

Julio César Alonso

Serie libros de texto

**Introducción al Modelo Clásico de
Regresión para Científico de Datos en R**

Primera Versión para comentarios 29 de agosto de 2021

Libro digital



Cienfi (<http://www.icesi.edu.co/centros-academicos/cienfi>).

Copyright© Introducción al Modelo Clásico de Regresión para Científico de Datos en R.

Correo Electrónico: jcalonso@icesi.edu.co

Escuela de Economía

Universidad Icesi

Calle 18 No 122-135, Cali, Colombia

Julio César Alonso C.

Introducción al Modelo Clásico de Regresión para Científico de Datos en R. 1ra ed.

– Universidad Icesi. 2021.

340 pp.

ISBN Obra Independiente: XXXXXXXX

1. Ciencia de Datos. 2. Modelo de Regresión 3. R.

Índice general

1	Introducción	9
1.1	Tareas del científico de datos	11
1.2	Tipos de analítica	15
1.3	Aproximación econométrica vs los científicos de datos	17
I	El modelo clásico de regresión múltiple	21
2	Modelo de regresión múltiple	23
2.1	Introducción	24
2.2	El modelo de regresión múltiple	30
2.2.1	Supuestos	32
2.2.2	Método de mínimos cuadrados ordinarios (MCO)	34
2.2.3	Propiedades de los estimadores MCO	39

2.3 Práctica en R	40
2.3.1 Lectura de datos	41
2.3.2 Estimación del modelo	43
2.4 Anexos	50
2.4.1 Derivación de los estimadores MCO	50
2.4.2 Demostración del Teorema de Gauss-Markov	51
2.4.3 Algunas propiedades importantes de los residuos estimados.	53
3 Inferencia y análisis de regresión	57
3.1 Introducción	58
3.2 Pruebas individuales sobre los parámetros	67
3.3 El ajuste del modelo (Fit del modelo)	72
3.4 Pruebas conjuntas sobre los parámetros	77
3.5 Prueba de Wald y su relación con la prueba F	79
3.6 Práctica en R: Explicando los rendimientos de una acción	80
3.6.1 Pruebas conjuntas sobre los parámetros	88
3.7 Anexo: Demostración de la ecuación 3.8	92
4 Comparación de Modelos	95
4.1 Introducción	96
4.2 Comparación de modelos empleando medidas de bondad de ajuste	96
4.3 Comparación de modelos empleando inferencia	98
4.3.1 Modelos anidados	98

4.3.2	Modelos no anidados	98
4.4	Práctica en R: Escogiendo el mejor modelo	100
4.4.1	Medidas de bondad de ajuste	103
4.4.2	Pruebas estadísticas	104
II	Extendiendo el modelo clásico de regresión múltiple	109
5	Variables dummy	111
5.1	Introducción	112
5.2	Usos de las variables dummy	117
5.2.1	Caso I. La función es la misma	117
5.2.2	Caso II. Cambio en intercepto.	118
5.2.3	Caso III. Cambio en pendiente	119
5.2.4	Caso IV. Cambio en intercepto y pendiente	120
5.2.5	Para tener en cuenta	121
5.3	Práctica en R	121
5.3.1	Creando variables dummy con paquetes de R	127
6	Selección automática de modelos	131
6.1	Introducción	132
6.2	Empleando “fuerza bruta”	134
6.3	Empleando estrategias inteligentes de detección de un mejor modelo 140	140
6.3.1	Regresión paso a paso (stepwise)	140
6.3.2	<i>Stepwise forward regression</i>	141

6.3.3	<i>Stepwise backward regression</i>	149
6.3.4	Combinando <i>forward</i> y <i>backward</i> (<i>step regression</i>)	152
6.4	Pongamos todo junto	155
6.4.1	Eliminando automáticamente variables no significativas	156
6.4.2	Comparación de modelos	160
6.5	Comentarios finales	163
7	Primer caso de negocio	165
7.1	Introducción	166
7.2	La pregunta de negocio	166
7.3	El plan	168
7.4	Detección de posibles modelos	168
7.4.1	Stepwise forward	169
7.4.2	Stepwise backward	170
7.4.3	Combinando forward y backward	171
7.5	Comparación de modelos	172
7.6	Identificación de la variable más importante	176
7.6.1	Coeficientes estandarizados	176
7.6.2	Aporte relativo de cada variable empleando el R cuadrado	179
7.7	Generación de las recomendaciones	181
7.8	Generación de visualizaciones de los resultados	182
7.8.1	Comentarios Finales	189

III Problemas econométricos	191
8 Multicolinealidad	193
8.1 Introducción	194
8.2 Los diferentes grados de multicolinealidad	196
8.2.1 Multicolinealidad perfecta	197
8.2.2 Multicolinealidad no perfecta	200
8.3 Pruebas para la detección de multicolinealidad	201
8.3.1 Factor de Inflación de Varianza (VIF)	202
8.3.2 Prueba de Belsley, Kuh y Welsh (1980)	202
8.4 Soluciones de la multicolinealidad (¡Si se necesitan!)	202
8.4.1 Regresión de Ridge	203
8.4.2 Componentes principales	203
8.4.3 Remover variables con alto VIF	204
8.5 Práctica en R	205
8.5.1 Pruebas de multicolinealidad	207
8.5.2 Solución del problema removiendo variables con alto VIF	208
8.5.3 Solución del problema empleando la regresión de Ridge	210
8.5.4 Solución del problema empleando componentes principales	212
8.6 Comentarios finales	216
8.7 Anexos	218
8.7.1 Demostración de la insesgadez del estimador MCO con variables explicativas aleatorias	218
8.7.2 Demostración de la eficiencia del estimador MCO con variables explicativas aleatorias	218

8.7.3	Demostración del sesgo del estimador de la regresión de Ridge	221
9	Heteroscedasticidad	223
9.1	Introducción	224
9.2	Pruebas para la detección de heteroscedasticidad	230
9.2.1	Prueba de Breusch-Pagan	232
9.2.2	Prueba de White	233
9.3	Solución a la heteroscedasticidad	234
9.3.1	Estimación consistente en presencia de heteroscedasticidad de los errores estándar.	
	234	
9.4	Práctica en R	236
9.4.1	Análisis gráfico de los residuos	236
9.4.2	Pruebas de heteroscedasticidad	238
9.4.3	Solución al problema de heteroscedasticidad con HC	242
9.5	Anexos	247
9.5.1	Demostración de la insesgadez de los estimadores en presencia de heteroscedasticidad	247
9.5.2	Demostración del sesgo de la matriz de varianzas y covarianzas en presencia de heteroscedasticidad	247
9.5.3	Solución por mínimos cuadrados ponderados	248
10	Primer caso de negocio (actualizado)	251
10.1	Introducción	252
10.2	El plan	252

10.3 Detección de posibles modelos	253
10.3.1 Stepwise forward	254
10.3.2 Stepwise backward	262
10.3.3 Combinando forward y backward	264
10.4 Comparación de modelos	266
10.5 Identificación de la variable más importante	269
10.6 Generación de visualizaciones de los resultados	270
10.6.1 Comentarios Finales	271
11 Autocorrelación	273
11.1 Introducción	274
11.2 Pruebas para la detección de autocorrelación	281
11.2.1 Prueba de Rachas (Runs test)	282
11.2.2 Prueba de Durbin-Watson	283
11.2.3 Prueba h de Durbin	284
11.2.4 Prueba de Box-Pierce y Ljung-Box	285
11.2.5 Prueba de Breusch-Godfrey	286
11.3 Solución a la autocorrelación	287
11.3.1 Estimación Consistente en presencia de Autocorrelación de los errores estándar.	
287	
11.4 Práctica en R: Explicando los rendimientos de una acción (continuación)	289
11.4.1 Construcción de la base de datos	289
11.4.2 Residuales del modelo y análisis gráfico de los residuales	292
11.4.3 Pruebas de Autocorrelación	293

11.4.4	Prueba de Rachas	294
11.4.5	Prueba de Durbin-Watson	295
11.4.6	Prueba de Box-Pierce y Ljung-Box	297
11.4.7	Prueba de Breusch-Godfrey	299
11.4.8	Solución al problema de autocorrelación con H.A.C.	301
11.5	Anexos	306
11.5.1	Demostración de la insesgadez de los estimadores en presencia de autocorrelación	306
11.5.2	Sesgo de la matriz de varianzas y covarianzas en presencia de autocorrelación	307
11.5.3	Límites de la prueba de Rachas para muestras pequeñas	308
11.5.4	Solución por el método de diferencias generalizadas	309
IV	Modelo de regresión y la analítica predictiva.	311
12	Predicciones con datos de corte transversal.	313
12.1	Introducción	314
12.2	Estrategias para la validación cruzada de modelos	315
12.2.1	Método de retención	316
12.2.2	Método LOOCV	316
12.2.3	Método de k iteraciones	317
12.3	Métricas para medir la precisión de las predicciones	318
12.4	Validación cruzada en R	320
12.4.1	Método de retención	322
12.4.2	Método LOOCV	324
12.4.3	Método de k iteraciones	325

12.5	Intervalo de confianza para la predicción	326
12.6	Comentarios finales	328
12.7	Anexo: Método de Bootstrapping para construcción de intervalos de confianza para las predicciones	329
13	Segundo caso de negocio	333
13.1	Introducción	334
13.2	La pregunta de negocio	334
13.3	El plan	338
13.4	Detección de posibles modelos	338
13.4.1	Stepwise forward	339
13.4.2	Stepwise backward	343
13.4.3	Combinando forward y backward	345
13.5	Selección del mejor modelo para predecir	346
13.6	Predicciones (escenarios)	349
V	Apéndices: Conceptos básicos álgebra matricial y estadística	353
14	Elementos de álgebra matricial	355
14.1	Introducción	356
14.2	Matriz triangular y diagonal	357
14.3	Adición, multiplicación por un escalar y multiplicación de matrices	358
14.4	La matriz identidad y la matriz de ceros.	362

14.5	La transpuesta de una matriz y la matriz simétrica.	362
14.6	Matriz idempotente y matrices ortogonales	363
14.7	Combinaciones lineales de vectores e independencia lineal	364
14.8	La Traza y el rango de una matriz	365
14.9	Determinante de una matriz	366
14.10	Valores propios de una matriz	368
14.11	La Matriz inversa	368
14.12	Elementos de cálculo matricial	371
14.13	Resultados especiales para el modelo de regresión múltiple	372
14.14	Empleando R para hacer operaciones matriciales	374
15	Elementos de Estadística	379
15.1	Introducción	380
15.2	Variables, vectores y matrices aleatorias	380
15.3	Distribución de probabilidad	381
15.4	Valor esperado de una variable aleatoria	382
15.5	Independencia (estadística) lineal	383
15.6	Varianza y momentos alrededor de la media de una variable aleatoria	383
15.7	Covarianza y Correlación entre dos variables aleatorias	386
15.8	Esperanza y Varianza de vectores aleatorios.	388
15.9	Estimadores puntuales y sus propiedades deseadas	390

Índice de figuras

1.1	Tarea de clustering	11
1.2	Tarea de clasificación	13
1.3	Tarea de regresión	14
1.4	Tipos de analítica	16
1.5	Ruta metodológica tradicional de la econometría	18
1.6	Ruta metodológica de los científicos de datos	20
2.1	Parte no estocástica del modelo lineal 2.7 con $\beta_0 = -3$, $\beta_1 = 2$ y $\beta_2 = 4$	27
2.2	Parte no estocástica del modelo lineal 2.7 con $\beta_0 = -3$, $\beta_1 = 2$ y $\beta_2 = 4$ con ejes rotados.	27
2.3	Parte no estocástica del modelo 2.8 con $\alpha_1 = 1$, $\alpha_2 = 2$ y $\alpha_3 = 4$	28
2.4	Parte no estocástica del modelo 2.8 reparametrizado y con $\alpha_1 = 1$, $\alpha_2 = 2$ y $\alpha_3 = 4$. .	29
2.5	Ejemplo de una muestra observada para dos variables (y y x)	35
2.6	Posibles líneas que se pueden trazar para una muestra observada de y y x	36
2.7	Ejemplo de la linea trazada con el método MCO para una muestra observada para dos variables (y y x)	37

2.8	Ejemplo de la suma de los cuadrados de los errores que implica el método MCO	38
2.9	Relación entre todas las variables de la base de datos	45
2.10	Relación entre las cantidades demandadas de azúcar y el precio	45
3.1	Relación entre los datos simulados	60
3.2	Histograma de los coeficientes estimados (en azul se presenta la media de los coeficientes estimados)	63
3.3	Pendientes estimadas para la primera muestra y sus intervalos de confianza	65
3.4	Distribución muestral de la media	66
3.5	Diferencia entre la distribución estándar normal z y la t con 1 grado de libertad t_1 . . .	68
3.6	Diferencia entre la distribución estándar normal z y la t con 10 grado de libertad t_{10} . .	69
3.7	Diferencia entre la distribución estándar normal z y la t con 30 grado de libertad t_{30} . .	69
3.8	Relación entre el p-valor y el nivel de significancia.	72
3.9	Ejemplo de la variación total para una observación de la variable dependiente (y_i)	73
3.10	Ejemplo de la variación explicada por el modelo para una observación de la variable dependiente (y_i)	74
3.11	Ejemplo de la variación no explicada por el modelo para una observación de la variable dependiente (y_i)	74
5.1	Ventas de boletas por partido (Y_i) y presencia de una estrella en el equipo visitante (X_i). .	112
5.2	Esquema del modelo estimado por MCO.	113
5.3	Tasa de desempleo mensual en Colombia	115
5.4	Tasa de desempleo mensual observada y desestacionalizada en Colombia	116
5.5	Caso I. No hay cambio	118
5.6	Caso II. Cambio en el intercepto	119
5.7	Caso III. Cambio en pendiente	120
5.8	Caso IV. Cambio en intercepto y pendiente	121

5.9	Evolución temporal de las unidades de Sell-in de la crema dental	122
5.10	Sell-in mensual de crema dental (en unidades) e inversión en material promocional . . .	123
6.1	Representación de las estrategias <i>stepwise forward</i> y <i>stepwise backward</i>	141
7.1	Coeficientes estandarizados del Modelo 3	179
7.2	Aporte relativo de cada variable al R^2 del Modelo 3	181
8.1	El supuesto de no multicolinealidad	195
8.2	Multicolinealidad entre dos variables del modelo	195
8.3	Multicolinealidad entre tres variables del modelo	196
8.4	Grados de multicolinealidad	197
9.1	Muestras con errores homoscedasticidad	225
9.2	Muestras con errores heteroscedasticidad	225
9.3	Distribución muestral de los estimadores bajo homocedasticidad y heteroscedasticidad (en azul se presenta la media de los coeficientes estimados)	230
9.4	Posibles patrones de comportamiento de los residuales que sugieren heteroscedasticidad	231
11.1	Comportamiento en el tiempo del error simulado de un proceos AR(1) con $\rho = 0,8$. .	276
11.2	Error simulado para el periodo t versus el mismo error en el periodo anterior (error simulado de un proceos AR(1) con $\rho = 0,8$)	276
11.3	Comportamiento de la autocorrelación para diferentes rezagos del error simulado de un proceos AR(1) con $\rho = 0,8$	277
11.4	Comportamiento en el tiempo del error simulado de un proceos AR(1) con $\rho = 0,8$. .	278
11.5	Error simulado para el periodo t versus el mismo error en el periodo anterior (error simulado de un proceos AR(1) con $\rho = -0,8$)	278
11.6	Comportamiento de la autocorrelación para diferentes rezagos del error simulado de un proceos AR(1) con $\rho = -0,8$	279
11.7	Comportamiento en el tiempo del error simulado sin autocorrelación	280

11.8	Error simulado no autocorrelacionado para el periodo t versus el mismo error en el periodo anterior ($t - 1$)	280
12.1	Diagrama del Método de retención para la evaluación cruzada de modelos	316
12.2	Diagrama del Método de validación cruzada de k iteraciones para la evaluación de modelos	317
12.3	Diagrama del Método de validación cruzada de k iteraciones para la evaluación de modelos	318
12.4	De la muestra completa a la muestra de evaluación (Ejemplo del Método de retención)	319
14.1	Diagrama de multiplicación de matrices	360
14.2	Diagrama de la matriz menor asociada al elemento a_{ij} de la matriz A	366
14.3	Ejemplo de una matriz adjunta de orden tres	369
15.1	Tipos de Asimetría de diferentes distribuciones	385
15.2	Tipos de Curtosis de diferentes distribuciones	386
15.3	Ejemplos de diferentes valores de la correlación	388

Índice de cuadros

2.1	Terminología de la regresión múltiple	32
2.2	Modelo estimado por MCO	48
3.1	Hipótesis individuales sobre los coeficientes del modelo MCO: región de rechazo y reglas de decisión.	70
3.2	Criterios para rechazar H_0 empleando el valor p.	72
3.3	Tabla ANOVA.	76
3.4	Tabla ANOVA con prueba de significancia global	79
3.5	Equivalencia entre el valor p, el reporte de R y los asteriscos en la literatura.	86
3.6	Modelo estimado por MCO para la acción de SURA	87
3.7	Modelos estimados por MCO para la acción de SURA	90
4.1	Modelos estimados por MCO	102
4.2	Medidas de bondad de ajuste para los tres modelos estimados	103
5.1	Estimación del modelo para diferentes especificaciones de la dummy	126

6.1	Modelos seleccionados por las métricas tras emplear fuerza bruta	139
6.2	Modelo seleccionado por el valor p con el algoritmo stepwise forward	146
6.3	Modelo seleccionado por AIC con el algoritmo stepwise forward	148
6.4	Modelo seleccionado por el valor p y el AIC con el algoritmo stepwise backward . . .	151
6.5	Modelo seleccionado por el R^2 ajustado con el algoritmo combinado	153
6.6	Modelo seleccionado por el valor p y el AIC con el algoritmo combinado	155
6.7	Modelos construidos hasta ahora con diferentes algoritmos y criterios	156
6.8	Comparación de modelo 3 antes y después de la función remueve.no.sinifica().	158
6.9	Modelos 3 al 6 tras emplear la función remueve.no.sinifica().	159
6.10	Modelos 7 al 10 tras emplear la función remueve.no.sinifica().	160
6.11	Mejor modelo seleccionado.	162
7.1	Variables presentes en la base de datos del caso de negocio.	167
7.2	Modelos a estimar con los diferentes algoritmos	169
7.3	Modelo seleccionado el algoritmo stepwise forward	170
7.4	Modelo seleccionado el algoritmo stepwise backward	171
7.5	Modelo seleccionado el algoritmo stepwise forward y backward	172
7.6	Variables explicativas incluidas en cada uno de los modelos calculados	173
7.7	Valores p de las pruebas J (H_0 : modelo de la fila es mejor que el de la columna) . . .	173
7.8	Medidas de bondad de ajuste para los 7 modelos comparados	174
7.9	Mejor modelo	175
8.1	Modelo de brecha salarial estimado por MCO	206
9.1	Modelo estimado por MCO y corrección HC	244
10.1	Modelos a estimar con los diferentes algoritmos	253

10.2	Modelo 1 estimado por MCO y corrección HC	259
10.3	Modelo seleccionado el algoritmo stepwise forward con corrección HC3	261
10.4	Modelos seleccionados con el algoritmo stepwise backward con corrección HC3 . . .	263
10.5	Modelo seleccionado el algoritmo stepwise forward y backward con corrección HC3 .	265
10.6	Variables explicativas incluidas en cada uno de los modelos calculados (con corrección HC3)	266
10.7	Valores p de las pruebas J (H_0 : modelo de la fila es mejor que el de la columna (con corrección HC3))	267
10.8	Medidas de bondad de ajuste para los 4 modelos comprados	267
10.9	Modelo estimado por MCO y corrección HC	268
11.1	Relación del estadístico DW y los casos de autocorrelación	283
11.2	Estadístico DW en casos de Autocorrelación	284
11.3	Prueba de Box-Pierce de los errores para los primeros 20 rezagos	298
11.4	Prueba de Ljung-Box de los errores para los primeros 20 rezagos	299
11.5	Prueba de Breusch-Godfrey de los errores	300
11.6	Modelo estimado por MCO y correcciones H.A.C.	305
11.7	Límite inferior de la prueba de Rachas. Nivel de confianza del 95 %	308
11.8	Límite superior de la prueba de Rachas. Nivel de confianza del 95 %	309
12.1	Modelos a valorar su capacidad predictiva.	321
12.2	Métricas de precisión de los 3 modelos en muestra de evaluación	324
12.3	Métricas de precisión de los 3 modelos con LOOCV	325
12.4	Métricas de precisión de los 3 modelos con 5 iteraciones	326
13.1	Modelos a estimar con los diferentes algoritmos	339
13.2	Modelo 1 estimado por MCO y corrección HC	341
13.3	Modelo seleccionado el algoritmo stepwise forward con corrección HC3	342

13.4 Modelos seleccionados con el algoritmo stepwise backward con corrección HC3	344
13.5 Modelo seleccionado el algoritmo stepwise forward y backward con corrección HC3 . .	345
13.6 Modelos seleccionados por los diferentes algoritmos con corrección HC3	347
13.7 Metricas de precisión de los 6 modelos con 5 iteraciones	348
15.1 Modelo del ejercicio estimado por MCO	395
15.2 Modelos del ejercicio estimados por MCO	396
15.3 Medidas de bondad de ajuste para los cinco modelos estimados	397
15.4 Modelos del ejercicio estimados por MCO	398
15.5 Modelo del ejercicio estimado por MCO	401

1 . Introducción

Objetivos del capítulo

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

- Explicar en sus propias palabras las diferentes tareas de analítica
- Explicar en sus propias palabras los tipos de analítica
- Explicar en sus propias palabras las diferencias entre la aproximación econométrica tradicional y la de análisis de los científicos de datos

Los modelos de analítica y la disponibilidad de grandes volúmenes de información están transformando el mundo de los negocios y la forma como las organizaciones generan valor para sus grupos de interés. La gran cantidad de datos disponibles genera grandes oportunidades para emplear herramientas de *business analytics* y *Big Data* que permitan tanto optimizar los procesos actuales de la organización, como generar características diferenciadoras que acompañen los nuevos productos que lanza la organización.

Los datos se han convertido en un recurso muy importante para las organizaciones, y el *data analytics* se ha convertido en la forma como las organizaciones pueden monetizar ese recurso. El *business analytics* es el proceso científico de transformar datos en *insights* con el propósito de tomar mejores decisiones. En últimas, el *business analytics* empodera a la organización para el logro de su misión.

En ese proceso científico de transformar datos en conclusiones con el propósito de tomar mejores decisiones existen diferentes actividades que van desde la recolección de datos y su almacenamiento hasta la toma de la decisión; pasando por la extracción, limpieza y preparación de los datos, su exploración y visualización y el modelado o experimentación o predicción o lo que sea que requiera para responder la pregunta de negocio planteada. Estas actividades no son desarrolladas por una sola persona. Normalmente existe un equipo con profesionales calificados que tienen diferentes competencias y roles en este proceso. En estos equipos está como nodo central el científico de datos quien estima y entrena modelos estadísticos y de inteligencia artificial o diseña experimentos para resolver las preguntas de negocio planteadas.

Material multimedia: roles en la analítica

Escanea el siguiente código o visita el siguiente enlace para ver un video sobre los tipos de analítica.



Enlace: <https://youtu.be/rhLWa-vOxyU>

El científico de datos exitoso necesita tener en su caja de herramientas diferentes aproximaciones estadísticas y de inteligencia artificial para poder emplear la herramienta adecuada para responder determinada pregunta de negocio. Este libro se centra en la tarea de regresión (que se explicará más adelante) y en una herramienta estadística muy potente y flexible: el modelo clásico de regresión múltiple.

Antes de entrar en materia con el modelo clásico de regresión, en este capítulo estudiaremos las diferentes tareas de la analítica, los tipos de analítica y la diferencia entre la aproximación tradicional de las ciencias y la ciencia de datos al momento de usar el modelos de regresión.

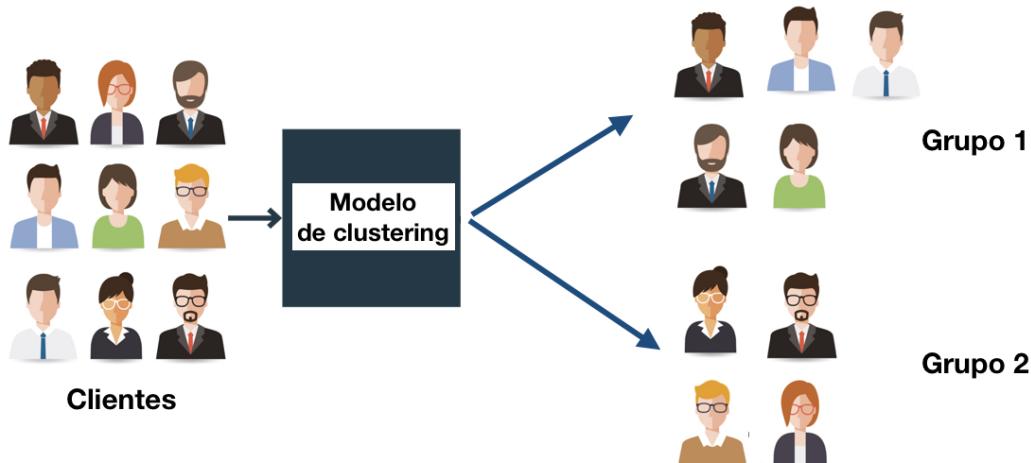
1.1 Tareas del científico de datos

El científico de datos en sus labores cotidianas se encuentra siempre en la necesidad de escoger la mejor herramienta para responder la pregunta de negocio planteada. Las respuestas a las preguntas de negocio implican diferentes tareas que se pueden clasificar en uno o varios de los siguientes tipos:

- *Resumir*
- *Visualizar*
- *Clusterizar (Agrupar)*
- *Clasificar*
- *Detectar excepciones*
- *Asociar*
- *Estimar regresiones*
- *Pronosticar*

Resumir implica simplificar la representación de los datos para generar información. La *Visualización* facilita la comprensión y el descubrimiento de los datos por medio de gráficos. El *Clustering* parte de una muestra para encontrar grupos de elementos similares. Por ejemplo, en la Figura 1.1 se presenta un conjunto de clientes que por medio de un modelo de clustering es distribuido en dos grupos de acuerdo con ciertas características.

Figura 1.1. Tarea de clustering



Para realizar las diferentes tareas de analítica empleamos modelos o algoritmos¹. En el ejemplo de la Figura 1.1 el modelo de clustering encuentra dos grupos: los que tienen gafas y los que no. Este tipo de modelo no tiene una variable a explicar e implica que el modelo o algoritmo aprenda sobre la estructura de los datos. A este tipo de algoritmos se les conoce como modelos de aprendizaje no supervisado.

Material multimedia: tarea de clustering

Escanea el siguiente código o visita el siguiente enlace para ver un video sobre la tarea de clustering.

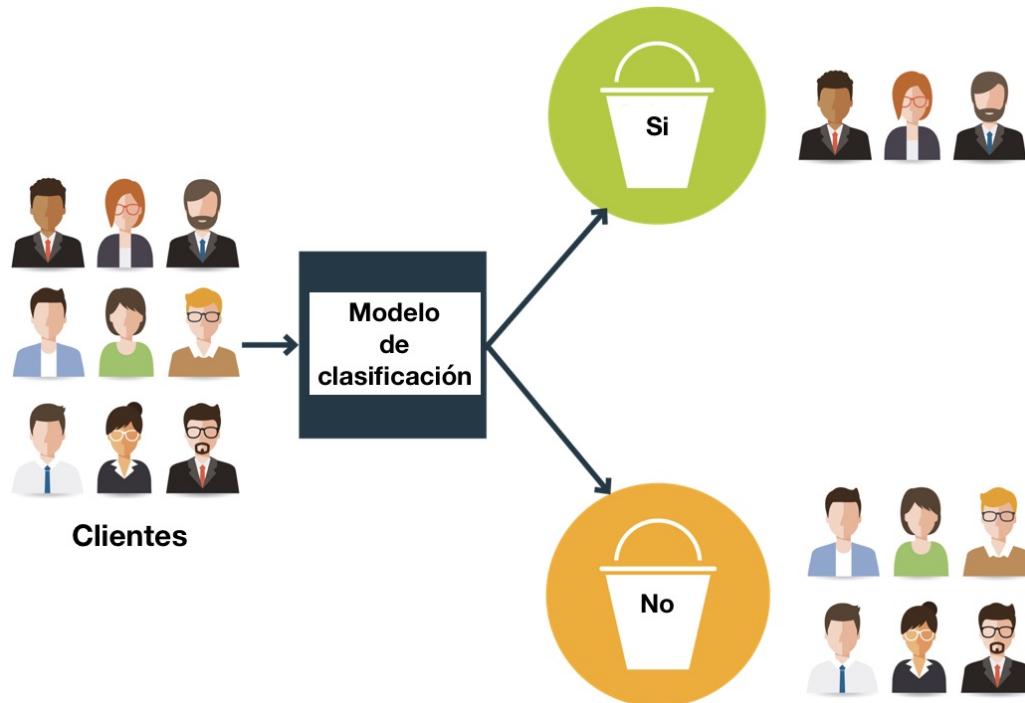


Enlace: <https://youtu.be/z0LX3sBSuXg>

La tarea de *Clasificación* tiene como finalidad predecir la categoría de un individuo. Por ejemplo, en algunas situaciones se deseará determinar si un nuevo cliente comprará o no nuestro producto. En este caso las categorías son compra o no compra. Otro tipo de preguntas que puede resolver esta tarea son: ¿se irá el cliente?, ¿pagará el crédito? y ¿será el individuo un buen *match* para la posición? En la Figura 1.2 se presenta una representación gráfica de esta tarea. Para esta tarea se emplean modelos o algoritmos que se estiman² empleando una muestra de individuos para los cuáles se tiene información de sus características y una variable dependiente que recoge si el individuo pertenece o no a una categoría. En este orden de ideas, los modelos que se emplean para esta tarea son modelos que intentan entender la relación entre unas variables y una variable categórica; relación que ya ocurrió en algún periodo. Este tipo de modelos se les conoce como algoritmos de aprendizaje supervisado, pues al modelo se le debe “enseñar” a qué categoría pertenece cada individuo.

¹Los modelos o algoritmos en algunos casos pueden ser útiles para hacer mas de una tarea, como veremos mas adelante.

²En el mundo de la estadística se emplea la expresión “estimar un modelo” para la construcción de un modelo a partir de una muestra. Por otro lado, en el mundo del inteligencia artificial se emplea la expresión “entrenar un modelo”.

Figura 1.2. Tarea de clasificación

Fuente: Elaboración propia

Material multimedia: tarea de clasificación

Escanea el siguiente código o visita el siguiente enlace para ver un video sobre la tarea de clasificación.



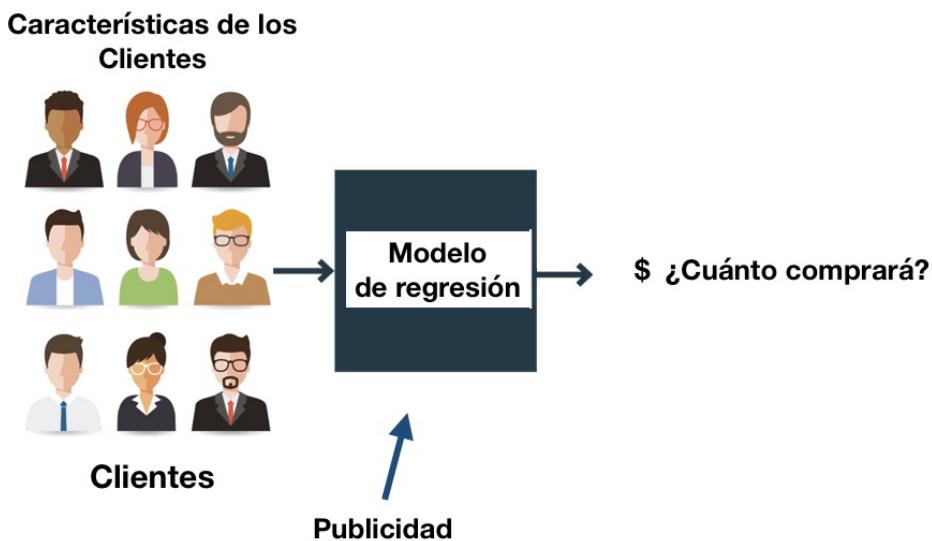
Enlace: <https://youtu.be/0K7ryP0uKGo>

La tarea de *Detección de excepciones* tiene como objetivo encontrar individuos con características o comportamiento diferentes. Esta tarea emplea modelos de aprendizaje no supervisado. La tarea de encontrar *Asociaciones* busca reglas de co-ocurrencia de productos en diferentes canastas. Es decir, busca encontrar cuáles productos son comprados regularmente al mismo tiempo que otros para poder sugerir composición de canastas. Estos modelos intentan encontrar la estructura de los datos sin la necesidad de enseñarle al algoritmo cuáles son las co-ocurrencias. Estos son modelos de aprendizaje no supervisado.

La tarea de *Estimar regresiones* implica encontrar relaciones entre muchas variables y una variable

cuantitativa de interés. Esto puede ser tanto para entender qué variables están asociadas a un fenómeno, como para simular el comportamiento en diferentes escenarios. En la Figura 1.3 se presenta un ejemplo en el que se tienen las características de diferentes clientes y variables como el gasto en publicidad (estas son las variables independientes) y el modelo de regresión determina la relación de estas variables con el monto (en dinero) de las compras (variable dependiente). Estos modelos son considerados modelos de aprendizaje supervisado.

Figura 1.3. Tarea de regresión



Fuente: Elaboración propia

La tarea de regresión se puede desarrollar empleando algoritmos de inteligencia artificial o modelos estadísticos. Los modelos estadísticos empleados para la estimación de las relaciones entre una variable dependiente y una o más variables independientes pueden ser clasificados en dos grandes categorías: los lineales y los no lineales. Los modelos de regresión que asumen una relación lineal entre las variables independientes y la dependiente corresponden al modelo clásico de regresión múltiple. Este tipo de modelo, aunque a primera vista parecería muy restrictivo el supuesto del comportamiento lineal, puede ser muy flexible como lo veremos a lo largo de este libro.

Finalmente, la tarea de generar *Pronósticos* implica predecir el comportamiento futuro de una variable cuantitativa³. Para esta tarea se emplean los patrones de comportamiento pasados para extrapolarlas al futuro. Estos modelos son considerados modelos de aprendizaje supervisado.

³Los modelos de regresión que se discuten a continuación pueden ser empleados para hacer la tarea de pronosticar

Este libro se concentra en el modelo de Clásico de Regresión Múltiple que permite realizar la tarea de *Estimar regresiones* y en algunos casos la tarea de *Pronosticar* si se emplea un componente temporal en el modelo. Este modelo fue desarrollado por la estadística, y los economistas han popularizado su uso para variables económicas y del mundo de los negocios dando origen a una disciplina conocida como la econometría. En este orden de ideas, en este libro nos concentraremos en este modelo econométrico. Estas herramientas de la econometría constituyen hoy un pilar importante de la caja de herramientas de los científicos de datos. En la sección 1.3 se discute la diferencia entre la aproximación de la econometría a resolver problemas empleando el modelo clásico de regresión múltiple y la ciencia de datos.

1.2 Tipos de analítica

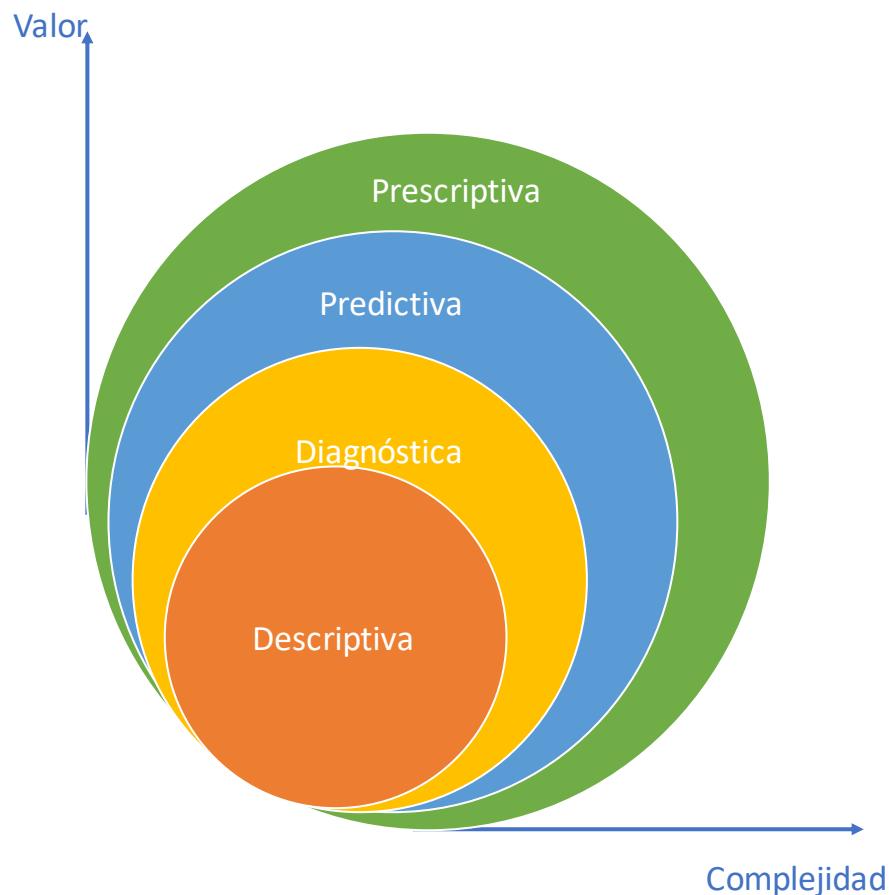
Antes de continuar es importante recordar que existen cuatro tipos de analítica: descriptiva, diagnóstica, predictiva y prescriptiva. (Ver Figura 1.4). Estos tipos de analítica engloban las tareas que discutimos anteriormente. No necesariamente un tipo de analítica es mejor que otra, cada una cumple una función diferente y responde a preguntas diferentes. Pero, es claro que a medida que pasamos de la analítica descriptiva a la prescriptiva se genera mayor valor a las organizaciones al tiempo que se está aumentando el grado de complejidad.

Material multimedia: tipos de analítica

Escanea el siguiente código o visita el siguiente enlace para ver un video sobre los tipos de analítica.



Enlace: https://youtu.be/ILoPGp6g_gI

Figura 1.4. Tipos de analítica

Fuente: Elaboración propia

La **analítica descriptiva** responde a la pregunta **¿qué está pasando en mi negocio?** Para esto emplea las bases de datos de la compañía, o las disponibles públicamente⁴ y emplea la estadística descriptiva para explorar los datos, resumir información en reportes y visualizaciones que sean útiles para entender qué ha pasado y qué está pasando con el negocio. Este tipo de analítica es la más utilizada en las empresas y es la que requiere menor grado de sofisticación técnica para aplicarlas.

La **analítica diagnóstica** típicamente quiere responder la pregunta de **por qué está pasando lo que está pasando en el negocio**. Este tipo de análisis tiene la capacidad de encontrar las causas de los problemas, la raíz de las relaciones y es capaz de aislar los efectos de diferentes fenómenos. El modelo de regresión que estudiaremos en este documento permite encontrar relaciones y aislar efectos.

La **analítica predictiva** busca responder la pregunta: **¿qué es posible que ocurra?** Para esto se

⁴Para ver una introducción rápida al tipo de datos que se emplean en el *business analytics* ver el video disponible en el siguiente enlace: https://youtu.be/2OxY2UTI_Bs.

apoya en datos históricos; es decir, con los datos que ya se tiene, se predicen comportamientos que aún no se conocen. En otras palabras, con datos ya existentes se predicen los datos que aún no se tienen o no han ocurrido. La predicción implica estimar los resultados de los datos no vistos. La creación de pronósticos (forecasting en inglés) es un área de la predicción en la que se realizan conjeturas sobre el futuro, basándonos en datos de series de tiempo⁵. La única diferencia entre la predicción y los pronósticos es que en esta última se considera la dimensión temporal. La analítica predictiva tiene como intención generar predicciones de variables cuantitativas.

El modelo de regresión múltiple puede emplearse tanto para hacer predicciones como pronósticos. Por ejemplo, si se emplea una muestra de muchos individuos en el mismo periodo (muestra de corte trasversal) para encontrar las variables asociadas a la cantidad de unidades que compra un cliente, el modelo podría ser empleado para responder la pregunta ¿cuánto compraría un nuevo cliente con determinadas características? En el Capítulo 12 se discutirá el uso del modelo de regresión múltiple para hacer predicciones. Si el modelo de regresión múltiple es estimado con series de tiempo (se observa uno o varios objetos de estudio periodo tras periodo), entonces podrá ser empleado para hacer pronósticos. Por ejemplo, si se cuenta con las ventas mensuales para muchos periodos y el modelo encuentra que variables están asociadas a estas ventas mes tras mes, el modelo podría responder la pregunta que ocurrirá en el futuro con las ventas. Con este tipo de modelos podemos responder preguntas como: ¿cuánto es lo más probable que venda un producto el próximo año?

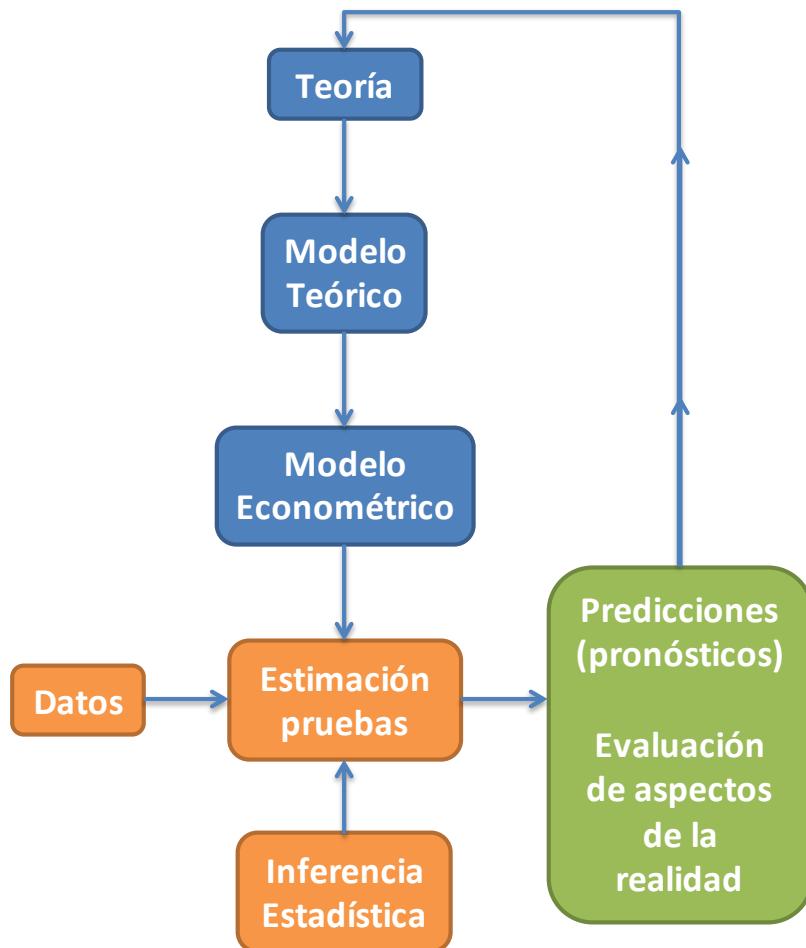
Finalmente, la **analítica prescriptiva** busca responder la pregunta: **¿qué necesito hacer?** Este tipo de analítica hace énfasis en el uso de técnicas de optimización para identificar cuál es la mejor alternativa para minimizar o maximizar algún objetivo así como modelos que permitan simular situaciones. El modelo de regresión múltiple pueden hacer parte de un análisis prescriptivo como se discutirá más adelante.

1.3 Consideraciones sobre la aproximación econométrica y los científicos de datos

La econometría es una disciplina que unifica las matemáticas, la estadística y la teoría económica con el objetivo de entender cuantitativamente las relaciones económicas (Frisch, 1933). Ésta 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 aplicación del modelo de regresión múltiple a problemas de los negocios y la economía se conoce se reconoce como una aproximación econométrica. Pero hay que tener cuidado y distinguir claramente las herramientas de la econometría y la aproximación econométrica a los problemas, pues como científico de datos emplearemos las herramientas de la econometría pero no la aproximación econométrica para hacer validar los modelos teóricos de la ciencia económica.

⁵Una serie de tiempo es una secuencia de datos que se producen en orden sucesivo a lo largo de un periodo de tiempo. Un ejemplo, es el precio del dólar para los últimos 100 días es una serie de tiempo. Otro ejemplo de datos de serie de tiempo son los datos de ventas mensuales, el precio mensual del producto y el gasto en publicidad mensual para los últimos 60 meses.

Figura 1.5. Ruta metodológica tradicional de la econometría



Fuente: Adaptado a partir de Spanos (1986).

Para lograr su objetivo, la econometría emplea una ruta metodológica que parte de un modelo teórico; es decir, de una formulación matemática de alguna teoría, como se muestra en la Figura 1.5.

