

Julio César Alonso

Introducción a los Modelos econométricos para un Científico de Datos en R

Primera Versión para comentarios

27 de marzo de 2019

Universidad Icesi

Introducción a los Modelos econométricos para un Científico de Datos en R / Julio C. Alonso.

p. cm.—(Wiley series in survey methodology)

“Wiley-Interscience.”

Includes bibliographical references and index.

ISBN XXXXXXXXXX (pbk.)

1. Modelos lineales—Methodology. 2. Social sciences—Research—Statistical methods. I. XXXXX II. Series.

HA31.2.S873 2007 001.4'33—dc22 2004044064

Índice general

1. Introducción	1
1.1. Consideraciones sobre la aproximación econométrica	1
1.2. R	3
1.2.1. Elementos Básicos	4

Parte I El modelo de regresión múltiple

2. Modelo de regresión múltiple	9
2.1. Introducción	10
2.2. El modelo de regresión múltiple	16
2.2.1. Supuestos	17
2.2.2. Método de mínimos cuadrados ordinarios (MCO)	18
2.2.3. Propiedades de los estimadores MCO	20
2.3. Práctica en R: Ley de Okun en Colombia	21
2.3.1. Lectura de datos	21
2.3.2. Estimación del modelo	23
2.4. Ejercicios	27
2.5. Apéndice	28
3. Inferencia y análisis de regresión	37
3.1. Introducción	38
3.2. Pruebas individuales sobre los parámetros	40
3.3. El ajuste del modelo (Fit del modelo)	43
3.4. Pruebas conjuntas sobre los parámetros	48
3.5. Prueba de Wald y su relación con la prueba F	51
3.6. Práctica en R: Explicando los rendimientos de una acción	52
3.6.1. Pruebas conjuntas sobre los parámetros	58
3.7. Apéndice	59

4. Comparación de Modelos	63
4.1. Introducción	64
4.2. Comparación de modelos empleando medidas de bondad de ajuste	64
4.3. Comparación de modelos empleando inferencia	66
4.3.1. Modelos anidados	66
4.3.2. Modelos no anidados	66
4.4. Práctica en R: Escogiendo el mejor modelo	68
4.4.1. Medidas de bondad de ajuste	70
4.4.2. Pruebas estadísticas	71
4.5. Ejercicios	73
5. Variables dummy	75
5.1. Introducción	76
5.2. Usos de las variables dummy	78
5.2.1. Caso I. La función es la misma	78
5.2.2. Caso II. Cambio en intercepto	79
5.2.3. Caso III. Cambio en pendiente	80
5.2.4. Caso IV. Cambio en intercepto y pendiente	81
5.3. Práctica en R	81
5.3.1. Relación entre la economía mundial y la colombiana	82
5.3.2. Creando variables dummy con el paquete dummies	86
6. Selección automática de modelos	89
6.1. Introducción	90
6.2. Empleando “fuerza bruta”	92
6.3. Empleando estrategias inteligentes de detección de un mejor modelo	98
6.3.1. Regresión paso a paso (Stepwise)	98
6.3.2. Stepwise Forward regression	99
6.3.3. Stepwise backward regression	105
6.3.4. Combinando forward y backward (step regression)	108
6.4. Pongamos todo junto	112
6.4.1. Eliminando automáticamente variables no significativas	113
6.4.2. Comparación de modelos	116
6.5. Comentarios finales	117

Parte II Problemas econométricos en los datos

7. Multicolinealidad	121
7.1. Introducción	122
7.2. Los diferentes grados de multicolinealidad	124
7.2.1. Multicolinealidad perfecta	125
7.2.2. Consecuencias de la multicolinealidad no perfecta	127
7.3. Pruebas para la detección de multicolinealidad	129
7.3.1. Factor de Inflación de Varianza (<i>VIF</i>)	130
7.3.2. Prueba de Belsley, Kuh y Welsh (1980)	130
7.4. Soluciones de la multicolinealidad (¡Sí se necesitan!)	131

7.4.1. Regresión de Ridge	131
7.4.2. Remover variables con alto <i>VIF</i>	131
7.5. Práctica en R: Análisis del efecto discriminatorio de género en las diferencias salariales en Colombia	132
7.5.1. Pruebas de multicolinealidad	134
7.5.2. Solución del problema removiendo variables con alto <i>VIF</i> ..	135
7.6. Ejercicios	136
7.7. Apéndice	137

Parte III Problemas econométricos en los datos

8. Heteroscedasticidad	143
8.1. Introducción	144
8.2. Pruebas para la detección de heteroscedasticidad	145
8.2.1. Prueba de Breusch-Pagan	147
8.2.2. Prueba de White	148
8.3. Solución a la heteroscedasticidad	148
8.3.1. Estimación Consistente en presencia de heteroscedasticidad de los errores estándar.	149
8.4. Práctica en R: Análisis del efecto discriminatorio de género en las diferencias salariales en Colombia	151
8.4.1. Análisis gráfico de los residuos	151
8.4.2. Pruebas de heteroscedasticidad	153
8.4.3. Solución al problema de heteroscedasticidad con HC	156
8.5. Apéndice	160
9. Autocorrelación	163
9.1. Introducción	164
9.2. Pruebas para la detección de autocorrelación	167
9.2.1. Prueba de Rachas (Runs test)	168
9.2.2. Prueba de Durbin-Watson	169
9.2.3. Prueba h de Durbin	171
9.2.4. Prueba de Box-Pierce y Ljung-Box	171
9.2.5. Prueba de Breusch-Godfrey	172
9.3. Solución a la autocorrelación	173
9.3.1. Estimación Consistente en presencia de Autocorrelación de los errores estándar.	174
9.4. Práctica en R: Explicando los rendimientos de una acción (continuación)	175
9.4.1. Construcción de la base de datos	175
9.4.2. Residuales del modelo y análisis gráfico de los residuales ..	177
9.4.3. Pruebas de Autocorrelación	179
9.4.4. Prueba de Rachas (Runs test)	180
9.4.5. Prueba de Durbin-Watson	181
9.4.6. Prueba de Box-Pierce y Ljung-Box	182

9.4.7. Prueba de Breusch-Godfrey	184
9.4.8. Solución al problema de heteroscedasticidad con H.A.C. ...	185
9.5. Ejercicios	189
9.6. Apéndice	189
Índice alfabético	193

Capítulo 1

Introducción

1.1. Consideraciones sobre la aproximación econométrica

La econometría es una disciplina que unifica las matemáticas, la estadística y la teoría económica con el objetivo de entender cuantitativamente las relaciones económicas **DefEconometrics**. Esta ha desarrollado unas técnicas estadísticas que le han permitido convertirse en una rama de la estadística. Si bien las técnicas que estudiaremos en este libro no son exclusivas de la econometría, la aproximación de la econometría provee una vía metodológica interesante para el científico de datos.

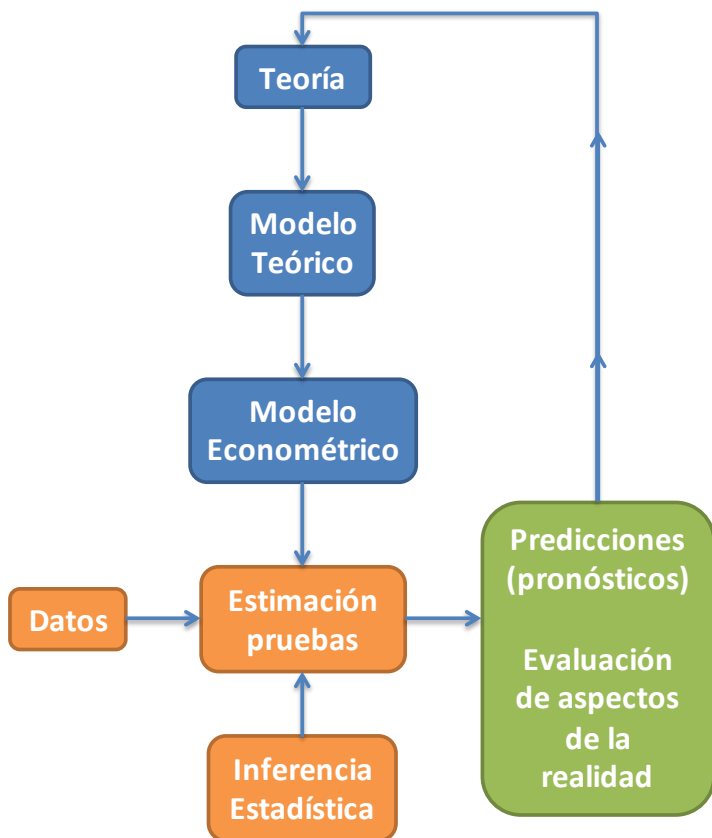
Para lograr su objetivo, una ruta metodológica que adopta la econometría es partir de un modelo teórico; es decir, de una formulación matemática de alguna teoría, como se muestra en la Figura 1.1.

Luego, transformamos este modelo teórico en un modelo estadístico o econométrico. Para ello, usamos algunos datos que, suponemos, representan adecuadamente las variables que conforman el modelo teórico y añadimos un término de error.

En un siguiente paso, estimamos el modelo usando métodos estadísticos que implican supuestos explícitos e implícitos. Si alguno de los supuestos del modelo no se cumple, debemos corregir el modelo al especificar adecuadamente el término de error o solucionando el problema que esté generando la violación del supuesto.

Una vez contamos con el modelo adecuado, podemos probar estadísticamente el cumplimiento o no de las restricciones planteadas por la teoría. Si el modelo que tenemos refleja el comportamiento real adecuadamente, podemos usar nuestras estimaciones para hacer predicciones y evaluaciones de los aspectos de la realidad relevantes para el científico de datos, en caso contrario, debemos empezar el proceso nuevamente, buscando un modelo teórico que explique adecuadamente las relaciones bajo estudio.

Sin embargo, esta metodología no tiene en cuenta cuál es el mecanismo que realmente generó los datos que observamos (el *Data Generating Process* o DGP), y que nos permitirían estimar el modelo econométrico. Así, surge una segunda ruta meto-

Figura 1.1 Primera ruta metodológica

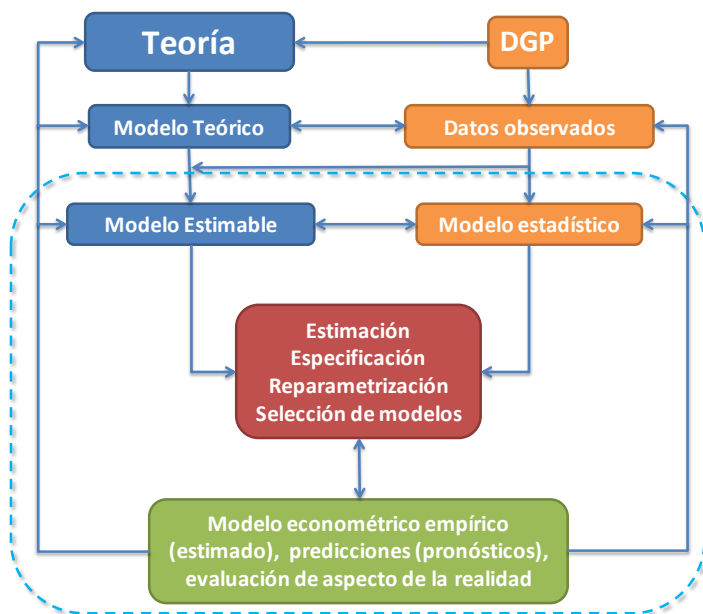
Adaptado a partir de Spanos86.

dológica menos tradicional pero más común a los científicos de datos que interpreta el modelo econométrico como una aproximación al verdadero DGP.

De acuerdo a esta segunda ruta metodológica, que se ilustra en la Figura 1.2, el modelo econométrico usa variables aleatorias para especificar una descripción de un mecanismo que generó los datos. Cuando el modelo teórico es el DGP y, por ello, el modelo econométrico difiere del teórico únicamente por razones puramente aleatorias, las dos metodologías coinciden.

Por ello, entenderemos el modelo teórico como una abstracción imperfecta que permite entender algún fenómeno particular. En este sentido, el modelo teórico no pretende explicar lo que ocurre en la realidad, sino que separa aquellos elementos que son útiles para explicar una situación específica.

Por supuesto, este modelo teórico está asociado con el verdadero DGP, no obstante, debido a que es imposible que el modelo teórico recoja toda la información

Figura 1.2 Segunda ruta metodológica

Adaptado a partir de Spanos86.

asociada al mecanismo que generó los datos, el modelo teórico no representa de forma precisa el DGP.

Asimismo, los datos que observamos en la realidad no siempre corresponden a las variables incluidas en el modelo teórico, sino que son muestras de datos generados por algún DGP real. De esta manera, en esta metodología no estimamos un modelo teórico al que le sumamos un término aleatorio de error, sino que estimamos un modelo estadístico, que tiene en cuenta la aleatoriedad real que dió lugar a los datos observados. Cuan parecidos sean el modelo econométrico y el modelo estadístico, depende de los datos disponibles y de la diferencia entre la teoría y el DGP real.

A lo largo de este libro, nos concentraremos en el área punteada de la Figura 1.2. La teoría y los modelos teóricos dependerán de cada problema específico bajo estudio del científico de datos y se encuentran por fuera del alcance de este libro.

1.2. R

En este libro emplearemos R, un lenguaje de programación estadístico que permite realizar desde cálculos muy sencillos, hasta programar diferentes rutinas que permiten realizar operaciones más complejas, a partir de líneas de código.

Este programa fue desarrollado por Kurt Hornik y es complementado frecuentemente por una amplia comunidad académica que contribuye al desarrollar nuevas y mejores funcionalidades para R.

Es un *software* particularmente popular entre la comunidad estadística y se distribuye gratuitamente bajo licencia GNU. Funciona en sistemas operativos Unix-Like, Windows y Mac.

Toda la información relacionada con R puede ser encontrada en CRAN (*Comprehensive R Archive Network*), a la cual se puede acceder a través de una serie de *mirrors* distribuidos en todo el mundo. Por ejemplo, la Universidad Icesi mantiene un *mirror* de R, al cual se accede a través de la página <http://www.icesi.edu.co/CRAN/>, y existen diferentes *mirrors* en los cinco continentes.

Para instalar el programa, puede ingresar a <http://cran.r-project.org/> y seguir las instrucciones que en esta página se muestran, de acuerdo a su sistema operativo. Debido a que la interfaz de R puede ser poco amigable con un usuario poco experimentado o relacionado con los códigos de programación, se han desarrollado diferentes interfaces para su visualización.

Ejemplos de ellas son R Commander, Tinn-R, Weka, RKWard, Deducer y RStudio. Todas ellas tienen licencia GNU y algunas están especialmente diseñadas para cumplir unas funciones específicas (como Deducer), mientras que otras, como Tinn-R o RStudio, no fueron pensadas para cumplir un conjunto de tareas específicas, por lo que facilitan el trabajo en R sin importar la tarea desarrollada¹.

1.2.1. Elementos Básicos

En R contamos con los siguientes elementos que nos permiten realizar diferentes cálculos y gráficos:

Script: Es un archivo en el cual podemos guardar las líneas de código que hemos escrito.

Consola: Es el espacio en el que se ejecutan las ordenes que damos a R a través de la línea de código. Es en este espacio en el que se muestran los resultados de los cálculos.

Workspace: Es un archivo en que se guardan los objetos creados en una sesión de R, como explicaremos más adelante.

Dispositivo Gráfico: Es la ventana en la que se muestran los gráficos realizados.

Generalmente, lo más conveniente es escribir y guardar el Script y, a partir de él ejecutar los comandos que sean necesarios. Estos comandos se dividen en dos grandes tipos:

Expresiones: Es una orden que se da a R. En general, son de la forma:
`comando(argumentos)`

¹ RStudio puede ser descargado desde <http://www.rstudio.com/ide/download/> para cualquier sistema operativo. Tinn-R sólo está disponible para Windows y puede descargarse desde <http://sourceforge.net/projects/tinn-r/>.

y le piden a R que ejecute inmediatamente la orden comando para los argumentos dados. Un comando básico es `help`, que sirve para visualizar la ayuda asociada a un comando y tiene como uno de sus argumentos el nombre del comando sobre el cual se quiere pedir ayuda. Por ejemplo, la línea `help(lm)` abre una ventana en un navegador con la página de ayuda para el comando `lm`. Si un comando tiene diferentes argumentos, estos se separan usando comas (,).

Asignaciones: Este tipo de comandos piden a R que guarde el resultado de ejecutar una orden en un elemento al que llamaremos objeto. Los objetos pueden recibir cualquier nombre alfa-numérico. Son de la forma `objeto<-comando(argumentos)` u `objeto=comando(argumentos)`. Así, después de realizar una asignación, en la consola de R no se muestra el resultado de comando. Por ejemplo, si asignamos el resultado de `help(lm)` al objeto `Ayudalm`, escribiendo la línea `Ayudalm<-help(lm)`, R no abre una ventana en el navegador hasta que se le pida que muestre el objeto `Ayudalm`. Para hacer esto, basta con escribir `Ayudalm`.

Vale la pena mencionar que R es sensible al uso de mayúsculas. Así, el objeto `Ayudalm` es diferente de `ayudalm`.

Los comandos de R se almacenan en librerías, que no son más que agrupaciones de comandos con fines similares. Para cargar una librería usamos el comando `library(nombre de la librería)`. Algunas de ellas se cargan automáticamente al abrir R, pero muchas otras no están instaladas por defecto.

Para instalar una librería, podemos usar el comando `install.packages('nombre de la librería')`. Probablemente, la primera vez que instalemos una librería, aparecerá un cuadro de diálogo en el que se nos pide que elegir el CRAN *mirror* más cercano. Una vez elegido, hacemos click en OK para finalizar el proceso de instalación.

Una vez instaladas estas librerías no es necesario volverlas a instalar. No obstante, para usarlas debemos pedir a R que las active en cada sesión. Así, para empezar a usar las librerías que anteriormente instalamos usamos las siguientes líneas:

Recuadro 1.1 Comandos básicos en R

Operación	Ejemplo	Comando
Crear un vector	$\mathbf{x} = [5, 15, 25, 3]$	<code>x<-c(5,15,25,3)</code>
Crear un vector con una secuencia	$\mathbf{x} = [1, 2, 3, 4, 5]$	<code>x<-1:5</code>
	$\mathbf{x} = [2, 4, 6, 8, 10]$	<code>x<-seq(2,10,by=2)</code>
Crear una matriz	$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$	<code>A<-matrix(c(1,2,3,4),</code>
Obtener el elemento i, j de una matriz	$\mathbf{A}_{1,2} = 3$	<code>A[1,2]</code>
Transponer una matriz	\mathbf{A}^T	<code>t(A)</code>
Hallar el determinante de una matriz	$\det(\mathbf{A})$	<code>det(A)</code>
Elegir una columna de la base de datos	Columna 3 de la base de datos Datos	<code>Datos[,3]</code>
	Si la variable X se encuentra en la columna 3	<code>Datos\$X</code>
Elegir una fila de la base de datos	Fila 5 de la base de datos Datos	<code>Datos[5,]</code>
Pedir ayuda	Ayuda sobre el comando seq	<code>help(seq)</code>

Parte I
El modelo de regresión múltiple

Capítulo 2

Modelo de regresión múltiple

Objetivos del capítulo

El lector, al finalizar este capítulo, estará en capacidad de:

- Estimar un modelo lineal con más de una variable explicativa empleando R .
- Identificar los coeficientes estimados por R.
- Transformar y crear variables en R.
- Interpretar los componentes de las tablas de salida que proporciona R.

2.1. Introducción

El científico de datos pocas veces se enfrenta con un problema bien definido en el que la teoría pueda aplicarse directamente, ya sea por la disponibilidad de información para construir determinada variable o por qué el problema no se encuentra acotado. Así típicamente éste se encuentra enfrentado a un problema en el que la variable a explicar (variable dependiente) es clara y existe un conjunto amplio de posibles variables explicativas. Antes de enfrentar ese problema un poco más complicado, veamos como es la aproximación clásica de construcción de modelos de las ciencias económicas. En este caso, la teoría es la base para la construcción de un modelo.

La teoría económica nos provee con diferentes relaciones entre variables económicas; por ejemplo, sabemos que la demanda de un bien Q depende del propio precio, del precio de los bienes relacionados (ya sean sustitutos o complementarios), del nivel de ingresos, de los gustos y del número de consumidores. Estas relaciones teóricas se pueden representar funcionalmente como:

$$Q = Q_x(p_x, p_{comp}, p_{sust}, I) \quad (2.1)$$

Además, la teoría nos permite conocer cuál sería el efecto de cambios de cada una de estas variables independientes, *ceteris paribus*, sobre las cantidades demandadas. En otras palabras, la teoría económica permite conocer el signo de $\frac{\partial Q_x}{\partial p_x}$, $\frac{\partial Q_x}{\partial p_{comp}}$, $\frac{\partial Q_x}{\partial p_{sust}}$, etc. Pero en general, la teoría no nos da indicios de la forma de la función de demanda; es decir, la teoría no puede brindar elementos para decidir entre diferentes formas funcionales. Por ejemplo,

$$Q_x(p_x, p_{comp}, p_{sust}, I) = A p_x^\alpha + \frac{1}{p_{comp}^\beta + C} + \ln(\varphi p_{sust}) + \alpha I \quad (2.2)$$

$$Q_x(p_x, p_{comp}, p_{sust}, I) = \beta_0 + \beta_1 p_x + \beta_2 p_{comp} + \beta_3 p_{sust} + \beta_4 I \quad (2.3)$$

Así, la teoría provee las relaciones funcionales teóricas, pero en general no provee la forma funcional explícita. Aún más, el carácter de la mayoría de las relaciones funcionales son determinísticas.¹ Relaciones funcionales como la expresada en la ecuación 2.1 corresponden a modelos matemáticos (exactos) y no estadísticos.

En la práctica se reconoce que las relaciones entre la variables independientes y la dependiente no tienen por qué ser exactas y se incluye un término aleatorio de error en los modelos económicos a estimar.² La inclusión de este término de error se justifica de diferentes formas. Por ejemplo,³

¹ No incluyen una variable aleatoria.

² Noten que si no existiera un término aleatorio en nuestro modelo, entonces estaríamos hablando de modelos determinísticos y no requeriríamos de métodos estadísticos para determinar los parámetros del modelo.

³ Para ver más justificaciones para la inclusión del término de error el lector puede consultar Gujarati, Damodar. 1997. *Econometría*: McGraw Hill. Pág. 38.

- Las respuestas humanas individuales a diferentes incentivos no son exactas y por tanto no son predecibles con total certidumbre; aunque se espera que en promedio si lo sean.
- En general, es imposible pretender que un modelo recoja todas y cada una de las variables que afectan directamente una variable, pues precisamente un modelo económico es una simplificación de la realidad y por tanto omite detalles de ella. Es importante anotar que en algunas ocasiones las variables que se omiten son conocidas, pero el investigador no cuenta con información para medir esas variables, por tanto deben ser omitidas.
- La variable dependiente puede estar medida con error, pues en la práctica los agregados económicos normalmente son estimados a partir de muestras. El error de medición en la variable dependiente estará recogido en el término de error, pues nuestro modelo no pretenderá explicar el error de medición sino el comportamiento promedio del agregado económico.
- En algunas oportunidades las relaciones entre variables anunciadas por la teoría económica son producto de un esfuerzo de resumir un conjunto de decisiones individuales. Así como las decisiones individuales son diferentes de individuo a individuo, cualquier intento por estimar estas relaciones a nivel agregado será simplemente una aproximación; por tanto la diferencia entre esta aproximación y el valor real será atribuida al término de error.

Entonces, todo modelo econométrico poseerá una parte aleatoria y una que no lo es (parte determinísticas). Por ejemplo, si se considera la relación funcional expresada en la ecuación 2.1, el correspondiente modelo estadístico corresponderá a $Q = Q_x(p_x, p_{comp}, p_{sust}, I) + \varepsilon$, donde $Q_x(p_x, p_{comp}, p_{sust}, I)$ corresponde a la porción determinística del modelo y ε representa el término aleatorio de error.

Como se mencionó antes, generalmente la teoría económica brinda las relaciones funcionales y de causalidad entre diferentes variables económicas. Pero la teoría no provee explícitamente el tipo de relación funcional entre las variables explicativas y la variable dependiente. En la práctica, los investigadores adoptan como estrategia tomar una primera aproximación a las relaciones funcionales expresadas en la teoría por medio de relaciones lineales o linealizables. En este capítulo concentraremos nuestra atención en los modelos lineales. Es decir, modelos de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (2.4)$$

En este modelo podemos distinguir varios componentes: la variable dependiente (Y_i), las variables independientes (X_{1i} y X_{2i}), los coeficientes o parámetros⁴ (β_0 , β_1 y β_2), y el término de error (ε_i). Adicionalmente, se presenta un subíndice i que representa que dicha relación se cumple para cualquier observación i que se realice. Típicamente i se encuentra entre 1 y n (el tamaño de la muestra); es decir, $i = 1, 2, \dots, n$. Es importante anotar que normalmente cuando se emplean datos de series de tiempo (o longitudinales) el subíndice i es cambiado por una t .

⁴ Los parámetros corresponden a números (constantes) que describen la relación entre las variables independientes y la variable dependiente. Más adelante se ampliará esta idea.

Antes de entrar en detalle, es importante aclarar qué se entiende por modelo lineal en este contexto. Para ser más precisos, cuando nos referimos a un modelo de regresión lineal, estamos hablando de un modelo que es lineal en sus parámetros y el término de error es aditivo. En otras palabras, los parámetros⁵ están multiplicando a las variables explicativas o representan un intercepto. Formalmente, *un modelo se considera un modelo lineal si la variable dependiente se puede expresar como una combinación lineal de las variables explicativas y un término de error, donde los parámetros son los coeficientes de la combinación lineal.*

Por ejemplo, el modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ es un modelo (estadístico) lineal, pues es lineal en los parámetros $\beta^T = (\beta_0, \beta_1, \beta_2)$; en este caso, estos representan un intercepto y pendientes (ver Ejemplo 1), y el término de error es aditivo. Por otro lado, un modelo como $Y_i = \beta_0 + (X_{1i})^{\beta_1} + \beta_2 X_{2i} + \varepsilon_i$ no es un modelo lineal, pues el modelo no es lineal respecto a β_1 , el cual representa una potencia.

Ahora bien, es importante resaltar que un modelo econométrico puede ser lineal en los parámetros y tener un término aleatorio aditivo, pero no representar una línea recta, un plano o sus equivalentes en dimensiones mayores. En estos casos, aunque no se trata de un modelo lineal desde el punto de vista matemático, aún tenemos un modelo estadístico lineal (Ejemplo 2.2).

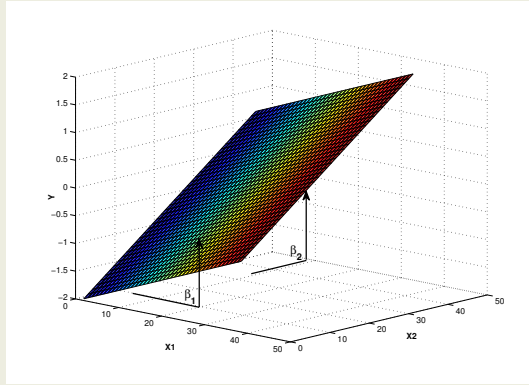
Ejemplo 2.1 Modelo de regresión lineal

Suponga la siguiente relación entre una variable dependiente (Y_i) y dos variables explicativas (X_{1i}, X_{2i}):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (2.5)$$

donde ε_i es un término aleatorio de error. Entendamos primero la naturaleza de esta relación funcional omitiendo el término aleatorio de error. En el siguiente gráfico se puede observar la función $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Noten que $\frac{\partial Y}{\partial X_1} = \beta_1$, es decir el coeficiente asociado a la variable independiente X_1 corresponden al cambio en la variable dependiente cuando la otra variable independiente se mantiene constante, es decir la pendiente de la función con respecto al plano formado por los ejes de X_1 y Y . Similar interpretación posee $\beta_2 = \frac{\partial Y_i}{\partial X_{2i}}$.

⁵ A excepción de la varianza del término de error.

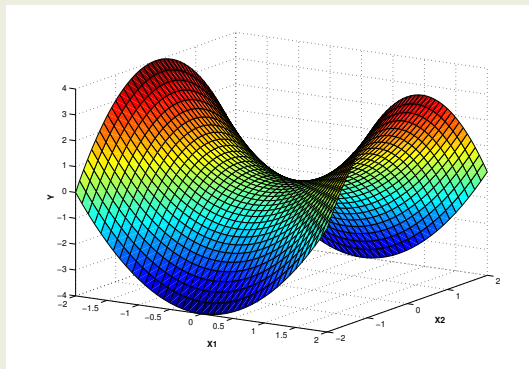


Ejemplo 2.2 Modelo intrínsecamente lineal

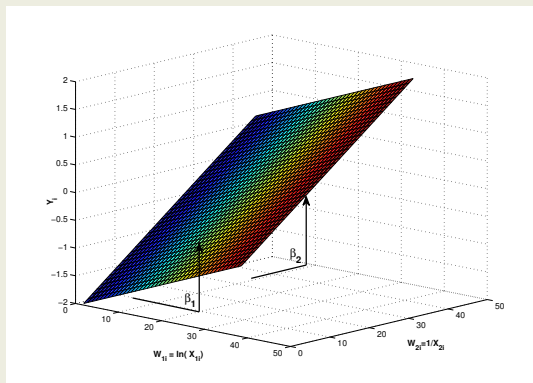
Suponga la siguiente relación entre una variable dependiente (Y_i) y dos variables explicativas (X_{1i}, X_{2i}):

$$Y_i = \alpha_1 + \alpha_2 \ln(X_{1i}) + \alpha_3 \left(\frac{1}{X_{2i}} \right) + \varepsilon_i \quad (2.6)$$

donde ε_i es un término aleatorio de error. En este caso, tenemos que el modelo es lineal en los parámetros α_1 , α_2 y α_3 , además el error es aditivo; por tanto este modelo es lineal. Es importante anotar que en este caso α_2 no es una pendiente, pues $\frac{\partial Y}{\partial X_1} = \frac{\alpha_2}{X_1}$; para α_3 ocurre algo similar. Por tanto este modelo no representa un plano (no es lineal desde el punto de vista matemático), como se puede observar en el siguiente gráfico en el cual se ha omitido el término de error, y por tanto solo se ha graficado la porción determinística del modelo



Pero esta función se puede expresar de tal forma que represente un plano; reparametrizando, podemos definir $W_{1i} = \ln(X_{1i})$ y $W_{2i} = \left(\frac{1}{X_{2i}}\right)$. Ahora, reemplazando estas dos nuevas variables en el modelo original, tenemos que $Y_i = \alpha_1 + \alpha_2 W_{1i} + \alpha_3 W_{2i} + \varepsilon_i$. Si graficamos esta nueva función, ignorando el término de error, en el espacio (Y, W_1, W_2) , se obtendrá lo siguiente:



Este modelo reparametrizado es un modelo lineal y representa un plano; y se puede estimar por medio de los métodos que estudiaremos en este capítulo.

Es importante anotar que la interpretación de los parámetros α_1 y α_2 es un poco diferente en este caso. Por ejemplo, α_2 representa el número de unidades en que cambiará la variable dependiente cuando W_1 cambia en una unidad, es decir cuando el $\ln(X_1)$ cambia en una unidad y no cuando X_1 cambia. Para interpretar α_2 en términos de X_1 es necesario derivar (2.6) con respecto a X_1 : $\frac{\partial Y_i}{\partial X_{1i}} = \frac{\alpha_2}{X_{1i}}$. Manipulando algebraicamente tenemos $\frac{\partial Y_i}{\frac{\partial X_{1i}}{X_{1i}}} = \alpha_2$. Multiplicando a ambos lados por $\frac{1}{100}$, obtenemos $\frac{\partial Y_i}{\frac{\partial X_{1i}}{X_{1i}} \times 100} = \frac{\alpha_2}{100}$. Es decir $\frac{\partial Y_i}{\Delta \% X_{1i}} = \frac{\alpha_2}{100}$. Por tanto, la interpretación de α_2 en términos de X_1 es la siguiente: cuando X_1 aumenta en uno por ciento, entonces Y_i aumentará en $\frac{\alpha_2}{100}$ unidades.

Existen otros modelos especiales que no son lineales en sus parámetros, pero mediante reparametrizaciones⁶ se convierten en modelos lineales. Estos se conocen con el nombre de *modelos linealizables* (o modelos intrínsecamente lineales) y también pueden ser estimados por los mismos métodos lineales de estimación de un modelo lineal.

⁶ Una reparametrización es un cambio de nombre de variables y/o parámetros del problema que mantiene la naturaleza de la relación inalterada

Ejemplo 2.3 Función de producción Cobb-Douglas

Una función de producción Cobb-Douglas se define como:

$$Y = AX_1^{\alpha_1} X_2^{\alpha_2} X_3^{\alpha_3} \quad (2.7)$$

donde Y corresponde a la producción, A denota el estado del conocimiento tecnológico actual, X_j representa la cantidad del insumo j empleado en el proceso productivo y α_j son los parámetros a estimar. Claramente, la expresión anterior no corresponde a un modelo estadístico, pues no presenta término aleatorio alguno. Como se discutió anteriormente, existen diferentes justificaciones para incluir un término estocástico de error en las relaciones que nos brinda la teoría económica. Así, un modelo econométrico a partir de la función de producción Cobb-Douglas sería:

$$Y_i = AX_{1i}^{\alpha_1} X_{2i}^{\alpha_2} X_{3i}^{\alpha_3} \varepsilon_i \quad (2.8)$$

donde $i = 1, 2, \dots, n$ y ε_i es un término aleatorio de error. Pero este modelo no es lineal, al no ser lineal en los parámetros. No obstante, éste se puede linealizar fácilmente; aplicando logaritmos a ambos lados de la expresión tendremos:

$$\ln(Y_i) = \ln(A) + \alpha_1 \ln(X_{1i}) + \alpha_2 \ln(X_{2i}) + \alpha_3 \ln(X_{3i}) + \ln(\varepsilon_i) \quad (2.9)$$

Definimos: $\beta = \ln(A)$ y $\mu_i = \ln(\varepsilon_i)$. Entonces la expresión anterior se puede reescribir de la siguiente forma:

$$\ln(Y_i) = \beta + \alpha_1 \ln(X_{1i}) + \alpha_2 \ln(X_{2i}) + \alpha_3 \ln(X_{3i}) + \mu_i \quad (2.10)$$

Un aspecto importante de este modelo es la interpretación singular de los coeficientes. En este caso, cada α_i corresponde a la elasticidad del producto con respecto al insumo i . Por otro lado, el intercepto β corresponde al valor esperado del logaritmo del valor de la producción cuando todos los insumos son iguales a una unidad. En otras palabras $A = e^\beta$. Así, el intercepto en este modelo tiene una interpretación difícil y en la mayoría de las aplicaciones carece de interpretación económica.

2.2. El modelo de regresión múltiple

Una vez se cuenta con un modelo lineal que representa la relación entre diferentes variables explicativas y la variable dependiente, se deseará conocer los valores de los parámetros del modelo. Para lograr este fin, se recopilan n observaciones de las variables explicativas y de la independiente. En general, un modelo lineal múltiple para el cual se cuenta con n observaciones y $k - 1$ variables explicativas está dado por:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \dots + \beta_k X_{k1} + \varepsilon_1 \\ y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \dots + \beta_k X_{k2} + \varepsilon_2 \\ &\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \dots + \beta_k X_{kn} + \varepsilon_n \end{aligned} \quad (2.11)$$

Para simplificar y ahorrar espacio, escribiremos el modelo 2.11 de una forma más abreviada, de tal modo que sólo describamos la observación i -ésima del modelo. Es decir:

$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (2.12)$$

para $i = 1, 2, \dots, n$. Otra forma de expresar el mismo modelo 2.12 de manera aún más abreviada es empleando matrices. Sean:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}_{k \times 1} \quad X = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix}_{n \times k} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

Entonces el modelo 2.12 se puede expresar de forma matricial así:

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\varepsilon}_{n \times 1} \quad (2.13)$$

Recuadro 2.1 Terminología de la regresión múltiple

y	X_2, X_3, \dots, X_k
Variable dependiente	Variables independientes
Variable explicada	Variables explicativas
Variables de respuesta	Variables de control
Variables predicha	Variables predictoras
Regresando	Regresores

2.2.1. Supuestos

Es importante estudiar en detalle el vector de errores ε . En general, esperamos que el error no sea predecible y por tanto trataremos de evitar cualquier componente determinístico o comportamiento sistemático del error. Así, asumiremos que en promedio el término de error es cero. En otras palabras, el valor esperado de cada término de error es cero. En caso que el modelo incluya un intercepto, cualquier componente determinístico del error es capturado por el intercepto. Formalmente, asumiremos que: $E[\varepsilon] = 0$ para todo $i = 1, 2, \dots, n$, o en forma matricial, $E[\varepsilon] = \mathbf{0}$.

Otro supuesto importante para garantizar que los errores son totalmente impredecibles es que cada uno de los errores sea linealmente independientes de los otros. En caso de existir dependencia lineal, habrá una forma de predecir errores futuros a partir de la historia de los errores. Por tanto, se asumirá que $E[\varepsilon_i \varepsilon_j] = E[\varepsilon_i] E[\varepsilon_j] = 0$, esto equivale a:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = E[\varepsilon_i \varepsilon_j] - E[\varepsilon_i] E[\varepsilon_j] = 0$$

A este supuesto se le conoce como el supuesto de no autocorrelación entre los errores.

Finalmente, asumiremos que cada uno de los errores tiene la misma varianza, es decir se asumirá que:

$$\text{Var}[\varepsilon_i] = \sigma^2$$

para todo $i = 1, 2, \dots, n$. Este supuesto se conoce como homoscedasticidad.

En resumen, asumiremos que el error cumple con los siguientes supuestos: 1) media cero, 2) varianza constante, y 3) independencia lineal entre los errores. Estos supuestos se acostumbra resumir de diferentes formas; por ejemplo, se pueden resumir los tres supuestos diciendo que los errores están linealmente independientemente distribuidos con media cero y varianza constante, denotado por $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$. Otra forma de escribir estos supuestos de forma matricial es $\varepsilon \sim (0, \sigma^2 \mathbf{I}_n)$, pues:

$$\begin{aligned} \text{Var}[\varepsilon] &= \begin{bmatrix} \text{Var}[\varepsilon_1] & \text{Cov}(\varepsilon_1, \varepsilon_2) & \cdots & \text{Cov}(\varepsilon_1, \varepsilon_n) \\ \text{Cov}(\varepsilon_2, \varepsilon_1) & \text{Var}[\varepsilon_2] & \cdots & \text{Cov}(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\varepsilon_n, \varepsilon_1) & \text{Cov}(\varepsilon_n, \varepsilon_2) & \cdots & \text{Var}[\varepsilon_n] \end{bmatrix} \\ \text{Var}[\varepsilon] &= \begin{bmatrix} \sigma^2 & 0 \\ & \ddots \\ 0 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 \\ & \ddots \\ 0 & 1 \end{pmatrix} = \sigma^2 \mathbf{I}_n \end{aligned} \quad (2.14)$$

Por otro lado, asumiremos que la información aportada por cada una de las variables explicativas al modelo es relevante. Es decir, no habrá ningún tipo de relación lineal entre las variables explicativas; pues en caso que una variable de control se pudiera expresar como una combinación lineal de otras variables explicativas, la información de la primera variable sería irrelevante pues ya está contenida en las otras. Así, asumiremos que X_2, X_3, \dots, X_k son linealmente independientes.

También, asumiremos que X_2, X_3, \dots, X_k son variables no estocásticas, pues se espera que el proceso de muestreo implícito en el modelo $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ debe poderse repetir numerosas veces y las variables exógenas no deben cambiar pues corresponden al diseño de nuestro “experimento”. Así, se asumirá que X_2, X_3, \dots, X_k son *determinísticas*⁷ (no aleatorias) y *linealmente independientes entre sí*.

Otros supuestos implícitos en nuestro modelo de regresión lineal $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ es que hay en efecto una *relación lineal* entre y y X_2, X_3, \dots, X_k . Asimismo, se supone que esta *relación es estable entre observaciones*; es decir, los parámetros del vector β son constantes a lo largo de la muestra.

Finalmente, cabe mencionar que desde el punto de vista estadístico el modelo de regresión no tiene una connotación de causalidad asociada a él. Así, para el método estadístico es igualmente válido considerar una de las variables explicativas como variable dependiente y la variable como una variable explicativa.

Recuadro 2.2 Supuestos del modelo de regresión múltiple

1. Relación lineal entre y y X_2, X_3, \dots, X_k
2. X_2, X_3, \dots, X_k son fijas y linealmente independientes (la matriz X tiene rango completo)
3. el vector de errores ε satisface:
 - Media cero $E[\varepsilon] = 0$
 - Varianza constante
 - No autocorrelación. Es decir: $\varepsilon_i \sim i.i.d(0, \sigma^2)$ o $\varepsilon_{n \times 1} \sim (\mathbf{0}_{n \times 1}, \sigma^2 \mathbf{I}_n)$

2.2.2. Método de mínimos cuadrados ordinarios (MCO)

Como se mencionó anteriormente, dadas unas observaciones de la variable dependiente e independientes, normalmente se deseará conocer el valor de los coeficientes o parámetros (β). Para lograr acercarnos al valor poblacional desconocido de los coeficientes (β) se emplean estimadores (fórmulas) que responden a una idea plausible para “adivinar” el valor adecuado de éstos.

Una manera muy común de aproximarse a encontrar el “mejor” valor para los coeficientes poblacionales desconocidos (β) es minimizar la suma de los errores

⁷ El supuesto de que las variables explicativas son determinísticas es un supuesto que se puede levantar sin muchas implicaciones. Pero por simplicidad, emplearemos este supuesto a lo largo del libro, a menos que se exprese lo contrario.

elevados al cuadrado;⁸ este método se conoce con el nombre de Mínimos Cuadrados Ordinarios (MCO, o en inglés OLS). Intuitivamente, el método de MCO minimiza la suma de la distancia entre cada una de las observaciones de la variable dependiente (y) y lo que el modelo “predice” ($\hat{y} = \mathbf{X}\hat{\beta}$).

Formalmente, el estimador de MCO para el vector de coeficientes β , denotado por $\hat{\beta}$, se encuentra solucionando el siguiente problema:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ [\mathbf{y} - \hat{\mathbf{y}}]^T [\mathbf{y} - \hat{\mathbf{y}}] \right\} \quad (2.15)$$

Es decir, minimizando el error del modelo cuadrado del modelo (la distancia entre el valor real y el estimado por el modelo). Esto es equivalente a:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ [\mathbf{y} - \mathbf{X}\hat{\beta}]^T [\mathbf{y} - \mathbf{X}\hat{\beta}] \right\} \quad (2.16)$$

donde $\mathbf{X}\hat{\beta}$ corresponde al vector de valores estimados por el modelo para la variable dependiente; es decir, el modelo estimado. Así, el estimador de MCO del vector de coeficientes β es:⁹

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (2.17)$$

La ecuación estimada estará dada por:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \quad (2.18)$$

La diferencia entre el vector de valores observados de la variable dependiente y y los correspondientes valores estimados \hat{y} se denomina el *error estimado* o residuos y se denota por $\hat{\varepsilon}$; en otras palabras:

$$\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} \quad (2.19)$$

En el Apéndice 2.3 se discuten algunas propiedades importantes del vector $\hat{\varepsilon}$

Por otro lado, el estimador de MCO para la varianza del error σ^2 es:

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-k} = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-k} = \frac{\mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y}}{n-k} \quad (2.20)$$

Y el estimador de MCO para la matriz de varianzas y covarianzas de $\hat{\beta}$:

$$\widehat{\text{Var}}[\hat{\beta}] = s^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (2.21)$$

⁸ Elevar al cuadrado el error tiene dos intenciones: 1) evitar que errores positivos y negativos se cancelen y 2) penalizar errores más grandes de manera más fuerte que errores pequeños.

⁹ En el Apéndice 2.1 se presenta la derivación de ésta fórmula.

Puesto que $(\mathbf{X}^T \mathbf{X})^{-1}$ es simétrica, la matriz de varianzas y covarianzas también lo será. Esta matriz tiene la varianza de los estimadores en la diagonal principal y las covarianzas en las posiciones fuera de la diagonal principal. En otras palabras,

$$\widehat{Var}[\beta] = \begin{bmatrix} \widehat{Var}[\beta_1] & \widehat{Cov}[\beta_1, \beta_2] & \cdots & \widehat{Cov}[\beta_1, \beta_k] \\ & \widehat{Var}[\beta_2] & \cdots & \widehat{Cov}[\beta_2, \beta_k] \\ & & \ddots & \vdots \\ & & & \widehat{Var}[\beta_k] \end{bmatrix} \quad (2.22)$$

2.2.3. Propiedades de los estimadores MCO

El estimador de MCO de β es el estimador lineal insesgado con la mínima varianza posible, por esto, el estimador de MCO se conoce como el *Mejor Estimador Lineal Insesgado (MELI)*.¹⁰ Este resultado se conoce como el Teorema de Gauss-Markov (Recuadro 2.3). La propiedad de que el estimador de MCO de β sea insesgado implica que en promedio el estimador obtendrá el valor real β , Formalmente:

$$E[\hat{\beta}] = \beta$$

Y la segunda propiedad que tiene el estimador MCO de β se denomina eficiencia. Es decir, que tiene la mínima varianza posible cuando se compara con todos los otros posibles estimadores lineales. En otras palabras, estas dos propiedades implican que el estimador MCO de β es el mejor estimador lineal disponible, pues en promedio no se equivoca y cuando éste se equivoca tiene la mínima desviación posible.

Recuadro 2.3 Teorema de Gauss-Markov

Si se considera un modelo lineal $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \beta_{k \times 1} + \epsilon_{n \times 1}$ y se supone que:

- Las X_2, X_3, \dots, X_k son fijas y linealmente independientes (es decir X tiene rango columna completo y es una matriz no estocástica).
- El vector de errores ϵ tiene media cero, varianza constante y no autocorrelación. Es decir: $E[\epsilon] = 0$ y $Var[\epsilon] = \sigma^2 I_n$

Entonces, el estimador de MCO $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ es el *Mejor Estimador Lineal Insesgado (MELI)*

¹⁰ Una demostración de este resultado se presenta en el Apéndice 2.2

En la práctica, el Teorema de Gauss-Markov implica que si garantizamos que se cumplen los supuestos entorno al error y de una matriz de regresores con rango completo, entonces el estimador de MCO será MELI.

2.3. Práctica en R: Ley de Okun en Colombia

En todos los capítulos de este libro, el lector encontrará ejercicios, los cuales serán resueltos paso a paso mediante el R (**Rcore**). En este capítulo explicaremos detenidamente desde cómo cargar los datos, hasta cómo estimar un modelo por MCO.

Para este ejercicio, nos basaremos en la ley de Okun para analizar la relación existente entre el crecimiento porcentual del PIB y el crecimiento de la tasa de desempleo en Colombia. Emplearemos datos desde el primer trimestre de 2001 hasta el tercer trimestre de 2018. Los datos están disponibles en el archivo adjunto (Data-Cap1.csv).

Antes de continuar consideremos la siguiente relación dada por la teoría económica:

$$\Delta \%U_t = \beta_0 + \beta_1 \Delta \%PIB_t + \varepsilon_t \quad (2.23)$$

donde $\Delta \%U_t$ y $\Delta \%PIB_t$ corresponden al cambio porcentual en el periodo t de la tasa de desempleo y del PIB de Colombia, respectivamente. Dada la información contenida en el archivo adjunto, se desea estimar el modelo 2.23.

Según **okun1963potential** deberá existir una relación negativa entre el crecimiento de la tasa de desempleo y la tasa de crecimiento del PIB. Adicionalmente, **okun1963potential** argumentaba que el crecimiento del PIB debería ser mucho más grande que la caída en el crecimiento del desempleo. De esta manera, se espera que β_1 sea negativo y mayor que uno. De hecho, Okun encontró que β_1 era de aproximadamente -3 para la economía norteamericana. Es más, existe una basta literatura que discute esos resultados (ver por ejemplo **prachowny1993okun**) para una reestimación de dichos parámetros para Estados Unidos. Así, el cumplimiento de la Ley de Okun implicará que $\beta_1 < -1$. Es decir tenemos una forma empírica de probar el cumplimiento de la Ley de Okun.

2.3.1. Lectura de datos

Recuerde que usted puede importar los datos de un archivo csv (archivo delimitado por comas) utilizando la función `read.csv`¹¹.

