo, en este caso el de las asesorías académicas, se integra la variable *asesoaca* en el modelo a estimar en función de evaluar la diferencia porcentual existente en el promedio individual (*promindiv*).

Para este modelo, la estimación del parámetro β_1 multiplicado por cien ($100\beta_1$) indicará la *diferencia porcentual que se predice del promedio individual entre estudiantes que sí asistieron a asesorías respecto a aquellos que no*, controlando otros factores tales como el rendimiento previo (*prombach*, *examingr*), motivación (*motiv*), calidad de la enseñanza en las aulas (*calense*), etc.

iii) **¿Cuál es la probabilidad de que un estudiante que no asistió a sesiones de tutoría repruebe al menos una materia de la DCB?**

$$\begin{aligned} \text{reprobo} &= \beta_0 + \beta_1 \text{tutoria} + \beta_2 \text{prombach} + \beta_3 \text{examingr} + \beta_4 \text{nivelmate} + \beta_5 \text{motiv} \\ &+ \beta_6 \text{nivautorr} + \beta_7 \text{hogar}_1 + \beta_8 \text{hogar}_2 + \beta_9 \text{ingreso}_1 + \beta_{10} \text{ingreso}_2 \\ &+ \beta_{11} \text{traslado}_1 + \beta_{12} \text{traslado}_2 + \beta_{13} \text{trabaja} + \beta_{14} \text{calense} + \beta_{15} \text{exper} + \mu \end{aligned} \quad (3.4)$$

Al igual que en los casos anteriores, se integran al modelo variables relacionadas al estudiante, al contexto y a la institución para controlar otros factores que afectan el que los estudiantes reprueben alguna materia.

En este caso, la variable dependiente es una variable binaria igual a uno si el estudiante reprobó al menos una materia de la DCB e igual a cero si no fue así (ver figura 3.5). Como se vio en el Capítulo II, este es un **modelo de probabilidad lineal**.

El interés en este caso radica en conocer el impacto que las sesiones de tutoría tienen en la probabilidad de reprobado al menos una materia. Esta información puede ser muy útil para evaluar la pertinencia de este tipo de programas de apoyo.

La estimación del parámetro β_1 indicará la *probabilidad de que un estudiante que no asistió a sesiones de tutoría repruebe al menos una materia de la DCB respecto a alguien que sí haya asistido a tutorías*.

Cabe destacar que en esta ocasión, el grupo de referencia corresponde a estudiantes que sí asistieron a sesiones de tutoría (*tutoria = 1*).

En los tres casos anteriores, se especificaron modelos con variables escogidas de las tablas 3.1 a la 3.5, para controlar los factores ampliamente documentados en el capítulo I. Sin embargo, esto no limita de manera alguna el número de variables que el usuario de la metodología pueda seleccionar para agregar a los modelos. Así mismo, la forma funcional especificada ha sido seleccionada en función de demostrar la utilidad y el alcance de modelos de regresión lineal en contextos educativos. Los tres casos no son más que ejemplos de aplicación.

Independientemente de la forma funcional y el número de variables explicativas seleccionadas, los pasos siguientes validarán (para el caso de variables de tipo cuantificable) si la especificación propuesta en este paso cumple con las condiciones y supuestos del MCO.

Al respecto, cabe destacar que los pasos 4 al 8 de esta metodología no se pueden aplicar en variables de tipo binario u ordinal. Esto limita en gran medida el análisis de los datos de entrada recolectados. En consecuencia, se sugiere que en la medida de lo posible y de la información que se tenga disponible, en la especificación

primaria de los modelos se utilicen variables de tipo cuantificable para controlar factores asociados al rendimiento académico.

3.1.3 Paso 3. Determinación del tamaño de muestra.

Una vez identificadas las variables de las cuales se debe recopilar información, es necesario conocer la cantidad de datos u observaciones que son necesarias; es decir, se debe decidir el tamaño de la muestra a analizar.

La determinación del tamaño de una muestra se basa en la probabilidad de incurrir en el error tipo II (β) al realizar una prueba de hipótesis, en el nivel de significancia α , así como en la sensibilidad deseada para detectar un valor de la media diferente al establecido en la hipótesis nula.

Del capítulo II se conoce que el nivel de significancia α depende completamente del juicio del analista, mientras que la probabilidad del error β depende tanto del verdadero valor del parámetro como del tamaño de la muestra. Dado que el error β y el tamaño de la muestra son dependientes entre sí, es posible también determinar un valor β de inicio, para conocer el tamaño de muestra necesario.

Para ello, el analista o investigador se debe guiar por la siguiente pregunta: **¿qué probabilidad de incurrir en un error tipo II al estimar el verdadero valor del parámetro de interés β_j , es aceptable?** En otras palabras, ¿qué probabilidad de “aceptar” la hipótesis nula $H_0: \beta_j = 0$ cuando en realidad ésta es falsa, será aceptable para el desarrollo de la investigación?

Para determinar el tamaño de muestra requerido, además de especificar las probabilidades α y β de cometer un error, se deberá especificar también “el tamaño relativo de la diferencia en las medias d que quiera detectarse” (Montgomery y Runger, 2013) en caso que la media real de β_j sea diferente de cero.

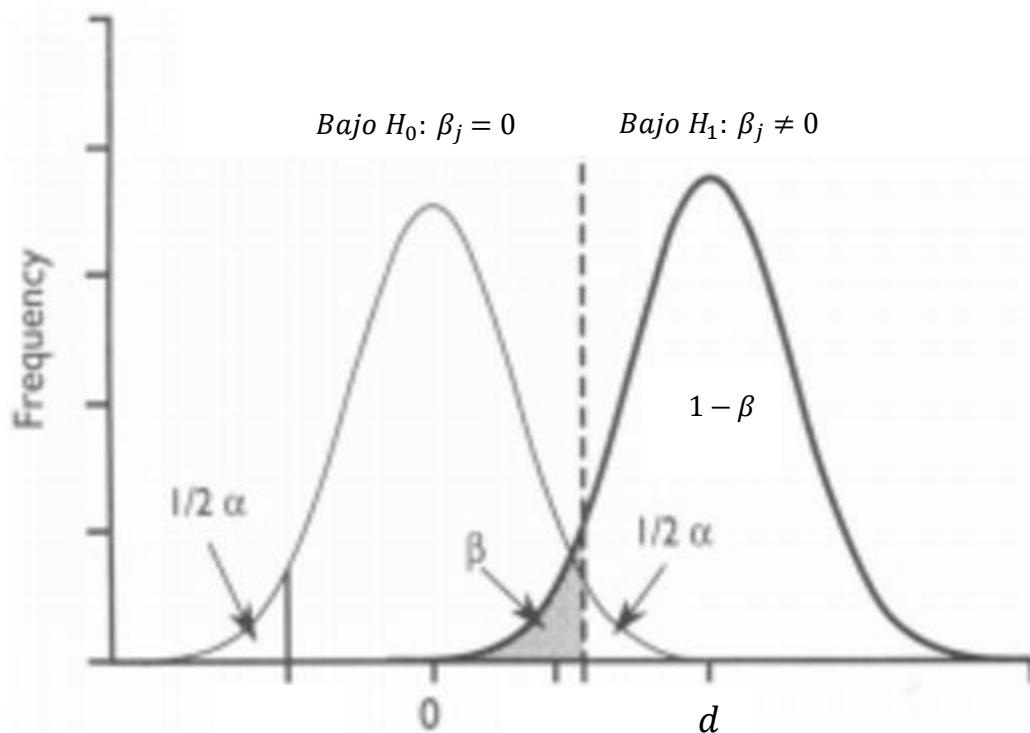


Figura 3.1 Superposición de dos distribuciones, la de la izquierda bajo H_0 y la de la derecha bajo H_1 . Se supone como verdadera la distribución bajo H_1 . Modificado de Florey, 1993.

El área sombreada en la figura 3.1 representa el porcentaje de muestras que llevarán a incurrir en un error tipo II; esto es, cualquier muestra cuyo valor promedio se encuentre en esta área no permitirá rechazar la hipótesis nula, aunque ésta sea falsa en realidad.

En esta figura, también se evidencia la importancia de definir una diferencia en las medias d que resulte de interés. Para el caso de cuantificar el impacto de programas de apoyo, la hipótesis nula establece que *no hay diferencia entre estudiantes que (por ejemplo) acuden a asesorías académicas y aquellos que no*. Al determinar el tamaño de muestra, se debe analizar la sensibilidad deseada, en función de que la prueba sea capaz de identificar diferencias *pequeñas* (p.e. $d \leq 1$) o solamente aquellas que sean moderadamente *grandes* (p.e. $d = 2$).

La distancia d se encuentra definida por la diferencia entre el valor del parámetro bajo análisis (que en este caso es cero) y otro valor propuesto arbitrariamente, en relación con la desviación estándar. Es decir:

$$d = \frac{|0 - \delta|}{\sigma} = \frac{|\delta|}{\sigma} \quad (3.5)$$

Dado que se desconoce la varianza poblacional de β_j es imposible determinar σ . Por ende, el valor d no se puede *calcular*, sino solamente *especificar*.

En la figura 3.1 se aprecia también la **potencia** de la muestra, definida como $1 - \beta$, para detectar una *verdadera diferencia* en cuanto a estudiantes que (por ejemplo) sí acuden a asesorías y aquellos que no.

La interpretación de los valores de α , β y d en la elección del tamaño de muestra es la siguiente: *la probabilidad de rechazar la hipótesis nula, siendo ésta verdadera, es del $100 * \alpha$ por ciento, mientras que la probabilidad de aceptar la hipótesis nula, siendo ésta falsa, es del $100 * \beta$ por ciento dada una media poblacional d diferente de cero.*

Una vez establecidos los parámetros α , β y d , es posible determinar el tamaño de muestra necesario. Para ello, se debe hacer uso de **curvas de operación** que grafiquen la probabilidad β para la prueba t contra un parámetro d para diferentes tamaños de muestra n .¹²

En el Apéndice A de Montgomery y Runger (2013) se pueden encontrar las curvas características de operación para diferentes valores de n , tanto para la prueba t como para la prueba normal estándar de una y dos colas.

¹² En caso de que se sepa de antemano que el tamaño de muestra tiene que ser relativamente grande ($n > 30$), es posible hacer uso de curvas de operación basadas en la distribución normal estándar.

Estas curvas se construyeron, al graficar la probabilidad de aceptar H_0 (cuando ésta es falsa) contra diferentes valores de d . Con el uso de estas curvas de operación, una vez definido el diseño de la muestra, es posible obtener el tamaño de la muestra n requerido.

La n obtenida será el tamaño de muestra **mínimo necesario** para que la potencia de la prueba sea $1 - \beta$.

A pesar de que lo ideal es obtener un tamaño de muestra n con base en un diseño previo, es muy probable que existan ocasiones en las que, debido a restricciones económicas o de otra índole, no sea posible utilizar el tamaño de muestra n calculado. En casos como este, se deberá utilizar el tamaño de muestra n más grande posible, y presentar la probabilidad β y la potencia de la prueba $1 - \beta$ asociada a esa n .

3.1.4 Paso 4. Identificación de posibles observaciones influyentes.

Antes de continuar con la construcción del modelo, se debe analizar la información que ha sido recolectada en función de identificar observaciones *poco usuales* respecto al resto de los datos, que puedan influir de manera importante tanto en la estimación como en la inferencia estadística a realizar.

Existen dos tipos de observaciones inusuales que deben ser identificadas, estas son: las observaciones *palanca* y las observaciones *influyentes*.

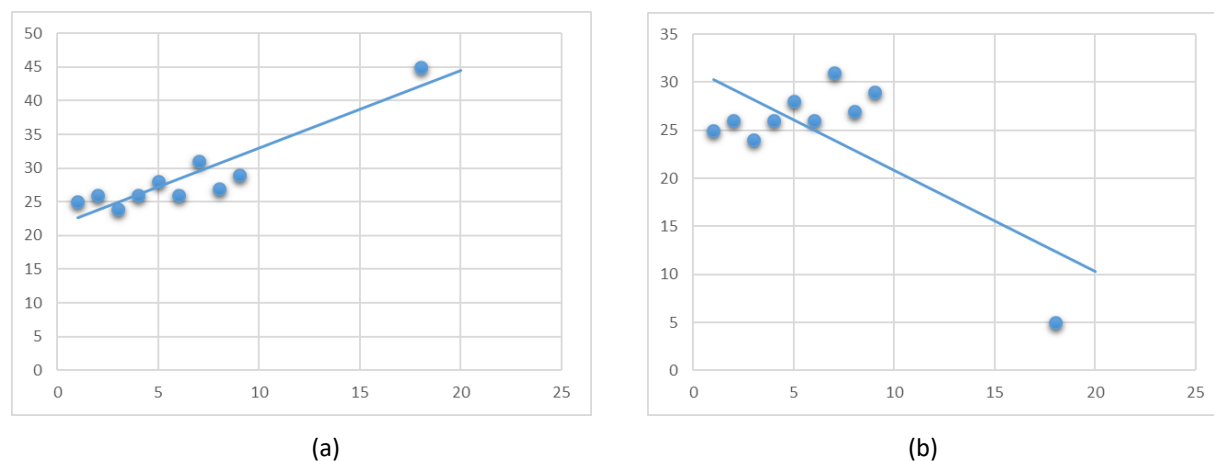


Figura 3.2 Tipos de observaciones: (a) Observación Palanca, (b) Observación Influyente. Elaboración propia con base en Montgomery *et al.*, 2012.

En la figura anterior se ejemplifica la diferencia entre este tipo de observaciones. La de la izquierda corresponde a una **observación palanca**. Esta observación se encuentra remota respecto a los valores de x , sin embargo su respuesta en la variable dependiente y es consistente con la predicción basada en las otras observaciones (la observación palanca se encuentra sobre la recta de regresión). Por lo general este tipo de observaciones tienen una importante influencia en el cálculo del coeficiente de determinación R^2 así como para la obtención de los errores estándar de $\hat{\beta}_j$; sin embargo, una observación palanca no tiene ningún efecto en la estimación de los coeficientes de regresión (Montgomery *et al.*, 2012).

Por otra parte, de lado derecho de la figura 3.2 se encuentra el ejemplo de una **observación influyente**. Este tipo de observación, además de encontrarse en un lugar remoto respecto a los valores de x de las demás observaciones, también se encuentra en un lugar que difiere del resto de las observaciones en cuanto a su respuesta en y .

Es más peligroso trabajar con una observación influyente que con una de tipo palanca debido a que la primera *atrae* hacia su ubicación a la recta de regresión, como se puede apreciar en la figura 3.2 (b), influyendo de manera importante la estimación de los coeficientes. Una observación influyente además, propicia una pobre capacidad de predicción del modelo.

La importancia de identificar (y eliminar en dado caso) observaciones influyentes radica en la capacidad que se desea tenga el modelo de regresión, de representar el comportamiento observado en *la mayoría* de las y los estudiantes; y en consecuencia, que las conclusiones obtenidas no sean significativamente diferentes al trabajar con cualquier otra muestra.

Al respecto, en este apartado se presentarán las condiciones bajo las cuales es posible eliminar observaciones influyentes de la muestra; sin embargo, se advierte que, en la mayoría de los casos, este tipo de observaciones pueden ser síntoma de cuestiones inobservadas y no incluidas en el modelo, o inclusive, indicios de información desconocida hasta el momento (Montgomery *et al.*, 2012). Por lo que, sean eliminadas o no, las observaciones inusuales deben identificarse y controlarse, en función de evaluar su impacto en la estimación del modelo.

Si entendemos como observación inusual, aquella que se encuentra en una región alejada del resto de los datos, ya sea respecto a los valores de alguna variable explicativa x_j o a la variable dependiente y , entonces lo más conveniente para identificar este tipo de observaciones es evaluar la **distancia** que existe entre cada observación, con el resto de las observaciones.

Una manera de hacer esto es a través de la **matriz H**, conocida también como la matriz “gorro” (Montgomery y Runger, 2013), la cual se calcula con operaciones matriciales (ver Anexo C). Esta matriz se define como:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.6)$$

La importancia de la matriz **H** radica en que ésta determina las varianzas y covarianzas de los valores ajustados, así como de los residuales. Así mismo, “[...]los elementos h_{ij} de la matriz **H** pueden ser interpretados como la **cantidad de apalancamiento** ejercida por la i -ésima observación y_i en el j -ésimo valor ajustado \hat{y}_j ” (Montgomery *et al.*, 2012).

Los elementos de la diagonal de la matriz \mathbf{H} , es decir:

$$h_{ii} = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \quad (3.7)$$

Donde \mathbf{x}'_i corresponde al i -ésimo renglón de la matriz \mathbf{X} , son los que requieren de mayor atención dado que estos corresponden a la **medida de la distancia estandarizada de la i -ésima observación con respecto al centro (o centroide) del espacio en x** . Montgomery y Runger (2013) señalan también que además de σ^2 , h_{ii} es la varianza del valor ajustado \hat{y}_i .