Luego, el economista transforma este modelo teórico en un modelo estadístico o económico. El modelo económico será construido con variables medibles que suponemos representan adecuadamente los conceptos teóricos y se le añade un término de error.

En un siguiente paso, el economista estima el modelo usando métodos estadísticos que implican supuestos explícitos e implícitos y datos reales. Si alguno de los supuestos del modelo no se cumple, el economista debe corregir el modelo al especificar adecuadamente el término de error o solucionando el problema que está generando la violación del supuesto. Esto se hace para que la herramienta estadística (económica) se comporte bien y permita tener confiabilidad en los

resultados⁶.

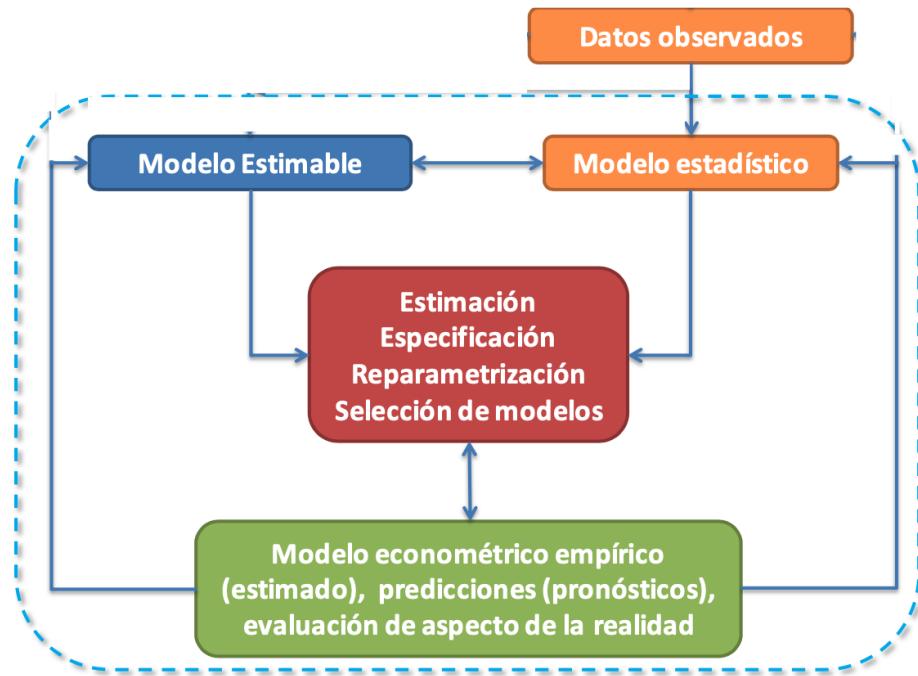
Una vez el economista tiene el modelo estadístico adecuado, procede a comprobar estadísticamente el cumplimiento o no de las restricciones planteadas por la teoría. Si el modelo estimado refleja el comportamiento real adecuadamente, el economista pues emplea las estimaciones para hacer predicciones y evaluaciones de los aspectos de la realidad relevantes; en caso contrario, el economista deberá empezar el proceso nuevamente, buscando un modelo teórico que explique adecuadamente las relaciones bajo estudio.

Sin embargo, la aproximación del científico de datos es diferente, si bien la herramienta (el modelo de clásico de regresión múltiple) es la misma. En la ciencia de datos no se cuenta con una teoría detrás, pero sí con una cantidad relativamente grande de variables candidatas a ser parte del mecanismo que realmente genera los datos que observamos (el *Data Generating Process* o DGP⁷). Es decir, tenemos datos y no es de interés probar un modelo teórico. En otras palabras, el científico de datos está interesado en encontrar patrones de comportamiento que permita responder preguntas de negocio y no realizar ciencia en su sentido estricto. Es decir, no es tarea del científico de datos validar modelos teóricos y no tiene como motivación del ejercicio empírico comprobar el cumplimiento o no de una teoría que presente hipótesis. La segunda ruta metodológica que es la que adopta el científico de datos, y menos tradicional en la econometría, implica partir de los datos para encontrar qué variable explican la variable dependiente. En otras palabras, en esta aproximación se interpreta el modelo econométrico como una aproximación al verdadero DGP.

⁶En el Capítulo 2 discutiremos que significa que los resultados sean confiables y cuáles son los supuestos de esta herramienta.

⁷El DGP es la “fórmula” o ley de movimiento que muestra la relación que existe entre las variables explicativas y la variable dependiente. En el siguiente capítulo discutiremos el DGP en detalle.

Figura 1.6. Ruta metodológica de los científicos de datos



Fuente: Adaptado a partir de Spanos (1986).

De acuerdo a esta segunda ruta metodológica, que se ilustra en la Figura 1.6, el modelo económico usa variables aleatorias disponibles en las bases de datos para especificar una descripción de un mecanismo que genera los datos⁸.

En la práctica, el científico de datos empleará todas las variables disponibles para explicar la variable de interés y dejará que sean los datos quienes “hablen” para así detectar cuáles variables hacen parte o no del *DGP*.

No obstante, debido a que es imposible tener variables para todas las características de la realidad⁹, el modelo estimado por esta ruta metodológica será en todo caso una simplificación de la realidad. A lo largo de este libro, nos concentraremos en el área punteada de la Figura 1.6.

⁸Cuando el modelo teórico es el DGP y, por ello, el modelo económico difiere del teórico únicamente por razones puramente aleatorias, las dos metodologías coinciden.

⁹Es decir, todas las variables asociadas al mecanismo que genera los datos.

Parte I

El modelo clásico de regresión múltiple

The screenshot shows a LaTeX editor interface with several floating windows. One window displays a warning message about the use of \text{...} instead of \textbf{...}. Another window shows a table with data for two companies. The main code area contains definitions for matrices M and ApMatriz1, and an equation for a regression model. A large blue circle highlights the number 2, which is part of the section title '2. Modelo de regresión múltiple'.

```
\begin{table}
\caption{Resumen de ventas y utilidades}
\begin{array}{l|llll}
\hline
& \multicolumn{2}{c}{Ventas} & \multicolumn{2}{c}{Utilidades} \\
\hline
\text{Compañía} & \text{Q1} & \text{Q2} & \text{Q1} & \text{Q2} \\
\hline
\text{A} & 100 & 120 & 20 & 30 \\
\text{B} & 150 & 180 & 30 & 40 \\
\hline
\text{Total} & 250 & 300 & 50 & 70 \\
\hline
\end{array}

```

```
\begin{matrix}
M = \begin{pmatrix} \text{&} & \text{&} & \text{&} & \text{&} \\ \text{Empres} & \text{1} & \text{2} & \text{3} & \text{4} \\ \text{Ventas} & 100 & 150 & 250 & 300 \\ \text{Costos} & 60 & 80 & 120 & 180 \\ \text{Utilidades} & 40 & 70 & 130 & 200 \end{pmatrix} \\ 
ApMatriz1 = \begin{pmatrix} \text{&} & \text{&} & \text{&} & \text{&} \\ \text{Empres} & \text{1} & \text{2} & \text{3} & \text{4} \\ \text{Ventas} & 100 & 150 & 250 & 300 \\ \text{Costos} & 60 & 80 & 120 & 180 \\ \text{Utilidades} & 40 & 70 & 130 & 200 \end{pmatrix}
\end{matrix}
```

```
\begin{aligned}
\text{Generalmente, se obtiene la ecuaci\'on:} \\
(\text{\ref{ApMatriz1}}) \text{ se puede reescribir como} \\
\begin{aligned}
&\begin{pmatrix} \text{&} & \text{&} & \text{&} & \text{&} \\ \text{Empres} & \text{1} & \text{2} & \text{3} & \text{4} \\ \text{Ventas} & 100 & 150 & 250 & 300 \\ \text{Costos} & 60 & 80 & 120 & 180 \\ \text{Utilidades} & 40 & 70 & 130 & 200 \end{pmatrix} \\
&\begin{pmatrix} \text{&} & \text{&} & \text{&} & \text{&} \\ \text{Empres} & \text{1} & \text{2} & \text{3} & \text{4} \\ \text{Ventas} & 100 & 150 & 250 & 300 \\ \text{Costos} & 60 & 80 & 120 & 180 \\ \text{Utilidades} & 40 & 70 & 130 & 200 \end{pmatrix} \\
&\begin{pmatrix} \text{&} & \text{&} & \text{&} & \text{&} \\ \text{Empres} & \text{1} & \text{2} & \text{3} & \text{4} \\ \text{Ventas} & 100 & 150 & 250 & 300 \\ \text{Costos} & 60 & 80 & 120 & 180 \\ \text{Utilidades} & 40 & 70 & 130 & 200 \end{pmatrix}
\end{aligned}
\end{aligned}
```

2 . Modelo de regresión múltiple

Objetivos del capítulo

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

- Explicar en sus propias palabras cuáles son los supuestos del Teorema de Gauss-Markov.
 - Explicar en sus propias palabras cuáles son las propiedades de los estimadores MCO.
 - 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 porque el problema no se encuentra acotado. Así, típicamente el científico de datos 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, supongamos que contamos con un número reducido de variables explicativas.

Por ejemplo, supongamos que queremos responder la pregunta de negocio: ¿de qué dependen las cantidades compradas de mi producto estrella Q ? Y supongamos además que se cuenta con una base de datos con las siguientes posibles variables explicativas: el precio del producto estrella p_x , el precio de un producto idéntico de la competencia que puede sustituir nuestro producto p_{sust} , el precio de un precio de un producto que acompaña (o complementa) el consumo de nuestro producto p_{comp} y el nivel de actividad económica que puede ser una variable que aproxime el ingreso que tienen los consumidores para comparar mi producto I . Así, la tarea del científico de datos será construir un modelo que cumpla la tarea de regresión que le permita hacer analítica diagnóstica. Estas variables disponibles podrían permitir inicialmente emplear un modelo que permita representar las cantidades compradas en función de las variables explicativas disponibles. En otras palabras,

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

Naturalmente, no hay nada que permita saber *a priori* (antes de usar los datos) si realmente son todas estas 4 variables importantes para explicar a Q . En el Capítulo 3 se discutirá como determinar con los datos cuáles variables afectan y cuáles no a la variable dependiente.

Por otro lado, es importante determinar la forma de la función $Q_x()$. Por ejemplo, la forma funcional puede ser

$$Q_x(p_x, p_{comp}, p_{sust}, I) = \gamma p_x^{\alpha_0} + \frac{1}{p_{comp}^{\beta} + C} + \ln(\varphi p_{sust}) + \alpha_0 I \quad (2.2)$$

o

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

Aún más, el carácter de las relaciones funcionales expresadas en (2.2) y (2.3) es determinístico¹. Estas dos expresiones 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 a estimar². La inclusión de este término de error se justifica de diferentes formas. Por ejemplo,³

¹No incluyen una variable aleatoria.

²Noten 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.

³Para ver más justificaciones para la inclusión del término de error el lector puede consultar Gujarati y Porter (2011).

- 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 sí 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 enunciadas 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 determinística). 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.

En la práctica, los científicos de datos adoptan como estrategia tomar una primera aproximación a las relaciones funcionales por medio de relaciones lineales (como en la expresión (2.3)) o linealizables. En este libro 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 el subíndice i es cambiado por una t . En otras palabras típicamente se emplea el subíndice i para cuando se cuenta con datos de corte trasversal, dado que i representa a los individuos que están bajo estudio en un período determinado. Y cuando se emplean series de tiempo, el subíndice que se emplea es t que representa el período (el tiempo) de la observación.

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

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

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 lineal⁶.

Por ejemplo, el modelo

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

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 2.2.1), y el término de error es aditivo. Por otro lado, un modelo como

$$Y_i = \beta_0 + (X_{1i})^{\beta_1} + \beta_2 X_{2i} + \varepsilon_i \quad (2.6)$$

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 estadístico 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 2.1).

Ejemplo 2.1 Modelo (matemáticamente) lineal

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

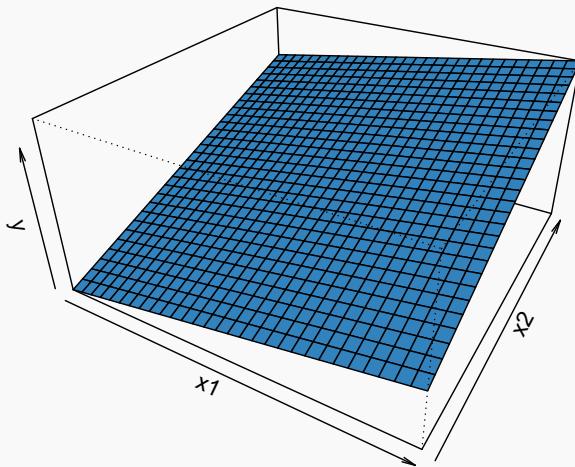
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad (2.7)$$

donde ε_i es un término aleatorio de error. Entendamos primero la naturaleza de esta relación funcional omitiendo el término aleatorio de error. En la Figura 2.1 se puede observar la función $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ con $\beta_0 = -3$, $\beta_1 = 2$ y $\beta_2 = 4$.

⁵A excepción de la varianza del término de error.

⁶En el Apéndice de Álgebra Matricial (sección 14.7) se puede refrescar este concepto.

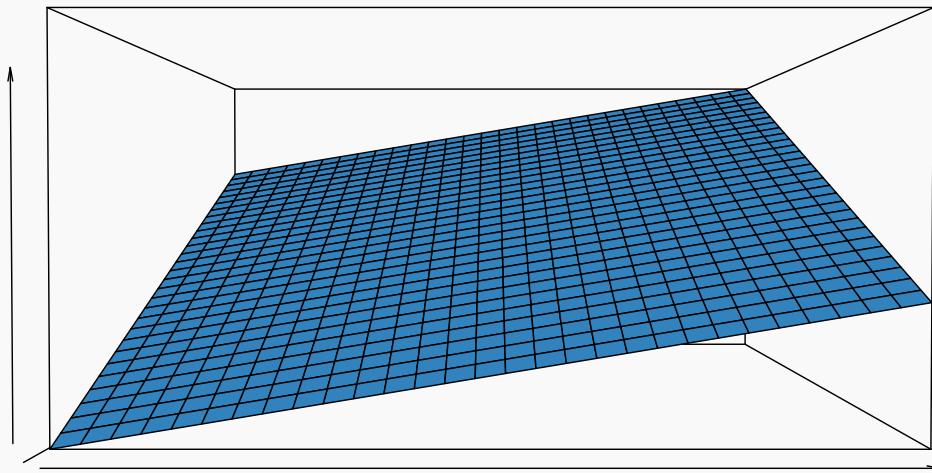
Figura 2.1. Parte no estocástica del modelo lineal 2.7 con $\beta_0 = -3$, $\beta_1 = 2$ y $\beta_2 = 4$



Fuente: Elaboración propia

Nota que $\frac{\partial Y}{\partial X_1} = \beta_1$; es decir, el coeficiente asociado a la variable independiente X_1 corresponde al cambio en la variable dependiente cuando la otra variable independiente se mantiene constante. En otras palabras, β_1 es la pendiente de la función con respecto al plano formado por los ejes de X_1 y Y . Esto lo podemos ver mas claramente si rotamos un poco los ejes (Ver Figura 2.2).

Figura 2.2. Parte no estocástica del modelo lineal 2.7 con $\beta_0 = -3$, $\beta_1 = 2$ y $\beta_2 = 4$ con ejes rotados.



Fuente: Elaboración propia

Similarmente, β_2 representa el cambio en la variable dependiente cuando la variable X_2 cambia en una unidad dejando X_1 constante ($\beta_2 = \frac{\partial Y_i}{\partial X_{2i}}$).

Ejemplo 2.2 Modelo intrínsecamente lineal

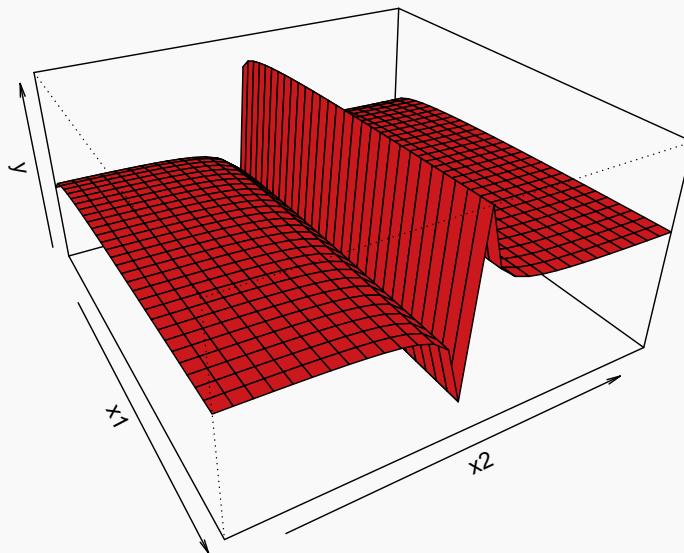
Suponga la siguiente relación entre una variable dependiente (Y_i) y dos variables explicativas (X_{1i}, X_{2i}):

$$Y_i = \alpha_1 + \alpha_2 \ln(X_{1i}) + \alpha_3 \left(\frac{1}{X_{2i}} \right) + \varepsilon_i, \quad (2.8)$$

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 estadísticamente lineal.

No obstante, el modelo no representa un plano (no es lineal desde el punto de vista matemático), como se puede observar en la Figura 2.3 (en la cual se ha omitido el término de error). 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.

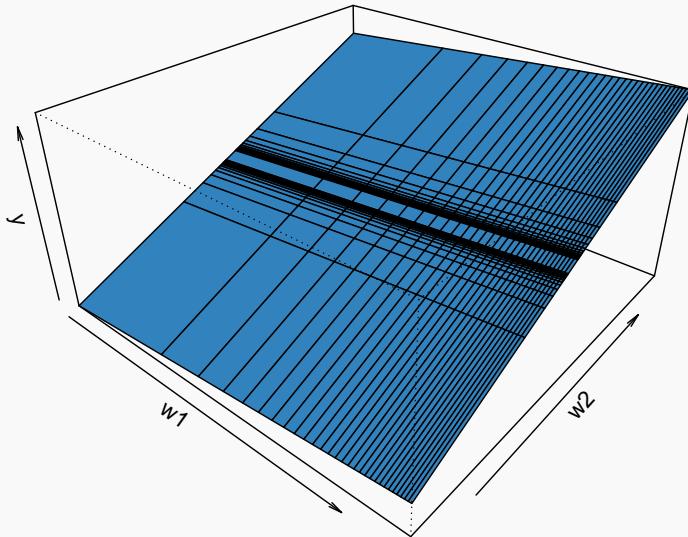
Figura 2.3. Parte no estocástica del modelo 2.8 con $\alpha_1 = 1$, $\alpha_2 = 2$ y $\alpha_3 = 4$



Fuente: Elaboración propia

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, remplazando estas dos nuevas variables en el modelo original, tenemos que $Y_i = \alpha_1 + \alpha_2 W_{1i} + \alpha_3 W_{2i} + \varepsilon_i$. La Figura 2.4 muestra esta nueva función, ignorando el término de error, en el espacio (Y, W_1, W_2).

Figura 2.4. Parte no estocástica del modelo 2.8 reparametrizado y con $\alpha_1 = 1$, $\alpha_2 = 2$ y $\alpha_3 = 4$



Fuente: Elaboración propia

Este modelo reparametrizado es un modelo matemáticamente lineal, representa un plano. Noten que el modelo 2.8 si es lineal desde el punto de vista estadístico 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 (2.8) con respecto a X_1 : $\frac{\partial Y_i}{\partial X_{1i}} = \frac{\alpha_2}{X_{1i}}$. Manipulando algebraicamente tenemos $\frac{\partial Y_i}{\partial X_{1i}} = \alpha_2$. Multiplicando a ambos lados por $\frac{1}{100}$, obtenemos $\frac{\partial Y_i}{\partial X_{1i} \times 100} = \frac{\alpha_2}{100}$. Es decir $\frac{\Delta Y_i}{\Delta \% X_{1i}} = \frac{\alpha_2}{100}$. 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 reparametrizaciones⁷ 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 de estimación de un modelo lineal.

⁷Una 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 Cobb-Douglas

Una función Cobb-Douglas se define como:

$$Y = \alpha_0^{\alpha_1} X_2^{\alpha_2} X_3^{\alpha_3} \quad (2.9)$$

donde Y corresponde a la variable dependiente y α_j para $j = 0, 1, 2, 3$ 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. Así, un modelo econométrico a partir de la función Cobb-Douglas sería:

$$Y_i = \alpha_0 X_{1i}^{\alpha_1} X_{2i}^{\alpha_2} X_{3i}^{\alpha_3} \varepsilon_i \quad (2.10)$$

donde^a $i = 1, 2, \dots, n$ 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:

$$\ln(Y_i) = \ln(\alpha_0) + \alpha_1 \ln(X_{1i}) + \alpha_2 \ln(X_{2i}) + \alpha_3 \ln(X_{3i}) + \ln(\varepsilon_i). \quad (2.11)$$

Definámos $\beta = \ln(\alpha_0)$ y $\mu_i = \ln(\varepsilon_i)$. Entonces la expresión anterior se puede reescribir de la siguiente forma:

$$\ln(Y_i) = \beta + \alpha_1 \ln(X_{1i}) + \alpha_2 \ln(X_{2i}) + \alpha_3 \ln(X_{3i}) + \mu_i \quad (2.12)$$

Un aspecto importante de este modelo es la interpretación singular de los coeficientes. En este caso, cada α_j (solo para $j = 1, 2, 3$) corresponde a la elasticidad de la variable dependiente con respecto a la variable explicativa j . Es decir el aumento porcentual en la variable dependiente dado un aumento en un 1% en la variable explicativa X_j . Para una demostración de esta afirmación se puede ver los ejercicios al final de este capítulo.

En otras palabras, en general cuando contamos con un modelo lineal con todas las variables expresadas en logaritmos, entonces las pendientes corresponderán a las elasticidades.

Por otro lado, el intercepto β corresponde al valor esperado del logaritmo del valor de Y cuando todos las variables explicativas son iguales a una unidad. En otras palabras, $\alpha_0 = e^\beta$. Así, el intercepto en este modelo tiene una interpretación difícil y en la mayoría de las aplicaciones carece de interpretación.

^aEs decir, el modelo es válido para cualquiera de las i observaciones.

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 dependiente. En general, un modelo lineal múltiple para el cual se cuenta con n observaciones y $k - 1$ variables explicativas está dado por:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \dots + \beta_k X_{k1} + \varepsilon_1 \\ y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \dots + \beta_k X_{k2} + \varepsilon_2 \\ &\vdots && \vdots && \vdots \\ y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \dots + \beta_k X_{kn} + \varepsilon_n \end{aligned} \quad (2.13)$$

Para este tipo de modelos, los datos pueden corresponder a dos tipos de estructura. Una estructura en la que se observan n individuos en el mismo período. A esta estructura se le conoce como datos de corte trasversal y es equivalente a tomar una foto de los n individuos. La segunda estructura posible es una serie de tiempo, en la que se observa un individuo (objeto de estudio) periodo tras periodo. En este caso es común emplear el subíndice t para denotar cada periodo que va desde el primero hasta el T . Para simplificar, en este libro siempre nos referiremos a n como el tamaño de la muestra, ya sea que empleemos datos de series de tiempo o corte transversal. Es decir, en este libro tendremos que n representará lo mismo que T . Existe una tercera estructura de datos que combina las dos anteriores, muchos individuos que se siguen en el tiempo. Esta última estructura se conoce como datos de panel. Las técnicas que se presentan en este libro solo aplican para las dos primeras estructuras de datos, pero se pueden extraer fácilmente para la estimación de modelos con datos de panel.

Regresando a la notación, para simplificar y ahorrar espacio, escribiremos el modelo 2.13 de una forma más abreviada, de tal modo que sólo describamos la observación i -ésima del modelo. Es decir:

$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (2.14)$$

para $i = 1, 2, \dots, n$.

Otra forma de expresar el mismo modelo 2.14 de manera aún más abreviada es empleando matrices. Sean:

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

Entonces, el modelo 2.14 se puede expresar de forma matricial así:

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\varepsilon}_{n \times 1} \quad (2.15)$$

Cuadro 2.1: Terminología de la regresión múltiple

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

Fuente: Elaboración propia

2.2.1 Supuestos

Es importante estudiar en detalle el vector de errores ε . En general, esperaremos 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 cero⁸. Formalmente, asumiremos que:

$$E[\varepsilon] = 0 \quad (2.16)$$

para todo $i = 1, 2, \dots, n$, o en forma matricial,

$$E[\varepsilon_i] = \mathbf{0}.$$

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

$$E[\varepsilon_i \varepsilon_j] = E[\varepsilon_i] E[\varepsilon_j] = 0, \quad (2.17)$$

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:

$$\text{Var}[\varepsilon_i] = \sigma^2 \quad (2.18)$$

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

En resumen, asumiremos que el error cumple con las siguientes condiciones:

- media cero,
- independencia lineal entre los errores

⁸En caso que el modelo incluya un intercepto, cualquier componente determinístico del error es capturado por el intercepto.

- varianza constante.

Estos supuestos se acostumbran resumir de diferentes formas; por ejemplo, se pueden resumir los tres supuestos expresando que los errores están linealmente independientemente distribuidos con media cero y varianza constante, denotado por

$$\varepsilon_i \sim l.i.d. (0, \sigma^2).$$

Otra forma de escribir estos supuestos de forma matricial es

$$\varepsilon \sim (0, \sigma^2 I_n),$$

donde

$$\begin{aligned} Var[\varepsilon] &= \begin{bmatrix} Var[\varepsilon_1] & Cov(\varepsilon_1, \varepsilon_2) & \cdots & Cov(\varepsilon_1, \varepsilon_n) \\ Cov(\varepsilon_2, \varepsilon_1) & Var[\varepsilon_2] & \cdots & Cov(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\varepsilon_n, \varepsilon_1) & Cov(\varepsilon_n, \varepsilon_2) & \cdots & Var[\varepsilon_n] \end{bmatrix} \\ Var[\varepsilon] &= \begin{bmatrix} \sigma^2 & 0 \\ 0 & \ddots \\ 0 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = \sigma^2 I_n \end{aligned} \quad (2.19)$$

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*. A este supuesto también se le conoce como el supuesto de no multicolinealidad (o no colinealidad) de las variables explicativas

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 determinísticas⁹(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 \mathbf{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 dependiente como una variable explicativa.

⁹El 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.

Supuestos del modelo de regresión múltiple

En resumen, los supuestos del modelo de regresión múltiple son:

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.

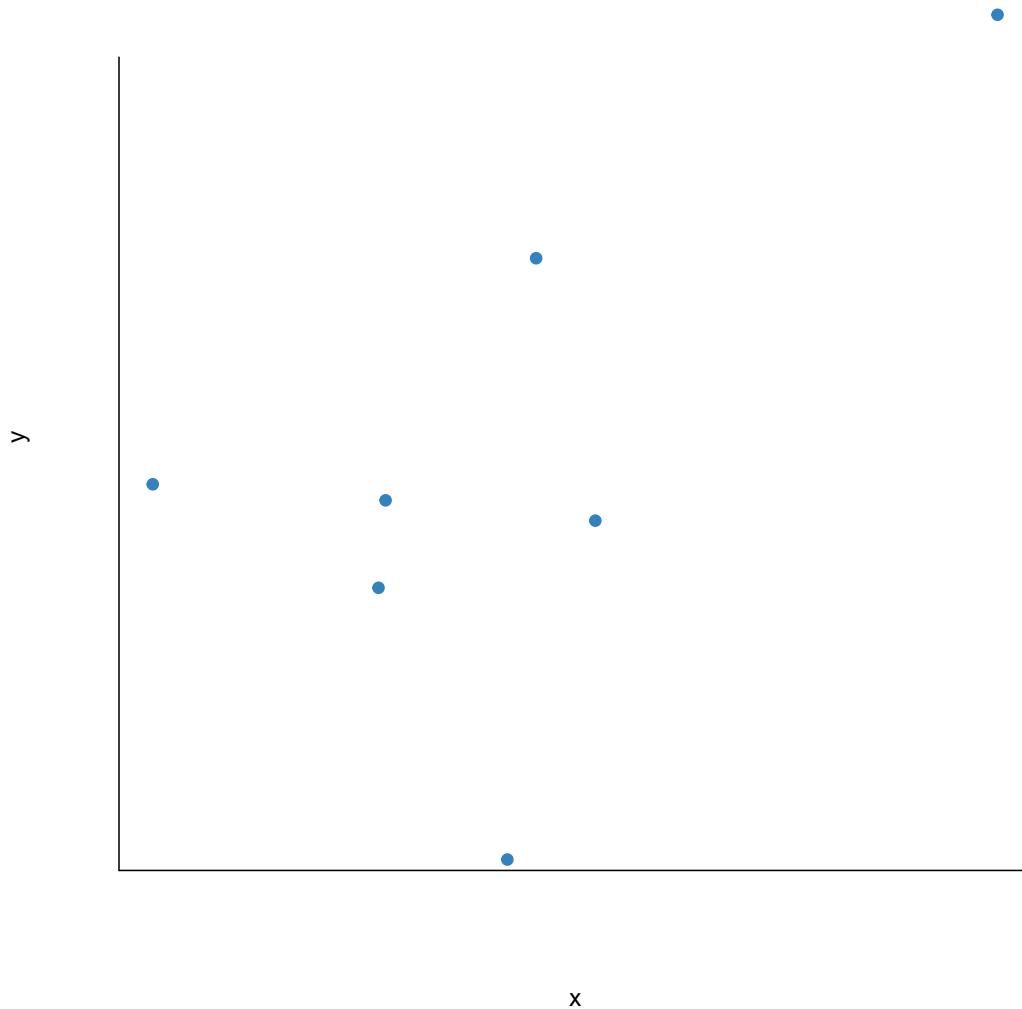
Es decir, $\varepsilon_i \sim i.i.d (0, \sigma^2)$ o $\varepsilon_{n \times 1} \sim (\mathbf{0}_{n \times 1}, \sigma^2 \mathbf{I}_n)$.

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 (vector 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 estos¹⁰.

Consideremos un ejemplo muy sencillo como el que se presenta en la Figura 2.5 donde tenemos una muestra de $n = 7$. En este caso se recopiló información para la variable dependiente y y para una variable explicativa x . Cada punto azul representa una observación (un par de y y x).

¹⁰Si el lector no se encuentra familiarizado o requiere un repaso con el lenguaje matricial y vectorial se recomienda revisar el Apéndice de álgebra matricial que se presenta en el Capítulo 14 al final del libro. Así mismo, en el Capítulo 15 se presenta una revisión de los conceptos estadísticos necesarios para comprender este libro.

Figura 2.5. Ejemplo de una muestra observada para dos variables (y y x)

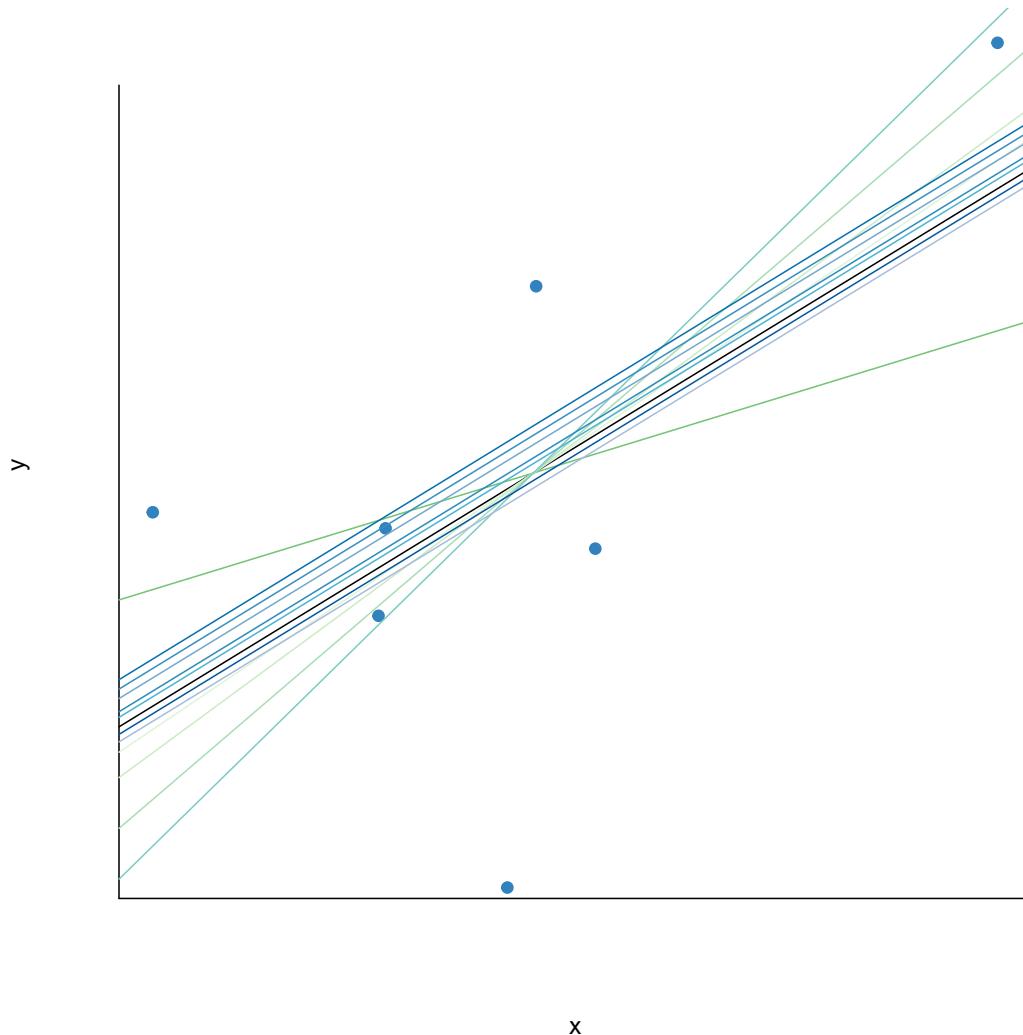
Fuente: Elaboración propia

En este caso, suponiendo la linealidad, el modelo estadístico con el que queremos describir la relación entre estas dos variables será:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i. \quad (2.20)$$

Así, el objetivo del científico de datos es encontrar el vector β que está formado por el intercepto (β_1) y la pendiente (β_2). En últimas encontrar el intercepto y la pendiente implica trazar una linea recta que aproxime todos los puntos de la Figura 2.5.

Existen muchas posibles rectas que podemos trazar. En la Figura 2.6 se muestran numerosas opciones. Cada línea corresponde a un conjunto “adivinado” (estimado) de intercepto y pendiente. La pregunta natural es ¿cuál de esas pendientes es la adecuada?.

Figura 2.6. Posibles líneas que se pueden trazar para una muestra observada de y y x 

Fuente: Elaboración propia

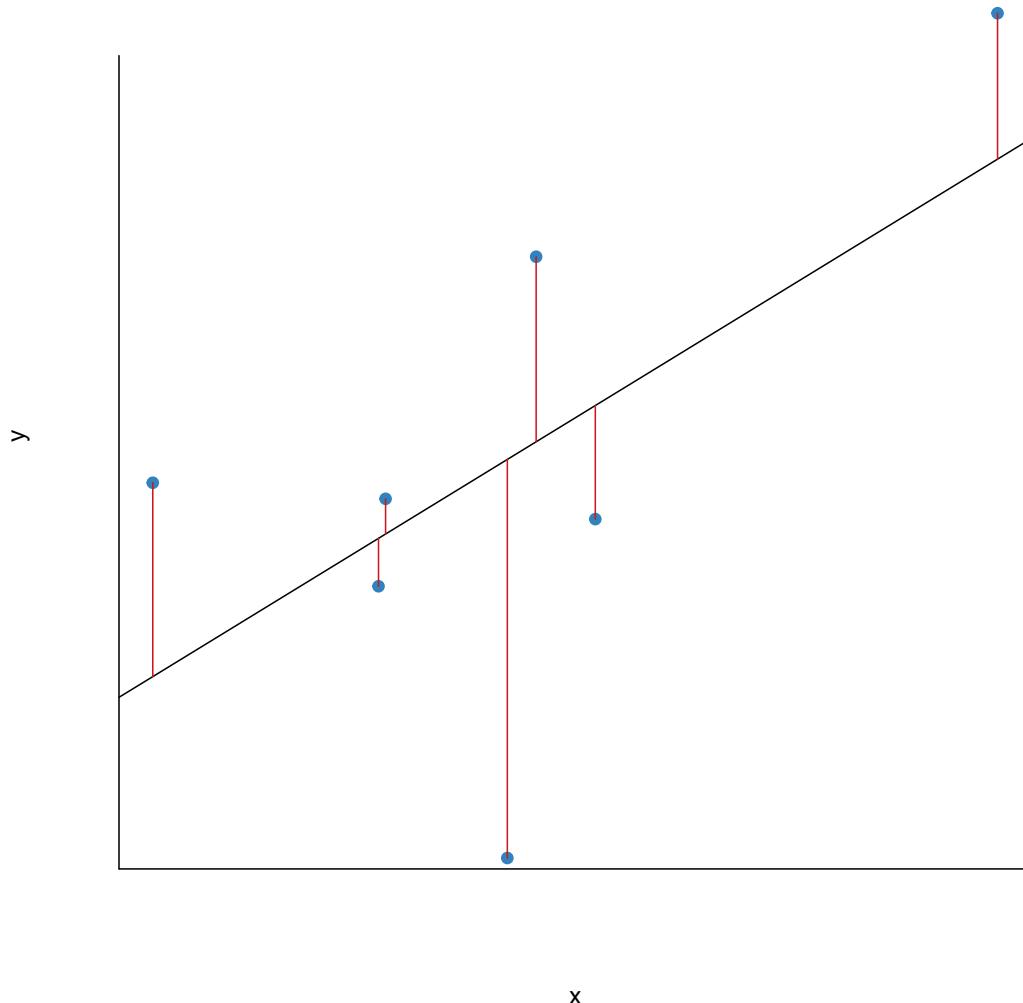
Una manera muy común de aproximarse a encontrar el “mejor” valor para los coeficientes poblacionales desconocidos (β) es minimizar la suma de los errores que produce el modelo elevados al cuadrado;¹¹ este método se conoce con el nombre de Mínimos Cuadrados Ordinarios (MCO, o en inglés OLS). Intuitivamente, 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” ($\hat{y} = \mathbf{X}\hat{\beta}$).

En la Figura 2.7 se presenta con líneas rojas cada una de las distancias entre el valor observado de la variable dependiente y_i y el valor que predice el modelo \hat{y}_i para un valor determinado de x_i . En otras palabras, las líneas rojas se presentan cada uno de los errores de modelo estimado. La línea (en negro) se ha trazado minimizando la suma al cuadrado de cada una de estas distancias. De todas las líneas

¹¹Elevar 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.

posibles (como las presentadas en la Figura 2.6), esta linea es la única que minimiza esa suma de errores al cuadrado.

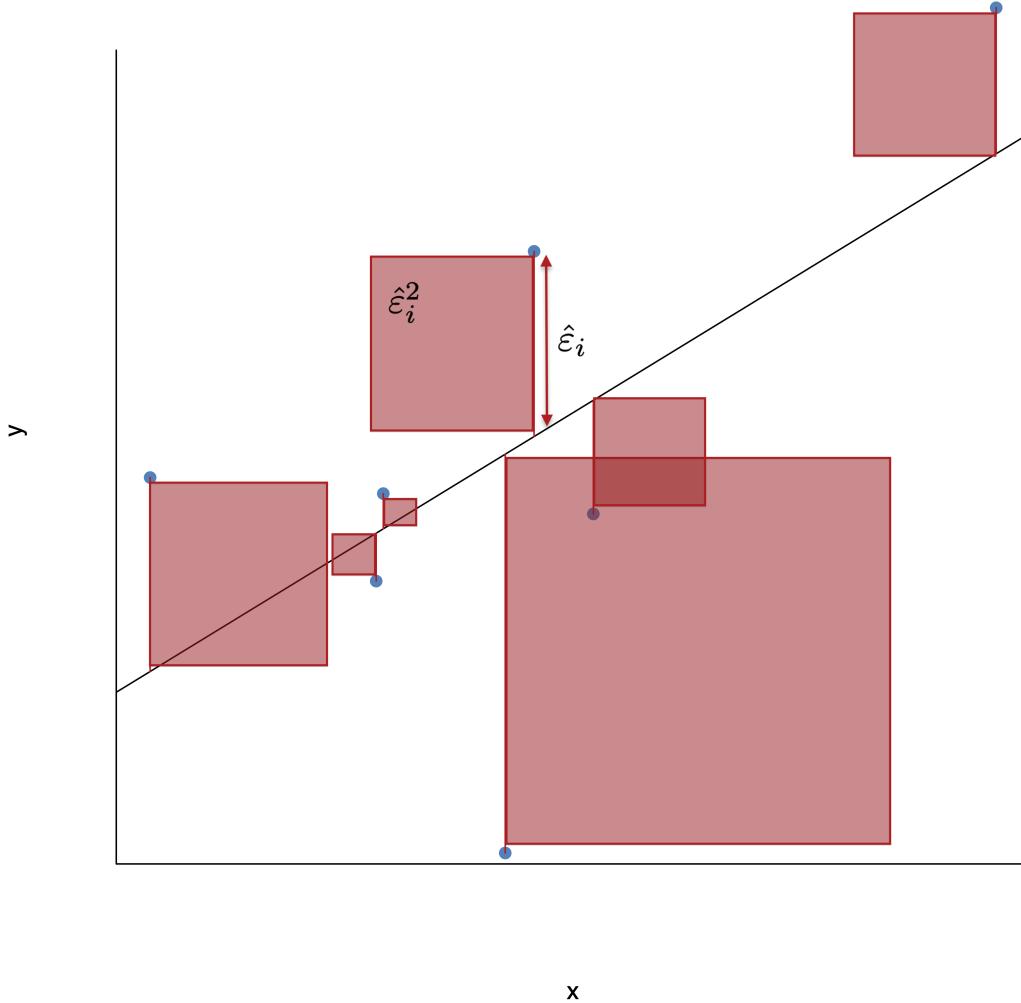
Figura 2.7. Ejemplo de la linea trazada con el método MCO para una muestra observada para dos variables (y y x)



Fuente: Elaboración propia

Esta idea de minimizar el cuadrado del error del modelo se puede ver graficado en la Figura 2.8. Las áreas de los cuadrados sombreadas en rojo representan cada uno de los errores del modelo ($\hat{\epsilon}_i$) elevado al cuadrado. El tamaño del lado del cuadrado es $\hat{\epsilon}_i$ y por tanto el área del cuadrado corresponde a $\hat{\epsilon}_i^2$. El método MCO lo que hace es encontrar la linea que minimiza la suma de las áreas del cuadrado; cualquier otra linea que se trace tendrá una suma de las áreas de los respectivos cuadrados mayor a la del método MCO.

Figura 2.8. Ejemplo de la suma de los cuadrados de los errores que implica el método MCO



Fuente: Elaboración propia

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

$$\underset{\hat{\beta}}{\text{Min}} \left\{ [\mathbf{y} - \hat{\mathbf{y}}]^T [\mathbf{y} - \hat{\mathbf{y}}] \right\} \quad (2.21)$$

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:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ [\mathbf{y} - \mathbf{X}\hat{\beta}]^T [\mathbf{y} - \mathbf{X}\hat{\beta}] \right\} \quad (2.22)$$

donde $\mathbf{X}\hat{\beta}$ corresponde al vector de valores estimados por el modelo para la variable dependiente; en

otras palabras, el modelo estimado. Así, el estimador de MCO del vector de coeficientes β es:¹²

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (2.23)$$

La ecuación estimada estará dada por:

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\beta}. \quad (2.24)$$

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

$$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X} \hat{\beta} \quad (2.25)$$

En el Anexo al final de este capítulo se discuten algunas propiedades importantes del vector $\hat{\epsilon}$.

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

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-k} = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n-k} = \frac{\mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y}}{n-k} \quad (2.26)$$

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

$$\widehat{Var}[\hat{\beta}] = s^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (2.27)$$

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,

$$\widehat{Var}[\hat{\beta}] = \begin{bmatrix} \widehat{Var}[\hat{\beta}_1] & \widehat{Cov}[\hat{\beta}_1, \hat{\beta}_2] & \cdots & \widehat{Cov}[\hat{\beta}_1, \hat{\beta}_k] \\ & \widehat{Var}[\hat{\beta}_2] & \cdots & \widehat{Cov}[\hat{\beta}_2, \hat{\beta}_k] \\ & & \ddots & \vdots \\ & & & \widehat{Var}[\hat{\beta}_k] \end{bmatrix}. \quad (2.28)$$

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)**.¹³ Este resultado se conoce como el Teorema de Gauss-Markov (ver recuadro abajo). La propiedad de que el estimador de MCO de β sea insesgado implica que en promedio el estimador obtendrá el valor real β , Formalmente:

$$E[\hat{\beta}] = \beta. \quad (2.29)$$

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

¹²En el anexo (ver sección 2.4) al final de este capítulo se presenta la derivación de ésta fórmula.