Para este ejemplo, descargue el archivo *DataCap1.csv* en su computador en el directorio de trabajo o en una ubicación que le sea conveniente (recuerde que si el archivo no se encuentra en el directorio de trabajo, usted tendrá que especificar toda

¹¹ Para una introducción de cómo importar datos en R se puede emitirse a la sección XXX

la ruta del folder donde se encuentra el archivo) e importe el archivo con la línea de código:

```
DataC1 <- read.csv("DataCap1.csv", sep=",")
```

Para asegurarse de que R haya leído sus datos correctamente, resulta importante mirar qué es lo que efectivamente el programa guardó en el objeto que hemos denominado *DataC1*. Para ello podemos ver las primeras filas del objeto y la clase de objeto de la siguiente manera:

```
head(DataC1)

##           X      TD    PIB  TD_cp PIB_cp
## 1 2001-03-01 16.70 77308      NA      NA
## 2 2001-06-01 14.70 81150 -11.98    4.97
## 3 2001-09-01 14.73 83299   0.23    2.65
## 4 2001-12-01 13.80 83757  -6.33    0.55
## 5 2002-03-01 16.43 84655  19.08    1.07
## 6 2002-06-01 15.80 84575  -3.85   -0.09

class(DataC1)

## [1] "data.frame"
```

Noten que el objeto *DataC1* fue leído como un *data.frame* que contiene una primera columna con las fechas (esta fue denominada *X*), y cuatro columnas mas que contienen la tasa de desempleo (*TD*), el PIB (*PIB*), el cambio porcentual de la tasa de desempleo (*TD_{cp}*) y el crecimiento porcentual del PIB (*PIB_{cp}*). También podemos observar que la primera observación para estas dos últimas variables no se encuentran disponibles (¿Por qué?). Procedamos a eliminar las tres primeras columnas que no serán necesarias y la primera observación que se encuentra perdida. Y constatemos que las variables que nos quedan en el *data.frame* son numéricas. Esto se puede hacer de la siguiente manera:

```
dim(DataC1)

## [1] 71  5

DataC1 <- DataC1[, -c(1:3)]
DataC1 <- DataC1[-1,]
head(DataC1)

##      TD_cp PIB_cp
## 2 -11.98   4.97
## 3   0.23   2.65
## 4  -6.33   0.55
## 5  19.08   1.07
## 6  -3.85  -0.09
## 7  -3.16   6.32

dim(DataC1)

## [1] 70  2
```

```
apply(DataC1, 2, class)

##      TD_cp      PIB_cp
## "numeric" "numeric"
```

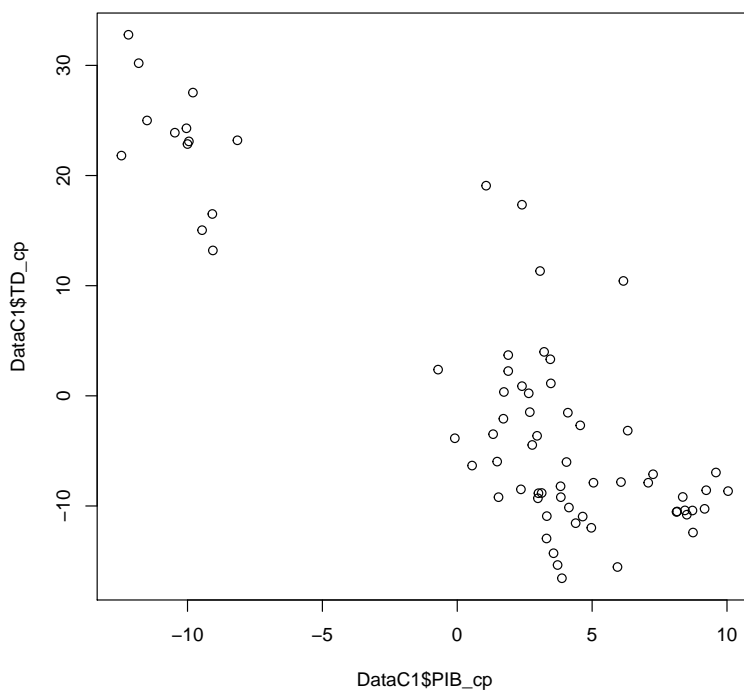
2.3.2. Estimación del modelo

Recordemos que se desea estimar el modelo:

$$\Delta \%U_t = \beta_0 + \beta_1 \Delta \%PIB_t + \varepsilon_t \quad (2.24)$$

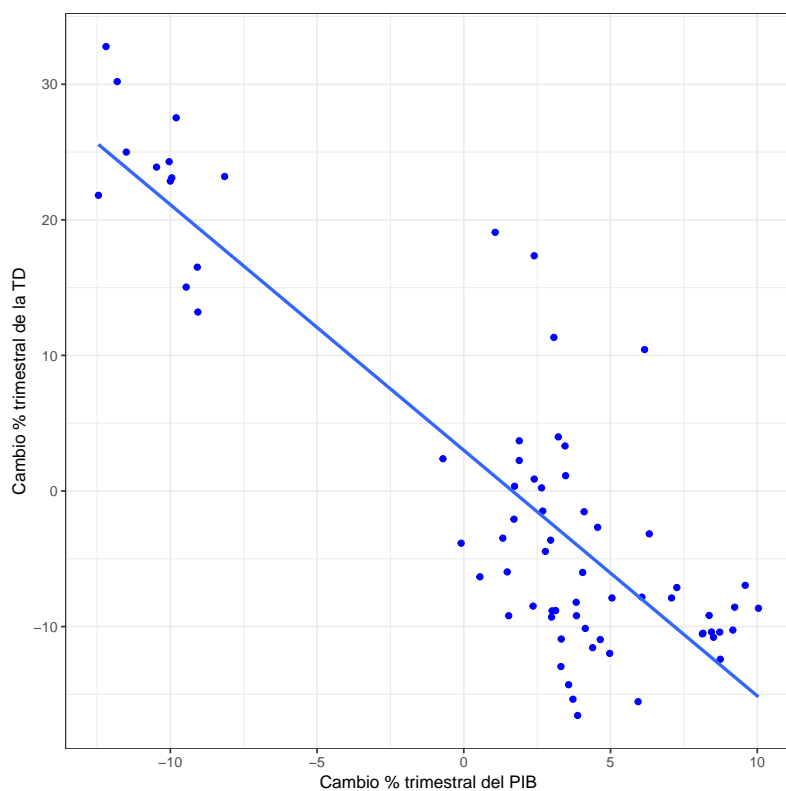
Antes de estimar el modelo, una buena práctica es graficar los datos para poder visualizar como se comportan los datos. Típicamente se emplea un diagrama de dispersión, en nuestro caso tendremos:

```
plot(DataC1$PIB_cp, DataC1$TD_cp)
```



Otra opción es emplear el paquete *ggplot* para tener una visualización más agradable como la siguiente:

Figura 2.1 Relación entre el crecimiento porcentual de la tasa de desempleo y del PIB para Colombia



La figura 2.1 nos sugiere la existencia de una relación lineal. Procedamos a encontrar dicha relación.

Para estimar este modelo por el método de mínimos cuadrados ordinarios la forma más simple es utilizando la función *lm*. Esta función se utiliza para ajustar modelos lineales y hace parte del paquete *stats* que a su vez hace parte del núcleo de paquetes pre instalados en R, motivo por el cual éste ya se encuentra cargado. Los dos principales argumentos de la función *lm* son dos: el modelo a estimar denominado fórmula y los datos que se emplearán. En nuestro caso:

```
R1 <- lm(formula = TD_cp ~ PIB_cp, data = DataC1)
```

En este caso hemos guardado en el objeto *R1* los resultados de estimar 2.24. Noten que se emplea la virgulilla (palito de la eñe) para representar el signo igual de la expresión 2.24, también es importante anotar que no fue necesario escribir todos los parámetros desados, R por defecto incluye un intercepto y los correspondientes

β s que representan pendientes. Para observar los resultados del modelo estimado se puede emplear la función *summary* de la siguiente manera:

```
summary(R1)

##
## Call:
## lm(formula = TD_cp ~ PIB_cp, data = DataC1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5319  -5.6147   0.8211   3.0796  18.6964
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0024     0.8393   3.577 0.000645 ***
## PIB_cp       -1.8120     0.1302 -13.922 < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.785 on 68 degrees of freedom
## Multiple R-squared:  0.7403, Adjusted R-squared:  0.7365
## F-statistic: 193.8 on 1 and 68 DF,  p-value: < 2.2e-16
```

Otras formas de expresar la fórmula que llevarían al mismo resultado (compruébelo) son:

```
R1a <- lm( TD_cp ~ PIB_cp, DataC1)
formulal <- TD_cp ~ PIB_cp
R2 <- lm( formulal , DataC1)
R3 <- lm( TD_cp ~ ., DataC1)
```

La última forma de expresar la fórmula con un punto implica emplear todas las variables que se encuentran en la base de datos (diferentes a la que se seleccionó para ser la dependiente) como variables explicativas. Dado que en nuestro caso solo tenemos una variable más, por eso el resultado es el mismo que en los casos anteriores.

Sin embargo, si lo que quiere es que el resultado de la estimación se muestre como la tabla que regularmente se utiliza en la presentación de documentos y está empleando \LaTeX , usted puede utilizar el paquete *estout* que deberá instalar como se indicó con el paquete anterior. Una vez instalado el paquete puede utilizar la siguiente secuencia de códigos para obtener la tabla adecuada:

```
library(estout)
estclear()
eststo(R1)
esttab(t.value = TRUE,  sig.levels = c(0.1, 0.05, 0.01),
       sig.sym=c("*", "**", "***"),
       caption="Modelo estimado por MCO",
       caption.top=TRUE, label="res1",
       var.rename=c("(Intercept)", "intercepto") )
```

Cuadro 2.1 Modelo estimado por MCO

	TD_{cp}
intercepto	3.002*** [3.577]
PIB_{CP}	-1.812*** [-13.922]
R^2	0.74
$adj.R^2$	0.736
N	70

t-values in brackets

* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)

Noten que en este caso la ecuación estimada será:

$$\Delta \%U_t = 3 - 1,81\Delta \%PIB_t \quad (2.25)$$

Así, el resultado implica que por cada punto porcentual que aumente el PIB de un trimestre a otro, la tasa de desempleo caerá en 1.81 puntos porcentuales.

Para obtener la matriz de varianzas y covarianzas del intercepto y la primera pendiente, que es otra cantidad que desconocíamos, se puede emplear la función *vcov* de la siguiente manera:

vcov (R1)

Esto quiere decir que en este caso, se tiene que:

$$\begin{aligned} \hat{\beta} &= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 3 \\ -1,81 \end{bmatrix} \\ \sqrt{\widehat{Var}(\beta_0)} &= S_{\hat{\beta}_0} = 0,84 \\ \sqrt{\widehat{Var}(\beta_1)} &= S_{\hat{\beta}_1} = 0,13 \end{aligned} \quad (2.26)$$

Como se discutirá en el próximo capítulo, del cuadro ?? podemos concluir que el coeficiente asociado a la variación del PIB es significativo ya que el estadístico t es relativamente alto (y el valor p asociado a este es muy pequeño). Así, este coeficiente es estadísticamente diferente de cero. Adicionalmente, aparentemente se cumple que $\hat{\beta}_1 < -1$ (Esto tendrá que ser demostrado por medio de una prueba tal como se discutirá en el siguiente capítulo). Por lo tanto aparentemente podemos concluir que la Ley de Okun se cumple para Colombia durante el periodo estudiado de acuerdo a la especificación empleada; es decir, los cambios en la tasa de crecimiento del PIB afectan la variación en la tasa de desempleo tal como lo postula la Ley de Okun.

2.4. Ejercicios

El gobierno de una pequeña República está reconsiderando la viabilidad del transporte ferroviario, para lo cual contrata un estudio que determine un modelo que permita comprender de una forma más precisa el comportamiento de los ingresos del sector (I medidos en millones de dólares). Un investigador sugiere el siguiente modelo:

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \alpha_4 LDies_t + \alpha_5 LEI_t + \alpha_6 V_t + \varepsilon_t \quad (2.27)$$

Donde, CE_t , CD_t , $LDies_t$, LEI_t y V_t representan el consumo de electricidad medido en millones de Kilovatios/hora, el consumo de diesel medido en millones de galones, el número de locomotoras diesel en el país, el número de locomotoras eléctricas y el número de viajeros (medido en miles de pasajeros) en el año t , respectivamente.

Para efectuar este estudio se cuenta con información para el período 1994-2018 (los datos se encuentran en el archivo regmult.xls).

1. De acuerdo con el enunciado anterior, estime el modelo 2.27 y reporte sus resultados en una tabla
2. Interprete los coeficientes estimados

2.5. Apéndice

Apéndice 2.1 Derivación de los estimadores MCO

Formalmente, el estimador de MCO para el vector de coeficientes β en el modelo 2.16, denotado por $\hat{\beta}$, se encuentra minimizando la distancia cuadrada entre cada valor observado del vector de realizaciones de la variable aleatoria dependiente (\mathbf{y}) y el vector de estimaciones del modelo ($\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$). Formalmente, el problema es:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ \left[\mathbf{y} - \mathbf{X}\hat{\beta} \right]^T \left[\mathbf{y} - \mathbf{X}\hat{\beta} \right] \right\} \quad (2.28)$$

Antes de encontrar las condiciones de primer orden y las de segundo orden para un mínimo, podemos simplificar un poco el problema expresado en la ecuación 2.28. Así, tenemos que:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ \left[\mathbf{y}^T - \hat{\beta}^T \mathbf{X}^T \right] \left[\mathbf{y} - \mathbf{X}\hat{\beta} \right] \right\}$$

Al multiplicar los elementos dentro de los corchetes cuadrados obtenemos:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \hat{\beta} + \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} \right\}$$

Observen que $\mathbf{y}^T \mathbf{1}_{1 \times n} \mathbf{X}_{n \times k} \hat{\beta}_{k \times 1} = \left(\hat{\beta}_{k \times 1}^T \mathbf{X}_{n \times k}^T \mathbf{y}_{1 \times n} \right)^T$ y además los productos $\hat{\beta}_{1 \times k}^T \mathbf{X}_{k \times n}^T \mathbf{y}_{n \times 1}$ y $\mathbf{y}_{1 \times n}^T \mathbf{X}_{n \times k} \hat{\beta}_{k \times 1}$ son escalares, por tanto $\mathbf{y}_{1 \times n}^T \mathbf{X}_{n \times k} \hat{\beta}_{k \times 1} = \hat{\beta}_{1 \times k}^T \mathbf{X}_{k \times n}^T \mathbf{y}_{n \times 1}$. Así, el problema 2.28 es equivalente a:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ \mathbf{y}^T \mathbf{y} - 2\hat{\beta}^T \mathbf{X}^T \mathbf{y} + \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} \right\} \quad (2.29)$$

Ya podemos regresar a nuestro problema de minimización y considerar la condición de primer orden para este problema, en este caso la condición es:¹²

$$\frac{\partial}{\partial \hat{\beta}} \{ \bullet \} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \hat{\beta} \equiv 0 \quad (2.30)$$

La expresión 2.30 se conoce como las ecuaciones normales. Ahora, despejando $\hat{\beta}$, obtenemos:

$$2\mathbf{X}^T \mathbf{X} \hat{\beta} = 2\mathbf{X}^T \mathbf{y}$$

Multiplicando a ambos lados por la inversa¹³ de $\mathbf{X}^T \mathbf{X}$, tendremos:

$$(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

¹² Hay que anotar que la derivada es con respecto a un vector y no a un escalar.

¹³ $(\mathbf{X}^T \mathbf{X})^{-1}$ existe pues \mathbf{X} tiene rango completo gracias al supuesto de que $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k$ son linealmente independientes

Así, el estimado de MCO del vector de coeficientes β es:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

La condición de segundo orden implica $\frac{\partial \{\bullet\}}{\partial \beta \partial \beta} = 2\mathbf{X}^T \mathbf{X}$. Noten que $(\mathbf{X}^T \mathbf{X})$ es una matriz positiva semi-definida lo que garantiza que $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ es un mínimo.

Apéndice 2.2 Demostración del Teorema de Gauss-Markov

El Teorema de Gauss-Markov implica los siguientes supuestos:

- Existe una relación lineal entre \mathbf{y} y las variables en la matriz \mathbf{X} que se puede representar por el modelo lineal $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$
- X_2, X_3, \dots, X_k son no estocásticas y linealmente independientes (es decir \mathbf{X} tiene rango completo y es una matriz no estocástica)
- El vector de errores $\boldsymbol{\varepsilon}$ tiene media cero, varianza constante y no autocorrelación. Es decir, $E[\boldsymbol{\varepsilon}] = 0$ y $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$

Entonces el estimador de MCO, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, es insesgado y eficiente. En otras palabras, $\hat{\boldsymbol{\beta}}$ es el *Mejor Estimador Lineal Insesgado* (MELI) para el vector de coeficientes poblacionales $\boldsymbol{\beta}$.

A continuación demostraremos este Teorema. Primero, demostraremos que $\hat{\boldsymbol{\beta}}$ es un estimador insesgado del vector de coeficientes poblacionales $\boldsymbol{\beta}$. Formalmente, tenemos que demostrar que $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. Para lograr tal fin calculemos el valor esperado del estimador MCO:

$$E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}]$$

Empleando el supuesto de que las variables explicativas son no estocásticas tenemos que,

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\mathbf{y}]$$

Y por tanto,

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}]$$

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\boldsymbol{\varepsilon}]$$

Recuerden que habíamos supuesto que $E[\boldsymbol{\varepsilon}] = 0$. Esto implica que

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{I} \cdot \boldsymbol{\beta}$$

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

Así, hemos demostrado que el estimador MCO es insesgado.

Antes de continuar con la demostración del Teorema de Gauss-Markov, encontremos la varianza del estimador de MCO. Es decir,

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \text{Var}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}]$$

$$\text{Var}[\hat{\boldsymbol{\beta}}] = ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \text{Var}[\mathbf{y}] ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T$$

$$\text{Var}[\hat{\boldsymbol{\beta}}] = ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \text{Var}[\mathbf{y}] (\mathbf{X} ((\mathbf{X}^T \mathbf{X})^{-1})^T)$$

$$\begin{aligned}
\text{Var} [\hat{\beta}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var} [\mathbf{y}] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
\text{Var} [\hat{\beta}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var} [\mathbf{X}\beta + \varepsilon] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
\text{Var} [\hat{\beta}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var} [\varepsilon] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
\text{Var} [\hat{\beta}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
\text{Var} [\hat{\beta}] &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
\text{Var} [\hat{\beta}] &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}
\end{aligned}$$

Ahora retornemos al Teorema de Gauss-Markov, para demostrarlo es necesario probar que este estimador tiene la mínima varianza entre todos los posibles estimadores lineales insesgados de β .

Sin perder generalidad, consideremos otro estimador lineal cualquiera $\tilde{\beta} = [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] \mathbf{y}$. Si $\tilde{\beta}$ es insesgado, se debe cumplir que $E [\tilde{\beta}] = \beta$. Es decir:

$$\begin{aligned}
E [\tilde{\beta}] &= [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] E [\mathbf{y}] \\
E [\tilde{\beta}] &= [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] E [\mathbf{X}\beta + \varepsilon] = [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] [\mathbf{X}\beta + 0] \\
E [\tilde{\beta}] &= [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\beta + C\mathbf{X}\beta] \\
E [\tilde{\beta}] &= \beta + C\mathbf{X}\beta
\end{aligned}$$

Para que este estimador sea insesgado, tiene que cumplirse que $C\mathbf{X} = 0$.

Ahora, analicemos la varianza de este nuevo estimador lineal insesgado.

$$\begin{aligned}
\text{Var} [\tilde{\beta}] &= \text{Var} \left[[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] \mathbf{y} \right] \\
\text{Var} [\tilde{\beta}] &= [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] \text{Var} [\mathbf{y}] [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C]^T \\
\text{Var} [\tilde{\beta}] &= [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] \sigma^2 I_n [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C]^T \\
\text{Var} [\tilde{\beta}] &= \sigma^2 [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + C] [\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + C^T] \\
\text{Var} [\tilde{\beta}] &= \sigma^2 [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + C\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T C^T + CC^T] \\
\text{Var} [\tilde{\beta}] &= \sigma^2 [(\mathbf{X}^T \mathbf{X})^{-1} + C\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} (C\mathbf{X})^T + CC^T]
\end{aligned}$$

Como la condición $C\mathbf{X} = 0$ para que $\tilde{\beta}$ sea insesgado tiene que cumplirse, entonces:

$$\text{Var} [\tilde{\beta}] = \sigma^2 \left[(\mathbf{X}^T \mathbf{X})^{-1} + CC^T \right]$$

CC^T es una matriz cuyos elementos en la diagonal principal serán positivos.¹⁴ (¿Por qué?) Ahora comparemos la varianza de nuestro estimador de MCO ($\hat{\beta}$) con $\tilde{\beta}$. Recuerden que $\text{Var} [\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$, adicionalmente observen que $\text{Var} [\tilde{\beta}] = \sigma^2 \left[(\mathbf{X}^T \mathbf{X})^{-1} + CC^T \right]$ no puede ser menor que $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$, pues:

- $\text{Var} [\tilde{\beta}_i] = \sigma^2 \left\{ \left[(\mathbf{X}\mathbf{X})_{ii}^{-1} + CC_{ii}^T \right] \right\}$, donde A_{ij} corresponde al elemento en la fila i y columna j de la matriz A , y
- CC_{ii}^T es positivo.

Por tanto $\text{Var} [\tilde{\beta}_i] = \sigma^2 \left\{ \left[(\mathbf{X}\mathbf{X})_{ii}^{-1} + CC_{ii}^T \right] \right\} > \text{Var} [\hat{\beta}_i] = \sigma^2 \left[(\mathbf{X}\mathbf{X})_{ii}^{-1} \right]$. En el mejor de los casos $\text{Var} [\tilde{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ y eso sólo ocurre cuando $C = 0$. En este caso $\tilde{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \hat{\beta}$, es decir la mínima varianza posible de un estimador lineal insesgado es igual a la varianza del estimador MCO. Por tanto $\hat{\beta}$ es MELI.

¹⁴ Una matriz positiva semi-definida

Apéndice 2.3 Algunas propiedades importantes de los residuos estimados.

Como se discutió anteriormente, se tiene que los residuos estimados están definidos como

$$\hat{\varepsilon} = y - \mathbf{X}\hat{\beta}$$

La primera propiedad que cumple el vector de residuos ($\hat{\varepsilon}$) es:

$$\mathbf{X}^T \hat{\varepsilon} = 0 \quad (2.31)$$

Para demostrar esta propiedad podemos partir de la definición de los residuos estimados. Es decir, tenemos que:

$$\mathbf{X}^T \hat{\varepsilon} = \mathbf{X}^T [y - \mathbf{X}\hat{\beta}] = \mathbf{X}^T y - \mathbf{X}^T \mathbf{X}\hat{\beta}$$

Además, sustituyendo $\hat{\beta}$ se tiene

$$\mathbf{X}^T \hat{\varepsilon} = \mathbf{X}^T y - \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

Por lo tanto,

$$\mathbf{X}^T \hat{\varepsilon} = \mathbf{X}^T y - \mathbf{X}^T y = 0$$

De esta propiedad se desprende un resultado muy interesante. Dado que esta propiedad implica que:

$$\mathbf{X}^T \hat{\varepsilon} = \begin{bmatrix} \sum_{i=1}^n X_{1i} \hat{\varepsilon}_i \\ \sum_{i=1}^n X_{2i} \hat{\varepsilon}_i \\ \vdots \\ \sum_{i=1}^n X_{ki} \hat{\varepsilon}_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Entonces, se desprende un resultado importante. *Si el modelo tiene intercepto la suma de los residuos estimados será cero.* Esta afirmación se puede demostrar fácilmente. Noten que en caso de tener intercepto el modelo, \mathbf{X} tendrá una columna de unos. Por ejemplo, $X_{1i} = 1$ y se desprende que:

$$\sum_{i=1}^n \hat{\varepsilon}_i = 0 \quad (2.32)$$

Este último resultado implica que *la media del residuo estimado sea cero.* En otras palabras:

$$\bar{\varepsilon} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i}{n} = 0 \quad (2.33)$$

La segunda propiedad de los residuos estimados que discutiremos es que:

$$E [\hat{\varepsilon}^T \hat{\varepsilon}] = E \left[\sum_{i=1}^n \hat{\varepsilon}_i^2 \right] = (n-2) \sigma^2 \quad (2.34)$$

Para demostrar esta propiedad, reescribamos de manera más conveniente los residuos estimados:

$$\begin{aligned} \hat{\varepsilon} &= y - \mathbf{X}\hat{\beta} = y - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y \\ \hat{\varepsilon} &= \left[I - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] y \end{aligned}$$

Ahora, definamos la matriz $\mathbf{M} = I - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Es muy fácil mostrar (hágalo) que \mathbf{M} es idempotente y simétrica. Entonces, se tiene que los residuos estimados se pueden expresar de la siguiente manera:

$$\hat{\varepsilon} = \mathbf{M}y = \mathbf{M}\mathbf{X}\beta + \mathbf{M}\varepsilon = \mathbf{M}\varepsilon \quad (2.35)$$

Es importante anotar que $\mathbf{M}\mathbf{X}\beta = 0$, dado que

$$\mathbf{M}\mathbf{X}\beta = \left[I - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{X}\beta = [\mathbf{X}\beta - \mathbf{X}\beta] = 0$$

Regresando a (2.35), se tiene que:

$$E [\hat{\varepsilon}^T \hat{\varepsilon}] = E [(\mathbf{M}\varepsilon)^T \mathbf{M}\varepsilon] = E [\varepsilon^T \mathbf{M}^T \mathbf{M}\varepsilon] = E [\varepsilon^T \mathbf{M}\varepsilon]$$

Y dado que el producto $\hat{\varepsilon}^T \hat{\varepsilon}$ es un escalar tendremos que:

$$E [\hat{\varepsilon}^T \hat{\varepsilon}] = \text{trace} (\varepsilon^T \mathbf{M}\varepsilon) = \text{trace} (\mathbf{M}\varepsilon^T \varepsilon)$$

Empleando las propiedades de la traza,

$$E [\hat{\varepsilon}^T \hat{\varepsilon}] = \text{trace} (\mathbf{M}\sigma^2 I) = \sigma^2 \text{trace} (\mathbf{M})$$

Noten que encontrar la traza de la matriz \mathbf{M} es muy sencillo pues,

$$\text{trace} (\mathbf{M}) = \text{trace} (I_n) - \text{trace} \left(\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right)$$

y empleando otras propiedades de la traza sabemos que:

$$\text{trace} (\mathbf{M}) = \text{trace} (I_n) - \text{trace} \left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \right)$$

De esta manera se tiene que

$$\text{trace} (\mathbf{M}) = \text{trace} (I_n) - \text{trace} (I_k) = n - k$$

Esto implica que

$$E [\hat{\varepsilon}^T \hat{\varepsilon}] = \sigma^2 \text{trace} (\mathbf{M}) = (n - k) \sigma^2 \quad (2.36)$$

Este último resultado es importante porque implica que *el estimador MCO para la varianza de los errores es insesgado*. En otras palabras:

$$E[s^2] = E\left[\frac{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}}{n-k}\right] = E\left[\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-k}\right] = \sigma^2 \quad (2.37)$$

Capítulo 3

Inferencia y análisis de regresión

Objetivos del capítulo

Al finalizar este capítulo, el lector estará en capacidad de:

- Realizar pruebas individuales y conjuntas sobre los coeficientes estimados de un modelo de regresión empleando R.
- Valorar mediante la salida de R la bondad estadística del modelo estimado.
- Interpretar y calcular con R de una regresión múltiple.

3.1. Introducción

El término inferencia se refiere a sacar conclusiones para la muestra a partir de la cual hemos estimado nuestro modelo. En este capítulo estudiaremos cómo comprobar restricciones que involucren uno o más parámetros del modelo estudiado. Y en especial discutiremos como determinar si una variable o un conjunto de estas afectan o no a la variable dependiente. En el capítulo anterior estudiamos el método de Mínimos Cuadrados Ordinarios (MCO) para estimar un modelo lineal de la forma:

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\varepsilon}_{n \times 1} \quad (3.1)$$

Como lo discutimos, el estimador de MCO para el vector de parámetro $\boldsymbol{\beta}$ está dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (3.2)$$

Adicionalmente, demostramos que si se supone que i) X_2, X_3, \dots, X_k son fijas y linealmente independientes, y ii) $E[\boldsymbol{\varepsilon}] = 0$, $Var[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$; entonces el estimador de MCO es el mejor estimador lineal insesgado.

Pero, antes de continuar es importante reconocer que la probabilidad de que el estimador de MCO ($\hat{\boldsymbol{\beta}}$) para una muestra dada sea exactamente igual al valor poblacional es cero. Aunque esta última afirmación suene a primera vista algo contradictoria; si se reflexiona un poco tendrá más sentido. Es muy difícil “adivinar” correctamente el valor real a partir de un solo intento, cuando el verdadero valor pertenece a un conjunto infinito de valores. (Ver ejercicio de simulación en R al final del capítulo para un ejemplo de esto)

Ustedes se deben estar preguntando: si eso es cierto, entonces ¿para qué empleamos estos estimadores si dada una muestra no hay chance que el estimador sea igual al valor real? La respuesta es clara, cómo se demostró en el capítulo anterior, en promedio el estimador de MCO es igual al valor poblacional (insesgado); por tanto, si recolectamos un número lo suficientemente grande de muestras podremos encontrar el valor promedio del estimador. Pero evidentemente en la práctica no nos podemos dar el “lujo” de tener más de una muestra para la población en estudio.

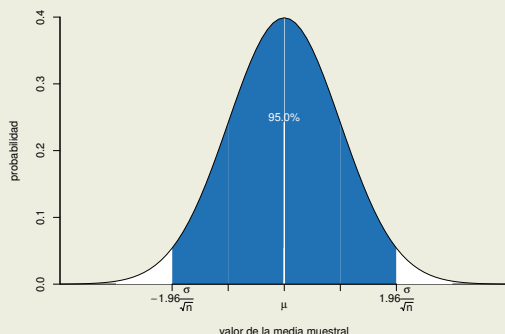
Es ahí donde la teoría estadística llega al auxilio del científico de datos. Si bien sabemos que un valor estimado a partir de una muestra tiene una probabilidad de cero de acertar el valor real, si podemos aumentar la certidumbre de nuestra estimación si conocemos la dispersión y la distribución del estimador (Ver Ejemplo 3.7).

Recuadro 3.1 Teorema del límite central

Supongamos que ustedes quieren conocer el peso medio de una población de estudiantes, es decir la media poblacional del peso de ellos. En este caso podrían hacer un censo del peso de todos los estudiantes de la universidad y una vez pesa-

dos todos calcular el promedio aritmético. Pero, hacer un censo es muy costoso y dispendioso, por tanto, como aprendieron en sus cursos introductorios de estadística, se acostumbra seleccionar una muestra aleatoria de tamaño n y a cada uno de esos n estudiantes se les pesa (W_i para $i = 1, 2, \dots, n$). A partir de esta muestra po-

demos calcular la media muestral, es decir $\bar{w} = \frac{\sum_{i=1}^n W_i}{n}$. ¿Cuál es la probabilidad que \bar{w} sea exactamente igual a la media poblacional?. Claramente esta probabilidad es cero; pero gracias al Teorema del Límite Central, si la muestra es lo suficientemente grande, entonces sabemos que \bar{w} sigue una distribución normal con una media igual a la media poblacional y varianza igual a la varianza poblacional dividida por el tamaño muestral ($\frac{\sigma^2}{n}$). Este resultado nos permite aumentar la certidumbre de nuestro pronóstico (\bar{w}), para que sea más exacto



Sabemos que si nos movemos dos desviaciones estándar (en este caso $\frac{\sigma}{\sqrt{n}}$) a la derecha y a la izquierda de nuestro valor estimado, entonces tenemos un 95% de seguridad de que el valor real está contenido en ese intervalo. Así, empleando la distribución de nuestro estimador podemos aumentar la certidumbre sobre nuestra estimación.

Si conocemos la distribución muestral de nuestro estimador de MCO podremos mejorar la certeza en torno a nuestras estimaciones. Estudiemos pues cuál es la distribución de nuestro estimador.

Sabemos que $E[\hat{\beta}] = \beta$ y $Var[\hat{\beta}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Por tanto, ahora necesitamos conocer la distribución exacta que sigue nuestro estimador de MCO, pues recuerden que una función de distribución no está caracterizada únicamente por su media y varianza. Por otro lado, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ es una combinación lineal de variables no aleatorias y variables aleatorias. Noten que $(\mathbf{X}^T\mathbf{X})^{-1}$ corresponde a una matriz no aleatoria, pues hemos asumido que \mathbf{X} es una matriz no estocástica. Además, dado que cada uno de los \mathbf{y}_i es una variable aleatoria, pues \mathbf{y}_i depende de ε_i ; tenemos que $\mathbf{X}^T\mathbf{y}$ es un vector cuyos elementos son sumas de variables aleatorias. Es decir,

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i \mathbf{X}_{2i} \\ \sum_{i=1}^n y_i \mathbf{X}_{3i} \\ \vdots \\ \sum_{i=1}^n y_i \mathbf{X}_{ki} \end{bmatrix}$$

Por lo tanto $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ también será un vector cuyos elementos corresponden a combinaciones lineales de sumas de variables aleatorias. Ahora bien, si tenemos una muestra lo suficientemente grande, podemos emplear el Teorema del Límite Central para concluir que el vector $\hat{\beta}$ al ser una combinación de suma de variables aleatorias seguirá una distribución normal.

Así, si hay una muestra lo suficientemente grande tendremos que:

$$\hat{\beta} \sim N_k \left(\beta, \left[\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \right]_{k \times k} \right) \quad (3.3)$$

En otras palabras, los estimadores de MCO seguirán una distribución asintóticamente normal (sigue una distribución Multivariada Normal de orden k) con media igual al valor poblacional y matriz de varianzas y covarianzas $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$.

De esta manera, cualquier estimador $\hat{\beta}_p$, para $1 \leq p \leq k$, seguirá una distribución asintótica normal con media igual al valor poblacional β_p y varianza igual al elemento en la columna y fila p de la matriz de varianzas y covarianzas de los estimadores de MCO. En otras palabras,

$$\hat{\beta}_p \sim N \left(\beta_p, \left[\sigma^2 \left\{ \text{Elemento de la matriz } (\mathbf{X}^T \mathbf{X})^{-1} \right\} \right] \right) \quad (3.4)$$

Una vez conocemos la distribución del estimador de MCO podremos hacer inferencia sobre los parámetros poblacionales. En lo que resta de este capítulo discutiremos cómo hacer inferencia respecto a un parámetro (pruebas individuales) y sobre un conjunto de ellos (pruebas conjuntas).

3.2. Pruebas individuales sobre los parámetros

Hay un pequeño inconveniente al emplear el resultado de la ecuación 3.4. En la práctica, no conocemos el valor poblacional de la varianza del error σ^2 , por tanto debemos estimarlo por medio de s^2 . Al estimar este parámetro, estamos introduciendo más incertidumbre en nuestro proceso de inferencia. Esta incertidumbre, agregada al problema de inferencia, implica que la distribución de los estimadores no será tan concentrada hacia la media, como lo es una distribución normal cuyas colas son

relativamente pequeñas. La mayor incertidumbre implicará una mayor probabilidad de estar lejos del valor de la media poblacional, que se refleja en unas colas más grandes (más altas). Estas colas más altas, aun manteniendo la simetría y gran masa de probabilidad cercana a la media, se pueden capturar con una distribución t.

Por tanto, cuando no conocemos la varianza poblacional del término aleatorio de error (¡que es siempre!), cada uno de los estimadores de MCO seguirá una distribución t, con $n - k$ grados de libertad. De esta manera, tenemos que un *intervalo de confianza* para β_k del $(1 - \alpha)$ 100% de confianza está dado por:

$$\hat{\beta}_p \pm t_{\frac{\alpha}{2}, n-k} s_{\hat{\beta}_p} \quad (3.5)$$

donde $s_{\hat{\beta}_p}$ corresponde a la raíz cuadrada de la varianza estimada del estimador $\hat{\beta}_k$. (Ver ejercicio de simulación en R al final del capítulo para un ejemplo de esto)

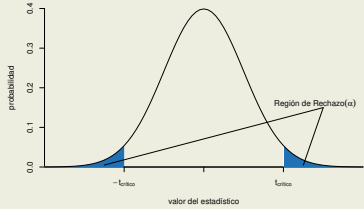
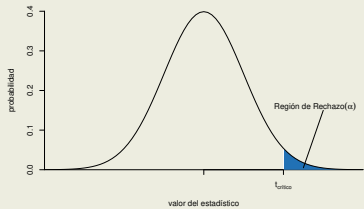
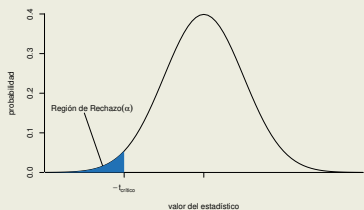
Además, la distribución del estimador nos permite probar la hipótesis nula que el valor poblacional β_p es igual a cualquier constante c , es decir $H_o : \beta_p = c$; versus la hipótesis alterna que β_p no es igual a la constante c , $H_A : \beta_p \neq c$. Para contrastar estas hipótesis, se emplea el siguiente estadístico t de prueba:

$$t_c = \frac{\hat{\beta}_p - c}{s_{\hat{\beta}_p}} \quad (3.6)$$

Este estadístico t_c (o t-calculado) sigue una distribución t con $n - k$ grados de libertad; por tanto, se rechazará la hipótesis nula si el valor absoluto del estadístico t_c es lo suficientemente grande. En otras palabras, si $|t_c| > t_{\frac{\alpha}{2}, n-k}$.

Un caso muy especial es cuando $c = 0$, en ese caso la hipótesis nula implicará $\beta_k = 0$; es decir, la variable X_k no explica la variable dependiente. Por otro lado, la hipótesis alterna implicará que $\beta_k \neq 0$; es decir, la variable X_k sí ayuda a explicar la variable dependiente. Cuando se concluye a partir de la muestra que $\beta_k \neq 0$, entonces se dice que la variable X_k es significativa para el modelo, alternativamente se dice que el coeficiente β_k es significativo. Por esto, este tipo de pruebas se conocen como pruebas de significancia individual.

Recuadro 3.2 Tipos de pruebas individuales

Hipótesis nula	Hipótesis alterna	Estadístico	Rechazar H_0 si	Zona de rechazo
$H_o : \beta_p = c$	$H_A : \beta_p \neq c$	$t_c = \frac{\hat{\beta}_p - c}{s_{\hat{\beta}_p}}$	$ t_c > t_{\frac{\alpha}{2}}$	
$H_o : \beta_p \leq c$	$H_A : \beta_p > c$	$t_c = \frac{\hat{\beta}_p - c}{s_{\hat{\beta}_p}}$	$t_c > t_{\alpha}$	
$H_o : \beta_p \geq c$	$H_A : \beta_p < c$	$t_c = \frac{\hat{\beta}_p - c}{s_{\hat{\beta}_p}}$	$t_c < t_{\alpha}$	

Recuadro 3.3 Pruebas de hipótesis y el valor p

Una manera alternativa de rechazar o no H_0 es empleando el valor p. Este valor corresponde a la probabilidad de obtener un estadístico t más grande en valor absoluto que el observado (Ver figura).

Empleando este criterio, la decisión de rechazar se puede tomar de la siguiente manera:

Nivel de significancia	Se rechaza si
10 %	valor $p < 0.1$
5 %	valor $p < 0.05$
1 %	valor $p < 0.01$

3.3. El ajuste del modelo (Fit del modelo)

Antes de continuar, es necesario preguntarnos ¿cómo determinar si el modelo estimado explica satisfactoriamente la muestra bajo estudio?. En otras palabras, si el modelo estadístico se ajusta bien a la muestra. En especial, ¿cómo podemos saber si el modelo lineal en efecto es válido para la muestra?, pues recuerde que nuestro primer supuesto es que la relación entre las variables independientes y la dependiente es lineal, supuesto simplificador que en la mayoría de los casos no tiene ningún soporte ni fáctico ni teórico.

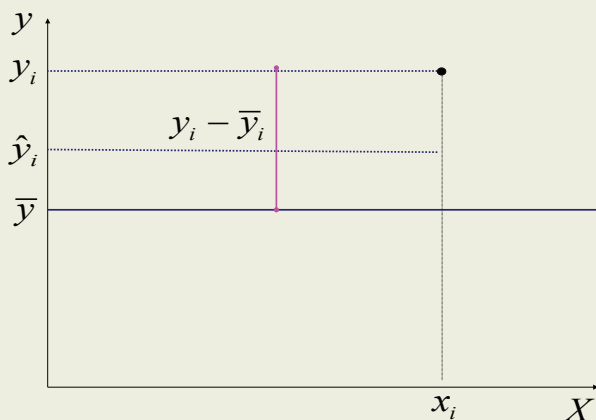
Antes de analizar qué tan bien explica nuestro modelo la variable dependiente, es necesario discutir cómo se puede descomponer la variabilidad de la variable dependiente. Inicialmente consideremos la variación total de la variable dependiente con respecto a su media ($y_i - \bar{y}$), esta variación puede ser descompuesta en dos partes¹. La primera de ellas es la parte de la variación que es explicada por el modelo; esta parte de la variación es la diferencia que existe entre lo que predice el modelo para la observación² i (\hat{y}_i) y nuestra mejor predicción si no contáramos con un modelo, es decir la media de la variable dependiente (\bar{y}). En otras palabras, la parte de la variación de la variable dependiente explicada por el modelo corresponde a $(\hat{y}_i - \bar{y})$. La segunda parte de la variación de la variable dependiente es la que no puede ser explicada por el modelo, que denominaremos el error o residuo, es decir $\hat{\epsilon}_i = (y_i - \hat{y}_i)$. Así, $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$. Esta descomposición se puede visualizar de forma más intuitiva con el recuadro 2.4.

Recuadro 3.4 Descomposición de la variación total de la variable dependiente

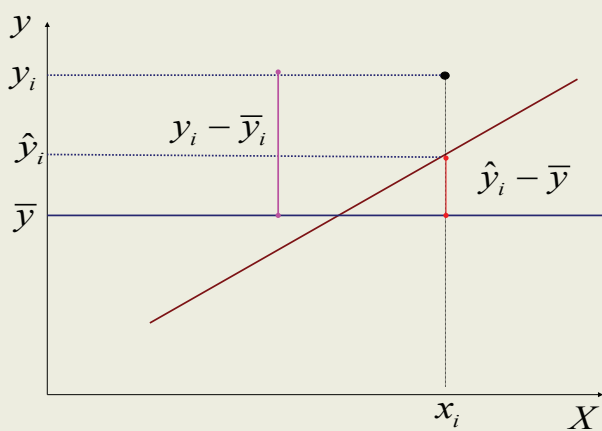
Si no contamos con un modelo, la mejor forma de explicar la variable dependiente es empleando su media. A la diferencia entre la media y cada observación se le conoce como la variación total.

¹ Noten que partimos de la desviación de cada observación con respecto a su media y la denominamos variación total, pues esta sería la desviación que tendría cada observación con respecto a lo esperado si usásemos el modelo más sencillo para explicar el comportamiento de una variable. Es decir, si usamos el modelo $y_i = \mu + \epsilon_i$, este será nuestro modelo de partida.

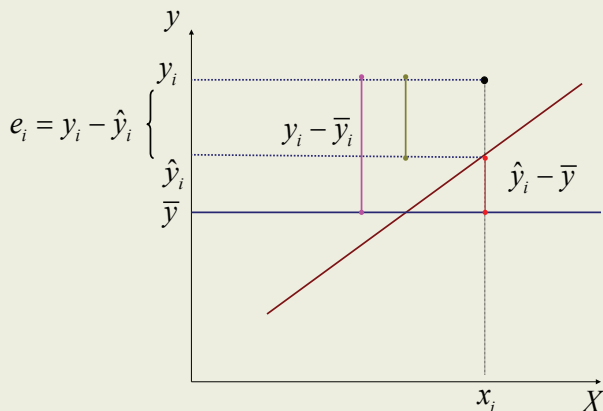
² $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki}$ para $i = 1, 2, \dots, n$, o en forma matricial $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.



La variación total se puede descomponer en: i) la parte explicada por el modelo ($\hat{y}_i - \bar{y}$), y



ii) la parte que el modelo no explica (el error) $\hat{\epsilon}_i = (y_i - \hat{y}_i)$



Si queremos conocer la variación total de la variable dependiente para toda la muestra, es buena idea considerar la suma de todas las variaciones al cuadrado para evitar el efecto contrario, de variaciones por encima y por debajo de la media. Llamaremos a $\sum_{i=1}^n (y_i - \bar{y})^2$ la Suma Total al Cuadrado (*SST* por su nombre en inglés: Sum of Squares Total). Es muy fácil mostrar que en presencia de un modelo con intercepto:³

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.7)$$

La parte de esta variación total al cuadrado que es explicada por el modelo será $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, la denotaremos como *SSR* (por su nombre en inglés: Sum of Squares of the Regression). Y la parte que no es explicada por la regresión se denotará por *SSE* (por su nombre en inglés: Sum of Squares Error). Estas sumas al cuadrado también se pueden expresar en forma matricial. Es relativamente fácil mostrar que:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \mathbf{y}^T \mathbf{y} - n\bar{y}^2$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y}$$

Así tendremos que

$$SST = SSR + SSE$$

³ En el apéndice 1 se presenta una demostración de esta afirmación.

Una evidente medida de qué tan bueno es nuestro modelo es examinar qué porcentaje de la variación total es explicada por el modelo, a esto se le conoce como el R^2 . Formalmente,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Dado que el R^2 es un porcentaje, éste estará entre cero y uno; es decir, $0 \leq R^2 \leq 1$. Si $R^2 = 1$, entonces toda la variación de la variable dependiente es explicada por el modelo; para el caso de dos variables explicativas, esto implicará que todos los puntos se encuentran sobre el plano estimado por el modelo, en caso de tener una sola variable explicativa, este R^2 implicará que todos los puntos están sobre la línea de regresión. Por otro lado, si $R^2 = 0$ tendremos que nuestro modelo no explica nada de la variación de la variable dependiente.

Recuadro 3.5 Resumiendo la descomposición de la variabilidad de la variable dependiente

Una forma de resumir la descomposición de la variabilidad de la variable dependiente⁴ es emplear una tabla conocida como la tabla ANOVA (Analysis of Variance.). La tabla ANOVA no sólo resume la descomposición de la variabilidad de la variable dependiente, sino que también resume información importante como los grados de libertad asociados con cada suma al cuadrado.

Intuitivamente, noten que para calcular el $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ se emplean n observaciones y se pierde un grado de libertad al emplear la media, por tanto los grados de libertad del SST son $n - 1$. Por otro lado, para calcular el $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ se emplean n observaciones y se pierde k grados de libertad al calcular \hat{y}_i , por tanto los grados de libertad del SSE son $n - k$. Finalmente, así como $SST = SSE + SSR$, también se debe cumplir que los grados de libertad del SST deben ser iguales a la suma de los grados de libertad del SSE y el SSR . Así por diferencia, podemos encontrar que los grados de libertad del SSR son $k - 1$. Una forma alternativa para llegar a este último resultado es advertir que para calcular $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ se emplean k grados de libertad en el cálculo de cada \hat{y}_i y se pierde un grado de libertad al emplear la media \bar{y} , por tanto los grados de libertad del SSR son $k - 1$.

Finalmente, en la tabla ANOVA se reporta la Media Cuadrada (MS) de los errores y de la regresión que corresponde a las respectivas Sumas al Cuadrado divididas por sus respectivos grados de libertad. En especial, el MS correspondiente al error, MSE , es exactamente igual al estimador de la varianza del error. Claramente de la tabla ANOVA, también se puede derivar rápidamente el R^2 , pues toda la información necesaria para el cálculo de este estadístico está disponible en la tabla.

Fuente de la Variación	SS	Grados de libertad	MS
Regresión	$SSR = \hat{\beta}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2$	$k - 1$	$MSRN = \frac{SSRN}{k-1}$
Error	$SSE = \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y}$	$n - k$	$MSR = s^2 = \frac{SSR}{n-k}$
Total	$SST = \mathbf{y}^T \mathbf{y} - n\bar{y}^2$	$n - 1$	

Antes de continuar, es importante anotar algunas consideraciones para tener en cuenta sobre la medida de bondad de ajuste brindada por el R^2 . Primero, es relativamente fácil mostrar que en caso de que el modelo estimado no contenga un intercepto, entonces $SST \neq SSRN + SSE$ (ver anexo 1). Por tanto en ese caso el R^2 no representará la parte de la variabilidad total de la variable dependiente explicada por el modelo, por eso se dice que el R^2 carece de interpretación cuando el modelo estimado no posee intercepto.