Por consiguiente, valores grandes de h_{ii} indican la existencia de observaciones inusuales al encontrarse lejos (con respecto a los valores en x_j) del resto. Es posible que estas observaciones inusuales sean además observaciones influyentes.

Montgomery *et al.* (2012) establecen como regla de oro que para cualquier observación donde el elemento de la diagonal $h_{ii} > 2\left(\frac{k-1}{n}\right)$, la observación en cuestión se encuentra lo suficientemente lejos de los demás datos para considerarse un punto u **observación palanca**.¹³

Sin embargo, es importante hacer notar que un valor alto de h_{ii} no necesariamente involucra una observación influyente. Dado que los elementos de la diagonal de la matriz \mathbf{H} sólo evalúan la distancia respecto a x , para encontrar evidencia de una observación influyente es necesario además un análisis que involucre a las observaciones Y_i y sus valores ajustados \hat{y}_i . El análisis del espacio en y conlleva al análisis de residuales.

La función en el lenguaje de programación estadístico R que permite obtener los valores h_{ii} de manera automática es `hatvalues(model, ...)` (John Fox *et al.*).

¹³ La regla de oro se deriva del hecho de que el tamaño promedio de la diagonal de la matriz "gorro" es: $\bar{h} = \frac{(k-1)}{n}$, por lo que un valor superior al doble del promedio es considerado un punto palanca. Para el caso en que la relación $2p/n$ sea mayor a 1, la regla de oro pierde su validez (Montgomery *et al.*, 2012).

El **análisis de residuales** se define como aquel en el que se presta atención a la distancia existente entre una observación real Y_i y el valor ajustado \hat{y}_i obtenido por la recta de regresión de MCO:

$$\hat{\mu}_i = Y_i - \hat{y}_i \quad (3.8)$$

Para la evaluación de datos extremos o influyentes se utilizan *residuales escalados*, los cuales permiten conocer en qué medida las observaciones se encuentran separadas del resto de los datos.

Uno de los residuales escalados más útiles es el **residual studentizado**, definido como:

$$r_i = \frac{\hat{\mu}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \quad i = 1, 2, \dots, n \quad (3.9)$$

En donde la estimación de la varianza $\hat{\sigma}^2$ es la que se presenta en el Anexo B.

Dado que la estimación $\hat{\sigma}^2$ se ha obtenido de una regresión que considera a todo el conjunto de observaciones n , el residual studentizado es conocido también como un tipo de **escalado interno** (Montgomery *et al.*, 2012). De tal manera, se vuelve interesante conocer cuál sería la magnitud del residual studentizado *sin* la i -ésima observación; esto es:

$$t_i = \frac{\hat{\mu}_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}} \quad i = 1, 2, \dots, n \quad (3.10)$$

En donde $\hat{\sigma}^2_{(i)}$ representa a la estimación de la varianza **sin** la i -ésima observación. Montgomery *et al.* (2012) demuestran que su valor se encuentra definido como:¹⁴

$$\hat{\sigma}^2_{(i)} = \frac{(n - k - 1)\hat{\sigma}^2 - \hat{\mu}_i^2 / (1 - h_{ii})}{n - k - 2} \quad (3.11)$$

Al estadístico t_i obtenido en (3.10) se le conoce como **R-student**, al cual se le considera también como un **residual studentizado externo** (Montgomery *et al.*, 2012).

El residual studentizado se obtiene fácilmente en el lenguaje de programación R con la función `rstudent(model, ...)`. Para el caso del residual studentizado externo, antes de aplicar esta función se deberá de obtener $\hat{\sigma}^2_{(i)}$ y posteriormente especificarla como la estimación de la varianza a utilizar en los argumentos de la función.

El análisis del residual studentizado tanto interno como externo, **en conjunto** con la evaluación de la diagonal h_{ii} permitirá identificar observaciones con alta probabilidad de ser influyentes.

3.1.5 Paso 5. Validación de la forma funcional.

El siguiente paso en el análisis de datos de entrada es encontrar evidencia que justifique la existencia de los supuestos utilizados para la construcción del modelo, especialmente aquellos que hacen referencia a la relación lineal entre y y x_j , así como a la constancia de la varianza (ver Anexo A).

Además, después de haber identificado posibles observaciones influyentes, cabe preguntarse si éstas son consecuencia de un problema mayor, como sería el caso de

¹⁴ Para la demostración matemática consultar a Montgomery, Peck y Vining, *Introduction to Lineal Regression Analysis*, 2012, Anexo C.8.

un supuesto que en realidad no se cumple. Esta posibilidad debe evaluarse antes de condenar cualquier observación como *influyente* (Montgomery *et al.*, 2012). En consecuencia, los pasos 5, 6 y 7 de esta metodología se enfocarán en la validación de los supuestos más importantes del modelo lineal clásico (MLC).

El primer supuesto a evaluar será el establecido en el paso 2; es decir, la *forma funcional* especificada para el modelo, en otras palabras, la relación *real* existente entre la variable dependiente y y las variables explicativas x_j . Para ello, en primera instancia se graficarán los residuales studentizados t_i contra los valores ajustados \hat{y}_i ¹⁵; esto permitirá encontrar evidencia de variables que necesiten una transformación (p.e. pasar de una forma *Nivel – ln* a una forma *ln – ln*). Posteriormente, se hará uso del estadístico t para comprobar la especificación inadecuada de la forma funcional, y en dado caso, hacer las modificaciones pertinentes al modelo.

En la siguiente figura se presentan dos ejemplos de patrones al graficar los residuales studentizados t_i contra los valores ajustados \hat{y}_i . La gráfica de lado izquierdo corresponde a una especificación adecuada de la forma funcional, mientras que la de lado derecho corresponde a una especificación inadecuada.

¹⁵ Al respecto cabe preguntarse: ¿por qué no utilizar al residual studentizado r_i ? La respuesta yace en el tipo de escalado de ambos residuales. Mientras que el cálculo de t_i se basa en la *omisión* de la i -ésima observación para estimar la varianza, el cálculo de r_i contempla la *inclusión* de todas las observaciones para la estimación.

Si se desea evaluar la posibilidad de que observaciones influyentes sean consecuencia de supuestos del MLC que no se cumplen, es necesario diferenciar cada observación del resto de los datos, para lo cual de entre los dos residuales studentizados el de escalado externo es el que cumple con esta condición.

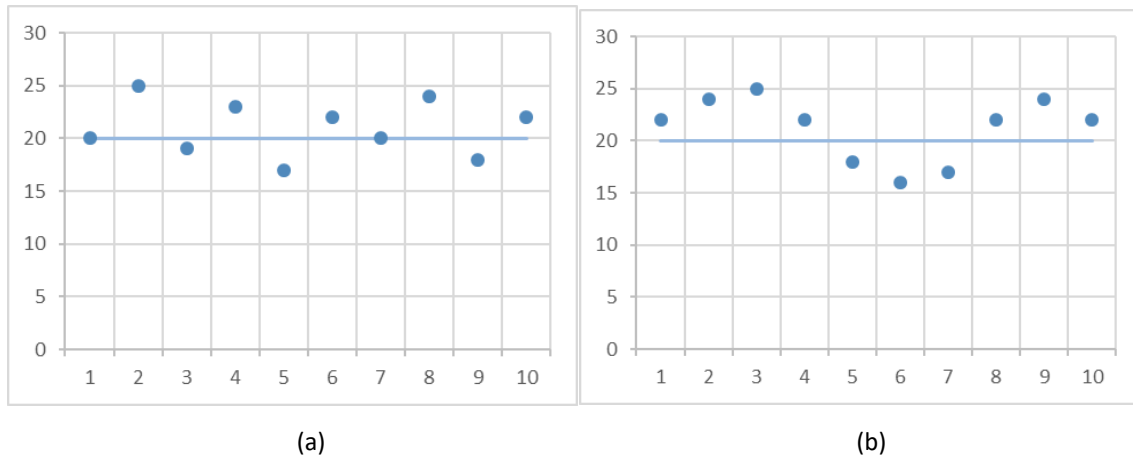


Figura 3.3 Patrones de gráficos de residuales: (a) Especificación de la forma funcional adecuada, (b) Especificación de la forma funcional inadecuada. Elaboración propia con base en Montgomery *et al.*, 2012.

Si el patrón encontrado en el análisis de residuales no corresponde a uno similar al de la figura 3.3 (a), entonces se deberá proceder al uso del estadístico t para comprobar la *no-linealidad* y evaluar la forma de la relación más apropiada.

El uso del estadístico t como prueba para la especificación incorrecta de la forma funcional se basa en la inclusión del valor ajustado \hat{y} , obtenido del modelo establecido en el *Paso 2*, como una variable explicativa más en un modelo como el siguiente:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta \hat{y}^2 + \mu \quad (3.12)$$

En donde \hat{y}^2 es una función particular de todos los cuadrados y productos cruzados de las x_j (Wooldridge, 2015). Al establecer la hipótesis nula $H_0: \delta = 0$, si el estadístico t resulta significativo (contra una alternativa de dos colas) entonces se comprueba que la forma funcional **es incorrecta** y se deberá considerar incluir al modelo términos cuadráticos y/o productos cruzados.

Cabe considerar en este punto, además de términos cuadráticos y productos cruzados, la posibilidad de tener que utilizar términos logarítmicos para explicar de mejor manera alguna relación entre la variable dependiente y las explicativas. Si bien la gráfica de los residuales studentizados t_i contra los valores ajustados \hat{y}_i proporcionará evidencia de esto, también se utilizará el estadístico t como prueba final para comprobar la necesidad o no, de incluir términos de este tipo.

Para lo anterior, el valor ajustado obtenido del modelo establecido en el *Paso 2*, se incluirá como variable explicativa de un nuevo modelo con términos logarítmicos; es decir:

$$Y = \beta_0 + \beta_1 \ln(x_1) + \dots + \beta_k \ln(x_k) + \theta \hat{y} + \mu \quad (3.13)$$

Se establecerá como hipótesis nula $H_0: \theta = 0$ y, en dado caso que el estadístico t resulte significativo (contra una alternativa de dos colas), se concluirá que el modelo requiere de la inclusión de términos logarítmicos para una adecuada especificación.

Tanto en la ecuación (3.12) como en la ecuación (3.13) se consideran las k variables explicativas involucradas en el modelo; sin embargo, esto no es una regla para poder validar la forma funcional. Por el contrario, al establecer las ecuaciones (3.12) y (3.13) se debe de analizar la naturaleza de las variables x_j e identificar aquellas para las cuales **una función no lineal tenga sentido práctico**. Tal es el caso del *rendimiento de los años de estudios*, en donde se espera que los años estudiados en una universidad tengan mayor valor que los años estudiados en cualquier bachillerato.

Finalmente, cabe destacar que este paso, establecido para validar la forma funcional del modelo, **sólo se ocupa de variables de tipo cuantificable**. Variables de tipo *ordinal* o *binario* no pueden ser sujetas a transformaciones, dado que éstas representan aspectos cualitativos. Si se transformara alguna variable binaria u

ordinal digamos, a un término cuadrático o logarítmico, el coeficiente $\hat{\beta}$ asociado a dicha variable carecería de sentido alguno.

3.1.6 Paso 6. Validación del supuesto de homocedasticidad.

El supuesto de homocedasticidad establece que la varianza del error μ , independientemente de los valores de las x_j , siempre es la misma; es decir, es constante (ver Anexo A).

Dada la importancia de este supuesto, se debe evaluar, con base a los datos recopilados, su cumplimiento. Para ello, en primera instancia se realizará un diagnóstico visual a través de gráficas que comparen al residual studentizado externo t_i contra cada variable explicativa x_j .

Una dispersión de puntos similar a la que se presenta en la siguiente figura es indicativo de *heterocedasticidad* existente en la variable x_j :

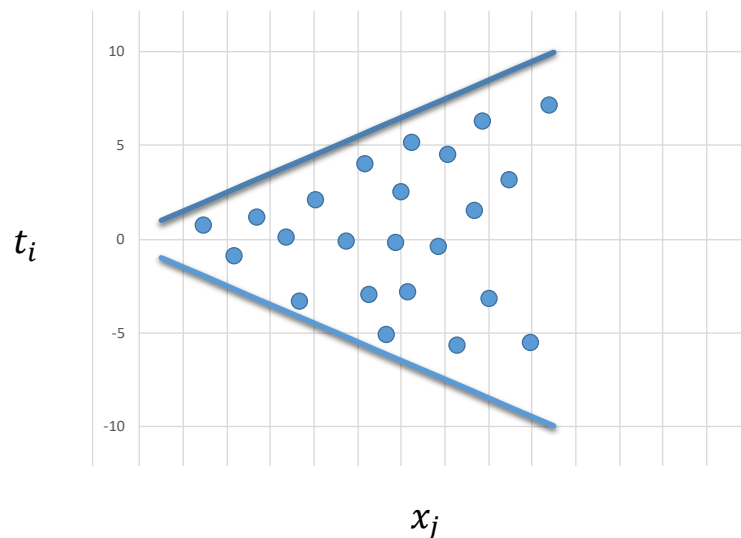


Figura 3.4 Gráfica de t_i contra x_j indicativa de heterocedasticidad existente. Elaboración propia con base en Montgomery *et al.*, 2012.

Si la dispersión de los datos no corresponde a una como la de la figura 3.3 (a), y más aún, se asemeja a una como la de la figura 3.4 (en otras palabras, si las observaciones presentan cierta tendencia) entonces es muy probable que la variable explicativa x_j en cuestión, no cumpla con el supuesto de homocedasticidad.

Al igual que con el supuesto de la forma funcional, se utilizarán pruebas estadísticas para comprobar la presencia de heterocedasticidad. En este caso, se hará uso de la *Prueba Breusch-Pagan*, así como de la *Prueba de White para heterocedasticidad* (Wooldridge, 2015). Los pasos para realizar cada prueba son los mismos que se presentan en el capítulo II.¹⁶

Para la prueba Breusch-Pagan, se establece la hipótesis nula $H_0: \delta_1 = \delta_2 = \dots = \delta_k = 0$ y se realizan los siguientes pasos:

- 1.- Se estima el modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \mu$ por MCO. Obtener los residuales cuadrados de MCO, $\hat{\mu}^2$ (uno para cada observación).
- 2.- Ejecutar la estimación del modelo (2.33) utilizando los residuales cuadrados del modelo original, es decir:

$$\hat{\mu}_i^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k + error$$

conservar la R-cuadrada de esta regresión, $R^2_{\hat{\mu}^2}$.

- 3.- Formar el estadístico F y calcular el valor-p (usando la distribución $F_{k, n-k-1}$). Si el valor-p es suficientemente pequeño, es decir, menor que el nivel de significancia elegido, se rechaza la hipótesis nula de homocedasticidad.

$$F = \frac{R^2_{\hat{\mu}^2}/k}{(1 - R^2_{\hat{\mu}^2})/(n - k - 1)}$$

¹⁶ Ver apartado *Heterocedasticidad* del Capítulo II.

Por otra parte, para la Prueba de White se establece la hipótesis nula $H_0: \delta_1 = \delta_2 = 0$ y se realizan los siguientes pasos:

1.- Estimar el modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \mu$ por MCO. Obtener los residuales $\hat{\mu}$ y los valores ajustados \hat{y} y calcular sus respectivos cuadrados, $\hat{\mu}^2$ y \hat{y}^2 .

2.- Ejecutar la regresión de la ecuación:

$$\hat{\mu}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error$$

conservar su R-cuadrada, $R^2_{\hat{\mu}^2}$.

3.- Formar el estadístico F y calcular el valor-p, empleando la distribución $F_{2,n-3}$.

Para el caso en que se compruebe la existencia de heterocedasticidad (y por tanto, el incumplimiento del supuesto de homocedasticidad) se procederá a utilizar estadísticos **robustos a la heterocedasticidad**.

El cálculo de la varianza de $\hat{\beta}_j$ se encontrará determinado por:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \hat{r}_{ij}^2 \hat{\mu}_i^2}{SRC_j} \quad (3.14)$$

Donde \hat{r}_{ij} denota el i -ésimo residual de regresar x_j sobre el resto de las variables independientes, $\hat{\mu}_i^2$ el i -ésimo residual cuadrado de la regresión inicial de Y sobre x_1, x_2, \dots, x_k y SRC_j es la suma de residuales cuadrados de esta regresión (Wooldridge, 2015).

La raíz cuadrada de $\widehat{Var}(\hat{\beta}_j)$; es decir el error estándar de $\hat{\beta}_j$, será utilizado para calcular el **estadístico t robusto a la heterocedasticidad**.

Por otra parte, se deberá calcular también un **estadístico F robusto a la heterocedasticidad**, mejor conocido como estadístico de Wald. Este estadístico es fácilmente calculado con el uso del lenguaje de programación R a través de la función `wald.test()`.