¹³Una demostración de este resultado se presenta en el anexo (Ver sección 2.4) al final de este capítulo.

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.

Teorema de Gauss-Markov

Si se considera un modelo lineal $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\epsilon}_{n \times 1}$ 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 $\boldsymbol{\epsilon}$ tiene media cero, varianza constante y no autocorrelación. Es decir: $E[\boldsymbol{\epsilon}] = 0$ y $Var[\boldsymbol{\epsilon}] = \sigma^2 I_n$

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)**.

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

En todos los capítulos de este libro, encontrarás ejercicios que serán resueltos paso a paso mediante el uso de R (R Core Team, 2018). En este capítulo explicaremos detenidamente desde cómo cargar los datos, hasta cómo estimar un modelo por MCO.

Para este ejercicio, nos enfrentamos a una situación en la que un gerente de mercadeo quiere determinar cuál es el efecto de aumentar en un dólar el precio del azúcar sobre la demanda de ésta, el principal producto de esta organización. En otras palabras, la pregunta de negocio es ¿qué efecto tiene un aumento de un dólar sobre la demanda de azúcar. Esta pregunta de negocio claramente implica emplear analítica diagnóstica y una tarea de estimar regresiones.

Cuando tenemos una pregunta de negocio a responder, de inmediato tenemos qué preguntarnos que datos tenemos. En este caso tenemos a nuestra disposición una base de datos relativamente limitada. Contamos con datos mensuales para los últimos 60 meses¹⁴ de las siguientes variables:

- Q = las toneladas de azúcar vendidas a los minoristas¹⁵ (Q_t)
- p = precio por libra en dólares que cobra la organización a sus distribuidores (p_t)
- pcomp = precio por libra en dólares que cobra el principal competidor a sus distribuidores ($p_{comp,t}$)
- IA = índice de actividad económica (IA_t).

¹⁴Noten que esto implica que contaremos con una serie de tiempo y por eso emplearemos el subíndice t en la formulación del modelo estadístico.

¹⁵Esto corresponde a lo que se conoce en la industria como el *Sell-in*. Es decir, *Sell-in* se domina la venta del fabricante al canal distribuidor o minorista. El otro término asociado es el *Sell-out* que corresponde a las ventas que se hacen al consumidor final después de pasar por el canal de distribución.

Los datos están disponibles en el archivo adjunto (*DataCap1.csv*)¹⁶. Lastimosamente no contamos con mas datos.

Con los datos disponibles podemos proceder a construir un primer posible modelo candidato a describir el comportamiento de las cantidades demandadas. El modelo sería:

$$Q_t = \beta_1 + \beta_2 p_t + \beta_3 p_{comp,t} + \beta_4 IA_t + \varepsilon_t. \quad (2.30)$$

De esta manera dados los datos disponibles en el archivo adjunto, se desea estimar el modelo 2.30. Y para poder responder la pregunta de negocio, será necesario concentrar nuestra atención sobre β_2 que representa el efecto de un aumento de un dolar en el precio de la organización por libra del azúcar sobre su demanda. Noten que es importante incluir más variables que p_t en el modelo de regresión, pues como se discutió en la sección 1.3 nuestro primer objetivo es encontrar el *DGP*. Y con este DGP responder las preguntas de negocio. Por otro lado, otra razón técnica para tratar de emplear todas aquellas variables que afectan a la variable dependiente es que esto minimizará el tamaño del error de nuestro modelo.

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()` del paquete base de R. Esta función como mínimo necesita los siguientes argumentos:

```
read.csv(file, header = TRUE, sep = ",")
```

donde:

- **file**: es el nombre del archivo y su ruta que se debe poner entre comillas.
- **header** : es un operador lógico que le dice a la función si la primera fila del archivo csv contiene el encabezado con el nombre de las variables. Por defecto `header = TRUE` se espera que la primera fila del archivo contenga el nombre de las variables.
- **sep**: Es el delimitador de los campos que emplea el archivo csv. Por defecto, se espera una coma.

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 la ruta del folder donde se encuentra el archivo). Importemos los datos empleando la función `read.csv()` y guardémos los datos en el objeto *DatosCap1*.

```
DatosCap1 <- read.csv("../Data/DataCap1.csv", sep = ",")
```

¹⁶Por razones de confidencialidad, los datos han sido modificados de sus valores originales.

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 DatosCap1. Para ello podemos ver las primeras filas del objeto y la clase de objeto de la siguiente manera:

```
head(DatosCap1)

##   X      Q    p pcomp  IA
## 1 1 529.904 8.81 8.40 1.8
## 2 2 1041.318 7.46 10.85 1.7
## 3 3 874.360 9.14 6.52 3.6
## 4 4 821.920 8.68 11.77 2.7
## 5 5 953.396 7.27 10.20 2.5
## 6 6 651.249 8.67 5.54 1.1

class(DatosCap1)

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

Noten que el objeto DatosCap1 fue leído como un **data.frame** que contiene una primera columna con un índice (esta fue denominada *X*), y cuatro columnas más que contienen las variables *Q*, *p*, *pcomp* e *IA*. Procedamos a eliminar la primera columna que no serán necesarias. Y constatemos que las variables que nos quedan en el **data.frame** son numéricas. Esto se puede hacer de diferentes maneras.

```
DatosCap1 <- DatosCap1[, -1]

head(DatosCap1)

##      Q    p pcomp  IA
## 1 529.904 8.81 8.40 1.8
## 2 1041.318 7.46 10.85 1.7
## 3 874.360 9.14 6.52 3.6
## 4 821.920 8.68 11.77 2.7
## 5 953.396 7.27 10.20 2.5
## 6 651.249 8.67 5.54 1.1

apply(DatosCap1, 2, class)

##      Q    p pcomp  IA
## "numeric" "numeric" "numeric" "numeric"

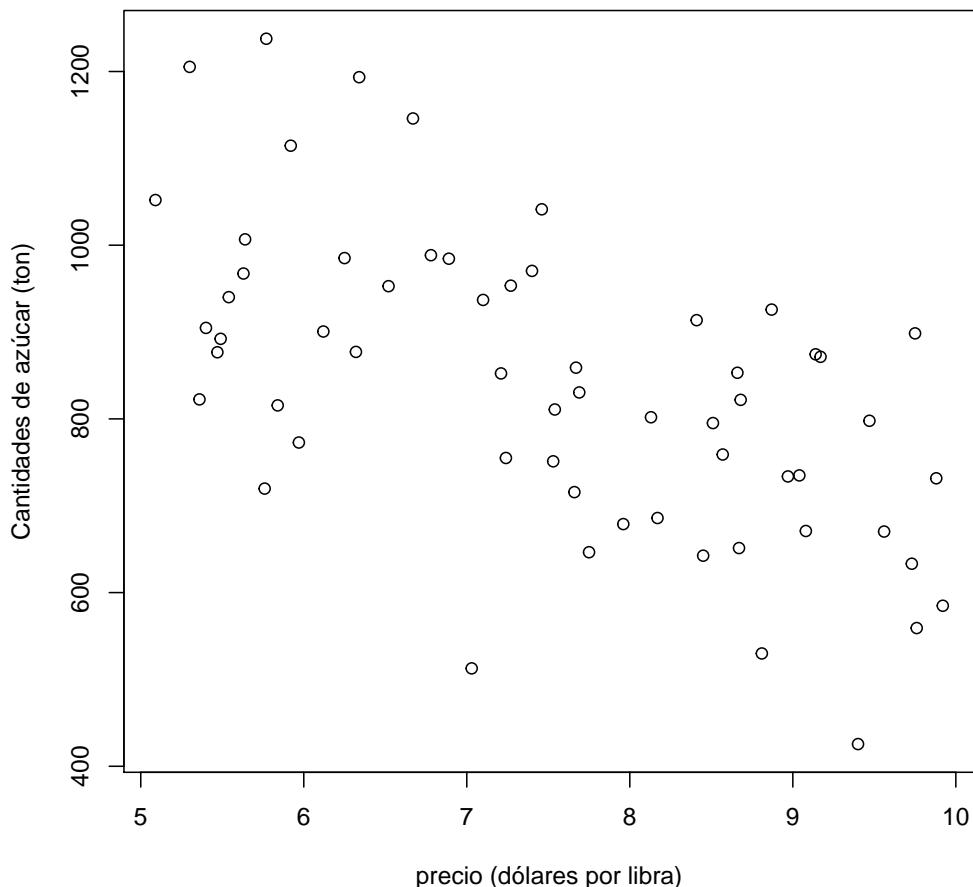
str(DatosCap1)
```

```
## 'data.frame': 60 obs. of 4 variables:  
## $ Q      : num  530 1041 874 822 953 ...  
## $ p      : num  8.81 7.46 9.14 8.68 7.27 8.67 7.96 7.53 6.34 9.56 ...  
## $ pcomp: num  8.4 10.85 6.52 11.77 10.2 ...  
## $ IA    : num  1.8 1.7 3.6 2.7 2.5 1.1 1.5 1.3 3.1 2 ...
```

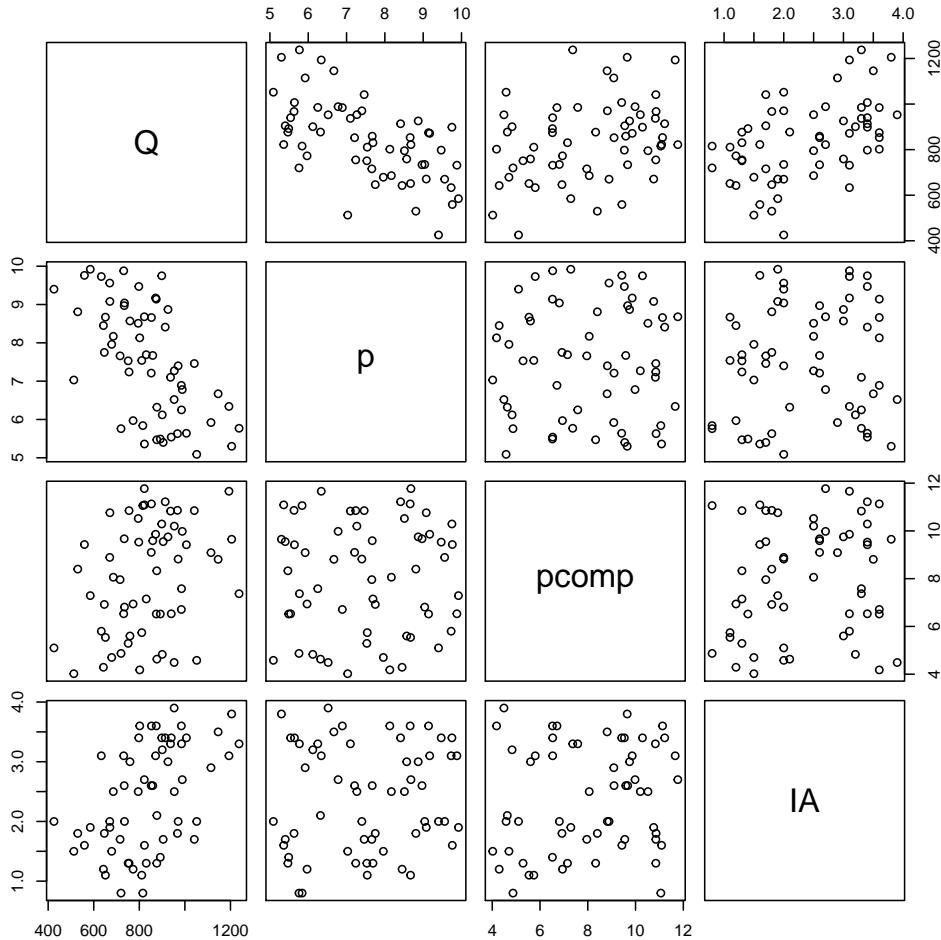
2.3.2 Estimación del modelo

Recordemos que se desea estimar el modelo presentado en 2.30. Antes de estimar el modelo, una buena práctica cuando se tiene pocas variables en la base de datos es graficar las variables y su relación con la variable dependiente. Típicamente se emplea un diagrama de dispersión, en nuestro caso tendremos.

```
plot(DatosCap1$p, DatosCap1$Q, xlab = "precio (dólares por libra)",  
      ylab = "Cantidades de azúcar (ton)")
```



```
plot(DatosCap1)
```



Otra opción es emplear los paquetes *ggplot2* (Wickham, 2016) y *GGally* (Schloerke y col., 2021) para tener una visualización más agradable como la que se presenta en las Figuras 2.9 y 2.10.

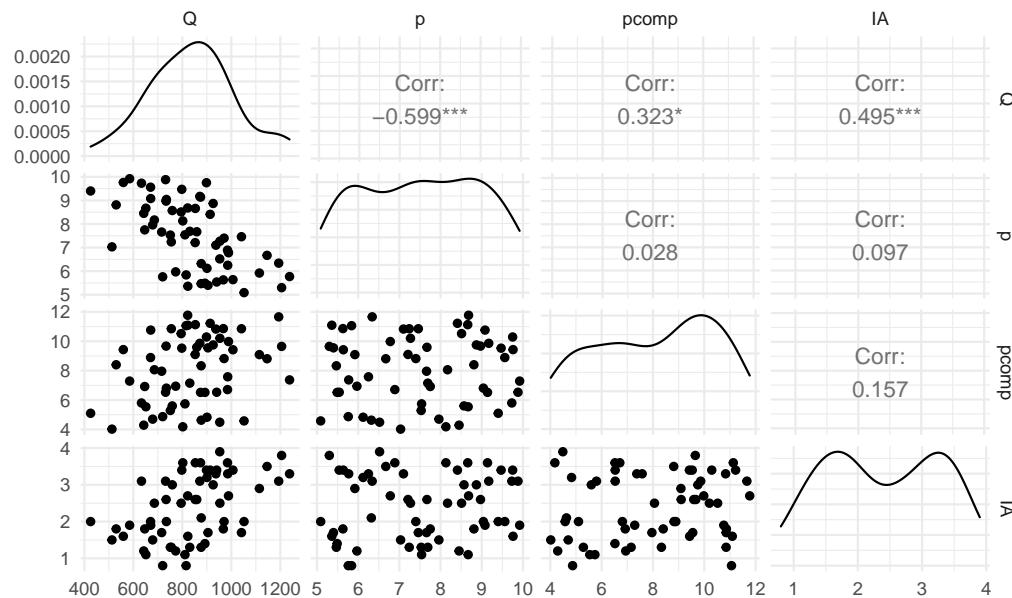
```
library(ggplot2)

library(GGally)
ggpairs(DatosCap1)+theme_minimal()

ggplot(DatosCap1, aes( Q, p)) + geom_point(color='blue') +
  labs(x = "precio (dólares por libra)",
       y = "Cantidades de azúcar (ton)") +
```

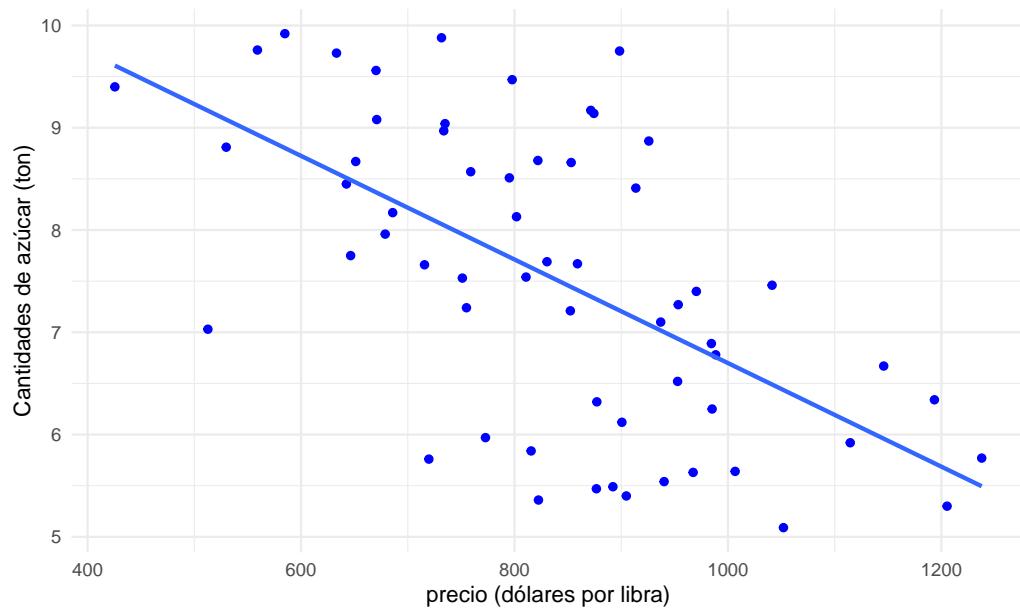
```
geom_smooth(method = "lm", se = FALSE) + theme_minimal()
```

Figura 2.9. Relación entre todas las variables de la base de datos



Fuente: Elaboración propia

Figura 2.10. Relación entre las cantidades demandadas de azúcar y el precio



Fuente: Elaboración propia

La figura 2.10 nos sugiere la existencia de una relación lineal por lo menos entre estas dos variables. Procedamos a estimar el modelo.

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 = Q ~ p + pcomp + IA, data = DatosCap1)
class(R1)

## [1] "lm"

R1

##
## Call:
## lm(formula = Q ~ p + pcomp + IA, data = DatosCap1)
##
## Coefficients:
## (Intercept)          p        pcomp         IA
## 1026.37      -77.62      19.27     100.46
```

En este caso hemos guardado en el objeto R1 los resultados de estimar 2.30. Este objeto será de clase **lm**. Noten que se empleo la virgulilla (palito de la eñe) para representar el signo igual de la expresión 2.30, también es importante anotar que no fue necesario escribir todos los parámetros deseados, R por defecto incluye un intercepto y los correspondientes coeficientes (elementos del vector β) 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 = Q ~ p + pcomp + IA, data = DatosCap1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -196.09   -50.92  -16.91    69.17   214.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1026.371    76.634 13.393 < 2e-16 ***
```

```

## p           -77.624    8.195  -9.472 3.15e-13 ***
## pcomp      19.270    5.177   3.722  0.00046 ***
## IA          100.458   13.606   7.384  8.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91.27 on 56 degrees of freedom
## Multiple R-squared:  0.7337, Adjusted R-squared:  0.7194
## F-statistic: 51.42 on 3 and 56 DF,  p-value: 4.264e-16

```

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

```

R1a <- lm(Q ~ p + pcomp + IA, DatosCap1)
formula1 <- Q ~ p + pcomp + IA
R2 <- lm(formula1, DatosCap1)
R3 <- lm(Q ~ ., DatosCap1)

```

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 las tres variables explicativas, 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 L^AT_EX, usted puede utilizar el paquete *stargazer* (Hlavac, 2018). Una vez instalado el paquete puede utilizar la función del mismo nombre para obtener la tabla en formato L^AT_EX.

```

library(stargazer)
stargazer(R1, t.auto = TRUE, p.auto = TRUE, title = "Modelo estimado por MCO",
          label = "res2.Hetero", header = FALSE, table.placement = "H",
          notes.align = "l")

```

Cuadro 2.2: Modelo estimado por MCO

<i>Dependent variable:</i>	
Q	
p	-77.624*** (8.195)
pcomp	19.270*** (5.177)
IA	100.458*** (13.606)
Constant	1,026.371*** (76.634)
<hr/>	
Observations	60
R ²	0.734
Adjusted R ²	0.719
Residual Std. Error	91.273 (df = 56)
F Statistic	51.417*** (df = 3; 56)

Note: *p<0.1; **p<0.05; ***p<0.01

Nota que en este caso la ecuación estimada será:

$$\hat{Q}_t = 1026,37 - 77,62p_t + 19,27p_{comp,t} + 100,46IA_t \quad (2.31)$$

Así, el resultado implica que por cada dólar por libra adicional en el precio del producto la demanda caerá en promedio en 77.62 toneladas mes.

Para obtener la matriz de varianzas y covarianzas del intercepto y las pendientes, que son otras cantidades que desconocíamos, se puede emplear la función función **vcov()** de la siguiente manera:

```
vcov(R1)
```

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

$$\hat{\beta}_1 = 1026,37 \quad (2.32)$$

$$\hat{\beta}_2 = -77,62 \quad (2.33)$$

$$\hat{\beta}_3 = 19,27 \quad (2.34)$$

$$\hat{\beta}_4 = 100,46 \quad (2.35)$$

$$\sqrt{\widehat{Var}(\hat{\beta}_1)} = S_{\hat{\beta}_1} = 76,63 \quad (2.36)$$

$$\sqrt{\widehat{Var}(\hat{\beta}_2)} = S_{\hat{\beta}_2} = 8,2 \quad (2.37)$$

$$\sqrt{\widehat{Var}(\hat{\beta}_3)} = S_{\hat{\beta}_3} = 5,18 \quad (2.38)$$

$$\sqrt{\widehat{Var}(\hat{\beta}_4)} = S_{\hat{\beta}_4} = 13,61 \quad (2.39)$$

Como se discutirá en el próximo capítulo, del Cuadro 2.2 podemos concluir que el coeficiente asociado al precio 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. Esto hace que nuestra respuesta a la pregunta de negocio sea correcta. Pero esto se discutirá en detalle en el próximo capítulo en el que discutiremos la inferencia a partir del modelo de regresión múltiple.

Ejercicios

2.1 Demuestre que los coeficientes de una función Cobb-Douglas como la representada en 2.10 se puede interpretar como elasticidades.

2.2 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 científico de datos, plantea el siguiente modelo, dada la disponibilidad de datos existente:

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \alpha_4 LDies_t + \alpha_5 LEI_t + \alpha_6 V_t + \varepsilon_t. \quad (2.40)$$

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 los datos disponibles en el archivo regmult.csv. Su misión es estimar el modelo 2.40 empleando el método MCO y reportar sus resultados en una tabla. Posteriormente, interprete cada uno de los coeficientes estimados.

2.4 Anexos

2.4.1 Derivación de los estimadores MCO

Formalmente, el estimador de MCO para el vector de coeficientes β en el modelo 2.22, denotado por $\hat{\beta}$, se encuentra minimizando la distancia cuadrada entre cada valor observado del vector de realizaciones de la variable aleatoria dependiente (y) y el vector de estimaciones del modelo ($\hat{y} = \mathbf{X}\hat{\beta}$). Formalmente, el problema es:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ \left[\mathbf{y} - \mathbf{X}\hat{\beta} \right]^T \left[\mathbf{y} - \mathbf{X}\hat{\beta} \right] \right\} \quad (2.41)$$

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 2.41. 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 $\mathbf{y}^T_{1 \times n} \mathbf{X}_{n \times k} \hat{\beta}_{k \times 1} = (\hat{\beta}_{k \times 1}^T \mathbf{X}_{n \times k}^T \mathbf{y}_{1 \times n})^T$ y además los productos $\hat{\beta}_{1 \times k}^T \mathbf{X}_{k \times n}^T \mathbf{y}_{n \times 1}$ y $\mathbf{y}^T_{1 \times n} \mathbf{X}_{n \times k} \hat{\beta}_{k \times 1}$ son escalares, por tanto $\mathbf{y}^T_{1 \times n} \mathbf{X}_{n \times k} \hat{\beta}_{k \times 1} = \hat{\beta}_{1 \times k}^T \mathbf{X}_{k \times n}^T \mathbf{y}_{n \times 1}$. Así, el problema 2.41 es equivalente a:

$$\underset{\hat{\beta}}{\text{Min}} \left\{ \mathbf{y}^T \mathbf{y} - 2\hat{\beta}^T \mathbf{X}^T \mathbf{y} + \hat{\beta}^T \mathbf{X}^T \mathbf{X}\hat{\beta} \right\} \quad (2.42)$$

Ya podemos regresar a nuestro problema de minimización y considerar la condición de primer orden para este problema, en este caso la condición es:¹⁷

$$\frac{\partial}{\partial \hat{\beta}} \{ \bullet \} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\hat{\beta} \equiv 0 \quad (2.43)$$

La expresión 2.43 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 inversa¹⁸ de $\mathbf{X}^T \mathbf{X}$, tendremos:

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

¹⁷Hay que anotar que la derivada es con respecto a un vector y no a un escalar.

¹⁸ $(\mathbf{X}^T \mathbf{X})^{-1}$ 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.

2.4.2 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 $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times k} \beta_{k \times 1} + \varepsilon_{n \times 1}$
- 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 ε tiene media cero, varianza constante y no autocorrelación. Es decir, $E[\varepsilon] = 0$ y $Var[\varepsilon] = \sigma^2 \mathbf{I}_n$

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

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

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

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

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

Y por tanto,

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

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

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

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

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

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,

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

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.

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

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

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

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

$C C^T$ es una matriz cuyos elementos en la diagonal principal serán positivos.¹⁹ (¿Por qué?) Ahora comparemos la varianza de nuestro estimador de MCO ($\hat{\beta}$) con $\tilde{\beta}$. Recuerden que $Var[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$, adicionalmente observen que $Var[\tilde{\beta}] = \sigma^2 \left[(\mathbf{X}^T \mathbf{X})^{-1} + C C^T \right]$ no puede ser menor que $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$, pues:

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

Por tanto $Var[\tilde{\beta}_i] = \sigma^2 \left\{ \left[(\mathbf{X} \mathbf{X})_{ii}^{-1} + C C_{ii}^T \right] \right\} > Var[\hat{\beta}_i] = \sigma^2 \left[(\mathbf{X} \mathbf{X})_{ii}^{-1} \right]$. En el mejor de los casos $Var[\tilde{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ 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 $\tilde{\beta}$ es MELI.

2.4.3 Algunas propiedades importantes de los residuos estimados.

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

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

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

$$\mathbf{X}^T \hat{\epsilon} = 0 \tag{2.44}$$

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

$$\mathbf{X}^T \hat{\epsilon} = \mathbf{X}^T \left[\mathbf{y} - \mathbf{X} \hat{\beta} \right] = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \hat{\beta}$$

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

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

¹⁹Una matriz positiva semi-definida

Por lo tanto,

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

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

$$\mathbf{X}^T \hat{\boldsymbol{\varepsilon}} = \left[\sum_{i=1}^n X_{1i} \hat{\varepsilon}_i \quad \sum_{i=1}^n X_{2i} \hat{\varepsilon}_i \quad \vdots \quad \sum_{i=1}^n X_{ki} \hat{\varepsilon}_i \right] = \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array} \right]$$

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:

$$\sum_{i=1}^n \hat{\varepsilon}_i = 0. \quad (2.45)$$

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

$$\bar{\boldsymbol{\varepsilon}} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i}{n} = 0. \quad (2.46)$$

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

$$E [\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}] = E \left[\sum_{i=1}^n \hat{\varepsilon}_i^2 \right] = (n-2) \sigma^2. \quad (2.47)$$

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

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

$$\hat{\boldsymbol{\varepsilon}} = \left[I - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] y.$$

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:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{M} y = \mathbf{M} \mathbf{X} \boldsymbol{\beta} + \mathbf{M} \boldsymbol{\varepsilon} = \mathbf{M} \boldsymbol{\varepsilon} \quad (2.48)$$

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

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

Regresando a (2.48), se tiene que:

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

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,

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

y empleando otras propiedades de la traza sabemos que:

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

De esta manera se tiene que

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

Esto implica que

$$E [\hat{\varepsilon}^T \hat{\varepsilon}] = \sigma^2 \text{trace} (\mathbf{M}) = (n - k) \sigma^2. \quad (2.49)$$

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

$$E [s^2] = E \left[\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - k} \right] = E \left[\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n - k} \right] = \sigma^2. \quad (2.50)$$

3 . Inferencia y análisis de regresión

Objetivos del capítulo

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

- Explicar en sus propias palabras porqué es necesario hacer inferencia sobre los valores estimados
 - Explicar en sus propias palabras el significado de las métricas de bondad de ajuste de una regresión múltiple
 - 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 población (o parámetros poblacionales) a partir de una muestra 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 cómo determinar si una variable o un conjunto de estas afectan o no a la variable dependiente.

En el Capítulo 2 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 parezca a primera vista algo contradictoria; si se reflexiona un poco tendrá más sentido. Es imposible “adivinar” correctamente el valor real (con todos los decimales posibles) a partir de un solo intento, cuando el verdadero valor pertenece a un conjunto infinito de valores.

Te debes 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, como 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.

Antes de continuar hagamos un ejercicio para aclarar este punto. Realicemos el siguiente experimento de simulación. Supongamos que se cuenta con el siguiente DGP conocido:

$$Y_i = 1 + 5X_{2i} + \varepsilon_i$$

Ahora supongamos que un científico de datos trata de encontrar ese DGP para lo cual cuenta con

observaciones para dos variables explicativas: $X2_i$ y $X3_i$. (Esta última variable sabemos que no afecta a Y_i pero el científico de datos no lo sabe).

Uno esperaría que al correr una regresión de Y_i en función de $X2_i$ y $X3_i$ y un intercepto, el coeficiente que acompaña $X2_i$ debería ser igual a 5 y el que acompaña a $X3_i$ debería ser cero. Veamos si esto ocurre así o no.

Generemos inicialmente una muestra de tamaño 100 ($n = 100$) del DGP real y además generaremos la variable $X3_i$. El primer paso será generar las variables explicativas. Esto lo haremos, generando números aleatorios de una distribución uniforme entre 2 y 2. Esto lo podemos hacer con la función **runif()** del paquete base de R. Esta función tiene como argumentos el número de números aleatorios que se quiere generar (**n**), el valor mínimo que puede tomar el valor aleatorio que se generará (**min**) y el valor máximo (**max**).

```
# se fija una semilla
set.seed(123456)
# se crea la variable explicativa X2
X2 <- runif(n = 100, min = -2, max = 2)
```

Noten que estamos empleando la función **set.seed()**. Esta función genera una semilla aleatoria (random seed en inglés) fija. La semilla aleatoria es un número empleado para inicializar un algoritmo generador de números aleatorios en los computadores. Esto garantiza que siempre que se genere un número aleatorio sea el mismo. De hecho los números aleatorios generados por los computadores no son realmente aleatorios, pues dependen de una función determinística que simula la aleatoriedad que podemos encontrar en la naturaleza. Por eso los números aleatorios generados por un computador se conocen como números pseudoaleatorios.

Regresando a nuestro experimento, ahora generemos el error del modelo. Supongamos que éste fue generado por una distribución normal con media cero y desviación estándar 3. Es decir,

$$\varepsilon_i \sim N(0, 3^2)$$

Los 100 errores que siguen esta distribución los podemos simular empleando la función **rnorm()** del paquete base de R. Esta función tiene como argumentos el número de números aleatorios que se quiere generar (**n**), la media (**mean**) y la desviación estándar (**sd**).

```
# se crea el término de error
error <- rnorm(100, 0, 3)
```

Ahora, construyamos la muestra para la variable dependiente empleando el DGP. Es decir,

```
# se crea la variable dependiente
Y <- 1 + 5 * X2 + error
```

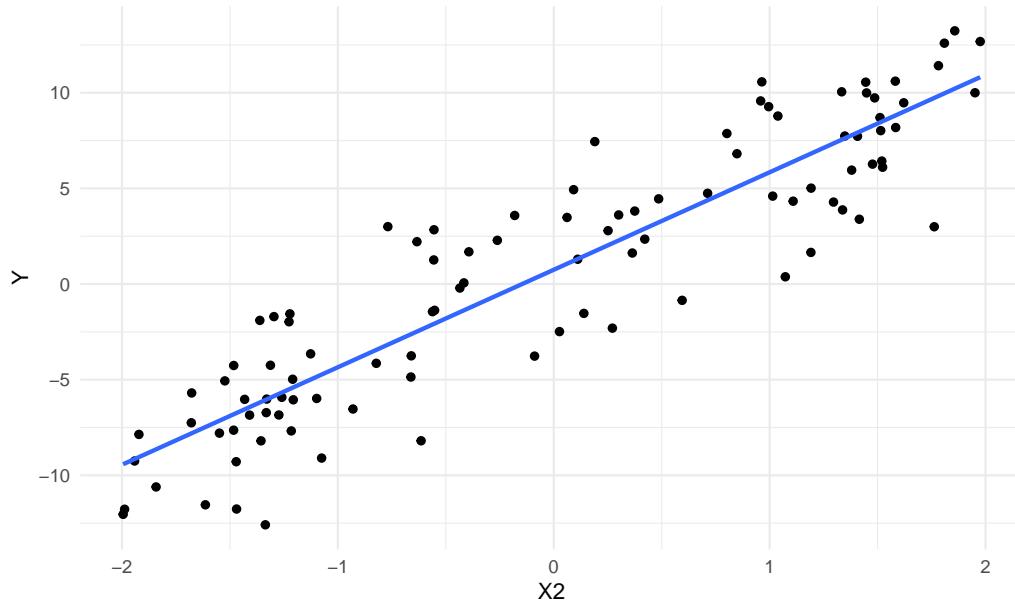
Noten que hemos generado 100 observaciones para Y_i y otras 100 para $X2_i$. Para completar nuestro experimento falta generar $X3_i$. Hagamos esto y construyamos un **data.frame** que contenga todas las variables.

```
# se crea la variable explicativa X3
X3 <- runif(100, 3, 6)
# se crea el data.frame
data <- data.frame(Y, X2, X3)
# chequeo del data.frame creado
str(data)

## 'data.frame': 100 obs. of  3 variables:
## $ Y : num  1.65 4.6 -0.21 2.21 2.84 ...
## $ X2: num  1.191 1.014 -0.435 -0.634 -0.555 ...
## $ X3: num  4.81 3.2 5.5 3.92 3.93 ...
```

Antes de continuar grafique los datos generados y encontrará una visualización similar a la Figura 3.1.

Figura 3.1. Relación entre los datos simulados



Fuente: Elaboración propia

Ahora estimemos el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Como lo dijimos anteriormente, se debería esperar que

- $\hat{\beta}_1 = 1$

- $\hat{\beta}_2 = 5$
- $\hat{\beta}_3 = 0$

Estimemos el modelo.

```
res1 <- lm(Y ~ X2 + X3, data = data)
summary(res1)

##
## Call:
## lm(formula = Y ~ X2 + X3, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.9684 -2.0363  0.1182  2.1905  6.2436
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.6305    1.5523    1.05   0.296
## X2          5.0950    0.2395   21.27  <2e-16 ***
## X3         -0.1997    0.3445   -0.58   0.563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.925 on 97 degrees of freedom
## Multiple R-squared:  0.8237, Adjusted R-squared:  0.82
## F-statistic: 226.5 on 2 and 97 DF,  p-value: < 2.2e-16
```

Noten que los valores estimados no coinciden con los valores reales (los del DGP). De aquí nace la afirmación que realizamos anteriormente sobre la probabilidad de cero de que con una muestra podamos obtener un valor estimado exactamente igual al valor poblacional.

Ahora, supongamos que tenemos el lujo de repetir este ejercicio 10 mil veces y guardamos los coeficientes estimados de cada repetición. Es decir, generaremos 10 mil muestras (empleando la misma variable $X2$ implica suponer que la variable explicatoria es fija) y para cada una de las muestras calculamos los coeficientes y los guardamos¹. Para hacer el experimento más interesante generemos el error de cada muestra de una distribución χ^2 con 5 grados de libertad².

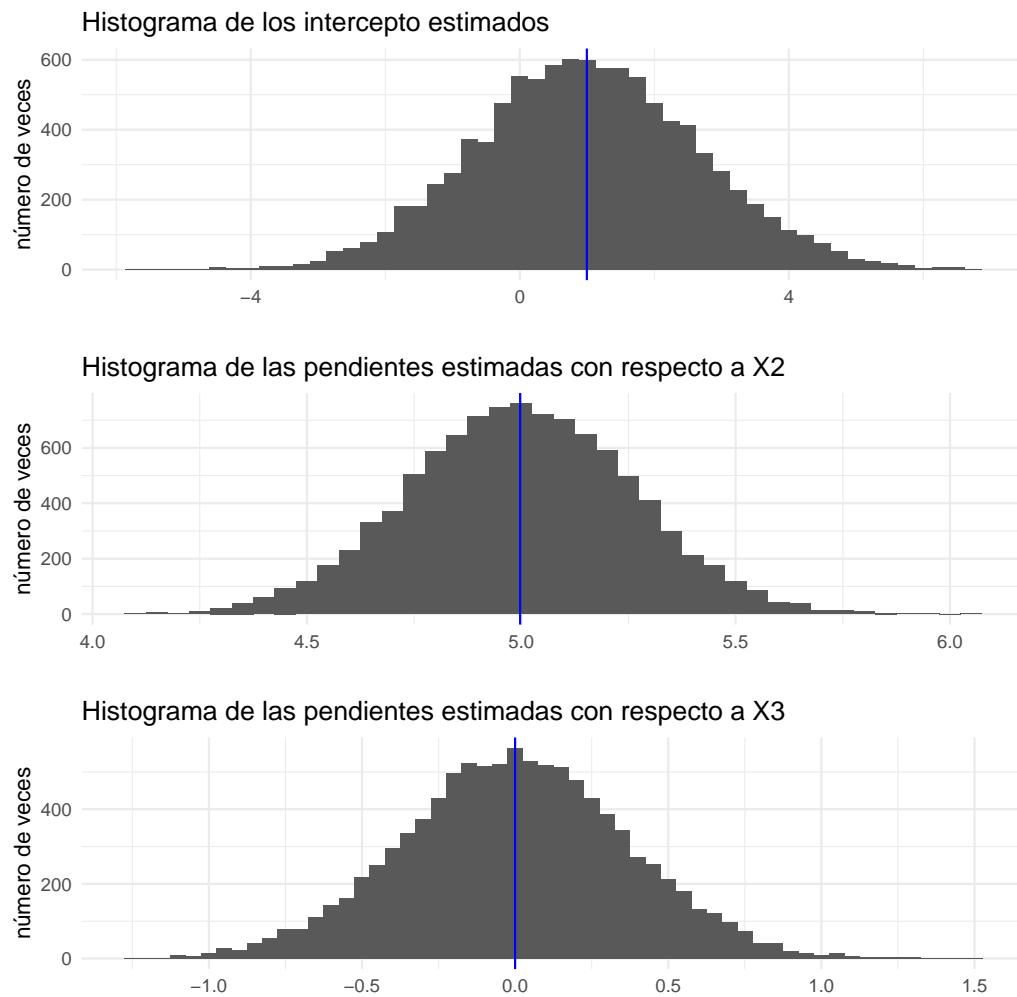
¹En este caso emplearemos un bucle (*loop*) para replicar este cálculo. Una breve introducción a los Lops en R se puede encontrar en Alonso (2021).

²Dado que una distribución χ^2 con g grados de libertad tiene como media sus grados de libertad, le restamos a cada valor generado aleatoriamente su media para centrar los errores y cumplir el supuesto del error.

```
# se crea objeto para guardar resultados
res <- matrix(, nrow = 10000, ncol = 3)
# se replican los cálculos 10 mil veces
for (N in 1:10000) {
  error <- rchisq(100, 5) - 5
  Y <- 1 + 5 * X2 + error
  data <- data.frame(Y, X2, X3)
  res2 <- lm(Y ~ X2 + X3, data = data)
  res[N, ] <- coef(res2)
}
```

Por otro lado, miremos por un momento los histogramas de la distribución de los coeficientes estimados (Ver Figura 3.2).

Figura 3.2. Histograma de los coeficientes estimados (en azul se presenta la media de los coeficientes estimados)



Fuente: Elaboración propia

Lo interesante es que la distribución de cada uno de estos coeficientes estimados (distribución muestral del estimador) tiene forma acampana, de hecho sigue una distribución normal. No obstante los residuales fueron creados a partir de una distribución χ^2 con 5 grados de libertad.

Ahora regresemos al caso en que se estima una sola vez el modelo. En ese caso se puede mejorar la precisión del modelo si se realizan pruebas de hipótesis o intervalos de confianza, empleando el resultado anterior.

Con todos los resultados de la simulación en el objeto `res` podemos por ejemplo encontrar la media de los valores estimados.