Segundo, dado que el R^2 es una medida del ajuste de un modelo lineal estimado a los datos, entonces el R^2 se puede emplear para escoger entre diferentes modelos lineales el que se ajuste mejor a los datos, pero si se cumplen algunas condiciones.

Por ejemplo, supongamos que se quiere estimar la función de demanda de un bien Q_i , para lo cual se cuenta con información de su precio, p_i , el precio de un bien sustituto, ps_i , y un bien complementario, pc_i y el nivel ingreso medio de los consumidores de este bien, I_i . Así, para estimar la demanda de este bien se emplean los siguientes modelos:

$$Q_i = \beta_0 + \beta_1 p_i + \beta_2 pc_i + \beta_3 ps_i + \beta_4 I_i + \varepsilon_i \quad (3.8)$$

$$\ln(Q_i) = \gamma_0 + \gamma_1 \ln(p_i) + \gamma_2 \ln(pc_i) + \gamma_3 \ln(ps_i) + \gamma_4 \ln(I_i) + \mu_i \quad (3.9)$$

$$Q_i = \beta_0 + \beta_1 p_i + \beta_2 pc_i + \beta_4 I_i + \varepsilon_i \quad (3.10)$$

$$Q_i = \beta_0 + \beta_1 p_i + \beta_2 ps_i + \beta_4 I_i + \varepsilon_i \quad (3.11)$$

Una vez estimados estos cuatro modelos se cuentan con su correspondiente R^2 . Un “impulso natural” es escoger como el “mejor” modelo aquel que tenga el R^2 más grande. Pero tal procedimiento en este caso es erróneo, pues al comparar el modelo 3.9 con los otros modelos nos damos cuenta que la variable dependiente es diferente en este modelo. Así, aún si se emplee la misma muestra para estimar los modelos, las SST no serán iguales para todos. La SST del modelo 3.9 es distinta a la de los otros tres modelos, pues para el modelo 3.9 $SST = \sum_{i=1}^n (\ln(Q_i) - \ln(\bar{Q}))^2$ mientras que para los otros modelos tenemos que $SST = \sum_{i=1}^n (Q_i - \bar{Q})^2$.

Por lo tanto, si comparamos el R^2 del modelo 3.9 con el de los otros tres modelos, no estaríamos comparando la explicación de la misma variabilidad total de la variable dependiente. Podemos concluir que *los R^2 entre diferentes modelos son*

comparables si y solamente si la variable dependiente es la misma en los modelos a comparar y se emplea la misma muestra en la estimación de los modelos.

Pero, aún si comparamos modelos con la misma variable dependiente y la misma muestra, como por ejemplo los modelos 3.8, 3.10 y 3.11, es necesario tener cuidado con esta comparación. Es relativamente fácil mostrar que el SSR, y por tanto el R^2 , es una función creciente del número de regresores ($k - 1$). Por tanto, a medida que consideramos modelos con más variables explicatorias, el R^2 será más grande.

Entonces, podríamos escoger cuál modelo explica mayor proporción de la variabilidad de la variable dependiente para los modelos 3.10 y 3.11 por medio del R^2 , pues en estos dos modelos se tiene el mismo número de variables explicativas (k) y la variable explicada es la misma. Pero, en general, el R^2 no es un buen criterio para escoger entre los modelos 3.8, 3.10 y 3.11.

3.4. Pruebas conjuntas sobre los parámetros

Hasta aquí hemos discutido cómo comprobar hipótesis individuales en torno a cada uno de los parámetros de un modelo de regresión. Pero, ¿qué hacer con estos resultados individuales? Supongamos la siguiente situación; hemos estimado con una muestra lo suficientemente grande el siguiente modelo:

$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

Además, suponga que hemos efectuado una serie de pruebas individuales sobre los parámetros estimados y se ha encontrado que:

- La hipótesis nula $\beta_2 = 0$ es rechazada a un nivel de significancia de α . Llamemos a este el resultado 1 (R_1)
- La hipótesis nula $\beta_3 = 0$ no se puede rechazar a un nivel de significancia de α . (R_2)
- La hipótesis nula $\beta_4 = 0$ no se puede rechazar a un nivel de significancia de α . (R_3)

Dados estos tres resultados, parece muy razonable tratar de unir estos tres resultados ($R_1 \cup R_2 \cup R_3$) y concluir que el modelo debería ser $y_i = \beta_1 + \beta_2 X_{2i} + \varepsilon_i$ con un nivel de significancia de α . Pero si reflexionamos un poco sobre las implicaciones de unir estos resultados, nos daremos cuenta rápidamente lo errado de unirlos. Noten que el primer resultado R_1 implica un error tipo I de tamaño α , mientras que los resultados R_2 y R_3 tienen asociados un error tipo II. Ahora, si unimos estos tres resultados, el nivel de significancia asociado a ($R_1 \cup R_2 \cup R_3$) no será simplemente α .

Para evitar ser muy conservador al unir diferentes resultados individuales, se han diseñado pruebas que tengan en cuenta estos múltiples errores. Tales pruebas se conocen como pruebas conjuntas. En esta sección las discutiremos en detalle.

Supongamos que usted quiere determinar si todos los coeficientes asociados a pendientes son o no iguales a cero conjuntamente, es decir $H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$ versus la hipótesis alterna no H_0 . Como lo discutimos anteriormente, este tipo de hipótesis no se puede probar con $k - 1$ hipótesis individuales. Noten que esta hipótesis puede reescribirse de la forma $\mathbf{R}_{(r) \times k} \beta_{k \times 1} = \mathbf{C}_{(r) \times 1}$, donde $\beta^T = [\beta_1 \ \beta_2 \ \dots \ \beta_k]$, $\mathbf{C}^T = [0 \ 0 \ \dots \ 0]$ y:

$$R_{(r-1) \times k} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Ahora, supongamos que quiere saber si dada una estimación se cumple una restricción que tiene sentido teórico. Por ejemplo, se desea emplear una función de producción Cobb-Douglas para determinar si se presenta rendimientos constantes de escala o no. Es decir, dada la siguiente ecuación estimada $\widehat{\ln(Y_i)} = \hat{\beta} + \hat{\alpha}_1 \ln(K_i) + \hat{\alpha}_2 \ln(L_i)$ se quiere comprobar si $\alpha_1 + \alpha_2$ es igual a uno o no. Esta hipótesis también se puede escribir de la forma $\mathbf{R}_{(r) \times k} \beta_{k \times 1} = \mathbf{C}_{(r) \times 1}$, donde $\beta^T = [\beta \ \alpha_1 \ \alpha_2]$, $R = [0 \ 1 \ 1]$ y $C = [1]$.

En general, cualquier hipótesis nula que pueda escribirse de la forma $R_{(r) \times k} \beta_{k \times 1} = C_{(r) \times 1}$ (versus la H_a : no H_0) se puede comprobar empleando el siguiente estadístico:

$$F_{\text{calculado}} = \frac{(\mathbf{C} - \mathbf{R}\hat{\beta})^T (\mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{R}^T)^{-1} (\mathbf{C} - \mathbf{R}\hat{\beta})}{\hat{\varepsilon}^T \hat{\varepsilon} / (n - k)} \Bigg/ r \quad (3.12)$$

Este estadístico sigue una distribución F con r grados de libertad en el numerador y $n - k$ grados de libertad en el denominador. Por tanto se podrá rechazar la hipótesis nula, con un nivel de significancia α si $F_{\alpha, (r, (n-k))} < F_{\text{calculado}}$.

Un caso especial de la hipótesis nula general $R_{(r) \times k} \beta_{k \times 1} = C_{(r) \times 1}$ (versus la H_a : no H_0) es $H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$ (todos los coeficientes asociados con pendientes son simultáneamente iguales a cero), es decir una prueba global de significancia. En este caso especial el estadístico $F_{\text{calculado}}$ se reduce a:

$$F_{\text{Global}} = \frac{\hat{\beta}^T \mathbf{X}^T \mathbf{y} - n \bar{y}^2}{\hat{\varepsilon}^T \hat{\varepsilon} / (n - k)} \Bigg/ (k - 1) = \frac{MSR}{MSE} \quad (3.13)$$

O de manera equivalente:

$$F_{\text{Global}} = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)} \quad (3.14)$$

Este F_{Global} sigue una distribución F con $k - 1$ grados de libertad en el numerador y $n - k$ grados de libertad en el denominador. Por tanto, se podrá rechazar la hipótesis

nula a un nivel de significancia α si $F_{\alpha,((k-1),(n-k))} < F_{Global}$. Noten que este test se puede derivar rápidamente de la tabla ANOVA (Recuadro 3.6).

Recuadro 3.6 Tabla ANOVA con prueba de significancia global

Fuente de la Variación	SS	Grados de libertad	MS	F-Global
Regresión	SSR	$k - 1$	$MSR = \frac{SSR}{k-1}$	$F_{Global} = \frac{MSR}{MSE}$
Error	SSE	$n - k$	$MSE = s^2 = \frac{SSE}{n-k}$	
Total	SST	$n - 1$		

Noten que la hipótesis nula de esta prueba de significancia global implica el modelo

$$y_i = \beta_1 + \varepsilon_i \tag{3.15}$$

mientras que la hipótesis alterna implica el modelo

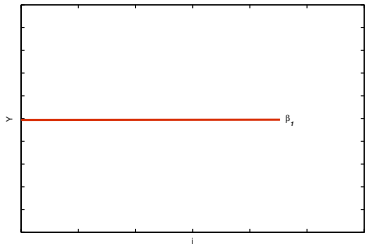
$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i \tag{3.16}$$

Es decir, si la hipótesis nula no es rechazada, esto implica que una mejor forma de explicar la variable dependiente es por medio de su media (una constante).⁵ Por tanto, esta prueba es otra forma de estudiar la bondad de ajuste del modelo, por eso no debe ser sorprendente la estrecha relación entre el F_{Global} y el R_2 .

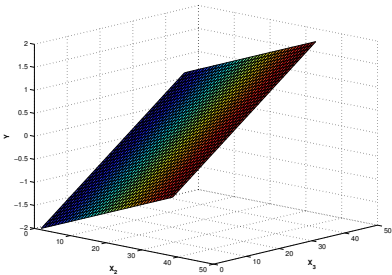
$$F_{Global} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{3.17}$$

Figura 3.1 Modelos 3.15 y 3.16

Modelo 3.15



Modelo 3.16



⁵ Es decir $Y_i = \mu + \varepsilon_i$, donde $\mu = \beta_1$ es la media.

3.5. Prueba de Wald y su relación con la prueba F

En la sección anterior se discutió cómo para comprobar una hipótesis nula que involucre restricciones lineales de la forma $\mathbf{R}\hat{\beta} = \mathbf{C}$ podemos emplear el estadístico $F_{calculado}$ siguiente:

$$F_{calculado} = \frac{(\mathbf{C} - \mathbf{R}\hat{\beta})^T (\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T)^{-1} (\mathbf{C} - \mathbf{R}\hat{\beta})}{\hat{\varepsilon}^T \hat{\varepsilon} / (n - k)} / r \quad (3.18)$$

Una alternativa para probar hipótesis que involucren restricciones de la forma $\mathbf{R}\hat{\beta} = \mathbf{C}$ es emplear el estadístico de Wald ($W_{calculado}$), el cual está relacionado con el $F_{calculado}$ descrito con anterioridad. Éste estadístico se puede calcular con la siguiente expresión:

$$W_{calculado} = (\mathbf{C} - \mathbf{R}\hat{\beta})^T (\mathbf{R}(S^2\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T)^{-1} (\mathbf{C} - \mathbf{R}\hat{\beta}) \quad (3.19)$$

El estadístico de Wald calculado sigue una distribución χ_r^2 donde r representa el número de restricciones lineales en la hipótesis nula. Por tanto, se podrá rechazar $H_0 : \mathbf{R}\beta = \mathbf{C}$ en favor de la hipótesis alterna (no H_0) si $W_{calculado} > \chi_r^2$.

Es muy fácil encontrar la relación entre el estadístico de Wald y el estadístico F. Por ejemplo, manipulando algebraicamente (3.19) tendremos que:

$$W_{calculado} = (\mathbf{C} - \mathbf{R}\hat{\beta})^T (\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T)^{-1} (\mathbf{C} - \mathbf{R}\hat{\beta}) (S^2)^{-1} \quad (3.20)$$

$$W_{calculado} = \frac{(\mathbf{C} - \mathbf{R}\hat{\beta})^T (\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T)^{-1} (\mathbf{C} - \mathbf{R}\hat{\beta})}{S^2} \quad (3.21)$$

y reemplazando S^2 tenemos que:

$$W_{calculado} = \frac{(\mathbf{C} - \mathbf{R}\hat{\beta})^T (\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T)^{-1} (\mathbf{C} - \mathbf{R}\hat{\beta})}{\hat{\varepsilon}^T \hat{\varepsilon} / (n - k)} \quad (3.22)$$

Al comparar la anterior expresión con (3.18) se puede concluir que $W_{calculado} = r \cdot F_{calculado}$. De hecho, para comprobar cualquier hipótesis que involucre restricciones lineales de la forma $\mathbf{R}\hat{\beta} = \mathbf{C}$ se podrá emplear una prueba F o una prueba Wald. Sin importar el estadístico que se emplee, la conclusión será la misma. Finalmente, esta prueba de Wald se puede calificar como una versión abreviada de una prueba de la razón de verosimilitud (Likelihood Ratio Test).

```
load("Data/RetornosDiarios.RData")
library(xts)
```

3.6. Práctica en R: Explicando los rendimientos de una acción

En esta sección encontraremos la relación del rendimiento diario⁶ de la acción de Suramericana con el rendimiento de otras acciones que también se tranzan en la Bolsa de Valores de Colombia. Se cuenta con información para el rendimiento diario de las siguientes acciones: GRUPOSURA, ECOPETROL, NUTRESA, EXITO, ISA, GRUPOAVAL, CONCONCRETO, VALOREM y OCCIDENTE. Se cuenta una base de datos que va desde 2012-01-02 hasta el 2019-01-14. La información se encuentra en el *working space* *RetornosDiarios.RData*.

La primera tarea como siempre será cargar el *working space*, para lo cual se puede emplear la siguiente línea de código:

```
load("Data/RetornosDiarios.RData")
```

El *working space* contiene el objeto *retornos.diarios*. Asegurémonos que los datos quedaron bien cargados y que cada variable se encuentre bien definida.

```
head(retornos.diarios)

##          GRUPOSURA ECOPETROL  NUTRESA  EXITO
## 2012-01-02  1.2779727 -0.3565066 -0.9216655 2.02184181
## 2012-01-03  2.5079684  2.0036027 -0.2781643 0.07695268
## 2012-01-04  0.4324999  1.0446990 -0.5586607 1.07116556
## 2012-01-05 -0.4324999 -0.2312140  0.6514681 0.45558165
## 2012-01-06  0.3091193  0.1156738 -0.5581410 0.00000000
## 2012-01-10 -1.5552413  3.6326355 -0.4675090 0.60423145
##          ISA  GRUPOAVAL CONCONCRET  VALOREM
## 2012-01-02 -1.983834  0.0000000  0.0000000  0.000000
## 2012-01-03  1.626052 -2.4292693  0.0000000 12.583905
## 2012-01-04  2.478003 -0.4106782  0.0000000  1.342302
## 2012-01-05  0.000000  0.4106782  0.0000000  0.000000
## 2012-01-06  0.000000 -1.2371292  0.0000000 -11.653382
## 2012-01-10 -2.120221  0.4140793  0.7220248  0.000000
##          OCCIDENTE
## 2012-01-02         0
## 2012-01-03         0
## 2012-01-04         0
## 2012-01-05         0
## 2012-01-06         0
## 2012-01-10         0

class(retornos.diarios)

## [1] "xts" "zoo"

sapply(retornos.diarios,class)

##          GRUPOSURA ECOPETROL NUTRESA EXITO ISA  GRUPOAVAL
## [1,] "xts"         "xts"         "xts"  "xts" "xts" "xts"
## [2,] "zoo"         "zoo"         "zoo"  "zoo" "zoo" "zoo"
```

⁶ El rendimiento diario se calcula como el crecimiento porcentual en el precio de la acción.

```
##          CONCRET VALOREM OCCIDENTE
## [1,]  "xts"      "xts"      "xts"
## [2,]  "zoo"      "zoo"      "zoo"
```

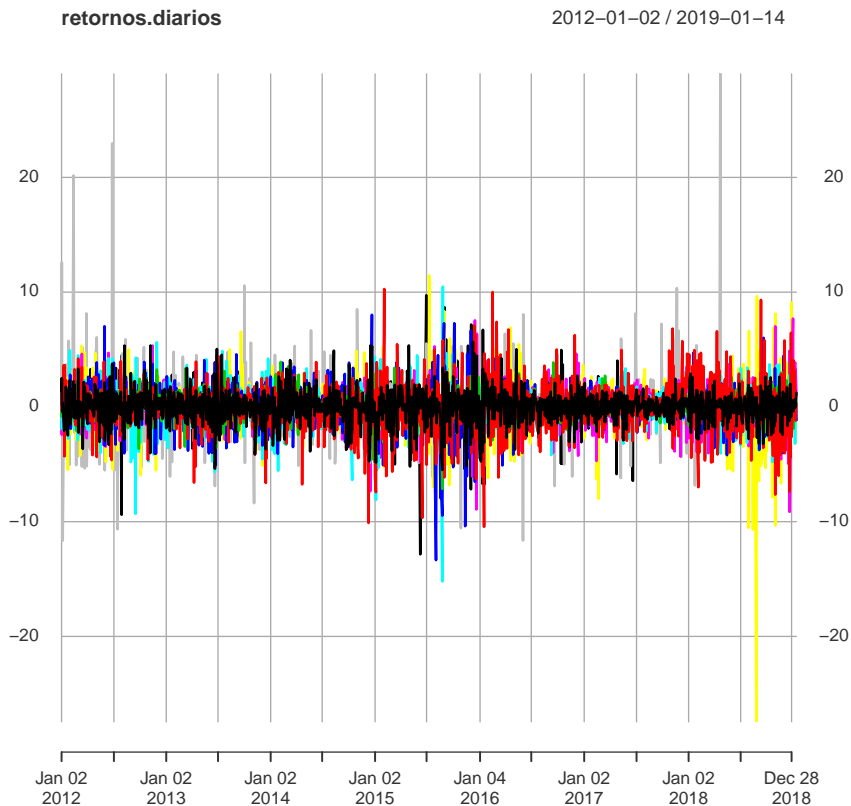
Esto significa que todas las series (variables), así como el objeto, son de clase *xts* y *zoo*. Esta es una clase de objeto que se emplea en R para manejar series de tiempo. Casi todas las funciones que operan sobre objetos de clase *data frame* también funcionarán sobre objetos *xts* y *zoo*. En especial la función *lm()* no tendrá ningún problema con este tipo de objetos. Para tener una idea de la base de datos antes de entrar a estimar el modelo, veamos las estadísticas descriptivas, un gráfico y la correlación entre las series. Empecemos por las estadísticas descriptivas.

```
summary(retornos.diarios)

##          Index          GRUPOSURA
## Min.      :2012-01-02   Min.      : -5.491576
## 1st Qu.:2013-09-25     1st Qu.: -0.605399
## Median :2015-06-30     Median : 0.000000
## Mean     :2015-07-04     Mean     : 0.002958
## 3rd Qu.:2017-04-07     3rd Qu.: 0.679216
## Max.     :2019-01-14     Max.     : 5.671623
## ECOPELTROL            NUTRESA
## Min.      : -10.44520   Min.      : -7.145896
## 1st Qu.: -0.96271      1st Qu.: -0.525538
## Median : 0.00000       Median : 0.000000
## Mean     : -0.02184     Mean     : 0.004678
## 3rd Qu.: 1.00884       3rd Qu.: 0.528802
## Max.     : 10.27128     Max.     : 4.807793
## EXITO                ISA
## Min.      : -13.35314   Min.      : -15.20399
## 1st Qu.: -0.76845      1st Qu.: -0.75520
## Median : 0.00000       Median : 0.00000
## Mean     : -0.03991     Mean     : 0.01399
## 3rd Qu.: 0.73004       3rd Qu.: 0.85536
## Max.     : 8.02808      Max.     : 10.48256
## GRUPOAVAL            CONCRET
## Min.      : -9.143421   Min.      : -27.44368
## 1st Qu.: -0.722025     1st Qu.: -0.44445
## Median : 0.000000      Median : 0.00000
## Mean     : -0.007806    Mean     : -0.07761
## 3rd Qu.: 0.719428      3rd Qu.: 0.36597
## Max.     : 7.696104     Max.     : 11.46629
## VALOREM              OCCIDENTE
## Min.      : -11.65338   Min.      : -12.84562
## 1st Qu.: 0.00000       1st Qu.: 0.00000
## Median : 0.00000       Median : 0.00000
## Mean     : 0.07358      Mean     : 0.01409
## 3rd Qu.: 0.00000       3rd Qu.: 0.00000
## Max.     : 29.02523     Max.     : 9.72907
```

Noten que todas las series tienen mediana cero y son relativamente volátiles. Esto se puede corroborar gráficamente. Grafiquemos las series empleando la función *plot()* del paquete *xts*.

```
#install.packages("xts")
library(xts)
plot(retornos.diarios)
```



Claramente las series (variables) de los rendimientos son bastante volátiles.
La correlación entre las variables en la base de datos se puede encontrar rápidamente de la siguiente manera.

```
#install.packages("xts")
cor(retornos.diarios)
```

##		GRUPOSURA	ECOPETROL	NUTRESA	EXITO
##	GRUPOSURA	1.00000000	0.31484667	0.30342557	0.325895624
##	ECOPETROL	0.31484667	1.00000000	0.22831629	0.211280942
##	NUTRESA	0.30342557	0.22831629	1.00000000	0.287789067
##	EXITO	0.32589562	0.21128094	0.28778907	1.000000000
##	ISA	0.38753552	0.23471847	0.31512773	0.342276871
##	GRUPOAVAL	0.21041357	0.13729742	0.21306950	0.165121466
##	CONCONCRET	0.06832743	0.04790886	0.08627644	0.054414525
##	VALOREM	0.08344384	0.03179496	0.04886834	0.024619737

```
## OCCIDENTE 0.04935665 0.02520985 0.04626136 0.003960018
##          ISA  GRUPOAVAL  CONCRET  VALOREM
## GRUPOSURA 0.38753552 0.21041357 0.0683274294 0.083443838
## ECOPETROL 0.23471847 0.13729742 0.0479088646 0.031794961
## NUTRESA 0.31512773 0.21306950 0.0862764384 0.048868338
## EXITO 0.34227687 0.16512147 0.0544145254 0.024619737
## ISA 1.00000000 0.17048382 0.0213234075 0.055440112
## GRUPOAVAL 0.17048382 1.00000000 0.0539074452 0.010420969
## CONCRET 0.02132341 0.05390745 1.0000000000 0.017061697
## VALOREM 0.05544011 0.01042097 0.0170616970 1.000000000
## OCCIDENTE 0.04018166 0.05118878 -0.0001436823 0.009036396
##          OCCIDENTE
## GRUPOSURA 0.0493566507
## ECOPETROL 0.0252098498
## NUTRESA 0.0462613552
## EXITO 0.0039600182
## ISA 0.0401816587
## GRUPOAVAL 0.0511887845
## CONCRET -0.0001436823
## VALOREM 0.0090363957
## OCCIDENTE 1.0000000000
```

¿Qué puede concluir?. Ahora procedamos a estimar el siguiente modelo de regresión:

$$\begin{aligned} \text{GRUPOSURA}_t = & \beta_1 + \beta_2 \text{ECOPETROL}_t + \beta_3 \text{NUTRESA}_t \\ & + \beta_4 \text{EXITO}_t + \beta_5 \text{ISA}_t + \beta_7 \text{GRUPOAVAL}_t \\ & + \beta_8 \text{CONCRET}_t + \beta_9 \text{VALOREM}_t \\ & + \beta_{10} \text{OCCIDENTE}_t + \varepsilon_t \end{aligned}$$

Donde GRUPOSURA_t representa el rendimiento diario de la acción del Grupo Sura. De manera análoga las otras variables representan los rendimientos de las otras acciones. El modelo se puede estimar y visualizar de la siguiente manera:

```
res1 <- lm(GRUPOSURA ~ ., retornos.diarios)
summary(res1)

##
## Call:
## lm(formula = GRUPOSURA ~ ., data = retornos.diarios)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.6534 -0.5737 -0.0133  0.6199  6.3444
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.006431   0.027001   0.238    0.812
## ECOPETROL    0.121058   0.014487   8.357 < 2e-16 ***
## NUTRESA      0.141435   0.027796   5.088 4.02e-07 ***
## EXITO        0.122278   0.018116   6.750 2.03e-11 ***
```

```
## ISA          0.188165    0.018689    10.068    < 2e-16 ***
## GRUPOAVAL    0.084403    0.020077     4.204    2.76e-05 ***
## CONCRET      0.021189    0.014785     1.433     0.152
## VALOREM      0.035659    0.014007     2.546     0.011 *
## OCCIDENTE    0.030972    0.026930     1.150     0.250
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.109 on 1687 degrees of freedom
## Multiple R-squared:  0.2617, Adjusted R-squared:  0.2582
## F-statistic: 74.75 on 8 and 1687 DF,  p-value: < 2.2e-16
```

Recuadro 3.7 Asteriscos en la literatura, en R y el valor p

R emplea una manera particular en poner asteriscos que no concuerda con la literatura científica (por lo menos en las ciencias sociales). En el siguiente cuadro se presenta una equivalencia

p <	en R	en literatura
0.1	.	*
0.05	*	**
0.01	**	*
0.001	***	No se emplea

Los resultados se resumen en el siguiente cuadro.

Los resultados muestran que los rendimientos de las acciones de ECOPETROL, NUTRESA, EXITO, ISA, GRUPOAVAL y VALOREM son variables significativas para explicar el rendimiento de la acción de GRUPOSURA. Las pruebas individuales de significancia permiten concluir que los rendimientos de las acciones de CONCRET y OCCIDENTE no tienen efecto sobre el rendimiento de la acción de GRUPOSURA. (Asegúrese que puede interpretar cada uno de los coeficientes significativos).

También podemos observar que el R^2 de la regresión es 0.262. Esto quiere decir que el modelo explica el 26.2% de la variación del rendimiento de la acción de GRUPOSURA. Noten que si bien parece pequeño este R^2 , no lo es. Dado que los rendimientos de la acción de GRUPOSURA es tan volátil (cambia tanto), explicar el 26.2% de esta variación no es despreciable. En el contexto de las finanzas, este R^2 no es bajo. Es importante tener en cuenta que en otras aplicaciones este R^2 puede ser considerado como muy bajo. Por eso es importante conocer el contexto bajo estudio para poder determinar si un R^2 obtenido es relativamente alto o bajo para el caso bajo estudio.

Cuadro 3.1 Modelo estimado por MCO

GRUPOSURA	
intercepto	0.006 [0.238]
ECOPETROL	0.121*** [8.357]
NUTRESA	0.141*** [5.088]
EXITO	0.122*** [6.75]
ISA	0.188*** [10.068]
GRUPOAVAL	0.084*** [4.204]
CONCONCRET	0.021 [1.433]
VALOREM	0.036** [2.546]
OCCIDENTE	0.031 [1.15]
R^2	0.262
$adj.R^2$	0.258
N	1696

t-values in brackets

* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)

Por otro lado el F_{Global} es 74.748. El respectivo p-valor (a próximamente 0) implica que se puede rechazar la hipótesis nula de que ninguna variable es importante para explicar los rendimientos de la acción de GRUPOSURA. Es decir, al menos una de las pendientes es diferente de cero.

Para el cálculo de la tabla ANOVA podemos emplear la función *anova* de la siguiente manera:

```
anova(res1)

## Analysis of Variance Table
##
## Response: GRUPOSURA
##          Df Sum Sq Mean Sq F value    Pr(>F)
## ECOPETROL  1  278.82  278.820  226.5073 < 2.2e-16 ***
## NUTRESA    1  159.09  159.086  129.2379 < 2.2e-16 ***
## EXITO      1  126.82  126.815  103.0217 < 2.2e-16 ***
## ISA        1  136.45  136.453  110.8514 < 2.2e-16 ***
## GRUPOAVAL  1   22.65   22.648   18.3989 1.893e-05 ***
## CONCONCRET 1    2.63    2.626    2.1335  0.14430
## VALOREM    1    8.02    8.019    6.5147  0.01079 *
## OCCIDENTE  1    1.63    1.628    1.3227  0.25027
## Residuals 1687 2076.62   1.231
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Noten que en este caso el SSR se presenta discriminado para cada variable.

3.6.1. Pruebas conjuntas sobre los parámetros

Ahora, consideremos la prueba conjunta de que las acciones de CONCRETTO y OCCIDENTE no tienen efecto sobre el rendimiento de la acción de GRUPOSURA. Es decir,

$$H_0 : \beta_8 = \beta_{10} = 0$$

La hipótesis alterna es no H_0 . Este tipo de restricciones se puede probar tanto con la prueba de Wald como con una prueba F . Esto se puede hacer de la siguiente manera empleando la función `linearHypothesis()` que se encuentra en el paquete `AER`. Esta función requiere al menos dos argumentos: El modelo y las restricciones. La función calcula por defecto la prueba F . Si se desea obtener el estadístico de Wald, entonces se requiere un tercer argumento `test="Chisq"`. Las siguientes líneas de código efectúan la respectiva prueba F y de Wald:

```
library(AER)
# prueba F
linearHypothesis( res1, c("CONCRET = 0", "OCCIDENTE = 0"))
## Linear hypothesis test
##
## Hypothesis:
## CONCRET = 0
## OCCIDENTE = 0
##
## Model 1: restricted model
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCRET +
## VALOREM + OCCIDENTE
##
## Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      1689 2080.8
## 2      1687 2076.6   2     4.1357 1.6799 0.1867

# prueba Wald test="Chisq"
linearHypothesis( res1, test="Chisq", c("CONCRET = 0", "OCCIDENTE = 0"))
## Linear hypothesis test
##
## Hypothesis:
## CONCRET = 0
## OCCIDENTE = 0
##
## Model 1: restricted model
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCRET +
## VALOREM + OCCIDENTE
##
## Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1      1689 2080.8
## 2      1687 2076.6   2     4.1357 3.3597    0.1864
```


Como se esperaba, el p-valor de ambas pruebas es el mismo. En este caso, no se puede rechazar la hipótesis nula; es decir, se puede concluir que las acciones de CONCRETO y OCCIDENTE no tienen efecto sobre el rendimiento de la acción de GRUPOSURA.

Ahora, noten que los coeficientes que acompañan los rendimientos de las acciones de ECOPETROL y EXITO parecen "a ojo" ser similares. Entonces probemos la siguiente hipótesis:

$$H_0 : \beta_2 = \beta_4 = 0,12$$

versus la hipótesis alterna de no H_0 . En este caso el código será:

```
## Linear hypothesis test
##
## Hypothesis:
## ECOPETROL = 0.12
## EXITO = 0.12
##
## Model 1: restricted model
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCRETO + VALOREM + OCCIDENTE
##
##      Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      1689 2076.7
## 2      1687 2076.6   2   0.028795 0.0117 0.9884
## Linear hypothesis test
##
## Hypothesis:
## ECOPETROL = 0.12
## EXITO = 0.12
##
## Model 1: restricted model
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCRETO + VALOREM + OCCIDENTE
##
##      Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
## 1      1689 2076.7
## 2      1687 2076.6   2   0.028795 0.0234      0.9884
```

Esto implica que la hipótesis nula no se puede rechazar. Es decir, se puede concluir que el efecto de un aumento de un punto porcentual en el rendimiento de las acciones de ECOPETROL o EXITO tiene el mismo efecto sobre la acción de GRUPOSURA. Es más dicho efecto no es estadísticamente diferente de 0.12 puntos porcentuales.

3.7. Apéndice

Apéndice 3.1 Demostración de la ecuación 3.7

La expresión 3.7 implica que en presencia de un modelo lineal con intercepto, podemos descomponer la variación total de la variable dependiente (SST) en la

parte explicada por el modelo de regresión (SSR) y la parte no explicada por el modelo (SSE). Formalmente, la proposición a probar es que si una de las variables explicativas es una constante (corresponde al intercepto), entonces:

$$SST = SSR + SSE$$

Para demostrar esta proposición, podemos partir del hecho que:

$$\begin{aligned}\hat{\varepsilon}^T \hat{\varepsilon} &= (y - \mathbf{X}\hat{\beta})^T (y - \mathbf{X}\hat{\beta}) \\ \hat{\varepsilon}^T \hat{\varepsilon} &= y^T y - 2\hat{\beta}^T \mathbf{X}^T y + \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} \\ \hat{\varepsilon}^T \hat{\varepsilon} &= y^T y - \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta}\end{aligned}$$

Remplazando, obtenemos

$$y^T y = \hat{\varepsilon}^T \hat{\varepsilon} + \hat{y}^T \hat{y}$$

Restando a ambos lados $n\bar{y}^2$, se obtiene:

$$y^T y - n\bar{y}^2 = \hat{\varepsilon}^T \hat{\varepsilon} - n\bar{y}^2 + \hat{y}^T \hat{y}$$

En otras palabras, tenemos que:

$$SST = SSE + \hat{y}^T \hat{y} - n\bar{y}^2$$

Ahora, será necesario demostrar que $SSR = \hat{y}^T \hat{y} - n\bar{y}^2$ siempre que el modelo lineal incluya un intercepto. Noten que:

$$\begin{aligned}SSR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \hat{y}_i^2 - 2 \sum_{i=1}^n \hat{y}_i \bar{y} + \sum_{i=1}^n \bar{y}^2 \\ SSR &= \hat{y}^T \hat{y} - 2\bar{y} \sum_{i=1}^n \hat{y}_i + n\bar{y}^2\end{aligned}$$

Remplazando el valor estimado de la variable dependiente, se tiene que:

$$\begin{aligned}SSR &= \hat{y}^T \hat{y} - 2\bar{y} \sum_{i=1}^n (y_i - \hat{\varepsilon}_i) + n\bar{y}^2 \\ SSR &= \hat{y}^T \hat{y} - 2\bar{y} \left[\sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\varepsilon}_i \right] + n\bar{y}^2 \\ SSR &= \hat{y}^T \hat{y} - 2\bar{y}n\bar{y} + 2\bar{y} \sum_{i=1}^n \hat{\varepsilon}_i + n\bar{y}^2\end{aligned}$$

Recordemos que si uno de los regresores es una constante, entonces ya se demostró que siempre se cumplirá que $\sum_{i=1}^n \hat{\epsilon}_i = 0$. Por eso,

$$SSR = \hat{y}^T \hat{y} - 2n\bar{y}^2 + n\bar{y}^2$$

$$SSR = \hat{y}^T \hat{y} - n\bar{y}^2$$

Es decir, se tiene que:

$$SST = SSE + SSR$$

Es importante anotar que si no se garantiza que $\sum_{i=1}^n \hat{\epsilon}_i = 0$, entonces no necesariamente se puede garantizar que la suma del SSR y el SSE sean iguales al SST . Y esto solo se puede asegurar cuando el modelo estimado tiene un intercepto.

Capítulo 4

Comparación de Modelos

Objetivos del capítulo

Al finalizar este capítulo, el lector estará en capacidad de:

- Emplear el R^2 ajustado, el AIC y el BIC para seleccionar el mejor modelo
- Realizar pruebas estadísticas que permitan seleccionar entre modelos anidado empleando R.
- Realizar pruebas estadísticas que permitan seleccionar entre modelos no anidados empleando R.

4.1. Introducción

En muchas ocasiones el científico de datos se enfrenta con el problema de tener que comparar diferentes modelos y escoger uno de ellos. En este capítulo nos concentraremos en dos tipos de aproximaciones, que no son excluyentes, para seleccionar modelos. La primera aproximación emplea medidas de bondad de ajuste y la segunda pruebas estadísticas. En lo que resta de este capítulo supondremos que queremos comparar modelos cuya variable dependiente es la misma y que emplean la misma muestra.

Pero antes de entrar en el detalle es importante anotar que al momento de comparar modelos que expliquen por ejemplo la variable y_i podemos enfrentar dos situaciones. Por ejemplo, consideremos los siguientes modelos para explicar las unidades vendidas de un SKU de una bebida carbonizada en el mes t (V_t):

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 pc_t + \beta_4 temp_t + \beta_5 PubliComp_t + \varepsilon_t \quad (4.1)$$

$$V_t = \beta_1 + \beta_2 p_t + \beta_3 pc_t + \varepsilon_t \quad (4.2)$$

$$V_t = \beta_1 + \beta_6 PubliTV_t + \beta_7 PubliRadio_t + \beta_8 PubliIntr_t + \varepsilon_t \quad (4.3)$$

donde p_t y pc_t denotan el precio por mililitro en el mes t de la bebida bajo estudio y del competidor mas cercano. $temp_t$ representa la temperatura promedio en el mes t en grados centígrados. $PubliComp_t$, $PubliTV_t$, $PubliRadio_t$ y $PubliIntr_t$ corresponden a la inversión en el mes t en millones de pesos en publicidad total del competidor mas cercano, la inversión propia en televisión, radio e Internet, respectivamente.

Noten que el modelo 4.1 es una versión restringida del modelo . Al modelo se le denomina modelo anidado (en el modelo) . Por otro lado, el modelo no se encuentra anidado en los modelos o .

4.2. Comparación de modelos empleando medidas de bondad de ajuste

En el capítulo anterior discutimos las limitaciones del R^2 al comparar modelos. Concluimos que el R^2 solo permite comparar modelos si estos tienen la misma variable dependiente y el mismo número de variables explicativas. Mostramos cómo el R^2 se infla con la inclusión de variables.

Para tener en cuenta la relación directa entre el SSR y el número de regresores (k), se ha diseñado el R^2 ajustado (\bar{R}^2). El \bar{R}^2 penaliza la inclusión de nuevas variables explicativas al modelo, esa penalización se da de la siguiente forma:

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

$\frac{n-1}{n-k}$ es el factor que penaliza el R^2 al incluir más variables explicativas, pues $\frac{n-1}{n-k}$ es creciente en k . En otras palabras, a medida que se incrementa el número de variables independientes en el modelo, $\frac{n-1}{n-k}$ aumenta al mismo tiempo que el R^2 aumenta y por tanto $(1 - R^2)$ disminuye. Así, el efecto de un aumento en el número de regresores no necesariamente implica un aumento en el \bar{R}^2 . El \bar{R}^2 aumentará únicamente si el aumento en el R^2 es lo suficientemente grande para compensar la penalización que se hace por incluir más variables. Así un \bar{R}^2 grande será mejor que uno pequeño.

De esta manera, con el \bar{R}^2 se pueden comparar los modelos 3.8, 3.10 y 3.11. El único problema que se presenta con el \bar{R}^2 , es que este carece de una interpretación así como la tiene el R^2 .

Existen otros problemas, que se estudiarán más adelante, que pueden “inflar” el R^2 . Así, aunque este estadístico presenta una interpretación muy clara e intuitiva, hay que tener cuidado antes de sacar conclusiones de este estadístico.

Existen otras medidas de bondad de ajuste que permiten comparar entre modelos (como el \bar{R}^2) pero que no tienen una interpretación como tal. Dos de estas dos medidas muy conocidas son el Criterio de Información de Akaike (por su nombre en inglés: *AIC*) y el Criterio de Información de Bayesiano¹ (por su nombre en inglés: *BIC*). Tanto el *AIC* como el *BIC* son estadísticos que fueron desarrollados en otra filosofía de estimación (Máxima Verosimilitud) y penalizan por la inclusión de mas variables explicativas como el \bar{R}^2 . Estas dos medidas adaptadas para el caso de la regresión múltiple estimada por MCO se definen como:

$$AIC = n + n \log(2\pi) + n \log\left(\frac{SSR}{n}\right) + 2(k+1)$$

$$BIC = n + n \log(2\pi) + n \log\left(\frac{SSR}{n}\right) + \log(n)(k+1)$$

Cuando empleamos el \bar{R}^2 para comparar modelos con la misma variable dependiente (y misma muestra) se selecciona el modelo que lo maximice. En el caso del *AIC* y *BIC*, se prefiere el modelo que minimice estos criterio. Por otro lado, la diferencia entre el *AIC* y el *BIC* es la forma en cómo se penaliza la inclusión de mas variables explicativas, por esto es posible que el modelo seleccionado sea diferente para estos dos criterios. Está muy documentado que el *AIC* tiene siempre una probabilidad de seleccionar modelos con más regresores que el *BIC*. Y por el otro lado, también se sabe que el *BIC* tiende a encontrar modelos relativamente pequeños en términos de variables explicativas. También es importante anotar que estas medidas de bondad de ajuste permiten comparar modelos anidados y no anidados.

En general, para la comparación de modelos se recomienda emplear los tres criterios (\bar{R}^2), *AIC*, *BIC*). Por ejemplo, supongamos un caso en el que *AIC* recomienda un modelo con 5 variables, el *BIC* recomienda 2 variables y el \bar{R}^2 sugiere 3 variables. En este caso se puede emplear los tres modelos seleccionados y comparar entre modelos empleando inferencia como veremos en la siguiente sección.

¹ Este criterio se conoce también como criterio de información de Schwarz y se reconocen por las siguientes siglas que vienen del inglés: SIC, SBC o SBIC.

4.3. Comparación de modelos empleando inferencia

4.3.1. Modelos anidados

En el capítulo anterior discutimos la inferencia para restricciones de la forma $R_{(r) \times k} \beta_{k \times 1} = C_{(r) \times 1}$. Discutimos como dichas restricciones se podían probar con una prueba F o con una prueba de Wald.

De hecho la comparación de modelos anidados es un caso especial de dichas pruebas. Es decir, cuando consideramos una hipótesis nula en la cual un grupo de coeficientes² es conjuntamente igual a cero, es decir $H_o : \beta_p = \beta_{p+1} = \dots = \beta_l = 0$, para $0 < p < l \leq k$ (versus la H_A : No H_o), esta prueba se conoce como una prueba de significancia conjunta. Es decir, la hipótesis nula es equivalente a $H_o : Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_{p-1} X_{p-1i} + \beta_{l+1} X_{l+1i} + \dots + \beta_k X_{ki} + \varepsilon_i$ (es decir un Modelo Restringido (R) del original) y la hipótesis alterna es $H_A : Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i$ (el Modelo sin restricción (U)). En este caso especial, el $F_{Calculado}$ de la expresión 3.12 es equivalente a:

$$F_c = \frac{(SSE_R - SSE_U)/r}{SSE_U/(n-k)}$$

donde $r = l - p$, SSE_R representa la suma de los cuadrados de los residuos estimados del modelo restringido y SSR_U representa la suma de los cuadrados de los residuos estimados del modelo sin restringir.

Así para probar la hipótesis nula que $H_o : \beta_p = \beta_{p+1} = \dots = \beta_l = 0$, para $0 < p < l \leq k$ (versus la H_A : No H_o), podemos estimar por MCO el modelo restringido, implicado por la hipótesis nula, como el modelo sin restringir y encontrar su correspondiente SSE . A partir de estas cantidades se puede calcular el $F_{Calculado}$, el cual se compara con $F_{\alpha, (r, (n-k))}$. A este tipo de test se le conoce como una prueba de modelo restringido versus modelo no restringido. Como se discusión en el capítulo anterior, este tipo de restricciones también se pueden probar con una prueba de Wald (función **waldtest**(del paquete **AER**).

4.3.2. Modelos no anidados

Formalmente, en este caso tendremos que la hipótesis nula de la prueba será

$$H_o : \mathbf{y} = \mathbf{X}\beta + \varepsilon_0$$

y la hipótesis alterna es

$$H_A : \mathbf{y} = \mathbf{Z}\gamma + \varepsilon_1$$

² Diferentes al término constante.

Para comprobar este tipo de hipótesis tenemos dos opciones: la prueba J y la prueba de Cox.

4.3.2.1. Prueba J

La idea de esta prueba es bastante evidente. La prueba J considera los dos modelos en uno. Es decir implica estimar el siguiente modelo:

$$\mathbf{y} = (1 - \lambda)\mathbf{X}\beta + (\lambda)\mathbf{Z}\gamma + \varepsilon$$

La idea de la prueba es primero estimar γ a partir de la regresión de \mathbf{y} en \mathbf{Z} . En un segundo paso correr una regresión de \mathbf{y} en \mathbf{X} y $\mathbf{Z}\hat{\gamma}$. Noten que si H_0 es verdad si $\lambda = 0$. Por tanto, la prueba implica el siguiente estadístico de prueba:

$$\frac{\hat{\lambda}}{s.e.(\hat{\lambda})}$$

Esta demostrado que este estadístico sigue una distribución aproximada a la estándar normal.

4.3.2.2. Prueba de Cox

La prueba de Cox es un tipo de prueba denominada razón de máxima verosimilitud (LR). Esta prueba implica el siguiente estadístico

$$q = \frac{c_{01}}{\sqrt{\frac{s_Z^2}{s_{ZX}^4 \mathbf{b}^T \mathbf{X}^T \mathbf{M}_Z \mathbf{M}_X \mathbf{M}_Z \mathbf{X} \mathbf{b}}}}$$

donde

$$c_{01} = \frac{n}{2} \ln \left[\frac{s_Z^2}{s_{ZX}^2} \right]$$

$$s_Z^2 = \frac{\mathbf{e}_Z^T \mathbf{e}_Z}{n}$$

$$s_X^2 = \frac{\mathbf{e}_X^T \mathbf{e}_X}{n}$$

$$s_{ZX}^2 = s_X^2 + \frac{\mathbf{b}^T \mathbf{X}^T \mathbf{M}_Z \mathbf{X} \mathbf{b}}{n}$$

Este estadístico sigue una distribución Chi-cuadrado con grados de libertad igual al número de parámetros en el

4.4. Práctica en R: Escogiendo el mejor modelo

Para mostrar como emplear R para seleccionar modelos, emplearemos unos datos simulados para una variable dependiente (y_i) y 10 variables explicativas X_j, i donde $j = 1, 2, \dots, 10$. Para cada variable se simulan 150 observaciones $i = 1, 2, \dots, 150$. Los datos están disponibles en el archivo *selModel.txt*. Carguemos los datos y constatemos que quedan bien cargados.

```
data <- read.table("Data/selModel.txt", header = TRUE, sep = ",")
head(data)

##      X      x1      x2      x3      x4      x5      x6      x7      x8      x9
## 1 1 5.925 6.512 6.883 5.157 6.420 5.833 5.356 6.626 5.730
## 2 2 4.565 3.638 4.378 5.056 3.165 4.598 4.576 4.171 4.259
## 3 3 5.436 5.501 6.596 4.476 4.992 5.524 6.235 5.402 5.465
## 4 4 4.912 5.849 6.545 5.715 4.668 5.259 5.786 5.070 5.499
## 5 5 5.696 6.330 5.279 5.017 4.229 5.769 4.468 4.771 6.062
## 6 6 4.009 4.584 4.220 6.463 5.511 4.917 5.920 3.492 3.219
##
##      x10      y
## 1 6.108 42.799
## 2 4.443 30.372
## 3 5.568 36.338
## 4 4.640 36.630
## 5 5.495 38.275
## 6 5.073 37.413

class(data)

## [1] "data.frame"

str(data)

## 'data.frame': 150 obs. of 12 variables:
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...
## $ x1 : num 5.92 4.57 5.44 4.91 5.7 ...
## $ x2 : num 6.51 3.64 5.5 5.85 6.33 ...
## $ x3 : num 6.88 4.38 6.6 6.54 5.28 ...
## $ x4 : num 5.16 5.06 4.48 5.71 5.02 ...
## $ x5 : num 6.42 3.17 4.99 4.67 4.23 ...
## $ x6 : num 5.83 4.6 5.52 5.26 5.77 ...
## $ x7 : num 5.36 4.58 6.24 5.79 4.47 ...
## $ x8 : num 6.63 4.17 5.4 5.07 4.77 ...
## $ x9 : num 5.73 4.26 5.46 5.5 6.06 ...
## $ x10: num 6.11 4.44 5.57 4.64 5.5 ...
## $ y : num 42.8 30.4 36.3 36.6 38.3 ...
```

Noten que se cargo una primera columna que corresponde al número de la observación, esto no lo necesitaremos. Eliminemos esa variable.

```
data <- data[, -1]
head(data, 3)

##      x1      x2      x3      x4      x5      x6      x7      x8      x9
```

```
## 1 5.925 6.512 6.883 5.157 6.420 5.833 5.356 6.626 5.730
## 2 4.565 3.638 4.378 5.056 3.165 4.598 4.576 4.171 4.259
## 3 5.436 5.501 6.596 4.476 4.992 5.524 6.235 5.402 5.465
##      x10      y
## 1 6.108 42.799
## 2 4.443 30.372
## 3 5.568 36.338
```

Ahora consideremos los siguientes tres modelos

$$y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \beta_5 X_{6,i} + \beta_6 X_{7,i} + \varepsilon_i \quad (4.4)$$

$$y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \varepsilon_i \quad (4.5)$$

$$y_i = \beta_1 + \beta_2 X_{4,i} + \beta_3 X_{5,i} + \beta_8 X_{8,i} + \beta_9 X_{9,i} + \beta_{10} X_{10,i} + \varepsilon_i \quad (4.6)$$

El modelo 4.5 esta anidado en el modelo 4.4. El modelo 4.6 no se encuentra anidado en ninguno de los otros modelos, ni al revés.

Procedamos a comparar los tres modelos empleando las métricas medidas de bondad de ajuste primero y luego empleando pruebas de hipótesis. Pero antes estimemos los tres modelos descritos arriba.