3.1.7 Paso 7. Validación del supuesto de normalidad.

Como se explica en el Anexo A, concerniente a los supuestos del modelo de regresión, el supuesto de normalidad es el más *fuerte* de todos dado que éste incluye de manera implícita a los demás. Al respecto, el supuesto de media condicional cero (supuesto 4 del modelo lineal clásico) se verifica de manera simultánea al evaluar la existencia de normalidad.

El gráfico que permite realizar esta validación es la del residual studentizado externo contra la probabilidad acumulada $P_i = \frac{(i-1)}{n}$, $i = 1, 2, \dots, n$. Un ejemplo de éste aparece en la siguiente figura:

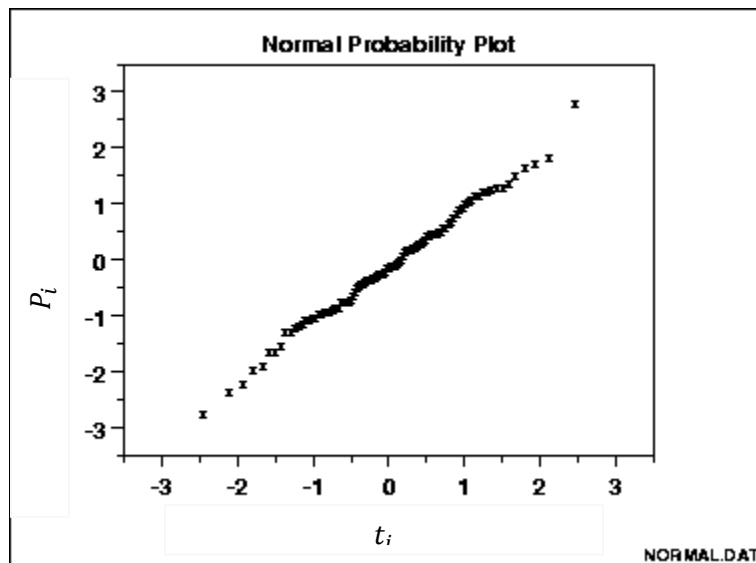


Figura 3.5 Gráfica de probabilidad normal de los residuales studentizados externos. Tomado de Wikimedia Commons.

En la figura 3.5 se aprecia una **gráfica de probabilidad normal de los residuales studentizados**. En ella, se grafican los t_i ordenados de menor a mayor contra el P_i de la observación i -ésima asociada.

Si los puntos de la gráfica se ajustan a una línea recta (como es el caso en la figura 3.5) entonces existe evidencia que justifica el cumplimiento del supuesto de normalidad. Cabe destacar que la pendiente de la gráfica de probabilidad normal es la medida de la desviación estándar mientras que el intercepto indica el valor esperado o promedio de los datos (Belsley, Kuh y Welsh, 1980).

Dado que al llegar a este paso ya se han validado dos de los supuestos más importantes (el supuesto de la forma funcional y el de homocedasticidad), así como también se ha recolectado una muestra aleatoria representativa (con lo que se cumple el supuesto de muestreo aleatorio), para la validación del supuesto de normalidad y, en consecuencia, la validación del supuesto de media condicional cero, es suficiente una evaluación visual de la gráfica de probabilidad normal.

3.1.8 Paso 8. Comprobación de observaciones influyentes.

Al término de los pasos 5, 6 y 7, el análisis derivará en dos conclusiones posibles:

1) Todos los supuestos se cumplen, por tanto, las observaciones identificadas en el paso 4 con un alto valor de h_{ii} (mayor a $2 \left(\frac{k-1}{n}\right)$) y también con un alto valor de los residuales studentizados t_i y r_i (mayor a 2)¹⁷, son consideradas como *influyentes*.

2) Alguno o varios de los supuestos del MLC no se cumplen, por tanto, se debe corroborar que las observaciones identificadas en el paso 4 sean efectivamente influyentes, y no observaciones inusuales consecuencia del incumplimiento de los supuestos.

¹⁷ Ambos residuales siguen el comportamiento de la distribución t , la cual mide el número de desviaciones estándar que la observación i se encuentra respecto de cero. Una observación que se encuentre más allá de dos desviaciones estándar se considera influyente ya que excede el valor crítico para un nivel de significancia del 5%.

En este paso se utilizarán dos medidas de influencia propuestas por Belsley, Kuh y Welsh (1980). La primera de ellas, denominada *DFBETAS* proporcionará una medida del efecto de observaciones consideradas como influyentes, en la magnitud de los coeficientes estimados $\hat{\beta}$. Es decir, para el caso en que todos los supuestos del MLC se cumplan y se hayan identificado observaciones influyentes, el estadístico *DFBETAS* proporcionará una **medida del impacto de dichas observaciones** en términos de la desviación estándar, en cuanto a los valores estimados de los coeficientes.

Por otra parte, para el caso en que alguno o varios de los supuestos no se cumplan, *DFBETAS* servirá como una **medida de comprobación** para determinar si las observaciones identificadas en el paso 4 son influyentes. Así mismo, se evaluará el impacto de dichas observaciones.

En conjunto con *DFBETAS* se utilizará otro estadístico, denominado *DFFITs*, el cual proporcionará también una medida del efecto de observaciones influyentes, pero ahora en función de los valores ajustados \hat{y} .

La importancia de este paso 8 radica en que, independientemente del cumplimiento o no de los supuestos del MLC, con el uso de estos estadísticos **se conocerá el efecto** en la estimación del modelo de regresión, **de las observaciones influyentes identificadas**. Así mismo, proveerá de evidencia complementaria para determinar si una observación es influyente o no, para el caso en que resultados de los pasos anteriores no hayan sido concluyentes.

El estadístico *DFBETA* se define como:

$$DFBETAS_{i,j} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)} C_{jj}}} \quad i = 1, 2, \dots, n \quad (3.15)$$

En donde C_{jj} es el j -ésimo elemento de la diagonal de $(X^T X)^{-1}$, $\hat{\sigma}^2_{(i)}$ es la estimación de la varianza sin la i -ésima observación y $\hat{\beta}_{j(i)}$ denota al j -ésimo coeficiente de regresión estimado sin la i -ésima observación.

Un valor *grande* de $DFBETAS_{i,j}$ indica que la observación i tiene una influencia considerable en el j -ésimo coeficiente de regresión (Montgomery *et al.*, 2012). Así mismo, una magnitud *grande* de este estadístico mostraría si las conclusiones obtenidas de la prueba de hipótesis para evaluar la significancia estadística, podrían verse afectadas (Belsley *et al.*, 1980).

La regla de oro de Belsley *et al.* (1980) establece que para $|DFBETAS_{i,j}| > 2/\sqrt{n}$, la observación i **se considera influyente** en la estimación del coeficiente β_j . En tanto, el valor absoluto del estadístico proporcionará información en cuanto al **efecto de dicha observación** en la magnitud del coeficiente estimado. Al respecto, un valor absoluto de $DFBETAS_{i,j}$ mayor a 2 se considera como un *efecto grande*.

El otro estadístico que permitirá analizar la sensibilidad de la ecuación de regresión a la presencia de observaciones influyentes es:

$$DFFITs_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} h_{ii}}} \quad i = 1, 2, \dots, n \quad (3.16)$$

En donde $\hat{y}_{(i)}$ representa al valor ajustado de y_i obtenido **sin** la observación i .

Este estadístico se define como el número de desviaciones estándar que el valor ajustado \hat{y}_i cambia *si* la observación i es removida (Montgomery *et al.*, 2012).

La regla de oro sugerida por Belsley *et al.* (1980) para establecer a la observación i como influyente es: $|DFFITs_i| > 2\sqrt{(k+1)/n}$. Al igual que con $DFBETAS_{i,j}$,

una magnitud mayor a 2 del estadístico $DFFITS_i$ indicará un *efecto grande* de la observación i en la magnitud del valor ajustado.

Los dos estadísticos presentados en esta sección proveen una adecuada medida de sensibilidad, dado que evalúan la influencia de una observación no sólo por su ubicación en X , sino también por su efecto en la variable dependiente. Su obtención se facilita en gran medida al hacer uso del lenguaje de programación R a través de las funciones `dfbetas(model, ...)` y `dffits(model, ...)` respectivamente (John Fox *et al.*).

Finalmente, cabe destacar que para tamaños de muestra grandes es muy difícil encontrarse con observaciones influyentes; no obstante, resulta sumamente útil descubrir observaciones que son *más influyentes* que otras en la estimación de los modelos de regresión lineal. Para ello, las reglas de oro de los estadísticos $DFBETAS_{i,j}$ y $DFFITS_i$ (que se encuentran en función del tamaño de la muestra) proveen una adecuada forma de identificación (Belsley *et al.*, 1980).

3.1.9 Paso 9. Diagnóstico de multicolinealidad.

En el capítulo II se mencionó la limitación que la matriz de correlación tiene para evaluar el grado de multicolinealidad presente en las variables de interés de un modelo de regresión lineal.

Un diagnóstico de multicolinealidad confiable debe proveer información respecto al **grado de dependencia lineal existente** entre las variables, así como poder identificar **qué variables se encuentran involucradas** (sean dos, tres o más).

Con base en lo anterior, en este paso se realizará un diagnóstico de multicolinealidad que hará uso de una matriz más *potente*, llamada **matriz de proporciones de varianza descompuesta** o Matriz π , propuesta por Belsley *et al.*, 1980; la cual permitirá identificar las variables que se encuentren altamente correlacionadas y el grado en que lo estén, considerando que pueden ser más de dos variables involucradas.

Primeramente, se deben identificar los elementos de la matriz T , la cual proviene de la descomposición de la matriz $X^T X$:

$$X^T X = T \Lambda T^T \quad (3.17)$$

Donde Λ es una matriz diagonal de orden $k \times k$ cuya diagonal principal son los valores propios¹⁸ λ_j ($j = 1, 2, \dots, k$) de $X^T X$ y T es una matriz ortogonal $k \times k$ cuyas columnas son los *vectores propios* de $X^T X$ (Montgomery *et al.*, 2012).

Es importante conocer que los elementos de las columnas de T (t_1, t_2, \dots, t_k) describen la *cantidad* de correlación existente entre las variables explicativas.

Para obtener los valores y vectores propios de cualquier conjunto de datos en el lenguaje de programación R, primeramente se tienen que convertir en una matriz con ayuda de la función `as.matrix()`, posteriormente se obtiene $X^T X$ con el comando `t(x) %*% x`. El resultado de esta operación será el argumento para la función `eigen()` la cual nos proveerá de los valores y vectores propios de $X^T X$ (Narayanachar, Ramaiah, Manjunath; 2016). Lo anterior se ejemplifica a continuación:

```
> datos <- as.matrix(datos)
> x1x <- t(datos) %*% datos
> datos_eigen <- eigen(x1x)
> x1x$values # Para obtener los valores propios
> x1x$vectors # Para obtener los vectores propios
```

¹⁸ Sea A una matriz de $n \times n$

- i) Un valor propio de A , también conocido como *eigenvalor*, es un escalar λ tal que:

$$\det(\lambda I - A) = 0$$

Ecuación Característica

- El valor del escalar λ será la raíz (o raíces) solución de la ecuación característica.
- ii) Los vectores propios de A correspondientes a un valor propio λ , también conocidos como *eigenvectores*, son soluciones diferentes de cero de:

$$(\lambda I - A)\bar{v} = \bar{0}$$

A continuación, se define al estadístico *Factor Inflacionario de Varianza* como:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (3.18)$$

Donde R_j^2 es el coeficiente de determinación de la regresión de la j -ésima variable explicativa sobre las demás variables independientes.

Este estadístico funciona como un **indicador del efecto que las otras variables independientes ($j - 1$) tienen en la varianza del coeficiente de regresión $\hat{\beta}_j$** (Chase Jr., 2013).

El factor inflacionario de varianza se puede obtener de manera sencilla en el lenguaje de programación R con el uso de la función `vif()` la cual requiere la integración de la librería `faraway`.

Para crear la matriz π , se definen las *proporciones de varianza descompuesta* como:

$$\pi_{ij} = \frac{t^2_{ji}/\lambda^2_i}{VIF_j}, \quad \begin{matrix} j = 1, 2, \dots, k \\ i = 1, 2, \dots, k \end{matrix} \quad (3.19)$$

De la expresión anterior se observa que π_{ij} es una matriz cuadrada de orden $k \times k$. Los elementos de cada columna de esta matriz son las proporciones de la varianza de cada $\hat{\beta}_j$ que son contribuidas por los valores propios. Lo anterior se observa mejor en una figura como la que se muestra a continuación:

Índice de Condición Asociado	Proporción de Varianza Descompuesta π					
	x_1	x_2	.	.	.	x_k
c_1	π_{11}	π_{12}	.	.	.	π_{1k}
c_2	π_{21}	π_{22}	.	.	.	π_{2k}
.
.
.
c_k	π_{k1}	π_{k2}	.	.	.	π_{kk}

Figura 3.6 Matriz de la proporción de varianza descompuesta π . Modificado de Belsley *et al.* (1980).

En la figura anterior se aprecia la matriz de proporción de varianza descompuesta, la cual la integran los elementos π_{ij} hasta para k variables explicativas. De lado izquierdo de la matriz se encuentra el *índice de condición asociado*, el cual se define como la razón entre el máximo valor propio de la diagonal de Λ y el valor propio asociado a la j -ésima variable independiente; es decir:

$$C_j = \frac{\lambda_{\text{máx}}}{\lambda_j} \quad (3.20)$$

El índice de condición asociado determinará el **número de relaciones lineales dependientes** que existan entre las variables explicativas.

La regla de oro establece que si $c_j > 1000$ existe una alta dependencia lineal en el renglón j de la matriz π (fuerte multicolinealidad). Valores mayores a 100 y menores a 1000 indican una moderada multicolinealidad entre algunas variables (Montgomery *et al.*, 2012).

Para obtener los índices de condición asociados en el lenguaje de programación R basta con extraer los valores propios obtenidos de la función `eigen()` y realizar la división expresada en (3.20) (Narayanachar *et al.*, 2016) esto es:

```
> datos_eigen <- eigen(x1x)
> max(datos_eigen$values)/datos_eigen$values
> which(max(datos_eigen$values)/datos_eigen$values>1000) # Para identificar
fuertes dependencias lineales
```

Una vez identificado el número de relaciones lineales dependientes, falta conocer **qué variables se encuentran correlacionadas**. Para ello se debe analizar el renglón j identificado con un alto índice de condición asociado. Aquellos elementos del renglón que cuenten con una proporción de varianza descompuesta mayor a 0.5; es decir $\pi_{ij} > 0.5$, determinarán las variables que se encuentran correlacionadas. Éstas, serán las variables x_j asociadas a la columna en la que $\pi_{ij} > 0.5$.¹⁹

Habiendo identificado el número de relaciones lineales dependientes, así como las variables involucradas en la correlación, el último paso en el proceso de diagnosticar la multicolinealidad es determinar las acciones correctivas a las variables encontradas como *problemáticas*.

Como se mencionó en el capítulo II, una técnica usual es eliminar las variables altamente correlacionadas del modelo. Sin embargo, esto no es para nada recomendable debido a la posible pérdida de información e incremento de incertidumbre, al agregar más variables explicativas al término de error μ . Además, cabe recordar la definición de Wooldridge (2015) en la que establece claramente que mientras las variables altamente correlacionadas no involucren a la variable de interés bajo estudio, la multicolinealidad *no* es un problema per se para los fines de la investigación.

¹⁹ El *cutoff* de 0.5 (recomendado por Montgomery *et. al*, 2012 y Belsley *et. al*, 1980) no es una regla estricta, sino una recomendación basada en trabajos empíricos anteriores. Lo que se debe tener en cuenta al evaluar la magnitud de las proporciones es que mientras su valor sea más cercano a uno, más fuerte será la correlación existente entre variables. Por el contrario, valores cercanos a cero indican poca o nula correlación existente entre las mismas.

Por consiguiente, lo primero que se tiene que analizar después del diagnóstico de multicolinealidad es si la o las variables que son de interés para la investigación son *problemáticas*. Para el caso en que dichas variables no se encuentren involucradas en problemas de multicolinealidad (aunque otras variables consideradas en el modelo sí lo estén), no hay razón para tomar acción alguna.

Por el contrario, si resulta que una o más variables de interés se encuentran altamente correlacionadas con otras, se deberá tomar alguna acción correctiva. El camino que se recomienda en esta metodología es tratar de **combinar** las variables que se encuentran altamente correlacionadas en una sola. Si este es el caso, la matriz π proporcionará una gran ayuda en determinar de **qué manera se pueden combinar** dichas variables.

Puede que una relación de dependencia lineal la integren sólo dos variables, mientras que por otro lado exista otra relación de dependencia lineal integrada por más de tres variables. La valiosa información proporcionada por la matriz π permitirá decidir *qué* variables pueden ser combinadas en una sola, así como *cuántas* variables se deben combinar (de ser posible) para que las estimaciones de los coeficientes de regresión sean sumamente confiables.