```
round(apply(res, 2, mean), 2)

## b1 b2 b3
## 1 5 0
```

Noten que obtenemos que la media de las 10 mil estimaciones de cada parámetro es igual al valor real. Es decir,

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

En otras palabras, este es un ejemplo de la insesgadez de los coeficientes, tal como se demostró en el capítulo anterior. Este experimento de simulación que acabamos de realizar se conoce como un Experimento o Simulación de Monte Carlo.

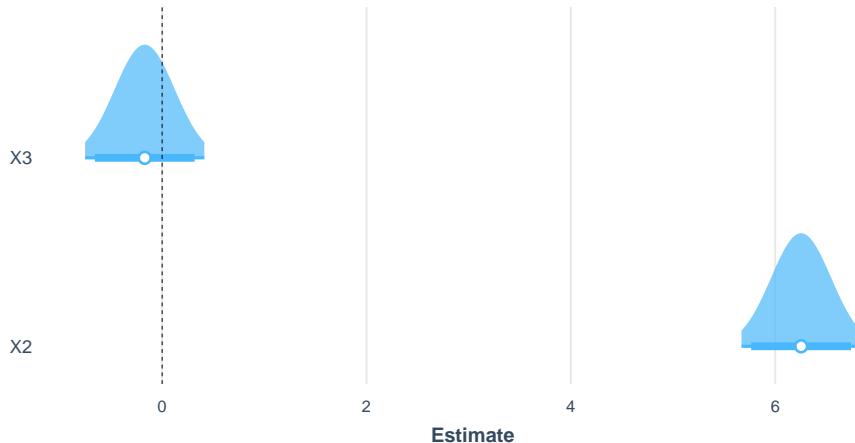
¿Qué es un Experimento de Monte Carlo

Un experimento de Monte Carlo o simulación de Monte Carlo o Método de Monte Carlo es una técnica estadística que se emplea para estimar los posibles resultados de un evento incierto. El método de Montecarlo fue inventado por John von Neumann y Stanislaw Ulam durante la Segunda Guerra Mundial para mejorar la toma de decisiones en condiciones de incertidumbre. Su nombre proviene de una conocida ciudad de casinos, llamada Mónaco, ya que el elemento de azar es el núcleo del enfoque de este ejercicio, similar al juego de la ruleta. En la estadística los experimentos de Monte Carlo son empleados, por ejemplo, para:

- Comparar el comportamiento y características de estimadores en condiciones de datos realistas.
- Proporcionar valores críticos de pruebas estadísticas que sean más ajustadas que aquellos derivados de distribuciones asintóticas.

Los experimentos de Montee Carlo en últimas nos permite entender el comportamiento de una técnica estadística en las condiciones muy específicas correspondientes al diseño del experimento. Es importante anotar que un experimento de Monte Carlo no corresponde a una demostración de que un resultado es universal. Y por tanto no reemplaza una demostración teórica.

En la Figura 3.3 podemos observar como las estimaciones del primer modelo (con una sola muestra) pueden ser mejoradas si empleamos un intervalo de confianza que tenga en cuenta la distribución muestral del estimador. Noten que los intervalos contienen los valores reales (poblacionales).

Figura 3.3. Pendientes estimadas para la primera muestra y sus intervalos de confianza

Fuente: Elaboración propia

Esto es lo que justifica emplear inferencia como lo veremos en este capítulo.

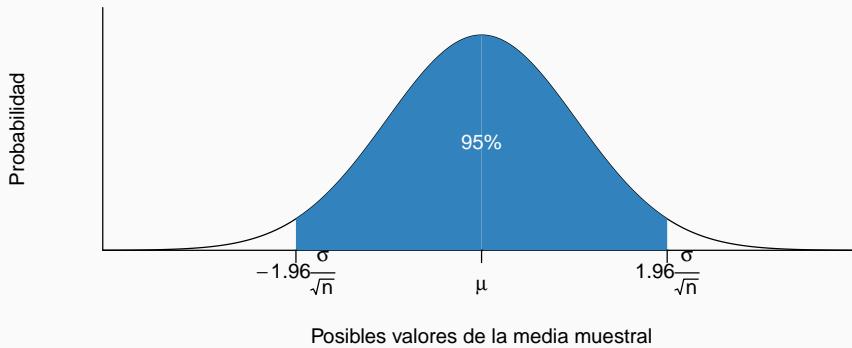
Teorema del Límite Central (TLC)

El **teorema del límite central** es uno de los resultados más poderosos y asombrosos de la ciencia estadística. La intuición detrás de este teorema es muy sencilla; si se suman un número suficientemente grande de variables aleatorias, entonces sin importar la distribución de las variables aleatorias sumadas, la sumatoria seguirá una distribución normal.

Tal vez, la aplicación mas sencilla del TLC es para hacer inferencia sobre la media. Veamos la intuición con un ejercicio mental. Supongamos que quieres conocer el peso medio de una población de estudiantes; es decir, la media poblacional del peso de ellos. En este caso podrán hacer un censo del peso de todos los estudiantes de la universidad y una vez pesados 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 a 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 podemos calcular la media muestral, $\bar{w} = \frac{\sum_{i=1}^n W_i}{n}$.

¿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.

Figura 3.4. Distribución muestral de la media

Fuente: Elaboración propia

Como la media corresponde a la suma de observaciones que fueron seleccionadas aleatoriamente dividida por una constante (el tamaño de la muestra), entonces el TLC aplica para la media muestral, siempre y cuando la muestra sea seleccionada aleatoriamente. En otras palabras, la distribución muestral de la media seguirá una distribución normal con media igual al valor de la media poblacional (μ) y desviación estándar (que para el caso de un estimador se denomina error estándar) $\frac{\sigma}{\sqrt{n}}$.

De esta manera, 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.

Ahora, empleemos el resultado de la estadística que nos permitirá construir intervalos de confianza y hacer pruebas de hipótesis. 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 y_i es una variable aleatoria, pues 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,

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

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 (¡qué 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:

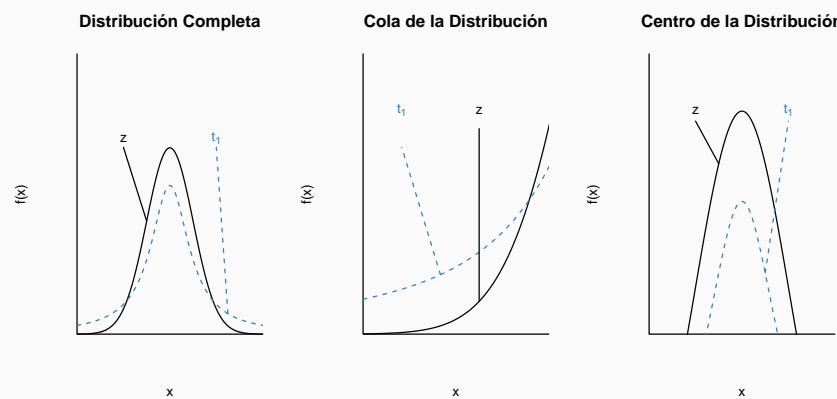
$$\hat{\beta}_p \pm t_{\frac{\alpha}{2}, n-k} s_{\hat{\beta}_p} \quad (3.5)$$

donde $s_{\hat{\beta}_p}$ corresponde a la raíz cuadrada de la varianza estimada del estimador $\hat{\beta}_p$.

La distribución t de Student y la distribución normal

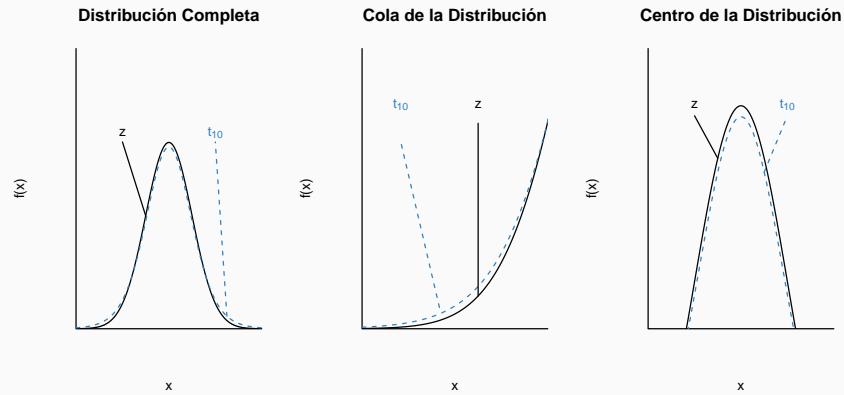
La distribución t posee una forma similar a la de la distribución normal. La diferencia está en las colas (los extremos) de las dos distribuciones. La distribución t tiende a tener colas más pesadas^a y por tanto un centro con menor probabilidad. Es decir, existe más probabilidad de caer lejos de la media en la distribución t que en la normal. Otra característica importante de la distribución t es que esta depende de un parámetro conocido como los grados de libertad. Los grados de libertad pueden ser entendidos como la cantidad de datos que están disponibles para hacer la inferencia y en este contexto para estimar la varianza poblacional del error. Así, entre más grande sea la muestra (o menor número de parámetros a estimar), o sea más grados de libertad, la precisión será mayor y por tanto existe una menor probabilidad de obtener valores lejanos del centro de la distribución. Es decir, entre más grados de libertad menos pesadas las colas. Y por tanto más se parecerán la distribución t y la estándar normal. En la Figuras 3.5, 3.6 y 3.7 se presenta este resultado.

Figura 3.5. Diferencia entre la distribución estándar normal z y la t con 1 grado de libertad t_1



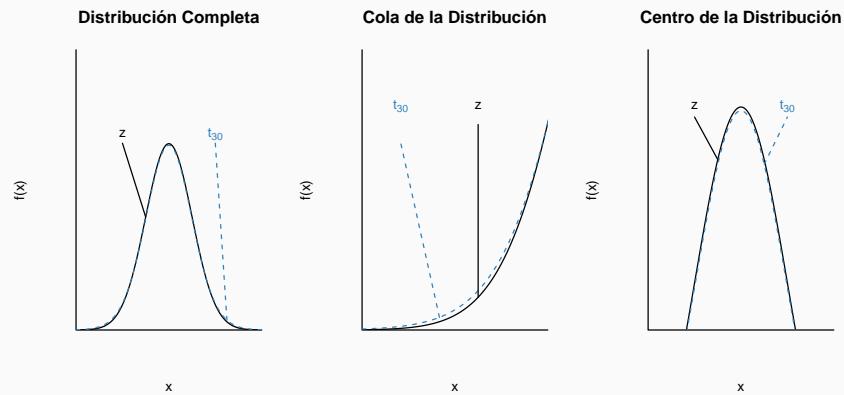
Fuente: Elaboración propia

Figura 3.6. Diferencia entre la distribución estándar normal z y la t con 10 grado de libertad t_{10}



Fuente: Elaboración propia

Figura 3.7. Diferencia entre la distribución estándar normal z y la t con 30 grado de libertad t_{30}



Fuente: Elaboración propia

Así, teniendo en cuenta la estimación de la varianza, tenemos que un intervalo de confianza del $100(1 - \alpha)\%$ para la media poblacional cuando la varianza es desconocida está dado por:

$$\bar{X} \mp t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}} \quad (3.6)$$

donde $t_{n-1, \frac{\alpha}{2}}$ corresponde al valor de la distribución t de Student con $n - 1$ grados de libertad tal que $P(|w| \leq t_{n-1, \frac{\alpha}{2}}) = \alpha$.

^aPor una cola pesada se entiende una probabilidad más alta de estar en las colas.

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 $H_0 : \beta_p = c$; 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:

$$t_c = \frac{\hat{\beta}_p - c}{s_{\hat{\beta}_p}} \quad (3.7)$$

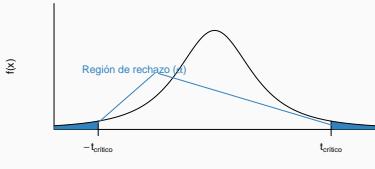
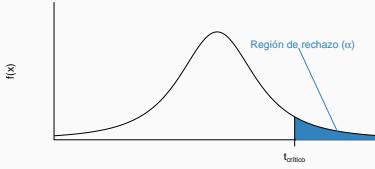
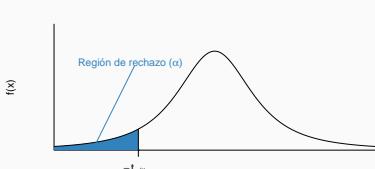
Este estadístico t_c (o t-calculado) sigue una distribución t con $n - k$ grados de libertad; por tanto, se rechazarán la hipótesis nula si el valor absoluto del estadístico t_c es lo suficientemente grande. En otras palabras, si $|t_c| > t_{\frac{\alpha}{2}, n-k}$.

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.

Tipos de pruebas individuales

Las pruebas para coeficientes individuales pueden ser de una o dos colas, dependiendo de la respectiva hipótesis nula y alterna. En el Cuadro 3.1 se presentan las correspondientes zonas de rechazo y reglas de decisión para cada una de los tres posibles casos.

Cuadro 3.1: Hipótesis individuales sobre los coeficientes del modelo MCO: región de rechazo y reglas de decisión.

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

Estructura de una prueba de hipótesis

En general una prueba de hipótesis siempre tiene la siguiente estructura

1. H_0 : Hipótesis Nula (hipótesis que se quiere refutar)
2. H_A : Hipótesis Alterna (hipótesis que se quiere aceptar)
3. Cálculo de un estadístico (depende de la distribución del estimador)
4. Decisión (Comparar el estadístico calculado con un valor crítico de la correspondiente función de distribución o tomar decisión con el valor p)

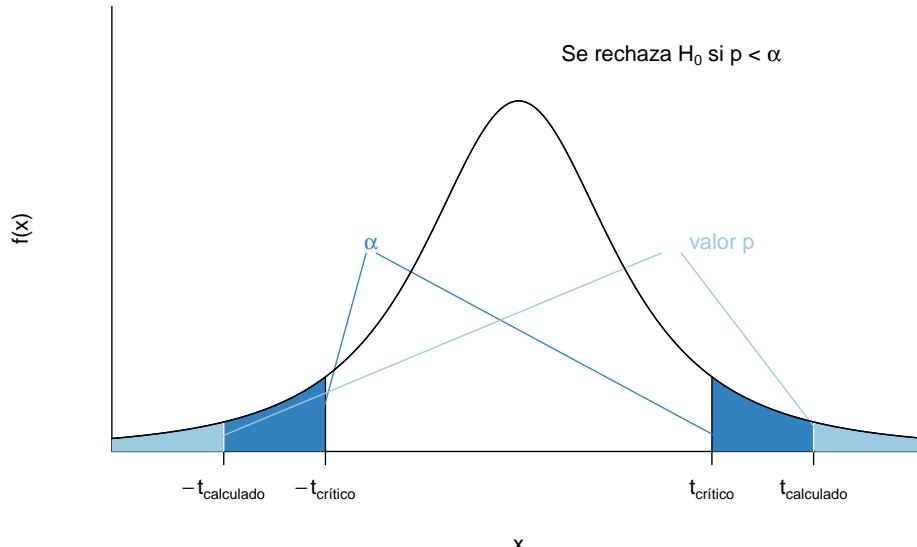
Es importante tener en cuenta que por razones técnicas la hipótesis alterna nunca puede contener un igual. Así, aún si lo que se desea probar es que un $\beta_i = 0$ no es posible asignar esta opción en la hipótesis alterna y tocará asignarla a la hipótesis nula.

Tipos de error de las pruebas de hipótesis

Dado un nivel de significancia α , si se rechaza la hipótesis nula es posible que incorrectamente hayamos rechazado la hipótesis nula cuando esta era verdadera. A este error lo llamamos **error tipo I**. Los estadísticos de pruebas están diseñados para que la probabilidad de ocurrencia del error tipo I sea $\alpha\%$. Supongamos ahora que no es posible rechazar la hipótesis nula, en este caso es factible que no estemos rechazando la hipótesis nula cuando esta en verdad es falsa. En ese caso el error se denomina **error tipo II** y su probabilidad se denominado con la letra griega β . Por tanto, lo ideal al diseñar una prueba de hipótesis es que tanto α como β sean lo más pequeño posible. Lastimosamente existe un compromiso (**trade-off**) entre el error tipo I y el error tipo II; pues cuando se disminuye el error tipo I, el error tipo II se aumenta. Típicamente, el tamaño de β no es conocido a priori.

Es por esta razón, que siempre que se plantea una prueba de hipótesis, se trata de construir la hipótesis nula y la alterna de tal forma que lo que se desea comprobar sea planteado en la hipótesis alterna y no en la hipótesis nula. Así, se pretende cometer un error tipo I mínimo que es más fácilmente controlable por el investigador, dado que los estadísticos para efectuar las pruebas de hipótesis son diseñados para controlar el error tipo I, dejando el error tipo II sin control.

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 3.8).

Figura 3.8. Relación entre el p-valor y el nivel de significancia.

Fuente: Elaboración propia

Empleando este criterio, la decisión de rechazar la hipótesis nula se puede tomar de la forma como se describe en el Cuadro 3.2.

Cuadro 3.2: Criterios para rechazar H_0 empleando el valor p.

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

Fuente: Elaboración propia

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 teórico.

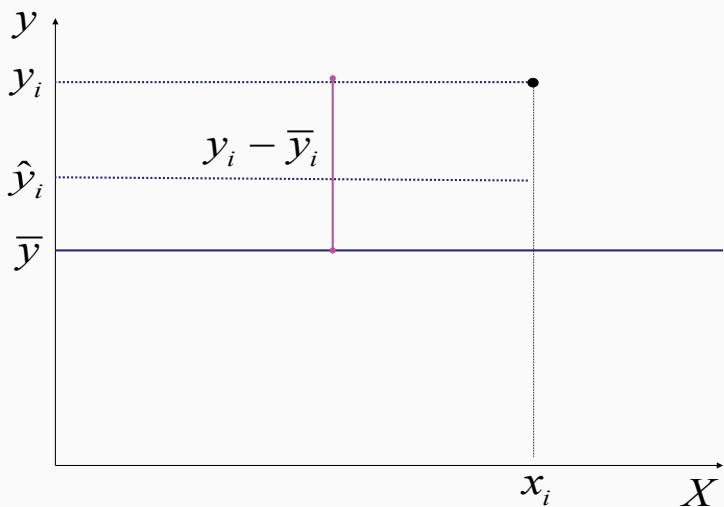
Antes de analizar qué tan bien explica nuestro modelo la variable dependiente, es necesario discutir cómo se puede descomponer la variabilidad de la variable dependiente. Inicialmente consideremos la variación total de la variable dependiente con respecto a su media ($y_i - \bar{y}$), esta variación puede

ser descompuesta en dos partes³. 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 observación⁴ 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. Formalmente, $\hat{\epsilon}_i = (y_i - \hat{y}_i)$. 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 la Figura 3.9.

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, representada en color rosado en la Figura 3.9.

Figura 3.9. Ejemplo de la variación total para una observación de la variable dependiente (y_i)



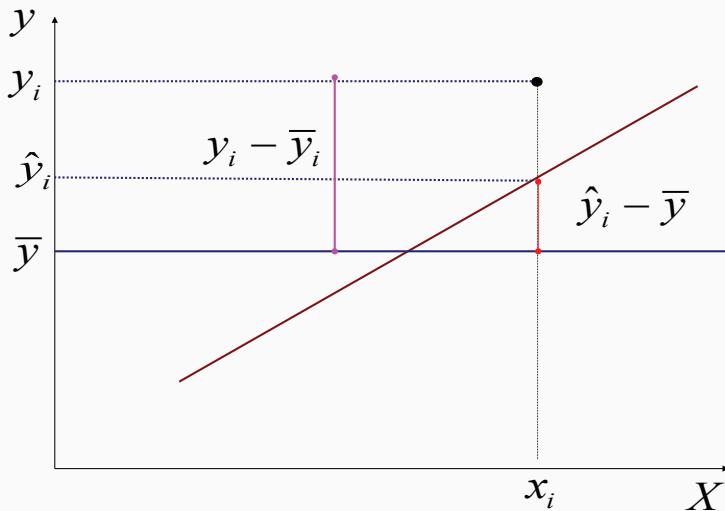
Fuente: Elaboración propia

La variación total se puede descomponer en: i) la parte explicada por el modelo ($\hat{y}_i - \bar{y}$) (Ver Figura 3.10), y ii) la parte que el modelo no explica (el error) $\hat{\epsilon}_i = (y_i - \hat{y}_i)$ (Ver Figura 3.11).

³Noten que partimos de la desviación de dada 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.

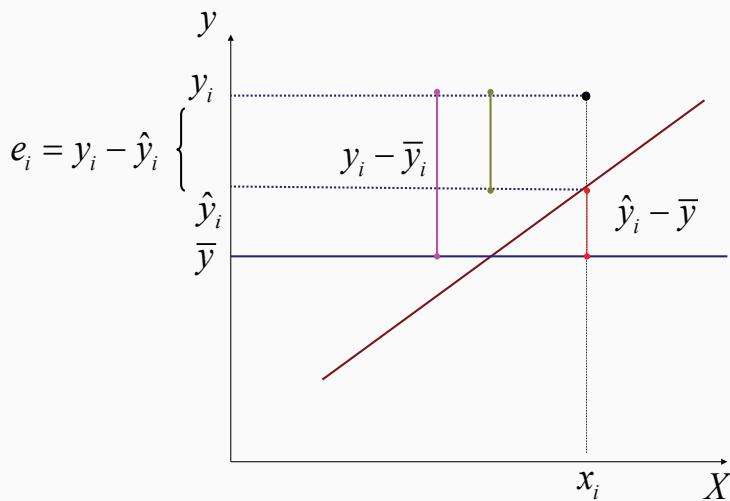
⁴ $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_k X_{ki}$ para $i = 1, 2, \dots, n$, o en forma matricial $\hat{y} = X\hat{\beta}$.

Figura 3.10. Ejemplo de la variación explicada por el modelo para una observación de la variable dependiente (y_i)



Fuente: Elaboración propia

Figura 3.11. Ejemplo de la variación no explicada por el modelo para una observación de la variable dependiente (y_i)



Fuente: Elaboración propia

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:⁵

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.8)$$

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 denominará 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:

$$\begin{aligned} SST &= \sum_{i=1}^n (y_i - \bar{y})^2 = \mathbf{y}^T \mathbf{y} - n\bar{y}^2 \\ SSR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2 \\ SSE &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{y} \end{aligned}$$

Así tendremos que

$$SST = SSR + SSE$$

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 implicaría que todos los puntos se encuentran sobre el plano estimado por el modelo, en caso de tener una sola variable explicativa, este R^2 implicaría 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.

Resumiendo la descomposición de la variabilidad de la variable dependiente

Una forma de resumir la descomposición de la variabilidad de la variable dependiente^a 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

⁵En el Anexo 3.7 al final de este capítulo se presenta una demostración de esta afirmación.

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.

Cuadro 3.3: Tabla ANOVA.

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

Fuente: Elaboración propia

^aEsta es una alternativa a la gráfica que no funcionaría para un modelo con más de dos variables explicativas. Aún en el caso de un modelo con una sola variable dependiente, el análisis gráfico se convierte en una situación inmanejable.

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 $SST \neq SSR + SSE$ (ver Anexo 3.7 al final del capítulo). 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 demanda de un bien Q_i , para lo cual se cuenta con datos 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:

$$Q_i = \beta_0 + \beta_1 p_i + \beta_2 pc_i + \beta_3 ps_i + \beta_4 I_i + \varepsilon_i \quad (3.9)$$

$$\ln(Q_i) = \gamma_0 + \gamma_1 \ln(p_i) + \gamma_2 \ln(pc_i) + \gamma_3 \ln(ps_i) + \gamma_4 \ln(I_i) + \mu_i \quad (3.10)$$

$$Q_i = \beta_0 + \beta_1 p_i + \beta_2 pc_i + \beta_3 ps_i + \beta_4 I_i + \varepsilon_i \quad (3.11)$$

$$Q_i = \beta_0 + \beta_1 p_i + \beta_2 pc_i + \beta_3 ps_i + \beta_4 I_i + \varepsilon_i \quad (3.12)$$

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 3.10 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 3.10 es distinta a la de los otros tres modelos, pues para el modelo 3.10 $SST = \sum_{i=1}^n (\ln(Q_i) - \bar{\ln}(Q))^2$ 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 3.10 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 3.9, 3.11 y 3.12, 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 explicativas, el R^2 será más grande.

Entonces, podríamos escoger cuál modelo explica mayor proporción de la variabilidad de la variable dependiente para los modelos 3.11 y 3.12 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 3.9, 3.11 y 3.12.

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_a . 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 $\mathbf{R}_{(r) \times k} \boldsymbol{\beta}_{k \times 1} = \mathbf{C}_{(r) \times 1}$, donde $\boldsymbol{\beta}^T = [\beta_1 \ \beta_2 \ \dots \ \beta_k]$, $\mathbf{C}^T = [0 \ 0 \ \dots \ 0]$ y

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

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 presentan 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 $\mathbf{R}_{(r) \times k} \boldsymbol{\beta}_{k \times 1} = \mathbf{C}_{(r) \times 1}$, donde $\boldsymbol{\beta}^T = [\beta \ \alpha_1 \ \alpha_2]$, $\mathbf{R} = [0 \ 1 \ 1]$ y $\mathbf{C} = [1]$.

En general, cualquier hipótesis nula que pueda escribirse de la forma $\mathbf{R}_{(r) \times k} \boldsymbol{\beta}_{k \times 1} = \mathbf{C}_{(r) \times 1}$ (versus la H_a : no H_0) se puede comprobar empleando el siguiente estadístico:

$$F_{calculado} = \frac{(\mathbf{C} - \mathbf{R}\hat{\boldsymbol{\beta}})^T (\mathbf{R}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{R}^T)^{-1} (\mathbf{C} - \mathbf{R}\hat{\boldsymbol{\beta}}) / r}{\hat{\varepsilon}^T \hat{\varepsilon} / (n - k)} \quad (3.13)$$

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 $F_{\alpha, (r, n-k)} < F_{Calculado}$.

Un caso especial de la hipótesis nula general $\mathbf{R}_{(r) \times k} \boldsymbol{\beta}_{k \times 1} = \mathbf{C}_{(r) \times 1}$ (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 $F_{calculado}$ se reduce a:

$$F_{Global} = \frac{\hat{\beta}^T \mathbf{X}^T \mathbf{y} - n\bar{y}^2 / (k-1)}{\hat{\epsilon}^T \hat{\epsilon} / (n-k)} = \frac{MSR}{MSE} \quad (3.14)$$

O de manera equivalente:

$$F_{Global} = \frac{R^2 / (k-1)}{(1-R^2) / (n-k)} \quad (3.15)$$

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 como se muestra en el Cuadro 3.4).

Cuadro 3.4: Tabla ANOVA con prueba de significancia global

Fuente de la Variación	SS	Grados de libertad	MS	F-Globals
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	$n-1$		

Fuente: Elaboración propia

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

$$y_i = \beta_1 + \varepsilon_i \quad (3.16)$$

mientras que la hipótesis alterna implica el modelo

$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (3.17)$$

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).⁶ 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 .

$$F_{Global} = \frac{R^2 / (k-1)}{(1-R^2) / (n-k)} \quad (3.18)$$

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:

⁶Es decir $Y_i = \mu + \varepsilon_i$, donde $\mu = \beta_1$ es la media.

$$F_{calculado} = \frac{\left(\mathbf{C} - \mathbf{R}\hat{\beta} \right)^T \left(\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T \right)^{-1} \left(\mathbf{C} - \mathbf{R}\hat{\beta} \right)}{\hat{\epsilon}^T\hat{\epsilon}/(n-k)} \quad (3.19)$$

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. Este estadístico se puede calcular con la siguiente expresión:

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

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 (3.20) tendremos que:

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

$$W_{calculado} = \frac{\left(\mathbf{C} - \mathbf{R}\hat{\beta} \right)^T \left(\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T \right)^{-1} \left(\mathbf{C} - \mathbf{R}\hat{\beta} \right)}{S^2} \quad (3.22)$$

y reemplazando S^2 tenemos que:

$$W_{calculado} = \frac{\left(\mathbf{C} - \mathbf{R}\hat{\beta} \right)^T \left(\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T \right)^{-1} \left(\mathbf{C} - \mathbf{R}\hat{\beta} \right)}{\hat{\epsilon}^T\hat{\epsilon}/(n-k)} \quad (3.23)$$

Al comparar la anterior expresión con (3.19) 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).

3.6 Práctica en R: Explicando los rendimientos de una acción

En esta sección aplicaremos los conceptos cubiertos hasta ahora. En esta ocasión en el área financiera de una organización se está evaluando la inversión de algunos excedentes de tesorería en acciones

del grupo SURA. Así, en el equipo de trabajo surgió la siguiente pregunta de negocio: ¿Cómo está relacionado el rendimiento de esta acción con el rendimiento de aquellas acciones en la que ya tenemos inversiones? Esta pregunta nace, pues se quiere diversificar el riesgo del portafolio. Para responder esta pregunta se cuenta con datos del rendimiento diario⁷ de la acción de Suramericana de otras acciones que también se tranzan en la Bolsa de Valores de Colombia y ya están en el portafolio de la organización. La base de datos contiene rendimientos diarios de las siguientes acciones: GRUPOSURA, ECOPETROL, NUTRESA, EXITO, ISA, GRUPOAVAL, CONCONCRETO, VALOREM y OCCIDENTE. La base de datos va desde 2012-01-02 hasta el 2019-01-14 y y todas las variables están medidas en puntos porcentuales. La base 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 función **load()** que se encuentra en el paquete base de R. El argumento mas importante para esta función es el nombre del archivo y su respectiva ruta. Procedamos a cargar el *working space*.

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

El *working space* contiene el objeto *retornos.diarios*. Miremos la clase de dicho objeto.

```
class(retornos.diarios)

## [1] "xts" "zoo"
```

Este objeto es de clase **xts** y **zoo**. Este tipo de clase es muy empleada cuando se trabaja con series de tiempo; en especial la clase **xts** es útil cuando las series tienen una frecuencia inferior al mes, como en este caso que tenemos datos diarios. Esta clase corresponde al paquete *xts* (Ryan y Ulrich, 2020) . Carguemos la librería (de ser necesario instale el paquete) y asegurémonos que los datos quedaron bien cargados y que cada variable se encuentre bien definida.

```
# install.packages('xts')

library(xts)
# chequeo de los primeros datos
head(retornos.diarios, 2)

##           GRUPOSURA ECOPETROL      NUTRESA      EXITO       ISA
## 2012-01-02  1.277973 -0.3565066 -0.9216655  2.02184181 -1.983834
## 2012-01-03  2.507968  2.0036027 -0.2781643  0.07695268  1.626052
##           GRUPOAVAL CONCONCRET VALOREM OCCIDENTE
## 2012-01-02  0.000000      0  0.0000      0
## 2012-01-03 -2.429269      0 12.5839      0
```

⁷El rendimiento diario se calcula como el crecimiento porcentual en el precio de la acción.

```

str(retornos.diarios)

## An 'xts' object on 2012-01-02/2019-01-14 containing:
##   Data: num [1:1696, 1:9] 1.278 2.508 0.432 -0.432 0.309 ...
##   - attr(*, "dimnames")=List of 2
##     ..$ : NULL
##     ..$ : chr [1:9] "GRUPOSURA" "ECOPETROL" "NUTRESA" "EXITO" ...
##   Indexed by objects of class: [Date] TZ: UTC
##   xts Attributes:
##   NULL

sapply(retornos.diarios, class)

##      GRUPOSURA ECOPETROL NUTRESA EXITO ISA   GRUPOAVAL
## [1,] "xts"      "xts"      "xts"      "xts" "xts"
## [2,] "zoo"       "zoo"       "zoo"       "zoo" "zoo" "zoo"
##      CONCONCRET VALOREM OCCIDENTE
## [1,] "xts"      "xts"      "xts"
## [2,] "zoo"       "zoo"       "zoo"

```

Noten que este objeto de clase **xts** reacciona un poco diferente a los objetos de clase **data.frame** a funciones como **str()**. Pero la mayoría de 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 su periodicidad (función **periodicity()**), las estadísticas descriptivas, un gráfico y la correlación entre las series.

```

periodicity(retornos.diarios)

## Daily periodicity from 2012-01-02 to 2019-01-14

summary(retornos.diarios)

##           Index          GRUPOSURA          ECOPETROL
## Min.   :2012-01-02   Min.   :-5.491576   Min.   :-10.44520
## 1st Qu.:2013-09-25  1st Qu.:-0.605399  1st Qu.: -0.96271
## Median :2015-06-30  Median : 0.000000  Median : 0.00000
## Mean   :2015-07-04  Mean   : 0.002958  Mean   : -0.02184
## 3rd Qu.:2017-04-07  3rd Qu.: 0.679216  3rd Qu.: 1.00884
## Max.   :2019-01-14  Max.   : 5.671623  Max.   : 10.27128
##           NUTRESA          EXITO          ISA
## Min.   :-7.145896   Min.   :-13.35314  Min.   :-15.20399
## 1st Qu.:-0.525538   1st Qu.: -0.76845  1st Qu.: -0.75520

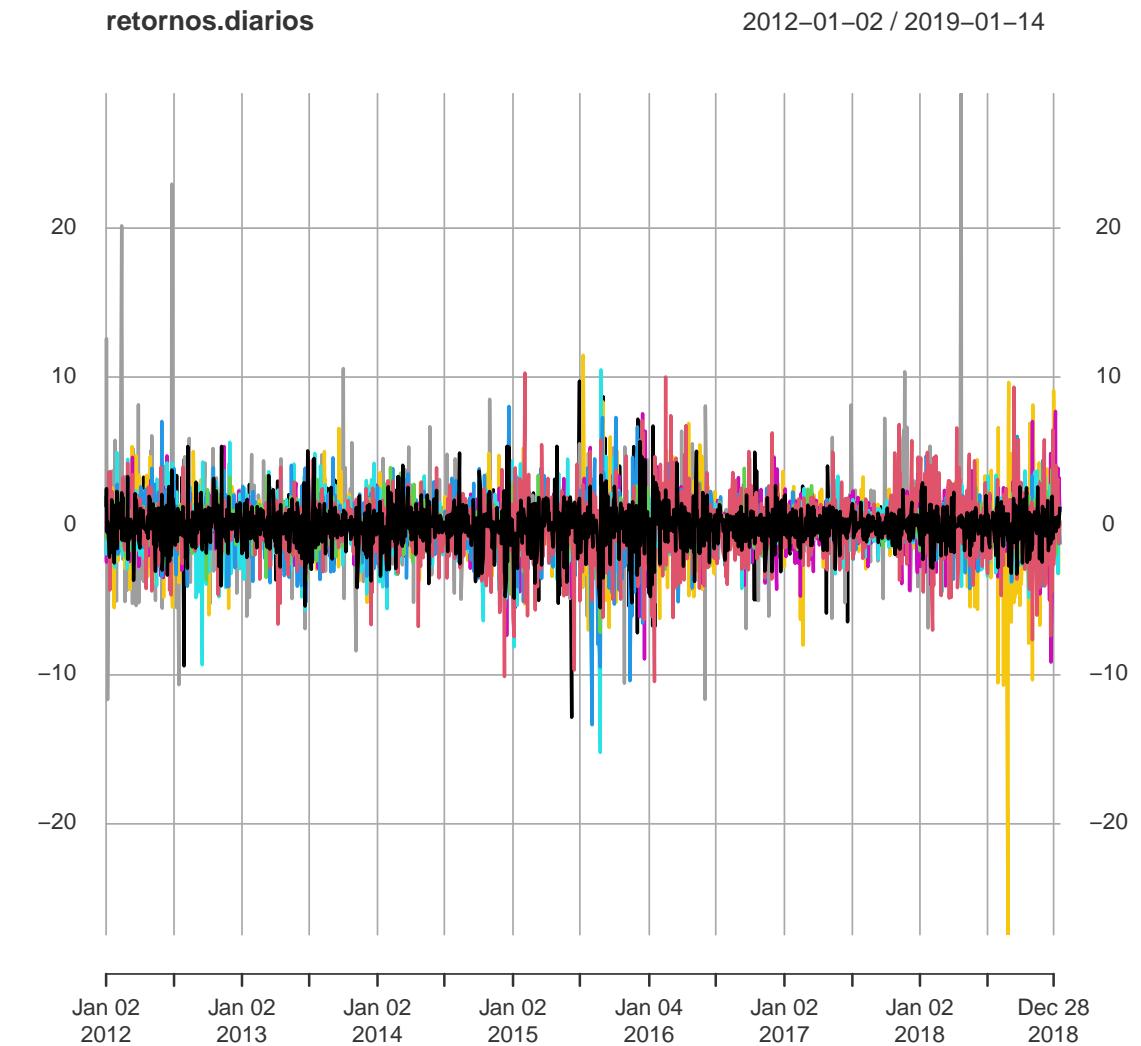
```

```
## Median : 0.000000 Median : 0.000000 Median : 0.000000
## Mean   : 0.004678 Mean   : -0.03991  Mean   : 0.01399
## 3rd Qu.: 0.528802 3rd Qu.: 0.73004  3rd Qu.: 0.85536
## Max.   : 4.807793 Max.   : 8.02808  Max.   : 10.48256
##      GRUPOAVAL      CONCONCRET      VALOREM
## Min.  :-9.143421  Min.  :-27.44368  Min.  :-11.65338
## 1st Qu.:-0.722025 1st Qu.:-0.44445  1st Qu.: 0.00000
## Median : 0.000000 Median : 0.000000 Median : 0.00000
## Mean   :-0.007806 Mean   :-0.07761  Mean   : 0.07358
## 3rd Qu.: 0.719428 3rd Qu.: 0.36597  3rd Qu.: 0.00000
## Max.   : 7.696104 Max.   : 11.46629  Max.   : 29.02523
##      OCCIDENTE
## Min.  :-12.84562
## 1st Qu.: 0.00000
## Median : 0.00000
## Mean   : 0.01409
## 3rd Qu.: 0.00000
## Max.   : 9.72907
```

Todas las series de los rendimientos tienen mediana cero y son relativamente volátiles. Esto se puede corroborar gráficamente. Grafiquemos las series empleando la función **plot.xts()** del paquete **xts**(Ryan y Ulrich, 2020). Al emplear la **plot()** sobre un objeto de clase **xts** automáticamente emplea la función **plot.xts()**.

```
# install.packages('xts')

plot(retornos.diarios)
```



Claramente las series (variables) de los rendimientos son bastante volátiles⁸. Ahora procedamos a

⁸Esta es una característica común en los rendimientos de las acciones. Para una mayor discusión ver Alonso (2006a) y Alonso y Torres (2014).

estimar el siguiente modelo de regresión:

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

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. En el capítulo anterior discutimos como estimar un modelo lineal. En este caso el modelo se puede estimar y visualizar de la siguiente manera:

```
# estimación del modelo
res1 <- lm(GRUPOSURA ~ ., retornos.diarios)
# imprime resultados
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 ***
## CONCONCRET 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
```

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 Cuadro 3.5 se presenta la equivalencia entre lo reportado en R y la literatura.

Cuadro 3.5: Equivalencia entre el valor p, el reporte de R y los asteriscos en la literatura.

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

Fuente: Elaboración propia

Los resultados se resumen en el siguiente Cuadro 3.6.

Cuadro 3.6: Modelo estimado por MCO para la acción de SURA

<i>Dependent variable:</i>	
	GRUPOSURA
ECOPETROL	0.121*** (0.014)
NUTRESA	0.141*** (0.028)
EXITO	0.122*** (0.018)
ISA	0.188*** (0.019)
GRUPOAVAL	0.084*** (0.020)
CONCONCRET	0.021 (0.015)
VALOREM	0.036** (0.014)
OCCIDENTE	0.031 (0.027)
Constant	0.006 (0.027)
Observations	1,696
R ²	0.262
Adjusted R ²	0.258
Residual Std. Error	1.109 (df = 1687)
F Statistic	74.748*** (df = 8; 1687)

Note: *p<0.1; **p<0.05; ***p<0.01

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 CONCONCRETO y OCCIDENTE no tienen efecto sobre el rendimiento de la acción de GRUPOSURA.

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.

Por otro lado el F_{Global} es 74.748. El respectivo p-valor (aproximadamente 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()** del paquete principal de R. El único argumento necesario, por ahora, es el objeto clase **lm** que tenga los resultados de la regresión.

```
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 CONCONCRETO 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 . Esta hipótesis nace de los resultados de las pruebas individuales. 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** (Kleiber y Zeileis, 2008). Esta función requiere al menos dos argumentos: El modelo

(objeto clase `lm`) 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("CONCONCRET = 0", "OCCIDENTE = 0"))

## Linear hypothesis test
##
## Hypothesis:
## CONCONCRET = 0
## OCCIDENTE = 0
##
## Model 1: restricted model
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCONCRET +
##           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("CONCONCRET = 0", "OCCIDENTE = 0"))

## Linear hypothesis test
##
## Hypothesis:
## CONCONCRET = 0
## OCCIDENTE = 0
##
## Model 1: restricted model
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + CONCONCRET +
##           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 los rendimientos de las acciones de CONCONCRETO y OCCIDENTE no tienen efecto sobre el rendimiento de la acción de GRUPOSURA. Así el nuevo

modelo será:

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

El nuevo modelo⁹ y el anterior se presentan en el Cuadro 3.7.

Cuadro 3.7: Modelos estimados por MCO para la acción de SURA

<i>Dependent variable:</i>		
GRUPOSURA		
	(1)	(2)
ECOPETROL	0.121*** (0.014)	0.122*** (0.014)
NUTRESA	0.141*** (0.028)	0.145*** (0.028)
EXITO	0.122*** (0.018)	0.123*** (0.018)
ISA	0.188*** (0.019)	0.188*** (0.019)
GRUPOAVAL	0.084*** (0.020)	0.086*** (0.020)
CONCONCRET	0.021 (0.015)	
VALOREM	0.036** (0.014)	0.036** (0.014)
OCCIDENTE	0.031 (0.027)	
Constant	0.006 (0.027)	0.005 (0.027)
Observations	1,696	1,696
R ²	0.262	0.260
Adjusted R ²	0.258	0.258
Residual Std. Error	1.109 (df = 1687)	1.110 (df = 1689)
F Statistic	74.748*** (df = 8; 1687)	99.025*** (df = 6; 1689)

Note: *p<0.1; **p<0.05; ***p<0.01

⁹El nuevo modelo lo nombraremos `res2`.

Ahora, noten que en este segundo modelo los coeficientes que acompañan los rendimientos de las acciones de ECOPETROL y EXITO parecen “a ojo” ser similares. Entonces probemos si realmente son iguales o no. Es decir, la hipótesis nula será:

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

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 + VALOREM
##
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1   1691 2080.8
## 2   1689 2080.8  2  0.046964 0.0191 0.9811
## Linear hypothesis test
##
## Hypothesis:
## ECOPETROL = 0.12
## EXITO = 0.12
##
## Model 1: restricted model
## Model 2: GRUPOSURA ~ ECOPETROL + NUTRESA + EXITO + ISA + GRUPOAVAL + VALOREM
##
##   Res.Df   RSS Df Sum of Sq Chisq Pr(>Chisq)
## 1   1691 2080.8
## 2   1689 2080.8  2  0.046964 0.0381      0.9811
```

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.

Finalmente noten que el intercepto del modelo no es significativo. No obstante no removeremos nunca el intercepto de un modelo para mantener las propiedades estadísticas del modelo y poder interpretar el R^2 .

Recordemos que la pregunta de negocio que queríamos responder es: ¿cómo está relacionado el rendimiento de la acción de SURA con el rendimiento de aquellas acciones en la que ya tenemos inversiones? Nuestros resultados permiten concluir que el rendimiento de la acción de SURA tienen relación con los rendimientos de las acciones de ECOPETROL, NUTRESA, EXITO, ISA, GRUPOAVAL y VALOREM.

Para terminar, interpretemos los coeficientes.

- Intercepto: Si el rendimiento de las otras acciones fuesen cero, entonces el rendimiento de la acción de SURA será cero.
- Por cada punto porcentual¹⁰ (pp) que aumente el rendimiento de la acción de ECOPETROL, el rendimiento de la acción de SURA aumentará en 0.12 pp.
- Por cada punto porcentual (p.p.) que aumente el rendimiento de la acción de NUTRESA, el rendimiento de la acción de SURA aumentará en 0.14 pp.
- Por cada punto porcentual (p.p.) que aumente el rendimiento de la acción de EXITO, el rendimiento de la acción de SURA aumentará en 0.12 pp.
- Por cada punto porcentual (p.p.) que aumente el rendimiento de la acción de ISA, el rendimiento de la acción de SURA aumentará en 0.19 pp.
- Por cada punto porcentual (p.p.) que aumente el rendimiento de la acción de GRUPOAVAL, el rendimiento de la acción de SURA aumentará en 0.19 pp.
- Por cada punto porcentual (p.p.) que aumente el rendimiento de la acción de VALOREM, el rendimiento de la acción de SURA aumentará en 0.19 pp.

Ejercicios

3.1 Replica el experimento de Monte Carlo que se realiza en la Introducción para muestras de tamaño 20, 50 y 200. ¿Qué puedes concluir?

3.7 Anexo: Demostración de la ecuación 3.8

La expresión 3.8 implica que en presencia de un modelo lineal con intercepto, podremos 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:

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

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

$$\hat{\epsilon}^T \hat{\epsilon} = y^T y - \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta}.$$

¹⁰Nota que no es lo mismo un aumento de un uno porciento que un aumento de un punto porcentual. Ten cuidado con esta diferencia al momento de interpretar coeficientes asociados con variables que están medidas en puntos porcentuales.

Reemplazando, obtenemos

$$\mathbf{y}^T \mathbf{y} = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} + \hat{\mathbf{y}}^T \hat{\mathbf{y}}.$$

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

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

En otras palabras, tenemos que:

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

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

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

$$SSR = \hat{\mathbf{y}}^T \hat{\mathbf{y}} - 2\bar{y} \sum_{i=1}^n \hat{y}_i + n\bar{y}^2.$$

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

$$SSR = \hat{\mathbf{y}}^T \hat{\mathbf{y}} - 2\bar{y} \sum_{i=1}^n (y_i - \hat{\boldsymbol{\varepsilon}}_i) + n\bar{y}^2$$

$$SSR = \hat{\mathbf{y}}^T \hat{\mathbf{y}} - 2\bar{y} \left[\sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_i \right] + n\bar{y}^2$$

$$SSR = \hat{\mathbf{y}}^T \hat{\mathbf{y}} - 2\bar{y}n\bar{y} + 2\bar{y} \sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_i + n\bar{y}^2.$$

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

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

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

Es decir, se tiene que:

$$SST = SSE + SSR.$$

Es importante anotar que si no se garantiza que $\sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_i = 0$, 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.

4 . Comparación de Modelos

Diseñado por Freepik

Objetivos del capítulo

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

- Explicar en sus propias palabras la intuición detrás de las métricas R^2 ajustado, el *AIC* y el *BIC* para comparar modelos.
- Emplear el R^2 ajustado, el *AIC* y el *BIC* para seleccionar el mejor modelo.
- Explicar en sus propias palabras la intuición detrás de las pruebas estadísticas que permitan seleccionar entre modelos anidados y no anidados empleando R.
- Realizar pruebas estadísticas que permitan seleccionar entre modelos anidados empleando R.
- Realizar pruebas estadísticas que permitan seleccionar entre modelos no anidados empleando R.

4.1 Introducción

Con frecuencia, 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_t 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 más 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 más cercano, la inversión propia en televisión, radio e Internet, respectivamente.

Noten que el modelo 4.2 es una versión restringida del modelo 4.1; pues si $\beta_4 = \beta_5 = 0$ en el modelo 4.1 entonces se obtiene el modelo restringido 4.2. Al modelo 4.2 se le denomina modelo anidado (en el modelo 4.1). Por otro lado, el modelo 4.3 no se encuentra anidado en los modelos 4.2 o 4.1. Es decir, no hay forma de restringir los modelos 4.1 o 4.2 para obtener el modelo 4.3.

4.2 Comparación de modelos empleando medidas de bondad de ajuste

En el capítulo anterior (ver 3.3) 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 además 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 ($k - 1$), 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. $\frac{n-1}{n-k}$ decrece con el aumento de k (inclusión de variables). En otras palabras, a medida que se incrementa el número de variables

independientes en el modelo, $\frac{n-1}{n-k}$ se hace más pequeño al mismo tiempo que el R^2 aumenta. Por tanto $(1 - R^2)$ disminuye a medida que se incluyen variables, pero el factor $\frac{n-1}{n-k}$ se hace más pequeño. 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 incrementará únicamente si el aumento en el R^2 es lo suficientemente grande para compensar la penalización que se hace por incluir más variables. Por lo tanto, un \bar{R}^2 grande será mejor que uno pequeño.

De esta manera, con el \bar{R}^2 se pueden comparar los modelos 3.9, 3.11 y 3.12, siempre y cuando se emplee la misma muestra para estimar los tres modelos (el mismo n y SST). El único problema que se presenta con el \bar{R}^2 , es que este carece de una interpretación a diferencia del R^2 .

Existen otros problemas, que se estudiarán más adelante, que pueden “inflar” el R^2 como la multicolinealidad (ver 8). De esta manera, aunque este estadístico presenta una interpretación muy clara e intuitiva, hay que tener cuidado antes de sacar conclusiones de este estadístico y del \bar{R}^2 .

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 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 más 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 más 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 alta de seleccionar modelos con mas 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. Adicionalmente, 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 las tres medidas (o métricas) (\bar{R}^2 , *AIC*, *BIC*) de bondad de ajuste; pero los resultados de cada uno de estos criterios no necesariamente concordarán. Por ejemplo, supongamos un caso en el que *AIC* recomienda un modelo con 5 variables,

¹Este 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.

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.

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 podrán probar con una prueba F o con una prueba de Wald (ver sección 3.5).

De hecho la comparación de modelos anidados es un caso especial de dichas pruebas. Es decir, estamos comparando modelos anidados cuando consideramos una hipótesis nula en la cual un grupo de coeficientes² es conjuntamente igual a cero, es decir $H_0 : \beta_p = \beta_{p+1} = \dots = \beta_l = 0$, para $0 < p < l \leq k$ (versus la $H_A : \text{No } H_0$). Esta prueba se conoce como una prueba de significancia conjunta. Es decir, la hipótesis nula es equivalente a $H_0 : Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_{p-1} X_{p-1i} + \beta_{l+1} X_{l+1i} + \dots + \beta_k X_{ki} + \varepsilon_i$ (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 $F_{\text{Calculado}}$ de la expresión 3.13 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 SSE_U representa la suma de los cuadrados de los residuos estimados del modelo sin restringir.

Así, para probar la hipótesis nula que $H_0 : \beta_p = \beta_{p+1} = \dots = \beta_l = 0$, para $0 < p < l \leq k$ (versus la $H_A : \text{No } H_0$), 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 $F_{\text{Calculado}}$, 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 discutió en el capítulo anterior, este tipo de restricciones también se pueden probar con una prueba de Wald (función `waldtest()` del paquete *AER* (Kleiber y Zeileis, 2008)).

4.3.2 Modelos no anidados

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

$$H_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$$

²Diferentes al término constante.

y la hipótesis alterna es

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

Para comprobar este tipo de hipótesis tenemos dos opciones: la prueba J y la prueba de Cox. Estas opciones se discuten a continuación.

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 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.

Prueba de Cox

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

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

donde

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

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

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

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

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

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.