```
res1 <- lm(y ~ x1 + x2 + x3 + x6 + x7, data)
res2 <- lm(y ~ x1 + x2 + x3 , data)
res3 <- lm(y ~ x4 + x5 + x8 + x9 + x10, data)
```

Los resultados de estos tres modelos se presentan en el cuadro ??.

Cuadro 4.1 Modelos estimados por MCO

	y	y	y
intercepto	12.39*** [8.303]	13.176*** [9.233]	14.033*** [8.811]
x1	0.953*** [3.53]	1.048*** [3.945]	
x2	1.936*** [6.556]	2.133*** [7.778]	
x3	1.038*** [3.576]	1.162*** [4.203]	
x6	0.412 [1.517]		
x7	0.167 [0.583]		
x4			1.429*** [4.67]
x5			1.533*** [5.122]
x8			0.443 [1.391]
x9			0.175 [0.556]
x10			0.589* [1.768]
R^2	0.651	0.643	0.566
$adj.R^2$	0.639	0.636	0.551
N	150	150	150
t-values in brackets			
* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)			

4.4.1. *Medidas de bondad de ajuste*

El \bar{R}^2 y los criterios de información *AIC* y *BIC* se pueden calcular fácilmente en R. Para calcular el \bar{R}^2 debemos emplear la función *summary* y extraer del objeto que crea *summary* el \bar{R}^2 de la siguiente manera:

```
R1 <- summary(res1)
R1$adj.r.squared
## [1] 0.6388607
```

Los criterios de información se calculan de la siguiente manera:

```
AIC(res1)
## [1] 733.9571
BIC(res1)
## [1] 755.0315
```

Ustedes pueden calcular estos indicadores para los otros modelos y encontrarán los resultados que se reportan en el siguiente cuadro.

Cuadro 4.2 Medidas de bondad de ajuste para los tres modelos estimados

	Modelo 1	Modelo 2	Modelo 3
R2.ajustado	0.639	0.636	0.551
AIC	733.957	733.323	766.718
BIC	755.032	748.376	787.792

El \bar{R}^2 sugiere que el mejor modelo es el 1 (4.4), mientras que los dos criterios de información sugieren que el mejor modelo es el 2 (4.5). Ahora veamos que decisión tomamos al emplear pruebas estadísticas.

4.4.2. Pruebas estadísticas

4.4.2.1. Modelos anidados

Como se mencionó anteriormente, el modelo 4.5 esta anidado en el modelo 4.4. Comparemos estos dos modelos empleando la función *anova()* empleando dos argumentos: primero el modelo restringido y segundo el modelo sin restringir. En este caso tenemos

```
anova(res2, res1)

## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x1 + x2 + x3 + x6 + x7
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      146 1091.1
## 2      144 1066.9   2    24.207 1.6337 0.1988
```

En este caso el estadístico F es igual a 1.63 y el respectivo valor p es 0.1988. Por tanto no se puede rechazar la hipótesis nula de que el modelo restringido (el modelo 2; es decir 4.5) es mejor que el modelo sin restringir (modelo 1; es decir 4.4).

4.4.2.2. Modelos no anidados

Ahora comparemos el modelo 3 (4.6) con los otros dos modelos. La prueba J se calcula de la siguiente manera

```
library(AER)
J.res1.3 <- jtest(res1, res3)
J.res1.3
```

```
## J test
##
## Model 1: y ~ x1 + x2 + x3 + x6 + x7
## Model 2: y ~ x4 + x5 + x8 + x9 + x10
##
##           Estimate Std. Error t value Pr(>|t|)
## M1 + fitted(M2)  0.48221    0.103789  4.646 7.606e-06 ***
## M2 + fitted(M1)  0.81939    0.095489  8.581 1.427e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El primer $t_{calculado}$ (4.65) es el que permite probar la hipótesis nula de que el modelo 1 (4.4) es mejor que el modelo 3 (4.6). En este caso se rechaza la nula (p valor de 0), lo cual significa que la prueba no puede concluir en favor del modelo 1.

Por otro lado, el segundo $t_{calculado}$ (8.58) es el que permite probar la hipótesis nula de que el modelo 3 (4.6) es mejor que el modelo 1 (4.5). En este caso también se rechaza la nula (p valor de 0), lo cual significa que la prueba tampoco puede concluir en favor del modelo 3. Es decir, esta prueba no es concluyente.

Para el caso de la comparación del modelo 2 (4.5) y el modelo 3 (??) se obtiene un resultado similar

```
J.res2.3 <- jtest(res2, res3)
J.res2.3

## J test
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x4 + x5 + x8 + x9 + x10
##
##           Estimate Std. Error t value Pr(>|t|)
## M1 + fitted(M2)  0.46880    0.093090  5.0360 1.391e-06 ***
## M2 + fitted(M1)  0.78589    0.090552  8.6788 8.131e-15 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora empleemos la prueba de Cox. que tiene una interpretación similar.

```
Cox.res1.3<-coxtest(res1, res3)
Cox.res1.3

## Cox test
##
## Model 1: y ~ x1 + x2 + x3 + x6 + x7
## Model 2: y ~ x4 + x5 + x8 + x9 + x10
##
##           Estimate Std. Error z value Pr(>|z|)
## fitted(M1) ~ M2 -20.255      4.4843 -4.5169 6.275e-06 ***
## fitted(M2) ~ M1 -44.422      4.1659 -10.6632 < 2.2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cox.res2.3<-coxtest(res2, res3)
Cox.res2.3
```

```
## Cox test
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x4 + x5 + x8 + x9 + x10
##
##           Estimate Std. Error   z value   Pr(>|z|)
## fitted(M1) ~ M2    -24.191      4.6851   -5.1634 2.425e-07 ***
## fitted(M2) ~ M1    -47.896      4.1226  -11.6180 < 2.2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados de esta prueba son los mismos que obtuvimos con la prueba J (no siempre coinciden los resultados de estas dos pruebas). No es posible concluir en favor de un modelo.

Así, uniendo todos los resultados encontramos que el modelo 3 (4.6) no es mejor que los modelos 1 (4.4) y 2 (4.5). Este resultado lo obtenemos con las pruebas Cox y J y las medidas de bondad de ajuste. Y entre el modelo 1 (4.4) y 2 (4.5), la prueba estadística y los dos criterios de información sugieren que el modelo 2 (4.5) es mejor modelo. En el caso del \bar{R}^2 no existe una diferencia muy grande entre los dos modelos. Por eso seleccionamos como mejor modelo el 2 (4.5).

4.5. Ejercicios

Continuando con el ejercicio del Capítulo 2, responda:

1. El modelo: $I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \alpha_4 LDies_t + \alpha_5 LEI_t + \alpha_6 V_t + \varepsilon_t$ generó mucha discusión en la pequeña República y en diversos foros se sugirieron diferentes modelos:

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \varepsilon_t \quad (4.7)$$

$$I_t = \alpha_1 + \alpha_2 LDies_t + \alpha_3 LEI_t + \varepsilon_t \quad (4.8)$$

$$I_t = \alpha_1 + \alpha_2 V_t + \varepsilon_t \quad (4.9)$$

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 LEI_t + \varepsilon_t \quad (4.10)$$

$$I_t = \alpha_1 + \alpha_2 CD_t + \alpha_3 LDies_t + \varepsilon_t \quad (4.11)$$

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 LDies_t + \alpha_4 LEI_t + \alpha_5 V_t + \varepsilon_t \quad (4.12)$$

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \alpha_4 V_t + \varepsilon_t \quad (4.13)$$

Determine cuál es el mejor modelo para explicar los ingresos del sector ferroviario en esta pequeña república (muestre todos los cálculos necesarios para tomar esta decisión y asegúrese que su respuesta está bien argumentada).

2. A partir del modelo que seleccionó en la pregunta anterior:
 - a) Analice la significancia individual de los coeficientes
 - b) Analice la significancia conjunta de los coeficientes

- c)* Determine cuál es el factor que más afecta el ingreso del sector ferroviario en esta nación
- d)* De acuerdo con su resultado, ¿qué puede sugerir al gobierno de esta nación para mejorar los ingresos del sector?

Capítulo 5

Variables dummy

Objetivos del capítulo

Al terminar la lectura de este capítulo el lector estará en capacidad de:

- Crear variables dummy utilizando R.
- Estimar modelos econométricos con variables dummy que permitan comprobar diferentes hipótesis.

5.1. Introducción

En los capítulos anteriores hemos discutido el uso de variables que toman valores en un dominio continuo en los modelos de regresión lineal. Sin embargo, en algunas ocasiones será necesario emplear variables que toman únicamente dos valores distintos. Por ejemplo, existirán algunas ocasiones en que para la estimación de un modelo sería relevante considerar si un individuo es mujer o no, si la observación es de un país que pertenece a determinado grupo económico o no o si una determinada variable superó un umbral o no.

Ejemplo 5.1 Un ejemplo sencillo

Suponga que un equipo de fútbol presenta dos volúmenes de ventas de entradas a los partidos (Y_i) bastante distintos:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

donde $X_i = 1$ si juega una estrella en el equipo contrario, cero en caso contrario.

Podemos entonces ver que las ventas de boletos están modeladas de diferente manera para las dos situaciones. En este caso, es como si tuviésemos dos modelos en uno. Si juega una estrella en el equipo contrario ($X_i = 1$) el modelo será

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Y si no juega una estrella ($X_i = 0$) el modelo será

$$Y_i = \beta_1 + \varepsilon_i$$

En este caso, β_2 representa la diferencia que presentan las ventas de entradas a los partidos cuando en el equipo contrario juega una súper estrella.

En este capítulo vamos a discutir las variables dummy, también conocidas como variables ficticias, variables indicador, variables ficticias o variables dicotómicas.

Las variables dummy pueden servir para:

- medir el efecto de “tratamientos” sobre la variable de respuesta
- “borrar” el efecto de una observación atípica (esta es una opción para no eliminar los outliers)
- incluir un conjunto de categorías
- agrupar las observaciones en diferentes grupos
- incluir variables categóricas
- efectos de umbral (Threshold Effects)
- detectar cambio estructural

Veamos un ejemplo de cómo las variables dummy permiten encontrar el efecto “tratamientos”. En los mercados accionarios se ha encontrado que su comportamiento en promedio no es igual todos los días; es decir, los días se convierten en un “tratamientos” diferente. En estos caso el efecto del día de la semana (DOW¹, por su sigla en inglés: Day Of the Week) se puede capturar creando variables dummy para los días de la semana. En este contexto el rendimiento diario de una acción (R_t) se puede modelar de acuerdo al modelo CAPM² empleando el rendimiento de un activo libre de riesgo ($R_t^{R.F.}$) y se puede incluir el efecto DOW de la siguiente manera:

$$R_t = \beta_1 + \beta_2 R_t^{R.F.} + \delta_1 D_{1t} + \delta_2 D_{2t} + \delta_3 D_{3t} + \delta_4 D_{4t} + \varepsilon_t$$

El lector puede constatar que este modelo implica un rendimiento promedio para cada día de la semana laboral diferente³.

Un ejemplo de cómo emplear las variables ficticias para medir efectos de umbral. Por ejemplo, los años de educación pueden no tener efecto directo sobre el ingreso de las personas ($income_i$), mas bien los umbrales de educación superados si pueden tener un efecto. El siguiente modelo refleja este efecto:

$$income_i = \beta_1 + \beta_2 age_i + \delta_1 B_i + \delta_2 M_i + \delta_3 P_i + \varepsilon_i$$

donde $B_i = 1$ si el máximo título es pregrado y cero en caso contrario. $M_i = 1$ si el máximo título es maestría y cero en caso contrario. $P_i = 1$ si el máximo título es Ph.D. y cero en caso contrario. El lector puede constatar cómo este modelo captura un efecto de umbral de los años de educación sobre el ingreso.

Finalmente, veamos un ejemplo de cómo emplear las variables dummy para capturar cambio estructural. Por cambio estructural se entiende un cambio en cómo se comporta el DGP a partir de un periodo determinado. Por ejemplo, el consumo en un país en el periodo t (C_t) puede cambia su comportamiento después de un periodo determinado t^* . El cambio (estructural) en la relación del C_t y del ingreso disponible de los hogares en el periodo t (Yd_t) se puede modelar empleando una variable dummy (D_t) que toma el valor de uno antes del periodo t^* y cero en caso contrario. El modelo sería el siguiente.

$$C_t = \beta_1 + \beta_2 Yd_t + \alpha D_t + \delta D_t Yd_t + \varepsilon_t$$

El lector debería corroborar que esto implica un cambio tanto en la parte del consumo que no depende del ingreso, así como de aquella que si depende de este.

En las siguientes secciones estudiaremos en detalle el efecto de incluir variables dummy. Pero antes de continuar es importante mencionar que cuando se crean variables dummy siempre se deben crear una menos de las opciones disponibles. Es decir, ustedes notarán que en el caso de los días de la semana laboral (efecto DOW)

¹ Ver **AlonsoC.2008b** para una introducción al modelo CAPM

² Ver **Alonso2013b** para una introducción al modelo CAPM.

³ Noten que esta es una alternativa a tener un modelo ANOVA de una o dos vías que permite incluir variables continuas.

se crearon 4 variables dicotómicas y no 5. Y así para todos los ejemplos descritos anteriormente. EN capítulos posteriores se explicará con mayor claridad por qué hacemos esto. Por ahora es importante recordar que si existen p posibilidades, entonces debemos crear $p - 1$ variables ficticias.

5.2. Usos de las variables dummy

Para comprender mejor el uso de las variables ficticias emplearemos un ejemplo sencillo de modelo que solo emplea una variable explicativa. Como ejemplo emplearemos la función de consumo keynesiana. Supondremos que esta relación se comporta de manera diferente en tiempos de contracción económica que en años de expansión. En este ejemplo podemos plantear los cuatro casos posibles en que el consumo agregado está determinado por el ingreso disponible.

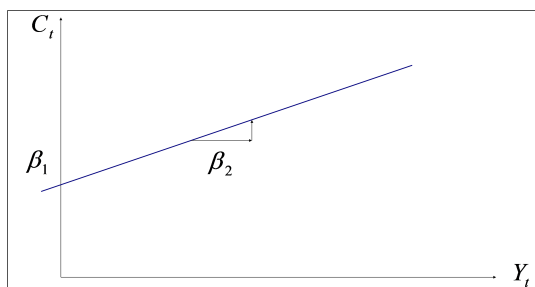
5.2.1. Caso I. La función es la misma

En este caso se asume que la función de consumo es igual tanto en tiempos de contracción como en tiempos de expansión. Esto implica la siguiente relación:

$$C_t = \beta_1 + \beta_2 Y_t + \varepsilon_t \quad (5.1)$$

Como podemos apreciar en la gráfica, tanto la pendiente como el intercepto de la función se mantienen inalterados en todo el período de estudio.

Figura 5.1 Caso I. No hay cambio



En este caso β_2 representa la propensión marginal a consumir y β_1 el consumo autónomo.

5.2.2. Caso II. Cambio en intercepto.

En el segundo caso, tenemos que el consumo es menor en tiempo de contracción que en tiempos de expansión (ceteris paribus). Esto se puede representar con una reducción del consumo autónomo (intercepto de la función) pero manteniéndose inalterada la propensión marginal al consumo (pendiente). Este efecto lo recoge el siguiente modelo:

$$C_t = \beta_1 + \beta_2 Y_t + \alpha D_t + \varepsilon_t \quad (5.2)$$

donde $D_t = 1$ si se trata de un periodo de contracción y cero en caso contrario.

Como podemos apreciar, la ecuación (5.2) sugiere que en periodos de contracción el valor que toma la variable dummy es de 1. Por lo tanto, en tiempo de contracción el modelo será:

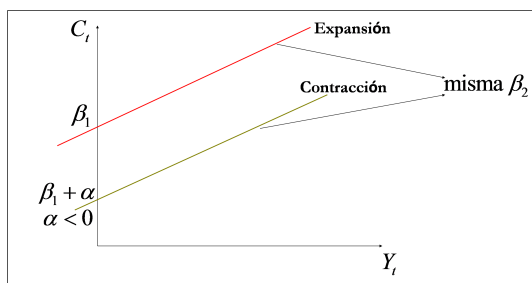
$$C_t = (\beta_1 + \alpha) + \beta_2 Y_t + \varepsilon_t$$

En tiempos de expansión, la variable dummy toma el valor de 0 y por lo tanto el modelo que describe el consumo se reduce a:

$$C_t = \beta_1 + \beta_2 Y_t + \varepsilon_t$$

Gráficamente esto se puede representar de la siguiente manera.

Figura 5.2 Caso II. Cambio en el intercepto



En este caso, α representa la diferencia del consumo autónomo entre el periodo de contracción y de expansión. Además, se espera que $\alpha < 0$. Noten que este caso genera rectas paralelas.

5.2.3. Caso III. Cambio en pendiente

En este caso, supongamos que el consumo es nuevamente menor en tiempo de contracción que en tiempos de expansión, pero a diferencia del Caso II en el que solo cambiaba el intercepto de la función, en este caso se presenta un cambio de pendiente (los individuos consumen menos por cada unidad monetaria adicional de ingreso disponible). Este hecho lo podemos representar con el siguiente modelo:

$$C_t = \beta_1 + \beta_2 Y_t + \gamma (D_t Y_t) + \varepsilon_t \quad (5.3)$$

donde D_t se define igual que antes; igual a uno si se trata de un periodo de contracción y cero en caso contrario.

La ecuación (5.4) nos muestra que en tiempo de contracción el valor que toma la variable dummy es de 1, produciéndose un cambio en pendiente. Es decir, tenemos que en tiempos de contracción el modelo que describe el consumo toma la siguiente forma:

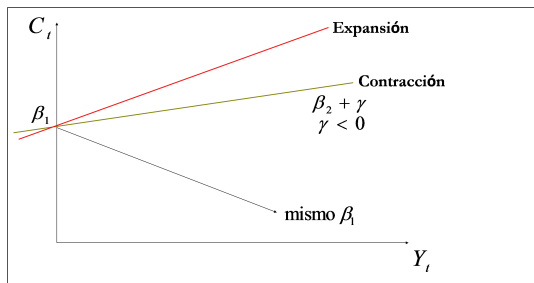
$$C_t = \beta_1 + (\beta_2 + \gamma) Y_t + \varepsilon_t$$

En tiempos de expansión la variable dummy toma el valor de 0 y por lo tanto el modelo que viene dado por:

$$C_t = \beta_1 + \beta_2 Y_t + \varepsilon_t$$

Gráficamente tenemos:

Figura 5.3 Caso III. Cambio en pendiente



En este caso γ corresponde a la diferencia entre la propensión marginal a consumir en periodos de contracción y expansión. En este caso se espera que $\gamma < 0$. Antes de pasar al siguiente caso, es importante mencionar que este tipo de modelos se denominan con interacción entre las variables. Es de decir, la variable dummy interactúa con la variable Y_d .

5.2.4. Caso IV. Cambio en intercepto y pendiente

En este último caso, el consumo se reduce en tiempos de contracción por las dos posibles vías; es decir, por una reducción tanto en el intercepto como en la pendiente. Este efecto es capturado por un modelo con la siguiente especificación:

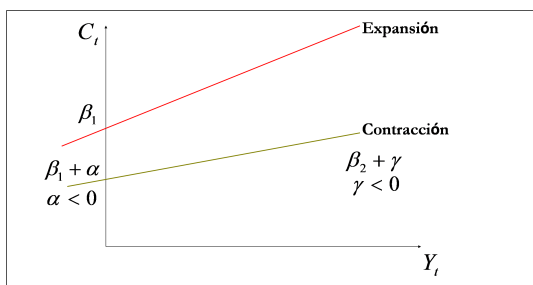
$$C_t = \beta_1 + \beta_2 Y_t + \alpha D_t + \gamma (D_t Y_t) + \varepsilon_t \quad (5.4)$$

La variable dummy se define igualmente que en los dos casos anteriores.

Análogamente, el modelo en periodos de contracción económica será: o de la siguiente forma:

$$C_t = (\beta_1 + \alpha) + (\beta_2 + \gamma) Y_t + \varepsilon_t \quad (5.5)$$

Figura 5.4 Caso IV. Cambio en intercepto y pendiente



En este caso α y γ representan respectivamente la diferencia del consumo autónomo y de la propensión marginal a consumir en periodos de contracción y de expansión. Se espera que α y γ sean negativos.

5.3. Práctica en R

Existen múltiples formas de crear variables dummy en R. Para mostrar diferentes formas de hacerlos emplearemos dos ejemplos. Por otro lado, es importante mencionar que si una variable es leída como un factor, la función `lm()` convierte automáticamente dicha variable en una dummy⁴.

⁴ Si se desea cambiar el grupo de referencia (el valor para el cual la dummy toma el valor de cero), se puede emplear la función `relevel(data.frame$variable, ref =)`, que toma como argumentos la variable del data frame que se quiera cambiar y ref con el nombre del nivel del factor (entre comillas) que se quiere emplear como la referencia.

5.3.1. Relación entre la economía mundial y la colombiana

Con este ejercicio ilustraremos cómo generar variables dummy con R en una serie de tiempo. Nos basaremos en modelar la denominada Ley Thirlwall formulada en 1979, la cual implica que para una economía abierta, la demanda impone restricciones al crecimiento dadas las diferentes dinámicas que ésta presenta en cada uno de los países y que ocasiona que crezcan a tasas diferentes.⁵

El modelo para probar la Ley de Thirlwall implica un modelo lineal del logaritmo del PIB (Y_t) del país bajo estudio en función del logaritmo del PIB del resto del mundo (f_t), es decir:

$$Y_t = \beta_0 + \beta_1 f_t + \varepsilon_t \quad (5.6)$$

donde β_0 representa el intercepto, β_1 la razón entre las elasticidades ingreso de las exportaciones e importaciones y ε_t un término de error aleatorio.

En su momento, Thirlwall demostró que si β_1 es igual a 1 significa que las tasas de crecimiento de largo plazo de la economía en estudio y las del resto de la economía son similares. Por otra parte, si β_1 es mayor que 1, entonces el desempeño de la economía en estudio está por encima de la del resto del mundo; es decir, el caso deseable. En consecuencia se obtendrá el resultado contrario cuando β_1 es menor que 1. Además encontró que si β_1 es mayor que 1, entonces la elasticidad ingreso de las exportaciones es superior a la de las importaciones; por tanto, a medida que se incrementa el ingreso de las familias, una mayor proporción del incremento del ingreso se destina a las importaciones. Desde otro punto de vista, esto implica que si la tasa de crecimiento del resto del mundo aumenta, entonces se puede elevar el crecimiento interno de la economía nacional en una mayor cuantía que el incremento que se dio en el resto del mundo. Por otra parte, si $0 < \beta_1 < 1$, entonces ante incrementos en el nivel de crecimiento de la economía mundial, la economía doméstica crecerá en menor cuantía a este aumento. Esta es otra razón por la cual es preferible que β_1 se aleje de 1.

Para estimar la ecuación (5.6) es necesario seleccionar una variable que refleje el comportamiento de la economía mundial, en nuestro caso seleccionamos el logaritmo del PIB de Estados Unidos (f_t) como una “proxy” del PIB mundial. Los datos de los logaritmos del PIB colombiano y el de Estados Unidos (a dólares constantes de 2010) se encuentran en el archivo *Thirlwall.csv*. Cargue los datos guárdelos en un objeto que llamaremos *data* y cuya clase sea *data.frame*.

Si analizamos los datos por medio de un gráfico de dispersión podemos ver que parece haber un cambio en la relación entre el PIB de Colombia con respecto al de Estados Unidos al comenzar la década de los noventa (figura 5.5). Sin embargo, el análisis gráfico no es suficiente para determinar si en efecto existe un cambio en el intercepto, en la pendiente o en ambos.

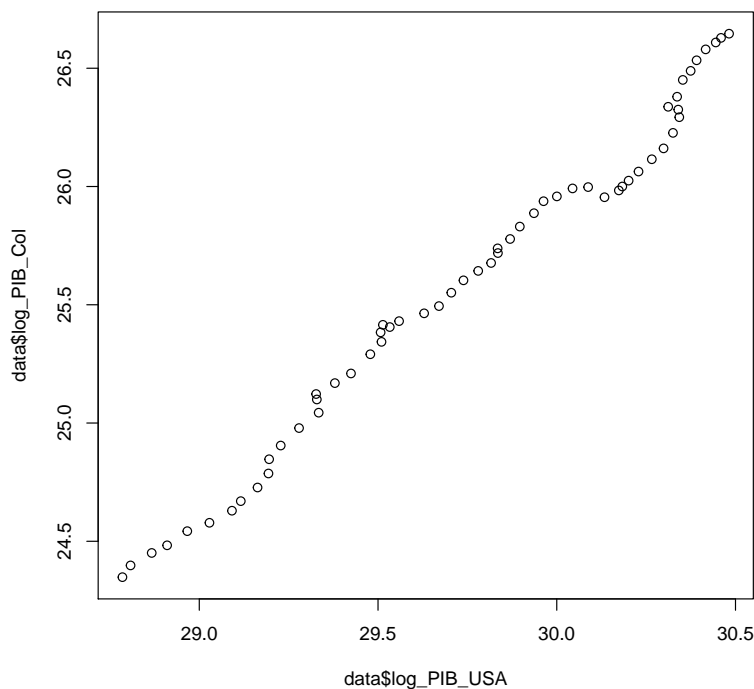
Para verificar la hipótesis de que la relación ha cambiado después de 1990 (período de la llamada apertura económica) se considera el siguiente modelo

$$Y_t = \beta_0 + \beta_1 f_t + \alpha_2 D_t + \beta_2 (D_t f_t) + \varepsilon_t$$

⁵ Para mayor información ver **thirlwall1**.

Figura 5.5 PIB real de Colombia con respecto al de Estados Unidos (en logaritmos)

```
plot(data$log_PIB_USA, data$log_PIB_Col)
```



donde D_t es 1 si la observación es del año es 1991 o posterior y cero en caso contrario. El lector deberá asegurarse de que este modelo genere un cambio tanto en intercepto como en pendiente.

Una vez las series del logaritmo del PIB de Colombia y de Estados Unidos sean leídas en R, necesitamos crear una variable dummy. R provee una variedad de posibilidades para la creación de variables dummy. En este caso emplearemos una prueba lógica sobre la variable *year* del data.frame *data*.

```
D <- data$year >= 1991
data$D <- as.numeric(D)
head(data)

##   X year log_PIB_Col log_PIB_USA D
## 1 1 1960    24.34832    28.78503 0
## 2 2 1961    24.39796    28.80777 0
```

```
## 3 3 1962      24.45067      28.86698 0
## 4 4 1963      24.48301      28.91004 0
## 5 5 1964      24.54285      28.96642 0
## 6 6 1965      24.57822      29.02846 0
```

```
tail(data)
```

```
##      X year log_PIB_Col log_PIB_USA D
## 53 53 2012      26.48895      30.37458 1
## 54 54 2013      26.53361      30.39121 1
## 55 55 2014      26.57981      30.41658 1
## 56 56 2015      26.60894      30.44479 1
## 57 57 2016      26.62837      30.45954 1
## 58 58 2017      26.64611      30.48202 1
```

```
data[29:36,]
```

```
##      X year log_PIB_Col log_PIB_USA D
## 29 29 1988      25.64314      29.78022 0
## 30 30 1989      25.67671      29.81637 0
## 31 31 1990      25.71864      29.83538 0
## 32 32 1991      25.73846      29.83464 1
## 33 33 1992      25.77811      29.86957 1
## 34 34 1993      25.83056      29.89666 1
## 35 35 1994      25.88708      29.93624 1
## 36 36 1995      25.93780      29.96307 1
```

Ahora sí podemos estimar el modelo deseado (asgúrese que puede estimarlo). Los resultados se presentan en la siguiente salida. Noten que la interacción entre la dummy y el logaritmo del PIB de los Estados Unidos se debe incluir en la fórmula de la siguiente manera $D \cdot \log_{PIB_{USA}}$.

```
##
## Call:
## lm(formula = log_PIB_Col ~ D + log_PIB_USA + D * log_PIB_USA,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14842 -0.04968  0.01334  0.06468  0.11968
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -15.56628    1.42221  -10.945 2.54e-15 ***
## D              0.92476    2.84473   0.325   0.746
## log_PIB_USA    1.38503    0.04846  28.580 < 2e-16 ***
## D:log_PIB_USA -0.03374    0.09486  -0.356   0.723
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08094 on 54 degrees of freedom
## Multiple R-squared:  0.9859, Adjusted R-squared:  0.9851
## F-statistic: 1260 on 3 and 54 DF, p-value: < 2.2e-16
```

Noten que ni el cambio en la pendiente (Coeficiente asociado a D_t) ni el cambio en la pendiente (asociado a $D_t f_t$) son individualmente significativos. Es decir, parece no existir ese cambio estructural después de los años noventa. Es más podemos emplear una prueba de modelos anidados que discutimos anteriormente para corroborar si estos dos coeficientes estimados son iguales a cero simultáneamente. Esto se logra de la siguiente manera.

```
anova(res2, res1)

## Analysis of Variance Table
##
## Model 1: log_PIB_Col ~ log_PIB_USA
## Model 2: log_PIB_Col ~ D + log_PIB_USA + D * log_PIB_USA
##   Res.Df    RSS Df Sum of Sq   F Pr(>F)
## 1      56 0.38276
## 2      54 0.35373   2    0.02903 2.2158 0.1189
```

Estos resultados muestran que no es posible rechazar la hipótesis nula de que los cambios en pendiente e intercepto son cero.

Ahora construya una dummy ($D2_t$) que tome el valor de uno únicamente para los años comprendidos entre 1991 y 2000. Determine si existe o no un cambio estructural⁶. Los resultados de los tres modelos se presentan en la siguiente tabla.

Cuadro 5.1 Modelos estimados por MCO para la Ley de Thirlwall

	$\log PIB_{Col}$	$\log PIB_{Col}$	$\log PIB_{Col}$
intercepto	-15.566*** [-10.945]	-13.134*** [-20.396]	-13.322*** [-20.924]
D	0.925 [0.325]		
$\log PIB_{USA}$	1.385*** [28.58]	1.302*** [60.146]	1.308*** [61.027]
D: $\log PIB_{USA}$	-0.034 [-0.356]		
D2			17.833** [2.557]
D2: $\log PIB_{USA}$			-0.595** [-2.558]
R^2	0.986	0.985	0.986
$adj.R^2$	0.985	0.984	0.986
N	58	58	58

t-values in brackets

* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)

Noten que en esta última ecuación estimada, el cambio en la pendiente e intercepto sí son significativos individualmente (con un nivel de confianza del 95%). Y conjuntamente también lo son con un nivel de significancia del 95%, como se muestra a continuación.

⁶ Ayuda: puede emplear el signo para hacer dos pruebas lógicas al mismo tiempo. Es el equivalente a τ^2 .

```
anova(res2, res3)

## Analysis of Variance Table
##
## Model 1: log_PIB_Col ~ log_PIB_USA
## Model 2: log_PIB_Col ~ D2 + log_PIB_USA + D2 * log_PIB_USA
##   Res.Df    RSS Df Sum of Sq   F    Pr(>F)
## 1      56 0.38276
## 2      54 0.34098   2   0.041776 3.308 0.04414 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Así, podemos rechazar la hipótesis nula que $\alpha_2 = \beta_2 = 0$ para la década de los 90. En otras palabras, efectivamente existe un cambio estructural en la relación en la elasticidad del PIB de Colombia con respecto al de Estados Unidos (ver figura ??) durante la década de los 90, pero dicha relación regreso a su nivel anterior después de ese periodo.

5.3.2. Creando variables dummy con el paquete dummies

El paquete *dummies* (**(dummiesR)**) es bastante flexible y permite crear variables dummy de diferentes maneras. Veamos un par de ejemplos. Creemos una variable que se llame letras y veamos su clase:

```
letras <- c( "a", "a", "b", "c", "d", "e", "f", "g", "h", "b", "b" )
letras

## [1] "a" "a" "b" "c" "d" "e" "f" "g" "h" "b" "b"

class(letras)

## [1] "character"
```

Ahora podemos emplear la función **dummies()** del paquete con el mismo nombre para crear variables dicotómicas. Esta función puede tomar argumentos que sean caracteres, o factores, no importa, en todo caso crea las variables dummy.

```
library(dummies)
dummy(letras)

##           book.Rnwa book.Rnwb book.Rnwc book.Rnwd book.Rnwe
## [1,]             1             0             0             0
## [2,]             1             0             0             0
## [3,]             0             1             0             0
## [4,]             0             0             1             0
## [5,]             0             0             0             1
## [6,]             0             0             0             1
## [7,]             0             0             0             0
## [8,]             0             0             0             0
```

```
## [9,] 0 0 0 0 0
## [10,] 0 1 0 0 0
## [11,] 0 1 0 0 0
##      book.Rnw:f book.Rnw:g book.Rnw:h
## [1,] 0 0 0
## [2,] 0 0 0
## [3,] 0 0 0
## [4,] 0 0 0
## [5,] 0 0 0
## [6,] 0 0 0
## [7,] 1 0 0
## [8,] 0 1 0
## [9,] 0 0 1
## [10,] 0 0 0
## [11,] 0 0 0

letras <- as.factor(letras)
letras

## [1] a a b c d e f g h b b
## Levels: a b c d e f g h

dummy(letras, sep=":")

##      book.Rnw:a book.Rnw:b book.Rnw:c book.Rnw:d
## [1,] 1 0 0 0
## [2,] 1 0 0 0
## [3,] 0 1 0 0
## [4,] 0 0 1 0
## [5,] 0 0 0 1
## [6,] 0 0 0 0
## [7,] 0 0 0 0
## [8,] 0 0 0 0
## [9,] 0 0 0 0
## [10,] 0 1 0 0
## [11,] 0 1 0 0
##      book.Rnw:e book.Rnw:f book.Rnw:g book.Rnw:h
## [1,] 0 0 0 0
## [2,] 0 0 0 0
## [3,] 0 0 0 0
## [4,] 0 0 0 0
## [5,] 0 0 0 0
## [6,] 1 0 0 0
## [7,] 0 1 0 0
## [8,] 0 0 1 0
## [9,] 0 0 0 1
## [10,] 0 0 0 0
## [11,] 0 0 0 0
```

Finalmente, es importante mencionar que se debe tener cuidado, pues este paquete creará p variables dummy si hay p posibilidades. Así al momento de emplear esta variables en un modelo de regresión se deberá omitir alguna de las variables

dummy para solo tener $p - 1$ variables⁷. Mas adelante se explicará la razón técnica para esto.

⁷ Otra opción es emplear la p variables dummy y omitir el intercepto. Pero en la mayoría de los casos esta no es una buena idea.

Capítulo 6

Selección automática de modelos

Objetivos del capítulo

Al finalizar este capítulo, el lector estará en capacidad de:

- Explicar en sus propias palabras las opciones que existen para seleccionar automáticamente el mejor modelo
- Seleccionar el mejor modelo de regresión dada una cantidad grande de regresores empleando R.

6.1. Introducción

En los capítulos anteriores hemos discutido cómo seleccionar el mejor modelo. Para ejemplificar la selección de modelos hemos empleado casos en los que se cuenta con un número relativamente reducido de posibles variables explicativas. Pero en la práctica es común que esto no ocurra, el científico de datos típicamente se encuentra con problemas en los que se conoce claramente la variables que se quiere explicar (variables dependiente) y un conjunto grande de posibles variables explicativas.

En general, si hay $k - 1$ variables independientes potenciales (además del intercepto) que pueden explicar a la variable dependiente, entonces hay $2^{(k-1)}$ subconjuntos de variables explicativas posibles para explicar a la variable dependiente. En la práctica desconocemos cuál de esos modelos es el correcto. Y por tanto tendríamos que probar todos los posibles modelos para encontrar el correcto. Ahora bien, cuando se cuenta con una teoría este problema no existe. Pero en la mayoría de los casos no contamos con teoría que nos apoye en el proceso. Entonces, por ejemplo si tenemos 10 ($(k - 1) = 10$) posibles variables explicativas entonces los posibles modelos serán $2^{(k-1)} = 2^{(10)} = 1024$. Si se tienen 20 variables candidatas, el número de posibles subconjuntos es de $1,048576 \times 10^6$ (mas de un millón de posibles modelos). Si son 25 posibles variables, entonces son $3,3554432 \times 10^7$. Es decir, aproximadamente 33 millones y medio de posibles modelos.

Así, si se cuenta con muchas variables no es viable calcular todos los posibles modelos. No obstante, si las variables son pocas podría ser viable estimar todos los posibles modelo. Es decir, si no tenemos teoría que soporte el modelaje y tenemos muchas variables, entonces tendremos muchos potenciales modelos a evaluar “manualmente”. Recordemos que en los capítulos anteriores vimos cómo comparar modelos para seleccionar el mejor empleando métricas como el R^2 ajustado, los criterios de información *AIC* o *BIC*, y pruebas estadísticas de modelos anidados y no anidados.

En este capítulo discutiremos diferentes algoritmos para encontrar el mejor modelo de regresión múltiple que se ajuste a unos datos determinados cuando se cuenta con un conjunto relativamente grande de posibles variables explicativas. Estos algoritmos emplean como base los conceptos que ya hemos estudiado en los capítulos anteriores. Así, entraremos directamente a una aplicación para mostrar estas aproximaciones al problema de la selección del mejor modelo.

Emplearemos unos datos simulados para una variable dependiente (y_i) y 25 posibles variables explicativas X_j , i donde $j = 1, 2, \dots, 25$. Para cada variable se simulan 150 observaciones $i = 1, 2, \dots, 150$. El modelo del que se simulan los datos incluye las variables x_1 a x_5 únicamente y los coeficientes son iguales a 1 para todos los casos (noten que en la vida real no conocemos esta información). La información se encuentra disponible en el archivo **DATOSautoSel.txt**.

Carguemos los datos en un objeto que llamaremos **data** y verifiquemos la clase del objeto leído y de cada una de las variables en el objeto. Esto se puede hacer de la siguiente manera.


```
data <- read.table("Data/DATOSautoSel.txt", header = TRUE)

str(data)

## 'data.frame': 150 obs. of 26 variables:
## $ x1 : num 5.92 4.57 5.44 4.91 5.7 ...
## $ x2 : num 6.51 3.64 5.5 5.85 6.33 ...
## $ x3 : num 6.88 4.38 6.6 6.54 5.28 ...
## $ x4 : num 5.16 5.06 4.48 5.71 5.02 ...
## $ x5 : num 6.42 3.17 4.99 4.67 4.23 ...
## $ x6 : num 5.83 4.6 5.52 5.26 5.77 ...
## $ x7 : num 5.36 4.58 6.24 5.79 4.47 ...
## $ x8 : num 6.63 4.17 5.4 5.07 4.77 ...
## $ x9 : num 5.73 4.26 5.46 5.5 6.06 ...
## $ x10: num 6.11 4.44 5.57 4.64 5.5 ...
## $ x11: num 4.82 5.58 5.04 6.33 3.37 ...
## $ x12: num 6.11 5.71 5.87 4.85 5.45 ...
## $ x13: num 6.61 5.07 5.02 6.13 6.35 ...
## $ x14: num 4.82 4.25 4.91 6.41 5.38 ...
## $ x15: num 6.3 4.55 4.83 4.22 5.75 ...
## $ x16: num 5.45 4.12 5.23 5.33 4.84 ...
## $ x17: num 6.42 4.2 4.85 6.06 5.41 ...
## $ x18: num 6.71 4.8 5.31 5.68 4.56 ...
## $ x19: num 5.56 4.16 5.03 5.29 5.54 ...
## $ x20: num 7.15 5.64 5.15 5.29 6.38 ...
## $ x21: num 6.47 3.93 4.93 5.04 5.45 ...
## $ x22: num 6.31 4.61 5.33 4.55 5.44 ...
## $ x23: num 5.41 4.1 4.58 5.4 4.65 ...
## $ x24: num 5.92 5.64 4.97 6.02 5.5 ...
## $ x25: num 5.12 5.62 5.3 5.87 5.12 ...
## $ y : num 42.8 30.4 36.3 36.6 38.3 ...

head(data, 2)

##      x1      x2      x3      x4      x5      x6      x7      x8      x9
## 1 5.925 6.512 6.883 5.157 6.420 5.833 5.356 6.626 5.730
## 2 4.565 3.638 4.378 5.056 3.165 4.598 4.576 4.171 4.259
##      x10     x11     x12     x13     x14     x15     x16     x17     x18
## 1 6.108 4.817 6.106 6.613 4.821 6.296 5.449 6.425 6.705
## 2 4.443 5.582 5.713 5.069 4.247 4.548 4.125 4.196 4.800
##      x19     x20     x21     x22     x23     x24     x25      y
## 1 5.564 7.154 6.466 6.309 5.407 5.915 5.119 42.799
## 2 4.163 5.636 3.931 4.612 4.097 5.638 5.616 30.372

class(data)

## [1] "data.frame"

sapply(data, class)

##      x1      x2      x3      x4      x5      x6
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##      x7      x8      x9     x10     x11     x12
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##      x13     x14     x15     x16     x17     x18
```

```
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##      x19      x20      x21      x22      x23      x24
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##      x25      y
## "numeric" "numeric"
```

En este caso, tenemos 25 posibles variables, entonces son $3,3554432 \times 10^7$ posibles modelos. Este capítulo está compuesto de dos partes. La primera, emplea un número reducido de variables ($(k - 1) = 10$) para mostrar aproximaciones que emplean “fuerza bruta”. En la segunda parte veremos varias técnicas para encontrar el mejor modelo cuando se tiene muchas posibles variables explicativas, emplearemos las 25 variables.

6.2. Empleando “fuerza bruta”

La primera, aproximación que estudiaremos es viable cuando se cuenta con pocas potenciales variables para explicar la variable dependiente (y). En este caso, se puede emplear la “fuerza bruta” de los computadores para encontrar el mejor modelo. Es decir, se puede emplear la capacidad de computo para calcular todos los posibles modelos y compararlos.

Supongamos que contamos con las 10 primeras variables ($(k - 1) = 10$) de nuestros datos simulados para explicar a y . No empleamos todas las variables explicativas para ahorrar tiempo en la estimación de todos los modelos. Creemos una base de datos solo con las variables que son de nuestro interés.

```
data2 <- data[, c(1:10, 26)]
head(data2, 2)

##      x1      x2      x3      x4      x5      x6      x7      x8      x9
## 1 5.925 6.512 6.883 5.157 6.420 5.833 5.356 6.626 5.730
## 2 4.565 3.638 4.378 5.056 3.165 4.598 4.576 4.171 4.259
##      x10      y
## 1 6.108 42.799
## 2 4.443 30.372
```

Empecemos por estimar un modelo lineal con todas las variables potenciales. Como sabemos esto se puede hacer con la función **lm()** del paquete básico de R. Recuerden que esta función incluirá siempre un intercepto a menos que se le indique lo contrario, empleando en la especificación del modelo un -1).

Estimemos un modelo con todas las posibles variables contenidas en el **data.frame** **data2** y guardemos los resultados de la estimación en un objeto llamado **model**.

```
model <- lm(y ~ ., data = data2)
summary(model)

##
```

```
## Call:
## lm(formula = y ~ ., data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.4754 -1.5134  0.0137  1.4273  8.0533
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.21741    1.40780   7.968 5.24e-13 ***
## x1           0.65111    0.26894   2.421 0.016768 *
## x2           1.72908    0.28714   6.022 1.46e-08 ***
## x3           0.94195    0.27727   3.397 0.000888 ***
## x4           0.92954    0.26551   3.501 0.000623 ***
## x5           0.94561    0.25956   3.643 0.000379 ***
## x6           0.07626    0.26472   0.288 0.773712
## x7           0.01759    0.28152   0.062 0.950278
## x8          -0.24513    0.27869  -0.880 0.380598
## x9          -0.54115    0.27812  -1.946 0.053705 .
## x10          0.22042    0.28886   0.763 0.446720
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.496 on 139 degrees of freedom
## Multiple R-squared:  0.7168, Adjusted R-squared:  0.6964
## F-statistic: 35.18 on 10 and 139 DF,  p-value: < 2.2e-16
```

Noten que de acuerdo a los resultados, las variables de x_1 a x_5 son estadísticamente significativas (nivel de confianza del 95%)¹. Las otras variables no son significativas.

Ahora podemos investigar todos los posible 1024 modelos. Esto lo podemos hacer con la función **all.possible.models** del paquete **olsrr**.

```
#install.packages("olsrr")
library(olsrr)
models <- ols_step_all_possible(model)
str(models)

## Classes 'ols_step_all_possible', 'tibble' and 'data.frame': 1023 obs. of
## $ mindex      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ n           : int  1 1 1 1 1 1 1 1 1 1 ...
## $ predictors: chr   "x2" "x5" "x4" "x3" ...
## $ rsquare     : num  0.55 0.403 0.382 0.367 0.305 ...
## $ adjr        : num  0.546 0.399 0.378 0.363 0.3 ...
## $ predrsq     : num  0.537 0.384 0.364 0.351 0.283 ...
## $ cp          : num  75.1 147.2 157.3 164.5 195 ...
## $ aic         : num  764 807 812 815 829 ...
## $ sbic        : num  337 378 383 387 401 ...
## $ sbc         : num  773 816 821 824 838 ...
```

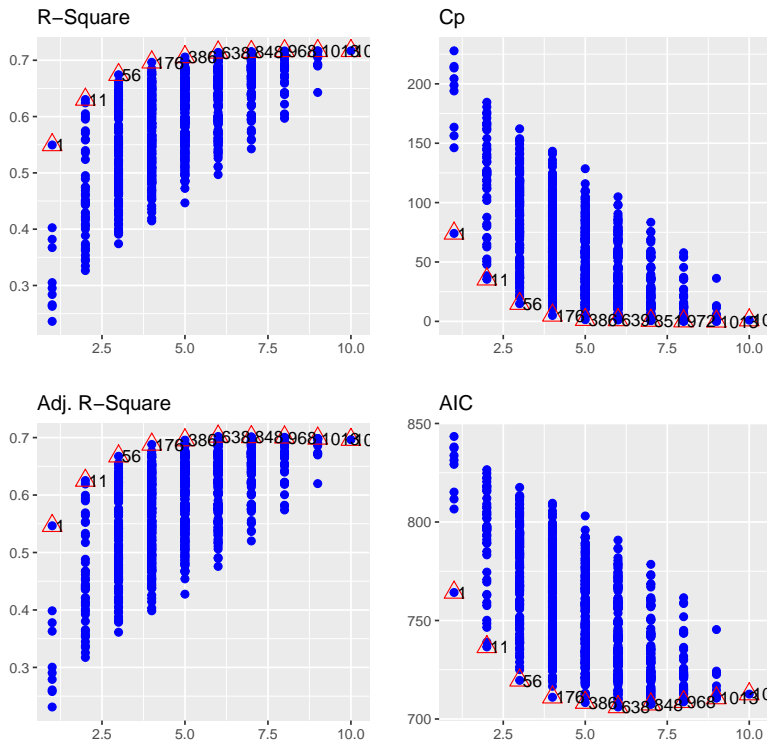
¹ La variable x_9 es significativa con un nivel de confianza del 90%. Usaremos un nivel de confianza del 95%.

```
## $ msep      : num  9.43 12.51 12.94 13.24 14.55 ...  
## $ fpe       : num  9.43 12.5 12.93 13.24 14.54 ...  
## $ apc       : num  0.463 0.614 0.635 0.65 0.714 ...  
## $ hsp       : num  0.0633 0.0839 0.0868 0.0889 0.0976 ...
```

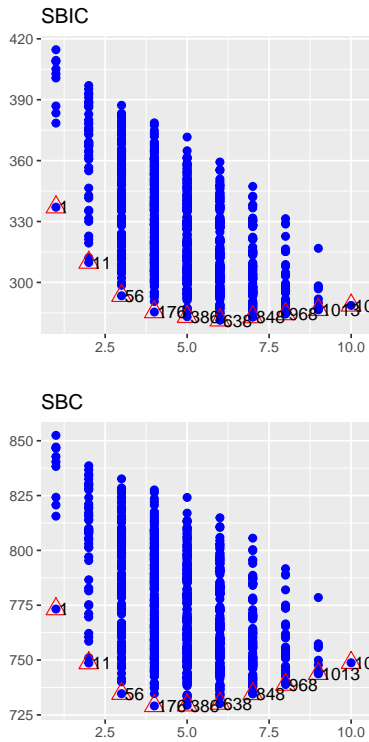
En el objeto **models** se encuentran varios estadísticos que permiten resumir las características estadísticas de los 1023 modelos estimados. Con un gráfico podemos resumir esta información.