3.1.10 Paso 10. Validación de la significancia práctica y estadística.

Hasta el inicio de este paso se han ejecutado varias regresiones en función de obtener información para el desarrollo de los pasos anteriores, sin embargo, no se le ha dado importancia alguna al significado de la magnitud de las $\hat{\beta}_j$. Esto ha sido totalmente intencional.

Antes de poner atención a la significancia práctica (magnitud de las $\hat{\beta}_j$) y estadística de las variables involucradas en el modelo, era necesario haber evaluado la *condición* en la que se encuentran los datos obtenidos, en función de identificar problemas tales como la multicolinealidad u observaciones influyentes, que ejerzan un efecto difícilmente identificable (y nada deseado) en la validación de las variables. Por lo anterior, es hasta este último paso en el que se prestará atención a la pertinencia de las variables involucradas (hasta ahora) en el modelo.

Lo primero que se validará será la **pertinencia estadística** de las variables; es decir, se buscará *evidencia suficiente* que justifique el argumento de que el coeficiente β asociado a la variable explicativa j es igual a cero. Si la evidencia es suficiente, estadísticamente la variable en cuestión no tiene ningún efecto en la variable dependiente.

Como se recordará del capítulo II, la técnica que permite realizar esta validación es la *prueba de hipótesis*. Para todas las variables independientes involucradas, el primer argumento que se debe evaluar es:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Además de éste, puede ser interesante poner a prueba otro argumento. Por ejemplo, para el segundo caso presentado en el *Paso 2*, referente al impacto de las asesorías académicas en el rendimiento académico, sería interesante poner a prueba la idea más lógica, en la que dicho impacto es positivo; es decir:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j > 0$$

Con este argumento, a través del rechazo de la hipótesis nula, se busca evidencia suficiente que justifique la afirmación de que el programa de asesorías académicas tiene un efecto positivo en los estudiantes de la Facultad.

El tercer caso del *Paso 2* también puede ser sujeto a prueba desde una diferente perspectiva:

$$H_0: \beta_j < 0.6$$

$$H_1: \beta_j \geq 0.6$$

Con este enunciado, se desea evaluar el *nivel de importancia* de las sesiones de tutoría, a través de *la probabilidad que tiene un estudiante que no asistió a tutoría de reprobación una materia de la DCB respecto a alguien sí asistió* (ver caso tres del Paso 2). Una probabilidad mayor a 0.6 es indicativo de la importancia de las sesiones de tutoría, mientras que una probabilidad menor a 0.6 puede indicar cierta relevancia, pero no en un nivel importante. La evidencia encontrada en esta prueba, en función de rechazar o no la hipótesis nula, permitirá determinar el nivel de importancia que este programa de apoyo tiene en cuanto a la posibilidad de reprobación alguna materia.

Respecto al nivel de significancia α y la probabilidad β , cabe destacar que estos valores ya han sido previamente escogidos en la determinación del tamaño de la muestra del Paso 3. En consecuencia, para la validación estadística de las variables explicativas se conoce con certeza la probabilidad de rechazar la hipótesis nula cuando ésta es verdadera, así como la probabilidad de aceptar dicha hipótesis cuando en realidad es falsa (dado un valor d de sensibilidad).

El estadístico de prueba t presentado en el capítulo II, permitirá evaluar los enunciados de la hipótesis nula y alternativa descritos en esta sección²⁰. A través de esta prueba y del valor crítico obtenido, se determinará si existe evidencia suficiente o no, para rechazar la hipótesis nula.

Una vez evaluada la significancia individual, para aplicaciones como la presentada en el primer caso del Paso 2, será importante evaluar así mismo la significancia conjunta de un grupo de variables de interés. En este caso, variables agrupadas por su dependencia al **estudiante**, a la **institución** o al **contexto**. La prueba

²⁰ A excepción del tercer caso del Paso 2, referente a la probabilidad de reprobación una materia. Al tratarse de un modelo de probabilidad lineal, existe heterocedasticidad en los datos, por ende se deben utilizar los estadísticos de prueba t y F robustos a heterocedasticidad.

estadística F permitirá determinar qué grupo de estos tres tiene un efecto mayor en el rendimiento académico.

Cabe recordar que junto a los resultados de la significancia estadística se debe presentar el *valor-p* asociado a cada prueba de hipótesis. Esto en función de que otros analistas no se limiten a la información proporcionada por el nivel de significancia seleccionado. Así mismo, en caso de que se trabaje con variables de tipo cuantificable y éstas presenten heterocedasticidad, se debe recordar utilizar las pruebas estadísticas t y F robustas.

Finalmente, una vez evaluada la significancia estadística, es momento ahora sí, de prestar atención a la magnitud de los coeficientes estimados $\hat{\beta}_j$. La significancia práctica permitirá cuantificar el impacto de cada una de las variables incluidas en el modelo. Valores $\hat{\beta}_j$ cercanos a cero indicarán muy poca o nula significancia de la variable independiente j asociada.

Para que una variable explicativa (o grupo de variables) pueda(n) permanecer en el modelo, deben contar con significancia estadística y práctica.

En este punto, cobra gran relevancia el uso del coeficiente de determinación ajustado \bar{R}^2 , que aportará información sobre la proporción de la varianza en la variable dependiente explicada por las variables independientes que se decida, permanezcan en el modelo.

3.2 Diagrama de flujo

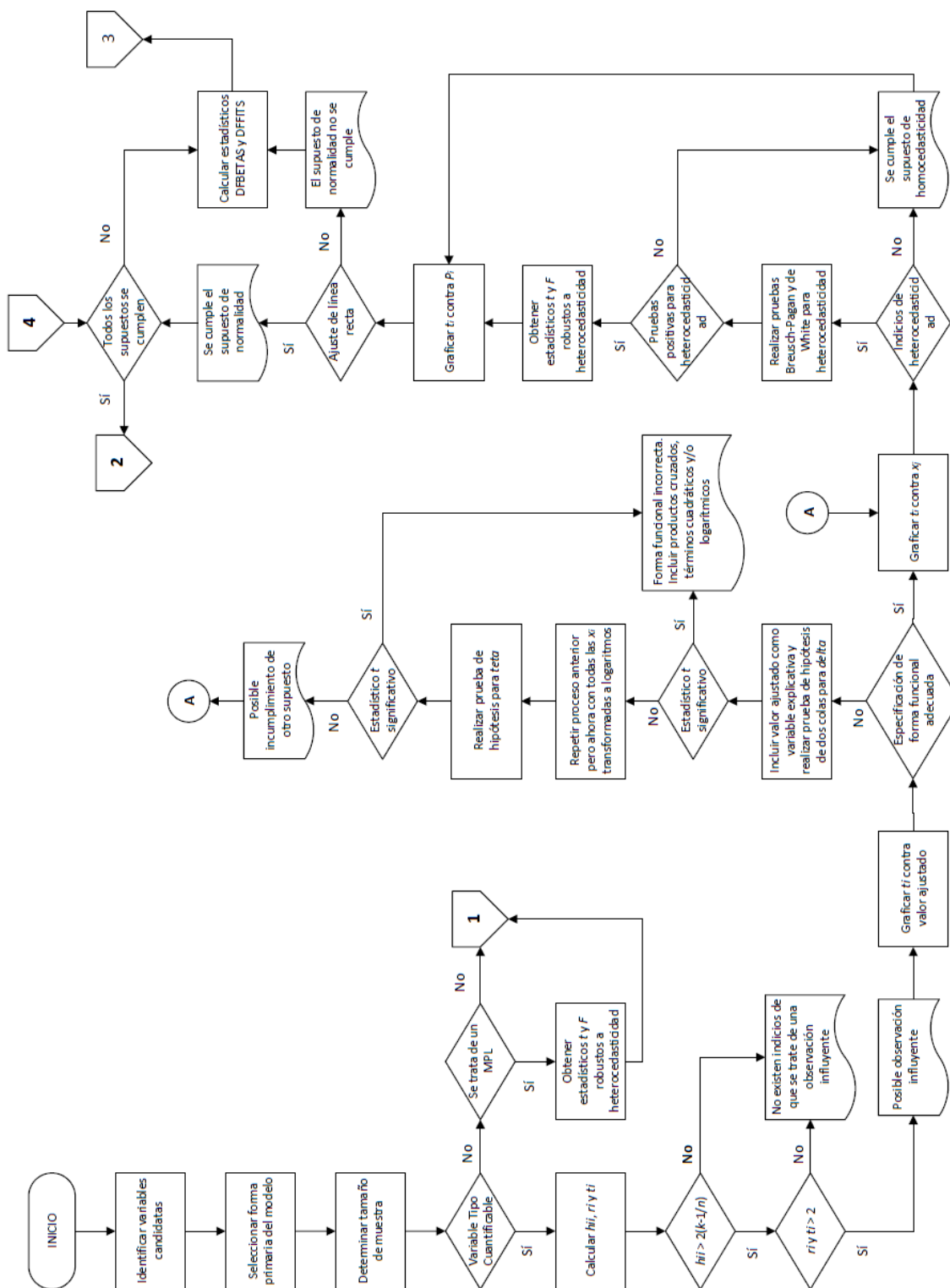


Figura 3.7 Diagrama de flujo para la creación del modelo. Elaboración propia.

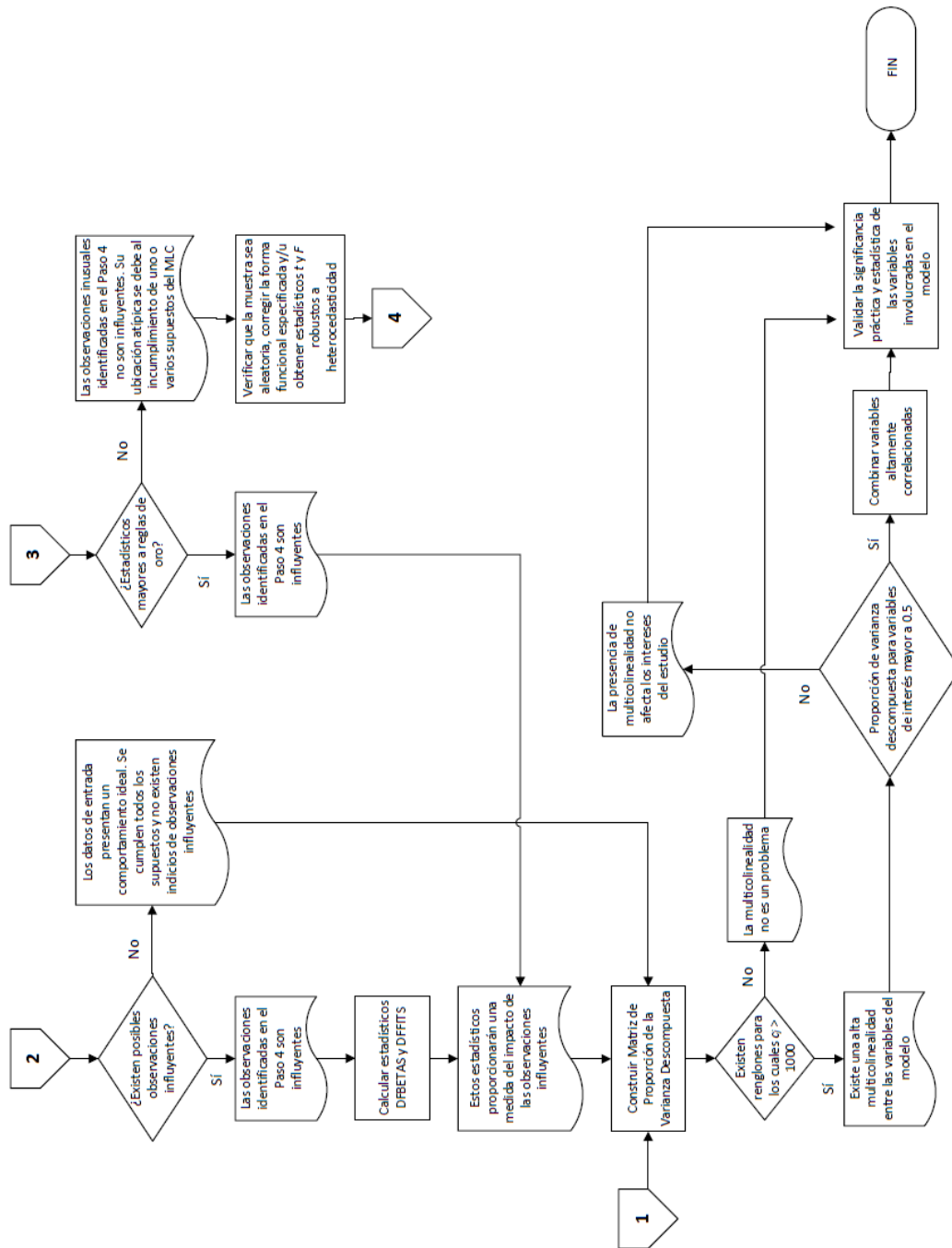


Figura 3.8 Diagrama de flujo para la creación del modelo (continuación). Elaboración propia.

3.3 Resumen del capítulo

En este capítulo se presentó una metodología para crear un modelo de regresión lineal que permita cuantificar el impacto de diversos factores (debidos al estudiante, al contexto o a la institución) en el rendimiento académico. La propuesta metodológica se integra por los siguientes diez pasos:

Paso 1: Identificación de variables candidatas. La función de este primer paso es contar con un panorama de la información disponible, considerando el tipo de variable con la que cuenta la institución educativa (cuantificable, binaria u ordinal), y su forma de medición. Así mismo, en este paso las variables se clasifican en función de su categoría. Las figuras 1.3, 1.4, 1.5 y 1.6 pueden usarse como guía en la clasificación de las variables.

Paso 2: Especificación primaria de la forma funcional del modelo. En este segundo paso se seleccionan las variables a incluir en el modelo, esto en función de la(s) pregunta(s) de investigación a la(s) que se desee encontrar respuesta. Para ello, se escoge la forma funcional que se crea más apropiada y/o que provea de formatos deseados para los resultados, tales como porcentajes o probabilidades.

Paso 3: Determinación del tamaño de muestra. El tercer paso consiste en la determinación del tamaño de la muestra. Para ello se debe especificar la probabilidad de cometer un error tipo I (α) y la de cometer un error tipo II (β), así como el tamaño relativo de la diferencia en las medias que quiera detectarse (d). Consecuentemente se deberán consultar las curvas características de operación que permitan encontrar el tamaño de muestra en función de los parámetros especificados.

Paso 4: Identificación de probables observaciones influyentes. El cuarto paso consiste en el análisis de los datos obtenidos, en función de identificar observaciones que difieran en gran medida del resto de los datos; es decir, que sean inusuales. Estas observaciones pueden ser de dos tipos: observaciones palanca u observaciones influyentes. Las primeras no representan gran problema en la estimación de los coeficientes; sin embargo, las segundas provocan estimaciones poco confiables además de una pobre capacidad de predicción del modelo.

Para la identificación de este tipo de observaciones se propone hacer uso de la diagonal de la matriz \mathbf{H} (valores h_{ii}) en conjunto con el residual studentizado tanto interno (r_i) como externo (t_i) para cada observación. Esto permitirá identificar observaciones con alta probabilidad de ser influyentes.

Paso 5: Validación de la forma funcional. Para tener certeza de que las observaciones identificadas en el paso anterior sean verdaderamente influyentes, se debe descartar la posibilidad de que su inusual ubicación en el espacio en x se deba al incumplimiento de algún supuesto del MLC. Por ello, el quinto paso se encarga de validar la forma funcional especificada con anterioridad; esto, a través del uso de una gráfica de dispersión que compare los valores ajustados \hat{y}_i obtenidos de la regresión del modelo especificado en el paso 2, contra los residuales studentizados t_i .

En caso de encontrar evidencia de una especificación incorrecta de la forma funcional, se propone realizar una prueba de hipótesis que involucre al valor ajustado \hat{y} como variable independiente en un modelo como el siguiente: $Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \delta \hat{y}^2 + \mu$. En caso de que el estadístico de prueba t resulte significativo bajo $H_0: \delta = 0$ contra una alternativa de dos colas, se comprobará que la especificación es incorrecta y además se obtendrá evidencia para justificar la inclusión de términos cuadráticos y/o productos cruzados al modelo.

Se establece un proceso similar para evaluar la necesidad de incluir términos logarítmicos, ahora incluyendo al valor ajustado en un modelo como el siguiente: $Y = \beta_0 + \beta_1 \ln(x_1) + \dots + \beta_k \ln(x_k) + \theta \hat{y} + \mu$. En caso de que el estadístico de prueba t resulte significativo bajo $H_0: \theta = 0$ contra una alternativa de dos colas, se podrá concluir que el modelo requiere de la inclusión de términos logarítmicos para una adecuada especificación.

Paso 6: Validación del Supuesto de homocedasticidad. El sexto paso consta de la validación de otro importante supuesto del MLC que es el de homocedasticidad. Para ello, se propone en primera instancia un diagnóstico visual a través de una gráfica de dispersión que compare al residual studentizado t_i contra cada variable explicativa x_j . Si las observaciones presentan cierta tendencia, esto será evidencia del incumplimiento del supuesto.