```
datos <- read.table("../Data/selModel.txt", header = TRUE, sep = ",")  
head(datos, 2)  
  
##   X     x1     x2     x3     x4     x5     x6     x7     x8     x9     x10  
## 1 1 5.925 6.512 6.883 5.157 6.420 5.833 5.356 6.626 5.730 6.108  
## 2 2 4.565 3.638 4.378 5.056 3.165 4.598 4.576 4.171 4.259 4.443  
##  
##      y  
## 1 42.799  
## 2 30.372  
  
class(datos)  
  
## [1] "data.frame"  
  
str(datos)  
  
## '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 cargó una primera columna que corresponde al número de la observación, esto no lo necesitaremos. Eliminemos esa variable.

```
datos <- datos[, -1]  
head(datos, 3)  
  
##     x1     x2     x3     x4     x5     x6     x7     x8     x9     x10
```

```

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

str(datos)

## 'data.frame': 150 obs. of 11 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 ...
## $ y   : num  42.8 30.4 36.3 36.6 38.3 ...

```

Todas las variables cargadas son numéricas. Ahora consideraremos 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 está 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 estimar los tres modelos y compararlos empleando las métricas 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, datos)
res2 <- lm(y ~ x1 + x2 + x3, datos)
res3 <- lm(y ~ x4 + x5 + x8 + x9 + x10, datos)

```

Los resultados de estos tres modelos se presentan en el Cuadro 4.1.

Cuadro 4.1: Modelos estimados por MCO

<i>Dependent variable:</i>			
	y		
	(1)	(2)	(3)
x1	0.953*** (0.270)	1.048*** (0.266)	
x2	1.936*** (0.295)	2.133*** (0.274)	
x3	1.038*** (0.290)	1.162*** (0.276)	
x6	0.412 (0.272)		
x7	0.167 (0.287)		
x4		1.429*** (0.306)	
x5		1.533*** (0.299)	
x8		0.443 (0.319)	
x9		0.175 (0.316)	
x10		0.589* (0.333)	
Constant	12.390*** (1.492)	13.176*** (1.427)	14.033*** (1.593)
Observations	150	150	150
R ²	0.651	0.643	0.566
Adjusted R ²	0.639	0.636	0.551
Residual Std. Error	2.722 (df = 144)	2.734 (df = 146)	3.036 (df = 144)
F Statistic	53.717*** (df = 5; 144)	87.677*** (df = 3; 146)	37.527*** (df = 5; 144)

Note: *p<0.1; **p<0.05; ***p<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 *slot* que contiene a \bar{R}^2 .

```
R1 <- summary(res1)
attributes(R1)

## $names
## [1] "call"          "terms"        "residuals"
## [4] "coefficients" "aliased"      "sigma"
## [7] "df"            "r.squared"    "adj.r.squared"
## [10] "fstatistic"   "cov.unscaled"
##
## $class
## [1] "summary.lm"

R1$adj.r.squared

## [1] 0.6388607
```

Los criterios de información se calculan empleando las funciones **AIC()** y **BIC()** del paquete base de R. Las dos funciones tienen como único argumento el objeto que contiene la regresión (objeto de clase **lm**).

```
AIC(res1)

## [1] 733.9571

BIC(res1)

## [1] 755.0315
```

Puedes calcular estos indicadores para los otros modelos. Los resultados que se reportan en el Cuadro 4.2.

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

Modelos anidados

Como se mencionó anteriormente, el modelo 4.5 está anidado en el modelo 4.4. Comparemos estos dos modelos empleando la función **anova()**. Esta función ya la había empleado en la sección 3.6, en ese momento solo empleamos un argumento (objeto de clase **lm**) para obtener la tabla ANOVA. Esta misma función permite efectuar una prueba F de modelo restringido versus modelo no restringido si se emplean dos argumentos: primero el modelo restringido (objeto de clase **lm**) y segundo el modelo sin restringir³ (objeto de clase **lm**). Empleando esta función encontramos lo siguiente⁴:

```
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).

Modelos no anidados

Ahora comparemos el modelo 3 (4.6) con los otros dos modelos. La prueba J se calcula empleando la función **jtest()** del paquete *AER* (Kleiber y Zeileis, 2008). Esta función solo tiene dos argumentos, dos objetos de clase **lm** que contienen los dos modelos no anidados. En este caso el orden de los argumentos no es importante para efectuar la prueba pero si para analizar sus resultados como veremos a continuación.

³En caso de cometer un error y asignar como primer argumento el modelo sin restringir, observaras que los grados de libertad serán negativos al igual que la suma cuadrada. Esto claramente es imposible. Nota que en todo caso el F estadístico es el correcto así como su respectivo valor p.

⁴Cuando emplees esta función hay que tener un poco de cuidado. Por defecto, esta función le asigna el nombre de “Model 1” al primer argumento (modelo con restricción) y “Model 2” al segundo argumento (modelo sin restricción). Claramente este es un rotulo arbitrario que es asignado por la función y nada tiene que ver que la numeración le hayas asignado a tus modelos en el análisis, como es nuestro caso en este ejemplo.

```

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 4.4 (este modelo fue el primer argumento de la función **jtest()**) es mejor que el modelo 4.6 (el segundo argumento empleado en dicha función). Este resultado permite rechazar la nula (valor p de 0), lo cuál significa que la prueba no puede concluir en favor del modelo 1. Es decir, hay suficiente evidencia para rechazar la hipótesis nula de que el modelo 4.4 (“Model 1” en la salida de R) es mejor que el modelo 4.6 (“Model 2” en la salida).

Por otro lado, el segundo $t_{calculado}$ (8.58) es el que permite probar la hipótesis nula de que el modelo 4.6 es mejor que el modelo 4.4. En este caso también se rechaza la nula (valor p de 0). Es decir, existe suficiente evidencia para rechazar la nula de que el modelo 4.6 (“Model 2” en la salida de R) es mejor que el modelo 4.4 (“Model 1” en la salida). Esto significa que la prueba presenta una contradicción o dicho de otra manera, esta prueba no es concluyente para este caso.

Para el caso de la comparación del modelo 4.5 y el modelo 4.6 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

```

En conclusión, para esta muestra la prueba J no es concluyente al momento de comparar los modelos no anidados.

Ahora empleemos la prueba de Cox. Esta prueba se puede calcular empleando la función **coxttest()** del paquete *AER* (Kleiber y Zeileis, 2008) . Esta función solo tiene dos argumentos, dos objetos de clase **lm** que contienen los dos modelos no anidados.

```
Cox.res1.3 <- coxttest(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 <- coxttest(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
```

La interpretación de la salida de esta prueba en R es similar a la prueba J. En este caso, pero no necesariamente siempre, los resultados de la prueba Cox son los mismos que obtuvimos con la prueba J (no siempre coinciden los resultados de estas dos pruebas). Por tanto, no es posible concluir en favor de un modelo.

De esta manera, uniendo todos los resultados encontramos que el 4.6 no es mejor que los 4.4 y 4.5. Este resultado lo obtenemos de las métricas de bondad de ajuste, pues con las pruebas Cox y J no podemos derivar una conclusión. Y entre el modelo 4.4 y 4.5, la prueba estadística y los dos criterios de información sugieren que el modelo 4.5 es mejor modelo. En el caso del \bar{R}^2 no existe una diferencia muy grande entre los dos modelos. Dado que las pruebas estadísticas nos permiten tener un 99 % de confianza de que el modelo 4.5 es mejor que el modelo 4.4, por eso es aconsejable seleccionar como mejor modelo 4.5. Es decir, el mejor modelo será:

$$y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \varepsilon_i.$$

Ejercicios

4.1 Continuemos con el ejercicio del final del Capítulo 2. En ese caso se estimó un modelo con todas las variables disponibles en la base de datos como posibles variables explicativas. No obstante diferentes científicos de datos han llegado a diferentes posibles modelos para explicar los ingresos del sector (I medidos en millones de dólares). Estos posibles modelos son:

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \alpha_4 LDies_t + \alpha_5 LEI_t + \alpha_6 V_t + \varepsilon_t. \quad (4.7)$$

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

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 LEI_t + \varepsilon_t \quad (4.9)$$

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \alpha_4 V_t + \varepsilon_t \quad (4.10)$$

$$I_t = \alpha_1 + \alpha_2 CE_t + \alpha_3 CD_t + \alpha_4 LDies_t + \alpha_5 LEI_t + \varepsilon_t. \quad (4.11)$$

Tu tarea es determinar cuál es el mejor modelo para explicar los ingresos del sector ferroviario e interpretar los resultados. Emplea los mismo datos que se encuentran en el archivo regmult.csv.

■

Parte II

Extendiendo el modelo clásico de regresión múltiple

5 . Variables dummy

Diseñado por Freepik

Objetivos del capítulo

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

- Explicar en sus propias palabras cómo se pueden emplear variables dummy en un modelo de regresión múltiple para probar diferentes hipótesis de trabajo.
- Crear variables dummy utilizando R.
- Estimar modelos econométricos con variables dummy que permitan comprobar diferentes hipótesis.

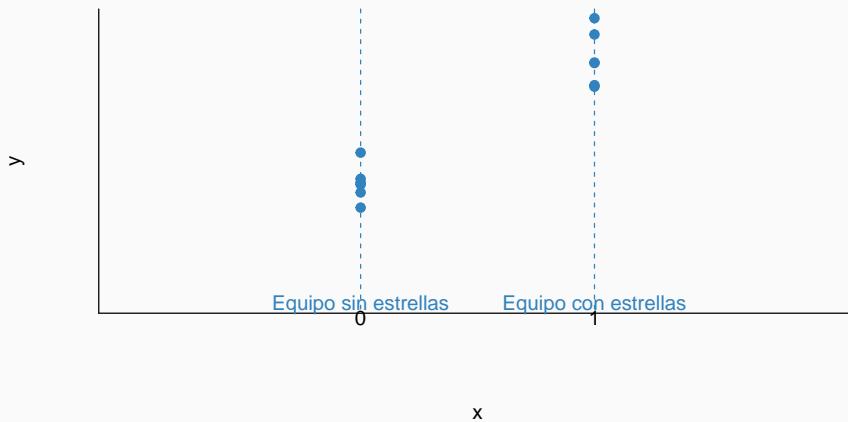
5.1 Introducción

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, podría ser necesario incluir en la estimación de un modelo una variable que represente 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, o si la observación corresponde a un trimestre o no.

Un ejemplo sencillo del uso de variables Dummy

Un gerente de un equipo de fútbol desea saber si el número de entradas vendidas para los partidos de local (Y_i) difiere si el equipo contrario tiene en su alineación una estrella o no. El científico de datos que tiene a su cargo dar respuesta a esta pregunta de negocio se cuenta con una muestra de n partidos y solamente dos variables: Y_i el número de boletas vendidas cuando el equipo es local y una variable X_i que toma el valor de uno si juega una estrella en el equipo contrario, cero en caso contrario. Una base de datos de estas características se vería algo similar a lo presentado en la Figura 5.1.

Figura 5.1. Ventas de boletas por partido (Y_i) y presencia de una estrella en el equipo visitante (X_i).



Fuente: Elaboración propia

Para este caso, se podría construir el siguiente modelo:

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

Recordemos que $X_i = 1$ si juega una estrella en el equipo contrario, cero en caso contrario. En este caso la variable explicativa es una variable cualitativa y no cuantitativa como lo habíamos visto en los capítulos anteriores.

Esta especificación del modelo nos permite modelar las ventas de boletos de diferente manera para las dos situaciones posibles. Es como si tuviésemos dos modelos en uno. Si juega una

estrella en el equipo contrario ($X_i = 1$) el modelo será

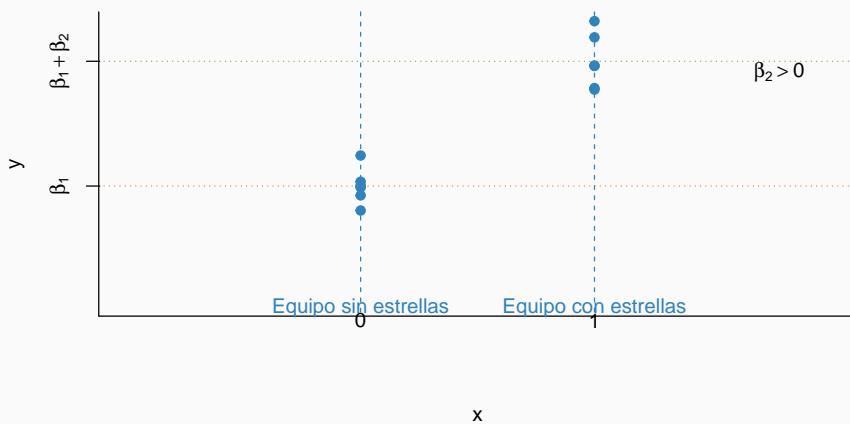
$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i.$$

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

$$Y_i = \beta_1 + \varepsilon_i.$$

Noten que β_2 representa la diferencia en las ventas (promedio) de entradas a los partidos cuando en el equipo contrario juega una súper estrella. De esta manera, la pregunta de negocio se puede resolver estimando el modelo planteado y comprobando inicialmente, si $\beta_2 = 0$ o no. Si se rechaza la nula de que $\beta_2 = 0$ se puede proceder a constatar (por medio de pruebas de hipótesis) que la diferencia entre la presentación de una estrella y no tenerla sí es positiva ($H_0 : \beta_2 \leq 0$). La Figura 5.2 muestra una representación gráfica de este modelo.

Figura 5.2. Esquema del modelo estimado por MCO.



Fuente: Elaboración propia

En este capítulo vamos a discutir las variables dummy, también conocidas como variables ficticias, variables indicador o variables dicotómicas. Este tipo de variables pueden ser muy útiles para modelar diferentes fenómenos. Por ejemplo, las variables dummy pueden servir para:

- medir el efecto de “tratamientos” o “intervenciones” sobre la variable de respuesta
- “borrar” el efecto de una observación atípica (ésta es una mejor opción que eliminar las observaciones anormales (outliers))
- incluir un conjunto de categorías
- agrupar las observaciones en diferentes grupos o regiones
- efectos de umbral (Threshold Effects)
- detectar un cambio estructural
- modelar la estacionalidad

Veamos un ejemplo de cómo las variables dummy permiten encontrar el efecto “tratamiento”. 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 “tratamientos” diferentes. 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:

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

Puedes constatar que este modelo implica un rendimiento promedio para cada día de la semana laboral diferente.

Ahora, concentrémonos en un ejemplo de cómo emplear las variables ficticias para medir el efecto de umbral. Por ejemplo, supongamos que se desea encontrar los determinantes del salario de los empleados de una compañía (salario_i). Típicamente se emplearía como variables explicativas la edad del empleado como una proxy de la experiencia y los años de educación. Pero podría encontrarse una situación en la cual los años de educación (como una variable continua) no tiene efecto directo sobre el salario, más bien superar unos umbrales de educación si pueden tener un efecto. En este caso el siguiente modelo refleja este efecto de umbral:

$$\text{salario}_i = \beta_1 + \beta_2 \text{age}_i + \delta_1 B_i + \delta_2 M_i + \delta_3 P_i + \varepsilon_i,$$

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. Puede demostrar fácilmente cómo este modelo captura un efecto de umbral de los años de educación sobre el ingreso.

Veamos otro ejemplo de efecto de umbral. En algunas ocasiones las ventas de un producto en el periodo t (V_t) pueden reaccionar de manera diferente si el precio (p_t) sube o baja. Esto se conoce como un comportamiento asimétrico. El modelo de regresión múltiple implica que la reacción de la variable dependiente (en este caso V_t) es el mismo cuando una de las variables explicativas aumenta o disminuye. Es decir, el comportamiento es simétrico.

Para capturar este comportamiento asimétrico podemos emplear el siguiente modelo:

$$V_t = \alpha_1 + \alpha_2 p_t + \delta_1 D_t p_t + \varepsilon_t,$$

donde $D_t = 1$ si en el periodo t el precio subió y cero en caso contrario. En este caso puedes confirmar que el efecto de un aumento en el precio (p_t) será diferente a cuando el precio disminuye. Este tipo de efectos también se conocen como un efecto de umbral, siendo el umbral cero. Es decir, $D_t = 1$ si $p_t - p_{t-1} > 0$. En otras palabras, el efecto de umbral corresponde a cuando es importante tener en cuenta en el modelo el hecho de que una variable supera un determinado valor (umbral).

¹Ver Alonso y Berggrun (2011) para una introducción al modelo CAPM y Alonso y Gallo (2013) para una discusión del DOW.

²Ver Alonso y Berggrun (2011) para una introducción al modelo CAPM.

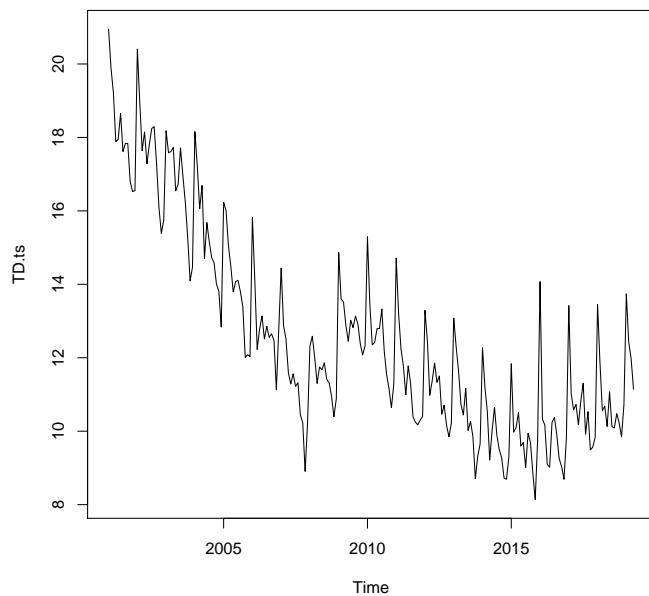
Continuemos con un ejemplo de cómo emplear las variables dummy para capturar cambio estructural. Por cambio estructural se entiende un cambio en cómo se comporta el DGP a partir de un periodo determinado. Por ejemplo, el consumo en un país en el periodo t (C_t) puede cambiar su comportamiento después 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:

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

Sería conveniente que corroboraras que este modelo implica un cambio tanto en la parte del consumo que no depende del ingreso (intercepto), así como de aquella que si depende de éste (pendiente).

Finalmente, veamos un ejemplo de cómo emplear variables dicotómicas para modelar la estacionalidad de una serie de tiempo (para una discusión amplia sobre este tema ver Alonso y Hoyos (2021)). En la Figura 5.3 se presenta la tasa de desempleo mensual para Colombia. En esta serie se observa una estacionalidad marcada. La estacionalidad es el comportamiento repetitivo que se encuentra en una serie de tiempo en el mismo periodo al interior de un año (para una discusión amplia sobre este tema ver Alonso y Hoyos (2021)).

Figura 5.3. Tasa de desempleo mensual en Colombia



Fuente: DANE y Elaboración propia

Podemos emplear las variables dicotómicas para desestacionalizar la tasa de desempleo mensual (TD_t) empleando el siguiente modelo:

$$TD_t = \beta_0 + \beta_1 Ene_t + \beta_2 Feb_t + \beta_3 Mar_t + \cdots + \beta_{12} Nov_t + \varepsilon_t,$$

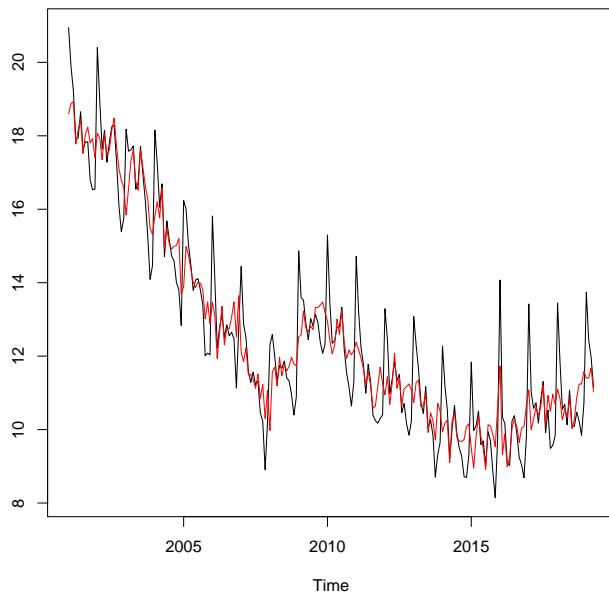
donde Ene_t es una variable dummy que toma el valor de uno si el periodo t coincide con el mes de enero y cero en caso contrario. De manera análoga, $Feb_t = 1$ si el mes es febrero y cero en caso contrario. Y de manera similar tendremos variables dummy para los otros meses.

Si se estima dicho modelo, la serie desestacionalizada ($TD_t^{desestacionalizada}$) se puede obtener restando el efecto de cada uno de los meses a la serie. Es decir,

$$TD_t^{desestacionalizada} = TD_t - [\hat{\beta}_1 Ene_t + \hat{\beta}_2 Feb_t + \hat{\beta}_3 Mar_t + \cdots + \hat{\beta}_{12} Nov_t].$$

En la Figura 5.4 se presenta la tasa de desempleo observada y la desestacionalizada que fue obtenida bajo el procedimiento descrito arriba³.

Figura 5.4. Tasa de desempleo mensual observada y desestacionalizada en Colombia



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. Esto se hace para evitar lo que se conoce como la “trampa de las variables dummy”.

Es decir, ustedes notarán que en el caso de los días de la semana laboral (efecto DOW) se crearon 4 variables dicotómicas y no 5. Y así para todos los ejemplos descritos anteriormente. En el Capítulo 8

³Existen otras formas de desestacionalizar una serie, pero estas están por fuera del alcance de este libro. Para una discusión de las técnicas de desestacionalizar, se puede consultar Alonso y Hoyos (2021).

se explicará con mayor claridad el porqué hacemos esto⁴. 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 que solo emplea una variable independiente. En esta ocasión emplearemos un modelo para explicar las cantidades vendidas (*sell-out*) de pasta dental de una marca famosa del país que llamaremos “DBLANCOS” (para mantener el anonimato de la información) por el canal tradicional (tiendas de barrio) en el mes t (Q_t). Para lograr este objetivo, el equipo de analítica solo cuenta con una variable continua: la inversión en material promocional que se entrega a las tiendas minoristas en el mes t (I_t). Por otro lado, el equipo de analítica al realizar su entendimiento del negocio, ha recibido información del departamento de mercadeo sobre un posible cambio en el comportamiento de las cantidades vendidas en meses donde se venden paquetes de pague 2 lleve 3 frente a períodos en los que no hay ninguna promoción. Afortunadamente se cuenta con información sobre aquellos meses en los que se han realizado promociones de pague 2 lleve 3. En este caso podemos identificar cuatro casos posibles, a continuación veremos cada uno estos casos.

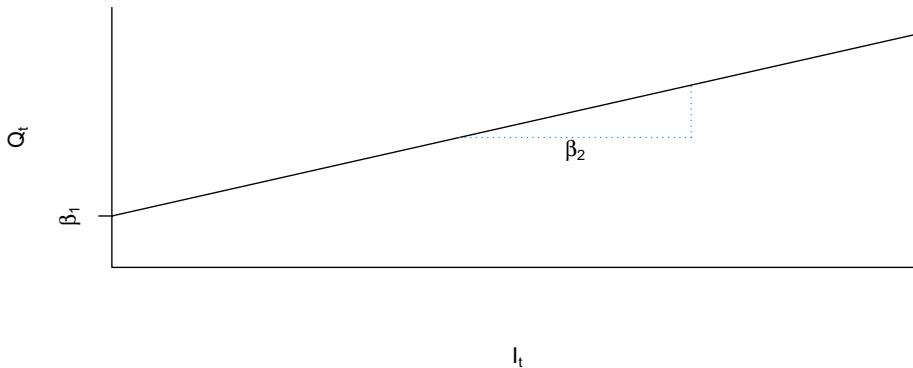
5.2.1 Caso I. La función es la misma

En este caso se asume que la cantidades vendidas es igual tanto en tiempos de promoción como en tiempos sin promoción. Esto implica la siguiente relación:

$$Q_t = \beta_1 + \beta_2 I_t + \varepsilon_t. \quad (5.1)$$

Como podemos apreciar en la Figura 5.5, tanto la pendiente como el intercepto de la función se mantienen inalterados en todo el período de estudio.

⁴Si se incluyen variables dummy para todas las opciones, una de estas variables no trae información nueva al modelo; en otras palabras esa variable será redundante. Esto implicará que las variables no son independientes y por tanto se violaría un supuesto del teorema de Gauss-Markov. Por ejemplo, supongamos que en un modelo se desea incluir el sexo del individuo como una variable explicativa. Para esto se crea una variable dummy M_i que toma el valor de uno si el individuo i es mujer y cero en caso contrario. Adicionalmente, se crea la variable H_i que toma el valor de uno si i es hombre y cero en caso contrario. Ahora noten que si para un determinado individuo i $M_i = 1$, entonces tiene que ser cierto que $H_i = 0$. Adicionalmente, si $M_i = 0$, entonces $H_i = 1$. Esto implica que H_i es una variable que sobra dado que su valor lo podemos conocer con seguridad del valor que tome M_i .

Figura 5.5. Caso I. No hay cambio

Fuente: Elaboración propia

En este caso β_2 representa cuántas unidades vendidas por cada peso adicional de ingreso y β_1 las cantidades vendidas que no dependen del ingreso. Es decir, un solo modelo para todas las situaciones.

5.2.2 Caso II. Cambio en intercepto.

En el segundo caso, queremos modelar que las unidades vendidas son mayores en períodos con promoción que en tiempos sin promoción (*ceteris paribus*⁵). Esto se puede representar con un incremento de las ventas sin importar el nivel de inversión (intercepto de la función) para los meses de promoción, al tiempo que se mantiene inalterada la pendiente. Este efecto lo recoge el siguiente modelo:

$$Q_t = \beta_1 + \beta_2 I_t + \alpha D_t + \varepsilon_t, \quad (5.2)$$

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

Como podemos apreciar, la ecuación 5.2 sugiere que en períodos de promoción ($D_t = 1$) el DGP (modelo) será:

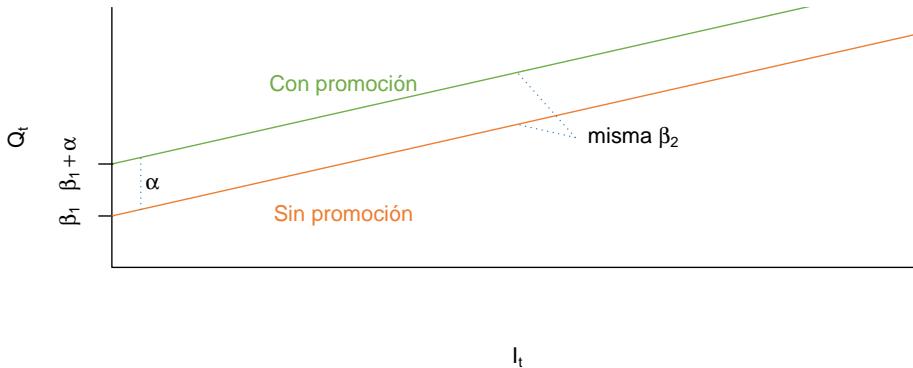
$$Q_t = (\beta_1 + \alpha) + \beta_2 I_t + \varepsilon_t$$

En tiempos sin promoción, la variable dummy toma el valor de 0 y por lo tanto el modelo que describe cantidades vendidas de cremas de dientes se reduce a:

$$Q_t = \beta_1 + \beta_2 I_t + \varepsilon_t.$$

La Figura 5.6 representa el modelo (5.2).

⁵El término *ceteris paribus* significa dejando todo lo demás constante.

Figura 5.6. Caso II. Cambio en el intercepto

Fuente: Elaboración propia

En este caso, α representa la diferencia del intercepto entre el periodo sin promoción y con promoción. En otras palabras, la diferencia entre las ventas que no dependen de la inversión entre los meses con promoción y aquellos que no tienen promoción. Además, se espera que $\alpha > 0$, como se representó en la Figura 5.6. Noten que este caso genera rectas paralelas.

5.2.3 Caso III. Cambio en pendiente

En este caso, supongamos que las ventas son nuevamente mayor en meses de promoción que en aquellos sin promoción, pero a diferencia del Caso II en el que solo cambiaba el intercepto de la función, en este caso se presenta un cambio de la pendiente (se vende más por cada unidad monetaria adicional de inversión en los meses de promoción). Este hecho lo podemos representar con el siguiente modelo:

$$Q_t = \beta_1 + \beta_2 I_t + \gamma(D_t I_t) + \varepsilon_t, \quad (5.3)$$

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

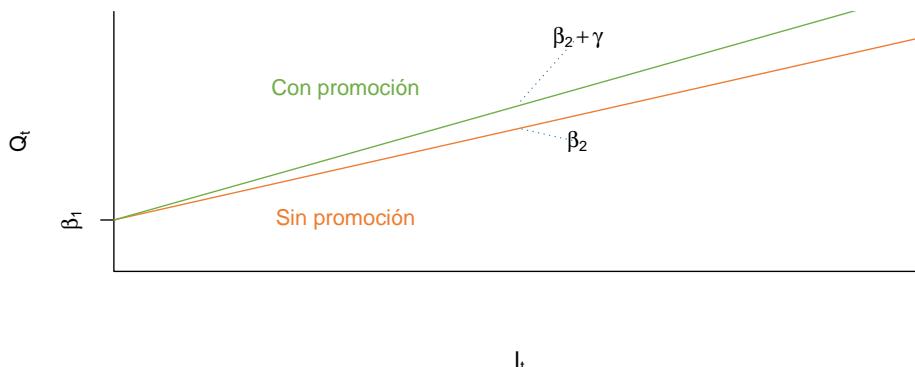
La ecuación 5.3 nos muestra que en meses con promoción (la variable dummy toma el valor de 1) se presenta un cambio en la pendiente. Es decir, tenemos que en tiempos con promoción el DGP que describe las unidades vendidas toma la siguiente forma:

$$Q_t = \beta_1 + (\beta_2 + \gamma) I_t + \varepsilon_t.$$

En tiempos sin promoción (la variable dummy toma el valor de 0) el DGP viene dado por:

$$Q_t = \beta_1 + \beta_2 I_t + \varepsilon_t.$$

Este DGP se representa en la Figura 5.7.

Figura 5.7. Caso III. Cambio en pendiente

Fuente: Elaboración propia

En este caso γ corresponde a la diferencia entre el efecto de una unidad más de inversión sobre las ventas en meses con promoción y sin promoción. En este caso se espera que $\gamma > 0$, como se representó en la Figura 5.7. Antes de pasar al siguiente caso, es importante mencionar que este tipo de modelos se denominan con interacción entre las variables. Es decir, la variable dummy interactúa con la variable I_t .

5.2.4 Caso IV. Cambio en intercepto y pendiente

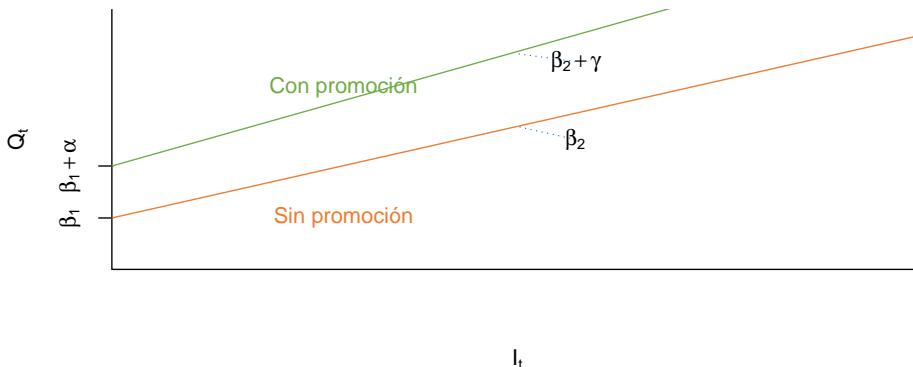
En este último caso, la venta de cremas dentales aumenta en tiempos con promoción por las dos posibles vías; es decir, por un aumento tanto en el intercepto como en la pendiente. Este efecto es capturado por un modelo con la siguiente especificación:

$$Q_t = \beta_1 + \beta_2 I_t + \alpha D_t + \gamma (D_t I_t) + \varepsilon_t \quad (5.4)$$

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

Análogamente, el modelo en períodos con promoción será de la siguiente forma:

$$Q_t = (\beta_1 + \alpha) + (\beta_2 + \gamma) I_t + \varepsilon_t \quad (5.5)$$

Figura 5.8. Caso IV. Cambio en intercepto y pendiente

Fuente: Elaboración propia

En este caso α y γ representan la diferencia del intercepto y de la pendiente en periodos con promoción y sin promoción, respectivamente. Se espera que α y γ sean positivos.

5.2.5 Para tener en cuenta

Antes de pasar a practicar en R, es importante recordar que en la práctica no podemos saber a cuál de los cuatro casos corresponde el DGP de los datos bajo estudio. Por eso, el científico de datos se ve en la necesidad de probar diferentes especificaciones. En los Capítulos 3 y 4 ya estudiamos métodos estadísticos que nos permitirán escoger entre uno de los casos, para una muestra determinada.

5.3 Práctica en R

Existen múltiples formas de crear variables dummy en R. Por un lado, 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 variable dummy⁶. Esto hace que en la mayoría de los casos no sea necesario generar una variable dummy cuando los datos son de corte transversal.

Con este ejercicio ilustraremos cómo generar variables dummy con R en una serie de tiempo. Continuando con el ejemplo de este capítulo, las cantidades vendidas de crema dental de la marca DBLANCOS por el canal tradicional depende de la inversión en material promocional que se entrega a las tiendas minoristas en el mes t . Por otro lado, se considera que la relación ha cambiado a partir del ingreso de las tiendas de descuento en los barrios, como D1 o Justo y Bueno, puesto que le han quitado participación a las tiendas de barrio. Así la pregunta de negocio es ¿qué tanto cambió el

⁶Si 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()** que toma como argumentos la variable del **data.frame** que se quiera cambiar y **ref** que permite escribir el nombre del nivel del factor (entre comillas) que se quiere emplear como la referencia. Por ejemplo, de la siguiente manera **relevel(data.frame\$variable, ref = "SI")**.

impacto de nuestra inversión en material promocional con la entrada de las tiendas de descuento en los barrios?

Como se discutió en la sección anterior, la relación entre estas dos variables implica un modelo lineal como el siguiente:

$$Q_t = \beta_0 + \beta_1 I_t + \varepsilon_t, \quad (5.6)$$

donde β_0 representa el intercepto, β_1 nos indica cómo cambian las cantidades vendidas de crema dental cuando cambia la inversión en el material promocional en una unidad y ε_t un término de error aleatorio.

Los datos para 228 meses se encuentran en el archivo *cremadental.csv*. Carga los datos y guárdalos en un objeto que llamaremos **datos.dummy** y cuya clase sea **datos.dummy.frame**.

Si graficamos las ventas mensuales en el tiempo, observamos un cambio en el comportamiento de estas al final del periodo de análisis (figura 5.9). Si graficamos la relación entre el *sell-in* y la inversión en material promocional (figura 5.10), se observa que ese cambio parece estar afectando tanto el intercepto de la relación como la pendiente. A esto se le denomina cambio estructural. 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.

Figura 5.9. Evolución temporal de las unidades de Sell-in de la crema dental

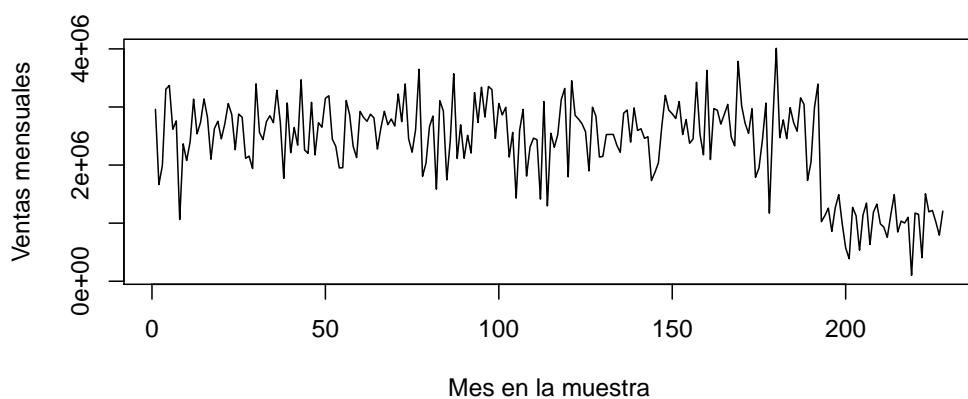
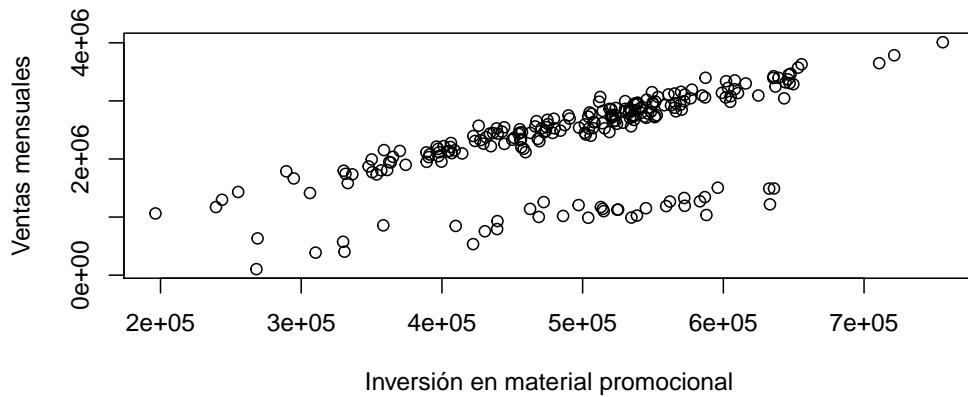


Figura 5.10. Sell-in mensual de crema dental (en unidades) e inversión en material promocional



Así parecería que existe una relación diferente para una parte de la muestra, ahora empleando el conocimiento del negocio sabemos que las tiendas de descuento aparecieron a partir de 2009. De esta manera, en este caso (a diferencia de la sección anterior) la variable dummy deberá ser definida como D_t es 1 si la observación es del año 2009 o posterior y cero en caso contrario.

Para verificar la hipótesis de que la relación ha cambiado después de 2009 se considera el siguiente modelo

$$Q_t = \beta_0 + \beta_1 \text{ingreso}_t + \alpha_2 D_t + \beta_2 (D_t \text{ingreso}_t) + \varepsilon_t. \quad (5.7)$$

Asegúrate que entiendes por qué este modelo genera un cambio tanto en intercepto como en la pendiente.

Una vez las series sean leídas en R, necesitamos crear esta variable dummy. R provee una variedad de posibilidades para la creación de variables dummy. En este caso emplearemos una prueba lógica sobre la variable *año* del data.frame *datos.dummy*.

```
# prueba lógica que crea la variable dummy
D <- datos.dummy$año >= 2009
# se convierte de clase lógica a numérica y se incluye en la
# base de datos
datos.dummy$D <- as.numeric(D)
# chequeando la creación de la variable.
head(datos.dummy, 3)

##    inversión      q  año mes D
## 1     566077 2959023 2000    1 0
## 2     294702 1664956 2000    2 0
## 3     350079 1992492 2000    3 0
```

```

tail(datos.dummy, 3)

##      inversión     q año mes D
## 226      486103 1019278 2018 10 1
## 227      439214  793211 2018 11 1
## 228      497122 1206111 2018 12 1

datos.dummy[106:115, ]

##      inversión     q año mes D
## 106      475444 2600517 2008 10 0
## 107      537867 2958311 2008 11 0
## 108      361114 1810287 2008 12 0
## 109      423541 2310373 2009  1 1
## 110      519009 2463626 2009  2 1
## 111      439715 2436638 2009  3 1
## 112      306218 1413677 2009  4 1
## 113      624870 3093221 2009  5 1
## 114      243714 1297751 2009  6 1
## 115      474318 2547605 2009  7 1

```

Ahora sí podemos estimar el modelo deseado (asegúrate que puedes estimarlo). Nota que la interacción entre la variable dummy y la inversión se debe incluir en la fórmula incluyendo $D * inversión$.

```

##
## Call:
## lm(formula = q ~ D + inversión + D * inversión, data = datos.dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1562566  -96055   74047  399299  764185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.217e+04  2.882e+05   0.250   0.802
## D          -2.055e+05  3.859e+05  -0.532   0.595
## inversión   5.125e+00  5.697e-01   8.996  <2e-16 ***
## D:inversión -5.209e-01  7.650e-01  -0.681   0.497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 550100 on 224 degrees of freedom
## Multiple R-squared:  0.4794, Adjusted R-squared:  0.4724
## F-statistic: 68.75 on 3 and 224 DF,  p-value: < 2.2e-16

```

El cambio en el intercepto (coeficiente asociado a D_t) y el cambio en la pendiente (asociado a $D_{t,inversion_t}$) no son individualmente significativos. Es decir, parece que no existe ese cambio estructural después de 2009 cuando llegan las tiendas de descuento a los barrios. No obstante ya sabemos que no es una buena idea sumar pruebas individuales para tomar una decisión. Entonces es mejor emplear una prueba de modelos anidados que discutimos en el Capítulo 4. En la sección 4.3.1 discutimos cómo corroborar si estos dos coeficientes estimados son iguales a cero simultáneamente. Esto se logra de la siguiente manera.

```
## Analysis of Variance Table
##
## Model 1: q ~ inversión
## Model 2: q ~ D + inversión + D * inversión
##   Res.Df      RSS Df  Sum of Sq    F    Pr(>F)
## 1     226 8.0141e+13
## 2     224 6.7795e+13  2 1.2346e+13 20.396 7.284e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estos resultados muestran que es posible rechazar la hipótesis nula de que los cambios en pendiente e intercepto son cero simultáneamente. Es decir, podemos concluir que existe un cambio estructural en la relación de las ventas y el ingreso de los tenderos en 2009. Puedes constatar que si se compara el modelo con variables dummy en el intercepto y pendiente con los modelos con solo cambio en intercepto o solo cambio en pendiente se llegaría a la conclusión que el mejor modelo es el que tiene el cambio tanto en intercepto como en pendiente⁷. No obstante individualmente ni el cambio en el intercepto ni el de la pendiente es individualmente significativo.

Ahora continuemos con nuestro análisis de los datos y nuestra tarea de encontrar un buen modelo para explicar las ventas. El modelo que acabamos de obtener no parece el mejor. Ahora, probemos una segunda hipótesis del departamento de mercadeo. En ese departamento se cree que para los años 2016, 2017 y 2018 la relación es diferente dado que las tiendas de descuento ya se habían consolidado y expandido por toda Colombia. Construyamos una dummy segunda ($D2_t$) que tome el valor de uno únicamente para los años comprendidos entre 2016 y 2018, periodo donde ya las tiendas de descuento se habían expandido en Colombia.

Olvidémonos por un momento de los resultados encontrados anteriormente y partamos del modelo expresado en 5.6 pero con la variable dummy expresada de la nueva forma. Determine si existe o no un cambio estructural⁸. Los resultados de los tres modelos se presentan en el Cuadro 5.1.

⁷Este corresponde al primer ejercicio de este capítulo.

⁸Ayuda: puedes emplear el signo & para hacer dos pruebas lógicas al mismo tiempo.

Cuadro 5.1: Estimación del modelo para diferentes especificaciones de la dummy

<i>Dependent variable:</i>			
	q		
	(1)	(2)	(3)
D	−205,487.400 (385,894.200)		
D2		−472,231.000*** (119,814.800)	
inversión	5.125*** (0.570)	4.889*** (0.411)	5.129*** (0.099)
D:inversión	−0.521 (0.765)		
D2:inversión		−2.240*** (0.239)	
Constant	72,169.520 (288,169.600)	−54,617.020 (207,448.000)	74,668.030 (49,871.470)
Observations	228	228	228
R ²	0.479	0.385	0.971
Adjusted R ²	0.472	0.382	0.970
Residual Std. Error	550,142.900 (df = 224)	595,489.400 (df = 226)	130,151.800 (df = 224)
F Statistic	68.749*** (df = 3; 224)	141.216*** (df = 1; 226)	2,487.744*** (df = 3; 224)

Note: *p<0.1; **p<0.05; ***p<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 99 %). Y conjuntamente también lo son con un nivel de significancia del 99 %, como se muestra a continuación.