```
library(ggplot2)  
plot(models)
```

page 1 of 2



page 2 of 2



En el eje horizontal podemos ver el número de variables empleadas en cada modelo, mientras que en el eje vertical encontramos el valor del estadístico. El triángulo muestra el modelo que maximiza o minimiza el valor del estadístico para cada uno de los posibles número de variables. Concentrémonos en el R^2 ajustado (se desea maximizar) y los criterios de información (se desean minimizar): AIC (Akaike information criteria), SBC (Schwarz bayes information criteria)². Recuerden que estas tres métricas penalizan la inclusión de más variables en el modelo.

Empleando el criterio del R^2 ajustado podemos llegar a la conclusión que el mejor modelo es uno que emplea 6 variables (x1 x2 x3 x4 x5 x9). De hecho, ese modelo corresponde al modelo número 638 de los 1023 estimados. Esta información se puede obtener de la siguiente manera.

```
models$mindex[which.max(models$predrsq)]
## [1] 638
models$n[which.max(models$predrsq)]
## [1] 6
models$predictors[which.max(models$predrsq)]
## [1] "x1 x2 x3 x4 x5 x9"
```

Empleando el criterio de información de AIC encontramos que el mejor modelo es el mismo.

```
models$mindex[which.min(models$aic)]
## [1] 638
models$n[which.min(models$aic)]
## [1] 6
models$predictors[which.min(models$aic)]
## [1] "x1 x2 x3 x4 x5 x9"
```

Para el caso del BIC el modelo seleccionado es diferente. Este modelo emplea 4 variables (x2 x3 x4 x5) y ese modelo corresponde al modelo número 176 de los 1023 estimados.

```
models$mindex[which.min(models$sbic)]
## [1] 176
models$n[which.min(models$sbic)]
## [1] 4
models$predictors[which.min(models$sbic)]
## [1] "x2 x3 x4 x5"
```

² Recuerden que este es otro nombre para el BIC.

Finalmente, podríamos chequear las características de los dos modelos. recuerden que con la función `summary()` se puede visualizar los resultados de la regresión.

```
modelo1 <- lm( y ~ x1 + x2 + x3 + x4 + x5 + x9, data = data2)
modelo2 <- lm( y ~ x2 + x3 + x4 + x5, data = data2)
```

Los resultados se reportan en el siguiente cuadro.

Cuadro 6.1 Modelos estimados por MCO

	y	y
intercepto	11.272*** [8.331]	11.924*** [9.088]
x1	0.675** [2.567]	
x2	1.692*** [6.246]	1.701*** [6.358]
x3	0.956*** [3.611]	0.857*** [3.255]
x4	0.978*** [3.92]	1.034*** [4.226]
x5	0.952*** [3.837]	0.997*** [4.041]
x9	-0.537** [-2.022]	
R ²	0.714	0.696
adj.R ²	0.702	0.688
N	150	150

t-values in brackets
* (p ≤ 0.1), ** (p ≤ 0.05), *** (p ≤ 0.01)

Noten que en este caso los modelos están anidados y por tanto se pueden comparar fácilmente empleando una prueba *F* que compare un modelo restringido con uno sin restringir.

```
library(AER)
anova(modelo2, modelo1)

## Analysis of Variance Table
##
## Model 1: y ~ x2 + x3 + x4 + x5
## Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x9
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      145 927.97
## 2      143 874.26  2    53.712  4.3928 0.01408 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No se puede rechazar la nula de que el modelo con restricciones (el pequeño) es mejor que el sin restringir con un 99% de confianza. Así, podemos concluir que el mejor modelo es el que incluye las variables (x2 x3 x4 x5). En este caso sabemos

que el modelo real que generó los datos incluye las variables x_1 a X_5 . Así, nuestra aproximación no encontró el modelo real, pero uno relativamente cercano. Con un 95 % de confianza se puede rechazar la nula en favor del modelo sin restringir. En este caso el modelo incluiría todas las variables de x_1 a X_5 pero también incluiría x_9 . tampoco el modelo exacto, pero lo suficientemente cercano.

6.3. Empleando estrategias inteligentes de detección de un mejor modelo

En algunas ocasiones es imposible encontrar el mejor modelo estimando todas las combinaciones (como el caso en el que se tienen 25 variables pues existen $3,3554432 \times 10^7$ posibles modelos). A continuación discutimos varios algoritmos que facilitan la tarea.

6.3.1. Regresión paso a paso (Stepwise)

La idea de la construcción de modelos por pasos es arribar a un modelo de regresión a partir de un conjunto de posibles variables explicativas basados en un criterio que permita adicionar variables (stepwise forward regression) o quitar variables (stepwise backwards regression).

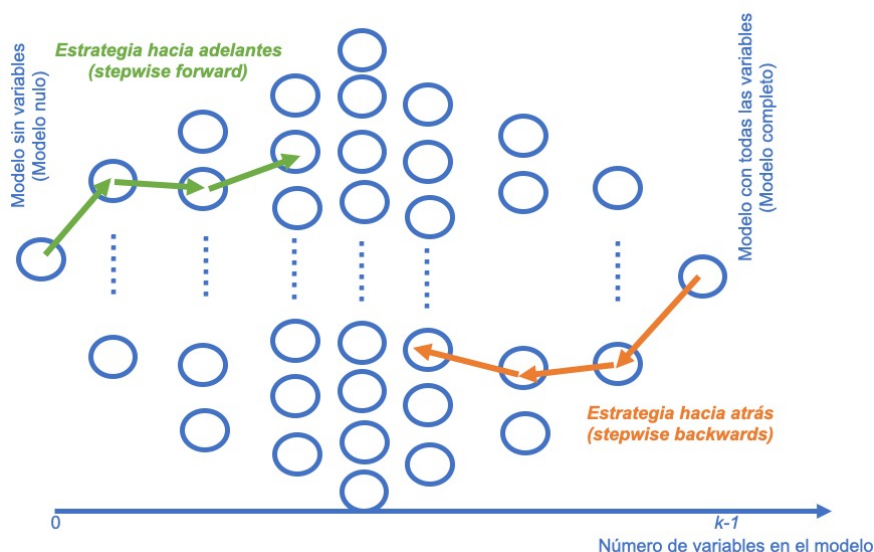
Por ejemplo, supongamos que empleamos un criterio como el valor p de la prueba de significancia individual de cada variable en el modelo. En el primer caso (stepwise Forward regression) se parte de un modelo sin variables. Se empieza adicionando al modelo la variable que tenga el mayor p valor. De forma gradual se incluye la siguiente variable que tenga el valor p más grande, se sigue de esta manera hasta que ya no quede ninguna variable para ingresar que sea significativa³.

En el segundo caso (stepwise backwards regression), se parte del modelo con todas las variables y se empieza a eliminar variables que tenga el valor p mas bajo. El proceso se repite hasta que no se puedan eliminar variables⁴. Es fácil imaginarse cómo funcionarán ambos métodos si se emplean criterios como el R^2 ajustado o criterios de información. La figura 6.1 muestra de manera esquemática estas dos aproximaciones.

A continuación veremos un ejemplo empleando la base de datos original con 25 variables explicativas.

³ Noten que esta aproximación tiene un problema práctico difícil de resolver. Se emplean múltiples pruebas individuales que acumulan el error tipo I. No existe almuerzo gratis, este es el costo de emplear esta aproximación

⁴ Noten que esta aproximación también tiene el problema mencionado para la aproximación forward. No existe almuerzo gratis, este es el costo de emplear esta aproximación

Figura 6.1 Representación de las estrategias stepwise forward y stepwise backwards

6.3.2. Stepwise Forward regression

La función **regsubsets()** del paquete **leap** permite encontrar los mejores subconjuntos de predictores utilizando R^2 ajustado partiendo de un modelo con todos los regresores (lo llamaremos el modelo máximo). Esta función no está diseñada para funcionar el valor p . La función **regsubsets()** calcula los mejores modelos para todos los posibles número de variables explicativas sin calcularlos de manera exhaustiva⁵.

Para usar esta función, tenemos que especificar

- la matriz que contiene todas las posibles variables explicativas ($x=$),
- el vector de la variable dependiente ($y=$)
- el método que se desea emplear, en este caso, (*method* = "forward")

Adicionalmente, se puede especificar el número máximo de modelos a ser examinados (*nvmax*). También en algunas ocasiones será útil especificar unas variables que independientemente del resultado del algoritmo deberían estar siempre presentes como variables explicativas en todos los modelos considerados. Esto último se

⁵ Dado que los criterios de información (*AIC* o *BIC*) solo difieren al comparar modelos con número diferentes de variables explicativas, el resultado final de los cálculos que realice esta función no depende del criterio de información que se emplee (**leaps**). Así, esta función se puede emplear también para escoger el mejor modelo empleando los criterios de información. En ese caso el código que se presenta mas adelante deberá ser modificado para emplear dichos criterios. Pero no es necesario modificar el código correspondiente a la función **regsubsets()** que se presenta a continuación.

puede hacer con el argumento *force.in* que debe tener el número de las columnas de la matriz X , también conocida como la matriz de diseño.

Para nuestro caso tenemos

```
library(leaps)

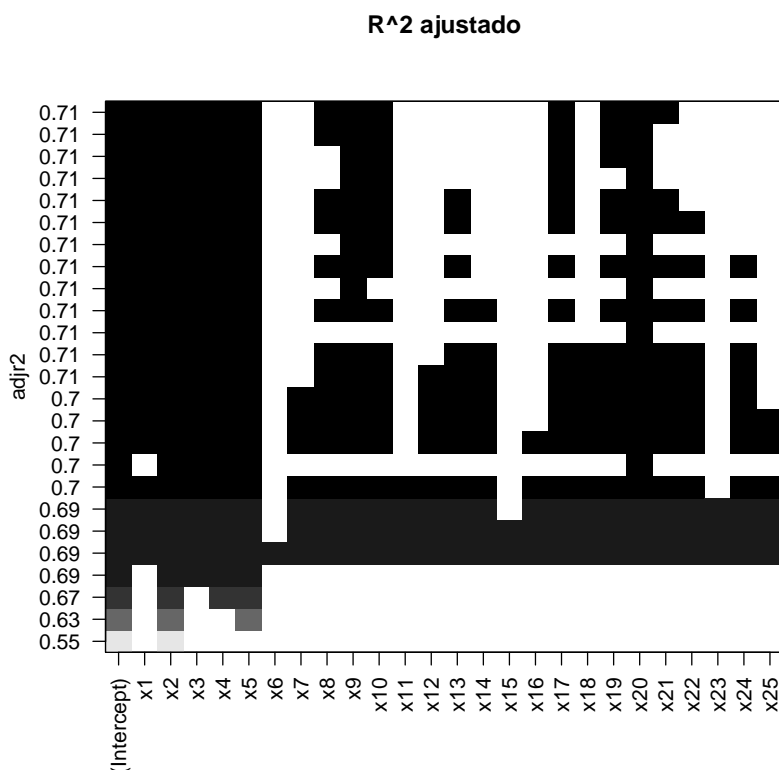
fwd.model <- regsubsets(x = data[,1:25], y = data[,26],
                       nvmax=1000, method = "forward")
```

En el objeto **fwd.model** se encuentran diferentes resultados. Veamos esto en detalle.

```
attributes(fwd.model)

## $names
## [1] "np"          "nrbar"       "d"           "rbar"
## [5] "thetab"      "first"       "last"        "vorder"
## [9] "tol"         "rss"         "bound"       "nvmax"
## [13] "ress"        "ir"          "nbest"       "lopt"
## [17] "il"          "ier"         "xnames"      "method"
## [21] "force.in"    "force.out"   "sserr"       "intercept"
## [25] "lindep"      "nullrss"     "nn"
##
## $class
## [1] "regsubsets"

plot(fwd.model, scale = "adjr2", main = "R^2 ajustado")
```



Este gráfico presenta el R^2 ajustado en el eje vertical y todas las potenciales variables evaluadas. El gráfico solo presenta los mejores modelos, en términos R^2 ajustado de los mejores modelos evaluados. Un cuadrado negro implica que la correspondiente variable es incluida en el modelo que produce ese correspondiente R^2 ajustado. Así, entre más “arriba” en el gráfico se muestre un modelo, mejor será este de acuerdo a esta métrica. El mejor modelo es el último que se presenta (fila superior)⁶. En este caso el modelo tiene las variables x1 a x5, x8 a x10, x17 y de x19 a x21. Estimemos ese modelo y guardémoslo en un objeto que llamaremos *modelo3*.

```
modelo3 <- lm( y ~ x1 + x2 + x3 + x4 + x5 +
                x8 + x9 + x10 + x17 + x19 +
                x20 + x21, data = data)
summary(modelo3)
```

⁶ Si se desea seleccionar el modelado empleando criterios de información, entonces la última línea de código debería ser modificada a `plot(fwd.model, scale = "bic")` o `plot(fwd.model, scale = "aic")`, según sea el caso

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x8 + x9 + x10 + x17 +
##      x19 + x20 + x21, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.3123 -1.5515 -0.0106  1.5859  6.5427
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.0686     1.3692   8.084 2.94e-13 ***
## x1           0.5791     0.2706   2.140 0.034097 *
## x2           1.6885     0.2803   6.024 1.48e-08 ***
## x3           0.8892     0.2746   3.238 0.001511 **
## x4           0.9751     0.2605   3.743 0.000267 ***
## x5           0.8822     0.2719   3.244 0.001480 **
## x8          -0.3131     0.2705  -1.158 0.249075
## x9          -0.5104     0.2787  -1.832 0.069184 .
## x10          0.4187     0.2833   1.478 0.141721
## x17          0.3070     0.2586   1.187 0.237254
## x19          0.3087     0.2894   1.067 0.287945
## x20         -0.7848     0.2850  -2.754 0.006696 **
## x21          0.3129     0.2926   1.069 0.286774
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.419 on 137 degrees of freedom
## Multiple R-squared:  0.7378, Adjusted R-squared:  0.7148
## F-statistic: 32.12 on 12 and 137 DF,  p-value: < 2.2e-16
```

Este modelo tiene variables no significativas individualmente que pueden ser removidas automáticamente como se discute mas adelante en este capítulo.

También podemos realizar una versión de regresión por pasos hacia adelante utilizando la función `ols_step_forward()` del paquete `olsrr`. Esta función nos permite usar el criterio del p valor de las pruebas de significancia individuales y criterios de información. Esta función solo necesita como argumento un objeto `lm` que contenga el modelo máximo (el que contienen todas las posibles variables. Por ejemplo

```
library(olsrr)
max.model <- lm( y ~ ., data = data)
fwd.model.2 <- ols_step_forward_p(max.model)
```

Los resultados los podemos explorar de muchas maneras, pero la mas sencilla es llamando al objeto. Esto nos mostrará cuáles son las variables que se incluyen en el mejor modelo

```
fwd.model.2

##
##                               Selection Summary
```

##	##	Variable		Adj.			
##	Step	Entered	R-Square	R-Square	C (p)	AIC	RMSE
##	1	x2	0.5495	0.5465	69.8613	764.2277	3.0502
##	2	x5	0.6303	0.6253	33.1629	736.5927	2.7727
##	3	x4	0.6742	0.6675	14.1062	719.6128	2.6116
##	4	x3	0.6964	0.6880	5.4785	711.0365	2.5298
##	5	x20	0.7069	0.6967	2.4563	707.7667	2.4943
##	6	x1	0.7200	0.7082	-1.8043	702.9271	2.4466
##	7	x9	0.7241	0.7105	-1.7852	702.6965	2.4370
##	8	x10	0.7281	0.7127	-1.7173	702.4882	2.4277
##	9	x17	0.7315	0.7142	-1.3183	702.6336	2.4214

El modelo seleccionado incluye las siguientes variables: x1 a x5, x9, x10, x17 y x20. (Asegúrese que entiende por qué se escoge dicho modelo según el gráfico). Estimemos este modelo y gurdemolo en el objeto *modelo4*.

```
modelo4 <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x9 + x10 + x17 +x20, data = data)
```

Los resultados se reportan en el cuadro ???. Este modelo también tiene variables no significativas individualmente.

Cuadro 6.2 Modelos estimados por MCO

	y
intercepto	11.174*** [8.284]
x1	0.663** [2.532]
x2	1.702*** [6.394]
x3	0.958*** [3.632]
x4	0.994*** [3.827]
x5	0.93*** [3.49]
x9	-0.501* [-1.825]
x10	0.392 [1.395]
x17	0.333 [1.32]
x20	-0.735** [-2.601]
R ²	0.731
adj.R ²	0.714
N	150

t-values in brackets
* (p ≤ 0.1), ** (p ≤ 0.05), *** (p ≤ 0.01)

El mismo paquete tiene una función que permite realizar el algoritmo empleando el criterio de información *AIC* (y otros como el mismo R^2 ajustado). Para el *AIC* la función es `ols_step_forward_aic`. El único argumento necesario es un objeto de clase **lm** que contenga el modelo máximo. En este caso tenemos.

```
fwd.model.3 <- ols_step_forward_aic(max.model)
fwd.model.3
```

Selection Summary						
Variable	AIC	Sum Sq	RSS	R-Sq	Adj.	R-Sq
x2	764.228	1679.790	1376.927	0.54954	0.54650	
x5	736.593	1926.638	1130.079	0.63030	0.62527	
x4	719.613	2060.953	995.764	0.67424	0.66754	
x3	711.036	2128.745	927.972	0.69642	0.68804	
x20	707.767	2160.781	895.936	0.70690	0.69672	
x1	702.927	2200.715	856.001	0.71996	0.70821	
x9	702.696	2213.351	843.366	0.72409	0.71049	
x10	702.488	2225.675	831.042	0.72813	0.71270	

El modelo seleccionado incluye las siguientes variables: x_1 a x_5 , x_9 , x_{10} y x_{20} . Noten que al llamar al objeto `fwd.model.3` podemos observar cuál variable fue adicionada en cada uno de los 8 pasos. En esta oportunidad, la primera variable adicionada al modelo fue la x_2 y la última x_{10} . Los resultados los podemos ver de manera gráfica.

En el *slot* denominado *predictors* podemos encontrar las variables que se seleccionaron en el mejor modelo. Así podemos estimar el mejor modelo según este algoritmo y el criterio de información *AIC* de la siguiente manera (esto evita tener que escribir manualmente la fórmula como lo habíamos hecho en los casos anteriores)

```
vars.modelo5 <- fwd.model.3$predictors
formula.modelo5 <- as.formula(
  paste("y ~ ", paste (vars.modelo5, collapse=" + "),
    sep="") )
formula.modelo5

## y ~ x2 + x5 + x4 + x3 + x20 + x1 + x9 + x10

modelo5 <- lm( formula.modelo5 , data = data)
```

En la siguiente tabla se reporta el resultado de este modelo.

Este modelo tiene dos variables no significativas individualmente. El lector ya conoce el procedimiento para eliminar estas variables que no son significativas para obtener un mejor modelo.

Cuadro 6.3 Modelo seleccionado por el algoritmo step forward empleando el criterio de informaci'on AIC. Estimado por MCO

	y
intercepto	11.348*** [8.431]
x1	0.686*** [2.619]
x2	1.722*** [6.46]
x3	0.95*** [3.593]
x4	1.051*** [4.096]
x5	1.051*** [4.193]
x9	-0.425 [-1.578]
x10	0.407 [1.446]
x20	-0.738** [-2.606]
R ²	0.728
adj.R ²	0.713
N	150

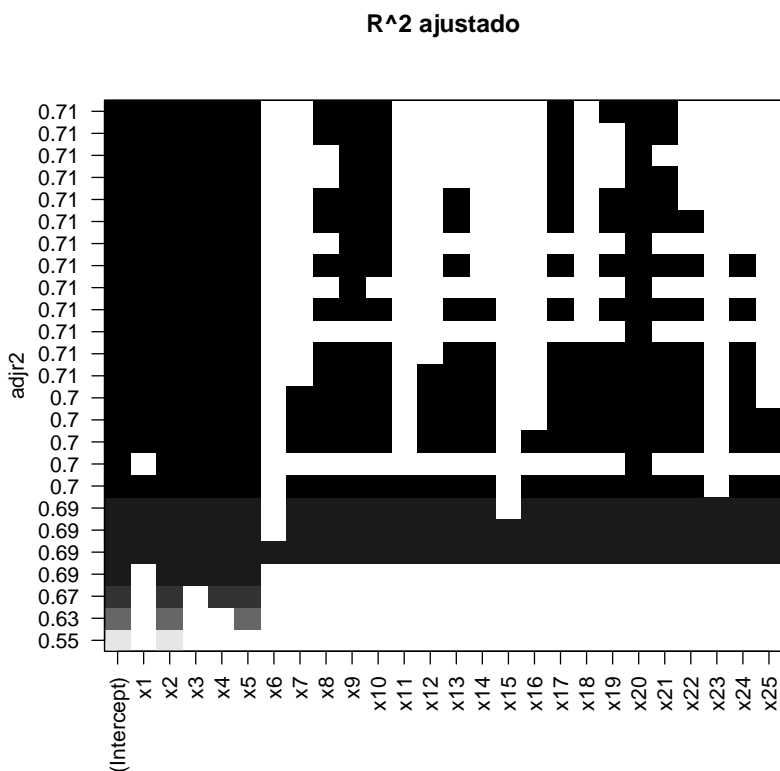
t-values in brackets
* (p ≤ 0.1), ** (p ≤ 0.05), *** (p ≤ 0.01)

6.3.3. Stepwise backward regression

De manera similar podemos emplear tanto el paquete *leaps* como el paquete *olsrr* para encontrar un modelo partiendo del modelo que incluye todas las variables y quitando una variable en cada paso. En este caso tendremos el siguiente resultado sí empleamos el criterio del R^2 ajustado.

```
back.model <- regsubsets(x = data[,1:25], y = data[,26],
                        nvmax=1000, method = "backward")

plot(back.model, scale = "adjr2", main = "R^2 ajustado")
```



El modelo tiene las variables x1 a x5, x8 a x10, x17 y de x19 a x21. Es decir, llega a la misma conclusión que el método forward.

De manera similar al algoritmo forward, podemos emplear la función `ols_step_backward_p()` del paquete `olsrr` para seleccionar el mejor modelo empleando el valor p de la prueba individual para eliminar variables. El argumento que necesita esta función es el objeto que contenga la estimación del modelo máximo.

```
back.model.2 <- ols_step_backward_p(max.model)
```

En este caso el modelo seleccionado incluye las siguientes variables: x1 a x5, x8, x9, x10, x17, x19 a x21. Calculemos dicho modelo y guardémoslo en el objeto `modelo6`.

```
modelo6 <- lm( y ~ x1 + x2 + x3 + x4 + x5 + x8 + x9
               + x10 + x17 + x19 + x20 + x21, data = data)
```

Noten que este modelo también tiene variables no significativas individualmente. Ustedes deberían emplear las técnicas que ya conocemos para encontrar un mejor modelo sin variables no significativas. Los resultados se reportan en el cuadro ??.

Nuevamente, como lo hicimos con el algoritmo forward, podemos emplear el paquete *olsrr* y la función *ols_step_backward_aic* para encontrar el mejor modelo de acuerdo con este algoritmo y el criterio de información *AIC*.

```
back.model.3 <- ols_step_backward_aic(max.model)

## Backward Elimination Method
## -----
##
## Candidate Terms:
##
## 1 . x1
## 2 . x2
## 3 . x3
## 4 . x4
## 5 . x5
## 6 . x6
## 7 . x7
## 8 . x8
## 9 . x9
## 10 . x10
## 11 . x11
## 12 . x12
## 13 . x13
## 14 . x14
## 15 . x15
## 16 . x16
## 17 . x17
## 18 . x18
## 19 . x19
## 20 . x20
## 21 . x21
## 22 . x22
## 23 . x23
## 24 . x24
## 25 . x25
##
##
## Variables Removed:
##
## - x6
## - x15
## - x23
## - x11
## - x16
## - x25
## - x7
## - x12
## - x18
## - x14
## - x24
## - x22
## - x13
## - x19
```

```
## - x8
## - x21
## - x17
##
## No more variables to be removed.
```

El modelo seleccionado incluye las siguientes variables: x6 a x8, x11 a x19 y x21 a x25. Estimemos el correspondiente modelo y guardémoslo en el objeto *modelo7*.

```
vars.modelo7 <- back.model.3$predictors
formula.modelo7 <- as.formula(paste("y ~ ", paste (vars.modelo7, collapse="
formula.modelo7
modelo7 <- lm( formula.modelo7 , data = data)
```

Este modelo también tiene variables no significativas individualmente, pero muchas mas que los modelos anteriores. Esto no es una característica de este algoritmo, solo es coincidencia. Los resultados se reportan en el cuadro ??.

6.3.4. *Combinando forward y backward (step regression)*

También podemos crear un modelo de regresión a partir de un conjunto de posibles variables explicativas ingresando y eliminando predictores basados en si se aumenta o no el R^2 ajustado, de forma escalonada hasta que ya no quede ninguna variable para ingresar o eliminar. El modelo de partida debe incluir todas las variables predictoras candidatas. Empleando el paquete *leaps* y la función que ya conocemos *regsubsets*.

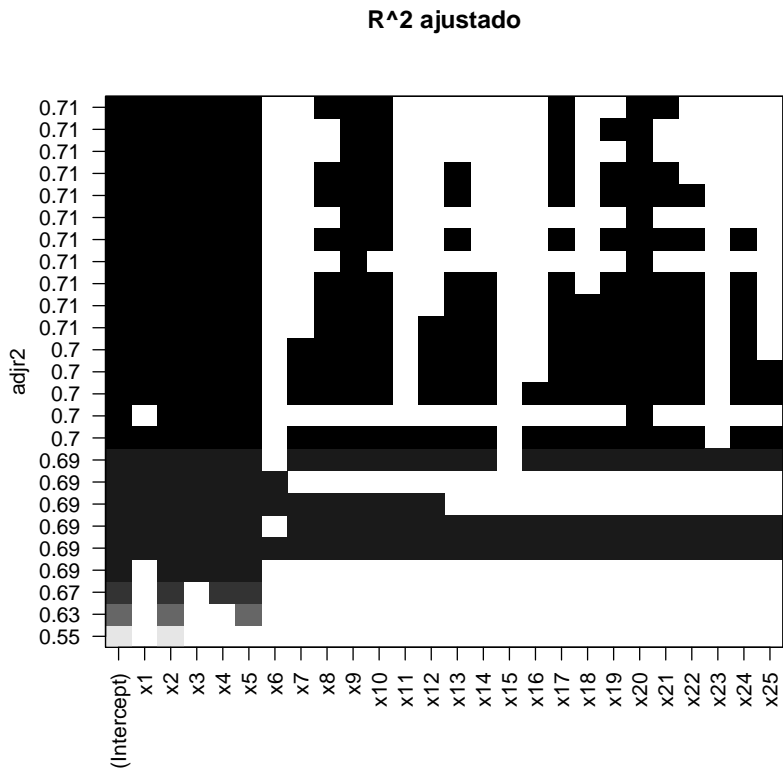
```
both.model <- regsubsets(x = data[,1:25], y = data[,26], nvmax=1000, method
plot(both.model, scale = "adjr2", main = "R^2 ajustado")
```

Cuadro 6.4 Modelo seleccionado por el algoritmo step backward empleando el valor p y el AIC. Estimado por MCO

	y	y
intercepto	11.069*** [8.084]	11.424*** [6.157]
x1	0.579** [2.14]	
x2	1.689*** [6.024]	
x3	0.889*** [3.238]	
x4	0.975*** [3.743]	
x5	0.882*** [3.244]	
x8	-0.313 [-1.158]	0.076 [0.202]
x9	-0.51* [-1.832]	
x10	0.419 [1.478]	
x17	0.307 [1.187]	0.678** [2.013]
x19	0.309 [1.067]	0.288 [0.743]
x20	-0.785*** [-2.754]	
x21	0.313 [1.069]	0.918** [2.484]
x6		0.157 [0.432]
x7		-0.114 [-0.3]
x11		0.591* [1.789]
x12		0.259 [0.661]
x13		0.381 [1.065]
x14		0.099 [0.278]
x15		-0.264 [-0.643]
x16		0.386 [1.056]
x18		0.249 [0.632]
x22		0.59 [1.601]
x23		0.029 [0.073]
x24		0.115 [0.304]
x25		0.255 [0.714]
R^2	0.738	0.579
$adj.R^2$	0.715	0.525
N	150	150

t-values in brackets

* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)



```
## y ~ x1 + x2 + x3 + x4 + x5 + x8 + x9 + x10 + x17 + x20 + x21
```

En este caso el modelo tiene las siguientes variables: x1, x2, x3, x4, x5, x8, x9, x10, x17, x20, x21. El lector puede estimar el correspondiente modelo (llámelo modelo8) los resultados de ese modelo se encuentran en la siguiente tabla.

Nuevamente, el modelo incluye variables no significativas.

Otra forma de emplear este método es usando el valor p como criterio para quitar o incluir variables. Esto se puede hacer empleando la función `ols_step_both_p()` del paquete `olsrr` el único argumento necesario para emplear la función es el modelo máximo.

```
both.model.2 <- ols_step_both_p(max.model)
```

En este caso el modelo seleccionado incluye las variables x1 a x5 y x20. El lector puede constatar que el correspondiente modelo es el reportado en la siguiente tabla.

```
## y ~ x2 + x5 + x4 + x3 + x20 + x1
```

Cuadro 6.5 Modelo seleccionado por el algoritmo combinado empleando el R^2 ajustado.

Estimado por MCO	
	y
intercepto	11.108*** [8.112]
x1	0.652** [2.489]
x2	1.694*** [6.041]
x3	0.903*** [3.29]
x4	0.976*** [3.745]
x5	0.93*** [3.467]
x8	-0.29 [-1.075]
x9	-0.475* [-1.716]
x10	0.426 [1.503]
x17	0.363 [1.434]
x20	-0.763*** [-2.682]
x21	0.329 [1.124]
R^2	0.736
$adj.R^2$	0.714
N	150
t-values in brackets	
* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)	

En este caso el modelo encontrado tiene todas las variables significativas (Ver cuadro ??).

Finalmente, empleando el criterio de AIC tendremos que el mejor modelo es:

```
both.model.3 <- ols_step_both_aic(max.model)
```

El modelo seleccionado con este algoritmo y criterio incluye las variables x1 a x5, x9, x10 y x20. El lector puede constatar que el correspondiente modelo (llámelo modelo10) es el reportado en la siguiente tabla.

```
## y ~ x2 + x5 + x4 + x3 + x20 + x1 + x9 + x10
```

Este modelo tiene dos variables no significativas.

Cuadro 6.6 Modelo seleccionado por el algoritmo combinado empleando el valor p. y el AIC. Estimado por MCO

	y	y
intercepto	11.464*** [8.534]	11.348*** [8.431]
x1	0.663** [2.583]	0.686*** [2.619]
x2	1.664*** [6.281]	1.722*** [6.46]
x3	0.932*** [3.618]	0.95*** [3.593]
x4	1.087*** [4.289]	1.051*** [4.096]
x5	1.056*** [4.202]	1.051*** [4.193]
x20	-0.715*** [-2.688]	-0.738** [-2.606]
x9		-0.425 [-1.578]
x10		0.407 [1.446]
R^2	0.72	0.728
$adj.R^2$	0.708	0.713
N	150	150

t-values in brackets

* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)

6.4. Pongamos todo junto

En la práctica queremos emplear un único modelo, para eso debemos comparar los modelos que hemos encontrado, ya sea que estos estén anidados o no. Pero antes es mejor comparar modelos que tengan solo variables explicativas significativas.

Recordemos que en este ejercicio de selección automática del mejor modelo hemos construido ya varias opciones como se resume en el cuadro ??.

Cuadro 6.7 Modelos construidos hasta ahora con diferentes algoritmos y criterios

Nombre del objeto	Algoritmo	Criterio
modelo3	Forward	R^2 ajustado
modelo4	Forward	valor p
modelo5	Forward	AIC
modelo6	Backward	R^2 ajustado
modelo7	Backward	AIC
modelo8	Both	R^2 ajustado
modelo9	Both	valor p
modelo10	Both	AIC

En la siguiente subsección veremos un método para limpiar las variables no significativas y en la segunda subsection compararemos los modelos.

6.4.1. *Eliminando automáticamente variables no significativas*

Como se discutió anteriormente, es posible que uno de los algoritmos nos arroje un "mejor" conjunto de variables no significativas. Es decir, los algoritmos y criterios no garantizan que el modelo tenga todas las variables estadísticamente significativas. Para eliminar de manera iterativa aquellas variables que no sean individualmente significativas, podemos emplear la siguiente función. El lector deberá seguir con detalle cada línea de la función para entender los *trucos* que se emplean.

```
remueve.no.sinifica <- function(modelo, p){
  # extrae el dataframe
  data <- modelo$model

  # extraer el nombre de todas las variables X
  all_vars <- all.vars(formula(modelo))[-1]
  # extraer el nombre de la variables y
  dep_var <- all.vars(formula(modelo))[1]
  # Extraer las variables no significativas
  # resumen del modelo
  summ <- summary(modelo)
  # extrae los valores p
  pvals <- summ[[4]][, 4]
  # creando objeto para guardar las variables no significativas
  not_signif <- character()
  not_signif <- names(which(pvals > p))

  # Si hay alguna variable no-significativa
  while(length(not_signif) > 0){
    all_vars <- all_vars[!all_vars %in% not_signif[1]]
    # nueva formula
    myForm <- as.formula(paste(paste(dep_var, "~ "),
                               paste(all_vars, collapse=" + "), sep=""))
    # re-escribe la formula
    modelo <- lm(myForm, data= data)

    # Extrae variables no significativas.
    summ <- summary(modelo)
    pvals <- summ[[4]][, 4]
    not_signif <- character()
    not_signif <- names(which(pvals > p))
    not_signif <- not_signif[!not_signif %in% "(Intercept)"]
  }
  modelo.limpio <- modelo
  return(modelo.limpio)
}
```

Para ver un ejemplo, regresemos al modelo construido por medio del algoritmo forward y el criterio del R^2 ajustado (ese modelo lo llamamos *modelo3*). La función que acabamos de construir (*remueve.no.sinifica*) tiene dos argumento, el primero es un objeto *lm* y el segundo el nivel de significancia al cuál se quiere que todas las variables sean significativas. Corramos esta función para el *modelo3* con un nivel de significancia del 95 % y guardemos los resultados en un objeto que denominaremos *modelo3.a*.

```
modelo3.a <- remueve.no.sinifica(modelo3, 0.05)
```

Antes de continuar, comparemos estos dos modelos en la siguiente tabla.

Cuadro 6.8 Comparación de modelos antes y después de la función *remueve.no.sinifica()*.

Estimado por MCO		
	y	y
intercepto	11.069*** [8.084]	12.444*** [9.472]
x1	0.579** [2.14]	
x2	1.689*** [6.024]	1.78*** [6.691]
x3	0.889*** [3.238]	0.946*** [3.601]
x4	0.975*** [3.743]	1.212*** [4.778]
x5	0.882*** [3.244]	1.156*** [4.567]
x8	-0.313 [-1.158]	
x9	-0.51* [-1.832]	
x10	0.419 [1.478]	
x17	0.307 [1.187]	
x19	0.309 [1.067]	
x20	-0.785*** [-2.754]	-0.608** [-2.269]
x21	0.313 [1.069]	
R^2	0.738	0.707
$adj.R^2$	0.715	0.697
N	150	150

t-values in brackets

* ($p \leq 0.1$), ** ($p \leq 0.05$), *** ($p \leq 0.01$)

Ahora todas las variables son significativas. Realicemos el mismo procedimiento para todos los modelos. El lector podrá constatar que todos los correspondientes modelos serán los que se reportan en la siguiente tabla.


```
## y ~ x1 + x2 + x3 + x4 + x5 + x20
## <environment: 0x7ffae497fe88>
## y ~ x1 + x2 + x3 + x4 + x5 + x20
## <environment: 0x7ffafce84e70>
## y ~ x2 + x3 + x4 + x5 + x20
## <environment: 0x7ffafd813380>
## y ~ x22 + x25 + x21
## <environment: 0x7ffae41af460>
## y ~ x1 + x2 + x3 + x4 + x5 + x20
## <environment: 0x7ffadfc5b870>
## y ~ x1 + x2 + x3 + x4 + x5 + x20
## y ~ x1 + x2 + x3 + x4 + x5 + x20
## <environment: 0x7ffafc36bc38>
```

Cuadro 6.9 Modelos 3 al 10 tras emplear la función `remueve.no.sinifica()`. Estimado por MCO

	y	y	y	y	y	y	y	y
intercepto	12.444*** [9.472]	11.464*** [8.534]	11.464*** [8.534]	12.444*** [9.472]	15.264*** [9.372]	11.464*** [8.534]	12.444*** [9.472]	11.464*** [8.534]
x2	1.78*** [6.691]	1.664*** [6.281]	1.664*** [6.281]	1.78*** [6.691]		1.664*** [6.281]	1.78*** [6.691]	1.664*** [6.281]
x3	0.946*** [3.601]	0.932*** [3.618]	0.932*** [3.618]	0.946*** [3.601]		0.932*** [3.618]	0.946*** [3.601]	0.932*** [3.618]
x4	1.212*** [4.778]	1.087*** [4.289]	1.087*** [4.289]	1.212*** [4.778]		1.087*** [4.289]	1.212*** [4.778]	1.087*** [4.289]
x5	1.156*** [4.567]	1.056*** [4.202]	1.056*** [4.202]	1.156*** [4.567]		1.056*** [4.202]	1.156*** [4.567]	1.056*** [4.202]
x20	-0.608** [-2.269]	-0.715*** [-2.688]	-0.715*** [-2.688]	-0.608** [-2.269]		-0.715*** [-2.688]	-0.608** [-2.269]	-0.715*** [-2.688]
x1		0.663** [2.583]	0.663** [2.583]			0.663** [2.583]		0.663** [2.583]
x17					1.082*** [3.542]			
x19					1.112*** [3.27]			
x21					1.712*** [5.366]			
R ²	0.707	0.72	0.72	0.707	0.504	0.72	0.707	0.72
adj.R ²	0.697	0.708	0.708	0.697	0.494	0.708	0.697	0.708
N	150	150	150	150	150	150	150	150

t-values in brackets
* (p ≤ 0.1), ** (p ≤ 0.05), *** (p ≤ 0.01)

Los resultados muestran que los modelos 3, 6 y 9 arriban al misma especificación, que implican las siguientes variables: x2, x3, x4, x5 y x20 (en el cuadro ?? se puede ver a qué algoritmo y criterio corresponde cada uno de esos modelos). Los modelos 4, 5, 8 y 10 implican emplear las mismas variables explicativas: x1, x2, x3, x4, x5 y x20 (en el cuadro ?? se puede ver a qué algoritmo y criterio corresponde cada uno de esos modelos). El modelo 7 emplea solamente las variables X17, X19 y x21.

Estos resultados nos llevan a comparar tres modelos; dos están anidados y el otro no.

6.4.2. Comparación de modelos

Finalmente, es importante comparar los modelos 3 modelos. Los tres modelos que compraremos son:

$$y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + \beta_6 x_{5i} + \beta_7 x_{20i} + \varepsilon_i \quad (6.1)$$

$$y_i = \beta_1 + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + \beta_6 x_{5i} + \beta_7 x_{20i} + \varepsilon_i \quad (6.2)$$

$$y_i = \beta_1 + \beta_8 x_{17i} + \beta_9 x_{19i} + \beta_{20} x_{21i} + \varepsilon_i \quad (6.3)$$

Por simplicidad y para evitar confusiones, llamemos a estos tres modelos A, B y C, respectivamente. Así, el modelo B se encuentra anidado en el A. El modelo C no está anidado en los modelos A o B.

```
modeloA <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x20, data)
modeloB <- lm(y ~ x2 + x3 + x4 + x5 + x20, data)
modeloC <- lm(y ~ x17 + x19 + x21, data)
```

El siguiente paso del científico de datos es escoger entre esos modelos. Para esto podemos emplear pruebas F para modelos anidados y la Prueba J para modelos no anidados. Pero antes, es importante anotar que el modelo A es el que tiene el R^2 ajustado más grande. Ahora procedamos a comparar los modelos A y B.

```
anova(modeloB, modeloA)