Como comprobación analítica se propone hacer uso de la prueba Breusch-Pagan y de la prueba de White para heterocedasticidad. En caso de que se compruebe el incumplimiento del supuesto de homocedasticidad, se deberán obtener los estadísticos t y F robustos a heterocedasticidad.

Paso 7: Validación del supuesto de normalidad. El séptimo paso consta de la validación del supuesto de normalidad, el *más fuerte* de todos dado que éste incluye de manera implícita a los demás. Al respecto, el supuesto de media condicional cero se verifica de manera simultánea al evaluar la existencia de normalidad. Dado que al llegar a este paso ya se han validado dos de los supuestos más importantes, así como también se ha recolectado una muestra aleatoria representativa, para la validación del supuesto de normalidad y en consecuencia, la validación del supuesto de media condicional cero, es suficiente una evaluación visual de la gráfica de probabilidad normal de los residuales studentizados, la cual compara los t_i contra la probabilidad acumulada $P_i = \frac{\binom{i-1}{2}}{n}$, $i = 1, 2, \dots, n$.

Paso 8: Comprobación de observaciones influyentes. Si todos los supuestos validados en pasos anteriores se cumplieron, entonces en este paso se comprueba que las observaciones identificadas en el cuarto paso son verdaderamente influyentes. En caso de que alguno o varios de los supuestos no se cumplan, se propone hacer uso de los estadísticos $DFBETAS$ y $DFFITs$ como medida de comprobación.

Independientemente del cumplimiento o no de los supuestos del MLC, se sugiere hacer el cálculo de los estadísticos mencionados ya que éstos permiten conocer el efecto en la estimación del modelo de regresión de las observaciones influyentes identificadas.

Paso 9: Diagnóstico de multicolinealidad. En este paso se evalúa el grado existente de multicolinealidad en el modelo. Para ello, se hace uso de la matriz de proporción de varianza descompuesta π que permite identificar variables que se encuentren altamente correlacionadas y el grado en que lo estén, considerando que pueden ser más de dos variables involucradas.

Para conocer el número de relaciones lineales dependientes que existan entre las variables explicativas se hace uso del índice de condición asociado c_j definido como la razón entre el máximo valor propio y el valor propio asociado a la j -ésima variable independiente. Si $c_j > 1000$ existe una alta dependencia lineal en el renglón j de la matriz π (fuerte multicolinealidad). Valores mayores a 100 y menores a 1000 indican una moderada multicolinealidad entre algunas variables.

Para conocer qué variables se encuentran correlacionadas se debe analizar el renglón j identificado con un alto índice de condición asociado. Aquellos elementos del renglón que cuenten con una proporción de varianza descompuesta mayor a 0.5; es decir $\pi_{ij} > 0.5$, evidenciarán a las variables que se encuentren correlacionadas.

Si la o las variables de interés bajo estudio se encuentran altamente correlacionadas con otras variables, se propone combinar dichas variables de interés para eliminar el efecto de la dependencia lineal y que las estimaciones de los coeficientes sean sumamente confiables. En caso de que las variables de interés no se encuentren involucradas en fuertes dependencias lineales, se considera que la multicolinealidad no es un problema.

Paso 10: Validación de la significancia práctica y estadística. En este último paso se evalúa la significancia práctica y estadística de las variables involucradas en el modelo. Para el caso de la significancia estadística se realiza la prueba de hipótesis bajo $H_0: \beta_j = 0$ contra una alternativa de dos colas. En caso de que el estadístico de prueba t resulte significativo, entonces estadísticamente la variable en cuestión no tiene efecto sobre la variable dependiente.

En este paso también se evalúa la significancia conjunta de un grupo de variables de interés; esto, con el uso del estadístico de prueba F para restricciones múltiples. En caso de trabajar con modelos de probabilidad lineal MPL o en presencia de heterocedasticidad, para la prueba de hipótesis se deben utilizar los estadísticos de prueba t y F robustos a heterocedasticidad. Así mismo, junto a los resultados de la significancia estadística se debe presentar el *valor-p* asociado a cada prueba de hipótesis. Finalmente, la significancia práctica estará dada por la magnitud de los coeficientes $\hat{\beta}_j$. Tanto la significancia práctica como estadística deben considerarse en la decisión de incluir o excluir variables independientes.



Principal, Facultad de Ingeniería UNAM
Foto: oferta.unam.mx/escuela-facultad/23/facultad-de-ingenieria

Una buena práctica en la construcción de modelos de regresión lineal MRL es su **validación**; esto es, la verificación de que el modelo funcionará de manera exitosa en la aplicación deseada. (Montgomery *et al.*, 2012).

Una *funcionalidad exitosa* para los MRL se basa principalmente en dos aspectos: un ajuste adecuado a las observaciones (residuales pequeños) y la capacidad de predecir nuevas observaciones de manera precisa.

Con base en lo anterior, un adecuado proceso de validación debe considerar el análisis de la magnitud y signo de los coeficientes en función de determinar si los $\hat{\beta}_j$ obtenidos pueden ser interpretados de manera razonable como una estimación del efecto de x_j ; así mismo, se debe analizar la estabilidad de los coeficientes al determinar si existen diferencias considerables al utilizar una muestra diferente. Por último, la validación requiere que se investigue la capacidad predictiva del modelo.

Cabe destacar que un modelo con ajuste adecuado a las observaciones no implica necesariamente que tenga también una buena capacidad predictiva. Puede ser el caso de que, a pesar de contar con un buen ajuste, la capacidad de predicción sea limitada.

Con base en Montgomery *et al.* (2012) existen tres técnicas para validar un modelo de regresión: (i) el análisis de los coeficientes y valores ajustados mediante la comparación de experiencias previas, (ii) la recolección de nuevas observaciones para investigar el desempeño en la predicción y (iii) la separación de los datos existentes al utilizar sólo una parte en la construcción del modelo y la otra para evaluar la capacidad predictiva.

A continuación, se explica con más detalle el uso de estas técnicas.

4.1 Análisis de coeficientes y valores ajustados

En el último paso de la metodología propuesta para la creación del modelo del capítulo anterior, se menciona la importancia de analizar la magnitud de los coeficientes estimados en función de ponderar la significancia práctica de las variables explicativas involucradas. Este análisis se encuentra íntimamente ligado con una parte del proceso de validación del modelo, el cual consiste en determinar si los signos y las magnitudes de las estimaciones $\hat{\beta}_j$ son razonables.

Dada la cantidad de información histórica con la que cuentan la mayoría de las instituciones educativas, en especial aquellas con vasta trayectoria tales como la Facultad de Ingeniería, es relativamente sencillo comparar los resultados obtenidos de una regresión lineal con **experiencias previas**, en función de validar si las estimaciones obtenidas concuerdan con lo experimentado en la realidad.

Un ejemplo de esto pudiera ser el efecto de un programa de apoyo, como el de las asesorías académicas. Al construir un modelo de regresión lineal para estimar el efecto de este factor, el analista se puede encontrar con la sorpresa de que el signo del coeficiente estimado sea negativo. Esto es contra intuitivo por lo que, una manera de *validar* el resultado, es a través de la comparación de las calificaciones obtenidas por los estudiantes, antes y después de la implementación de las asesorías académicas. Si la tendencia indica una mejora después de la implementación de este programa de apoyo, entonces el signo de la estimación **no** es razonable.

Estimaciones $\hat{\beta}_j$ de magnitud y/o signo poco razonables, son indicios de una estimación poco confiable causada por el incumplimiento de algún supuesto del MRL así como también de problemas de multicolinealidad.

Como se recordará del capítulo anterior, en los pasos para la creación del modelo se toma en cuenta esta posibilidad y se plantean líneas de acción para que las estimaciones obtenidas sean sumamente confiables.

Una de estas líneas de acción involucra al **factor inflacionario de varianza** (VIF) por sus siglas en inglés, utilizado en el *Paso 9* para el diagnóstico y valoración de multicolinealidad. Este estadístico también funge como una importante guía para la validación de un modelo. En caso de que se encuentre con un VIF_j mayor a 5 o a 10, el coeficiente asociado a la j -ésima variable independiente se encuentra pobremente estimado, debido a una alta dependencia lineal entre los regresores (Montgomery *et al.*, 2012).

Otra manera de validar el modelo a través de las estimaciones de los coeficientes es evaluando su **estabilidad**. Esto es, obtener estimaciones muy parecidas al realizar el procedimiento con una muestra completamente diferente. Este tipo de validación se garantiza en el *Paso 3* de la propuesta metodológica al diseñar un tamaño de muestra **representativo** de la población.

Dado que en el proceso de la metodología propuesta para la creación del modelo se realiza un análisis completo para eliminar cualquier posibilidad de obtener estimaciones poco confiables, es muy poco probable que, siguiendo los pasos establecidos, el analista se encuentre con estimaciones $\hat{\beta}_j$ no razonables.

El análisis de los valores ajustados \hat{y} provee otra medida de validación del modelo. Al igual que las estimaciones de los coeficientes, valores ajustados poco razonables (negativos para observaciones de naturaleza positiva, o que caen fuera del rango de respuesta de la variable) son indicios del incumplimiento de algún supuesto del MLC o de problemas de multicolinealidad.

Para un modelo creado bajo el cumplimiento de los diez pasos establecidos en el tercer capítulo, esta técnica de validación, más que para *validar* los resultados obtenidos, puede servir para proporcionar evidencia que demuestre la confiabilidad de las estimaciones, debido a que el proceso de creación del modelo ya la involucra de manera implícita.

4.2 Recolección de nuevas observaciones

La recolección de nuevas observaciones es la técnica más efectiva en cuanto a la validación de la capacidad predictiva del modelo. Si el modelo proporciona predicciones precisas de nuevas observaciones, el usuario o analista incrementará su confianza tanto en el modelo como en su proceso de construcción (Montgomery *et al.*, 2012).

Los mismos autores establecen que un rango de entre 15 a 20 nuevas observaciones es deseable en la recolección de más datos, para proporcionar una evaluación confiable de la capacidad predictiva del modelo.

Una manera muy sencilla de comparar lo real contra lo predicho por el modelo es a través de la diferencia absoluta de las observaciones: $|y_i - \hat{y}_i|$. El escenario ideal sería que las diferencias fueran muy cercanas a cero.

De manera general, la capacidad predictiva de un modelo se puede evaluar a través del **cuadrado medio de error de predicción (CMEP)** encontrado en las nuevas observaciones (Montgomery *et al.*, 2012), es decir:

$$CMEP = \frac{\sum_{i=n+1}^N (y_i - \hat{y}_i)^2}{p} \quad (4.1)$$

Donde $p =$ cantidad de nuevas observaciones, $n =$ tamaño de muestra, y $N = n + p$.

Por otra parte, definiendo el **cuadrado medio de error o residual (CME)** del modelo como:

$$CME = \frac{\sum_{i=1}^n \hat{\mu}_i^2}{n - k - 1} \quad (4.2)$$

Es posible comparar la capacidad de ajuste del modelo contra su capacidad predictiva, a través del *CMEP* y el *CME*. Si $CMEP > CME$ la capacidad predictiva del modelo *es menor* a la capacidad de ajuste con las observaciones existentes. En caso de que $CME > CMEP$ entonces la capacidad predictiva *es mayor* a la capacidad de ajuste. Que una conclusión obtenida de esta comparación sea buena o mala, depende del (los) objetivo(s) y de la(s) pregunta(s) de investigación planteada(s) en la construcción del modelo.

Si bien la técnica de recolección de nuevas observaciones es la más efectiva, el costo de su uso es tener que volver a campo para obtener la nueva información. Puede que, para algunas aplicaciones, esto no sea posible, para lo cual el uso de la siguiente técnica de validación sería más apropiado.

4.3 Separación de los datos existentes

Si no es posible recolectar nueva información debido a ciertas restricciones (como pudieran ser el tiempo disponible y el costo asociado) lo más viable es utilizar los datos existentes. La manera de hacer esto es separar el total de observaciones en dos grupos, el primero de ellos denominado *datos de estimación*, y el segundo denominado *datos de predicción* (Snee, 1977).

El primer grupo es usado para construir el modelo de regresión, mientras que el segundo es usado para analizar la capacidad predictiva de dicho modelo. Lo más importante en el uso de esta técnica de validación es determinar la manera en la que se escogerán las observaciones en función de asignarlas al primer o segundo grupo.

Para datos de series de tiempo, es posible utilizar el *tiempo* como medida de separación. Esto es, establecer una *fecha de corte*, para la cual, todas las observaciones anteriores a dicha fecha pertenecerán al grupo de datos de estimación, mientras que las observaciones recolectadas en fechas posteriores a la de corte, pertenecerán al grupo de datos de predicción.

Sin embargo, para datos de corte transversal (en los cuales se basa la metodología del capítulo III) muchas veces la medida de separación no es tan obvia. En este caso, lo más sencillo sería utilizar cualquier regla arbitraria en la asignación de las observaciones a alguno de ambos grupos.

Al respecto, Montgomery *et al.* (2012) advierten sobre el uso de reglas arbitrarias en la separación de los datos, ya que éstas presentan una desventaja en cuanto a que no se tiene ninguna seguridad de que el grupo de datos de predicción sean lo suficientemente “estrictos” para poder evaluar al modelo. Esto significa que no hay garantía en que las observaciones del grupo de predicción sean **extrapolaciones**, por lo que, el esfuerzo de validación proporcionaría escasa información en cuanto a la capacidad predictiva del modelo.

Ante estas dificultades, para establecer una medida de separación, Snee (1977) publicó un algoritmo cuyo objetivo es “dividir los datos en dos grupos que cubran aproximadamente la misma región y cuenten con propiedades estadísticas similares”. El algoritmo, denominado DUPLEX, fue originalmente desarrollado por R. W. Kennard; su procedimiento utiliza la distancia entre pares de todas las observaciones.

El algoritmo DUPLEX consta de los siguientes pasos:

1.- Se estandarizan todas las observaciones i para cada variable explicativa j . Esto se logra a través de la fórmula:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{jj}^{1/2}}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, k \quad (4.3)$$

Donde $S_{jj} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ es la suma de cuadrados ajustada de la j-ésima variable independiente (Montgomery *et al.*, 2012).

2.- Se ortonormaliza a los regresores estandarizados. Esto se logra factorizando la matriz $\mathbf{Z}^T\mathbf{Z}$ como: $\mathbf{Z}^T\mathbf{Z} = \mathbf{T}^T\mathbf{T}$, donde \mathbf{T} es una matriz triangular superior. Los elementos de \mathbf{T} se encuentran utilizando el método Cholesky.²¹

3.- Utilizando \mathbf{T} se realiza la transformación $\mathbf{W} = \mathbf{Z}\mathbf{T}^{-1}$. El nuevo grupo de variables w serán ortogonales y tendrán varianza unitaria; **esta transformación permite que las observaciones se esparzan espacialmente de manera uniforme** (Montgomery *et al.*, 2012; Kennard y Stone, 1969).

4.- Se calcula la distancia Euclidiana entre todos los pares posibles $\binom{n}{2}$ de puntos u observaciones.

5.- Las dos observaciones que se encuentren más alejadas entre sí se asignan al grupo *datos de estimación*. Posteriormente, de las observaciones restantes, el par de observaciones que se encuentren más alejadas entre sí se asignan al grupo *datos de predicción*. De los datos restantes, la observación que se encuentre más apartada del par de *datos de estimación* se asigna a éste mismo grupo. De manera análoga, de los datos restantes, aquella observación que se encuentre más apartada de los *datos de predicción* se asigna a este mismo grupo. El algoritmo continúa hasta que todas las observaciones hayan sido clasificadas en alguno de los dos grupos (Snee, 1977; Montgomery *et al.*, 2012).

Para hacer uso del algoritmo DUPLEX como técnica de validación, Snee (1977) recomienda **no** considerar la separación de datos a menos que $n \geq 2k + 25$; así mismo, en función de “contar con adecuados grados de libertad para el error, y así proporcionar significativas pruebas de hipótesis y análisis de residuales”, el tamaño del grupo *datos de estimación* debe ser mayor a $k + 10$ (o $k + 15$), recordando que k representa el número de variables independientes consideradas.

²¹ Para mayor referencia en cuanto al método Cholesky consultar *Theory and application of the linear model* (1976) de Franklin A. Graybill.

Por otra parte, a pesar de que los conjuntos de datos de estimación y de predicción pueden ser del mismo tamaño, también es posible que el primero de ellos contenga un mayor número de datos con respecto al segundo. Esto resulta lógico al considerar que mientras más datos se tengan para crear el modelo, las estimaciones serán más precisas.

En caso de decidir utilizar un conjunto de datos de estimación más grande que el de datos de predicción, el algoritmo DUPLEX no cambia de manera importante. La única diferencia sería que, al alcanzar el tamaño máximo del conjunto de datos de predicción, todas las observaciones restantes se asignarían al conjunto de datos de estimación. Una proporción razonable sería que al conjunto de datos de estimación se le asignara el 70% del total del tamaño de la muestra, y que al conjunto de datos de predicción se le asignara el 30% restante. Al respecto, Snee (1977) advierte que, para contar con una validación confiable de la capacidad predictiva del modelo, el conjunto de datos de predicción debe contener al menos 15 observaciones.