```
## Analysis of Variance Table
##
## Model 1: q ~ inversión
## Model 2: q ~ D2 + inversión + D2 * inversión
##   Res.Df      RSS Df Sum of Sq    F    Pr(>F)
## 1     226 8.0141e+13
## 2     224 3.7944e+12  2 7.6347e+13 2253.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Así, podemos rechazar la hipótesis nula que no existió cambio en la pendiente y el intercepto para los últimos 3 años. La evidencia muestra que este último modelo explica mucho mejor la muestra que tenemos al frente y por tanto debería ser el indicado para tomar decisiones.

5.3.1 Creando variables dummy con paquetes de R

En algunas ocasiones será necesario crear las variables dummy en vez de dejar que la función **lm()** las cree automáticamente. Por ejemplo, suponga que tenemos datos para todos los días de la semana en la base de datos y solo queremos crear variables dummy para los cinco días laborales. Si empleamos esta variable con la función **lm()**, ésta creará seis variables dummy para todos los días de la semana (y descarta una). En ese caso podemos emplear paquetes de R para hacer esta tarea más rápidamente. La verdad no existe un mejor paquete para crear variables dummy, solo unos funcionan mejor que otros en diferentes circunstancias. En las siguientes subsecciones veremos dos opciones de paquetes.

El paquete dummies

El paquete *dummies* (Brown, 2012) es bastante flexible y permite crear variables dummy de diferentes maneras. Veamos un par de ejemplos. Creamos un **data.frame** con datos de ventas, la ciudad (variable tipo factor) y el año.

```
ventas <- data.frame(ciudades = c("Cali", "Medellin", "Bogotá",
  "Cali"), año = c(2019, 2020, 2020, 2020), ventas = c(10,
  30, 40, 15))
ventas$ciudades <- as.factor(ventas$ciudades)

str(ventas)

## 'data.frame': 4 obs. of  3 variables:
## $ ciudades: Factor w/ 3 levels "Bogotá","Cali",...: 2 3 1 2
## $ año      : num  2019 2020 2020 2020
## $ ventas   : num  10 30 40 15
```

Ahora podemos emplear la función **dummy()** de este paquete para crear variables dicotómicas para la ciudad. Esta función puede tomar como argumentos variables que sean caracteres, o factores, no importa, en todo caso crea las variables dummy. Otro argumento importante es el carácter que se quiere emplear como separador para el nombre de las variables dicotómicas que se crean.

```
library(dummies)
dummy(ventas$ciudades)

##      LibroRegresion.RnwBogotá LibroRegresion.RnwCali
## [1,]                 0                  1
## [2,]                 0                  0
## [3,]                 1                  0
## [4,]                 0                  1
##      LibroRegresion.RnwMedellin
## [1,]                 0
## [2,]                 1
```

```

## [3,]          0
## [4,]          0

dummy(ventas$ciudades, sep = "_")

##      LibroRegresion.Rnw_Bogotá LibroRegresion.Rnw_Cali
## [1,]                  0                  1
## [2,]                  0                  0
## [3,]                  1                  0
## [4,]                  0                  1
##      LibroRegresion.Rnw_Medellin
## [1,]                  0
## [2,]                  1
## [3,]                  0
## [4,]                  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 $p - 1$ variables⁹. En el Capítulo 8 se explicará la razón técnica para esto.

El paquete `fastDummies`

Otra opción es el paquete *fastDummies* (Kaplan, 2020) . Este paquete tiene la función **dummy_cols()** que permite al mismo tiempo que se crea la variable dummy descartar una de ellas. Esta función implica por lo menos los siguientes argumentos:

```
dummy_cols( .data, select_columns = NULL, remove_most_frequent_dummy = FALSE, ignore_na = FALSE)
```

donde:

- **.data:** es un objeto de clase matriz o *data.frame*.
- **select_columns:** si se emplea un *data.frame* este es un vector con los nombres de las variables a las que se les desea crear las variables dicotómicas.
- **remove_most_frequent_dummy:** Si `remove_most_frequent_dummy = TRUE` se elimina la dummy para la categoría en la que exista más observaciones. Si este argumento es igual a `FALSE`, entonces no se remueve ninguna variable dummy. Esta última opción es la que se ejecuta por defecto.

⁹Otra opción es emplear la p variables dummy y omitir el intercepto. Pero en la mayoría de los casos esta no es una buena idea.

- **ignore_na**: Cuando `ignore_na = FALSE` (opción por defecto) se crea una variable dummy cuando se encuentra un “NA” en la variable a ser transformada. Si `ignore_na = TRUE` entonces se omite las observaciones con “NA”, no se crea una nueva variable y a esa observación se le asigna “NA” en todas las dummies que se creen. (Esta es la opción recomendada en la mayoría de los casos).

Empleemos esta función con un **data.frame** similar al anterior, pero adicionemos datos perdidos en la variable que transformaremos.

```
ventas <- data.frame(ciudades = c("Cali", "Medellin", "Bogotá",
  "Cali", NA), año = c(2019, 2020, 2020, 2020, 2019), ventas = c(10,
  30, 40, 15, NA))
ventas$ciudades <- as.factor(ventas$ciudades)
str(ventas)

## 'data.frame': 5 obs. of  3 variables:
## $ ciudades: Factor w/ 3 levels "Bogotá","Cali",...: 2 3 1 2 NA
## $ año      : num  2019 2020 2020 2020 2019
## $ ventas   : num  10 30 40 15 NA
```

Ahora, veamos diferentes opciones al aplicar esta función.

```
library(fastDummies)
dummy_cols(ventas$ciudades)

##      .data .data_Bogotá .data_Cali .data_Medellin .data_NA
## 1      Cali          0         1          0          0
## 2 Medellin          0         0          1          0
## 3  Bogotá          1         0          0          0
## 4      Cali          0         1          0          0
## 5     <NA>         NA        NA         NA         1
```

Quitemos la variable dummy para la opción que más se repite y descartemos la opción de crear una dummy para las observaciones no disponibles (“NA”).

```
dummy_cols(ventas$ciudades, remove_most_frequent_dummy = TRUE,
  ignore_na = FALSE)

##      .data .data_Bogotá .data_Medellin
## 1      Cali          0         0
## 2 Medellin          0         1
## 3  Bogotá          1         0
## 4      Cali          0         0
## 5     <NA>         NA        NA
```

Hagamos lo mismo, pero empleando el **data.frame** y no la columna

```
dummy_cols(ventas, select_columns = "ciudades", remove_most_frequent_dummy = TRUE,  
ignore_na = FALSE)
```

```
##   ciudades año ventas ciudades_Bogotá ciudades_Medellin  
## 1     Cali 2019     10          0          0  
## 2 Medellin 2020     30          0          1  
## 3  Bogotá 2020     40          1          0  
## 4     Cali 2020     15          0          0  
## 5      <NA> 2019      NA         NA         NA
```

Ejercicios

5.1 Continuemos con el ejemplo realizado en este capítulo y estudiemos en mas detalle los posibles modelos que se podían estimar con la dummy definida como: $D_t = 1$ si t es del año 2009 o posterior y cero en caso contrario.

Estima un modelo sin variables dummy, con solamente cambio en la pendiente, otro con solo cambio en el intercepto y con cambio en ambos. Ahora, decide cuál modelo es mejor: sin cambios, con dummy solo en intercepto, solo en pendiente, o con dummy en ambos.



6 . Selección automática de modelos

Diseñado por Freepik

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 exemplificar 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 variable que se quiere explicar¹ o predecir² (variable dependiente) y un conjunto grande de posibles variables explicativas. Pero no sabemos cuáles variables si son importantes y cuáles no al momento de explicar nuestra variable de interés.

En general, si hay $k - 1$ variables independientes potenciales (además del intercepto) que pueden explicar o predecir a la variable dependiente, entonces hay $2^{(k-1)}$ subconjuntos de variables explicativas posibles para explicar³ los cambios en la variable dependiente. Es decir, existen $2^{(k-1)}$ modelos que pueden ser candidatos a ser el mejor modelo para explicar 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 correcto⁴. Por ejemplo si tenemos 10 posibles variables explicativas ($(k - 1) = 10$) entonces los posibles modelos serán $2^{(k-1)} = 2^{(10)} = 1024$. Si se tienen 20 variables candidatas, el número de posibles subconjuntos es de $1,048576 \times 10^6$ (más 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 candidatas a ser explicativas no es viable calcular todos los posibles modelos y compararlos. No obstante, si las variables son relativamente pocas podría ser viable estimar todos los posibles modelo y compararlos. Recordemos que en el capítulo anterior estudiamos cómo comparar modelos para seleccionar al mejor empleando métricas como el R^2 ajustado, los criterios de información *AIC* o *BIC* (Ver Sección 4.2) y pruebas estadísticas de modelos anidados (Ver Sección 4.3.1) y no anidados (Ver Sección 4.3.2).

En este capítulo discutiremos diferentes algoritmos para encontrar el mejor modelo (o por lo menos el modelo para ser un buen candidato a 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 seleccionar el mejor modelo.

¹Si la respuesta a la pregunta de negocio implica un tipo de analítica diagnóstica.

²Si la respuesta a la pregunta de negocio implica un tipo de analítica predictiva.

³De aquí en adelante se omitirá la alusión a la posibilidad de emplear las técnicas de este capítulo para hacer analítica predictiva. Esto se hace solamente para ahorrar tiempo y espacio. Tienes que tener en cuenta que las técnicas que estudiaremos aquí pueden ser aplicadas para encontrar el mejor modelo de regresión independientemente si este se emplea o no para hacer analítica diagnóstica o predictiva y aún podría ser prescriptiva.

⁴Nota que cuando se cuenta con una teoría que se desea probar, este problema no existe. Es decir, en la aproximación econométrica tradicional discutida en la sección 1.3 la teoría nos dirá cuál debería ser el modelo a probar y este problema de seleccionar el modelo con los datos desaparece. Pero como se discutió en esa misma sección la aproximación del científico de datos es muy diferente y por tanto este problema de la selección del mejor modelo se encuentra en el centro del quehacer diario del científico de datos que emplea el modelo de regresión.

Emplearemos unos datos simulados para una variable dependiente (y_i) y 25 posibles variables explicativas x_{ji} 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_{25} únicamente y los coeficientes son iguales a 1 para todos los casos (nota 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 *datos* y verifiquemos la clase del objeto leído y de cada una de las variables en el objeto.

```
# lectura de datos
datos <- read.table("../Data/DATOSautoSel.txt", header = TRUE)
# clase de las columnas
str(datos)

## '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 ...

# las dos primeras filas de datos
head(datos, 2)

##      x1     x2     x3     x4     x5     x6     x7     x8     x9     x10
```

```

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

# clase del objeto
class(datos)

## [1] "data.frame"

```

En este caso, tenemos 25 posibles variables, entonces son $3,3554432 \times 10^7$ (aproximadamente 33.6 millones) de posibles modelos.

Este capítulo está compuesto de dos partes. La primera, emplea un número reducido de variables ($(k - 1) = 10$) para mostrar una aproximación que emplea “fuerza bruta”. Es decir, se evalúan todos los posibles modelos y se comparan. En la segunda parte veremos diferentes algoritmos para encontrar el un modelo para ser candidato a mejor modelo cuando se tiene muchas posibles variables explicativas, y para esta sección sí 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 cómputo 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 . Así, en este caso se compararán 1024 modelos. No empleamos todas las variables explicativas para ahorrar tiempo en la estimación de todos los modelos y para hacer viable este ejercicio. Creemos un **data.frame** con solo las variables que son de nuestro interés.

```

datos2 <- datos[, c(1:10, 26)]
head(datos2, 2)

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

```

```
## 1 42.799
## 2 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. Recuerda 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 objeto **datos2** y guardemos los resultados de la estimación en un objeto llamado **modelo**.

```
# estimación del modelo con todas las variables
modelo <- lm(y ~ ., data = datos2)
summary(modelo)

##
## Call:
## lm(formula = y ~ ., data = datos2)
##
## 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 **x1** a **x5** son estadísticamente significativas (nivel de confianza del 95%)⁵. Las otras variables no son significativas.

⁵La variable **x9** es significativa con un nivel de confianza del 90 %. Usaremos un nivel de confianza del 95 %.

Ahora podemos investigar todos los 1024 posibles modelos. Esto lo podemos hacer con la función **all.possible.modelos()** del paquete **olsrr**(Hebbali, 2020). Esta función requiere típicamente de un solo argumento que corresponde a un objeto de un objeto de clase **lm** que contenga el modelo más grande posible. Guardemos los resultados de ejecutar esta función en el objeto **mod**⁶.

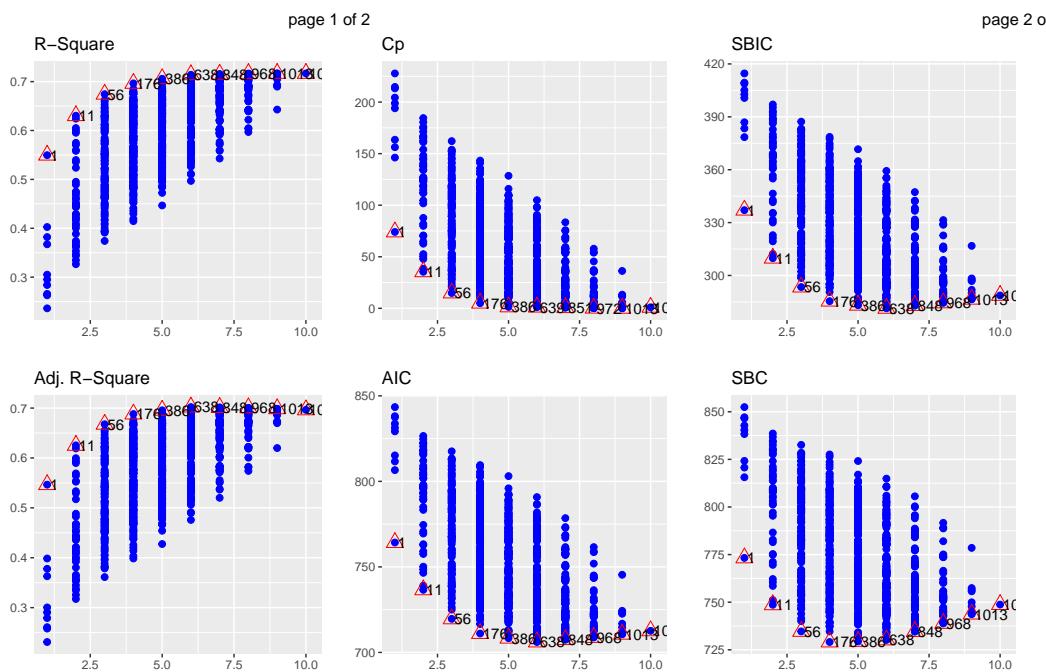
```
# install.packages('olsrr') cargar librería
library(olsrr)
# se estiman todos los posibles modelos
mod <- ols_step_all_possible(modelo)
# atributos del nuevo objeto
str(mod)

## Classes 'ols_step_all_possible' and 'data.frame': 1023 obs. of 14 variables:
## $ 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 ...
## $ msep       : num 1396 1851 1914 1960 2153 ...
## $ 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 **mod** se encuentran varios estadísticos que permiten resumir las características estadísticas de todos los modelos estimados. Con un gráfico podemos resumir esta información.

```
library(ggplot2)
plot(mod)
```

⁶Esto puede tomar un rato, dado que se están estimando 1024 modelos.



En el eje horizontal podemos ver el números de variables empleadas en cada modelo, mientras que en el eje vertical encontramos el valor de la métrica. El triángulo muestra el modelo que maximiza o minimiza el valor de la métrica 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, SBC (o también conocido como BIC). Recuerda 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 ($x_1 x_2 x_3 x_4 x_5 x_9$). De hecho, ese modelo corresponde al modelo número 638 de todos los estimados. Esta información se puede obtener de la siguiente manera.

```
mod$minindex[which.max(mod$adjr)]
## [1] 638

mod$n[which.max(mod$adjr)]
## [1] 6

mod$predictors[which.max(mod$adjr)]
## [1] "x1 x2 x3 x4 x5 x9"
```

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

```
mod$minindex[which.min(mod$aic)]  
  
## [1] 638  
  
mod$n[which.min(mod$aic)]  
  
## [1] 6  
  
mod$predictors[which.min(mod$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.

```
mod$minindex[which.min(mod$sbc)]  
  
## [1] 176  
  
mod$n[which.min(mod$sbc)]  
  
## [1] 4  
  
mod$predictors[which.min(mod$sbc)]  
  
## [1] "x2 x3 x4 x5"
```

Finalmente, tenemos todos modelos candidatos a ser el mejor modelo. Estimemos los dos modelos y guardemos los resultados en los objetos (modelo1) y (modelo2). Los resultados se reportan en el Cuadro 6.1.

```
# estimación del primer modelo candidato  
modelo1 <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x9, data = datos2)  
# estimación del segundo modelo candidato  
modelo2 <- lm(y ~ x2 + x3 + x4 + x5, data = datos2)
```

Cuadro 6.1: Modelos seleccionados por las métricas tras emplear fuerza bruta

<i>Dependent variable:</i>		
	y Modelo 1 (1)	Modelo 2 (2)
x1	0.675** (0.263)	
x2	1.692*** (0.271)	1.701*** (0.268)
x3	0.956*** (0.265)	0.857*** (0.263)
x4	0.978*** (0.249)	1.034*** (0.245)
x5	0.952*** (0.248)	0.997*** (0.247)
x9	−0.537** (0.266)	
Constant	11.272*** (1.353)	11.924*** (1.312)
Observations	150	150
R ²	0.714	0.696
Adjusted R ²	0.702	0.688
Residual Std. Error	2.473 (df = 143)	2.530 (df = 145)
F Statistic	59.496*** (df = 6; 143)	83.157*** (df = 4; 145)

Note: *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 (Ver 4.3.1 para una discusión del tema). Hagamos dicha comparación.

```
library(AER)
# comparación de modelos anidados
```

```
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
```

Según los resultados, 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 (con un nivel de confianza del 99 %), 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 x1 a x5. 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 x1 a x5 pero también incluiría x9; tampoco es el modelo exacto, pero es lo suficientemente cercano.

6.3 Empleando estrategias inteligentes de detección de un mejor modelo

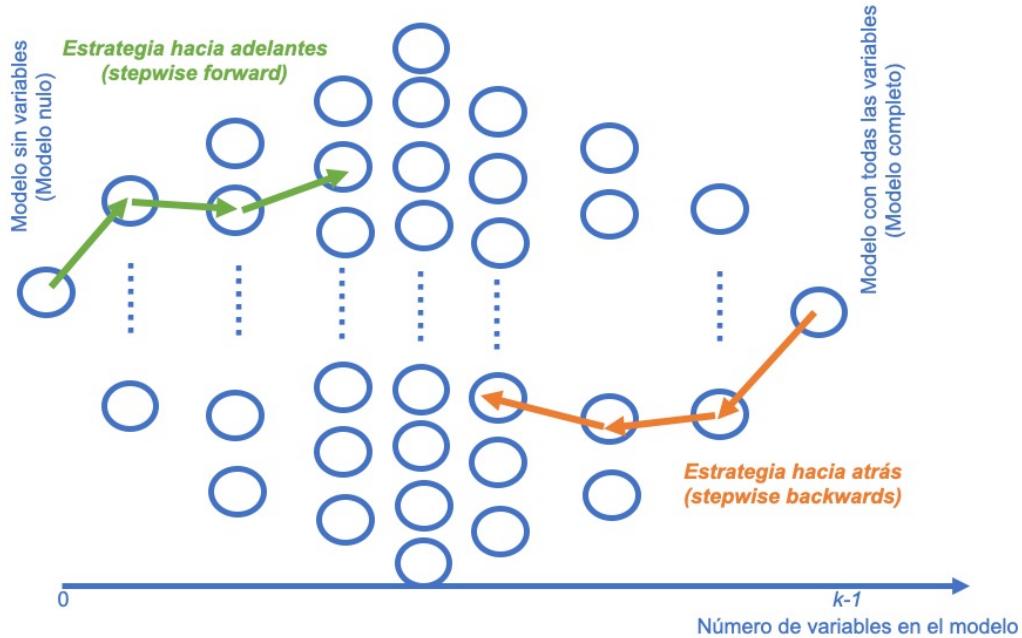
En algunas ocasiones es imposible encontrar el mejor modelo estimando todas las combinaciones (como el caso en el que se tienen 25 variables pues existen $3,3554432 \times 10^7$ posibles modelos). A continuación discutimos varios algoritmos que facilitan la tarea.

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 menor valor p. De forma gradual se incluye la siguiente variable que tenga el valor p más pequeño, se sigue de esta manera hasta que ya no quede ninguna variable para ingresar que sea significativa⁷.

⁷Noten 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. Adicionalmente, en presencia de heteroscedasticidad (Ver Capítulo 9) o autocorrelación (Ver Capítulo 11) estos valores p no serían válidos por no tener en cuenta el problema y por tanto las conclusiones podrían ser erradas.

Figura 6.1. Representación de las estrategias *stepwise forward* y *stepwise backward*

Fuente: Elaboración propia

En el segundo caso (*stepwise backward regression*), se parte del modelo con todas las variables y se empieza a eliminar variables que tenga el valor p más alto. El proceso se repite hasta que no se puedan eliminar variables⁸. 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.

6.3.2 Stepwise forward regression

La función **regsubsets()** del paquete **leaps** (Lumley, 2020) permite encontrar los mejores subconjuntos de variables explicativas utilizando el R^2 ajustado partiendo de un modelo con todos los regresores (lo llamaremos el modelo máximo). Esta función no está diseñada para funcionar con el valor p. La función **regsubsets()** calcula los mejores modelos para todos los posibles número de variables explicativas sin calcularlos de manera exhaustiva⁹.

⁸Noten que esta aproximación también tiene el problema mencionado para la aproximación forward.

⁹Dado 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 (Lumley, 2020). 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 más 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.

Los argumentos más importantes de esta función son

```
regsubsets(x, y, method=method=c("backward", "forward", "seqrep"), nvmax, force.in)
```

donde:

- **x**: la matriz que contiene todas las posibles variables explicativas
- **y**: el vector de la variable dependiente (**y**)
- **method**: el método que se desea emplear: (**method =“backward”**) para el método backward, (**method =“forward”**) para el método forward y (**method =“seqrep”**) para el método combinado.
- **nvmax**: el número máximo de variables a incluir en un modelo a ser examinado. Por ejemplo, si **nvmax = 15** entonces solo se evaluarán hasta modelos con 15 variables explicativas.
- **force.in**: El número de la columna de la variable explicativa que se desee incluir siempre en el modelo. Por defecto es igual a NULL pero en algunas ocasiones es útil especificar una o unas variables que independientemente del resultado del algoritmo deberían estar siempre presentes como variables explicativas en todos los modelos considerados.

Empleemos esta función para nuestro caso y evaluando modelos con todas las variables.

```
# cargando la libreria
library(leaps)

fwd.model <- regsubsets(x = datos[, 1:25], y = datos[, 26], nvmax = 250,
method = "forward")
```

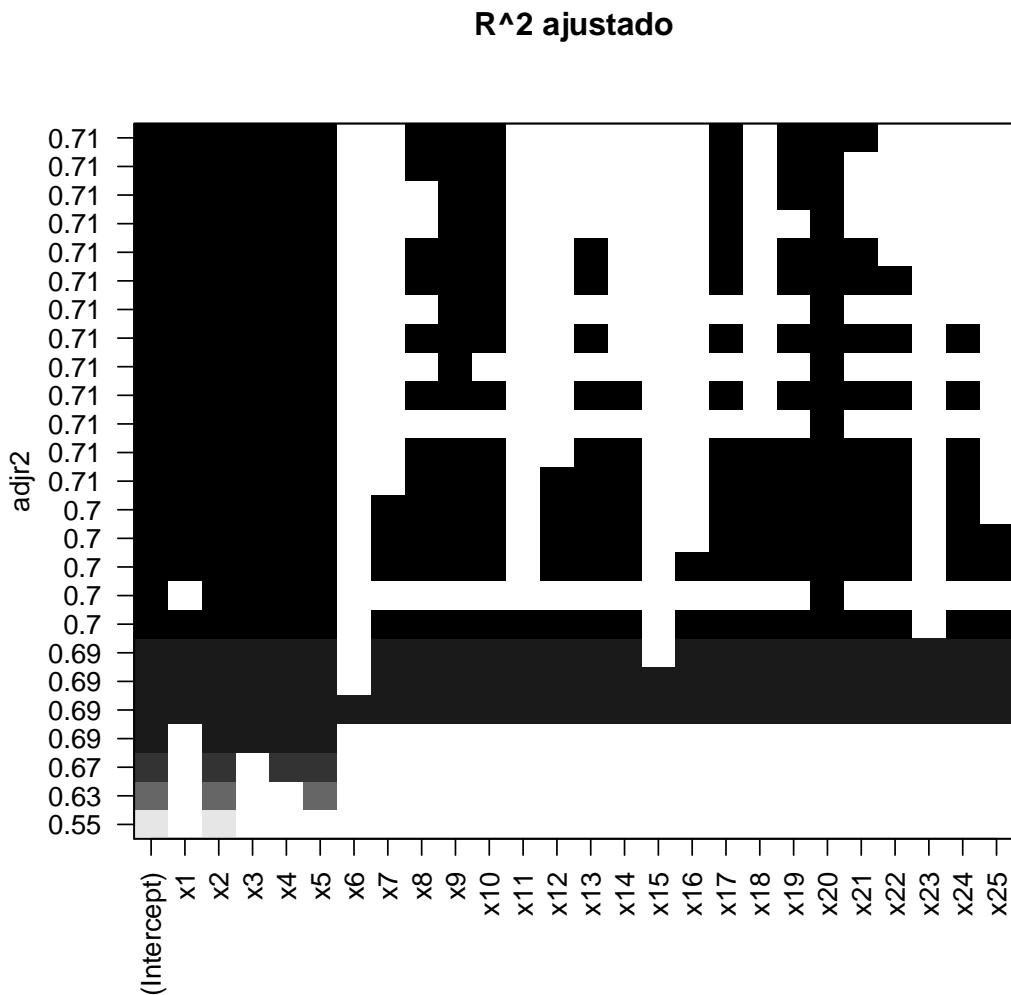
En el objeto **fwd.model** se encuentran diferentes resultados. Veamos esto en detalle y generemos un gráfico que trae por defecto esta librería.

```
# atributos del objeto
attributes(fwd.model)

## $names
## [1] "np"          "nrbar"       "d"           "rbar"        "thetab"
## [6] "first"       "last"        "vorder"      "tol"         "rss"
## [11] "bound"       "nvmax"       "ress"        "ir"          "nbest"
## [16] "lopt"        "il"          "ier"         "xnames"      "method"
## [21] "force.in"    "force.out"   "sserr"       "intercept"   "lindep"
## [26] "nullrss"     "nn"

##
## $class
## [1] "regsubsets"

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



Esta visualización presenta el R^2 ajustado en el eje vertical y todas las potenciales variables evaluadas. El gráfico solo presenta los mejores modelos, en términos del R^2 ajustado, entre todos los mejores modelos evaluados. Una fila corresponde a un modelo y 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 (una fila), mejor será éste de acuerdo a esta métrica. El mejor modelo es el último que se presenta (fila superior)¹⁰. En este caso el modelo tiene intercepto y las variables x1 a x5, x8 a x10, x17 y de x19 a x21. Estimemos ese modelo y

¹⁰Si se desea seleccionar el modelo empleando criterios de información, entonces la última línea de código debería ser modificada a `plot(fwd.model, scale = "bic")` para el caso de SBC y para el caso del AIC se debe emplear `plot(fwd.model, scale = "Cp")`. Cp corresponde al criterio de información Cp de Mallows. La literatura ha demostrado que los resultados obtenido con el Cp de Mallows (en orden) son los mismos que los del AIC para los modelos lineales (Boisbunon y col., 2013). Es decir, el criterio de información Cp y el AIC son equivalentes en el orden que generan.

guardémoslo en un objeto que llamaremos `modelo3`.

```
# estimación del modelo 3
modelo3 <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x8 + x9 + x10 + x17 +
  x19 + x20 + x21, data = datos)
summary(modelo3)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x8 + x9 + x10 + x17 +
##     x19 + x20 + x21, data = datos)
##
## 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 más adelante en la sección 6.4.1 de este capítulo.

También podemos realizar una versión de regresión por pasos hacia adelante utilizando la función función `ols_step_forward_p()` del paquete `olsrr`(Hebbali, 2020). Esta función nos permite usar el criterio del valor *p* de las pruebas de significancia individuales y criterios de información. El único argumento de esta función es un objeto de la clase `lm`. Ese modelo debe incluir todas las variables explicativas que se desean explorar; es decir el modelo máximo. En este caso esta función puede ser empleada de la siguiente manera:

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

Los resultados los podemos explorar de muchas maneras, pero la más 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: x_1 a x_5 , x_9 , x_{10} , x_{17} y x_{20} (asegúrese que entiende el porqué se escoge dicho modelo según el gráfico). Estimemos este modelo y guardemóslo en el objeto `modelo4`.

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

Los resultados se reportan en el Cuadro 6.2. Este modelo también tiene variables no significativas individualmente.

Cuadro 6.2: Modelo seleccionado por el valor p con el algoritmo stepwise forward

<i>Dependent variable:</i>	
	y Modelo 4
x1	0.663** (0.262)
x2	1.702*** (0.266)
x3	0.958*** (0.264)
x4	0.994*** (0.260)
x5	0.930*** (0.266)
x9	-0.501* (0.275)
x10	0.392 (0.281)
x17	0.333 (0.252)
x20	-0.735** (0.282)
Constant	11.174*** (1.349)
<hr/>	
Observations	150
R ²	0.731
Adjusted R ²	0.714
Residual Std. Error	2.421 (df = 140)
F Statistic	42.372*** (df = 9; 140)

Note: *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 *R²* 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.

```
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} . Nota 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 si se emplea la función `plot()` sobre el objeto `fwd.model.3`.

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).

```
# se extraen las variables del mejor modelo según el
# algoritmo
vars.modelo5 <- fwd.model.3$predictors
# se construye la fórmula
formula.modelo5 <- as.formula(paste("y ~ ", paste(vars.modelo5,
  collapse = " + "), sep = """))
# constata que la fórmula es correcta
formula.modelo5

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

# se estima el modelo con la fórmula construida
modelo5 <- lm(formula.modelo5, data = datos)
```

En el Cuadro 6.3 se reporta el resultado de este modelo.

Cuadro 6.3: Modelo seleccionado por AIC con el algoritmo stepwise forward

<i>Dependent variable:</i>	
	y Modelo 5
x1	0.686*** (0.262)
x2	1.722*** (0.267)
x3	0.950*** (0.264)
x4	1.051*** (0.257)
x5	1.051*** (0.251)
x9	-0.425 (0.269)
x10	0.407 (0.282)
x20	-0.738** (0.283)
Constant	11.348*** (1.346)
<hr/>	
Observations	150
R ²	0.728
Adjusted R ²	0.713
Residual Std. Error	2.428 (df = 141)
F Statistic	47.203*** (df = 8; 141)

Note: *p<0.1; **p<0.05; ***p<0.01

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 modelo¹¹.

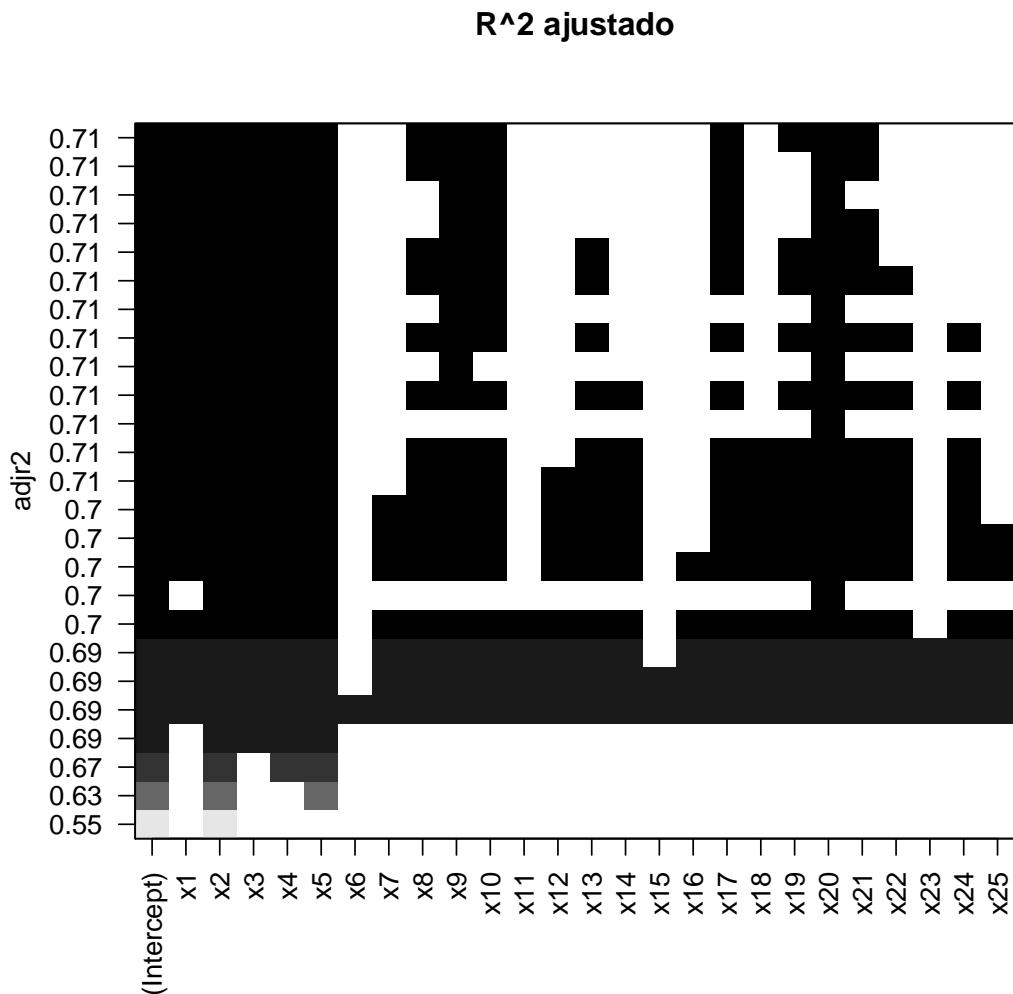
¹¹Más adelante en la sección 6.4.1 de este capítulo se discute como remover estas variables empleando una función que automatiza el proceso.

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 si empleamos el criterio del R^2 ajustado.

```
back.model <- regsubsets(x = datos[, 1:25], y = datos[, 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, 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.

```
library(olsrr)
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. En el compartimiento *\$model* del objeto que acabamos de crear quedan almacenado el mejor modelo guardémoslo en el objeto *modelo6*.

```
modelo6 <- back.model.2$model
summary(modelo6)
```

Nota que este modelo también tiene variables no significativas individualmente. Ustedes deberían emplear las técnicas que ya conocen para encontrar un mejor modelo sin variables no significativas¹². Los resultados se reportan en el cuadro 6.4.

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)
```

El modelo seleccionado incluye las siguientes variables: x1 a x5, x9, x10 y X20. Nuevamente podemos recuperar el mejor modelo con el compartimiento *\$model* del objeto que acabamos de crear. Guardémoslo en le objeto *modelo7*.

```
modelo7 <- back.model.3$model
```

Este modelo también tiene variables no significativas individualmente, pero muchas más que los modelos anteriores. Esto no es una característica de este algoritmo, solo es coincidencia. Los resultados se reportan en el Cuadro 6.4.

¹²Más adelante en la sección 6.4.1 de este capítulo se discute como remover estas variables empleando una función que automatiza el proceso.

Cuadro 6.4: Modelo seleccionado por el valor p y el AIC con el algoritmo stepwise backward

<i>Dependent variable:</i>		
	response	
	Modelo 6	Modelo 7
	(1)	(2)
x1	0.579** (0.271)	0.686*** (0.262)
x2	1.689*** (0.280)	1.722*** (0.267)
x3	0.889*** (0.275)	0.950*** (0.264)
x4	0.975*** (0.261)	1.051*** (0.257)
x5	0.882*** (0.272)	1.051*** (0.251)
x8	−0.313 (0.271)	
x9	−0.510* (0.279)	−0.425 (0.269)
x10	0.419 (0.283)	0.407 (0.282)
x17	0.307 (0.259)	
x19	0.309 (0.289)	
x20	−0.785*** (0.285)	−0.738** (0.283)
x21	0.313 (0.293)	
Constant	11.069*** (1.369)	11.348*** (1.346)
Observations	150	150
R ²	0.738	0.728
Adjusted R ²	0.715	0.713
Residual Std. Error	2.419 (df = 137)	2.428 (df = 141)
F Statistic	32.117*** (df = 12; 137)	47.203*** (df = 8; 141)

Note:

*p<0.1; **p<0.05; ***p<0.01

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 ingresándolas y eliminándolas 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 (método combinado). El modelo de partida debe incluir todas las variables explicativas candidatas. Empleando el paquete paquete *leaps* y la función que ya conocemos `textbf{regsubsets()}`. Para emplear este algoritmo solo se necesita cambiar el argumento **method**. En este caso se requiere `method = "seqrep"`.

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

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 este modelo se encuentran en Cuadro 6.5. Nuevamente, el modelo incluye variables no significativas.

Cuadro 6.5: Modelo seleccionado por el R^2 ajustado con el algoritmo combinado

<i>Dependent variable:</i>	
	y Modelo 8
x1	0.652** (0.262)
x2	1.694*** (0.280)
x3	0.903*** (0.274)
x4	0.976*** (0.261)
x5	0.930*** (0.268)
x8	-0.290 (0.270)
x9	-0.475* (0.277)
x10	0.426 (0.283)
x17	0.363 (0.253)
x20	-0.763*** (0.284)
x21	0.329 (0.292)
Constant	11.108*** (1.369)
Observations	150
R ²	0.736
Adjusted R ²	0.714
Residual Std. Error	2.420 (df = 138)
F Statistic	34.898*** (df = 11; 138)

Note: *p<0.1; **p<0.05; ***p<0.01

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*. Esta función tiene como único argumento un objeto de la clase **lm** que corresponde al modelo máximo. Para este contexto tendremos el siguiente código.

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

En este caso el modelo seleccionado incluye las variables x_1 a x_5 y x_{20} . El lector puede constatar que el correspondiente modelo es el reportado en el Cuadro 6.6. Estimemos este modelo y guardémolo en el objeto `modelo9`.

Finalmente, empleando el criterio de AIC tendremos que el mejor modelo incluye las variables x_1 a x_5 , x_9 , x_{10} y x_{20} . Este modelo se reporta en el Cuadro 6.6 y se guardó en el objeto `modelo10`. Este modelo tiene dos variables no significativas.

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

Cuadro 6.6: Modelo seleccionado por el valor p y el AIC con el algoritmo combinado

<i>Dependent variable:</i>		
	y Modelo 9 (1)	Modelo 10 (2)
x1	0.663** (0.257)	0.686*** (0.262)
x2	1.664*** (0.265)	1.722*** (0.267)
x3	0.932*** (0.258)	0.950*** (0.264)
x4	1.087*** (0.253)	1.051*** (0.257)
x5	1.056*** (0.251)	1.051*** (0.251)
x9		−0.425 (0.269)
x10		0.407 (0.282)
x20	−0.715*** (0.266)	−0.738** (0.283)
Constant	11.464*** (1.343)	11.348*** (1.346)
Observations	150	150
R ²	0.720	0.728
Adjusted R ²	0.708	0.713
Residual Std. Error	2.447 (df = 143)	2.428 (df = 141)
F Statistic	61.274*** (df = 6; 143)	47.203*** (df = 8; 141)

Note:

*p<0.1; **p<0.05; ***p<0.01

6.4 Pongamos todo junto

En la práctica queremos emplear un único modelo, para eso debemos comparar los modelos que hemos encontrado, anidados o no. Pero antes es mejor comparar modelos que tengan solo variables explicativas significativas. Es decir, modelos que no tengan variables que no son estadísticamente importantes para explicar la variable dependiente.

Recordemos que en este ejercicio de selección automática del mejor modelo hemos construido ya varios modelos candidatos a mejor modelo como se resume en el cuadro 6.7.

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 de manera automática y en la segunda subsección 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 modelo candidato a ser el “mejor” modelo que tenga 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 función `remueve.no.sinifica()` que crearemos a continuación. Puedes seguir cada línea de la función para entender los *trucos* que se emplean.

```
remueve.no.sinifica <- function(modelo, p) {
  # extrae el datosframe
  datos <- 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 = datos)  
  
    # 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 guardamos en el objeto `modelo3`). La función que acabamos de construir (`remueve.no.sinifica()`) tiene dos argumentos, el primero es un objeto de clase `lm` y el segundo el nivel de significancia que indica el nivel por debajo del cual se considera que una variable es significativa. Corramos esta función para el objeto `modelo3` con un nivel de significancia del 5% 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 reportando en el Cuadro 6.8.

Cuadro 6.8: Comparación de modelo 3 antes y después de la función remueve.no.sinifica().

<i>Dependent variable:</i>		
	y Modelo 3 (1)	y Modelo 3a (2)
x1	0.579** (0.271)	
x2	1.689*** (0.280)	1.780*** (0.266)
x3	0.889*** (0.275)	0.946*** (0.263)
x4	0.975*** (0.261)	1.212*** (0.254)
x5	0.882*** (0.272)	1.156*** (0.253)
x8	−0.313 (0.271)	
x9	−0.510* (0.279)	
x10	0.419 (0.283)	
x17	0.307 (0.259)	
x19	0.309 (0.289)	
x20	−0.785*** (0.285)	−0.608** (0.268)
x21	0.313 (0.293)	
Constant	11.069*** (1.369)	12.444*** (1.314)
Observations	150	150
R ²	0.738	0.707
Adjusted R ²	0.715	0.697
Residual Std. Error	2.419 (df = 137)	2.494 (df = 144)
F Statistic	32.117*** (df = 12; 137)	69.459*** (df = 5; 144)

Note:

*p<0.1; **p<0.05; ***p<0.01

Ahora todas las variables son significativas. Realicemos el mismo procedimiento para todos los modelos. Asegúrate que puedes obtener los modelos que se reportan en los Cuadros 6.9 y 6.10.

Cuadro 6.9: Modelos 3 al 6 tras emplear la función remueve.no.sinifica().

	<i>Dependent variable:</i>			
	y			
	Modelo 3.a (1)	Modelo 4.a (2)	Modelo 5.a (3)	Modelo 6.a (4)
x1		0.663** (0.257)	0.663** (0.257)	
x2	1.780*** (0.266)	1.664*** (0.265)	1.664*** (0.265)	1.780*** (0.266)
x3	0.946*** (0.263)	0.932*** (0.258)	0.932*** (0.258)	0.946*** (0.263)
x4	1.212*** (0.254)	1.087*** (0.253)	1.087*** (0.253)	1.212*** (0.254)
x5	1.156*** (0.253)	1.056*** (0.251)	1.056*** (0.251)	1.156*** (0.253)
x20	-0.608** (0.268)	-0.715*** (0.266)	-0.715*** (0.266)	-0.608** (0.268)
Constant	12.444*** (1.314)	11.464*** (1.343)	11.464*** (1.343)	12.444*** (1.314)
Observations	150	150	150	150
R ²	0.707	0.720	0.720	0.707
Adjusted R ²	0.697	0.708	0.708	0.697
Residual Std. Error	2.494 (df = 144)	2.447 (df = 143)	2.447 (df = 143)	2.494 (df = 144)
F Statistic	69.459*** (df = 5; 144)	61.274*** (df = 6; 143)	61.274*** (df = 6; 143)	69.459*** (df = 5; 144)

Note:

*p<0.1; **p<0.05; ***p<0.01

Cuadro 6.10: Modelos 7 al 10 tras emplear la función remueve.no.sinifica().