## Analysis of Variance Table
##
## Model 1: y ~ x2 + x3 + x4 + x5 + x20
## Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x20
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      144 895.94
## 2      143 856.00   1    39.935 6.6713 0.0108 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La prueba F permite rechazar la hipótesis nula de que el modelo B es mejor que el modelo A. Es decir, el modelo A es mejor. Ahora comparemos este modelo con el no anidado

```
library(AER)
J.resA.C <- jtest(modeloA, modeloC)
J.resA.C
```

```
## J test
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x20
## Model 2: y ~ x17 + x19 + x21
##
##           Estimate Std. Error t value Pr(>|t|)
## M1 + fitted(M2)  0.17644    0.111276  1.5856   0.1151
## M2 + fitted(M1)  0.90322    0.084069 10.7438  <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con un 95% de significancia se puede rechazar la hipótesis nula de que el modelo C es mejor que el A. Y no es posible rechazar la nula de que el modelo A es mejor que el C. Así llegamos a la conclusión que el mejor modelo es el A. En otras palabras, el mejor modelo es:

$$y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + \beta_6 x_{5i} + \beta_7 x_{20i} + \varepsilon_i \quad (6.4)$$

Para finalizar, recordemos que los datos fueron simulados de un modelo real en el que las variables explicativas eran de x_1 a x_5 . Las otras variables no se empleaban para simular y . Nuestra selección automática nos lleva a encontrar un modelo muy cercano al real.

6.5. Comentarios finales

Existen otros métodos de selección de modelos menos tradicionales. Por ejemplo, el paquete *subselect* cuenta con algoritmos genéticos (GA) para la selección de modelos (ver función *anneal()*). También se puede explorar la función *RegBest()* del paquete *FactoMineR* que emplea otras técnicas de inteligencia artificial para la selección de modelos.

Parte II

Problemas econométricos en los datos

Capítulo 7

Multicolinealidad

Objetivos del capítulo

Al finalizar este capítulo, el lector estará en capacidad de:

- Identificar los diferentes síntomas que presenta un modelo estimado en presencia de multicolinealidad.
- Efectuar con R diferentes pruebas formales, con el fin de detectar multicolinealidad en el modelo.

7.1. Introducción

Como lo habíamos discutido en capítulos anteriores, si el modelo de regresión múltiple cumple con los supuestos que se resumen en el recuadro 7.1, entonces el Teorema de Gauss-Markov demuestra que los estimadores MCO son MELI (Mejor Estimador Lineal Insesgado) y por lo tanto tienen la menor varianza posible cuando se comparan con todos los estimadores lineales posibles.

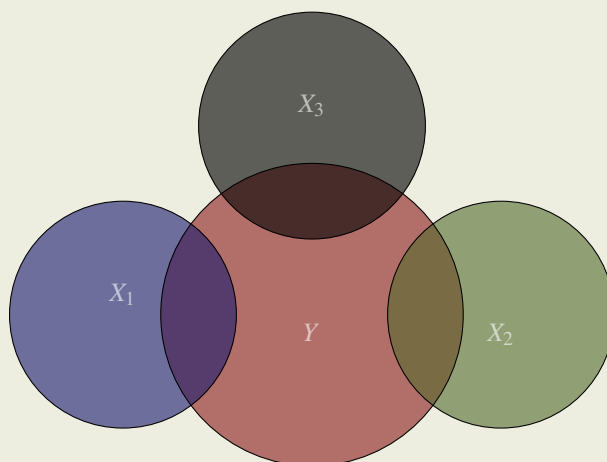
Recuadro 7.1 Supuestos del modelo de regresión múltiple

1. Relación lineal entre y y X_2, X_3, \dots, X_k
2. Las X_2, X_3, \dots, X_k son fijas y linealmente independientes (i.e. la matriz X tiene rango completo)
3. el vector de errores ε satisface:
 - Media cero ($E[\varepsilon] = 0$),
 - Varianza constante
 - No autocorrelación
 Es decir: $\varepsilon_i \sim i.i.d(0, \sigma^2)$ ó $\varepsilon_{n \times 1} \sim (0_{n \times 1}, \sigma^2 I_n)$

Ahora veamos qué ocurre si se viola una parte del supuesto 2. En especial, las variables explicativas (X 's) no sean linealmente independientes entre sí. Es importante anotar que la violación de la otra parte del supuesto no tiene grandes implicaciones sobre el resultado que los estimadores MCO sean MELI (El lector interesado puede consultar el Apéndice al final de este capítulo para una demostración de la insesgadez y eficiencia de este estimador). La violación de este supuesto se puede entender gráficamente con el ejemplo 7.1

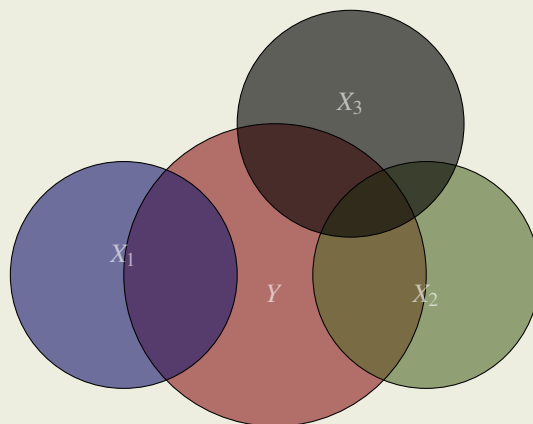
Ejemplo 7.1 Relación entre las variables

Los supuestos 1 y 2 del teorema del Gauss Markov señalan la existencia de una relación lineal entre las variables explicativas y la dependiente, además de una relación linealmente independiente entre las variables explicativas (X 's). El siguiente gráfico representa estos supuestos. Podemos observar que existe una relación entre la variable explicada y las independientes, pero a su vez independencia entre estas últimas. Los círculos representan las variables y sus intersecciones la relación entre ellas.

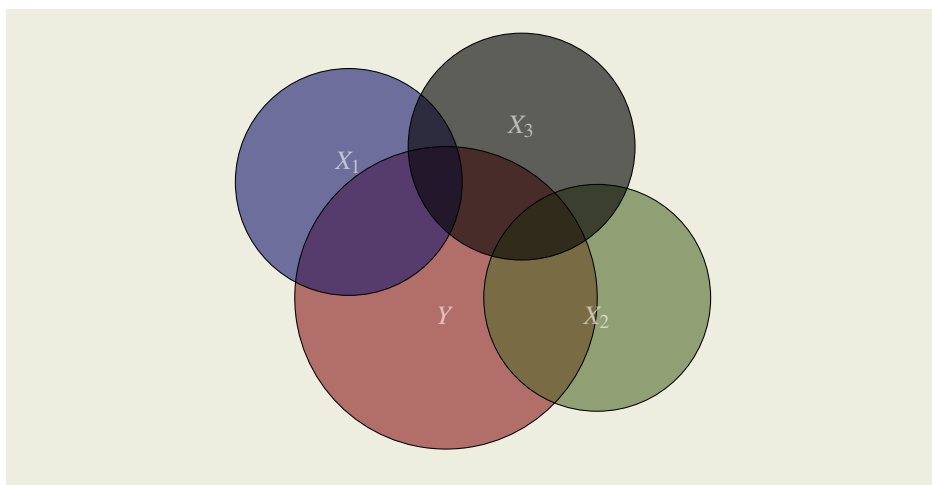


El problema aparece cuando tenemos algún tipo de relación lineal entre las variables independientes o entre un subconjunto de ellas.

En el siguiente gráfico podemos apreciar la presencia de una relación entre las variables X_2 y X_3 , es decir, una parte de la información proporcionada por X_2 está a su vez contenida en la variable X_3 . En este caso, sí X_2 cambia, esto provoca un cambio directo en Y (representado por β_2) y también un cambio indirecto; pues al cambiar X_2 , X_3 cambia y esto a su vez provoca el cambio en Y .



O podría darse el caso en el que todas las variables independientes se encuentren relacionadas; es decir, todas ellas comparten una parte de la información contenida en cada una.

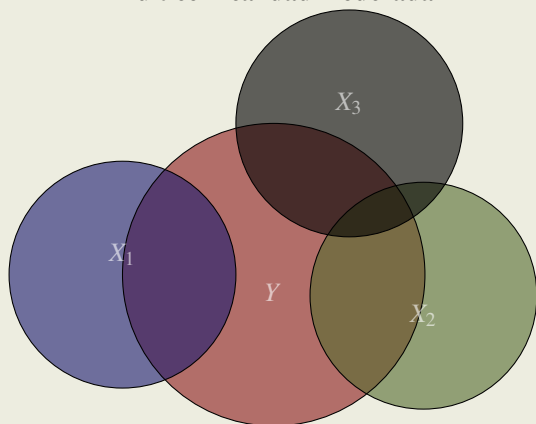


7.2. Los diferentes grados de multicolinealidad

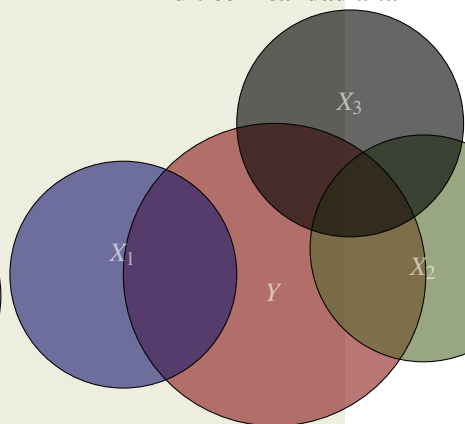
En general, cuando hay cierto grado de relación lineal entre las variable independientes decimos que existe multicolinealidad (o colinealidad). En la práctica, tendremos diferentes grados de multicolinealidad; en el Ejemplo 2 se presentan los cuatro posibles grados o tipos de multicolinealidad.

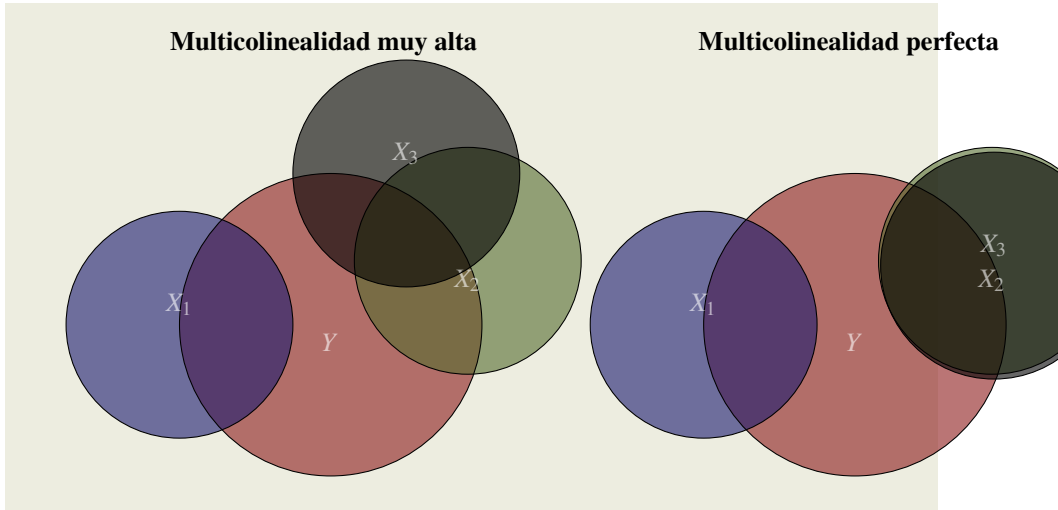
Ejemplo 7.2

Multicolinealidad moderada



Multicolinealidad alta





Discutiremos primero la multicolinealidad perfecta y sus efectos.

7.2.1. *Multicolinealidad perfecta*

Partamos de un ejemplo. Supongamos que queremos explicar la relación entre el peso en kilogramos de un individuo (kg_i) con las horas diarias promedio de actividad física y las calorías consumidas. Veamos qué sucede si empleamos el siguiente modelo:

$$kg_i = \beta_0 + \beta_1 ha_i + \beta_2 cd_i + \beta_3 cs_i + e_i$$

donde las variables ha_i , cd_i y cs_i corresponden a los promedios de horas diarias de actividad física, calorías consumidas por día y calorías consumidas por semana por el individuo i , respectivamente.

Como podemos apreciar, las variables cd_i y cs_i presentan una relación lineal perfecta (y por tanto multicolinealidad perfecta) debido a que siempre vamos a tener que $7cd_i = cs_i$ y por lo tanto las X 's no son linealmente independientes entre sí.

Matricialmente este hecho se puede representar de la siguiente manera:

$$\mathbf{y} = \begin{bmatrix} kg_1 \\ kg_2 \\ \vdots \\ kg_n \end{bmatrix}_{n \times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & ha_1 & cd_1 & 7cd_1 \\ 1 & ha_2 & cd_2 & 7cd_2 \\ 1 & ha_3 & cd_3 & 7cd_3 \\ 1 & ha_4 & cd_4 & 7cd_4 \\ 1 & ha_5 & cd_5 & 7cd_5 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{n \times 4} \quad (7.1)$$

Las consecuencias de esta relación entre dos columnas de la matriz \mathbf{X} es que ésta no tendrá rango columna completo y por tanto $\mathbf{X}^T\mathbf{X}$ no tendrá rango completo. Es decir, $\det(\mathbf{X}^T\mathbf{X}) = 0$. Así, $(\mathbf{X}^T\mathbf{X})^{-1}$ no existirá y por tanto $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ no existirá. En este caso, si queremos eliminar la multicolinealidad perfecta entre cd_i y cs_i , debemos eliminar cualquiera de las dos variables, ya que ambas están aportando la misma información al modelo.

En resumen, el problema que se presenta en presencia de multicolinealidad perfecta es que las columnas de la matriz de las X 's no son linealmente independientes y esto implicará que el estimador de MCO no existirá. Por lo tanto, si un modelo presenta multicolinealidad perfecta, entonces cualquier paquete estadístico reportará un error y no podrá estimar los coeficientes. Finalmente, aunque es imposible estimar un modelo con este problema, de todas formas por el mismo diseño del modelo es fácil de detectar y de resolver el problema.

Intuitivamente, el problema de multicolinealidad perfecta implica que la información contenida en una variable es redundante pues esa información ya se recoge en otras variables explicativas. Así el problema de multicolinealidad perfecta es un problema de cómo se plantea el modelo y en consecuencia es fácil de solucionar.

También es posible estar expuesto a la presencia de multicolinealidad perfecta al utilizar variables dummy. Por ejemplo, supongamos que queremos ver el efecto del sector de la economía donde se emplea un individuo sobre el salario (w_i). Supongamos que el modelo, sin tener en cuenta el efecto del sector económico, es:

$$w_i = \lambda_1 + \lambda_2 (E_i) + \lambda_3 (C_i) + \mu_i \quad (7.2)$$

donde E_i y C_i representan los años de educación y los años de capacitación del individuo i respectivamente

Ahora, supongamos que la economía tiene tres diferentes sectores: primario (Agricultura, minería, ganadería, etc.), secundario (Manufacturas, etc.), terciario (Comercio, servicios, etc.) y además un individuo solo puede recibir un salario si trabaja en uno de esos tres sectores.

Esto implica generar las siguientes variables dummy:

$$D_{1i} = \begin{cases} 1 & \text{si } i \in \text{Sec. Primario} \\ 0 & \text{o.w.} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{si } i \in \text{Sec. Secundario} \\ 0 & \text{o.w.} \end{cases}$$

$$D_{3i} = \begin{cases} 1 & \text{si } i \in \text{Sec. Terciario} \\ 0 & \text{o.w.} \end{cases}$$

entonces, nuestro modelo se convierte en:¹

$$w_i = \lambda_1 + \lambda_2 (E_i) + \lambda_3 (C_i) + \lambda_4 D_{1i} + \lambda_5 D_{2i} + \lambda_6 D_{3i} + \mu_i$$

$$E[w_i] = \lambda_1 + \lambda_2 E_i + \lambda_3 C_i + \lambda_4 E[D_{1i}] + \lambda_5 E[D_{2i}] + \lambda_6 E[D_{3i}]$$

Pero si analizamos el anterior modelo mediante su expresión matricial, nos damos cuenta rápidamente del problema que aparece. Matricialmente el modelo será:

$$\mathbf{y} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}_{n \times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & E_1 & C_1 & 1 & 0 & 0 \\ 1 & E_2 & C_2 & 0 & 0 & 1 \\ 1 & E_3 & C_3 & 0 & 1 & 0 \\ 1 & E_4 & C_4 & 0 & 0 & 1 \\ 1 & E_5 & C_5 & 0 & 0 & 1 \\ 1 & E_6 & C_6 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{n \times 6}$$

Para todas las observaciones ($\forall i$), tenemos que $D_{1i} + D_{2i} + D_{3i} = 1$. Además, la columna de la constante será igual a 1 en cada observación. Por lo anterior, concluimos que las X 's no son linealmente independientes, al existir una combinación lineal de las columnas 4, 5 y 6 que es exactamente igual a la primera columna.

Si queremos no tener multicolinealidad perfecta entre las variables dummy y el intercepto debemos eliminar una de las dummy obteniendo el siguiente modelo:

$$w_i = \lambda_1 + \lambda_2 (E_i) + \lambda_3 (C_i) + \lambda_4 D_{1i} + \lambda_5 D_{2i} + \mu_i$$

Ahora,

$$D_{1i} + D_{2i} \neq 1 \quad \forall i$$

y por tanto, no hay relación lineal entre las dummy y el intercepto.

En este caso $E[w_i] = \lambda_1 + \lambda_2 E_i + \lambda_3 C_i + \lambda_4 E[D_{1i}] + \lambda_5 E[D_{2i}]$. Es decir:

$$E[w_i] = \begin{cases} (\lambda_1 + \lambda_4) + \lambda_2 E_i + \lambda_3 C_i & \text{si } i \in \text{sec. Prim} \\ (\lambda_1 + \lambda_5) + \lambda_2 E_i + \lambda_3 C_i & \text{si } i \in \text{sec. Sec} \\ \lambda_1 + \lambda_2 E_i + \lambda_3 C_i & \text{si } i \in \text{sec. Ter} \end{cases}$$

En general cuando usamos variables dummy tenemos que tener cuidado si existen j posibilidades diferentes entonces usamos $j - 1$ variables dummy ó j variables dummy y quitamos el intercepto, así evitamos multicolinealidad perfecta.

7.2.2. Consecuencias de la multicolinealidad no perfecta

En la práctica, el problema de multicolinealidad perfecta es muy raro, pero sí es común contar con modelos con variables independientes altamente corre-

¹ Por simplicidad solo consideraremos cambios en el intercepto.

lacionadas entre sí, sin que esta relación sea perfecta. Supongamos que dos o más variables están relacionadas (pero no de manera perfecta), en este caso se tendrá que el $\det(\mathbf{X}^T\mathbf{X}) = |\mathbf{X}^T\mathbf{X}|$ existe, pero tiende a cero. Por lo tanto, la matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ si existe pero en general tendrá valores muy grandes. Recuerde que $(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{|\mathbf{X}^T\mathbf{X}|} \text{Adj}(\mathbf{X}^T\mathbf{X})$.

Como la matriz de varianzas y covarianzas de los coeficientes estimados depende de la matriz² $(\mathbf{X}^T\mathbf{X})^{-1}$, entonces las varianzas de los β' s tenderán a ser a su vez muy grandes. Esto implica que los t-calculados de los β' s para probar la significancia de los coeficientes estimados sean relativamente bajos,³ y así se tenderá a no rechazar la hipótesis nula de no significancia individual de los coeficientes. Así mismo una matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ con valores relativamente grandes, implicará que la suma cuadrada de la regresión $(SSR = ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T \mathbf{X}^T\mathbf{y} - n\bar{y}^2)$ será relativamente grande y por tanto el R^2 y el $F - \text{Global}$ serán relativamente grandes ya que estos dependen del SSR.

Así, los síntomas más comunes de la multicolinealidad no perfecta se pueden resumir de la siguiente manera:

1. t calculados bajos acompañados de $F - \text{Global}$ y R^2 altos
2. Sensibilidad de los β' s estimados a cambios pequeños en la muestra⁴
3. Sensibilidad de los β' s a la inclusión o exclusión de regresores⁵

No obstante la multicolinealidad no perfecta provoca estos síntomas, en muchos casos este problema es ignorado por los científicos de datos. Existen varias razones. Primero, si se parte de un modelo teórico que claramente determina cuáles son los variables a incluir, entonces la existencia de multicolinealidad no perfecta es problema de los datos y no un problema del modelo o del método de estimación. Es decir, la muestra que se emplea es la que tiene el problema y de pronto en otras muestras el problema no está presente. Una segunda razón para ignorar la existencia de la multicolinealidad no perfecta es que se privilegie un R^2 alto para el problema bajo estudio. En especial, si la interpretación de los coeficientes no es importante, pero si se desea generar buenos pronósticos, entonces se podría privilegiar la existencia de este problema en vez de entrar a solucionarlo. En todo caso, siempre es mejor saber si este problema existe o no, ya sea que se ignore o no.

Por otro lado, existen algunas técnicas estadísticas que permiten lidiar con el problema de la multicolinealidad no perfecta, como por ejemplo la “ridge regression” o el método de componentes principales para condensar la información de las variables relacionadas en una sola. Pero en la realidad las prácticas más empleadas por los científicos de datos para “lidiar” con la multicolinealidad no perfecta son:

² $\text{Var}[\hat{\beta}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

³ Recordemos que $t_c = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$.

⁴ Esto ocurre porque la matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ cambiará mucho con incluir o eliminar una fila de \mathbf{X} .

⁵ Esto ocurre porque $(\mathbf{X}^T\mathbf{X})^{-1}$ cambiará mucho con incluir o eliminar una columna de \mathbf{X} .

- Descartar una variable de las que presentan el problema. Esta aproximación no será aceptable para el investigador si la especificación del modelo inicial proviene directamente de la teoría económica. La omisión de una variable relevante (según la teoría) para explicar la variable dependiente provoca un sesgo de omisión. Problema que es mas grave que el problema causado por la multicolinealidad no perfecta. Así, este remedio puede ser mas costoso que el mismo problema de multicolinealidad.
- Transformar las variables relacionadas, de tal forma que el modelo involucre razones de las variables y no las variables en sus niveles. Esta aproximación también puede no ser aceptable, pues la teoría económica sugiere los cambios de los niveles de la variable dependiente, dados cambios en los niveles de las variables explicativas y no en las razones de estas variables.
- Aumentar la muestra. Al aumentar la muestra, se puede brindar más información que permita mejorar la precisión en la estimación de los coeficientes y por tanto la disminución del error estándar de estos. El problema con esta aproximación es que en la mayoría de los casos el investigador emplea la mayor cantidad de datos disponibles para el problema y aumentar la muestra es una misión casi imposible.
- No hacer nada. Si bien se identifica la presencia de éste problema, en muchos casos no se realiza algún procedimiento para resolverlos, pues la multicolinealidad no perfecta no afecta las propiedades de MELI de los MCO.⁶ El problema aparece en la interpretación de los coeficientes y en algunos casos en la reducción de los t-calculados que podrían ser lo suficientemente grandes como para rechazar la nula de no significancia si la multicolinealidad no perfecta no estuviese presente. Así, si esta es la opción que se adopta para “lidiar” con la multicolinealidad, se tendrá que ser muy cuidadoso con la interpretación de los coeficientes y con la inferencia respecto a estos.

En conclusión, de encontrarse multicolinealidad no perfecta en un modelo se debe ser muy cauteloso en la interpretación del R^2 y de los coeficientes, así como su significancia. Por otro lado, si el objetivo del modelo estimado con multicolinealidad es la de producir pronósticos, entonces el problema de multicolinealidad no es muy importante, siempre y cuando la relación entre las variables explicativas se mantenga en el período proyectado.

7.3. Pruebas para la detección de multicolinealidad

Además de chequear los síntomas, en la práctica es necesaria la utilización de pruebas más formales con el fin de detectar la presencia de multicolinealidad no

⁶ En el Apéndice 2 del Capítulo 2, se presenta la demostración del teorema de Gauss-Markov. En esta demostración, es fácil notar que las propiedades de insesgadez y mínima varianza de los estimadores MCO dependen de los supuestos asociados al término de error y no a qué tan grande o pequeño sea el determinante de la matriz $X^T X$.

perfecta. A continuación se describen tres de las pruebas más utilizadas para este fin.

7.3.1. *Factor de Inflación de Varianza (VIF)*

Dado que uno de los síntomas de la multicolinealidad no perfecta es que los errores estándar de los coeficientes estimados por el método de MCO (y por tanto su varianza) es más grande de lo que debería ser, una forma de detectar la existencia de este problema es mirar como se “infla” la varianza de un coeficiente por no ser este independiente a los demás. En otras palabras, el Factor de Inflación de Varianza (*VIF* por su sigla en inglés que viene del término Variance inflation factor) compara cuál hubiese sido la varianza de un coeficiente determinado si la correspondiente variable fuera totalmente independiente a las demás con el valor realmente observado de dicha varianza. Así, el *VIF* mide en cuantas veces se aumentó la varianza de un coeficiente por la existencia de un posible problema de multicolinealidad no perfecta.

EL *VIF* para el coeficiente j se define como

$$VIF_j = \frac{1}{1 - R_j^2} \quad (7.3)$$

donde R_j^2 es el R^2 de la regresión de X_j en función de los demás regresores. Así, el *VIF* muestra el aumento en $Var[b_j]$ que puede atribuirse al hecho que esa variable no es ortogonal a las otras variables del modelo.

Algunos autores como **Babin2014** argumentan que si el VIF_j excede 3.0, entonces se considera que existe un problema de multicolinealidad. Otros autores como **sheather2009modern** afirman que un VIF_j mayor a 4 es síntoma de un problema grande. No obstante una regla empírica (rule of thumb) muy común en la práctica es considerar el problema de multicolinealidad alta si el VIF_j es mayor a 10 (Ver por ejemplo **Kutner2004**).

7.3.2. *Prueba de Belsley, Kuh y Welsh (1980)*

Esta prueba diseñada por **Belsley1980a** también es conocida con el nombre de prueba Kappa. Esta prueba se basa en los valores propios de la matriz $\mathbf{X}^T \mathbf{X}$. Ellos demostraron que empleando el valor máximo y mínimo de esta matriz es posible detectar la multicolinealidad. La prueba se construye de la siguiente manera:

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_k}} \quad (7.4)$$

donde λ_1 es el valor propio más grande de $\mathbf{X}^T \mathbf{X}$ y λ_k es el valor propio más pequeño. Los autores demostraron que si $\kappa > 20$ entonces existe un problema de multicolinealidad.

7.4. Soluciones de la multicolinealidad (¡Sí se necesitan!)

7.4.1. Regresión de Ridge

Una forma de solucionar el problema de la multicolinealidad es emplear la regresión de Ridge, pero esta solución trae un costo asociado. Esta aproximación es una variante de los MCO, cuyo objetivo es evitar el problema de Multicolinealidad modificando a la matriz $\mathbf{X}^T \mathbf{X}$. La modificación se realiza de tal manera que $\det(\mathbf{X}^T \mathbf{X})$ se aleja de cero.

El costo de esta aproximación es que los nuevos parámetros estarán sesgados (en el Apéndice de este capítulo se presenta una demostración del sesgo que presenta este estimador), pero la varianza es más pequeña. es decir, existe un “tradeoff” entre varianza y sesgo. Como dicen los economistas, “no hay un almuerzo gratis” pues este método implica la aparición de un nuevo parámetro l .

Como establecer este nuevo parámetro se convierte en la gran pregunta de este método. Antes de continuar es importante anotar que la regresión de Ridge es un recurso de última instancia. Solo se emplea cuando la multicolinealidad es casi perfecta.

El estimador de Ridge está definido como

$$\mathbf{b}(l) = (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (7.5)$$

Esto implica que

$$\mathbf{b}(l) = \left(\mathbf{I} + l(\mathbf{X}^T \mathbf{X})^{-1} \right)^{-1} \hat{\beta} \quad (7.6)$$

donde $l > 0$ y es no estocástico. l se conoce como el parámetro de reducción de sesgo, y este parámetro se escoge con algoritmos automáticos como por ejemplo: HKB (Hoerl, Kennard y Baldwin (1975)), L-W o GCV (Generalized Crossvalidation). Este último algoritmo es el más común.

l también se puede escoger graficando los coeficientes estimados en función de l y se selecciona l mas pequeño que produce estimadores estables.

7.4.2. Remover variables con alto VIF

Es común que los científicos de datos empleen una aproximación diferente a las dos anteriores para solucionar el problema de la multicolinealidad. Dado que los científicos de datos típicamente se enfrentan a tener una clara variable dependiente

y un grupo grande de variables explicativas, una aproximación natural es descartar de manera recursiva las variables del modelo que tengan un *VIF* mas grande de un determinado umbral (típicamente mayor a 4).

Así, esta aproximación implica los siguientes pasos:

- **Paso1:** Calcular el *VIF* para cada variable explicativa del modelo
- **Paso2:** Identificar la variable explicativa con el mayor *VIF*
- **Paso3:** Si el mayor *VIF* es inferior al umbral determinado (normalmente 4), parar. En caso contrario, re-estimar el modelo sin dicha variable
- **Paso4:** Regresar al primer paso

De esta manera, se arriba a un modelo con todas las variables con un *VIF* relativamente pequeño. Esta aproximación tiene sentido en un contexto en el cual no se cuenta con una teoría de tras que guíe el análisis de se realiza. Adicionalmente, esta solución es muy fácil de automatizar.

7.5. Práctica en R: Análisis del efecto discriminatorio de género en las diferencias salariales en Colombia

El objetivo de este ejercicio es aplicar las tres pruebas de multicolinealidad anteriormente descritas y de ser necesario solucionar dicho problema. Para este fin y con la misma metodología seguida por **luisa1** se analizan las brechas salariales entre hombres y mujeres en el municipio de Santiago de Cali en el año 2003. Los datos se encuentran disponibles en el archivo *DatosMultiColinealidad.csv*.

Uno de los métodos más utilizados para determinar el efecto de la discriminación en la brecha salarial es la propuesta por **oaxaca1**, quienes sugieren una técnica para medir la diferencia que tiene sobre el ingreso, características como el capital humano entre géneros. Para estimar este efecto, se emplea un modelo de ecuaciones Mincerianas (**mincer1**) que fueron pensadas para estimar la rentabilidad de la educación. Aunque originalmente el autor propone estimar dos ecuaciones por separado, una para hombres y otra para mujeres, en nuestro caso sólo estimaremos una ecuación y añadiremos una variable dummy para capturar el efecto que tiene el género sobre el salario.

De esta manera, el modelo a estimar es el siguiente:

$$\begin{aligned} \ln(ih_i) = & \beta_0 + \beta_1 yedu_i + \beta_2 exp_i + \beta_3 exp_i^2 + \beta_4 D_i \\ & + \beta_5 Dyedu_i + \beta_6 Dexp_i + \beta_7 Dexp_i^2 + \varepsilon_i \end{aligned} \quad (7.7)$$

donde

$$D_i = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{o.w.} \end{cases}$$

$\ln(ih_i)$ representa el logaritmo natural del ingreso por hora del individuo i , $yedu_i$ y exp_i denotan los años de educación y de experiencia del individuo i . Los datos se

encuentran en el archivo DatosMultiColinealidad.csv. Antes de analizar los datos, es claro que la especificación de este modelo incluye dos variables que están relacionadas, pero no de manera lineal; es decir, exp y exp^2 . Como la relación es cuadrática y no lineal, no hay razón por la cual esperar, a priori, la presencia de multicolinealidad perfecta o no perfecta.

Como siempre, el primer paso será cargar los datos y constatar que estos quedan bien cargados. Noten que la variable exp_i^2 no está en el data frame. Si el nombre de las variables no queda bien leído, cámbielos. El modelo lo podemos estimar de la siguiente manera sin necesidad de crear una nueva variable:

La notación $I()$ en la fórmula implica que R hará la operación presentada dentro del paréntesis y se agregará con una variable explicativa. Los resultados de la estimación del modelo se reportan en el cuadro ?? . Antes de entrar a calcular los estadísticos para determinar la presencia de multicolinealidad, es importante anotar que los síntomas no están presentes, pero esto implica que no exista multicolinealidad. Pero recuerden que antes de analizar cualquier resultado es importante estar seguro que no existen problemas de multicolinealidad.

Cuadro 7.1 Modelo estimado por MCO

	Lnih
intercepto	5.856*** [77.641]
yedu	0.131*** [32.801]
exp	0.034*** [6.222]
I(exp ²)	0*** [-2.888]
sexomujer	-0.117 [-1.386]
exp:sexomujer	-0.004 [-0.503]
I(exp ²) : sexomujer	0 [0.345]
R ²	0.454
adj.R ²	0.452
N	1415

t-values in brackets

* (p ≤ 0.1), ** (p ≤ 0.05), *** (p ≤ 0.01)

7.5.1. Pruebas de multicolinealidad

7.5.1.1. VIF

Para calcular el *VIF* emplearemos el paquete *car*. Este paquete tiene la función *vif()* cuyo único argumento es un objeto *lm* con el modelo estimado.

```
library(car)
vif(res1)

##          yedu          exp          I(exp^2)          sexo
##    1.249719    19.643504    18.080978    7.589628
##    exp:sexo I(exp^2):sexo
##    43.865651    25.438134
```

En este caso el *VIF* para las variables *exp*, exp^2 , *sexo* y la interacción de esta última con las dos anteriores son muy grandes. por ejemplo, *exp* tiene una varianza aproximadamente 20 veces más grande que si las variables no presentan colinealidad.

Claramente, los resultados implican la existencia de un problema delicado de multicolinealidad, no obstante los síntomas no estaban muy claramente presentes.

7.5.1.2. Prueba de Belsley, Kuh y Welsh (1980)

Para realizar esta prueba es necesario encontrar la matriz $\mathbf{X}^T \mathbf{X}$ y calcular sus valores propios. Esto se puede hacer empleando la función *model.matrix()* que permite obtener la matriz *X*. Esta función solo tiene como argumento un objeto *lm* con el modelo estimado. Los valores propios de una matriz se pueden encontrar empleando la función *eigen*. A continuación se presenta el código que calcula esta prueba.

```
XTX <- model.matrix(res1)
e <- eigen(t(XTX) %*% XTX)
e$val

## [1] 1.214720e+09 1.856133e+08 2.232308e+05 2.870475e+04
## [5] 1.723882e+04 1.191735e+02 3.274535e+01

lambda.l <- max(e$val)
lambda.k <- min(e$val)
kappa <- sqrt(lambda.l/lambda.k)
kappa

## [1] 6090.644
```

Este estadístico es muy grande ($\kappa = 6090.644165$). Esta prueba también coincide en la existencia de un problema serio de multicolinealidad.

Finalmente y teniendo en cuenta los resultados de todas las pruebas efectuadas, podemos concluir que el modelo presenta una alta correlación entre las variables *exp* y exp^2 . De la definición de las variables sabemos que no existe una relación

lineal perfecta, pero nuestro hallazgo puede ser síntoma de que los valores de la variable *exp* corresponden a un rango relativamente corto, para el cual la relación exponencial puede ser aproximada por una relación lineal. Ahora bien, no estamos frente a un problema de multicolinealidad perfecta, y por lo tanto no se está violando el segundo supuesto del teorema de Gauss Markov. Por otro lado, el modelo teórico necesita incluir en la especificación ambas variables por razones bien fundamentadas en **oaxaca1** y sería incorrecto eliminar alguna de las dos variables.

El lector puede proceder a determinar si existe o no diferencias salariales entre las mujeres y los hombres para esta muestra.

7.5.2. Solución del problema removiendo variables con alto VIF

Veamos la siguiente función que permite eliminar de manera automática e iterativa las variables cuyos respectivos coeficientes tengan un *VIF* superior a un umbral determinado (*u*)

El lector deberá seguir con detalle cada línea de la función para entender los trucos que se emplean.

```
remueve.VIF.grande <- function(modelo, u){
  require(car)
  # extrae el dataframe
  data <- modelo$model
  # Calcula todos los VIF
  all_vifs <- car::vif(modelo)
  # extraer el nombre de todas las variables X
  names_all <- names(all_vifs)
  # extraer el nombre de la variables y
  dep_var <- all.vars(formula(modelo))[1]

  # Remover las variables con VIF > u
  # y reestimar el modelo con las otras variables

  while(any(all_vifs > u)){
    # elimina variable con max vif
    var_max_vif <- names(which(all_vifs == max(all_vifs)))
    # remueve la variable
    names_all <- names_all[!(names_all) %in% var_max_vif]
    # nueva formula
    myForm <- as.formula(paste(paste(dep_var, "~ "),
                               paste(names_all, collapse=" + "), sep=""))
    # re-build model with new formula
    modelo.prueba <- lm(myForm, data= data)
    all_vifs <- car::vif(modelo.prueba)
  }
  modelo.limpio <- modelo.prueba
  return(modelo.limpio)
}
```

La función que acabamos de construir (*remueve.VIF.grande*) tiene dos argumentos, el primero es un objeto *lm* y el segundo el umbral para el *VIF*.

Dado que no es necesario solucionar el problema de multicolinealidad en el caso del modelo estudiado. EL lector puede emplear esta función con la regresión del ejercicio al final del capítulo

7.6. Ejercicios

Un investigador es contratado para estimar un modelo sencillo que explique, empleando la teoría económica, la tasa de crecimiento de las importaciones de un pequeño país del Pacífico Sur. Para elaborar la tarea se cuenta con 27 observaciones (la información se encuentra en el archivo *multi.xls*). Las observaciones corresponden a las series de Estados Unidos desde 1990 hasta el 2016 de las tasas de crecimiento de las importaciones (Y_t), el producto interno bruto (X_{1t}) y la inflación (X_{2t}). Al investigador que de inmediato se le asigna esta tarea, realiza una exhaustiva revisión bibliográfica que le lleva a plantear el siguiente modelo para explicar el comportamiento del crecimiento de las exportaciones:

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \varepsilon_t \quad (7.8)$$

1. Estime el modelo 7.8 y explique si existen problemas de multicolinealidad
2. Determine si existe o no multicolinealidad por medio de las pruebas estudiadas en este capítulo
3. Estime dos regresiones simples a partir del modelo 7.8; es decir, la tasa de crecimiento de las importaciones en función de sólo una de las variables independientes mediante los siguientes modelos:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad (7.9)$$

$$Y_i = \beta_0 + \beta_1 X_{2i} + \varepsilon_i \quad (7.10)$$

¿Qué ocurre con la significancia y el ajuste del modelo?

4. Teniendo en cuenta los resultados hasta ahora encontrados, corrija (de ser posible) el problema de multicolinealidad si es que existe. Explique.

7.7. Apéndice

Apéndice 7.1 Demostración de la insesgadez del estimador MCO con \mathbf{X} aleatoria

Recordemos que el estimador MCO ($\hat{\beta}$) esta dado por

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

. Este estimador se puede reescribir de la siguiente manera:

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \varepsilon) \\ \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \\ \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \\ \hat{\beta} &= \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \end{aligned} \tag{7.11}$$

Ahora consideremos el valor esperado para un valor dado de \mathbf{X} no aleatorio.

$$\begin{aligned} E[\hat{\beta} | \mathbf{X}] &= E[\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon | \mathbf{X}] \\ E[\hat{\beta} | \mathbf{X}] &= \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\varepsilon | \mathbf{X}] \end{aligned}$$

Empleando el supuesto de que los errores tienen media cero ($E[\varepsilon | \mathbf{X}] = 0$) se obtiene que

$$E[\hat{\beta} | \mathbf{X}] = \beta$$

Por otro lado, recordemos la *Law of iterated expectations* (Ley de valor esperado iterado) que implica

$$E[W] = E_Z[E[W | Z]]$$

donde $E_Z[\cdot]$ es el valor esperado considerando todos los posibles valores de Z y $E[W | Z]$ es una función de los valores de Z . Entonces, si \mathbf{X} es aleatorio:

$$E[\hat{\beta}] = E_{\mathbf{X}}[E[\hat{\beta} | \mathbf{X}]] = E_{\mathbf{X}}[\beta] = \beta$$

Es decir, el estimador MCO $\hat{\beta}$ es insesgado aún si las variables explicativas son aleatorias.

Apéndice 7.2 Demostración de la eficiencia del estimador MCO con \mathbf{X} aleatoria

Recordemos que el estimador MCO ($\hat{\beta}$) es un estimador lineal dado por

$$\hat{\beta} = \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon$$

Ahora por simplicidad reescribamos el estimador de la siguiente manera:

$$\hat{\beta} = \beta + \mathbf{A}\varepsilon$$

donde $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

Ahora calculemos la matriz de varianzas y covarianzas del estimador MCO.

$$\begin{aligned} \text{Var} [\hat{\beta} | \mathbf{X}] &= E \left[\left(\hat{\beta} - E [\hat{\beta}] \right) \left(\hat{\beta} - E [\hat{\beta}] \right)^T | \mathbf{X} \right] \\ \text{Var} [\hat{\beta} | \mathbf{X}] &= E \left[\left(\hat{\beta} - \beta \right) \left(\hat{\beta} - \beta \right)^T | \mathbf{X} \right] \\ \text{Var} [\hat{\beta} | \mathbf{X}] &= E \left[\left(\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon - \beta \right) \left(\hat{\beta} - \beta \right)^T | \mathbf{X} \right] \\ \text{Var} [\hat{\beta} | \mathbf{X}] &= E \left[\left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \right) \left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \right)^T | \mathbf{X} \right] \\ \text{Var} [\hat{\beta} | \mathbf{X}] &= E \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \varepsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} | \mathbf{X} \right] \\ \text{Var} [\hat{\beta} | \mathbf{X}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E [\varepsilon \varepsilon^T | \mathbf{X}] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ \text{Var} [\hat{\beta} | \mathbf{X}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\sigma^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ \text{Var} [\hat{\beta} | \mathbf{X}] &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \end{aligned}$$

Ahora, debemos demostrar que dicha varianza es la minima posible. Para esto, partamos de escribir de otra manera el estimador MCO

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\hat{\beta} = \mathbf{A} \mathbf{y}$$

Supongamos ahora que existe otro estimador lineal

$$\tilde{\beta} = \mathbf{C} \mathbf{y}$$

Para ser comparable, el nuevo estimador debe ser insesgado

$$E [\tilde{\beta} | \mathbf{X}] = E [\mathbf{C} \mathbf{y} | \mathbf{X}] = \mathbf{C} E [\mathbf{X} \beta + \varepsilon | \mathbf{X}]$$

$$E [\tilde{\beta} | \mathbf{X}] = \mathbf{C} \mathbf{X} \beta + \mathbf{C} E [\varepsilon | \mathbf{X}]$$

$$E [\tilde{\beta} | \mathbf{X}] = \mathbf{C} \mathbf{X} \beta$$

Esto quiere decir que para ser insesgado se necesita

$$\mathbf{C} \mathbf{X} = \mathbf{I}$$

Ahora miremos la varianza, pero antes definamos:

$$\mathbf{D} = \mathbf{C} - \mathbf{A} = \mathbf{C} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

Por lo tanto, tenemos que

$$\mathbf{D}\mathbf{y} = \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$$

y

$$\mathbf{C} = \mathbf{D} + \mathbf{A}$$

Entonces,

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = \text{Var} [(\mathbf{D} + \mathbf{A}) \mathbf{y} | \mathbf{X}]$$

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = \text{Var} [(\mathbf{D} + \mathbf{A}) (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) | \mathbf{X}]$$

Por lo tanto, tenemos que la varianza del otro estimador es

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = \text{Var} [(\mathbf{D} + \mathbf{A}) \mathbf{X}\boldsymbol{\beta} + (\mathbf{D} + \mathbf{A}) \boldsymbol{\varepsilon} | \mathbf{X}]$$

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = \text{Var} [(\mathbf{D} + \mathbf{A}) \boldsymbol{\varepsilon} | \mathbf{X}]$$

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = (\mathbf{D} + \mathbf{A}) \text{Var} [\boldsymbol{\varepsilon} | \mathbf{X}] (\mathbf{D} + \mathbf{A})^T$$

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = (\mathbf{D} + \mathbf{A}) \sigma^2 (\mathbf{D}^T + \mathbf{A}^T)$$

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = \sigma^2 (\mathbf{D}\mathbf{D}^T + \mathbf{D}\mathbf{A}^T + \mathbf{A}\mathbf{D}^T + \mathbf{A}\mathbf{A}^T)$$

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = \sigma^2 (\mathbf{D}\mathbf{D}^T + \mathbf{D}\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}^T + \mathbf{A}\mathbf{A}^T)$$

Entonces, recordemos que para que $\tilde{\boldsymbol{\beta}}$ sea insesgado se necesita

$$\mathbf{C}\mathbf{X} = \mathbf{I}$$

y dado que

$$\mathbf{C} = \mathbf{D} + \mathbf{A}$$

Esto implica que

$$\mathbf{C}\mathbf{X} = \mathbf{I} = \mathbf{D}\mathbf{X} + \mathbf{A}\mathbf{X}$$

Por lo tanto

$$\mathbf{C}\mathbf{X} = \mathbf{I} = \mathbf{D}\mathbf{X} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}$$

$$\mathbf{D}\mathbf{X} = \mathbf{0}$$

Regresando a la varianza del estimador, tendremos que

$$\text{Var} [\tilde{\boldsymbol{\beta}} | \mathbf{X}] = \sigma^2 (\mathbf{D}\mathbf{D}^T + \mathbf{D}\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}^T + \mathbf{A}\mathbf{A}^T)$$

$$\begin{aligned}
\text{Var} [\tilde{\beta} | \mathbf{X}] &= \sigma^2 (\mathbf{D}\mathbf{D}^T + \mathbf{A}\mathbf{A}^T) \\
\text{Var} [\tilde{\beta} | \mathbf{X}] &= \sigma^2 \mathbf{D}\mathbf{D}^T + \sigma^2 \left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right) \left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right)^T \\
\text{Var} [\tilde{\beta} | \mathbf{X}] &= \sigma^2 \mathbf{D}\mathbf{D}^T + \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\
\text{Var} [\tilde{\beta} | \mathbf{X}] &= \sigma^2 \mathbf{D}\mathbf{D}^T + \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \\
\text{Var} [\tilde{\beta} | \mathbf{X}] &= \text{Var} [\hat{\beta} | \mathbf{X}] + \sigma^2 \mathbf{D}\mathbf{D}^T \\
\mathbf{q}^T \mathbf{D}\mathbf{D}^T \mathbf{q} &= \mathbf{z}^T \mathbf{z} \geq 0 \\
\text{Var} [\tilde{\beta} | \mathbf{X}] &\geq \text{Var} [\hat{\beta} | \mathbf{X}]
\end{aligned}$$

Q.E.D. Por lo tanto no es posible obtener un estimador insesgado con una varianza menor, aún en presencia de regresores aleatorios.

Apéndice 7.3 Demostración del sesgo del estimador de la regresión de Ridge

Partamos de la definición del estimador de Ridge

$$\mathbf{b}(l) = (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Ahora calculemos el valor esperado de dicho estimador

$$\begin{aligned}
E[\mathbf{b}(l) | \mathbf{X}] &= E \left[(\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} | \mathbf{X} \right] \\
E[\mathbf{b}(l) | \mathbf{X}] &= (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T E[\mathbf{y} | \mathbf{X}] \\
E[\mathbf{b}(l) | \mathbf{X}] &= (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T E[\mathbf{X}\beta + \varepsilon | \mathbf{X}] \\
E[\mathbf{b}(l) | \mathbf{X}] &= (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T \mathbf{X}\beta
\end{aligned}$$

Es fácil ver que el sesgo será mayor si l es más grande. Por otro lado, es fácil mostrar que la varianza de este estimador es:

$$\text{Var}[\mathbf{b}(l) | \mathbf{X}] = \sigma^2 (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1}$$

Así, la varianza del estimador tiende a ser más pequeña entre más grande sea l . De hecho, $\text{Var}[\hat{\beta} | \mathbf{X}] > \text{Var}[\mathbf{b}(l) | \mathbf{X}]$

Parte III

Problemas econométricos en los datos

Capítulo 8

Heteroscedasticidad

Objetivos del capítulo

Al finalizar este capítulo, el lector estará en capacidad de:

- Realizar diferentes tipos de análisis gráficos que revelen la posibilidad de heteroscedasticidad en los residuos empleando R.
- Efectuar utilizando R las pruebas estadísticas necesarias para detectar la violación del supuesto de homoscedasticidad en los residuos. En especial las pruebas de Breusch-Pagan y la prueba de White.
- Corregir el problema de heteroscedasticidad empleando estimadores consistentes para los errores estándar en R.

8.1. Introducción

Como lo hemos discutido en capítulos anteriores, si los supuestos del modelo de regresión que se resumen en el recuadro 8.1 se cumplen, entonces el Teorema de Gauss-Markov demuestra que los estimadores MCO son MELI (Mejor Estimador Lineal Insesgado). En el capítulo anterior analizamos las consecuencias de la violación del supuesto de independencia lineal entre variables explicativas en un modelo de regresión múltiple. También discutimos cómo dicho problema era un problema de los datos. En este Capítulo nos concentramos en la violación del supuesto que el término de error tiene varianza constante.

Recuadro 8.1 Teorema de Gauss-Markov

Si se considera un modelo lineal $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$ y se supone que:

- Las X_2, X_3, \dots, X_k son fijas y linealmente independientes (es decir X tiene rango completo y es una matriz no estocástica).
- El vector de errores $\boldsymbol{\varepsilon}$ tiene media cero, varianza constante y no autocorrelación. Es decir: $E[\boldsymbol{\varepsilon}] = 0$ y $Var[\boldsymbol{\varepsilon}] = \sigma^2 I_n$

Entonces el estimador de MCO $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ es el *Mejor Estimador Lineal Insesgado (MELI)*

En ocasiones el supuesto de homoscedasticidad o varianza del término de error constante no tiene mucho sentido. Un ejemplo de esto se presenta al analizar el consumo en función del ingreso con una muestra de hogares (corte transversal). Supongamos que se desea emplear un modelo en que el consumo del hogar i depende de su nivel de ingresos y otras variables que caracterizan la condición socio demográfica del hogar. Los hogares con ingresos bajos presentan un comportamiento en su consumo mucho menos variable que el consumo de los hogares con altos ingresos. Es decir, es de esperarse que para ingresos altos el comportamiento del consumo se disperse más con respecto a lo esperado (su media). Por otro lado, a medida que el ingreso sea más bajo los hogares no podrán tener una dispersión muy grande con respecto a lo esperado para su nivel de ingresos. Esto implica que las observaciones correspondientes al consumo de hogares con ingresos bajos tendrán una varianza con respecto a su valor esperado (varianza del error) mucho menor que aquellos hogares con ingresos altos. También existen otras razones para que se presente la heteroscedasticidad, como por ejemplo el aprendizaje sobre los errores y mejoras en la recolección de la información a medida que ésta se realiza.

En general, este problema econométrico es muy común en datos de corte transversal, aunque es posible que el problema también se presente con series de tiempo; especialmente si se están modelando rendimientos de activos o el valor de activos financieros como las acciones. Formalmente, en presencia de este problema, la matriz de varianzas y covarianzas de los errores será:

$$Var[\varepsilon] = E[\varepsilon^T \varepsilon] = \Omega = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \neq \sigma^2 I$$

En presencia de heteroscedasticidad, los estimadores MCO siguen siendo insesgados, pero ya no tienen la mínima varianza posible¹. Es decir, los estimadores MCO no son MELI.

De hecho, el estimador MCO de la matriz de varianzas y covarianzas ($Var[\hat{\beta}] = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$) será sesgado. En presencia de heteroscedasticidad la matriz de varianzas y covarianzas del estimador MCO del vector β es:²

$$Var[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Omega \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

Esto implica que el estimador de la matriz de varianzas y covarianzas $\widehat{Var}[\hat{\beta}] = s^2(\mathbf{X}^T \mathbf{X})^{-1}$ sea sesgado; y por lo tanto, si usamos este último estimador en pruebas de hipótesis o intervalos de confianza para los coeficientes estimados obtendremos conclusiones erróneas en torno a los verdaderos β s. Esto ocurrirá en las pruebas individuales y conjuntas.

Así, si se emplea el estimador MCO en presencia de heteroscedasticidad, entonces los estimadores de los coeficientes serán insesgados, pero los errores estándar no serán los adecuados. Por tanto, cualquier conclusión derivada de la inferencia a partir de los estimadores MCO será incorrecta.

8.2. Pruebas para la detección de heteroscedasticidad

En general, una buena práctica cuando se estiman modelos econométricos es emplear gráficos que permitan intuir que está ocurriendo con los residuos estimados. Esto permite intuir si existen síntomas de la presencia de heteroscedasticidad. Obviamente, los gráficos no proveen evidencia contundente para concluir, pero sí proveen la intuición necesaria para iniciar las pruebas formales.

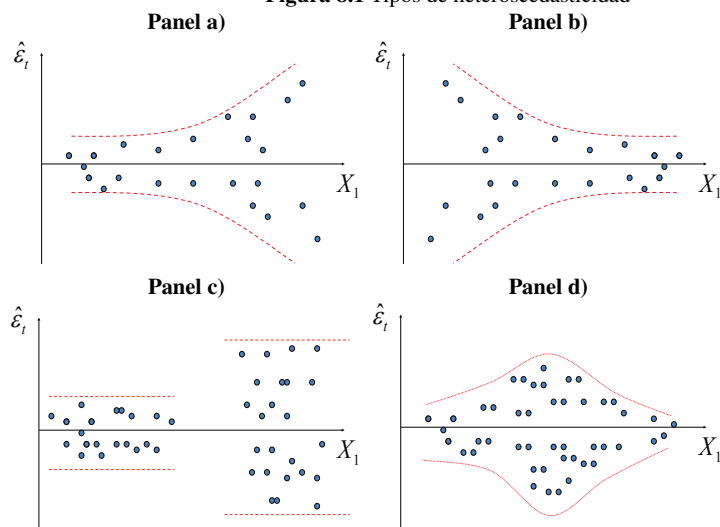
Dado que la heteroscedasticidad implica una variabilidad no constante del error, entonces lo más adecuado sería poder graficar el vector de errores (ε). Lastimosamente, este vector no es observable, pero la mejor aproximación que tenemos para conocer ese vector sería el error estimado ($\hat{\varepsilon}$).

Así, el primer paso para detectar la presencia de heteroscedasticidad es graficar los residuos del modelo estimado. Los gráficos más empleados son gráficos de dispersión de:

¹ En el Apéndice 8.1 se presenta una demostración.

² En el Apéndice 8.2 se presenta una demostración.

1. Los errores estimados versus las observaciones ($\hat{\varepsilon}_i$ vs $i = 1, 2, \dots, n$)
2. Los errores estimados versus cada una de las variables explicativas ($\hat{\varepsilon}_i$ vs cada una de las columnas de bfX)

Figura 8.1 Tipos de heteroscedasticidad

En estos gráficos se busca algún tipo de regularidad como la que se presenta en la figura 8.1. En el panel a) observamos que los residuos se vuelven más grandes a medida que la variable dependiente crece. En este caso la dispersión con respecto a la media (cero) crece a medida que la variable explicativa crece y esto sucede de manera exponencial. Por tanto, existe heteroscedasticidad que podría ser del tipo $\sigma_i^2 = X_{1i} \sigma^2$. En el panel b) el comportamiento de la varianza es opuesto al presentado en el panel a), heteroscedasticidad que podría ser del tipo $\sigma_i^2 = \frac{1}{X_{1i}} \sigma^2$. En el panel c) claramente existen dos grupos de datos; los de varianza grande y los de baja varianza. Por último en el panel d) los residuos son menores para las mediciones grandes y pequeñas pero crecen en los valores intermedios de la misma. En los cuatro casos hay síntomas de heteroscedasticidad.

Antes de entrar a considerar las pruebas formales, es importante aclarar que la heteroscedasticidad implica una varianza diferente del error para al menos una observación, así existen muchas formas de heteroscedasticidad. Es decir, la heteroscedasticidad puede depender de una variable explicativa, de una variable que no se incluye en el modelo o de cualquier función de las observaciones. Esto hace difícil la tareas de identificar el problema con gráficos y con pruebas.

Por otro lado, en la práctica los científicos de datos enfrentan comúnmente problemas en los que los modelos incluyen muchas variables explicativas, así realizar un análisis gráfico puede ser un trabajo que tome mucho tiempo y no genere mu-

cha intuición sobre este problema. En esas situaciones, el análisis gráfico puede ser omitido.

A continuación consideraremos dos pruebas, cada una diseñada para detectar diferentes "tipos" de heteroscedasticidad. En todos los casos la hipótesis nula de estas pruebas es la presencia de homoscedasticidad $H_0 : \sigma_i^2 = \sigma^2; \forall i$ versus la hipótesis alterna de que existe algún tipo de heteroscedasticidad.

8.2.1. Prueba de Breusch-Pagan

Esta prueba, permite la posibilidad de determinar si más de una variable causa el problema de heteroscedasticidad.

breush1 diseñaron una prueba que permite detectar si existe una relación entre la varianza del error y un grupo de variables (recogidas en un vector Z). El tipo de relación de esta prueba no está suscrita a algún tipo de relación funcional. En este caso, la prueba está diseñada para detectar la heteroscedasticidad de la forma $\sigma_i^2 = f(\gamma + \delta Z_i)$. Donde $Z_{i(g \times 1)}$ es un grupo de g variables que afectan a la varianza que se organizan en forma vectorial y $\delta_{(1 \times g)}$ corresponde a un vector de constantes.

Por ejemplo, supongamos que el investigador cree que el problema de heteroscedasticidad está siendo causado por las variables W_i y V_i . En ese caso tendremos que $Z_{i(2 \times 1)} = [W_i, V_i]^T$ y $\delta_{(1 \times 2)} = [\delta_1, \delta_2]$. Por tanto, $\sigma_i^2 = f(\gamma + \delta_1 W_i + \delta_2 V_i)$.

Recapitulando, la prueba de Breusch-Pagan implica las siguientes hipótesis nula y alterna:

$$\begin{aligned} H_0 : \sigma_i^2 &= \sigma^2; \forall i \\ H_A : \sigma_i^2 &= f(\gamma + \delta Z_i) \end{aligned}$$

Los pasos para efectuar esta prueba son los siguientes:

1. Corra el modelo original $Y = X\beta + \varepsilon$ y encuentre la serie de los residuos $\hat{\varepsilon}$.
2. Calcule³ $\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n}$.
3. Posteriormente estime la siguiente regresión auxiliar: $\frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2} = \gamma + \delta Z_i + \mu_i$.
4. Calcule la suma de los cuadrados de la regresión auxiliar ($SSR = SST - SSE$).
5. Calcule el estadístico $BP = \frac{SSR}{2}$.

breush1 demostraron que el estadístico BP sigue una distribución Chi-cuadrado con g grados de libertad (χ_g^2), bajo el supuesto de que los errores se distribuyen normalmente. Por tanto, se rechazará la hipótesis nula de homoscedasticidad a favor de una heteroscedasticidad de la forma $\sigma_i^2 = f(\gamma + \delta Z_i)$ si el estadístico BP es mayor que $\chi_{g, \alpha}^2$, con un nivel de confianza del $(1 - \alpha)\%$.

Sin embargo, se ha documentado mucho que esta prueba no funciona bien cuando el supuesto de normalidad no se cumple (Ver por ejemplo **koenker1981note**).

³ Noten que esto corresponde al estimador de la varianza del error del Método de Máxima Verosimilitud.

koenker1981note propone una modificación de la prueba BP que implica transformación de los residuos. Esta modificación se conoce como la versión studentizada de la prueba BP.

8.2.2. Prueba de White

white1 desarrolló una prueba más general que las anteriores, con la ventaja de no requerir que ordenemos los datos en diferentes grupos, ni tampoco depende de que los errores se distribuyan normalmente. Esta prueba implica las siguientes hipótesis nula y alterna:

$$\begin{aligned} H_0 : \sigma_i^2 &= \sigma^2; \forall i \\ H_A : &No H_0 \end{aligned} \quad (8.1)$$

Los pasos para efectuar esta prueba son los siguientes:

1. Corra el modelo original $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ y encuentre la serie de los residuos $\hat{\varepsilon}$.
2. Ahora corra la siguiente regresión auxiliar: $\hat{\varepsilon}_i^2 = \gamma + \sum_{m=1}^k \sum_{j=1}^k \delta_{mj} X_{mi} X_{ji} + \mu_i$. Por ejemplo, si el modelo original es $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$, entonces el modelo para la regresión auxiliar sería:⁴ $\hat{\varepsilon}_i^2 = \gamma + \delta_1 X_{2i} + \delta_2 X_{3i} + \delta_3 X_{2i}^2 + \delta_4 X_{3i}^2 + \delta_5 X_{2i} X_{3i} + \vartheta_i$.
3. Calcule el R^2 de la regresión auxiliar.
4. Calcule el estadístico de White $W_a = n \times R^2$.

white1 demostró que su estadístico W_a sigue una distribución Chi-cuadrado con un número de grados de libertad igual al número de regresores que se emplean en la regresión auxiliar (g). Por tanto se rechazará la hipótesis nula de no heteroscedasticidad con un nivel de confianza de $(1 - \alpha)\%$, cuando el estadístico de esta prueba sea mayor que $\chi_{g,\alpha}^2$.

8.3. Solución a la heteroscedasticidad

En la sección pasada se discutió cómo detectar la presencia de heteroscedasticidad; pero, ¿qué hacer si ésta está presente en un modelo de regresión? A continuación, existen dos soluciones que se emplean comúnmente en la estadística y econometría tradicional para solucionar el problema. La primera solución es tratar de resolver de raíz el problema modificando la muestra. Este método se conoce como el método de Mínimos Cuadrados Ponderados (MCP) que hace parte de la

⁴ El último término se conoce con el nombre de término cruzado. En algunas oportunidades cuando el modelo original cuenta con muchas variables explicatorias y/o el número de observaciones no es mucho, puede ocurrir que no existan los grados de libertad necesarios para correr la regresión auxiliar incluyendo los términos cruzados. En esos casos, algunos autores acostumbran correr la regresión auxiliar sin los términos cruzados, si bien esto le resta poder a la prueba.

familia de los Mínimos Cuadrados Generalizados (MCG). Esta aproximación implica conocer exactamente cómo es la heteroscedasticidad; algo que típicamente es difícil para el científico de datos. Esta aproximación será discutida en un Capítulo más adelante.

La segunda opción implica solucionar los síntomas de la heteroscedasticidad, tratando de estimar de manera consistente la matriz de varianzas y covarianzas de los coeficientes estimados. Esto permite corregir los errores estándar de los coeficientes, y de esta manera los t calculados serán recalculados y por tanto los valores p son diferentes.

8.3.1. *Estimación Consistente en presencia de heteroscedasticidad de los errores estándar.*

Es muy probable que en la práctica no podamos encontrar la forma exacta de la heteroscedasticidad y por tanto no podremos “corregir” la muestra, de tal manera que la heteroscedasticidad desaparezca, tal como lo hace el método de MCP. Para dar solución a esta dificultad, **white1** ideó una forma de corregir el estimador y no la muestra.

En especial, **white1** mostró que es posible aún encontrar un estimador apropiado para la matriz de varianzas y covarianzas de los β 's obtenidos por MCO.⁵ Recordemos que en presencia de heteroscedasticidad, la varianza de los coeficientes tiene la siguiente estructura:

$$\text{Var} [\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Omega \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

white1 sugiere usar el siguiente estimador consistente para la matriz de varianzas y covarianzas de los β 's obtenidos por MCO

$$\text{Est.Var} [\hat{\beta}] = n(\mathbf{X}^T \mathbf{X})^{-1} S_0 (\mathbf{X}^T \mathbf{X})^{-1}$$

donde:

$$S_0 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i^T \quad x_i^T = (1 \ x_{1i} \ x_{2i} \ \dots \ x_{ki})$$

Esa matriz de varianzas y covarianzas deberá ser empleada para calcular los estadísticos de las pruebas de hipótesis tanto individuales como conjuntas. De esta manera, el problema del sesgo en la matriz de varianzas y covarianzas es solucionado. De hecho, **white1** demostró que ese estimador es consistente; es decir, i.e. no funciona bien en muestras pequeñas. Por esta razón este estimador (y otros similares) es conocido como estimador H.C(heteroskedasticity consistent).

⁵ Recuerden que el estimador MCO del vector β sigue siendo insesgado, el problema se presenta en el estimador de la matriz de varianzas y covarianzas.

Es importante mencionar, que esto no hace que al emplear esta solución se obtengan un estimador MELI. La varianza sigue siendo más grande que por ejemplo lode GLS que se discutirán en capítulo posterior.

Por otro lado, **DavidsonRussellandMacKinnon1993** y Cribari-Neto (2004) propusieron modificaciones a la propuesta de White (1980). Es decir, existen varios estimadores H.C. disponibles

El estimador de White para la matriz de varianzas y covarianzas de $\hat{\beta}$ se tiene

$$\frac{1}{n} \left(\frac{\mathbf{X}^T \mathbf{X}}{n} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \omega_i \mathbf{x}_i \mathbf{x}_i^T \right) \left(\frac{\mathbf{X}^T \mathbf{X}}{n} \right)^{-1}$$

La diferencia entre los métodos H.C. propuestos están en ω_i (ponderación de los datos). Por ejemplo,

- White (1980) propone $\omega_i = e_i^2$ (A este método se le conoce como HC0).
- Davidson y MacKinnon (1993) proponen : $\omega_i = \frac{n}{n-k} e_i^2$ (HC1)
- Davidson y MacKinnon (1993) también proponen otro estimador: $\omega_i = \frac{e_i^2}{1-p_{ii}}$ (HC2)

donde p_{ii} es el elemento i de la matriz de proyecciones conocida como **P** gorro.

- Davidson y MacKinnon (1993) sugieren un tercer estimador: $\omega_i = \frac{e_i^2}{(1-p_{ii})^2}$ (HC3)
- Cribari-Neto (2004) proponen: $\omega_i = \frac{e_i^2}{(1-p_{ii})^{\delta_i}}$ (HC4)

donde $\delta_i = \min \{4, p_{ii} / \bar{p}\}$ y $\bar{p} = \frac{\sum_{i=1}^n p_{ii}}{n}$

Davidson y MacKinnon (1993) sugerían nunca usar White (1980) (HC0) pues podemos encontrar un mejor estimador. En esa misma dirección, Long y Ervin (2000) encontraron con simulaciones de Monte Carlo que HC3 se comporta mejor en muestras pequeñas y grandes. Cribari-Neto (2004) mostró que HC4 se comporta mejor en muestras pequeñas, especialmente si hay observaciones influyentes.

Por otro lado, también es importante reconocer que al emplear un estimador H.C. también debemos modificar nuestras pruebas conjuntas empleando la correspondiente matriz H.C. (ya se había mencionado que las pruebas individuales se debían modificar.)

Por ejemplo, la prueba de Wald para una restricción de la forma $\mathbf{R}\beta = \mathbf{q}$ será:

$$W = \left(\mathbf{R}\hat{\beta} \right)^T \left[\mathbf{R} \left(\text{Est.Asy.Var} \left[\hat{\beta} \right] \right) \mathbf{R}^T \right]^{-1} \left(\mathbf{R}\hat{\beta} \right)$$

8.4. Práctica en R: Análisis del efecto discriminatorio de género en las diferencias salariales en Colombia

El objetivo de este ejercicio es aplicar las diferentes pruebas de heteroscedasticidad y de ser el caso aplicar una solución. Continuaremos con el análisis del efecto discriminatorio de género en las diferencias salariales en Colombia. En el anterior ya habíamos realizado una aproximación teórica a este modelo y aplicado las distintas pruebas de multicolinealidad. El modelo a estimar es el siguiente:⁶

$$\ln(ih_i) = \beta_0 + \beta_1 yedu_i + \beta_2 exp_i + \beta_3 exp_i^2 + \beta_4 D_i + \beta_5 Dyedu_i + \beta_6 Dexp_i + \beta_7 Dexp_i^2 + \varepsilon_i$$

donde

$$D_i = \begin{cases} 1 & \text{si el individuo } i \text{ es hombre} \\ 0 & \text{o.w.} \end{cases}$$

$\ln(ih_i)$ representa el logaritmo natural del ingreso por hora del individuo i , $yedu_i$ y exp_i denotan los años de educación y de experiencia del individuo i . Los datos se encuentran en el archivo *DatosMultiColinealidad.csv*.

8.4.1. Análisis gráfico de los residuos

Como se mencionó anteriormente, una práctica común para detectar intuitivamente problemas de heteroscedasticidad en el término de error es emplear gráficas de dispersión de los errores estimados ($\hat{\varepsilon}$) y de las variables independientes, al igual que los errores estimados y el valor estimado de la variable dependiente. Algunos autores también sugieren emplear gráficos de dispersión del cuadrado de los residuos ($\hat{\varepsilon}^2$) y las variables explicativas. Estos gráficos son empleados para determinar la existencia de algún patrón en la variabilidad del error. En este caso sólo graficaremos los errores contra las variables explicatorias $yedu_i$, exp_i y exp_i^2 . (Pero el lector debería efectuar todos los otros gráficos)

Cargue los datos, si el nombre de las variables no queda bien leído, cámbielos y estime el modelo. Posteriormente, extraigamos los residuales del objeto *LM* con la función *resid()* y guardémoslo con el objeto *e*.

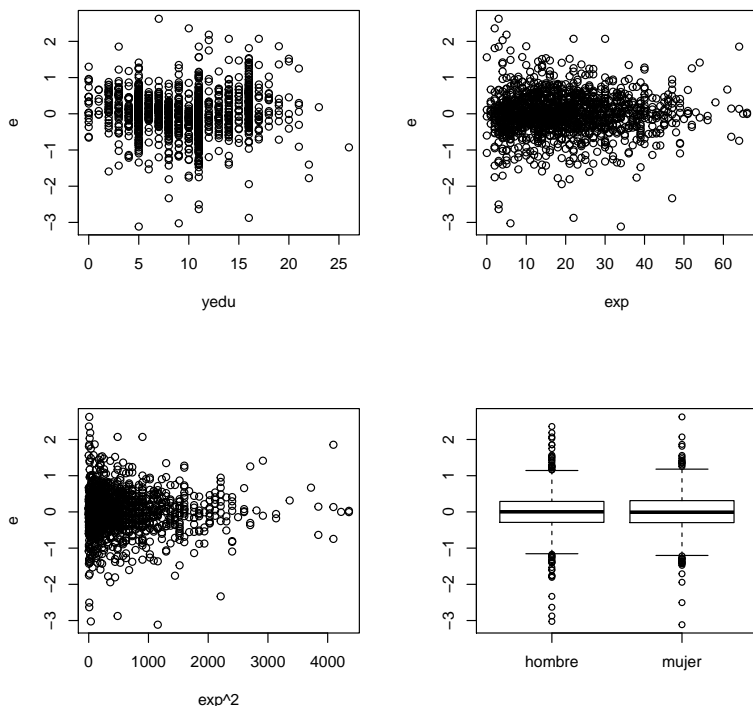
```
data <- read.csv("Data/DatosMultiColinealidad.csv", sep=",")
names(data)[1] <- "sexo"
res1 <- lm(lnih ~ yedu + exp + I(exp^2) + sexo
          + sexo * exp + sexo * I(exp^2), data)
e <- resid(res1)
```

El correspondiente modelo estimado se reportó en el capítulo anterior. Pero noten que esos resultados no son interesantes hasta que estemos seguros que no existe

⁶ Si desea información teórica acerca de este modelo, remítase al capítulo anterior.

un problema de heteroscedasticidad. Ahora grafiquemos los residuos versus las variables explicativas del modelo.