La última recomendación de Snee (1977) concerniente al uso del algoritmo DUPLEX, establece que es necesario eliminar *observaciones replicas* o *pseudoreplicas*. Las observaciones replicas son aquellas que cuentan con valores idénticos en x_1, x_2, \dots, x_k , mientras que las observaciones pseudoreplicas, también conocidas como *near neighbors*, son aquellas que cuentan con valores casi idénticos en sus variables independientes.

Si este tipo de observaciones no se identifican y se usan en la ejecución del algoritmo DUPLEX, los conjuntos de datos de estimación y predicción serán muy similares, lo que provocará una deficiente validación del modelo. En palabras de Snee (1977): “Dado que [al utilizar observaciones replicas o pseudoreplicas] el conjunto de datos de predicción cuenta con la misma estructura de correlación que el conjunto de datos de estimación, la desviación estándar predicha será aproximadamente igual que la desviación estándar de los residuales del conjunto de datos de estimación, independientemente de si el modelo es “válido” o no”.

Al respecto, Montgomery *et al.* (2012) proponen el uso de la suma ponderada de la distancia cuadrada (WSSD por sus siglas en inglés) entre dos observaciones,

como medida para la identificación de observaciones replicas o *near neighbors*. La WSSD se define como:

$$D_{ii'}^2 = \sum_{j=1}^k \left[\frac{\hat{\beta}_j (x_{ij} - x_{i'j})}{\sqrt{\hat{\sigma}^2}} \right]^2 \quad (4.4)$$

Donde $D_{ii'}^2$ representa la suma ponderada de la distancia cuadrada entre la observación i y la observación i' . Pares de observaciones con “pequeños” valores de $D_{ii'}^2$ se consideran *near neighbors*, lo que implica que se encuentran *relativamente* cerca en el espacio en x . Por otra parte, pares de observaciones con valores de $D_{ii'}^2$ “grandes” (por ejemplo $D_{ii'}^2 \gg 1$) no se consideran *near neighbors* (Montgomery *et al.*, 2012).

Existe una dificultad al utilizar la WSSD como medida para identificar *near neighbors*, y es que quienes la proponen no establecen **qué tan pequeño debe ser $D_{ii'}^2$** para clasificar al par de observaciones como *near neighbors*.

Existen otras medidas para identificar este tipo de observaciones. Johnson (1967) por ejemplo, hace uso de una matriz jerárquica basada en la distancia entre puntos para identificar *near neighbors*.²²

Una vez identificados *near neighbors* en los datos de la muestra, para poder proseguir con el algoritmo DUPLEX, dichas observaciones **se deben promediar en sus valores en x** . Los promedios para cada x_1, x_2, \dots, x_k serán los datos a utilizar en el algoritmo.

Finalmente, al determinar a qué conjunto pertenecerá cada observación (si al de datos de estimación o de predicción), el último paso es crear el modelo de regresión con base al primer conjunto y, utilizando los datos de predicción, validar su

²² Para mayor referencia consultar *Hierarchical Clustering Schemes*, Stephen C. Johnson (1967).

capacidad predictiva. Para esta tarea es posible utilizar los estadísticos *CMEP* y *CME*.

4.4 Resumen del capítulo

Este último capítulo se enfocó en las metodologías o técnicas existentes de validación para un modelo de regresión lineal.

Entendiendo como *validación*, la verificación de que el modelo funcionará de manera exitosa en la aplicación deseada, se encontró que una *funcionalidad exitosa* para los MRL se basa principalmente en dos aspectos: un ajuste adecuado a las observaciones y la capacidad de predecir nuevas observaciones de manera precisa.

Se destacó que, un modelo con ajuste adecuado a los datos existentes, no implica necesariamente que tenga también una buena capacidad predictiva. Puede ser el caso de que, a pesar de contar con un buen ajuste, la capacidad de predicción del modelo sea limitada.

En consecuencia, se analizaron las tres principales técnicas de validación. La primera de ellas, concerniente al **análisis de los coeficientes y valores ajustados**, se enfoca en el ajuste del modelo a los datos existentes. Esta técnica de validación se basa en la magnitud y signo de los coeficientes $\hat{\beta}_j$, así como de los valores ajustados \hat{y} , en función de determinar si éstos pueden ser interpretados de manera razonable, como una estimación del efecto de x_j en el primer caso, y como una respuesta de la variable dependiente y en el segundo.

Se estableció que, el comparar la información obtenida de los coeficientes $\hat{\beta}_j$ con experiencias previas, es una buena forma de validar si los resultados obtenidos son razonables; así mismo, valores ajustados negativos para observaciones de naturaleza positiva, o que caen fuera del rango de respuesta de la variable, son indicios de una limitada funcionalidad del modelo.

Además, se presentó a la comparación de estimaciones obtenidas de muestras diferentes, como otra manera de validar el modelo a través del análisis de los coeficientes. Si las estimaciones son similares, entonces se dice que el modelo presenta *estabilidad*.

La segunda técnica analizada fue la **recolección de nuevas observaciones**. Con base en la literatura, esta técnica es la más efectiva en cuanto a la validación de la capacidad predictiva del modelo.

Así mismo, para que la evaluación de dicha capacidad sea confiable, se identificó que es necesario recolectar, al menos, de entre 15 a 20 nuevas observaciones.

Respecto a la capacidad predictiva, se presentó al cuadrado medio de error de predicción (CMEP), como un estadístico a utilizar para evaluar la precisión de predicción del modelo.

Si bien la técnica de recolección de nuevas observaciones es la más efectiva, el costo de su uso es tener que volver a campo para obtener la nueva información. Para varias aplicaciones, es posible que existan restricciones en cuanto al tiempo disponible y/o el costo asociado, para lo cual esta técnica de validación no resulte viable. En consecuencia, se presentó a la técnica de validación basada en la **separación de los datos existentes**, como una alternativa capaz de evaluar la capacidad predictiva del modelo.

Esta técnica separa el total de observaciones en dos grupos, el primero de ellos denominado *datos de estimación*, y el segundo denominado *datos de predicción*. El objetivo de esta separación es utilizar al primer grupo para construir el modelo de regresión, mientras que el segundo es usado para analizar la capacidad predictiva de dicho modelo. Lo más importante en el uso de esta técnica de validación, es determinar la manera en la que se escogerán las observaciones, en función de asignarlas al primer o segundo grupo.

Al respecto, se presentó al algoritmo DUPLEX como una forma metódica de separar los datos en ambos conjuntos, garantizando grupos que cubran aproximadamente la misma región y cuenten con propiedades estadísticas similares (Snee, 1977).

Respecto al uso del algoritmo DUPLEX, se encontró que es muy importante eliminar del conjunto de datos, observaciones replicas o pseudoreplicas (es decir, observaciones con valores idénticos o casi idénticos en x_1, x_2, \dots, x_k). Esto, debido a que, si se usan en la ejecución del algoritmo, los conjuntos de datos de estimación y predicción serán muy similares, lo que provocará una deficiente validación del modelo.

Para identificar a este tipo de observaciones, se analizó la propuesta de Montgomery *et al.* (2012), basada en el uso de la suma ponderada de la distancia cuadrada $D_{ii'}^2$, entre dos observaciones. Este estadístico pondera los valores en x_1, x_2, \dots, x_k entre la observación i e i' , en razón de la desviación estándar. Sin embargo, se identificó una dificultad al usar este estadístico, ya que los autores de la propuesta no establecen qué tan pequeño debe ser $D_{ii'}^2$ para considerar a un par de observaciones como *near neighbors* o pseudoreplicas.

Finalmente, en consecuencia a esta dificultad, se presentó una referencia a la propuesta metodológica de otro autor, para identificar a este tipo de observaciones.

Conclusiones.

Este trabajo de investigación aborda un procedimiento para construir y estimar modelos de regresión lineal con base en una muestra aleatoria obtenida en un punto dado en el tiempo. En tanto que existe una miríada de investigaciones relacionadas con el rendimiento académico y con los factores que influyen en este, que utilizan modelos de regresión lineal, pocos autores se han preocupado por utilizar una metodología que permita obtener estimaciones confiables e insesgadas.

Con base en la revisión de la literatura, no se encontraron investigaciones relacionadas con el objeto de estudio de este trabajo, que consideraran en la estimación de modelos de regresión lineal la influencia de observaciones influyentes, el cumplimiento de los supuestos del modelo lineal clásico, o la existencia de una alta multicolinealidad entre las variables independientes involucradas.

La importancia de contar con estimaciones confiables e insesgadas radica en que, una vez obtenida la estimación de los parámetros poblacionales $\hat{\beta}$ (la cual se considera la medida del impacto de programas de apoyo a estudiantes, entre otros) se puedan establecer metas que sean realistas y alcanzables (con base en la información obtenida de la estimación de los MRL). Esto, a su vez, permitirá diseñar indicadores de desempeño que sirvan como herramienta a los procesos de mejora continua de las instituciones educativas.

Así mismo, he identificado que al menos para esta aplicación (la valoración de factores que afectan el rendimiento académico de los estudiantes) el uso del modelo de regresión lineal se aferra a una vertiente mecánica del mismo, en donde lo que importa es encontrar la curva que mejor se ajuste a los datos. Esto deja de lado una aplicación de los MRL más enriquecedora, la cual integra aspectos relacionados con la **teoría estadística**, aspectos relacionados con la **teoría de la disciplina a analizar** y **los datos**. Desde 1980 David A. Belsley, Edwin Kuh y Roy E. Welsch ya habían identificado la importancia de analizar estos tres elementos en conjunto, sin embargo, al menos para el análisis del rendimiento académico, esta idea aún no se

ha generalizado. Es mi intención que, con el uso de la metodología presentada en este trabajo, esto pueda cambiar.

En consecuencia a lo anterior, de este trabajo de investigación se desprenden 10 pasos que, ejecutados en secuencia, permiten construir y estimar modelos de regresión lineal, considerando el impacto de observaciones influyentes, del incumplimiento de algún supuesto del MLC o de la existencia de altas correlaciones entre las variables independientes.

Finalmente, se establecen a continuación tres líneas de investigación o ejes de interés que he identificado, para los cuales la metodología propuesta en esta tesis puede proporcionar información de utilidad. Éstas son:

a. La cuantificación del impacto de programas de apoyo

En esta línea, el interés radica en poder cuantificar el impacto de programas de apoyo institucionales, tales como las asesorías académicas, las tutorías, los cursos extraordinarios, etc. Las preguntas a las que se les podrá dar respuesta son, por ejemplo:

- ¿Qué diferencia existe entre estudiantes que acuden a asesorías académicas y aquellos que no, dados los mismos niveles en rendimiento previo, motivación, situación económica, etc., en cuanto a su rendimiento académico?
- ¿Las sesiones de tutoría marcan un cambio positivo en la tendencia de estudiantes con un pobre rendimiento previo?
- ¿Los cursos extraordinarios han disminuido el tiempo en el que un alumno vuelve a ser regular, después de haber reprobado una o varias materias? Si así fuera, ¿en qué medida?
- ¿Qué diferencia existe entre estudiantes que presentan exámenes extraordinarios en tres etapas y exámenes con taller de preparación? ¿Cuál de los dos programas incrementa la tasa de aprobación de la mejor manera?
- ¿En qué medida las asesorías psicopedagógicas influyen en el rendimiento de los estudiantes?

- Dados los niveles socioeconómicos y sociodemográficos, en promedio, en qué medida la obtención de una beca influye en la probabilidad de aprobar todas las materias de un semestre.

b. Comparación entre factores que influyen en el rendimiento académico

Este eje de interés concierne en comparar factores que influyen en el rendimiento académico, en función de encontrar cuáles tienen mayor peso, e inclusive poder ponderarlos de mayor a menor influencia. Las preguntas a las que se les puede dar respuesta son, por ejemplo:

- ¿Qué factor (motivacional, socioeconómico, debido al docente, etc.) en promedio, tiene mayor influencia en el rendimiento de los estudiantes?
- En el caso de estudiantes de bajos recursos, en promedio ¿una mejora en el factor docente podría contrarrestar o inhibir la influencia del factor socioeconómico? Si así fuera, ¿en qué medida?
- ¿Qué tan importante es el factor psicosocial en cuanto al rendimiento académico?

c. Análisis proactivos

La última línea de investigación o eje de interés, concierne a los análisis proactivos que son posibles de hacer con el uso de modelos de probabilidad lineal (MPL). Estos análisis permitirán identificar *riesgos* (de reprobar, de desertar) asociados al perfil de los estudiantes. Con esta información las instituciones educativas pueden actuar de manera proactiva en el apoyo a los estudiantes que así lo requieran. En este tipo de análisis será posible dar respuesta a cuestiones como:

- Con base al factor motivacional y psicosocial, en qué alumnos la probabilidad de necesitar asesoría psicopedagógica es mayor.
- Con base al rendimiento previo, en qué alumnos la probabilidad de necesitar asesorías académicas es mayor.

- Con base al factor socioeconómico y sociodemográfico, en qué alumnos la probabilidad de necesitar una beca económica o en especie es mayor.

Ventajas de la propuesta metodológica.

- Se realiza un análisis detallado para identificar y controlar observaciones influyentes que, de otra manera, causarían una estimación sesgada y afectarían en gran medida la capacidad predictiva del modelo. Se mide además, el efecto que tienen estas observaciones en la estimación del modelo.
- Se comprueba, tanto de manera gráfica como analítica, si la forma funcional especificada para el modelo es correcta.
- Se comprueba, tanto de manera gráfica como analítica, la existencia del supuesto de homocedasticidad.
- Se comprueba de manera gráfica la existencia del supuesto de normalidad.
- Se hace uso de la matriz de proporciones de varianza descompuesta o matriz π , propuesta por Belsley *et al.*, 1980, como medida de diagnóstico de multicolinealidad. A diferencia de otras herramientas estadísticas, como la matriz de correlación, la matriz π permite identificar correlaciones entre más de dos variables independientes, así como el total de relaciones lineales dependientes en todo el conjunto de regresores.

Limitaciones de la propuesta metodológica.

- Cabe la posibilidad de que, en la especificación primaria de la forma del modelo, se utilicen demasiados regresores. En la metodología no se establece alguna acción a tomar en cuanto a un exceso de regresores.
- En caso de comprobarse una especificación incorrecta de la forma funcional, la metodología no establece qué tipo de relación es la más adecuada entre la variable dependiente y las independientes.
- Para el análisis de variables dependientes binarias, sólo se considera el uso de modelos de probabilidad lineal MPL.

- Para la estimación del modelo de regresión lineal MRL, aún en presencia de multicolinealidad, sólo se considera el método de mínimos cuadrados ordinarios MCO.
- El procedimiento para la identificación de observaciones influyentes mide el impacto *individual* de cada observación. Por lo que no se considera una posible influencia conjunta ocasionada por dos observaciones o más que, de manera individual, no tienen comportamiento influyente.

Investigaciones futuras.

- *Jerarquización de las instituciones educativas con base en el valor añadido a sus estudiantes.*

En el primer capítulo se presentó el término de *eficacia escolar* acuñado por Murillo (2003), el cual corresponde a una línea de investigación que busca conocer, entre otros aspectos, qué capacidad tienen las escuelas para incidir en el desarrollo de los alumnos. Con base en Murillo (2003), de entre otras características, una escuela se puede clasificar como eficaz si el valor que le aporta la institución al alumno es alto, y si dicho valor, además, es agregado a cualquier alumno, no importando sus diferencias socio-económicas, lingüísticas, de género, o de cualquier otra índole.

Al respecto cabe preguntarse, ¿cómo medir el valor que le aporta una institución educativa a sus estudiantes?, ¿cómo saber si el valor agregado es alto o bajo?, ¿cómo identificar si dicho valor es agregado independientemente, o a pesar de, diferencias en el rendimiento previo de los alumnos, diferencias en los niveles de motivación, etc.?

Con base en la investigación realizada en este trabajo, una forma de dar respuesta a estas interrogantes es a través del **análisis de residuales** y la **estimación de MRL**.

En consecuencia, de este trabajo de tesis se desprende una línea de investigación futura que puede enfocarse en la jerarquización de las instituciones educativas con base en el valor añadido a sus estudiantes, medido a través de los residuales obtenidos de la estimación de un modelo de regresión lineal, en el cual se utilice como variable dependiente alguna de las propuestas en la tabla 3.5, y como variables independientes algunas de las propuestas en las tablas 3.1, 3.2 y 3.3.

- *Diseño de metas óptimas a través de la programación lineal.*

En la justificación para el desarrollo de este trabajo se hizo mención de la importancia que tiene medir el impacto de diversos factores en el rendimiento

académico, con la finalidad de definir metas e indicadores que permitan evaluar y/o crear procesos de mejora continua para las instituciones educativas.