<i>Dependent variable:</i>					
	Modelo 7.a	Modelo 8.a	y	Modelo 9.a	Modelo 10.a
	(1)	(2)	(3)	(4)	
x1		0.663** (0.257)	0.663** (0.257)	0.663** (0.257)	0.663** (0.257)
x2	1.701*** (0.268)	1.664*** (0.265)	1.664*** (0.265)	1.664*** (0.265)	1.664*** (0.265)
x3	0.857*** (0.263)	0.932*** (0.258)	0.932*** (0.258)	0.932*** (0.258)	0.932*** (0.258)
x4	1.034*** (0.245)	1.087*** (0.253)	1.087*** (0.253)	1.087*** (0.253)	1.087*** (0.253)
x5	0.997*** (0.247)	1.056*** (0.251)	1.056*** (0.251)	1.056*** (0.251)	1.056*** (0.251)
x20		−0.715*** (0.266)	−0.715*** (0.266)	−0.715*** (0.266)	−0.715*** (0.266)
Constant	11.924*** (1.312)	11.464*** (1.343)	11.464*** (1.343)	11.464*** (1.343)	11.464*** (1.343)
Observations	150	150	150	150	150
R ²	0.696	0.720	0.720	0.720	0.720
Adjusted R ²	0.688	0.708	0.708	0.708	0.708
Residual Std. Error	2.530 (df = 145)	2.447 (df = 143)			
F Statistic	83.157*** (df = 4; 145)	61.274*** (df = 6; 143)			

Note:

*p<0.1; **p<0.05; ***p<0.01

Los resultados muestran que los modelos 3 y 6 arriban a la misma especificación: x2, x3, x4, x5 y x20 (en el cuadro 6.7 se puede ver a qué algoritmo y criterio corresponde cada uno de esos modelos). Los modelos 4, 5, 8, 9 y 10 implican emplear las mismas variables explicativas: x1, x2, x3, x4, x5 y x20 (en el cuadro 6.7 se puede ver a qué algoritmo y criterio corresponde cada uno de esos modelos). El modelo 7 emplea solamente las variables x2, x3, X4, y X5. Estos resultados nos llevan a comparar tres modelos que están anidados.

6.4.2 Comparación de modelos

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

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

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

$$y_i = \beta_1 + \beta_3 x_2 i + \beta_4 x_3 i + \beta_5 x_4 i + \beta_6 x_5 i + \varepsilon_i \quad (6.3)$$

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 está anidado en el modelo B y por tanto también en el C.

```
modeloA <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x20, datos)
modeloB <- lm(y ~ x2 + x3 + x4 + x5 + x20, datos)
modeloC <- lm(y ~ x2 + x3 + x4 + x5, datos)
```

El siguiente paso del científico de datos es escoger entre estos modelos. Para esto podemos emplear pruebas F para modelos anidados empleando la función **anova()**. 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 (con un 95 % de confianza). Es decir, el modelo A es mejor. Ahora continuemos con las comparaciones el modelo A y C

```
anova(modeloC, modeloA)

## Analysis of Variance Table
##
## Model 1: y ~ x2 + x3 + x4 + x5
## Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x20
##   Res.Df   RSS Df Sum of Sq    F   Pr(>F)
## 1     145 927.97
## 2     143 856.00  2    71.971 6.0115 0.003113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con un 99 % de confianza se puede rechazar la hipótesis nula de que el modelo C es mejor que el A. En otras palabras, el mejor modelo es el A:

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

Los resultados de la estimación de este modelo se presentan en el Cuadro 6.11.

Cuadro 6.11: Mejor modelo seleccionado.

<i>Dependent variable:</i>	
	y Mejor modelo
x1	0.663** (0.257)
x2	1.664*** (0.265)
x3	0.932*** (0.258)
x4	1.087*** (0.253)
x5	1.056*** (0.251)
x20	-0.715*** (0.266)
Constant	11.464*** (1.343)
Observations	150
R ²	0.720
Adjusted R ²	0.708
Residual Std. Error	2.447 (df = 143)
F Statistic	61.274*** (df = 6; 143)

Note: *p<0.1; **p<0.05; ***p<0.01

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* (Orestes Cerdeira y col., 2020) cuenta con algoritmos genéticos (GA) para la selección de modelos (ver función **anneal()**). También se puede explorar la función **RegBest()** del paquete *FactoMineR* (Lê y col., 2008) que emplea otras técnicas de inteligencia artificial para la selección de modelos.

7 . Primer caso de negocio

Diseñado por Freepik

Objetivos del capítulo

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

- Emplear las herramientas estudiadas en los capítulos anteriores para responder una pregunta de negocio que implique analítica diagnóstica
- Presentar los resultados de una regresión de manera gráfica empleando R.
- Determinar cuál variable tiene más efecto sobre la variable explicativa empleando R.

7.1 Introducción

En los capítulos anteriores hemos estudiado las bases del modelo clásico de regresión múltiple y cómo encontrar el mejor modelo para hacer analítica diagnóstica o analítica predictiva. En este capítulo pondremos todos los elementos juntos para resolver un caso de negocio que implica analítica diagnóstica. Adicionalmente, discutiremos cómo determinar cuál es la variable que tiene más impacto sobre la variable dependiente y cómo presentar los resultados de un modelo de regresión de manera visual. Por otro lado, es importante aclarar que aún no verificaremos el cumplimiento de los supuestos del modelo de regresión múltiple que se discutirán en la segunda parte de este libro. Nuestro análisis no estará completo hasta que se haga dicho supuesto. Pero este es un buen momento para hacer un alto en el camino y aplicar todo lo que hemos estudiado hasta el momento.

7.2 La pregunta de negocio

Mashable (<https://mashable.com>) es un portal de noticias en Internet que está interesado en entender de qué depende el número de veces que es compartido (*shares*) en redes sociales un artículo publicado por el portal para poder determinar políticas editoriales. La pregunta de negocio que tiene el editor es ¿De qué depende el número de *shares* de un artículo? Es más, esta pregunta, como de costumbre, implica otra pregunta ¿Existe alguna variable accionable¹ que pueda ser modificada para generar una recomendación a los escritores? Nota que entre más veces se comparta un artículo más ingresos generará al portal y de ahí el interés de tener una guía para que los escritores generen artículos que sean muy compartidos. Nuestra tarea en este Capítulo es responder esa pregunta de negocio y hacer recomendaciones prácticas a los escritores. Esto implicará presentar nuestros resultados de una manera amigable a diferentes audiencias.

Para responder esta pregunta contamos con una base de datos con los artículos publicados en un periodo de dos años suministrada por Fernandes y col. (2015). La base de datos se encuentra en el archivo *DatosCaso1.csv*. Estos datos son reales y fueron descargados de la siguiente página <https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>.

La base de datos contiene 39644 observaciones y 61 variables que se presentan en el Cuadro 7.1.

¹Por una variable accionable en la jerga de los negocios, es aquella que le permite a la organización desarrollar estrategias y campañas que permitan el logro de un objetivo.

Cuadro 7.1: Variables presentes en la base de datos del caso de negocio.

Variable	Descripción
url	URL del artículo
timedelta	Días entre la publicación del artículo y la fecha de corte de la base de datos
n_tokens_title	Número de palabras del título
n_tokens_content	Número de palabras en el contenido
n_unique_tokens	Tasa de palabras únicas en el contenido
n_non_stop_words	Tasa de palabras <i>non-stop words</i> ² .
n_non_stop_unique_tokens	Tasa de palabras únicas <i>non-stop</i> en el contenido
num_hrefs	Número de enlaces
num_self_hrefs	Número de enlaces a otros artículos publicados por Mashable
num_imgs	Número de imágenes
num_videos	Número de videos
average_token_length	Longitud promedio de las palabras del contenido
num_keywords	Número de palabras clave en los metadatos
data_channel_is_lifestyle	¿Es el canal de datos estilo de vida?
data_channel_is_entertainment	¿Es el canal de datos entretenimiento?
data_channel_is_bus	¿Es el canal de datos Business?
data_channel_is_socmed	¿Es el canal de datos Redes sociales?
data_channel_is_tech	¿Es el canal de datos Tech?
data_channel_is_world	¿Es el canal de datos Mundo?
kw_min_min	Peor palabra clave (min. shares)
kw_max_min	Peor palabra clave (max. shares)
kw_avg_min	Peor palabra clave (avg. shares)
kw_min_max	Mejor palabra clave (min. shares)
kw_max_max	Mejor palabra clave (max. shares)
kw_avg_max	Mejor palabra clave (avg. shares)
kw_min_avg	Promedio palabra clave (min. shares)
kw_max_avg	Promedio palabra clave (max. shares)
kw_avg_avg	Promedio palabra clave (avg. shares)
self_reference_min_shares	Mínimos de <i>shares</i> de artículos referenciados en Mashable
self_reference_max_shares	Máximos de <i>shares</i> de artículos referenciados en Mashable
self_reference_avg_shares	Promedio de <i>shares</i> de artículos referenciados en Mashable
weekday_is_monday	¿Se publicó el artículo un lunes?
weekday_is_tuesday	¿Se publicó el artículo un martes?
weekday_is_wednesday	¿Se publicó el artículo un miércoles?
weekday_is_thursday	¿Se publicó el artículo un jueves?
weekday_is_friday	¿Se publicó el artículo un viernes?
weekday_is_saturday	¿Se publicó el artículo un sábado?
weekday_is_sunday	¿Se publicó el artículo un fin de semana?
is_weekend	Cercanía al tema 0 del LDA ³ .
LDA_00	Cercanía al tema 1 del LDA
LDA_01	Cercanía al tema 2 del LDA
LDA_02	Cercanía al tema 3 del LDA
LDA_03	Cercanía al tema 4 del LDA
LDA_04	Índice de Subjetividad del texto
global_subjectivity	Polaridad de sentimientos del texto
global_sentiment_polarity	Tasa de palabras positivas en el contenido
global_rate_positive_words	Tasa de palabras negativas en el contenido
global_rate_negative_words	Tasa de palabras positivas entre los tokens no neutrales
rate_positive_words	Tasa de palabras negativas entre los tokens no neutrales
rate_negative_words	Polaridad media de las palabras positivas
avg_positive_polarity	Polaridad mínima de las palabras positivas
min_positive_polarity	Polaridad máxima de las palabras positivas
max_positive_polarity	Polaridad media de las palabras negativas
avg_negative_polarity	Polaridad mínima de las palabras negativas
min_negative_polarity	Polaridad máxima de las palabras negativas
max_negative_polarity	Subjetividad del título
title_subjectivity	Polaridad del título
title_sentiment_polarity	Nivel de subjetividad absoluta
abs_title_subjectivity	Nivel de polaridad absoluta
abs_title_sentiment_polarity	Número de acciones (Variable dependiente)
shares	

Fuente: Fernandes y col. (2015)

¹Las *stop words* o palabras vacías son palabras comunes de un idioma que no aportan al análisis como por ejemplo: los, las, tendremos, etc. Para discusión introductoria al análisis de textos se puede consultar Alonso Cifuentes (2020).

²La LDA (*Latent Dirichlet Allocation*) en este contexto es una variable generada por modelo estadístico que asocia palabras recogidas en documentos y las asocia con un pequeño número de temas. Los modelos que generan los LDA pertenece al campo del aprendizaje de máquina.

7.3 El plan

La primera tarea del científico de datos y de todo el equipo de analítica de una organización es precisar al máximo la pregunta de negocio que se desea responder. En este caso ya la pregunta de negocio está clara. Así mismo, de la mano de la definición de la pregunta de negocio va la identificación de los datos disponibles y la técnica o modelo a emplear. En este caso también esto es muy claro, contamos con una base de datos definida y limpia y la técnica a emplear es la regresión múltiple. Ahora debemos trazar una ruta analítica para responder la pregunta de negocio.

Los pasos que podemos desarrollar en este caso son:

1. Encontrar diferentes modelos candidatos a ser el mejor modelo y limpiarlos de variables no significativas
2. Comparar los modelos candidatos para seleccionar un único modelo
3. Identificar la variable más importante para explicar la variable dependiente
4. Generar las recomendaciones
5. Generar visualizaciones de los resultados

Empecemos a ejecutar esa ruta analítica para resolver la pregunta de negocio

7.4 Detección de posibles modelos

En este caso tenemos que explicar la variable *shares* para lo cuál contamos con 59 potenciales variables explicativas. Nota que la primera variable en la base de datos no es relevante (*url*), ésta corresponde al enlace del artículo. Esto implica que tendremos $5,7646075 \times 10^{17}$ posibles modelos. Un número muy grande de modelos como para emplear la fuerza bruta. Esto implica la necesidad de emplear estrategias inteligentes de detección de un mejor modelo.

Empecemos por leer los datos y eliminar la primera variable que no es relevante.

```
datos.caso1 <- read.csv("../Data/DatosCaso1.csv", sep = ",")  
datos.caso1 <- datos.caso1[, -1]
```

Nota que los datos quedaron bien cargados y las clases de las variables son las correctas. Tu puedes constatar que no existen datos perdidos y que la base está lista para iniciar a trabajar.

Procedamos a encontrar los mejores modelos empleando las estrategias de regresión paso a paso forward, backward y combinada con el AIC, con el valor p y el R^2 ajustado. Empleando las 9 opciones de algoritmos que se presentan en el Cuadro 7.2 estimaremos los modelos y eliminaremos aquellas variables que no sean significativas. Los modelos que obtenemos las guardaremos con los nombres que se presentan el Cuadro 7.2.

Cuadro 7.2: Modelos a estimar con los diferentes algoritmos

Nombre del objeto	Algoritmo	Criterio
modelo1	Forward	R^2 ajustado
modelo2	Forward	valor p
modelo3	Forward	AIC
modelo4	Backward	R^2 ajustado
modelo5	Backward	valor p
modelo6	Backward	AIC
modelo7	Both	R^2 ajustado
modelo8	Both	valor p
modelo9	Both	AIC

Antes de iniciar este proceso, partamos de estimar los modelos lineales con todas las variables potenciales (`max.model`) y sin variables (`min.model`).

```
# modelo con todas las variables
max.model <- lm(shares ~ ., data = datos.caso1)
# modelo sin variables
min.model <- lm(shares ~ 1, data = datos.caso1)
```

Nota que la variable `weekday_is_sunday` genera el fenómeno conocido como la trampa de las variables dummy (ver 5), pues es redundante al tener las otras 6 variables dummy para los otros días de la semana. R detecta esto y si bien se incluye en la fórmula no se incluye en la regresión. Lo mismo ocurre con la variable `is_weekend..`.

Ahora procedamos a encontrar modelos candidatos para ser los mejores modelos. Empecemos con la estrategia *stepwise Forward*.

7.4.1 Stepwise forward

Empleando lo aprendido en el Capítulo 6 podemos obtener los modelos reportados en el Cuadro 7.3 tras limpiar las variables no significativas (con un 95 % de confianza). Estas estimaciones pueden tomar un tiempo considerable. Tienes que tener paciencia para obtener estos resultados.

Cuadro 7.3: Modelo seleccionado el algoritmo stepwise forward

	Dependent variable:		
	shares		
	Modelo 1 (R^2 aj)	Modelo 2 (valor p)	Modelo 3 (AIC)
	(1)	(2)	(3)
timedelta	1.989*** (0.289)	1.956*** (0.298)	1.934*** (0.301)
n_tokens_title	115.292*** (28.484)	114.826*** (28.508)	112.215*** (28.518)
num_hrefs	32.445*** (5.996)	32.014*** (6.093)	28.069*** (6.318)
num_self_hrefs	-50.164*** (16.821)	-56.079*** (17.044)	-50.839*** (17.026)
n_tokens_content			0.313** (0.148)
LDA_02		-736.567*** (245.926)	-655.120*** (245.728)
global_rate_positive_words		-9.889.320** (4.102.061)	-10.560.820*** (4.098.269)
min_positive_polarity		-2.079.226* (897.810)	-1.882.670* (925.530)
num_imgs	18.426** (7.588)	16.852** (7.621)	
average_token_length	-465.212*** (91.107)	-375.242*** (94.275)	-336.700*** (95.554)
data_channel_is_entertainment	-720.981*** (155.162)	-844.317*** (160.579)	-885.439*** (162.484)
LDA_03			711.475*** (252.221)
kw_min_max	-0.003** (0.001)	-0.003*** (0.001)	-0.003*** (0.001)
weekday_is_saturday		584.363** (242.131)	583.713** (242.093)
kw_min_avg	-0.412*** (0.070)	-0.393*** (0.070)	-0.366*** (0.071)
kw_max_avg	-0.208*** (0.020)	-0.197*** (0.020)	-0.186*** (0.021)
kw_avg_avg	1.814*** (0.106)	1.743*** (0.112)	1.655*** (0.123)
self_reference_min_shares	0.023*** (0.003)	0.022*** (0.003)	0.023*** (0.003)
self_reference_max_shares	0.003** (0.002)	0.003** (0.002)	0.003** (0.002)
data_channel_is_lifestyle		-519.056** (264.402)	
weekday_is_monday	500.355*** (157.046)	467.261*** (155.696)	470.091*** (155.693)
global_subjectivity	2.257.169*** (674.289)	2.751.031*** (749.067)	2.545.088*** (751.661)
avg_negative_polarity	-1.762.793*** (513.995)	-1.675.100*** (518.389)	-1.517.854*** (520.352)
is_weekend	399.482** (174.688)		
Constant	-2.273.864*** (542.108)	-1.863.821*** (563.070)	-1.947.287*** (565.875)
Observations	39.644	39.644	39.644
R ²	0.022	0.023	0.023
Adjusted R ²	0.022	0.022	0.022
Residual Std. Error	11.500.240 (df = 39626)	11.498.180 (df = 39622)	11.497.820 (df = 39622)
F Statistic	52.669*** (df = 17; 39626)	43.519*** (df = 21; 39622)	43.641*** (df = 21; 39622)

Note:

*p<0.1; **p<0.05; ***p<0.01

Los resultados presentados en el Cuadro 7.3 muestran tres modelos que no se encuentran anidados.

7.4.2 Stepwise backward

De manera similar en el Cuadro 7.4 se presentan los resultados de emplear el algoritmo *stepwise backward* y tras limpiar las variables no significativas (con un 95 % de confianza).

Cuadro 7.4: Modelo seleccionado el algoritmo stepwise backward

	Dependent variable:		
	Modelo 4 (R^2 aj)	shares	Modelo 5 (AIC)
		(1)	(2)
timedelta	2.061*** (0.290)	2.005*** (0.300)	
n_tokens_title	109.600*** (28.515)	117.813*** (28.496)	
n_tokens_content	0.344** (0.143)	0.392*** (0.144)	
num_hrefs	29.114*** (6.263)	28.770*** (6.193)	
num_self_hrefs	-46.305*** (16.876)	-48.638*** (16.926)	
average_token_length	-270.978*** (77.560)		
data_channel_is_lifestyle		-548.165** (264.778)	
data_channel_is_entertainment	-806.415*** (158.995)	-820.385*** (160.393)	
kw_min_max	-0.003*** (0.001)	-0.003*** (0.001)	
kw_min_avg	-0.394*** (0.071)	-0.402*** (0.070)	
kw_max_avg	-0.202*** (0.021)	-0.200*** (0.021)	
kw_avg_avg	1.764*** (0.118)	1.785*** (0.113)	
self_reference_min_shares	0.023*** (0.003)	0.026*** (0.003)	
self_reference_max_shares	0.003** (0.002)		
weekday_is_monday	498.951*** (157.051)		
LDA_03	831.991*** (248.729)		3,823.354*** (260.552)
weekday_is_tuesday		-501.242*** (170.702)	
weekday_is_wednesday		-353.003** (170.514)	
weekday_is_thursday		-520.139*** (171.619)	
weekday_is_friday		-467.728** (185.668)	
LDA_02		-829.469*** (244.930)	
global_subjectivity		2,573.787*** (762.226)	
global_rate_positive_words		-9,992.733*** (4,561.701)	
LDA_04			1,586.553*** (266.837)
rate_positive_words		-1,925.147*** (538.102)	-1,090.105*** (394.075)
LDA_01			974.634*** (306.466)
data_channel_is_bus			-673.282*** (256.431)
rate_negative_words		-2,213.520*** (566.001)	-1,432.505*** (509.275)
avg_negative_polarity	-2,051.808*** (493.987)	-1,809.646*** (541.052)	
is_weekend	399.337** (174.717)		
kw_avg_max			501.066*** (172.564)
LDA_00			0.002*** (0.0005)
max_negative_polarity			2,285.665*** (389.917)
min_negative_polarity			-1,329.028** (631.231)
self_reference_avg_shares			-1,000.127*** (231.201)
Constant	-2,346.261*** (548.107)	-1,580.902*** (579.263)	0.024*** (0.002)
Observations	39.644	39.644	39.644
R ²	0.022	0.022	0.012
Adjusted R ²	0.022	0.022	0.012
Residual Std. Error	11,500.570 (df = 39626)	11,499.790 (df = 39621)	11,556.590 (df = 39631)
F Statistic	52.534*** (df = 17; 39626)	41.072*** (df = 22; 39621)	41.347*** (df = 12; 39631)

Note:

*p<0.1; **p<0.05; ***p<0.01

Los resultados presentados en el Cuadro 7.4 muestran también tres modelos que no se encuentran anidados.

7.4.3 Combinando forward y backward

Y finalmente, el Cuadro 7.5 se presentan los resultados de emplear el algoritmo combinado y tras limpiar las variables no significativas (con un 95 % de confianza).

```
## Reordering variables and trying again:
```

Cuadro 7.5: Modelo seleccionado el algoritmo stepwise forward y backward

	Dependent variable:		
	shares		
	Modelo 7 (R^2 aj)	Modelo 8 (valor p)	Modelo 9 (AIC)
	(1)	(2)	(3)
timedelta	1.813*** (0.296)	1.934*** (0.301)	1.934*** (0.301)
n_tokens_title	112.667*** (28.518)	112.215*** (28.518)	112.215*** (28.518)
num_hrefs	35.948*** (5.852)	28.069*** (6.318)	28.069*** (6.318)
num_self_hrefs	-45.905*** (16.816)	-50.839*** (17.026)	-50.839*** (17.026)
n_tokens_content		0.313** (0.148)	0.313** (0.148)
average_token_length	-262.432*** (77.592)	-336.700*** (95.554)	-336.700*** (95.554)
global_subjectivity		2,545.088*** (751.661)	2,545.088*** (751.661)
data_channel_is_entertainment	-862.271*** (161.735)	-885.439*** (162.484)	-885.439*** (162.484)
kw_min_max	-0.003*** (0.001)	-0.003*** (0.001)	-0.003*** (0.001)
weekday_is_saturday		583.713** (242.093)	583.713** (242.093)
kw_min_avg	-0.368*** (0.071)	-0.366*** (0.071)	-0.366*** (0.071)
kw_max_avg	-0.185*** (0.021)	-0.186*** (0.021)	-0.186*** (0.021)
kw_avg_avg	1.648*** (0.122)	1.655*** (0.123)	1.655*** (0.123)
self_reference_min_shares	0.023*** (0.003)	0.023*** (0.003)	0.023*** (0.003)
self_reference_max_shares	0.003** (0.002)	0.003** (0.002)	0.003** (0.002)
weekday_is_monday	487.990*** (157.080)	470.091*** (155.693)	470.091*** (155.693)
LDA_02	-643.086*** (235.856)	-655.120*** (245.728)	-655.120*** (245.728)
global_rate_positive_words		-10,560.820*** (4,098.269)	-10,560.820*** (4,098.269)
min_positive_polarity		-1,882.670** (925.530)	-1,882.670** (925.530)
LDA_03	670.391*** (247.454)	711.475*** (252.221)	711.475*** (252.221)
avg_negative_polarity	-2,160.921*** (492.394)	-1,517.854*** (520.352)	-1,517.854*** (520.352)
is_weekend	414.940** (174.657)		
Constant	-1,815.480*** (562.734)	-1,947.287*** (565.875)	-1,947.287*** (565.875)
Observations	39.644	39.644	39.644
R ²	0.022	0.023	0.023
Adjusted R ²	0.022	0.022	0.022
Residual Std. Error	11,500.330 (df = 39626)	11,497.820 (df = 39622)	11,497.820 (df = 39622)
F Statistic	52.633*** (df = 17; 39626)	43.641*** (df = 21; 39622)	43.641*** (df = 21; 39622)

Note:

*p<0.1; **p<0.05; ***p<0.01

Los resultados presentados en el Cuadro 7.5 muestran que los modelos seleccionados por los criterios de valor p (Modelo 8) y AIC (Modelo 9) y el algoritmo combinado son el mismo. El modelo obtenido con este algoritmo y el criterio de R^2 aj no está anidado en estos dos modelos anteriores.

7.5 Comparación de modelos

En resumen, contamos con 9 modelos con las variables explicativas que se representan con una X en el Cuadro 7.6. Los modelos 3, 8 Y 9 son los mismos. Los otros seis modelos no se encuentran anidados. Por eso tendremos que comparar estos modelos con pruebas de modelos no anidados.

Cuadro 7.6: Variables explicativas incluidas en cada uno de los modelos calculados

	Forward			Backward			Both		
	modelo 1	modelo 2	modelo 3	modelo 4	modelo 5	modelo 6	modelo 7	modelo 8	modelo 9
timedelta	X	X	X	X	X		X	X	X
n_tokens_title	X	X	X	X	X		X	X	X
n_tokens_content				X	X		X	X	X
num_href	X	X	X	X	X		X	X	X
num_self_href	X	X	X	X	X		X	X	X
average_token_length					X		X	X	X
num_imgs average_token_length	X	X							
data_channel_is_lifestyle					X				
data_channel_is_entertainment	X	X	X	X	X		X	X	X
kw_min_max				X	X		X	X	X
kw_min_avg		X	X	X	X		X	X	X
kw_max_avg	X	X	X	X	X		X	X	X
kw_avg_avg	X	X	X	X	X		X	X	X
kw_avg_max					X				
self_reference_min_shares	X	X	X	X	X		X	X	X
self_reference_max_shares	X	X	X	X			X	X	X
self_reference_avg_shares					X				
data_channel_is_lifestyle				X					
data_channel_is_bus					X				
weekday_is_monday	X	X	X	X			X	X	X
weekday_is_tuesday					X				
weekday_is_wednesday						X			
weekday_is_thursday						X			
weekday_is_friday						X			
weekday_is_saturday	X	X	X				X	X	
is_weekend	X			X		X			
LDA_00					X				
LDA_01						X			
LDA_02		X	X		X		X	X	X
LDA_03	X	X	X	X	X	X	X	X	
LDA_04					X				
global_rate_positive_words	X	X			X		X	X	
global_subjectivity	X	X	X		X		X	X	
min_positive_polarity	X	X					X	X	
avg_negative_polarity	X	X	X	X	X		X	X	X
max_negative_polarity						X			
min_negative_polarity						X			
rate_positive_words					X	X			
rate_negative_words					X	X			

Empecemos comparando todos los modelos con la prueba J. En el Cuadro 7.7 se reportan los valores p de las pruebas J que permiten probar la hipótesis nula de que el modelo de la fila es mejor que el de la columna.

Cuadro 7.7: Valores p de las pruebas J (H_0 : modelo de la fila es mejor que el de la columna)

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7
Modelo 1		0.000	0.000	0.002	0.000	0.002	0.001
Modelo 2	0.243		0.006	0.003	0.084	0.028	0.015
Modelo 3	0.079	0.026		0.282	0.054	0.775	0.282
Modelo 4	0.001	0.000	0.000		0.000	0.043	0.005
Modelo 5	0.002	0.000	0.000	0.000		0.005	0.000
Modelo 6	0.000	0.000	0.000	0.000	0.000		0.000
Modelo 7	0.001	0.000	0.000	0.013	0.000	0.853	

Si miramos la primera fila, con un 99 % de confianza, podemos concluir que el modelo 1 no es mejor que los otros modelos⁴. Si miramos la primera columna, podemos ver cómo la nula de que el

⁴Nota que la hipótesis nula asociada a los valores p reportados en la primera fila del Cuadro 7.7 corresponde a que el

modelo 2 es mejor que el 1 no se puede rechazar y lo mismo ocurre para el modelo 3 comparado con el modelo 1. Es decir, con un 99% de confianza, podemos concluir que los modelos 2 y 3 son mejores que el 1. Para los otros modelos no se puede afirmar algo similar, y por tanto la prueba no es concluyente. Al comparar el modelo 2 con el 3, se encuentra que se puede rechazar la nula de que el modelo 2 es mejor que el tres, pero no que el modelo 3 es mejor que el 2. Es decir, el modelo 3 es mejor. La prueba no puede concluir al comparar el modelo 2 con el 4; y permite concluir que el modelo 2 es mejor que el 5, 6 y 7.

Para el modelo 3, no se puede rechazar que este modelo sea mejor que cada uno de los otros 6 modelos, pero las hipótesis nulas opuestas si se pueden rechazar (con un 99% de confianza). Podemos encontrar que el modelo 4 es mejor que el 6. El modelo 5 no es mejor que los otros modelos, lo mismo ocurre con el modelo 6. Y el modelo 7 es mejor que el 4 y el 6. Para las otras comparaciones que no se mencionan la prueba no es concluyente. Es decir, poniendo todo junto el modelo 3 es el mejor.

Ahora empleemos las métricas AIC y BIC y el R^2 ajustado, para comparar los modelos. Los resultados se reportan en el Cuadro 7.8 y BIC()

Cuadro 7.8: Medidas de bondad de ajuste para los 7 modelos comparados

	R2.ajustado	AIC	BIC
Modelo 1	0.022	853877.384	854040.550
Modelo 2	0.022	853867.163	854064.680
Modelo 3	0.022	853864.662	854062.179
Modelo 4	0.022	853879.621	854042.788
Modelo 5	0.022	853879.238	854085.342
Modelo 6	0.012	854259.948	854380.176
Modelo 7	0.022	853877.986	854041.152

El \bar{R}^2 y el AIC sugieren que el mejor modelo es el 3, mientras que el BIC selecciona el 1. Poniendo todo junto, el mejor modelo será el modelo 3 (que es igual al 8 y 9) el cuál se reporta en el Cuadro 7.9.

modelo 1 es mejor al modelo de la respectiva columna. Y esa hipótesis nula se puede rechazar en todos los casos.

Cuadro 7.9: Mejor modelo

<i>Dependent variable:</i>	
	shares Modelo 3
kw_avg_avg	1.655*** (0.123)
self_reference_min_shares	0.023*** (0.003)
kw_max_avg	−0.186*** (0.021)
kw_min_avg	−0.366*** (0.071)
timedelta	1.934*** (0.301)
num_hrefs	28.069*** (6.318)
n_tokens_title	112.215*** (28.518)
data_channel_is_entertainment	−885.439*** (162.484)
LDA_03	711.475*** (252.221)
avg_negative_polarity	−1.517.854*** (520.352)
average_token_length	−336.700*** (95.554)
global_subjectivity	2.545.088*** (751.661)
weekday_is_monday	470.091*** (155.693)
kw_min_max	−0.003*** (0.001)
weekday_is_saturday	583.713** (242.093)
num_self_hrefs	−50.839*** (17.026)
n_tokens_content	0.313** (0.148)
LDA_02	−655.120*** (245.728)
global_rate_positive_words	−10.560.820*** (4,098.269)
min_positive_polarity	−1.882.670** (925.530)
self_reference_max_shares	0.003** (0.002)
Constant	−1,947.287*** (565.875)
Observations	39.644
R ²	0.023
Adjusted R ²	0.022
Residual Std. Error	11,497.820 (df = 39622)
F Statistic	43.641*** (df = 21; 39622)

Note:

*p<0.1; **p<0.05; ***p<0.01

En todo el proceso que desarrollemos es importante no olvidar cuál es la pregunta de negocio que queremos responder. Nuestra pregunta de negocio inicial era ¿De qué depende el número de *shares* de un artículo? Esta pregunta ya la podemos responder con el mejor modelo encontrado en este caso las variables que explican los *shares* son:

```
kw_avg_avg, self_reference_min_shares, kw_max_avg, kw_min_avg, timedelta,
↪ num_hrefs, n_tokens_title, data_channel_is_entertainment, LDA_03,
↪ avg_negative_polarity, average_token_length, global_subjectivity,
↪ weekday_is_monday, kw_min_max, weekday_is_saturday, num_self_hrefs,
↪ n_tokens_content, LDA_02, global_rate_positive_words,
↪ min_positive_polarity, self_reference_max_shares,
↪ data_channel_is_lifestyle, num_keywords, kw_max_min,
↪ abs_title_sentiment_polarity, abs_title_subjectivity, kw_avg_min,
↪ kw_min_min.
```

De esas variables podemos denotar que aquellas que al aumentarse aumentan los *shares* son:

```
kw_avg_avg, self_reference_min_shares, timedelta, num_hrefs,
↪ n_tokens_title, LDA_03, global_subjectivity, weekday_is_monday,
↪ weekday_is_saturday, n_tokens_content, self_reference_max_shares.
```

Las otras variables tienen una relación inversa con los *shares*.

Ahora, este listado de variables es útil, pero contar con 28 para generar sugerencias a los escritores puede implicar una tarea ardua. Si recuerdan, la segunda pregunta de negocio derivada que teníamos es ¿Existe alguna variable accionable que pueda ser modificada para generar una recomendación a los escritores? En la siguiente sección discutiremos cómo identificar las variables más importantes al momento de explicar la variable dependiente.

7.6 Identificación de la variable más importante

Una pregunta habitual cuando estamos haciendo analítica diagnóstica es ¿cuál variable es la más importante para explicar la variable dependiente? Existen varias formas de responder esta pregunta que discutiremos a continuación.

7.6.1 Coeficientes estandarizados

Tal vez la primera respuesta que salta a la mente a la pregunta cuál variable explicativa es la más importante es emplear los coeficientes estimados ($\hat{\beta}$). Pero, ¡esta respuesta no es correcta! No se puede ver el valor del coeficiente estimado para determinar cuál variable es la más importante, dado que estos coeficientes estimados están en las unidades en las que se expresa tanto la variable

dependiente como la independiente. Así el tamaño de los coeficientes dependen de las unidades en que esté medida la variable dependiente y cada una de las variables explicativas.

Si comparamos coeficientes estimados estaríamos comparando peras con manzanas. Para resolver este problema se emplean los coeficientes estandarizados. Estos implican expresar los coeficientes estimados en términos de desviaciones estándar. Es decir, el coeficiente estandarizado para la variable explicativa j (con⁵ $j = 2, 3, \dots, k$) será:

$$\hat{\beta}_{j,estand} = \hat{\beta}_j \frac{s_j}{s_y}, \quad (7.1)$$

donde s_j y s_y representan la desviación estándar muestral del regresor j y de la variable dependiente, respectivamente.

Una vez los coeficientes se encuentran estandarizados, podremos comparar el efecto de un aumento de una desviación estándar de cada una de las variables explicativas sobre la variable dependiente; este efecto también medido en desviaciones estándar.

Una forma de calcular rápidamente los coeficientes estandarizados en R es emplear la función `calc.relimp()` del paquete `relaimpo` (Grömping, 2006). Esta función necesita dos argumentos para calcular los coeficientes estandarizados al cuadrado: un objeto de clase `lm` y el tipo de medida que se desea (argumento `type`). Para el caso de los coeficientes estandarizados, `type = "betasq"`.

```
# install.packages('relaimpo')
library(relaimpo)
# cálculo de los coeficientes estandarizados al cuadrado
coef.estandarizados <- calc.relimp(modelo3, type = "betasq")
# coeficientes estandarizados
coef.estandarizados <- sqrt(coef.estandarizados$betasq)
coef.estandarizados

##                               kw_avg_avg      self_reference_min_shares
##                         0.18759294                      0.03841779
##                               kw_max_avg          kw_min_avg
##                         0.09764905                      0.03576353
##                               timedelta        num_hrefs
##                         0.03562264                      0.02735715
##                               n_tokens_title data_channel_is_entertainment
##                           0.02040320                      0.02913085
##                               LDA_03      avg_negative_polarity
##                           0.01806327                      0.01667411
##       average_token_length      global_subjectivity
##                           0.02445276                      0.02554175
##       weekday_is_monday          kw_min_max
##                           0.01511678                      0.01413186
```

⁵Al intercepto no se le calcula este tipo de coeficientes.

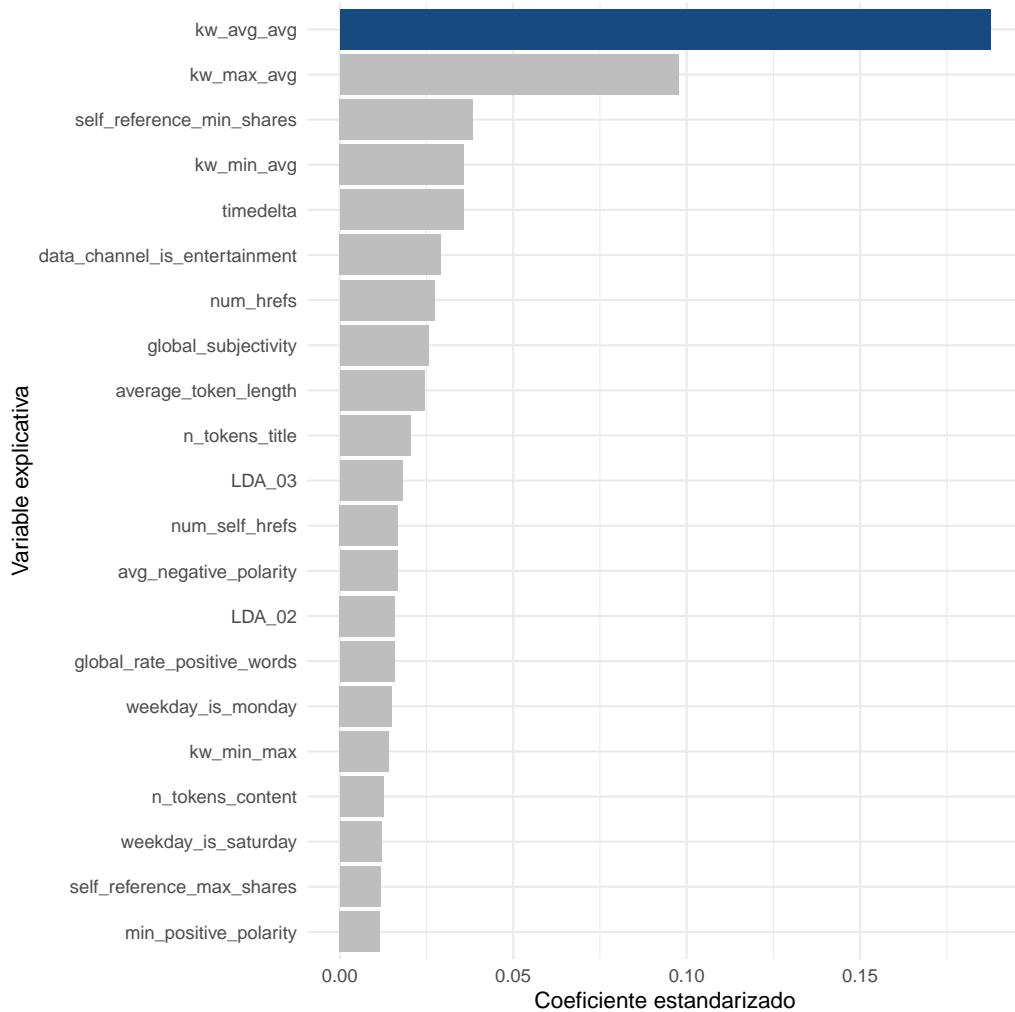
```
##      weekday_is_saturday           num_self_hrefs
##                  0.01209565          0.01685682
##      n_tokens_content              LDA_02
##                  0.01266509          0.01589746
##      global_rate_positive_words    min_positive_polarity
##                  0.01583054          0.01154752
##      self_reference_max_shares
##                  0.01169232

# encontrar el coeficiente mas grande
coef.estandarizados[which.max(coef.estandarizados)]
```



```
## kw_avg_avg
## 0.1875929
```

Ahora podríamos afirmar que la variable que más afecta a los *shares* es la variable *kw_avg_avg*. Esto lo podemos mostrar de una manera visual empleando la Figura 7.1.

Figura 7.1. Coeficientes estandarizados del Modelo 3

7.6.2 Aporte relativo de cada variable empleando el R cuadrado

Otra forma de medir el aporte relativo de cada una de las variables explicativas es determinar la proporción adicional de la variación de la variable dependiente que es explicada por cada una de las variables, dado que los otros regresores ya están incluidos en el modelo. Es decir, el aumento en el R^2 que se obtiene al adicionar el respectivo regresor dado que ya están en el modelo las otras $k - 2$ variables explicativas .

Esta medida de importancia relativa se puede calcular con la función `calc.relimp()` del paquete `relaimpo` que ya habíamos empleado. En este caso, debemos camiar el valor del argumento `type` a “`last`” .

```

aporte_relativo1 <- calc.relimp(modelo3, type = "last")

aporte_relativo1

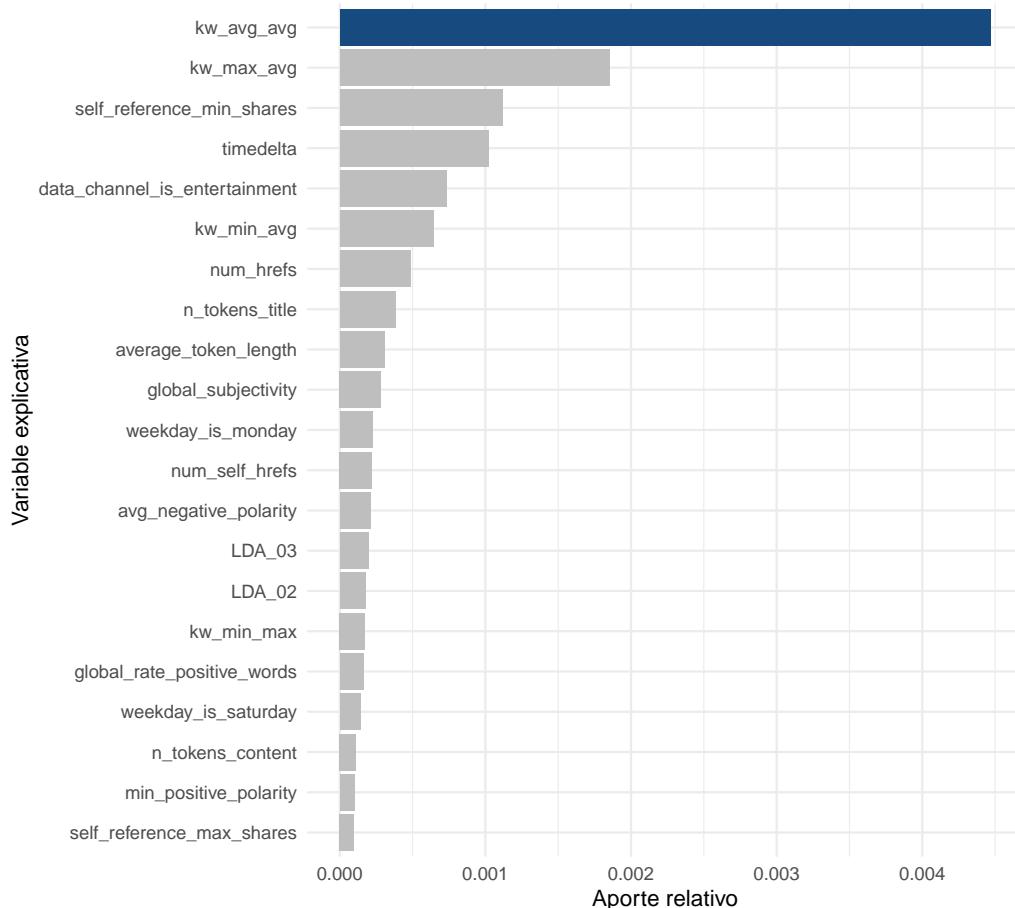
## Response variable: shares
## Total response variance: 135185984
## Analysis based on 39644 observations
##
## 21 Regressors:
## kw_avg_avg self_reference_min_shares kw_max_avg kw_min_avg timedelta num_hrefs n_tokens_title dat
## Proportion of variance explained by model: 2.26%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##                               last
## kw_avg_avg                  4.468595e-03
## self_reference_min_shares   1.115899e-03
## kw_max_avg                  1.851655e-03
## kw_min_avg                  6.450879e-04
## timedelta                   1.019827e-03
## num_hrefs                   4.869298e-04
## n_tokens_title              3.819461e-04
## data_channel_is_entertainment 7.325371e-04
## LDA_03                      1.962854e-04
## avg_negative_polarity      2.098929e-04
## average_token_length        3.062830e-04
## global_subjectivity         2.828098e-04
## weekday_is_monday          2.248855e-04
## kw_min_max                 1.737811e-04
## weekday_is_saturday         1.434062e-04
## num_self_hrefs              2.199541e-04
## n_tokens_content            1.104386e-04
## LDA_02                      1.753337e-04
## global_rate_positive_words 1.638049e-04
## min_positive_polarity      1.020706e-04
## self_reference_max_shares  9.890499e-05

aporte_relativo1$last[which.max(aporte_relativo1$last)]

##  kw_avg_avg
## 0.004468595

```

Según este método, podemos afirmar que la variable que más afecta a los *shares* es la variable *kw_avg_avg*. Esto lo podemos mostrar de una manera visual empleando la Figura 7.2.

Figura 7.2. Aporte relativo de cada variable al R^2 del Modelo 3

7.7 Generación de las recomendaciones

Ahora podemos proceder a generar las recomendaciones. Ya conocemos la respuesta a la pregunta ¿De qué depende el número de *shares* de un artículo? y ¿Existe alguna variable accionable que pueda ser modificada para generar una recomendación a los escritores? Es mas ya sabemos cuáles son las variables más importantes. Así podemos proceder a realizar recomendaciones.

Sabemos que las variables de las que dependen los *shares* son las siguientes 28:

```

kw_avg_avg, self_reference_min_shares, kw_max_avg, kw_min_avg, timedelta,
↳ num_hrefs, n_tokens_title, data_channel_is_entertainment, LDA_03,
↳ avg_negative_polarity, average_token_length, global_subjectivity,
↳ weekday_is_monday, kw_min_max, weekday_is_saturday, num_self_hrefs,
↳ n_tokens_content, LDA_02, global_rate_positive_words,
↳ min_positive_polarity, self_reference_max_shares,
↳ data_channel_is_lifestyle, num_keywords, kw_max_min,
↳ abs_title_sentiment_polarity, abs_title_subjectivity, kw_avg_min,
↳ kw_min_min.

```

De esas variables la mas importante es `kw_avg_avg` (independientemente del método que empleemos).

Entonces, el promedio de palabras claves debe ser lo más grande posible, pues esta es la variable mas importante al momento de explicar los *shares*.

Tu puedes continuar generando recomendaciones con los resultados. Por ejemplo, nota que existen variables no accionables como `timedelta` sobre la cual no se puede actuar, no obstante en este caso la mayoría de variables son accionables.

7.8 Generación de visualizaciones de los resultados

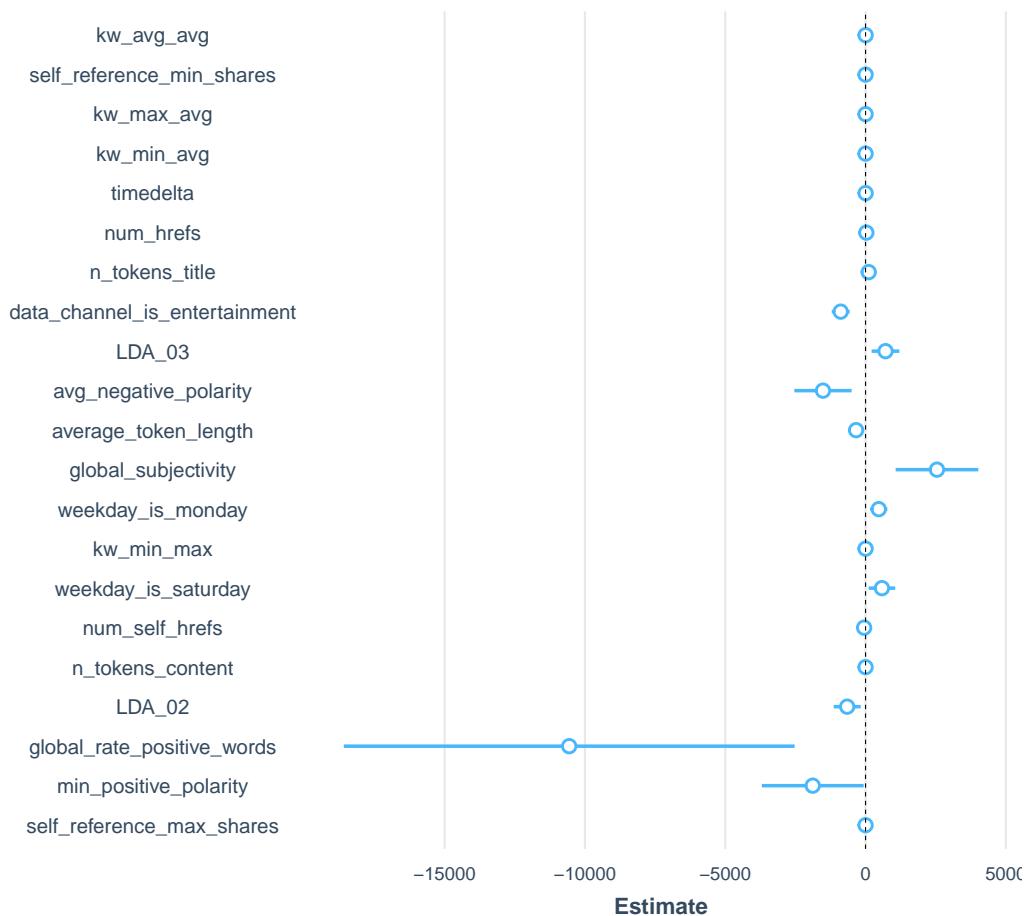
Ahora veamos cómo se pueden presentar los resultados de una manera mas amigable que emplear cuadros. El Cuadro 7.9 puede no ser la mejor opción de presentar los resultados para la mayoría de públicos, en especial para los tomadores de decisiones. En esos casos podríamos emplear gráficos para mostrar los resultados. Por ejemplo, el paquete *jtools* (Long, 2020) permite visualizar los resultados de un objeto `lm`. Veamos rápidamente como funciona este paquete.

La función `plot_summs()` permite visualizar rápidamente un objeto `lm`. Si solo usamos como argumento un objeto `lm` obtendremos una visualización rápida.