```
attach(data)
par(mfrow=c(2,2))
plot(yedu, e)
plot(exp, e)
plot(exp^2, e)
plot(sexo, e)
```



En el gráfico no es posible apreciar una relación clara entre la variabilidad de los errores y los años de educación al igual que con el *sexo*. Con los años de experiencia no es muy claro, pero parece existir menos dispersión a medida que los años de experiencia aumentan. Pero este fenómeno se hace evidente cuando se grafican los residuos versus los años de experiencia al cuadrado. Se observa claramente que la dispersión de los residuos disminuye a medida que se incrementa el cuadrado de la experiencia. Este es un síntoma de heteroscedasticidad.

8.4.2. Pruebas de heteroscedasticidad

8.4.2.1. Prueba de Breusch-Pagan

Hay muchas funciones que permiten calcular esta prueba en R. Por ejemplo, existe la función `bptest()` del paquete `lmtest` que nos da la opción de calcular la versión studentizada de la prueba propuesta por Koenker (1981) para el caso en que los residuos no sigan una distribución normal. Otra función con unas opciones muy útiles es `ols_test.breusch.pagan()` del paquete `olsrr`. Esta última función no permite calcular la versión studentizada de la prueba pero permite hacer simultáneamente muchas pruebas de Breusch-Pagan y corrige los valores p para tener en cuenta las múltiples pruebas que se pueden realizar al mismo tiempo. Las correcciones que incluye esta función son las de Bonferroni, Sidak y Holm.

Empecemos empleando la función `bptest()` del paquete `lmtest`. Esta función tiene tres argumentos importantes, uno de ellos indispensable. El primer argumento es un objeto *LM*, el segundo permite determinar si se empleará la versión studentizada de la prueba o no (`studentize =`). El valor por defecto es *TRUE*, es decir que si no se especifica este argumento se calculará la versión studentizada. Si no se especifican más argumentos, la función realizará la prueba con la hipótesis alterna de que todas las variables causan el problema. Por ejemplo

```
library(lmtest)
bptest(res1, studentize = FALSE)

##
## Breusch-Pagan test
##
## data: res1
## BP = 49.546, df = 6, p-value = 5.798e-09
```

En este caso se puede rechazar la nula de homoscedasticidad en favor de la alterna que la varianza es función de todas las variables. En otras palabras, existe un problema de heteroscedasticidad. Otro argumento que se puede emplear en esta función es una fórmula que incluya las variables que se crean están provocando el problema de heteroscedasticidad. Por ejemplo,

```
bptest(res1, ~ exp, studentize = FALSE)

##
## Breusch-Pagan test
##
## data: res1
## BP = 2.3968, df = 1, p-value = 0.1216
```

En este caso la hipótesis nula de homoscedasticidad no es rechazada y por tanto se concluiría que existe homoscedasticidad (por lo menos por lo menos no existe heteroscedasticidad causada únicamente por la variable *exp*), contrario a como lo intuimos de los gráficos. Noten que esto llama la atención a la necesidad de hacer múltiples pruebas con hipótesis alternas diferentes (por lo menos para cada variable y el total de estas).

La función `ols_test.breusch.pagan()` del paquete `olsrr`, permite probar al mismo tiempo si cada una de las variables está causando el problema de heteroscedasticidad o todas al mismo tiempo. Esta función requiere de cuatro argumentos. El objeto `lm` al que se le realizará la prueba. Un argumento denominado `rhs` (right hand side) que si es igual a `TRUE` implica que se emplearán todas las variables explicativas del modelo para hacer las pruebas; si `rhs = FALSE` entonces la hipótesis alterna será que la varianza es función de los valores estimados de la variable dependiente (esto comúnmente no tiene mucho sentido). El tercer argumento permite realizar todas las pruebas en las que la varianza depende de cada una de las variables explicativas y de todas ella (`multiple = TRUE`). Finalmente, está el argumento (`p.adj`) que permite especificar qué tipo de corrección estadística aplicarle al valor p para tener en cuenta que se están realizando multiples comparaciones al tiempo. Por ejemplo, el siguiente código realiza pruebas de Breusch-Paga para cada una de las variables explicativas y para todas al tiempo, y corrige los respectivos valores p con el método de Bonferroni.

```
library(olsrr)
ols_test_breusch_pagan(res1, rhs = TRUE, multiple = TRUE,
                        p.adj = 'bonferroni')

##
## Breusch Pagan Test for Heteroskedasticity
## -----
## Ho: the variance is constant
## Ha: the variance is not constant
##
##                               Data
## -----
## Response : Lnhi
## Variables: yedu exp I(exp^2) sexomujer exp:sexomujer I(exp^2):sexomujer
##
## Test Summary (Bonferroni p values)
## -----
## Variable                chi2      df      p
## -----
## yedu                    23.13058768    1  9.081815e-06
## exp                     2.39681316    1  7.294963e-01
## I(exp^2)                 0.06323815    1  1.000000e+00
## sexomujer               3.46086767    1  3.770240e-01
## exp:sexomujer           3.96265681    1  2.791214e-01
## I(exp^2):sexomujer      1.73988308    1  1.000000e+00
## -----
## simultaneous           49.54557532    6  5.797772e-09
## -----
```

Los resultados muestran que con un 99 % de confianza podemos rechazar la nula de homoscedastidad para todos los casos. es importante anotar que el estadísticos no son iguales entre la función `bptest()` y `ols_test.breusch.pagan()` dado que esta última no emplea como estadístico de prueba $BP = \frac{SSR}{2}$ sino nR^2 . En todo caso los resultados son equivalentes.

Antes de concluir que existe un problema de heteroscedasticidad, es importante estar seguros que el supuesto de normalidad de los errores que requiere esta prueba se cumple. Existen varias pruebas de normalidad como se discute en Alonso y Montenegro (2015), todas tienen la característica de que la hipótesis nula es la normalidad y la alterna es la no normalidad. Una forma rápida de realizar varias pruebas de normalidad es emplear la función `ols_test_normality()` del paquete `olsrr`.

```
ols_test_normality(res1)

## -----
##          Test              Statistic      pvalue
## -----
## Shapiro-Wilk              0.9583         0.0000
## Kolmogorov-Smirnov         0.0636         0.0000
## Cramer-von Mises          174.9537         0.0273
## Anderson-Darling          12.4223         0.0000
## -----
```

En todos los casos las pruebas de normalidad permiten concluir que los residuos no siguen una distribución normal. Por eso no son confiables de los resultados de la prueba de Breusch-Pagan tradicional. Deberíamos entonces emplear la versión studentizada de la prueba propuesta por Koenker (1981) para el caso en que los residuos no sigan una distribución normal. Es decir,

```
bptest(res1, studentize = TRUE)

##
## studentized Breusch-Pagan test
##
## data: res1
## BP = 18.412, df = 6, p-value = 0.00528
```

Por lo tanto los resultados de esta prueba indican que existe un problema de heteroscedasticidad.

8.4.2.2. Prueba de White

Por último, realicemos la prueba de White para contrastar la hipótesis nula de no heteroscedasticidad versus la alterna de heteroscedasticidad. Para llevar a cabo esta prueba se puede emplear nuevamente la función `bptest()`. Waldman (1983) mostró que si las variables en la hipótesis alterna son las mismas que las usadas en la prueba de White, entonces esta prueba es algebraicamente igual a la versión studentizada de Breusch-Pagan (con todas las variables de la regresión auxiliar de white como causantes de la heteroscedasticidad). Es decir, se deben emplear todas las variables explicativas en el modelo, sus cuadrados y los productos cruzados de estos.

```
bptest(res1, ~ yedu + exp + I(exp^2) + sexo + sexo*yedu +
  sexo*exp + sexo * I(exp^2) + I(yedu^2) + I(exp^4) +
  I(yedu*exp) + I(yedu*exp^2) + sexo * I(yedu*exp) +
```

```

sexo *I(exp^3) + I(exp^3)

##
## studentized Breusch-Pagan test
##
## data:  res1
## BP = 37.863, df = 14, p-value = 0.0005452

```

En esta ocasión no se emplea la variable dummy al cuadrado pues sería igual a la variable sin elevar al cuadrado y esto generaría un problema de multicolinealidad perfecta en las regresiones auxiliares de la prueba. El resultado implica que se puede rechazar la nula de homoscedasticidad en favor de la alterna de que existe algún tipo de heteroscedasticidad. Por tanto podemos concluir que existe un problema de heteroscedasticidad al unir los resultados de las dos pruebas.

8.4.3. Solución al problema de heteroscedasticidad con HC

Como se discutió anteriormente, una forma de solucionar el problema es empleando estimadores consistentes en presencia de heteroscedasticidad para la (H.C.) para la matriz de varianzas y covarianzas. Esto se puede hacer empleando el paquete *sandwich* y la función *vcovHC*. Esta función requiere dos argumentos: el objeto *lm* al que se le quiere corregir la matriz de varianzas y covarianzas y el tipo de corrección que por defecto es la que denominamos HC3 (la propuesta por Davidson y MacKinnon (1993) y sugerida por Cribari-Neto (2004)).

```

library(sandwich)
vcovHC(res1) # HC3 Davidson y MacKinnon (1993)

##
## (Intercept) yedu
## (Intercept) 8.317882e-03 -2.271974e-04
## yedu -2.271974e-04 2.059261e-05
## exp -4.947456e-04 1.825652e-06
## I(exp^2) 7.713180e-06 1.795725e-08
## sexomujer -5.423997e-03 -3.514125e-05
## exp:sexomujer 4.554557e-04 1.742774e-06
## I(exp^2):sexomujer -7.815056e-06 -8.851575e-09
## exp I(exp^2)
## (Intercept) -4.947456e-04 7.713180e-06
## yedu 1.825652e-06 1.795725e-08
## exp 4.626380e-05 -8.473137e-07
## I(exp^2) -8.473137e-07 1.718700e-08
## sexomujer 4.723621e-04 -7.955854e-06
## exp:sexomujer -4.609217e-05 8.527277e-07
## I(exp^2):sexomujer 8.506639e-07 -1.721951e-08
## sexomujer exp:sexomujer
## (Intercept) -5.423997e-03 4.554557e-04
## yedu -3.514125e-05 1.742774e-06
## exp 4.723621e-04 -4.609217e-05
## I(exp^2) -7.955854e-06 8.527277e-07

```

```
## sexomujer          9.727628e-03 -8.059566e-04
## exp:sexomujer      -8.059566e-04  8.079147e-05
## I(exp^2):sexomujer 1.354197e-05 -1.502262e-06
##                    I(exp^2):sexomujer
## (Intercept)        -7.815056e-06
## yedu               -8.851575e-09
## exp                8.506639e-07
## I(exp^2)           -1.721951e-08
## sexomujer          1.354197e-05
## exp:sexomujer      -1.502262e-06
## I(exp^2):sexomujer 3.066016e-08
```

Ahora, como no es muy útil la matriz de varianzas y covarianzas sola, sino mas bien los respectivos t individuales y sus correspondientes valores p, podemos emplear la función `coefTest()` para realizar las pruebas individuales. Por ejemplo:

```
coefTest(res1, vcov = (vcovHC(res1)))

##
## t test of coefficients:
##
##              Estimate Std. Error t value
## (Intercept)  5.8557e+00 9.1202e-02 64.2056
## yedu         1.3064e-01 4.5379e-03 28.7882
## exp         3.3948e-02 6.8017e-03 4.9911
## I(exp^2)     -2.9550e-04 1.3110e-04 -2.2540
## sexomujer    -1.1666e-01 9.8629e-02 -1.1828
## exp:sexomujer -3.9387e-03 8.9884e-03 -0.4382
## I(exp^2):sexomujer 5.3015e-05 1.7510e-04 0.3028
##
##              Pr(>|t|)
## (Intercept) < 2.2e-16 ***
## yedu        < 2.2e-16 ***
## exp         6.754e-07 ***
## I(exp^2)     0.02435 *
## sexomujer    0.23707
## exp:sexomujer 0.66131
## I(exp^2):sexomujer 0.76211
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# HC3 Davidson y MacKinnon (1993)
```

La función `coefTest()` necesita dos argumentos: el modelo y la matriz de varianzas y covarianzas que se quiera emplear. Si no se especifica una matriz de varianzas y covarianzas, entonces se empleara la de los MCO. Para este caso podemos ver como individualmente, los coeficientes de las variables asociadas a la dummy de sexo no son significativos individualmente.

A manera de ejemplo se muestra las otras posibles correcciones.

```
coefTest(res1, vcov = (vcovHC(res1, "HC0")))# White (1980)
##
```

```
## t test of coefficients:
##
##              Estimate Std. Error t value
## (Intercept)    5.8557e+00  9.0230e-02  64.8974
## yedu           1.3064e-01  4.5104e-03  28.9636
## exp            3.3948e-02  6.6308e-03   5.1197
## I(exp^2)       -2.9550e-04  1.2605e-04 -2.3443
## sexomujer      -1.1666e-01  9.7408e-02 -1.1977
## exp:sexomujer  -3.9387e-03  8.7934e-03 -0.4479
## I(exp^2):sexomujer  5.3015e-05  1.6940e-04  0.3130
##              Pr(>|t|)
## (Intercept)    < 2.2e-16 ***
## yedu           < 2.2e-16 ***
## exp            3.483e-07 ***
## I(exp^2)       0.0192 *
## sexomujer      0.2312
## exp:sexomujer  0.6543
## I(exp^2):sexomujer  0.7544
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

coeftest(res1, vcov = (vcovHC(res1, "HC1")) # Davidson y MacKinnon (1993)

##
## t test of coefficients:
##
##              Estimate Std. Error t value
## (Intercept)    5.8557e+00  9.0454e-02  64.7367
## yedu           1.3064e-01  4.5216e-03  28.8919
## exp            3.3948e-02  6.6473e-03   5.1070
## I(exp^2)       -2.9550e-04  1.2637e-04 -2.3385
## sexomujer      -1.1666e-01  9.7650e-02 -1.1947
## exp:sexomujer  -3.9387e-03  8.8152e-03 -0.4468
## I(exp^2):sexomujer  5.3015e-05  1.6982e-04  0.3122
##              Pr(>|t|)
## (Intercept)    < 2.2e-16 ***
## yedu           < 2.2e-16 ***
## exp            3.721e-07 ***
## I(exp^2)       0.0195 *
## sexomujer      0.2324
## exp:sexomujer  0.6551
## I(exp^2):sexomujer  0.7549
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

coeftest(res1, vcov = (vcovHC(res1, "HC2")) # Davidson y MacKinnon (1993)

##
## t test of coefficients:
##
##              Estimate Std. Error t value
## (Intercept)    5.8557e+00  9.0708e-02  64.5556
## yedu           1.3064e-01  4.5241e-03  28.8759
```

```
## exp          3.3948e-02  6.7141e-03  5.0562
## I(exp^2)     -2.9550e-04  1.2851e-04 -2.2994
## sexomujer    -1.1666e-01  9.8006e-02 -1.1904
## exp:sexomujer -3.9387e-03  8.8884e-03 -0.4431
## I(exp^2):sexomujer 5.3015e-05  1.7217e-04  0.3079
##
## Pr(>|t|)
## (Intercept) < 2.2e-16 ***
## yedu        < 2.2e-16 ***
## exp         4.839e-07 ***
## I(exp^2)     0.02163 *
## sexomujer    0.23411
## exp:sexomujer 0.65774
## I(exp^2):sexomujer 0.75819
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

coeftest(resl, vcov = (vcovHC(resl, "HC4")) # Cribari-Neto (2004)

##
## t test of coefficients:
##
##              Estimate Std. Error t value
## (Intercept)  5.8557e+00 9.1595e-02 63.9306
## yedu         1.3064e-01 4.5321e-03 28.8250
## exp          3.3948e-02 6.9554e-03  4.8808
## I(exp^2)     -2.9550e-04 1.3620e-04 -2.1697
## sexomujer    -1.1666e-01 9.9245e-02 -1.1755
## exp:sexomujer -3.9387e-03 9.1539e-03 -0.4303
## I(exp^2):sexomujer 5.3015e-05 1.8070e-04  0.2934
##
## Pr(>|t|)
## (Intercept) < 2.2e-16 ***
## yedu        < 2.2e-16 ***
## exp         1.177e-06 ***
## I(exp^2)     0.0302 *
## sexomujer    0.2400
## exp:sexomujer 0.6671
## I(exp^2):sexomujer 0.7693
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso se puede observar que los resultados son robustos a la aproximación que se emplee. En todo caso parece ser mejor emplear el HC3 de acuerdo a las simulaciones de Montecarlo realizadas por Long y Ervin (2000). Ahora miremos si conjuntamente todos los coeficientes que acompañan a las dummy son cero o no con la corrección de heteroscedasticidad. Eso lo podemos hacer de manera muy fácil dado que todas las funciones estudiadas previamente soportan la inclusión de una matriz de varianzas y covarianzas H.C. Por ejemplo,

```
library(AER)
res2 <- lm(Lnih ~ yedu + exp + I(exp^2) , data)
```

```
waldtest(res2, res1, vcov = vcovHC( res1))

## Wald test
##
## Model 1: Lnih ~ yedu + exp + I(exp^2)
## Model 2: Lnih ~ yedu + exp + I(exp^2) + sexo + sexo * exp + sexo * I(exp^2)
##      Res.Df Df       F      Pr(>F)
## 1      1411
## 2      1408   3 10.149 1.293e-06 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados muestran qué se puede rechazar la hipótesis nula de que el modelo restringido es mejor que no restringido. Es decir, al menos uno de los coeficientes asociados a la dummy de sexo es significativo. Es decir, si existe diferencia en como se trata a los hombres y a las mujeres.

8.5. Apéndice

Apéndice 8.1 Demostración de la insesgadez de los estimadores en presencia de heteroscedasticidad

En presencia de heteroscedasticidad los estimadores MCO siguen siendo insesgados. Esta afirmación se puede demostrar fácilmente. Consideremos un modelo lineal con un término de error heteroscedástico y no autocorrelación. Es decir,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Donde $E[\boldsymbol{\varepsilon}_i] = 0$, $Var[\boldsymbol{\varepsilon}_i] = \sigma_{\varepsilon}^2$ y $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j] = 0$, $\forall i \neq j$. Ahora determinemos si $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ sigue siendo insesgado o no. Así,

$$E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE[\mathbf{y}]$$

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE[\boldsymbol{\varepsilon}]$$

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{I} \bullet \boldsymbol{\beta}$$

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

Apéndice 8.2 Sesgo de la matriz de varianzas y covarianzas en presencia de heteroscedasticidad

En presencia de heteroscedasticidad el estimador de la matriz de varianzas y covarianzas de MCO ($\widehat{Var}[\hat{\boldsymbol{\beta}}] = s^2(\mathbf{X}^T\mathbf{X})$) es sesgado. Es más, el estimador MCO

para los coeficientes ($\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$) no es eficiente; es decir no tiene la mínima varianza posible. Esta afirmación se puede demostrar fácilmente.

Continuando con el modelo considerado en el Apéndice anterior, en este caso tenemos que:

$$\text{Var}[\varepsilon] = E[\varepsilon^T \varepsilon] = \Omega = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Ahora podemos calcular la varianza de los estimadores MCO. Es decir,

$$\text{Var}[\hat{\beta}] = \text{Var}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}]$$

Por tanto tendremos que

$$\text{Var}[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var}[\mathbf{y}] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

$$\text{Var}[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var}[\varepsilon] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

$$\text{Var}[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Omega \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

Por tanto la varianza no es la mínima posible. Y por otro lado, al emplear el estimador MCO para la matriz de varianzas y covarianzas de los betas ($\widehat{\text{Var}}[\hat{\beta}] = s^2 (\mathbf{X}^T \mathbf{X})^{-1}$) en presencia de heteroscedasticidad se obtendrá un estimador cuyo valor esperado no es igual a la varianza real; es decir, será insesgado.

Capítulo 9

Autocorrelación

Objetivos del capítulo

Al finalizar este capítulo, el lector estará en capacidad de:

- Realizar en R diferentes tipos de análisis gráficos que revelen la posibilidad de autocorrelación en los residuos.
- Efectuar en R las pruebas estadísticas necesarias para detectar la violación del supuesto de no autocorrelación en los residuos. En especial las pruebas de Rachas, Durbin Watson, Box-Pierce y de Ljung-Box.
- Corregir el problema de autocorrelación empleando estimadores consistentes para los errores estándar en R.

9.1. Introducción

En los dos capítulos anteriores hemos analizado las consecuencias de violar algunos de los supuestos del modelo de regresión. Analizamos las consecuencias de la violación del supuesto de independencia lineal entre variables explicativas en un modelo de regresión múltiple, posteriormente nos enfocamos en las consecuencias de un error homoscedástico.

Finalmente, y para concluir nuestra discusión de la violación de los supuestos que garantizan el cumplimiento del Teorema de Gauss-Markov (Ver recuadro 9.1) nos concentraremos en los efectos que tiene la violación del supuesto de no autocorrelación (no existe relación entre los diferentes errores). Este supuesto garantiza la no presencia de un patrón predecible en el comportamiento de los errores. Cuando este supuesto es violado, lo cual ocurre comúnmente cuando trabajamos con datos de series de tiempo, se dice que los errores presentan autocorrelación (o correlación serial); en otras palabras, están relacionados entre sí.

Recuadro 9.1 Teorema de Gauss-Markov

Si se considera un modelo lineal $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$ y se supone que:

- Las X_2, X_3, \dots, X_k son fijas y linealmente independientes (es decir X tiene rango completo y es una matriz no estocástica).
- El vector de errores $\boldsymbol{\varepsilon}$ tiene media cero, varianza constante y no autocorrelación. Es decir: $E[\boldsymbol{\varepsilon}] = 0$ y $Var[\boldsymbol{\varepsilon}] = \sigma^2 I_n$

Entonces el estimador de MCO $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ es el *Mejor Estimador Lineal Insesgado (MELI)*

Si existe autocorrelación entre los errores, entonces los estimadores MCO siguen siendo insesgados pero no son eficientes (ver 1 y 2 para la demostración de estos resultados).

Veamos más en detalle qué significa la autocorrelación. Y para simplificar, estudiemos inicialmente el caso más sencillo. Cuando existe una relación lineal “grande” entre las observaciones adyacentes, pero esta relación (lineal) tiende a desaparecer a medida que se consideran errores más lejanos. Formalmente tenemos:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (9.1)$$

donde el término del error tiene media cero y se encuentra correlacionado con el error del periodo anterior: $E[\varepsilon_t] = 0$; $\varepsilon_t = \rho \varepsilon_{t-1} + v_t \quad \forall t$ con $0 \leq |\rho| < 1$ y $Var(v_t) = \sigma_v^2$.

Este tipo de autocorrelación es conocido como un proceso auto-regresivo de orden uno o AR(1) para abreviar. Si seguimos asumiendo que la varianza de los errores es constante, entonces se puede probar fácilmente (hágalo) que:

$$\sigma_{\varepsilon}^2 = \frac{\sigma_v^2}{(1 - \rho^2)}$$

Por otro lado, dado que el valor esperado del error es cero se puede mostrar fácilmente (hágalo):

$$\text{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \rho \sigma_{\varepsilon}^2$$

Entonces, la correlación entre los errores adyacentes¹ será:

$$\frac{\text{Cov}(\varepsilon_t, \varepsilon_{t-1})}{\sqrt{\text{Var}(\varepsilon_t) \text{Var}(\varepsilon_{t-1})}} = \frac{\text{Cov}(\varepsilon_t, \varepsilon_{t-1})}{\sigma_{\varepsilon}^2} = \rho$$

Similarmente es relativamente sencillo demostrar que:

$$\text{Cov}(\varepsilon_t, \varepsilon_{t-2}) = \rho^2 \sigma_{\varepsilon}^2$$

$$\text{Cov}(\varepsilon_t, \varepsilon_{t-3}) = \rho^3 \sigma_{\varepsilon}^2$$

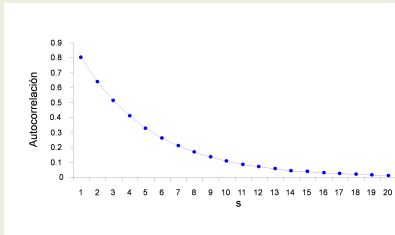
En términos generales tenemos que en este caso de errores con un proceso AR(1) la autocorrelación para diferentes rezagos está dada por:²

$$\rho(s) = \rho^s \sigma_{\varepsilon}^2$$

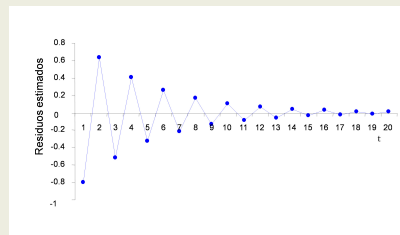
Es decir, en el caso de un proceso AR(1) a medida que se alejan en el tiempo las observaciones (se consideran más rezagos) la correlación entre los errores es menor. Esto lo podemos observar en el ejemplo 9.1. Si no existiese un problema de autocorrelación, la autocorrelación para los diferentes rezagos será cero.

Ejemplo 9.1 Autocorrelación de los errores con un proceso AR(1)

$\rho=0.8$



$\rho=-0.8$



¹ La correlación entre errores inmediatamente adyacentes, es decir separados por únicamente un periodo, también se denomina autocorrelación a un rezago. Si se considera la relación entre errores separados por dos periodos se denomina autocorrelación para a dos rezagos, etc.

² A la función que muestra correlación para diferentes rezagos de un proceso se le denomina función de autocorrelación.

Regresando al problema de autocorrelación, su origen puede ser causado porque las relaciones entre las variables pueden ser dinámicas. Es decir, todo el ajuste entre las variables no se hace en un mismo período; un ejemplo de este comportamiento son las expectativas adaptativas. La autocorrelación también se puede deber a que la información no está disponible instantáneamente y por tanto, la variable dependiente puede depender de errores previos. Por ejemplo, la información de las utilidades no se encuentra disponible sino después de varios períodos, y los agentes tendrán que esperar unos periodos para ajustar sus decisiones. En general, la autocorrelación es un problema muy común en modelos que emplean series de tiempo.

La autocorrelación entre los errores puede tomar muchas formas. En general, se dirá que los errores siguen un proceso autorregresivo de orden p ($AR(p)$) cuando el error depende de los p períodos anteriores. Por ejemplo, si el término de error tiene el siguiente comportamiento $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + v_t$ ($\forall t$), entonces se dirá que el error es auto-regresivo de orden 2 ($AR(2)$). Si el comportamiento es $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + v_t$, entonces los errores seguirán un proceso $AR(3)$. En general, si el comportamiento del error es $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + \dots + \rho_p \varepsilon_{t-p} + v_t$ entonces el error sigue un proceso $AR(p)$.

Matricialmente, la presencia de autocorrelación implica que la matriz de varianzas y covarianzas del término de error no es $\sigma^2 I_n$, sino una matriz cuadrada con la misma constante sobre la diagonal pero por fuera de la diagonal ya no se tienen ceros. Por ejemplo, en el caso de un error que sigue un proceso $AR(1)$ la siguiente matriz de varianzas y covarianzas de los errores será:

$$E[\varepsilon^T \varepsilon] = \Omega = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho & 1 \end{bmatrix} \neq \sigma^2 I_n$$

Como se mencionó anteriormente, en presencia de autocorrelación los estimadores MCO continúan siendo insesgados pero no tienen la mínima varianza posible.³ Es más, en presencia de autocorrelación, el estimador MCO de la matriz de varianzas y covarianzas del vector β será sesgado.⁴ Por lo tanto, si usamos este último estimador en pruebas de hipótesis (individuales o conjuntas) o intervalos de confianza para los coeficientes estimados, entonces obtendremos conclusiones erróneas en torno a los verdaderos β s.

³ En el apéndice 9.1 se presenta una demostración de esta afirmación.

⁴ En el apéndice 9.2 se presenta una demostración de esta afirmación.

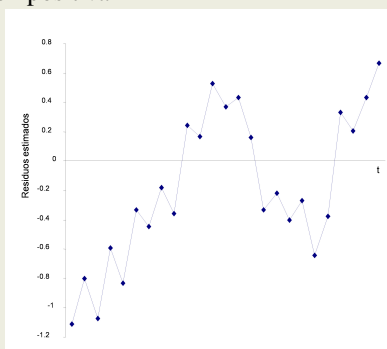
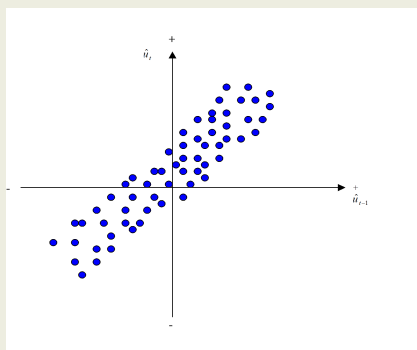
9.2. Pruebas para la detección de autocorrelación

Ejemplo 9.2 Tipos de autocorrelación de primer orden

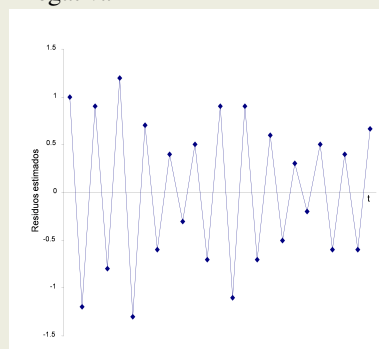
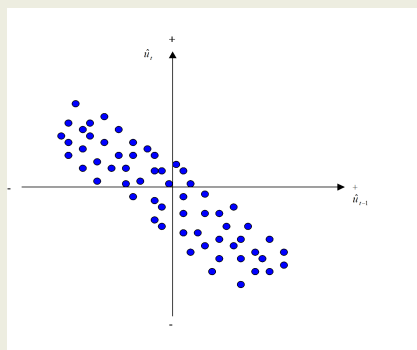
A continuación nos concentraremos en procesos $AR(1)$ por simplicidad. Cuando los errores siguen un proceso $AR(1)$, se pueden distinguir dos tipos de autocorrelación:

1. Autocorrelación positiva ($0 < \rho < 1$). En este caso, los errores de períodos adyacentes tienden a tener el mismo signo.
2. Autocorrelación negativa ($-1 < \rho < 0$). En este caso, los errores de períodos adyacentes tienden a tener diferente signo.

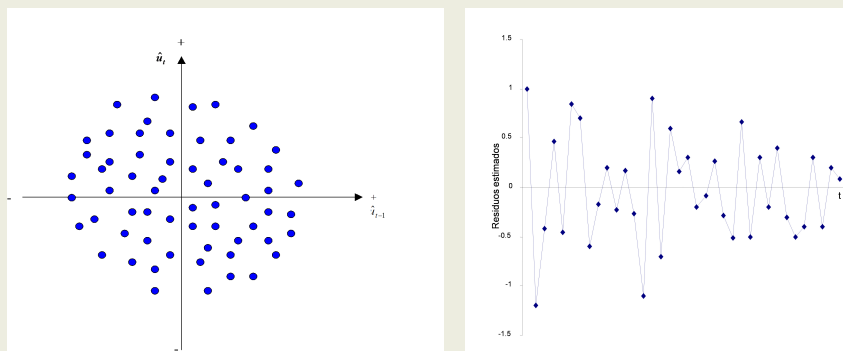
Autocorrelación positiva



Autocorrelación negativa



No autocorrelación



El primer paso, antes de realizar pruebas estadísticas, es verificar si los síntomas de este problema están presentes. Una vez más, estos síntomas pueden ser vistos en el vector de errores. Pero dicho vector no es observable, por tanto la mejor aproximación es examinar los errores estimados. Las gráficas más comunes son:

1. Los errores contra el tiempo \hat{e}_t vs $t = 1, 2, \dots, n$
2. Los errores contra los errores del periodo anterior \hat{e}_t vs \hat{e}_{t-1}

Estos gráficos permiten identificar algún tipo de regularidad como los que se presentan en el ejemplo 9.1. En los dos primeros gráficos se observa el comportamiento típico de errores con un proceso $AR(1)$ y con autocorrelación positiva, observamos que el signo de los residuos persiste en periodos prolongados de tiempo. Por otro lado en el caso de presencia de autocorrelación negativa sucede lo contrario, ya que el signo de los residuos cambia de un periodo a otro, mientras que en los gráficos de no autocorrelación, no podemos observar ningún patrón claro de comportamiento en el signo de los residuales.

Así como en el caso de la heteroscedasticidad, el análisis gráfico es un análisis informal que nos permite intuir la presencia del problema, sin embargo para un análisis más formal existen diferentes pruebas para identificar la autocorrelación.

9.2.1. Prueba de Rachas (Runs test)

La prueba de rachas, propuesta por **wald1**, es una prueba de independencia lineal no paramétrica cuya idea es relativamente sencilla. Si no hay autocorrelación, entonces no deberían haber muchos errores seguidos con el mismo signo (autocorrelación positiva), ni tampoco muchos cambios de signo seguido (autocorrelación negativa). En otras palabras, debe existir la cantidad adecuada de cambios de signo en una serie de datos: ni muchos, ni pocos.

Esta prueba tiene además la ventaja de no necesitar suponer una distribución de los errores. Para probar la hipótesis nula de que los errores son totalmente aleatorios ($H_0 : \rho = 0$) versus la alterna de que existe algún tipo de autocorrelación en los errores, se requiere seguir los siguientes pasos a partir de los errores estimados:

1. Cuente el número de errores con signo positivo (N_+) y con signo (N_-)
2. Cuente el número rachas (k), es decir de "seguidillas" de signo. Por ejemplo, si tenemos que los signos de los errores son: - - - - + + + - + + + - + +. Entonces se tendrán seis rachas ($k = 6$). (- - - -)(+ + +)(-)(+++)(-)(++). Note que el número de rachas es igual al número de cambios de signo
3. Si N_+ y/o N_- son menores que 20, entonces se puede construir un intervalo de confianza del 95 % para el número de rachas "razonable" bajo la hipótesis nula a partir de los valores críticos que se presentan en el apéndice 9.?? que se presentan al final de este capítulo
4. Si N_+ y/o N_- son mayores que 20, entonces se puede construir un intervalo de confianza del $(1 - \alpha)100\%$ para el número de rachas "razonable" bajo la hipótesis de la siguiente manera:

$$\left[E[k] \pm z_{\frac{\alpha}{2}} \sqrt{Var[k]} \right] \quad (9.2)$$

donde, el valor esperado y la varianza de k (las rachas) son:

$$E(k) = \frac{2N_+N_-}{N_+ + N_-} + 1 \quad (9.3)$$

$$Var(k) = \frac{2N_+N_- (2N_+N_- - N_+ - N_-)}{(N_+ + N_-)^2 (N_+ + N_- - 1)} \quad (9.4)$$

La hipótesis nula se puede rechazar si el número de rachas observadas no están contenidas en el intervalo de confianza.⁵

9.2.2. Prueba de Durbin-Watson

durbin4 diseñaron una prueba de autocorrelación con gran poder para detectar errores con autocorrelación de primer orden. Esta prueba se ha convertido en la más común para detectar este problema por ser relativamente intuitiva. Los autores definen el siguiente estadístico de prueba a partir de los errores estimados:

$$DW = \frac{\sum_{t=2}^n (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\hat{\epsilon}^T \hat{\epsilon}}$$

⁵ Una manera alternativa para comprobar la hipótesis nula, cuando se tienen más de 20 observaciones de un mismo signo es emplear como estadístico de prueba $RA = \frac{k - E(k)}{\sqrt{Var(k)}}$. La hipótesis nula puede ser rechazada si $|RA| > z_{\frac{\alpha}{2}}$.

Si la muestra es lo suficientemente grande es posible demostrar que $DW \approx 2(1 - \hat{\rho})$. De esta expresión se puede deducir que este estadístico estará acotado entre cero y 4 ($0 \leq DW \leq 4$). De hecho, como se muestra en el recuadro 9.2, intuitivamente se puede conocer el tipo de problema presente en la regresión a partir del valor del estadístico DW . Naturalmente, será necesario efectuar una prueba formal para determinar con mayor certeza si existe o no autocorrelación.

Recuadro 9.2 Estadístico DW en casos de Autocorrelación

No correlación	$\rho = 0$	$\hat{\rho} \approx 0$	$DW \approx 2$
Correlación positiva	$1 > \rho > 0$	$1 > \hat{\rho} > 0$	$DW < 2$
Correlación negativa	$-1 < \rho < 0$	$-1 < \hat{\rho} < 0$	$DW > 2$

Naturalmente la regla que se presenta en el recuadro 2 es únicamente intuitiva. Para tener una decisión con mayor grado de certidumbre se deberá efectuar una prueba de hipótesis.

El estadístico DW nos permite contrastar tres diferentes hipótesis nulas, como se reportan en el recuadro 9.3.

Recuadro 9.3

$H_0 : \rho = 0$	$\rho = 0$	$H_A : \rho \neq 0$
H_0 : No autocorrelación Positiva	$1 > \rho > 0$	$H_A : \rho > 0$
H_0 : No autocorrelación Negativa	$-1 < \rho < 0$	$H_A : \rho < 0$

durbin4 encontraron la distribución de su estadístico DW y la tabularon. Tradicionalmente se empleaba una tabla para poder tomar la decisión de rechazar o no la hipótesis nula de no autocorrelación. En la actualidad, es más común que la decisión se tome empleando un valor p .

Sobre esta prueba es importante destacar varios aspectos:

- El DW no tiene sentido si no hay intercepto (ver **durbin4**).
- El DW depende del supuesto que las X 's sean no estocásticas (ver **durbin4**).
- La prueba tiene un mayor poder ante procesos $AR(1)$.
- Esta prueba tampoco aplica en los casos en que existen variables dependientes rezagadas en la derecha del modelo; por ejemplo, para el modelo $Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \varepsilon_t$ este estadístico no aplica.

9.2.3. Prueba *h* de Durbin

Como se mencionó anteriormente, si el modelo emplea la variable dependiente rezagada como explicativa, la prueba de Durbin-Watson no aplica. Para solucionar este problema **durbin2** sugirió el siguiente estadístico:

$$h = \hat{\rho} \sqrt{\frac{n}{1 - n(\widehat{Var}(\hat{\alpha}))}}$$

donde:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \dots + \beta_k X_{kt} + \alpha Y_{t-1} + \varepsilon_t$$

Por lo tanto, es relativamente sencillo demostrar que:

$$h = \left(1 - \frac{DW}{2}\right) \sqrt{\frac{n}{1 - n(\widehat{Var}(\hat{\alpha}))}} \quad (9.5)$$

durbin2 demostró que este estadístico de prueba sigue una distribución estándar normal ($h \sim N(0, 1)$) y por lo tanto, si se cumple que $|h| > z_{\frac{\alpha}{2}}$ entonces es posible rechazar $H_0 : \rho = 0$ a favor de la $H_A : \rho \neq 0$. Finalmente, como podemos apreciar en la ecuación 9.5, esta prueba no es válida en los casos en que $n(\widehat{Var}(\hat{\alpha})) \geq 1$.

9.2.4. Prueba de Box-Pierce y Ljung-Box

Otra aproximación para comprobar la existencia o no de autocorrelación es determinar si las autocorrelaciones a diferentes rezagos son o no iguales a cero. **box1** diseñan una prueba basada en la autocorrelación muestral de los errores que permite detectar la existencia de errores con procesos más persistentes que AR(1). Recordemos que la autocorrelación poblacional se define de la siguiente forma:

$$\gamma_j = \frac{Cov(\varepsilon_t, \varepsilon_{t-j})}{\sqrt{Var(\varepsilon_t) Var(\varepsilon_{t-j})}} = \frac{Cov(\varepsilon_t, \varepsilon_{t-j})}{\sigma_\varepsilon^2}$$

Y la correspondiente autocorrelación muestral es:

$$\hat{\gamma}_j = \frac{\sum_{t=k+1}^n (\hat{\varepsilon}_t - \bar{\varepsilon})(\hat{\varepsilon}_{t-j} - \bar{\varepsilon})}{\sum_{t=1}^n (\hat{\varepsilon}_t - \bar{\varepsilon})^2} = \frac{\sum_{t=j+1}^n \hat{\varepsilon}_t \hat{\varepsilon}_{t-j}}{\sum_{t=1}^n (\hat{\varepsilon}_t)^2}$$

box1 definen una prueba que permite determinar si las primeras s autocorrelaciones son conjuntamente iguales a cero o no. Es decir, permite comprobar la hipótesis nula de un error no autocorrelacionado (las correlaciones a los s rezagos son cero),

versus la hipótesis alterna de la existencia de algún tipo de autocorrelación (por lo menos una autocorrelación no es cero). Para comprobar esta hipótesis nula, **box1** sugieren el estadístico Q :

$$Q = n \sum_{j=1}^s r_k^2 \sim_a \chi_s^2$$

donde s corresponden al número de rezagos que se desean considerar dentro de la prueba. Ellos demuestran que su estadístico sigue una distribución Chi-cuadrado con s grados de libertad (χ_s^2). Por lo tanto, será posible rechazar la H_0 (error no autocorrelacionado) si se cumple que $Q > \chi_s^2$.

Sin embargo, la prueba de Box-Pierce sólo es válida para muestras grandes ($n > 20$), para resolver este inconveniente **ljung1** proponen una modificación del estadístico anterior para que presente un mejor comportamiento en muestras pequeñas. El estadístico de Ljung-Box corresponde a:

$$Q' = n(n+2) \sum_{k=1}^s \frac{r_j^2}{n+j}$$

Este estadístico funciona y posee la misma distribución que el de la prueba de Box-Pierce.

Finalmente, es importante mencionar que una práctica muy común es realizar esta prueba para un número relativamente grande de rezagos. Es decir, hacer las correspondientes pruebas para diferentes rezagos; por ejemplo, se calculan los correspondientes estadísticos para comprobar las siguientes hipótesis alternas:

$$\begin{aligned} H_0 : \gamma_1 &= 0 \\ H_0 : \gamma_1 &= \gamma_2 = 0 \\ &\vdots \\ H_0 : \gamma_1 &= \gamma_2 = \dots \gamma_{m/3} = 0 \end{aligned}$$

La decisión de si los errores están o no autocorrelacionados se toma teniendo en cuenta las decisiones de cada una de estas pruebas.

9.2.5. Prueba de Breusch-Godfrey

breush2; **godfrey1** diseñaron una prueba que permite comprobar la hipótesis nula de no autocorrelación versus la alterna de que el error sigue un proceso autorregresivo de orden p . Esta prueba también es conocida como la prueba del multiplicador de Lagrange o prueba *LM* (por su sigla en inglés: Lagrange multiplier test).

Esta prueba se basa en una idea muy sencilla. Si existe autocorrelación en los errores, entonces éstos son explicados por sus valores pasados, pero si no hay auto-

correlación entonces los valores pasados de los errores no pueden explicar el comportamiento actual del error.

Así, para probar la hipótesis nula de no autocorrelación versus la alterna de unos errores con un proceso $AR(s)$, se pueden emplear los residuos estimados de la regresión bajo estudio para comprobar si los valores pasados del error sirven o no para explicar el error del periodo t . Es decir, la prueba LM implica los siguientes pasos:

1. Estime el modelo de regresión original:

$$y_t = \beta_1 + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \dots + \beta_k X_{k,t} + \varepsilon_t$$

y obtenga la serie de los errores estimados ($\hat{\varepsilon}_t$).

2. Estime la siguiente regresión auxiliar:

$$\hat{\varepsilon}_t = \alpha_1 + \alpha_2 X_{2,t} + \alpha_3 X_{3,t} + \dots + \alpha_k X_{k,t} + \omega_1 \hat{\varepsilon}_{t-1} + \omega_2 \hat{\varepsilon}_{t-2} + \dots + \omega_s \hat{\varepsilon}_{t-s} + \xi_t$$

3. Empleando el R^2 de la regresión auxiliar calcule el estadístico LM de la siguiente manera:⁶

$$LM = (n - s) \times R^2$$

4. Compare el estadístico LM con el valor crítico de la distribución Chi-cuadrado con s grados de libertad. Se rechazará la hipótesis nula si $LM > \chi^2_{s,\alpha}$.

Al igual que la prueba de Box-Pierce, cuando se emplea esta prueba normalmente se realizan las pruebas para diferentes hipótesis nulas y se toma la decisión basándose en el conjunto de los resultados.

9.3. Solución a la autocorrelación

Así como en el caso de la heteroscedasticidad (Ver Capítulo 8), existen dos formas de solucionar la existencia de autocorrelación. La primera solución es tratar de resolver de raíz el problema modificando la muestra. Este método se conoce como el método de Diferencias generalizadas que hace parte de la familia de los Mínimos Cuadrados Generalizados (MCG). Esta aproximación implica conocer exactamente cómo es la autocorrelación; algo que típicamente es difícil para el científico de datos. Esta aproximación será discutida en un capítulo más adelante.

La segunda opción implica solucionar los síntomas de la autocorrelación, tratando de estimar de manera consistente la matriz de varianza y covarianzas de los coeficientes estimados. Esto permite corregir los errores estándar de los coeficientes, y de esta manera los t calculados serán recalculados y por tanto los valores p son diferentes. Una aproximación muy similar al a de White para solucionar el problema de heteroscedasticidad.

⁶ Algunos paquetes estadísticos calculan el estadístico LM multiplicando el R^2 por n y no por $(n - s)$. Si el tamaño de la muestra es grande, estas dos aproximaciones son equivalentes, en caso contrario es mejor multiplicar por $(n - s)$.

9.3.1. Estimación Consistente en presencia de Autocorrelación de los errores estándar.

De manera similar a la solución de White (1980) para la heteroscedasticidad, Newey 1987 sugieren un estimador para la matriz de varianzas y covarianzas.

Recordemos, que en presencia de perturbaciones no esféricas tendremos que:

$$\Psi = \text{Var} \left[\hat{\beta} | \mathbf{X} \right] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Omega \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

Esto se puede reescribir como

$$\Psi = \left(\frac{1}{n} \mathbf{X}^T \mathbf{X} \right)^{-1} \frac{1}{n} \Phi \left(\frac{1}{n} \mathbf{X}^T \mathbf{X} \right)^{-1}$$

donde

$$\Phi = \frac{1}{n} \mathbf{X}^T \Omega \mathbf{X}$$

Y eso se puede escribir de la siguiente manera:

$$\Phi = \frac{1}{n} \sum_{i,j=1}^n w_{|i-j|} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_j^T$$

donde $w = [w_0 \dots w_{n-1}]^T$ corresponde a un vector de pesos. Newey-West (1987) sugieren

$$w_\ell = 1 - \frac{\ell}{L+1}$$

donde L corresponde al número máximo de rezagos. Generalmente $L \approx N^{1/4}$

Este estimador se caracteriza por ser consistente; es decir ser sesgado en muestras pequeñas, pero éste desaparece cuando la muestra se vuelve grande. Por eso este estimador es conocido como estimador H.A.C. (heteroskedasticity and autocorrelation consistent). De manera similar a lo que discutido para el caso de los H.C., estos estimadores no hace que los MCO se conviertan nuevamente en MELI. La varianza del estimador sigue siendo más grande que los GLS.

De manera similar al caso de la heteroscedásticas, el estimador de Newey-West genera varianzas de los estimadores que son relativamente más pequeñas y por tanto los t más grandes. Es decir, los t calculados y los correspondientes valores p deben ser corregidos.

En otras palabras, debemos modificar nuestras pruebas individuales y conjuntas empleando la correspondiente matriz H.A.C. Por ejemplo, la prueba de Wald para una restricción de la forma $\mathbf{R}\beta = \mathbf{q}$ será:

$$W = (\mathbf{R}\hat{\beta})^T \left[\mathbf{R} \left(\text{Est.Asy.Var} \left[\hat{\beta} \right] \right) \mathbf{R}^T \right]^{-1} (\mathbf{R}\hat{\beta})$$

Por otro lado, así como en el caso de la corrección H.C. diversos autores han intentado mejorar la aproximación provista por Newey-West:

- Andrews (1991) sugiere: $w_\ell = \frac{3}{z^2} \left(\frac{\sin(z)}{z} - \cos(z) \right)$, donde $z = 6\pi/5 \cdot \ell/B$ (`kernHAC()`)
- Lumley and Heagerty (1999) sugieren otra forma de pesar los datos (`weave()`)

Las dos últimas aproximaciones son las más usadas en la actualidad. Pero existe poca documentación de cuándo es mejor uno u otro caso. Esto es aún materia de investigación.

9.4. Práctica en R: Explicando los rendimientos de una acción (continuación)

En el capítulo 3 construimos un modelo para explicar el rendimiento de la acción de Suramericana empleando el rendimiento de otras acciones que se transan en la Bolsa de Valores de Colombia. Las acciones empleadas fueron: GRUPOSURA, ECOPETROL, NUTRESA, EXITO, ISA, GRUPOAVAL, CONCONCRETO, VALOREM y OCCIDENTE. Empleamos una base de datos que va desde 2012-01-02 hasta el 2019-01-14. La información se encuentra en el *working space* `RetornosDiarios.RData`.

En esta ocasión también incluiremos la información de la tasa de depósitos a término fijo a 90 días (DTF) como una variable que captura el rendimiento de un activo libre de riesgo. Esta información está disponible en el archivo `DataCDTs.xlsx`. Nuestro objetivo será estimar el siguiente modelo y solucionar la autocorrelación del modelo, si se encuentra. El modelo será:

$$\begin{aligned} \text{GRUPOSURA}_t = & \beta_1 + \beta_2 \text{ECOPETROL}_t + \beta_3 \text{NUTRESA}_t \\ & + \beta_4 \text{EXITO}_t + \beta_5 \text{ISA}_t + \beta_7 \text{GRUPOAVAL}_t \\ & + \beta_8 \text{CONCONCRETO}_t + \beta_9 \text{VALOREM}_t \\ & + \beta_{10} \text{OCCIDENTE}_t + \beta_{11} \text{DTF}_t + \varepsilon_t \end{aligned}$$

9.4.1. Construcción de la base de datos

Nuestra primera tarea es cargar los datos y consolidar todos los datos (los de los rendimientos y los de la *DTF* en un solo `data.frame`). Los datos de los rendimientos se encuentran en un archivo `.RData`, así que lo podemos cargar empleando la función `load()`.