El alcance de este trabajo de investigación se limitó a una forma de medir dicho impacto, sin involucrarse en la formulación de metas e indicadores. En consecuencia, de este trabajo de tesis se desprende una segunda línea de investigación futura que puede enfocarse en la formulación de metas, haciendo uso de la programación lineal.

Del modelo de regresión lineal estimado:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \hat{\mu}$$

Y considerando que éste se considera lineal en sus parámetros, cabe analizar si para modelos *nivel – nivel* como el de la ecuación anterior, es posible considerar a la ecuación estimada como la **función objetivo** de un problema de programación lineal, en el que se busque maximizar o minimizar la respuesta en la variable dependiente \hat{y} , en donde las variables regresoras funjan como **variables de decisión**.

Habiendo establecido las **restricciones** económicas o de cualquier otra índole, existentes para las variables de decisión (antes variables regresoras o independientes), en caso de cumplir con los supuestos del método de solución del programa lineal (por ejemplo, del algoritmo simplex), las soluciones encontradas para las variables de decisión, indicarían los niveles a los cuales dichas variables deben encontrarse para que el avance promedio en créditos de los estudiantes (por ejemplo) sea el mayor posible, dadas las restricciones existentes.

En caso de que lo anterior fuera viable, se tendría la información necesaria (el máximo o menor valor posible para la variable de interés relacionada con la medición del rendimiento académico y, los niveles idóneos de factores que influyen

en la variación de dicha variable) para establecer **metas óptimas**²³ y, en consecuencia, indicadores de desempeño.

²³ Es decir, metas lo más ambiciosas posible que a la vez sigan siendo realistas.

Referencias.

- Aiteco Consultores. (2016). *Artículos: Lo que no se mide, no se puede mejorar*. Obtenido de Aiteco Consultores Web site: <https://www.aiteco.com/lo-que-no-se-mide/>
- Alvarado Guerrero, I. R., Cepeda Islas, M. L., Del Bosque Fuentes, A. E., & Vega Valero, Z. (2014). Comparación de estrategias de estudio y autorregulación en universitarios. *Revista Electrónica de Investigación Educativa*, XVI(1). Recuperado el 15 de Marzo de 2017, de <https://redie.uabc.mx/redie/article/view/730/897>
- Asturias, M. Á. (1967). (E. Poniatowska, Entrevistador) Recuperado el 22 de Octubre de 2017, de <http://www.jornada.unam.mx/2017/10/22/opinion/a03a1cul>
- Bailey, M., Taasoobshirazi, G., & Carr, M. (23 de Abril de 2014). A Multivariate Model of Achievement in Geometry. *The Journal of Educational Research*, CVII(6), 440-461. Recuperado el 13 de Febrero de 2017, de <http://www.tandfonline.com/doi/citedby/10.1080/00220671.2013.833073?scroll=top&needAccess=true>
- Bárceñas Escobar, M. (s.f.). *El perfil real del alumno de la Facultad de Ingeniería UNAM, en el momento del ingreso*. Facultad de Ingeniería, Universidad Nacional Autónoma de México, México.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New Jersey, Estados Unidos: John Wiley & Sons, Inc.
- Bolívar, A., Freires, S., González, E., & Salas, L. (Junio de 2015). *Distribuciones fundamentales en las metodologías inferenciales ("T" Student, Chi cuadrado y Fisher)*. Recuperado el 14 de Agosto de 2017, de Presentación en SlideShare: https://es.slideshare.net/jonatan0106/distribucion-de-fisher-jic cuadrado?qid=7cbb034a-9252-4259-b60e-f29aca4b876d&v=&b=&from_search=3

- Bonnefoy, J. C. (2006). Indicadores de Desempeño en el Sector Público. *Curso-Seminario "Políticas Presupuestarias y Gestión por Resultados"*. República Dominicana: CEPAL.
- Castejón, J. L. (2014). *Aprendizaje y rendimiento académico*. Alicante, España: Editorial Club Universitario.
- Chase Jr., C. W. (2013). *Demand-Driven Forecasting: A Structured Approach to Forecasting*. Hoboken, New Jersey, Estados Unidos: John Wiley & Sons, Inc.
- Conway, T., Mackay, S., & Yorke, D. (1994). Strategic Planning in Higher Education: Who Are the Customers? *The International Journal of Educational Management*, VIII(6), 29-36. Recuperado el 10 de Marzo de 2017, de [https://scholar.google.com.mx/scholar?hl=es&as_sdt=0%2C5&as_vis=1&q=s+tragic+planning+in+higher+education+tony+conway&btnG=](https://scholar.google.com.mx/scholar?hl=es&as_sdt=0%2C5&as_vis=1&q=s+trategic+planning+in+higher+education+tony+conway&btnG=)
- Coordinación de Evaluación Educativa. (2016). *Perfil de ingreso de los alumnos de la generación 2017*. Facultad de Ingeniería, Universidad Nacional Autónoma de México, Secretaría de Apoyo a la Docencia, México. Recuperado el 20 de Marzo de 2017, de http://www.ingenieria.unam.mx/evaluacioneducativa/informacion_primer_ingreso.html
- Crampton, F. E., & Thompson, D. C. (Noviembre de 2011). When money matters: School Infrastructure Funding and Student Achievement. *School Business Affairs*, LXXVII(10), 14-18. Recuperado el 22 de Marzo de 2017, de <https://eric.ed.gov/?q=when+money+matters&id=EJ967506>
- Cuellar Gaxiola, A., Monje Martínez, J., Osornio Castillo, L., & Valadez Nava, S. (Diciembre de 2008). Variables sociodemográficas que influyen en el rendimiento académico de estudiantes de medicina en la FESI-UNAM. *Revista Electrónica de Psicología Iztacala*, XI(4), 1-14. Recuperado el 3 de Marzo de 2017, de <http://revistas.unam.mx/index.php/repi/article/view/18591>

- Edel Navarro, R. (2003). El rendimiento académico: Concepto, Investigación y Desarrollo. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, REICE, I(2)*. Recuperado el 6 de Febrero de 2017, de <https://revistas.uam.es/index.php/reice/article/view/5354>
- Education Resources Information Center. (s.f.). Recuperado el 10 de Febrero de 2017, de ERIC Institute of Education Sciences Web site: <https://eric.ed.gov/>
- Escalante, C. A. (2017). *Segundo Informe de Actividades 2016*. Facultad de Ingeniería, Universidad Nacional Autónoma de México, México.
- Facultad de Ingeniería. (2016). Recuperado el 2 de Abril de 2017, de Sitio web de la Coordinación de Evaluación Educativa: <http://www.ingenieria.unam.mx/evaluacioneducativa/mision.html>
- Facultad de Ingeniería. (2016). Recuperado el 2 de Abril de 2017, de Sitio web de la Coordinación de Evaluación Educativa: <http://www.ingenieria.unam.mx/evaluacioneducativa/funciones.html>
- Ferris, C. D., Grubbs, F. E., & Chalmers, L. (Junio de 1946). Operating Characteristics for the Common Statistical Tests of Significance. *The Annals of Mathematical Statistics, XVII(2)*, 178-197. Recuperado el 20 de Enero de 2018, de http://www.jstor.org/stable/2236037?seq=1#page_scan_tab_contents
- Florey, C. (1 de Mayo de 1993). Sample size for beginners. *National Institutes of Health, 306(6886)*, 1181-1184. Recuperado el 25 de Agosto de 2017, de <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1677669/>
- Florez, N., Luna, M., Salazar Elena, R., & Valenti, G. (2009). *Factores asociados al logro educativo, un enfoque centrado en el estudiante*. Facultad Latinoamericana de Ciencias Sociales - SEDE México. México: FLACSO México.
- Gala-Emma Peñalba, E. (2003). Introducción. En F. J. Murillo, *La Investigación sobre Eficacia Escolar en Iberoamerica. Revisión internacional sobre el estado del arte*. (pág. 467). Bogotá, Colombia: Convenio Andrés Bello y el Centro de Investigación y Documentación Educativa del MECD español.

- Garbanzo, V., & Guiselle, M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Red de Revistas Científicas de América Latina, el Caribe, España y Portugal*, XXXI(1), 43-63.
- García Jiménez, M. V., Alvarado Izquierdo, J. M., & Jiménez Blanco, A. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema*, XII(2), 248-252. Recuperado el 15 de Marzo de 2017, de <http://www.psicothema.com/pdf/558.pdf>
- Gráfico de probabilidad normal*. (s.f.). Recuperado el 12 de Octubre de 2017, de Wikimedia Commons: https://es.wikipedia.org/wiki/Gr%C3%A1fico_de_probabilidad_normal
- Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Duxbury Press.
- Ibarra García, E., & Medina Mora, P. (s.f.). *Perfil de estudiantes que aprueban todas las materias en el primer semestre de acuerdo a sus hábitos de estudio*. Facultad de Ingeniería, Universidad Nacional Autónoma de México.
- Johnson, S. C. (Septiembre de 1967). Hierarchical Clustering Schemes. *Psychometrika*, XXXII(3).
- Kennard, R. W., & Stone, L. A. (Febrero de 1969). Computer Aided Design of Experiments. *Technometrics*, XI(1), 137-148. Recuperado el 12 de Diciembre de 2017, de <https://www.jstor.org/stable/1266770>
- Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1983). *Forecasting: Methods and Applications*. Jhon Wiley & Sons, Inc.
- Montgomery, D. C., & Runger, G. (2013). *Probabilidad y Estadística aplicadas a la Ingeniería*. México: Limusa Wiley.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis, Quinta Edición*. Hoboken, New Jersey, Estados Unidos: John Wiley & Sons, Inc.

- Murillo Torrecilla, F. J. (2016). *Estudios sobre eficacia escolar en Iberoamérica. 15 buenas investigaciones*. Bogotá, Colombia: Convenio Andrés Bello.
- Murillo, F. J. (2003). Una panorámica de la investigación iberoamericana sobre eficacia escolar. *Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en la Educación*, I(1). Recuperado el 20 de Febrero de 2017, de <http://www.ice.deusto.es/rinace/reice/vol1n1/Murillo.pdf>
- Murillo, F. J. (2007). *Investigación Iberoamericana sobre eficacia escolar*. Bogotá, Colombia: Convenio Andrés Bello.
- Narayanachar Tattar, P., Ramaiah, S., & Manjunath, B. (2016). *A course in statistics with R*. Chichester, West Sussex, Reino Unido: John Wiley & Sons, Ltd.
- Pérez López, C. (2007). *Econometría Básica, técnicas y herramientas*. Madrid, España: Pearson Educación, S.A.
- Planck, B. (2014). Factores determinantes del rendimiento académico de los estudiantes de la Universidad de Atacama. *Revista Estudios Pedagógicos*, XV(1), 25-39.
- Pusser, B. (2014). Forces in Tension: The State, Civil Society and Market in the Future of the University. *Revista de la Educación Superior*, XLIII(170), 9-35. Recuperado el 15 de Marzo de 2017, de http://publicaciones.anuias.mx/pdfs/revista/Revista170_S2A1EN.pdf
- R Core, T. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Viena, Austria. Obtenido de <https://www.R-project.org/>
- Reyes Carreto, R., Godinez Jaimes, F., Ariza Hernández, F. J., Sánchez Rosas, F., & Torreblanca Ignacio, O. F. (2014). Un modelo empírico para explicar el desempeño académico de estudiantes de bachillerato. *Revista Perfiles Educativos*, XXXVI(146). Recuperado el 13 de Febrero de 2017, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-26982014000400004

- Rodríguez-Ayán, M. N. (2007). *Análisis Multivariado del Desempeño Académico de Estudiantes Universitarios de Química*. Madrid, España: Tesis de Doctorado, Facultad de Psicología, Universidad Autónoma de Madrid.
- Snee, R. D. (Enero de 1977). Validation of Regression Models: Methods and Examples. *Technometrics*, 415-428. Recuperado el 20 de Diciembre de 2017, de <http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1977.10489581#.Wo91t3biaUk>
- Tomás-Miquel, J.-V., Expósito-Langa, M., & Sempere-Castelló, S. (2014). Determinantes del rendimiento académico en los estudiantes de grado. Un estudio en administración y dirección de empresas. *Revista de Investigación Educativa*, XXXII(2), 379-392.
- Velázquez Velázquez, D. (1994). *Evaluación del Programa de Carrera de Ingeniero Civil de la ENEP Aragón, UNAM: con énfasis en los egresados (1980-1990)*. México: UNAM, Tesis de Maestría.
- Velázquez Velázquez, D. (2002). El Perfil del Personal Docente de Ingeniería. *Memorias de la XXIX Conferencia Nacional de Ingeniería*. Cancún, Quintana Roo.
- Velázquez Velázquez, D. (2016). Elementos que influyen en la formación de estudiantes de ingeniería. En O. E. Arango Tobón, *Ética profesional y responsabilidad social universitaria: universidad, sociedad y sujeto* (págs. 7-21). Medellín, Colombia: Fundación Universitaria Luis Amigó.
- Velázquez, V. D. (1999). *Determinación de las Variables Cuantitativas y Cualitativas que influyen en el Rendimiento Académico de los Alumnos de la Carrera de Ingeniería Civil del Campus Aragón-UNAM*. México: Informe anual de trabajo, UNAM.
- Wooldridge, J. M. (2015). *Introducción a la econometría. Quinta edición*. México: Cengage Learning Editores, S.A. de C.V.

Anexo A. Supuestos del modelo de regresión lineal.

Existen un conjunto de supuestos sólo aplicables para datos de corte transversal (aunque con muchas semejanzas para otro tipo de datos) bajo los cuales el estimador obtenido por MCO se vuelve un estimador insesgado del parámetro poblacional.

Los primeros cinco, conocidos como *supuestos de Gauss-Markov* (Wooldridge, 2015) son los siguientes:

Supuesto 1. (Lineal en los parámetros)

El modelo poblacional se expresa como:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \mu$$

donde β_0, \dots, β_k son los parámetros poblacionales desconocidos de interés y μ es un error aleatorio o término de perturbación no observable.

Supuesto 2. (Muestreo aleatorio)

Se tiene una muestra aleatoria de n observaciones, $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, 2, \dots, n\}$ que sigue el modelo poblacional del primer supuesto.

Supuesto 3. (No hay colinealidad perfecta)

En la muestra (y por tanto en la población) ninguna de las variables independientes es constante y no hay ninguna relación *lineal* exacta entre las variables independientes.

Supuesto 4. (Media condicional cero)

El valor esperado del error μ , dados los valores de las variables independientes, es cero. En otras palabras:

$$E(\mu | x_1, x_2, \dots, x_k) = 0$$

Bajo estos cuatro supuestos se establece el siguiente teorema (Wooldridge, 2015):

TEOREMA A.1. Insesgamiento de los estimadores de MCO

El valor esperado de una estimación por MCO es:

$$E(\widehat{\beta}_j) = \beta_j \quad j = 0, 1, \dots, k,$$

para cualquier valor del parámetro poblacional β_j . En otras palabras, los estimadores de MCO son estimadores **insesgados** de los parámetros poblacionales.

Al hablar de una estimación insesgada por MCO, se hace referencia al *procedimiento* mediante el cual se obtienen las estimaciones cuando se le considera aplicado a todas las muestras aleatorias posibles (Wooldridge, 2015).

Una vez establecida la manera por la que la estimación por MCO es insesgada, falta analizar la *varianza* de estos estimadores $Var(\widehat{\beta}_j)$. Esto servirá para tener una medida de dispersión en su distribución de muestreo. Para obtenerla, es necesario agregar un quinto supuesto:

Supuesto 5. (Homocedasticidad)

Dado cualquier valor de las variables explicativas, el error μ tiene la misma varianza. En otras palabras:

$$Var(\mu|x_1, \dots, x_k) = \sigma^2$$

Este quinto supuesto establece que la varianza en el término de error μ , condicional en las variables explicativas, es la *misma* para todas las combinaciones de valores de las variables independientes (Wooldridge, 2015).

Bajo los cinco supuestos anteriores, es posible establecer una manera de calcular la varianza de los estimadores por MCO:

TEOREMA A.2. Varianza de Muestreo de los Estimadores de Pendiente de MCO

Bajo los *supuestos de Gauss-Markov* condicionales en los valores muestrales de las variables independientes,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{STC_j(1 - R_j^2)}$$

para $j = 1, 2, \dots, k$ donde $STC_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ es la variación muestral total en x_j , R_j^2 es la R cuadrada de regresión de x_j sobre todas las otras variables independientes (incluyendo un intercepto) y σ^2 es la varianza poblacional (Wooldridge, 2015).

Sobre la magnitud de $Var(\hat{\beta}_j)$ cabe destacar que una varianza grande significa un estimador menos preciso, esto se traduce en intervalos de confianza grandes y pruebas de hipótesis menos exactas (Wooldridge, 2015).