```
load("Data/RetornosDiarios.RData")
class(retornos.diarios)
```

El objeto *retornos.diarios* es de la clase *xts*, lo cual permite manejar fácilmente las fechas.

La información de la *DTF* la podemos cargar con la función *read_excel()* del paquete *readxl*. Esta función tiene una característica importante para este ejercicio, pues nos permite cargar un archivo y especificarle el tipo de variable que deberá aplicar a cada columna al momento de cargar los datos. La primera columna del archivo de Excel corresponde a las fechas y la segunda a los datos como tal de la *DTF*. Para evitar que se pierda la información de la fecha podemos emplear el siguiente código:

```
library(readxl)
DTF <- read_excel("Data/DataCDTs.xlsx", col_types = c("date", "numeric") )
head(DTF, 2)

## # A tibble: 2 x 2
##   `Fecha(dd/mm/aaaa)` DTF90dias
##   <dtm>                <dbl>
## 1 2012-01-02 00:00:00    0.0513
## 2 2012-01-03 00:00:00    0.0558

class(DTF)

## [1] "tbl_df"      "tbl"        "data.frame"

colnames(DTF)[1] <- "Fecha"
```

Noten que el nombre de la primera columna fue modificado para hacer más fácil la manipulación de esa variable. Por otro lado, la clase del objeto *DTF* no es *xts*. Procedamos a cambiar dicha clase para hacer más fácil la unión de las dos bases de datos.

```
library(xts)
DTF$Fecha <- as.Date(DTF$Fecha)
class(DTF$Fecha)

## [1] "Date"

DTF90dias <- xts(DTF$DTF90dias, order.by = DTF$Fecha)
class(DTF90dias)

## [1] "xts" "zoo"
```

Antes de unir los dos objetos Podemos constatar que ambos objetos tienen la misma periodicidad y que cubran el mismo periodo.

```
periodicity(DTF90dias)

## Daily periodicity from 2012-01-02 to 2019-01-14

periodicity(retornos.diarios)

## Daily periodicity from 2012-01-02 to 2019-01-14
```

Ya podemos unir los dos objetos por medio de la función *merge()* del paquete *xts* (es decir, *merge.xts()*). Esta función permite pegar dos objetos *xts* de diferentes

maneras empleando el argumento *join*. Si *join = "outer"*, se crea una base de datos con todas las fechas incluidas en los dos objetos. Si en uno de los objetos no existía una fecha, entonces los valores faltantes se remplazan por "NA"⁷. Si *join = "inner"* se construirá una nueva base de datos únicamente con las filas (fechas) que estén en común en ambos objetos. Si *join = "left"* el nuevo objeto tendrá solo las fechas del primer objeto. Si el segundo objeto no tiene información para una de esas fechas, se rellenará esa información con un "NA". De manera similar, Si *join = "right"*, el nuevo objeto tendrá las fechas del segundo objeto.

En nuestro caso, el lector puede constatar que los dos objetos, si bien cubren el mismo periodo, no tienen la misma cantidad de datos. Esto ocurre porque hay unos días hábiles en los que la Bolsa de Valores no se encuentra abierta, pero si se recoge información para la DTF. Así, que dado nuestro objetivo será mas conveniente unir los objetos de tal manera que tengamos observaciones para los días en los que la Bolsa de Valores estuvo abierta. Es decir,

```
data <- merge(retornos.diarios, DTF90dias, join = "left")
head(data, 3)
```

		GRUPOSURA	ECOPETROL	NUTRESA	EXITO
##	2012-01-02	1.2779727	-0.3565066	-0.9216655	2.02184181
##	2012-01-03	2.5079684	2.0036027	-0.2781643	0.07695268
##	2012-01-04	0.4324999	1.0446990	-0.5586607	1.07116556
		ISA	GRUPOAVAL	CONCONCRET	VALOREM
##	2012-01-02	-1.983834	0.0000000	0	0.0000000
##	2012-01-03	1.626052	-2.4292693	0	12.583905
##	2012-01-04	2.478003	-0.4106782	0	1.342302
		OCCIDENTE	DTF90dias		
##	2012-01-02	0	0.05131078		
##	2012-01-03	0	0.05576149		
##	2012-01-04	0	0.04769692		

9.4.2. Residuales del modelo y análisis gráfico de los residuales

Ya podemos correr el modelo y examinar los respectivos residuales. Esto lo podemos hacer con un gráfico de líneas de los residuos en función del tiempo y uno de dispersión de los residuales en el periodo actual versus los residuos rezagados⁸.

```
modelo1 <- lm(GRUPOSURA ~ ., data)
summary(modelo1)
```

```
##
## Call:
```

⁷ La función permite cambiar como se rellena los datos faltantes empleando el argumento *fill*. Por defecto *fill = NA*.

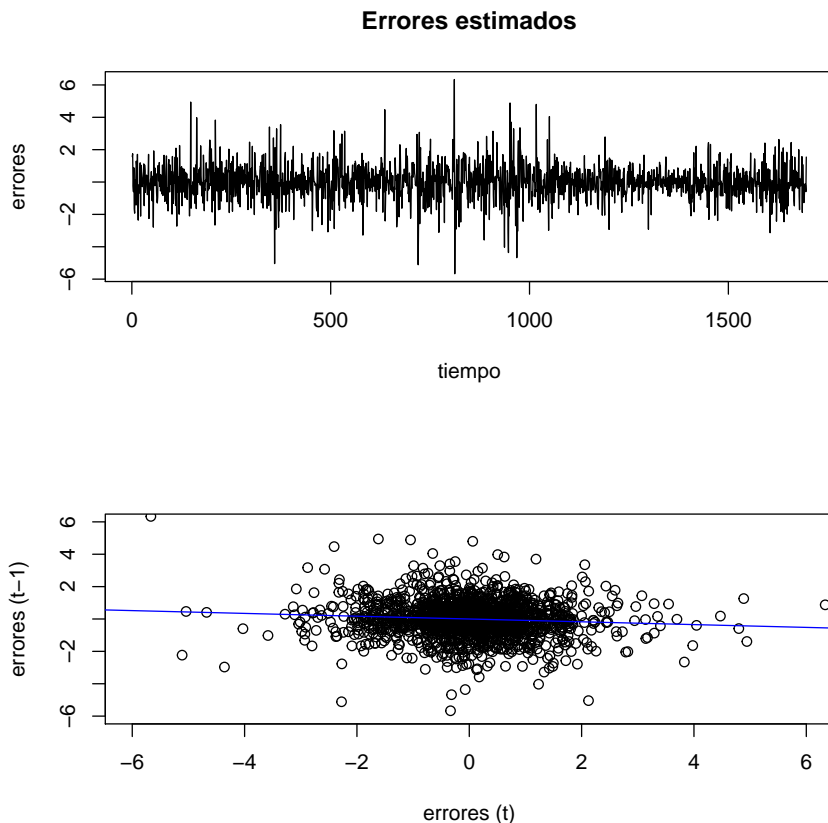
⁸ El termino rezagado implica que la variable se observa en un periodo anterior. Por ejemplo, cuando hablamos de la variable y_t rezagada un periodo, nos estamos refiriendo a y_{t-1} . Esto se puede hacer en R empleando la función *lag.xts()*

```
## lm(formula = GRUPOSURA ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.6689 -0.5673 -0.0126  0.6206  6.3370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.09227    0.14058   0.656   0.5117
## ECOPETROL     0.12139    0.01450   8.372 < 2e-16 ***
## NUTRESA       0.14137    0.02780   5.085 4.09e-07 ***
## EXITO         0.12255    0.01812   6.762 1.87e-11 ***
## ISA           0.18844    0.01870  10.078 < 2e-16 ***
## GRUPOAVAL     0.08437    0.02008   4.201 2.79e-05 ***
## CONCONCRET    0.02138    0.01479   1.446   0.1484
## VALOREM       0.03552    0.01401   2.535   0.0113 *
## OCCIDENTE     0.03076    0.02694   1.142   0.2536
## DTF90dias    -1.69580    2.72526  -0.622   0.5339
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.11 on 1686 degrees of freedom
## Multiple R-squared:  0.2619, Adjusted R-squared:  0.2579
## F-statistic: 66.46 on 9 and 1686 DF,  p-value: < 2.2e-16

e <- residuals(modelo1)
class(e)

## [1] "numeric"

par(mfrow=c(2,1))
ts.plot(e, main= "Errores estimados", xlab= "tiempo",
        ylab= "errores")
plot(e, lag.xts(e), xlab= "errores (t)", ylab= "errores (t-1)",
     xlim=c(-6, 6), ylim=c(-6, 6))
reg <- lm(e ~ lag.xts(e))
abline(reg, col="blue")
```

El primer gráfico muestra uno errores que alternan "mucho" su signo, esto puede ser muestra de una autocorrelación negativa. Por otro lado, el segundo gráfico no presenta una fuerte relación negativa entre los errores. Naturalmente, los gráficos nunca serán concluyentes, pero si nos permiten tener una intuición de lo que está ocurriendo con los residuales. A continuación se presentan las pruebas de autocorrelación discutidas anteriormente.

9.4.3. Pruebas de Autocorrelación

Recordemos que todas las pruebas descritas anteriormente tienen como hipótesis nula el cumplimiento del supuesto (no autocorrelación) y la alterna la violación de alguna manera del supuesto. Procedamos a efectuar dichas pruebas.

9.4.4. Prueba de Rachas (*Runs test*)

Para realizar la prueba de rachas, podemos emplear la función *runs.test()* del paquete *tseries*. Esta función tiene dos argumentos. El primer argumento es obligatorio y corresponde a un objeto que contenga una variable de clase *factor* que muestre si el error es positivo o negativo. El segundo argumento (*alternative*) no es obligatorio y determina cuál es la hipótesis alterna que se desea probar. Si *alternative* = "two.sided", la alterna será que existe algún tipo de autocorrelación. Esta es la opción por defecto, es decir si no se especifica este argumento, se efectuará esta prueba. Si *alternative* = "less" la alterna es que la autocorrelación es positiva (menos cambios de signos que los esperados) y si *alternative* = "greater", la alterna es que existe autocorrelación negativa (mas cambios de signos que los esperados). Así, nuestro primer paso para efectuar esta prueba es convertir los residuos estimados en una variable dicotómica que muestre el signo del residual para cada periodo.

```
library(tseries)
signo.error <- factor(e>0)
head(signo.error, 3)

## 2012-01-02 2012-01-03 2012-01-04
##          TRUE      TRUE      FALSE
## Levels: FALSE TRUE

runs.test(signo.error)

##
##  Runs Test
##
## data:  signo.error
## Standard Normal = 2.7324, p-value = 0.006288
## alternative hypothesis: two.sided

runs.test(signo.error, alternative="less")

##
##  Runs Test
##
## data:  signo.error
## Standard Normal = 2.7324, p-value = 0.9969
## alternative hypothesis: less

runs.test(signo.error, alternative="greater")

##
##  Runs Test
##
## data:  signo.error
## Standard Normal = 2.7324, p-value = 0.003144
## alternative hypothesis: greater
```

Los resultados permiten rechazar con un 99% de confianza (valor p de 0.0063) la no autocorrelación y por tanto se puede concluir que existe algún tipo de autocorrelación. Por otro lado, no se puede rechazar la hipótesis nula de que no existe autocorrelación o autocorrelación (en favor de la alterna de autocorrelación positi-

va) dado que el correspondiente valor p es 0.9969. Finalmente, podemos concluir con esta prueba que existe autocorrelación negativa con un 99% de confianza (valor p de 0.0031).

9.4.5. Prueba de Durbin-Watson

Esta prueba se puede implementar empleando la función `dwtest()` del paquete *AER*. Esta función tiene dos argumentos principales. El primero es un objeto de clase *lm* que contenga el modelo al cual se le quiere efectuar la prueba. este argumento es obligatorio. El segundo argumento no es obligatorio y permite escoger cuál es la hipótesis alterna que se desea probar. Si `alternative = "two.sided"`, la alterna será que existe algún tipo de autocorrelación. Si `alternative = "greater"` la alterna es que la autocorrelación es positiva. Esta es la opción por defecto, es decir si no se especifica este argumento, se efectuará esta prueba. Y si `alternative = "less"`, la alterna es que existe autocorrelación negativa. Noten que este argumento funciona algo diferente a lo descrito con la prueba de rachas, y por tanto requiere mucho cuidado al momento de emplearlas.

La prueba se puede implementar de la siguiente manera:

```
library(AER)
dwtest(modelo1, alternative = "two.sided")

##
## Durbin-Watson test
##
## data: modelo1
## DW = 2.1703, p-value = 0.0004734
## alternative hypothesis: true autocorrelation is not 0

dwtest(modelo1, alternative = "greater")

##
## Durbin-Watson test
##
## data: modelo1
## DW = 2.1703, p-value = 0.9998
## alternative hypothesis: true autocorrelation is greater than 0

dwtest(modelo1, alternative = "less")

##
## Durbin-Watson test
##
## data: modelo1
## DW = 2.1703, p-value = 0.0002367
## alternative hypothesis: true autocorrelation is less than 0
```

Estos resultados son similares a los obtenidos por la prueba de rachas. Se rechaza la nula de no autocorrelación (valor p de 5×10^{-4}) y también se concluye que existe autocorrelación negativa (valor p de 2×10^{-4})

Antes de continuar con la siguiente prueba, es importante mencionar que al no contar este modelo con la variable explicativa rezaga como variable explicativa, no es relevante la prueba h de Durbin. Cuando sea necesario, esta prueba se puede efectuar empleando la función *durbinH()* del paquete *ecm*.

9.4.6. Prueba de Box-Pierce y Ljung-Box

La prueba de Box-Pierce y la modificación de Ljung-Box pueden calcularse empleando la misma función de la base de R: *Box.test*. Esta función tiene tres argumentos importantes,

- el vector al que se le quiere hacer la prueba
- el número de rezagos (*lag*) para incluir en la hipótesis nula y alterna. Por defecto $lag = 1$.
- el tipo de prueba (*type*). Las opciones son *type = "Box-Pierce"* (la opción por defecto de la función) y *type = "Ljung-Box"*. En el segundo caso se realiza la prueba de Ljung-Box que introduce la corrección para muestras pequeñas.

Entonces para probar la hipótesis $H_0 : \gamma_1 = 0$ con la prueba de Box-Pierce podemos emplear el siguiente código

```
Box.test(e, lag = 1)

##
##   Box-Pierce test
##
## data:  e
## X-squared = 12.642, df = 1, p-value = 0.0003772
```

Así, podemos rechazar la hipótesis nula de no autocorrelación (para el primer rezago). Como se mencionó anteriormente, es común probar esta hipótesis para los primeros rezagos, por lo menos los primeros 20. Es decir,

$$\begin{aligned} H_0 : \gamma_1 &= 0 \\ H_0 : \gamma_1 &= \gamma_2 = 0 \\ &\vdots \\ H_0 : \gamma_1 &= \gamma_2 = \dots \gamma_{20} = 0 \end{aligned}$$

Para realizar estas prueba podemos crear una función para construir una tabla con todas las pruebas que se deseen

```
tabla.Box.Pierce <- function(residuo, max.lag = 20,
                             type = "Box-Pierce"){
  # se crean objetos para guardar los resultados
```

```
BP.estadistico <- matrix(0,max.lag,1)
BP.pval <-matrix(0,max.lag,1)

# se calcula la prueba para los diferentes rezagos
for (i in 1:max.lag) {
  BP<- Box.test(residuo, lag = i, type = type)
  BP.estadistico[i]<-BP$statistic
  BP.pval[i]<-round(BP$p.value,5)
}
labels<- c( "Rezagos", type, "p-valor")

Cuerpo.Tabla <- cbind(matrix(1:max.lag,max.lag,1),
                      BP.estadistico, BP.pval)
TABLABP <- data.frame(Cuerpo.Tabla)
names(TABLABP) <- labels
return(TABLABP)
}
```

Ahora podemos emplear la función para crear el siguiente cuadro.

Cuadro 9.1 Prueba de Box-Pierce de los errores para los primeros rezagos

Rezagos	Box-Pierce	p-valor
1.00	12.64	0.00
2.00	17.53	0.00
3.00	17.62	0.00
4.00	18.53	0.00
5.00	22.76	0.00
6.00	25.18	0.00
7.00	26.34	0.00
8.00	28.83	0.00
9.00	28.90	0.00
10.00	29.92	0.00
11.00	30.07	0.00
12.00	31.09	0.00
13.00	31.09	0.00
14.00	37.82	0.00
15.00	37.82	0.00
16.00	39.31	0.00
17.00	42.41	0.00
18.00	43.89	0.00
19.00	43.91	0.00
20.00	43.96	0.00

Los resultados nos permiten demos concluir que las autocorrelaciones de los errores no son cero. Así podemos concluir que existe autocorrelación.

El lector puede generar fácilmente la siguiente tabla que contienen los resultados de la prueba de Ljung-Box aplicada a los mismos residuos.

No es sorprendente que los resultados de esta prueba sean los mismo que los obtenidos con la prueba de Box-pierce, pues en este caso la muestra es grande. Así, la corrección para muestras pequeñas no era importante.

Cuadro 9.2 Prueba de Ljung-Box de los errores para los primeros rezagos

Rezagos	Box-Pierce	p-valor
1.00	12.64	0.00
2.00	17.53	0.00
3.00	17.62	0.00
4.00	18.53	0.00
5.00	22.76	0.00
6.00	25.18	0.00
7.00	26.34	0.00
8.00	28.83	0.00
9.00	28.90	0.00
10.00	29.92	0.00
11.00	30.07	0.00
12.00	31.09	0.00
13.00	31.09	0.00
14.00	37.82	0.00
15.00	37.82	0.00
16.00	39.31	0.00
17.00	42.41	0.00
18.00	43.89	0.00
19.00	43.91	0.00
20.00	43.96	0.00

9.4.7. Prueba de Breusch-Godfrey

Esta prueba se puede realizar empleando la función *bgtest* del paquete *lmtest*. Similar a las anteriores pruebas, esta función tiene dos argumentos. El primero es el objeto *lm* al que se le quiere hacer la prueba. El segundo argumento es el orden (*order*) de la autocorrelación que se desea probar. Por defecto este argumento es igual a uno. Para probar la hipótesis nula de no autocorrelación versus la alterna de unos errores con un proceso $AR(1)$ podemos emplear el siguiente código

```
library(lmtest)
bgtest(modelo1, order = 1)

##
## Breusch-Godfrey test for serial correlation of order
## up to 1
##
## data: modelo1
## LM test = 12.818, df = 1, p-value = 0.0003434
```

Los resultados muestran qué se puede rechazar la hipótesis nula. En otras palabras, podríamos concluir que los errores pueden seguir un proceso $AR(1)$, o lo que es equivalente, existe autocorrelación. De manera similar a la anterior prueba, es usual realizar la prueba para diferentes ordenes del proceso AR. En la práctica no es muy común que esta prueba se realice para muchos rezagos. A continuación se presenta una función que permite realizar la prueba para los rezagos deseados.

```
tabla.Breusch.Godfrey <- function(modelo, max.order = 5){
  # se crean objetos para guardar los resultados
```

```

BG.estadistico <- matrix(0, max.order, 1)
BG.pval <-matrix(0, max.order, 1)

# se calcula la prueba para los diferentes rezagos
for (i in 1:max.order) {
  BG<- bgtest(modelo, order = i)
  BG.estadistico[i]<--BG$statistic
  BG.pval[i]<-round(BG$p.value,5)
}

labels<- c( "Orden AR(s)", "Breusch-Godfrey", "p-valor")

Cuerpo.Tabla <- cbind(matrix(1:max.order,max.order,1),
                      BG.estadistico, BG.pval)
TABLABP <- data.frame(Cuerpo.Tabla)
names(TABLABP) <- labels
return(TABLABP)
}

```

Ahora podemos emplear la función para crear el siguiente cuadro.

Cuadro 9.3 Prueba de Breusch-Godfrey de los errores

Orden AR(s)	Breusch-Godfrey	p-valor
1.00	-12.82	0.00
2.00	-19.29	0.00
3.00	-19.84	0.00
4.00	-21.31	0.00
5.00	-26.85	0.00

Esta prueba también permite concluir que existe un problema de autocorrelación.

9.4.8. Solución al problema de heteroscedasticidad con H.A.C.

Todas las pruebas nos llegan a concluir que tenemos un problema de autocorrelación. Este problema lo podemos solucionar empleando estimadores H.A.C. para la matriz de varianzas y covarianzas. Al igual que lo hicimos con la heteroscedasticidad, esto se puede hacer empleando el paquete *sandwich* y las siguientes funciones:

- *NeweyWest()*: para obtener la corrección de Newey-West (1987)
- *kernHAC()*: para obtener la corrección de Andrews (1991)
- *weave()*: para obtener la corrección de Lumley and Heagerty (1999)

Las tres funciones tienen como argumento el objeto *lm* al que se le quiere corregir la matriz de varianzas y covarianzas. Por ejemplo, para obtener la matriz de varianzas y covarianzas con la corrección de Newey-West (1987) de la siguiente manera:

```
library(sandwich)
NeweyWest(modelol)

## (Intercept)          ECOPETROL          NUTRESA
## (Intercept)  9.251428e-03 -1.868819e-04  7.256654e-05
## ECOPETROL   -1.868819e-04  3.123882e-04  3.780966e-05
## NUTRESA      7.256654e-05  3.780966e-05  1.226507e-03
## EXITO       -6.857704e-05  1.318300e-06 -4.168024e-05
## ISA         2.647414e-04 -1.091065e-04 -2.535839e-04
## GRUPOAVAL   -5.466807e-05  3.521443e-05 -1.406924e-04
## CONCRET     -1.218345e-05 -5.385873e-06 -6.182498e-05
## VALOREM     -1.496096e-04 -6.511587e-06 -5.742468e-05
## OCCIDENTE   -3.419532e-05 -5.897218e-06  8.906229e-06
## DTF90dias   -1.665297e-01  4.001086e-03 -3.423908e-04
##          EXITO          ISA          GRUPOAVAL
## (Intercept) -6.857704e-05  2.647414e-04 -5.466807e-05
## ECOPETROL   1.318300e-06 -1.091065e-04  3.521443e-05
## NUTRESA     -4.168024e-05 -2.535839e-04 -1.406924e-04
## EXITO       7.578591e-04 -1.779598e-04  3.037862e-05
## ISA        -1.779598e-04  7.044014e-04 -5.601628e-05
## GRUPOAVAL   3.037862e-05 -5.601628e-05  2.960277e-04
## CONCRET     -1.947145e-05 -3.467629e-05  2.009650e-08
## VALOREM     -9.317952e-05  3.369104e-05  6.656075e-06
## OCCIDENTE   -1.145985e-04 -3.396612e-05 -1.318819e-06
## DTF90dias   1.317957e-03 -6.602825e-03  1.326814e-03
##          CONCRET          VALOREM          OCCIDENTE
## (Intercept) -1.218345e-05 -1.496096e-04 -3.419532e-05
## ECOPETROL   -5.385873e-06 -6.511587e-06 -5.897218e-06
## NUTRESA     -6.182498e-05 -5.742468e-05  8.906229e-06
## EXITO       -1.947145e-05 -9.317952e-05 -1.145985e-04
## ISA        -3.467629e-05  3.369104e-05 -3.396612e-05
## GRUPOAVAL   2.009650e-08  6.656075e-06 -1.318819e-06
## CONCRET     2.094780e-04  5.700092e-06  4.102452e-07
## VALOREM     5.700092e-06  2.199244e-04  5.590307e-05
## OCCIDENTE   4.102452e-07  5.590307e-05  9.564615e-04
## DTF90dias   5.690392e-04  1.835562e-03 -8.530728e-04
##          DTF90dias
## (Intercept) -0.1665296873
## ECOPETROL   0.0040010864
## NUTRESA     -0.0003423908
## EXITO       0.0013179571
## ISA        -0.0066028252
## GRUPOAVAL   0.0013268136
## CONCRET     0.0005690392
## VALOREM     0.0018355623
## OCCIDENTE   -0.0008530728
## DTF90dias   3.1537629951
```

Ahora, como no es muy útil la matriz de varianzas y covarianzas sola, sino mas bien los respectivos t individuales y sus correspondientes valores p, podemos emplear la función *coefest()* para realizar las pruebas individuales. Esta función ya la habíamos estudiado en el Capítulo 8. Por ejemplo:


```

coeftest(modelo1, vcov = NeweyWest(modelo1))

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.092275   0.096184   0.9594   0.33752
## ECOPETROL    0.121392   0.017675   6.8682 9.112e-12 ***
## NUTRESA      0.141372   0.035022   4.0367 5.664e-05 ***
## EXITO        0.122549   0.027529   4.4516 9.082e-06 ***
## ISA          0.188444   0.026541   7.1002 1.829e-12 ***
## GRUPOAVAL    0.084369   0.017205   4.9036 1.032e-06 ***
## CONCRET      0.021385   0.014473   1.4775   0.13973
## VALOREM      0.035525   0.014830   2.3955   0.01671 *
## OCCIDENTE    0.030765   0.030927   0.9948   0.31999
## DTF90dias   -1.695802   1.775884  -0.9549   0.33976
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Para este caso podemos ver como individualmente, los coeficientes asociados a los rendimientos de CONCRET, OCCIDENTE y la DTF a 90 días no son significativos individualmente.

Ahora hagamos lo mismo para las correcciones de Andrews (1991) y Lumley and Heagerty (1999).

```

coeftest(modelo1, vcov = kernHAC(modelo1))

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.092275   0.121312   0.7606   0.44698
## ECOPETROL    0.121392   0.018125   6.6975 2.879e-11 ***
## NUTRESA      0.141372   0.035605   3.9705 7.472e-05 ***
## EXITO        0.122549   0.027337   4.4829 7.860e-06 ***
## ISA          0.188444   0.024900   7.5679 6.204e-14 ***
## GRUPOAVAL    0.084369   0.019500   4.3266 1.603e-05 ***
## CONCRET      0.021385   0.016233   1.3174   0.18790
## VALOREM      0.035525   0.015976   2.2236   0.02631 *
## OCCIDENTE    0.030765   0.035191   0.8742   0.38212
## DTF90dias   -1.695802   2.268825  -0.7474   0.45490
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

coeftest(modelo1, vcov = weave(modelo1))

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.092275   0.133149   0.6930   0.48839

```

```
## ECOPETROL      0.121392    0.016945    7.1639 1.167e-12 ***
## NUTRESA        0.141372    0.033778    4.1853 2.995e-05 ***
## EXITO          0.122549    0.024031    5.0997 3.786e-07 ***
## ISA            0.188444    0.023324    8.0795 1.227e-15 ***
## GRUPOAVAL      0.084369    0.018810    4.4853 7.774e-06 ***
## CONCRETET      0.021385    0.015415    1.3872 0.16556
## VALOREM        0.035525    0.016485    2.1550 0.03131 *
## OCCIDENTE      0.030765    0.036069    0.8529 0.39381
## DTF90dias      -1.695802    2.488200   -0.6815 0.49562
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Las tres correcciones coinciden en concluir que las siguientes variables no son significativas: CONCRETET, OCCIDENTE y la DTF a 90 días.

Ahora miremos si conjuntamente todos los coeficientes que acompañan a dichas variables son no significativos. Estimemos el correspondiente modelo anidado y comparemos los modelos

```
modelo2 <- lm(GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA
              + GRUPOAVAL + VALOREM, data)
```

```
waldtest(modelo2, modelo1, vcov = NeweyWest(modelo1))
```

```
## Wald test
```

```
##
```

```
## Model 1: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + VALOREM
```

```
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCRETET
```

```
## VALOREM + OCCIDENTE + DTF90dias
```

```
## Res.Df Df      F Pr(>F)
```

```
## 1      1689
```

```
## 2      1686  3 1.3719 0.2496
```

```
waldtest(modelo2, modelo1, vcov = kernHAC(modelo1))
```

```
## Wald test
```

```
##
```

```
## Model 1: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + VALOREM
```

```
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCRETET
```

```
## VALOREM + OCCIDENTE + DTF90dias
```

```
## Res.Df Df      F Pr(>F)
```

```
## 1      1689
```

```
## 2      1686  3 0.9793 0.4016
```

```
waldtest(modelo2, modelo1, vcov = weave(modelo1))
```

```
## Wald test
```

```
##
```

```
## Model 1: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + VALOREM
```

```
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCRETET
```

```
## VALOREM + OCCIDENTE + DTF90dias
```

```
## Res.Df Df      F Pr(>F)
```

```
## 1      1689
```

```
## 2      1686  3 1.055 0.3672
```

Los resultados muestran, con todas las correcciones que las tres variables son conjuntamente no significativas. En otras palabras el modelo restringido es mejor que él sin restringir. Así, podemos afirmar que la DTF a 90 días, los rendimientos de CONCRETO y OCCIDENTE no afectan el rendimiento de la acción de Suramericana.

9.5. Ejercicios

El gobierno de un pequeño país caribeño está interesado en conocer la relación existente entre el ingreso disponible y el consumo privado, con el fin de establecer la política impositiva que regirá en el año siguiente. Se cuenta con datos sobre el consumo privado y la renta disponible recolectada desde el primer trimestre de 1959 hasta el tercer trimestre del año 2002, ambos valores medidos en millones de moneda local. Los datos se encuentran en el archivo auto.xls.

De acuerdo con la siguiente información responda:

1. Estime el modelo que explica la situación. Reporte sus resultados en una tabla.
2. Efectúe el análisis gráfico de los errores estimados. ¿Qué tipo de problema puede intuir a partir de este análisis? Explique.
3. Realice las pruebas que considere necesarias para determinar la existencia o no de un problema de autocorrelación en el modelo. Especifique siempre las hipótesis que sustentan cada prueba y muestre claramente la conclusión a la que llega.
4. Según las conclusiones que extrajo del punto anterior:
 - a) ¿A qué conclusión llega? ¿explique por medio de la teoría econométrica cómo solucionaría el problema si es que este existe?
 - b) Ahora, demuestre que el problema ha desaparecido (realice las pruebas pertinentes). Además estime el nuevo modelo e interprete los coeficientes estimados (siempre y cuando haya encontrado un problema econométrico en el modelo original)

9.6. Apéndice

Apéndice 9.1 Demostración de la insesgadez de los estimadores en presencia de autocorrelación

En presencia de autocorrelación los estimadores MCO siguen siendo insesgados. Esta afirmación se puede demostrar fácilmente. Sin perder generalidad consideremos un modelo lineal con un término de error homoscedástico y con autocorrelación de orden uno. Es decir,

$$y = X\beta + \varepsilon$$

Donde $E[\varepsilon_t] = 0$, $Var[\varepsilon_t] = \sigma_\varepsilon^2$ y $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, $\forall t$. Ahora determinemos si $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ sigue siendo insesgado o no. Así,

$$E[\hat{\beta}] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE[\mathbf{y}]$$

$$E[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE[\mathbf{X}\beta + \varepsilon] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE[\varepsilon]$$

$$E[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = I \bullet \beta$$

$$E[\hat{\beta}] = \beta$$

Apéndice 9.2 Sesgo de la matriz de varianzas y covarianzas en presencia de autocorrelación

En presencia de autocorrelación el estimador de la matriz de varianzas y covarianzas de MCO ($\widehat{Var}[\hat{\beta}] = s^2(\mathbf{X}^T\mathbf{X})$) es sesgado. Es más el estimador MCO para los coeficientes ($\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$) no es eficiente; es decir no tiene la mínima varianza posible. Esta afirmación se puede demostrar fácilmente.

Continuando con el modelo considerado en el Apéndice anterior (error con una autocorrelación de orden uno), en este caso tenemos que:

$$Var[\varepsilon] = E[\varepsilon^T\varepsilon] = \Omega = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho & 1 \end{bmatrix} \quad (9.6)$$

Ahora podemos calcular la varianza de los estimadores MCO. Es decir,

$$Var[\hat{\beta}] = Var[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \quad (9.7)$$

Por tanto tendremos que

$$Var[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TVar[\mathbf{y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \quad (9.8)$$

$$Var[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TVar[\varepsilon]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \quad (9.9)$$

$$Var[\hat{\beta}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Omega\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \quad (9.10)$$

Por tanto la varianza no es la mínima posible. Y por otro lado, al emplear el estimador MCO para la matriz de varianzas y covarianzas de los betas ($\widehat{Var}[\hat{\beta}] =$

$s^2 (\mathbf{X}^T \mathbf{X})^{-1}$) en presencia de autocorrelación se obtendrá un estimador cuyo valor esperado no es igual a la varianza real; es decir, será insesgado.

Índice alfabético

R^2 , 12

distribución de los MCO, 4

distribución t, 5

F-global, 14

inferencia, 2

mínimos cuadrados ordinarios, 2

prueba global de significancia, 13, 14

pruebas conjuntas sobre los parámetros, 12

pruebas individuales sobre los parámetros, 4

R-cuadrado, 10

SSE, 9

SSR, 9

SST, 9

suma total al cuadrado, 9

t-calculado, 5

tabla ANOVA, 10, 14

teorema del límite central, 3, 4

valor p, 6

variación total al cuadrado, 9