Con el teorema A.2 se puede observar que a una mayor correlación entre variables independientes ($R_j^2 \rightarrow 1$) la varianza del estimador tiende a infinito ($Var(\hat{\beta}_j) \rightarrow \infty$). Este problema se define en el apartado *Multicolinealidad* del Capítulo II.

Conocer los primeros dos momentos (valor esperado y varianza) no es suficiente para justificar pruebas de inferencia estadística, ya que para esto se requiere conocer toda la distribución muestral de los $\hat{\beta}_j$ (Wooldridge, 2015). Por consiguiente, se debe establecer un supuesto más que permita realizar inferencias sobre la distribución muestral de los estimadores de MCO.

Supuesto 6. (Normalidad)

El error poblacional μ se encuentra distribuido normalmente, con media cero y varianza constante; es decir: $\mu \sim Normal(0, \sigma^2)$.

El supuesto de normalidad es el más *fuerte* de todos. En conjunto con los supuestos de Gauss-Markov, a los seis supuestos presentados en este apartado se les conoce como **supuestos del modelo lineal clásico**. De manera que un modelo construido bajo estos seis supuestos es conocido como un **modelo lineal clásico** (Wooldridge, 2015).

Anexo B. Estimación de la varianza poblacional.

Uno de los parámetros necesarios para obtener la varianza de los estimadores de MCO es la varianza poblacional σ^2 . Sin embargo, al ser un parámetro poblacional éste no se conoce con certeza y por tanto se deberá de estimar.

Un estimador insesgado natural de la varianza poblacional sería el promedio muestral de los errores cuadrados: $n^{-1} \sum_{i=1}^n \mu_i^2$. Sin embargo, este estimador no resulta ser viable ya que el factor μ_i es un *factor inobservable* y por tanto desconocido. Por consiguiente, un estimador insesgado de la varianza poblacional se deberá calcular a través de un estimador insesgado del error o factor inobservable; es decir, del residual $\hat{\mu}_i^2$:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\mu}_i^2}{(n - k - 1)}$$

Donde $\hat{\sigma}^2$ es el estimador de la varianza poblacional, $\sum_{i=1}^n \hat{\mu}_i^2$ es el promedio muestral de los residuales al cuadrado y el término $(n - k - 1)$ son los grados de libertad.

A la raíz cuadrada de la expresión del Teorema A.2 (es decir, la varianza de $\hat{\beta}_j$) calculada con σ^2 se le conoce como **desviación estándar de $\hat{\beta}_j$** . Sin embargo, dicho valor jamás será conocido (dado que σ^2 se desconoce).

Al sustituir σ^2 por $\hat{\sigma}^2$ en $Var(\hat{\beta}_j)$ y obtener la raíz cuadrada de ésta última, se obtiene el **error estándar de $\hat{\beta}_j$** (el cual funge como estimador de la desviación estándar de $\hat{\beta}_j$):

$$ee(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{STC_j(1 - R_j^2)}}$$

Cabe destacar un aspecto importante, la obtención de la varianza del estimador $\hat{\beta}_j$ como su respectivo error estándar, se basa en el Supuesto 5 del Anexo A, que refiere a la *Homocedasticidad*.

Para el caso en el que el error μ no presente la misma varianza para cualesquiera fueran los valores de las variables independientes (es decir, que exista *Heterocedasticidad*) el Supuesto 5 no se cumplirá y entonces el procedimiento para obtener $Var(\hat{\beta}_j)$ y $ee(\hat{\beta}_j)$ ya no serán válidos.

Anexo C. Forma matricial de la estimación por MCO.

El enfoque matricial de la estimación por MCO facilita el entendimiento de las operaciones involucradas y disminuye el espacio requerido por las expresiones matemáticas.

Un modelo de regresión múltiple puede expresarse como:

$$Y = X\beta + \mu$$

donde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{y} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

El vector Y de orden $(n \times 1)$ representa a las observaciones, X es una matriz $(n \times p)$ de los niveles de las variables independientes, en donde $p = k + 1$, β es un vector $(p \times 1)$ de los coeficientes de regresión, y μ es un vector $(n \times 1)$ de los residuales (Montgomery y Runger, 2013).

La finalidad es encontrar el vector que minimice al residual cuadrado:

$$L = \sum_{i=1}^n \mu_i^2 = \mu^T \mu = (Y - X\beta)^T (Y - X\beta)$$

El estimador de mínimos cuadrados $\hat{\beta}$ es la solución para β en las ecuaciones (Montgomery y Runger, 2013):

$$\frac{\partial L}{\partial \beta} = 0$$

El sistema de ecuaciones escalar en (2.7) se representa matricialmente como:

$$\mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{Y}$$

Al despejar al estimador de MCO se tiene:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Es importante hacer notar la estructura de la matriz $\mathbf{X}^T \mathbf{X}$ en la que los elementos de su diagonal son las sumas de los cuadrados de los elementos de las columnas de \mathbf{X} , y los elementos que están fuera de la diagonal son las sumas de los productos cruzados de los elementos de las columnas de \mathbf{X} ; por su parte, los elementos de $\mathbf{X}^T \mathbf{Y}$ son las sumas de los productos cruzados de las columnas de \mathbf{X} y las observaciones y_i (Montgomery y Runger, 2013).

Por su parte, el modelo ajustado en notación matricial se expresa como:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Al sustituir el valor de $\hat{\boldsymbol{\beta}}$ en la ecuación anterior se tiene:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{Y}$$

La matriz \mathbf{H} transforma los valores observados en un vector de valores ajustados $\hat{\mathbf{y}}$ (Montgomery y Runger, 2013). La diagonal h_{ii} de esta matriz; es decir:

$$h_{ii} = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

tiene gran importancia en la determinación de observaciones influyentes.

Anexo D. Funciones del lenguaje de programación estadístico R.

El uso del lenguaje de programación estadístico R permite calcular las operaciones matemáticas presentadas y discutidas en este trabajo, de manera ágil y eficiente, disminuyendo el tiempo requerido de cálculo a cuestión de segundos.

A continuación, se presentan las funciones que permiten realizar dichas operaciones en este lenguaje de programación. Su descripción ha sido obtenida con el uso de la función `help()` en la consola de RStudio.

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although [aov](#) may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Arguments

<code>formula</code>	an object of class " formula " (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.
<code>data</code>	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in <code>data</code> , the variables are taken from <code>environment(formula)</code> , typically the environment from which <code>lm</code> is called.
<code>subset</code>	an optional vector specifying a subset of observations to be used in the fitting process.
<code>weights</code>	an optional vector of weights to be used in the fitting process. Should be <code>NULL</code> or a numeric vector. If non- <code>NULL</code> , weighted least squares is used with weights <code>weights</code> (that is, minimizing $\sum(w \cdot e^2)$); otherwise ordinary least squares is used. See also 'Details',
<code>na.action</code>	a function which indicates what should happen when the data contain <code>NA</code> s. The default is set by the <code>na.action</code> setting of options , and is na.fail if that is unset. The 'factory-fresh' default is na.omit . Another possible value is <code>NULL</code> , no action. Value na.exclude can be useful.

<code>method</code>	the method to be used; for fitting, currently only <code>method = "qr"</code> is supported; <code>method = "model.frame"</code> returns the model frame (the same as with <code>model = TRUE</code> , see below).
<code>model, x, y, qr</code>	logicals. If <code>TRUE</code> the corresponding components of the fit (the model frame, the model matrix, the response, the QR decomposition) are returned.
<code>singular.ok</code>	logical. If <code>FALSE</code> (the default in S but not in R) a singular fit is an error.
<code>contrasts</code>	an optional list. See the <code>contrasts.arg</code> of model.matrix.default .
<code>offset</code>	this can be used to specify an <i>a priori</i> known component to be included in the linear predictor during fitting. This should be <code>NULL</code> or a numeric vector of length equal to the number of cases. One or more offset terms can be included in the formula instead or as well, and if more than one are specified their sum is used. See model.offset .
<code>...</code>	additional arguments to be passed to the low level regression fitting functions (see below).

Details

Models for `lm` are specified symbolically. A typical model has the form `response ~ terms` where `response` is the (numeric) response vector and `terms` is a series of terms which specifies a linear predictor for `response`. A terms specification of the form `first + second` indicates all the terms in `first` together with all the terms in `second` with duplicates removed. A specification of the form `first:second` indicates the set of terms obtained by taking the interactions of all terms in `first` with all terms in `second`. The specification `first*second` indicates the cross of `first` and `second`. This is the same as `first + second + first:second`.

If the formula includes an [offset](#), this is evaluated and subtracted from the response.

If `response` is a matrix a linear model is fitted separately by least-squares to each column of the matrix.

See [model.matrix](#) for some further details. The terms in the formula will be re-ordered so that main effects come first, followed by the interactions, all second-order, all third-order and so on: to avoid this pass a `terms` object as the formula (see [aov](#) and `demo(glm.vr)` for an example).

A formula has an implied intercept term. To remove this use either `y ~ x - 1` or `y ~ 0 + x`. See [formula](#) for more details of allowed formulae.

Non-NULL `weights` can be used to indicate that different observations have different variances (with the values in `weights` being inversely proportional to the variances); or equivalently, when the elements of `weights` are positive integers w_i , that each response y_i is the mean of w_i unit-weight observations (including the case that there are w_i observations equal to y_i and the data have been summarized).

`lm` calls the lower level functions [lm.fit](#), etc, see below, for the actual numerical computations. For programming only, you may consider doing likewise.

All of `weights`, `subset` and `offset` are evaluated in the same way as variables in `formula`, that is first in `data` and then in the environment of `formula`.

Value

`lm` returns an object of [class](#) "lm" or for multiple responses of class `c("mlm", "lm")`.

The functions `summary` and [anova](#) are used to obtain and print a summary and analysis of variance table of the results. The generic accessor functions `coefficients`, `effects`, `fitted.values` and `residuals` extract various useful features of the value returned by `lm`.

An object of class "lm" is a list containing at least the following components:

<code>coefficients</code>	a named vector of coefficients
<code>residuals</code>	the residuals, that is response minus fitted values.
<code>fitted.values</code>	the fitted mean values.
<code>rank</code>	the numeric rank of the fitted linear model.
<code>weights</code>	(only for weighted fits) the specified weights.
<code>df.residual</code>	the residual degrees of freedom.
<code>call</code>	the matched call.
<code>terms</code>	the terms object used.
<code>contrasts</code>	(only where relevant) the contrasts used.
<code>xlevels</code>	(only where relevant) a record of the levels of the factors used in fitting.
<code>offset</code>	the offset used (missing if none were used).
<code>y</code>	if requested, the response used.
<code>x</code>	if requested, the model matrix used.
<code>model</code>	if requested (the default), the model frame used.
<code>na.action</code>	(where relevant) information returned by model.frame on the special handling of NAs.

In addition, non-null fits will have components `assign`, `effects` and (unless not requested) `qr` relating to the linear fit, for use by extractor functions such as `summary` and [effects](#).

Regression Deletion Diagnostics

Description

This suite of functions can be used to compute some of the regression (leave-one-out deletion) diagnostics for linear and generalized linear models discussed in Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), etc.

Usage

```
influence.measures(model)

rstandard(model, infl = lm.influence(model, do.coef = FALSE),
           sd = sqrt(deviance(model)/df.residual(model)),
           type = c("sd.1", "predictive"), ...)

rstudent(model, infl = lm.influence(model, do.coef = FALSE),
          res = infl$wt.res, ...)

dffits(model, infl = , res = )

dfbeta(model, infl = lm.influence(model, do.coef = TRUE), ...)

dfbetas(model, infl = lm.influence(model, do.coef = TRUE), ...)

covratio(model, infl = lm.influence(model, do.coef = FALSE),
          res = weighted.residuals(model))

cooks.distance(model, infl = lm.influence(model, do.coef = FALSE),
               res = weighted.residuals(model),
               sd = sqrt(deviance(model)/df.residual(model)),
               hat = infl$hat, ...)

hatvalues(model, infl = lm.influence(model, do.coef = FALSE), ...)

hat(x, intercept = TRUE)
```

Arguments

<code>model</code>	an R object, typically returned by lm or glm .
<code>infl</code>	influence structure as returned by lm.influence or influence (the latter only for the <code>glm</code> method of <code>rstudent</code> and <code>cooks.distance</code>).
<code>res</code>	(possibly weighted) residuals, with proper default.
<code>sd</code>	standard deviation to use, see default.
<code>hat</code>	hat values $H[i, i]$, see default.
<code>type</code>	type of residuals for <code>rstandard</code> , with different options and meanings for <code>lm</code> and <code>glm</code> . Can

be abbreviated.
x the X or design matrix.
intercept should an intercept column be prepended to x?
... further arguments passed to or from other methods.

Details

The primary high-level function is `influence.measures` which produces a class "infl" object tabular display showing the DFBETAS for each model variable, DFFITS, covariance ratios, Cook's distances and the diagonal elements of the hat matrix. Cases which are influential with respect to any of these measures are marked with an asterisk.

The functions `dfbetas`, `dffits`, `covratio` and `cooks.distance` provide direct access to the corresponding diagnostic quantities. Functions `rstandard` and `rstudent` give the standardized and Studentized residuals respectively. (These re-normalize the residuals to have unit variance, using an overall and leave-one-out measure of the error variance respectively.)

Values for generalized linear models are approximations, as described in Williams (1987) (except that Cook's distances are scaled as F rather than as chi-square values). The approximations can be poor when some cases have large influence.

The optional `infl`, `res` and `sd` arguments are there to encourage the use of these direct access functions, in situations where, e.g., the underlying basic influence measures (from `lm.influence` or the generic `influence`) are already available.

Note that cases with `weights == 0` are *dropped* from all these functions, but that if a linear model has been fitted with `na.action = na.exclude`, suitable values are filled in for the cases excluded during fitting.

For linear models, `rstandard(*, type = "predictive")` provides leave-one-out cross validation residuals, and the "PRESS" statistic (**P**REdictive **S**um of **S**quares, the same as the CV score) of model `model` is

```
PRESS <- sum(rstandard(model, type="pred")^2)
```

The function `hat()` exists mainly for S (version 2) compatibility; we recommend using `hatvalues()` instead.

Student's t-Test

Description

Performs one and two sample t-tests on vectors of data.

Usage

```
t.test(x, ...)

## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)

## S3 method for class 'formula'
t.test(formula, data, subset, na.action, ...)
```

Arguments

<code>x</code>	a (non-empty) numeric vector of data values.
<code>y</code>	an optional (non-empty) numeric vector of data values.
<code>alternative</code>	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter.
<code>mu</code>	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
<code>paired</code>	a logical indicating whether you want a paired t-test.
<code>var.equal</code>	a logical variable indicating whether to treat the two variances as being equal. If <code>TRUE</code> then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.
<code>conf.level</code>	confidence level of the interval.
<code>formula</code>	a formula of the form <code>lhs ~ rhs</code> where <code>lhs</code> is a numeric variable giving the data values and <code>rhs</code> a factor with two levels giving the corresponding groups.
<code>data</code>	an optional matrix or data frame (or similar: see model.frame) containing the variables in the formula <code>formula</code> . By default the variables are taken from <code>environment(formula)</code> .
<code>subset</code>	an optional vector specifying a subset of observations to be used.
<code>na.action</code>	a function which indicates what should happen when the data contain <code>NA</code> s. Defaults to <code>getOption("na.action")</code> .
<code>...</code>	further arguments to be passed to or from methods.

Details

The formula interface is only applicable for the 2-sample tests.

`alternative = "greater"` is the alternative that `x` has a larger mean than `y`.

If `paired` is `TRUE` then both `x` and `y` must be specified and they must be the same length. Missing values are silently removed (in pairs if `paired` is `TRUE`). If `var.equal` is `TRUE` then the pooled estimate of the variance is used. By default, if `var.equal` is `FALSE` then the variance is estimated separately for both groups and the Welch modification to the degrees of freedom is used.

If the input data are effectively constant (compared to the larger of the two means) an error is generated.

Value

A list with class `"htest"` containing the following components:

<code>statistic</code>	the value of the t-statistic.
<code>parameter</code>	the degrees of freedom for the t-statistic.
<code>p.value</code>	the p-value for the test.
<code>conf.int</code>	a confidence interval for the mean appropriate to the specified alternative hypothesis.
<code>estimate</code>	the estimated mean or difference in means depending on whether it was a one-sample test or a two-sample test.
<code>null.value</code>	the specified hypothesized value of the mean or mean difference depending on whether it was a one-sample test or a two-sample test.
<code>alternative</code>	a character string describing the alternative hypothesis.
<code>method</code>	a character string indicating what type of t-test was performed.
<code>data.name</code>	a character string giving the name(s) of the data.

A continuación, se presentan enlaces a la documentación existente en [rdocumentation.org](https://www.rdocumentation.org) de otros estadísticos mencionados en este trabajo:

Estadístico de Wald.

<https://www.rdocumentation.org/packages/aod/versions/1.3/topics/wald.test>

Prueba Breusch-Pagan.

<https://www.rdocumentation.org/packages/lmtest/versions/0.9-35/topics/bptest>