```

library(jtools)
plot_summs(modelo3)

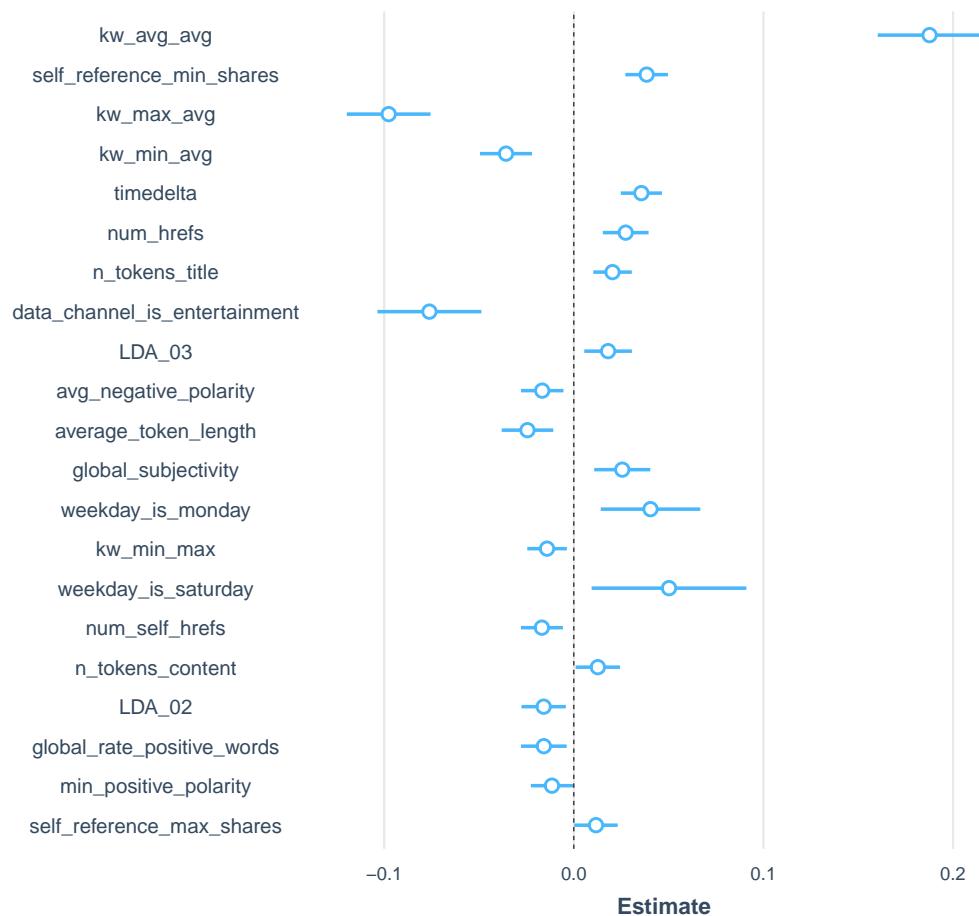
```



Este gráfico permite ver los coeficientes y sus respectivos intervalos de confianza. El intervalo por defecto es del 95 % y emplea los errores estándar estimados por MCO.

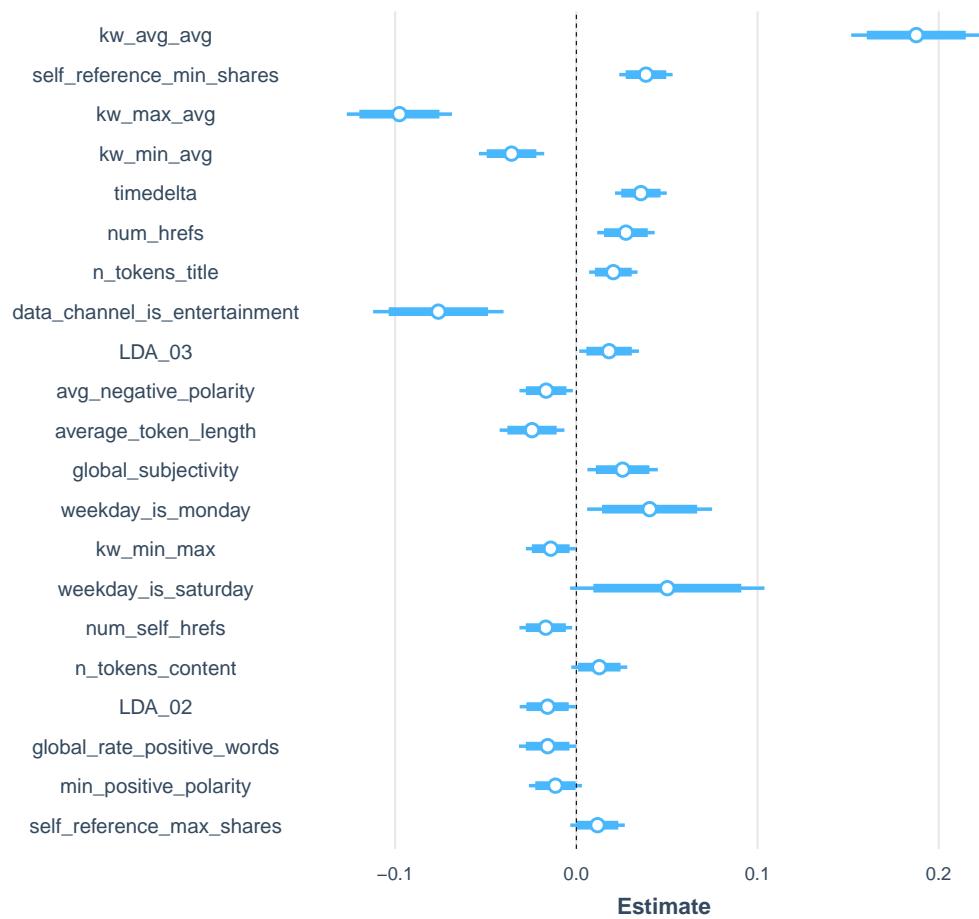
Un problema de este gráfico es que los coeficientes están en diferentes escalas y el estimador del coeficiente asociado a “global_rate_positive_words” es grande (en valor absoluto) respecto a los otros coeficientes. Algo similar ocurre con el coeficiente asociado a “global_subjectivity”. Cómo lo discutimos en la sección 7.6.1, los coeficientes estimados ($\hat{\beta}$) dependen de las unidades en que se midan las variables explicativas (y la dependiente). Como lo vimos, una forma de evitar esto es graficar los coeficientes estandarizados empleando los argumentos **scale** y **transform.response**. Por defecto estos dos argumentos son fijados en FALSE, si los pasamos a TRUE las variables explicativas (**scale**) y la dependiente (**transform.response**) son estandarizadas. En nuestro caso podemos obtener los coeficientes estandarizados y los correspondientes intervalos de confianza de la siguiente manera.

```
plot_summs(modelo3, scale = TRUE, transform.response = TRUE)
```



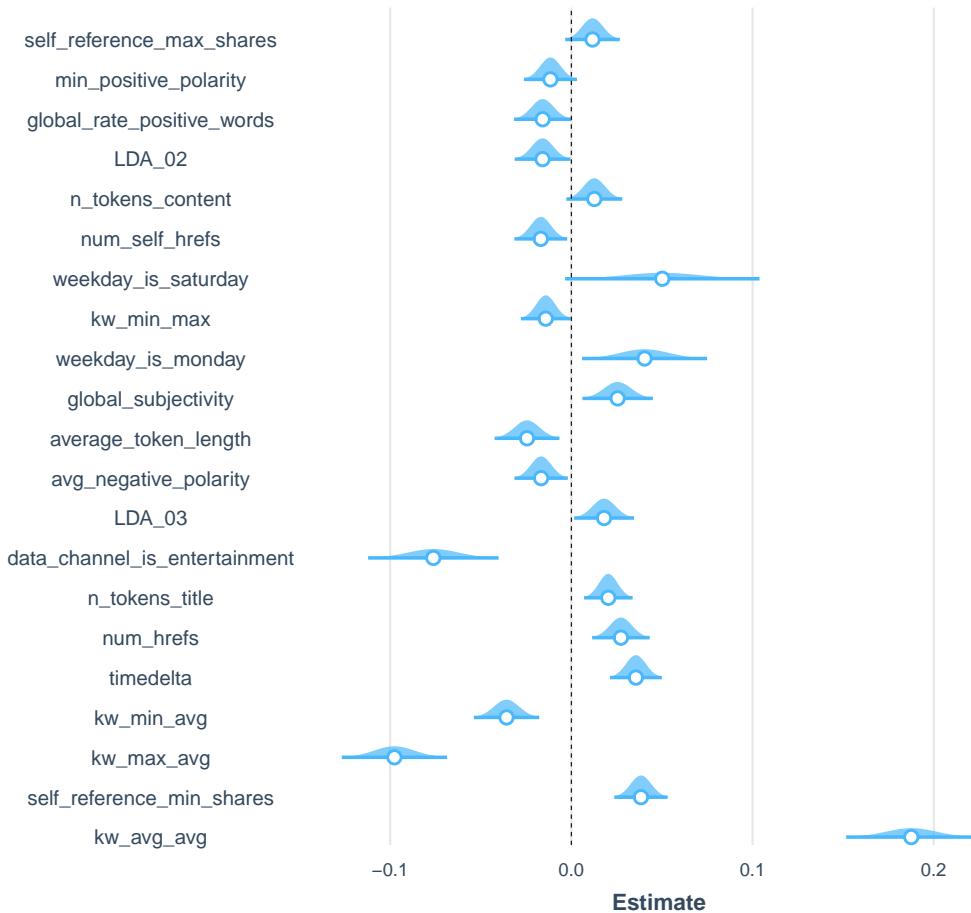
Podemos cambiar otras opciones del gráfico. Por ejemplo, podemos incluir un intervalo de confianza del 99 % y del 95 %.

```
plot_summs(modelo3, scale = TRUE, transform.response = TRUE,
           ci_level = 0.99, inner_ci_level = 0.95)
```



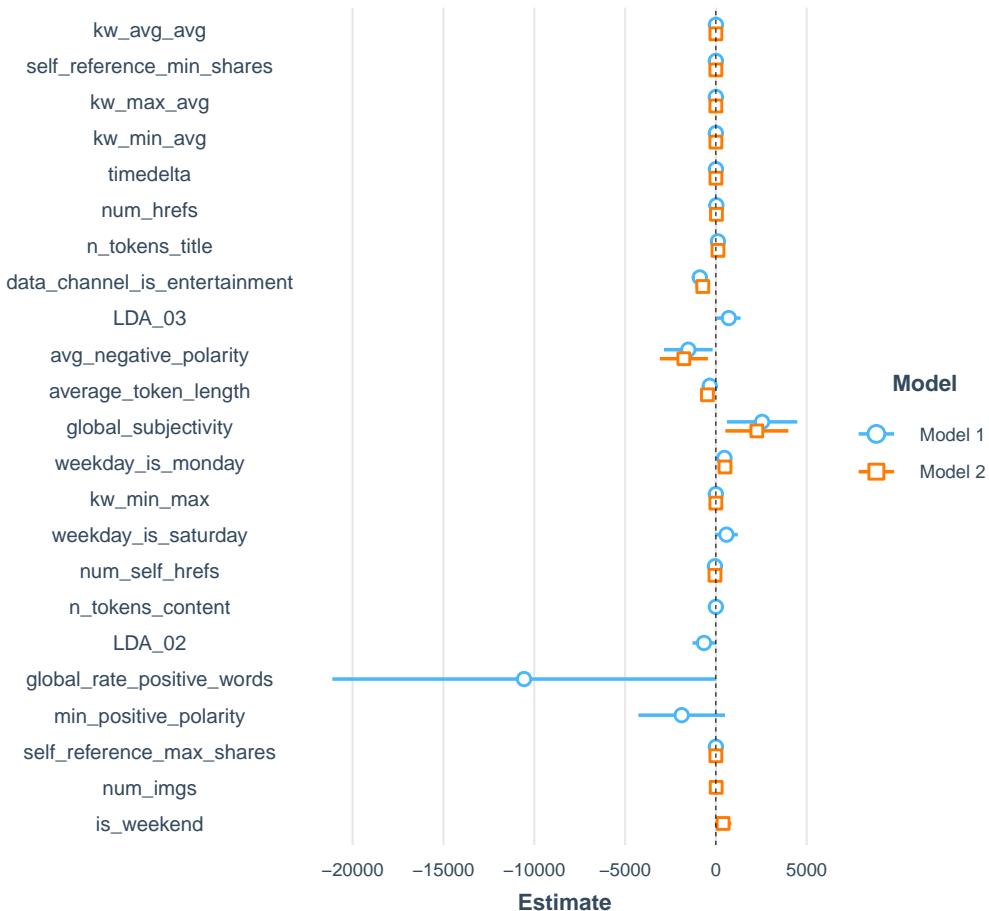
También podemos incluir la distribución asintótica de los estimadores.

```
plot_summs(modelo3, scale = TRUE, transform.response = TRUE,
            ci_level = 0.99, plot.distributions = TRUE)
```

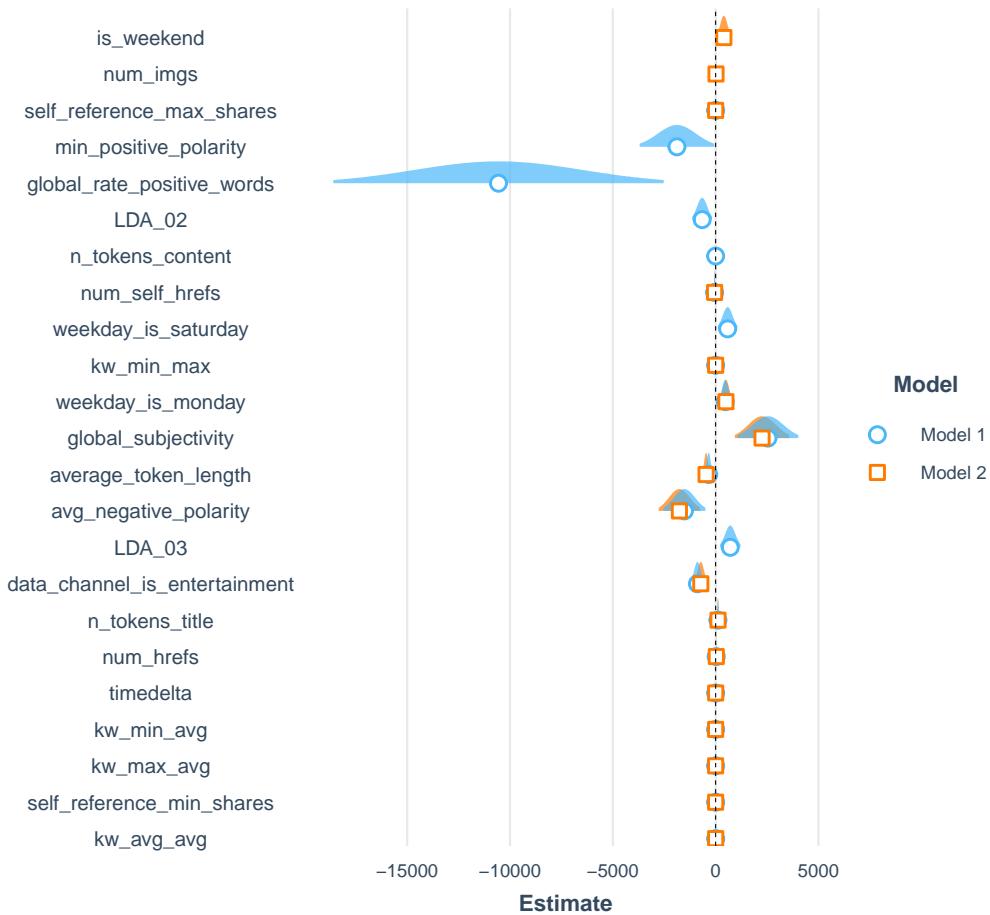


Además podemos mostrar dos modelos al mismo tiempo. Por ejemplo,

```
plot_summs(modelo3, modelo1, ci_level = 0.99)
```



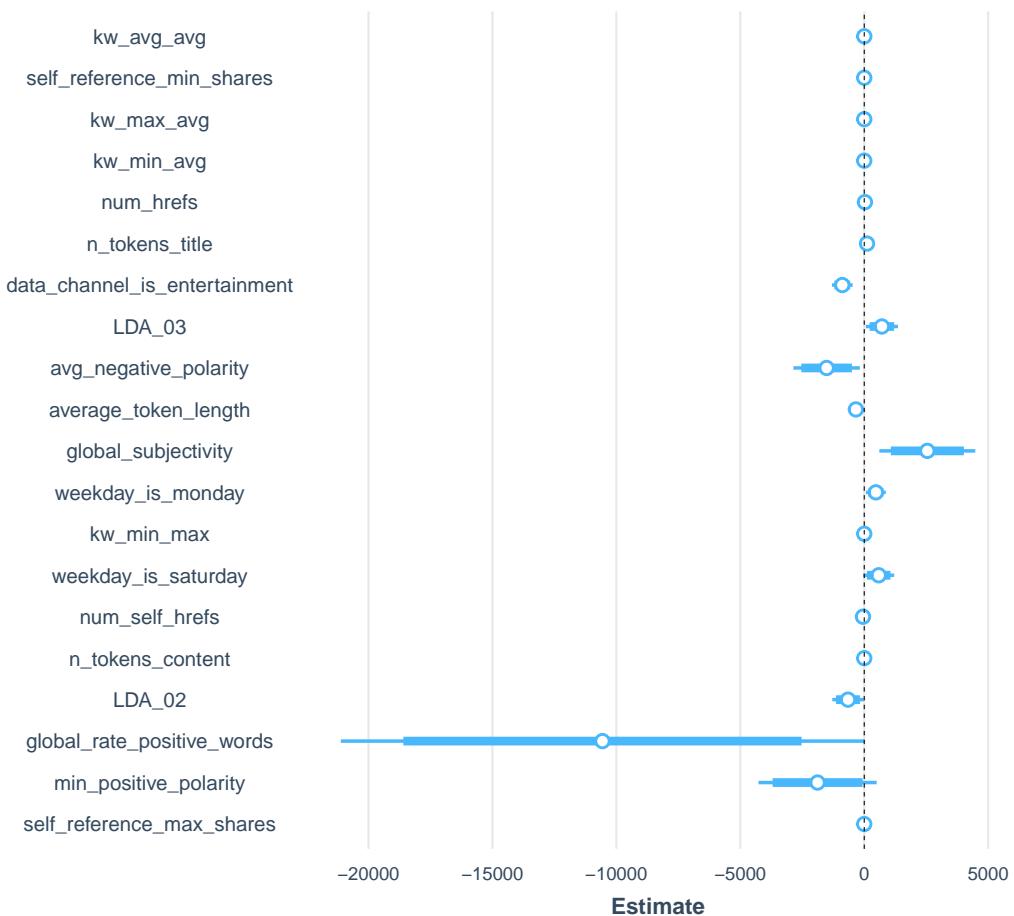
```
plot_summs(modelo3, modelo1, plot.distributions = TRUE,
            rescale.distributions = TRUE)
```



También podemos omitir una variable si no queremos que cree “ruido” al momento de la presentación⁶.

```
plot_summs(modelo3, ci_level = 0.99, inner_ci_level = .95,
            omit.coefs = c("timedelta", "(Intercept)"))
```

⁶Por ejemplo si no se quieren mostrar variables no accionables o quitar el intercepto.



7.8.1 Comentarios Finales

En este capítulo hemos seguido el proceso paso a paso para responder una pregunta de negocio empleando un modelo de regresión para hacer analítica diagnóstica . Adicionalmente, discutimos cómo encontrar la variable más importante y cómo visualizar los resultados. Antes de continuar, es importante resaltar que no se han chequeado los supuestos del modelo de regresión, y por tanto no podemos estar seguros que el método de MCO empleado nos provee estimadores MELI.

En los siguientes capítulos estudiaremos cómo constatar si los supuestos se cumplen y en caso que estos no se cumplan, cómo solucionar el problema. En el Capítulo 10 retomaremos este caso de negocio y constataremos si se cumplen los supuestos del modelo.

Es importante recalcar que aún no podemos sacar conclusiones finales para tomar decisiones, pues no estamos seguros si el modelo que estimamos es bueno (cumple los supuestos del Teorema de Gauss-Markov). En el Capítulo 10 retomaremos esta pregunta de negocio.

Parte III

Problemas econométricos

8 . 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.
- Decidir si los problemas generados por la multicolinealidad se deben o no solucionar.
- Solucionar, de ser necesario, el problema de multicolinealidad empleando R.

8.1 Introducción

Como lo habíamos discutido en capítulos anteriores, si el modelo de regresión múltiple cumple con los supuestos que se resumen en el recuadro abajo, 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.

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, $\varepsilon_i \sim i.i.d (0, \sigma^2)$ o en forma matricial $\varepsilon_{n \times 1} \sim (0_{n \times 1}, \sigma^2 I_n)$

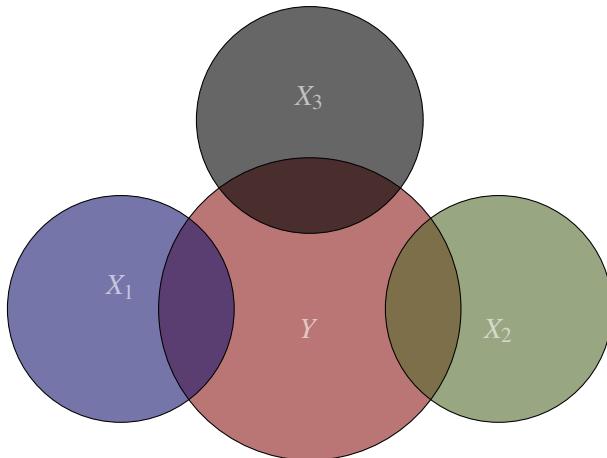
Ahora veamos qué ocurre si no se cumple una parte del supuesto 2: Las X_2, X_3, \dots, X_k son fijas y linealmente independientes. En especial, que las variables explicativas (Xs) no sean linealmente independientes entre sí. Es importante anotar que la violación de la otra parte del supuesto no tiene grandes implicaciones sobre el resultado que los estimadores MCO sean MELI. En el Anexo al final de este capítulo se presenta la demostración de la insesgadez y eficiencia (Ver sección 8.7) de este estimador si las variables explicativas son estocásticas.

En este capítulo nos concentraremos en la violación del supuesto de independencia lineal de las variables explicativas. Este problema que puede presentar un conjunto de datos se conoce con el nombre de multicolinealidad o colinealidad.

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). La Figura 8.1 representa estos supuestos. Los círculos representan las variables y sus intersecciones la relación entre ellas. Podemos observar que existe una relación entre la variable explicada y las independientes¹, pero a su vez existe independencia entre estas últimas².

¹Esta relación se representa por el área en común entre el círculo que representa la variable explicativa y la dependiente.

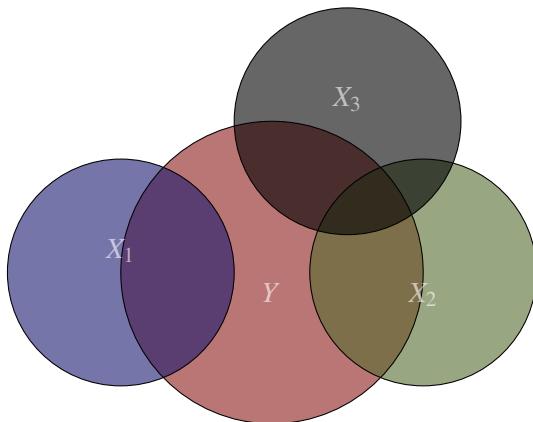
²No existe área en común entre las variables explicativas.

Figura 8.1. El supuesto de no multicolinealidad

Fuente: Elaboración propia

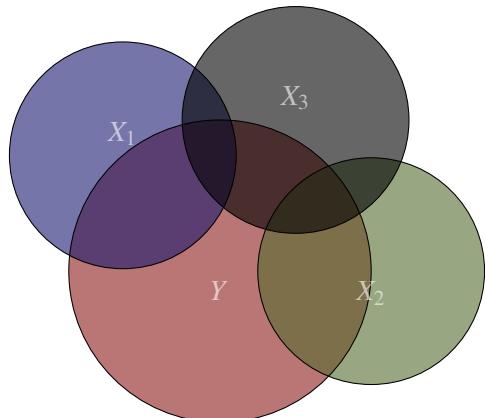
El problema de multicolinealidad aparece cuando tenemos algún tipo de relación lineal entre las variables independientes o entre un subconjunto de ellas.

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, si 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 . La Figura 8.2 representa esta posibilidad donde el supuesto no se cumple porque existe una relación entre dos variables.

Figura 8.2. Multicolinealidad entre dos variables del modelo

Fuente: Elaboración propia

También podría darse el caso en el que todas las variables independientes o un grupo de ellas se encuentren relacionadas. En la Figura 8.3 se presenta un ejemplo donde todas las variables explicativas comparten la información contenida en cada una de ellas.

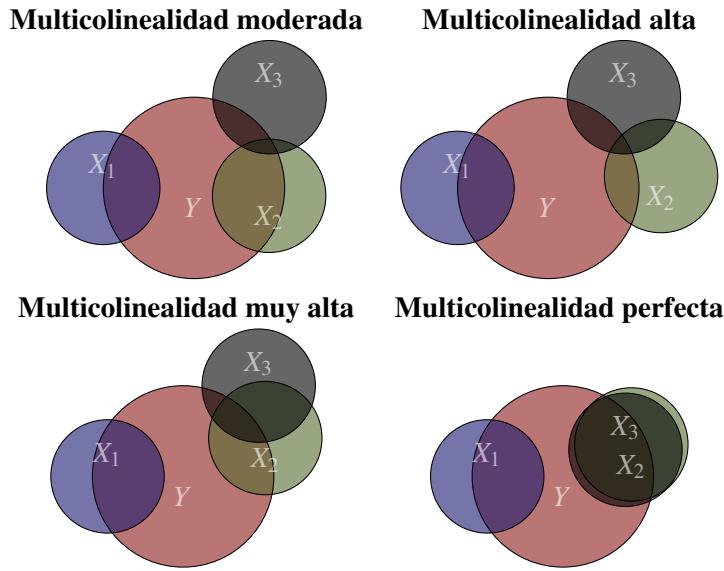
Figura 8.3. Multicolinealidad entre tres variables del modelo

Fuente: Elaboración propia

Todos los casos anteriores son ejemplos de multicolinealidad. En la siguiente sección se describen los diferentes grados de este problema. Posteriormente se discute como detectar la existencia de este problema y como solucionar el problema. Finalmente se presenta una aplicación en R.

8.2 Los diferentes grados de multicolinealidad

En general, cuando hay cierto grado de relación lineal entre las variables independientes decimos que existe multicolinealidad (o colinealidad). En la práctica, tendremos diferentes grados de multicolinealidad. Por ejemplo en la Figura 8.4 se presentan los cuatro posibles grados o tipos de multicolinealidad.

Figura 8.4. Grados de multicolinealidad

Fuente: Elaboración propia

A continuación discutiremos las implicaciones de la multicolinealidad, primero la multicolinealidad perfecta y sus efectos.

8.2.1 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 (hai) y las calorías consumidas. Veamos qué sucede si empleamos el siguiente modelo:

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

donde cd_i y cs_i corresponden a las calorías consumidas por día y calorías consumidas por semana por el individuo i , respectivamente.

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 $7 \times cd_i = cs_i$ y por lo tanto las X's no son linealmente independientes entre sí.

Matricialmente este hecho se puede representar de la siguiente manera:

$$\mathbf{y} = \begin{bmatrix} kg_1 \\ kg_2 \\ \vdots \\ kg_n \end{bmatrix}_{n \times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & ha_1 & cd_1 & 7cd_1 \\ 1 & ha_2 & cd_2 & 7cd_2 \\ 1 & ha_3 & cd_3 & 7cd_3 \\ 1 & ha_4 & cd_4 & 7cd_4 \\ 1 & ha_5 & cd_5 & 7cd_5 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{n \times 4} \quad (8.1)$$

Nota que la cuarta columna de la matriz \mathbf{X} es una combinación lineal del vector columna 3 (ver la sección 14.7 para una discusión del concepto de combinación lineal de vectores). Cuando existe multicolinealidad perfecta, una columna es combinación lineal de otra y otras columnas.

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 existirá³. En este caso, si queremos eliminar la multicolinealidad perfecta entre cd_i y cs_i , debemos eliminar cualquiera de las dos variables. No importa cuál se elimine, 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 los X 's no son linealmente independientes y esto implicará que el estimador de MCO no existirá. Si un modelo presenta multicolinealidad perfecta, entonces la función **lm()** elimina automáticamente una de las variables involucradas en el problema.

Intuitivamente, el problema de multicolinealidad perfecta implica que la información contenida en una variable es redundante pues esta 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:

$$w_i = \lambda_1 + \lambda_2 (E_i) + \lambda_3 (C_i) + \mu_i \quad (8.2)$$

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:

³En el Capítulo 14 puedes encontrar un repaso de los conceptos de álgebra matricial para entender esta argumentación.

$$D_{1i} = \begin{cases} 1 & \text{si } i \in \text{Sec. Primario} \\ 0 & \text{o.w.} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{si } i \in \text{Sec. Secundario} \\ 0 & \text{o.w.} \end{cases}$$

$$D_{3i} = \begin{cases} 1 & \text{si } i \in \text{Sec. Terciario} \\ 0 & \text{o.w.} \end{cases},$$

entonces, nuestro modelo se convierte en:⁴

$$w_i = \lambda_1 + \lambda_2 (E_i) + \lambda_3 (C_i) + \lambda_4 D_{1i} + \lambda_5 D_{2i} + \lambda_6 D_{3i} + \mu_i.$$

Si analizamos la expresión matricial de este modelo, 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, 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.

⁴Por simplicidad solo consideraremos cambios en el intercepto.

En general cuando usamos variables dummy tenemos que tener cuidado si existen j posibilidades diferentes entonces usamos $j - 1$ variables dummy, o j variables dummy y quitamos el intercepto, así evitamos la multicolinealidad perfecta. Esto permite evitar la “trampa de las variables dummy”.

8.2.2 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 correlacionadas 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 matriz⁵ $(\mathbf{X}^T \mathbf{X})^{-1}$, 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,⁶ y así se tendrá 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 $F - Global$ serán relativamente grandes ya que estos dependen del SSR .

Así, los síntomas más comunes de la multicolinealidad no perfecta⁷ se pueden resumir de la siguiente manera:

1. t calculados bajos acompañados de $F - Global$ y R^2 altos
2. Sensibilidad de los β 's estimados a cambios pequeños en la muestra⁸
3. Sensibilidad de los β 's a la inclusión o exclusión de regresores⁹

No obstante la multicolinealidad no perfecta provoca estos síntomas, en muchos casos este problema es ignorado por los científicos de datos. Una 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 buenas predicciones (se está haciendo analítica predictiva), 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.

⁵ $\text{Var}[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$.

⁶ Recordemos que $t_c = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$.

⁷ Antes de continuar es importante resaltar que la multicolinealidad no perfecta no es un problema del modelo que se estima sino de la muestra con la que se cuenta. Es decir, se puede estimar el mismo modelo, pero con una muestra diferentes y no tener multicolinealidad. Por eso al problema de multicolinealidad se le clasifica como un problema de los datos.

⁸ Esto ocurre porque la matriz $(\mathbf{X}^T \mathbf{X})^{-1}$ cambiaría mucho con incluir o eliminar una fila de \mathbf{X} .

⁹ Esto ocurre porque $(\mathbf{X}^T \mathbf{X})^{-1}$ cambiaría mucho con incluir o eliminar una columna de \mathbf{X} .

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:

- Descartar una variable de las que presentan el problema.
- Transformar las variables relacionadas, de tal forma que el modelo involucre razones de las variables y no las variables en sus niveles.
- 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 científico de datos 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 este problema, en muchos casos no se realiza algún procedimiento para resolverla, pues la multicolinealidad no perfecta no afecta las propiedades MELI de los MCO.¹⁰ 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 ésta es la opción que se adopta para “lidiar” con la multicolinealidad, entonces 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 predicciones, entonces el problema de multicolinealidad no es muy importante, siempre y cuando la relación entre las variables explicativas se mantenga para la muestra que se desea predecir.

8.3 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 perfecta. A continuación se describen tres de las pruebas más utilizadas para este fin¹¹.

¹⁰En la sección 2.4, 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 $X^T X$.

¹¹Formalmente, las pruebas disponibles para detectar la multicolinealidad son mas unas métricas que sirven como indicador de la presencia del problema. Lo que veremos a continuación no son pruebas en el sentido estricto al no involucrar un estadístico de prueba y una comparación con un valor de una distribución o un valor p asociado a la decisión.

8.3.1 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 cómo 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 cuántas 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

$$VIF_j = \frac{1}{1 - R_j^2} \quad (8.3)$$

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 $Var[\hat{\beta}_j]$ que puede atribuirse al hecho que esa variable no es ortogonal a las otras variables del modelo.

Algunos autores como Hair y col. (2014) argumentan que si el VIF_j excede 3.0, entonces se considera que existe un problema de multicolinealidad. Otros autores como Sheather (2009) 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 Kutner, M. H.; Nachtsheim, C. J.; Neter (2004)). .

8.3.2 Prueba de Belsley, Kuh y Welsh (1980)

Esta prueba diseñada por Belsley y col. (1980) también es conocida con el nombre de prueba Kappa. Esta prueba se basa en los valores propios¹² de la matriz $\mathbf{X}^T \mathbf{X}$. Ellos demostraron que empleando el valor propio máximo y mínimo de esta matriz es posible detectar la multicolinealidad. La prueba se construye de la siguiente manera:

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_k}} \quad (8.4)$$

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.

8.4 Soluciones de la multicolinealidad (¡Si se necesitan!)

¹²Para la discusión de cómo se calculan los valores propios de una matriz puedes ver la sección 14.10.

8.4.1 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 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 sesgados¹³, 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 .

Cómo 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

$$\mathbf{b}(l) = (\mathbf{X}^T \mathbf{X} + l\mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (8.5)$$

Esto implica que

$$\mathbf{b}(l) = (\mathbf{I} + l(\mathbf{X}^T \mathbf{X})^{-1})^{-1} \hat{\beta}, \quad (8.6)$$

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 el propuesto por Cule y De Iorio (2012). l también se puede escoger graficando los coeficientes estimados en función de l y se selecciona l más pequeño que produce estimadores estables.

8.4.2 Componentes principales

El análisis de componentes principales (PCA) es un procedimiento matemático que sirve para resumir muchas variables en menos variables y que no requiere supuestos. Se emplea cuando se desea incluir el máximo de información posible contenida en un conjunto de variables en un número inferior de variables, las cuales reciben el nombre de componentes principales. Por esto el PCA es conocido como un método de reducción de variables o de reducción de dimensionalidad. Una de sus limitaciones es que solo aplica a variables cuantitativas.

Material multimedia: PCA

Escanea el siguiente código o visita el siguiente enlace para ver un video sobre la técnica PCA.

¹³En el Anexo de este capítulo, sección 8.7, se presenta una demostración del sesgo que presenta este estimador.



Enlace: https://youtu.be/-EQFB_iiqd4

Es decir, que lo que se quiere obtener es una matriz W , de p filas con k columnas, de tal forma que proyectada sobre la matriz X , con n filas y p columnas, permita obtener una matriz Y , con n filas y k columnas, que contenga el máximo de información posible y un menor número de variables:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} \quad (8.7)$$

Para un mayor discusión de la técnica de PCA se puede consultar Alonso (2020). Esta aproximación implica los siguientes pasos:

- **Paso 1:** Estandarizar las variables cuantitativas y explicativas contenidas en la matriz \mathbf{X} . Es decir, a cada observación de cada variable restarle su respectiva media, y dividirla por su respectiva desviación estándar.
- **Paso 2:** Calcular la matriz de varianzas y covarianzas, $S_X = \frac{1}{n-1} ((\mathbf{X} - \bar{\mathbf{x}})(\mathbf{X} - \bar{\mathbf{x}})^T)$
- **Paso 3:** Calcular los vectores propios y valores propios de la matriz de covarianzas y varianzas. El número de vectores propios y valores propios es igual al número de variables explicativas seleccionadas para el análisis.
- **Paso 4:** Escoger el número de componentes principales k , en este paso es donde sucede la reducción de dimensionalidad de los datos. Para esto organizamos los vectores propios de mayor a menor en función de su valor propio y seleccionamos aquellos que explican una mayor proporción de la varianza total (dividimos la varianza explicada por cada componente por la suma de la varianza total).
- **Paso 5:** Construir la matriz de proyección W a partir de los k vectores propios seleccionados.
- **Paso 6:** Transformar el conjunto de datos original X a través de W para obtener Y .
- **Paso 7:** Correr el modelo de regresión inicial empleando el conjunto de datos transformado.

8.4.3 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* más 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 solución es muy fácil de automatizar como lo discutiremos en la siguiente sección.

8.5 Práctica en R

El objetivo de este ejercicio es aplicar las tres pruebas de multicolinealidad anteriormente descritas y de ser necesario solucionar dicho problema. En este caso el gerente del hospital de nivel tres “SALUD PARA LOS COLOMBIANOS”, ubicada en la ciudad de Bogotá, se encuentra al frente de una reestructuración de los salarios que recibe su equipo de trabajo con el objetivo de lograr equidad laboral con enfoque de género. Se presume que existen brechas salariales entre hombres y mujeres¹⁴. Los datos se encuentran disponibles en el archivo *DatosMultiColinealidad.csv*.

La pregunta de negocio es si existe o no una brecha salarial entre hombres y mujeres. Uno de los métodos más utilizados para determinar el efecto de la discriminación en la brecha salarial es la propuesta por Oaxaca (1973), quienes sugieren una técnica para medir la diferencia que tiene sobre el ingreso, características como el capital humano entre géneros. El modelo a estimar es el siguiente:

$$\ln(ih_i) = \beta_0 + \beta_1 yedu_i + \beta_2 exp_i + \beta_3 exp_i^2 + \beta_4 D_i + \beta_5 Dexp_i + \beta_6 Dexp_i^2 + \varepsilon_i \quad (8.8)$$

donde

$$D_i = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{o.w.} \end{cases}$$

Además, $\ln(ih_i)$ 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 . 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** leído y no es necesaria crearla para estimar el modelo. Empleemos la función **lm()** del paquete básico de R para estimar el modelo descrito en 8.8.

La notación $I()$ en la fórmula implica que R hará la operación presentada dentro del paréntesis y se agregará como una variable explicativa. Los resultados de la estimación del modelo se reportan en el

¹⁴Los datos y la aproximación de esta sección corresponden a Bernat (2004).

Cuadro 8.1. Antes de entrar a calcular los estadísticos para determinar la presencia de multicolinealidad, es importante anotar que los síntomas no están presentes, pero esto no implica la ausencia de multicolinealidad. Recuerden que antes de analizar cualquier resultado es importante estar seguros que no existen problemas de multicolinealidad.

Cuadro 8.1: Modelo de brecha salarial estimado por MCO

<i>Dependent variable:</i>	
	Lnih Modelo original
yedu	0.131*** (0.004)
exp	0.034*** (0.005)
I(exp^2)	-0.0003*** (0.0001)
sexomujer	-0.117 (0.084)
exp:sexomujer	-0.004 (0.008)
I(exp^2):sexomujer	0.0001 (0.0002)
Constant	5.856*** (0.075)
<hr/>	
Observations	1,415
R ²	0.454
Adjusted R ²	0.452
Residual Std. Error	0.574 (df = 1408)
F Statistic	195.348*** (df = 6; 1408)

Note: *p<0.1; **p<0.05; ***p<0.01

8.5.1 Pruebas de multicolinealidad

VIF

Para calcular el *VIF* emplearemos el paquete *car* (Fox y Weisberg, 2019) . Este paquete tiene la función **vif()** cuyo único argumento es un objeto de clase **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*, *exp²*, *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 claramente presentes.

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()** del paquete básico de R que permite obtener la matriz \mathbf{X} . Esta función solo tiene como argumento un objeto de clase **lm** con el modelo estimado. Los valores propios de una matriz se pueden encontrar empleando la función **eigen()** .

```
# matriz X
XTX <- model.matrix(res1)
# se calculan los valores propios
e <- eigen(t(XTX) %*% XTX)
# se muestran los valores propios
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

# se crea el valor propio mas grande
lambda.1 <- max(e$val)
# se crea el valor propio mas pequeño
lambda.k <- min(e$val)
# se calcula kappa
```

```

kappa <- sqrt(lambda.1/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 un alto grado de multicolinealidad. Esto probablemente se debe a la 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 que se emplea para calcular las brechas salariales necesita incluir en la especificación ambas variables por razones bien fundamentadas en Oaxaca (1973) y sería incorrecto eliminar alguna de las dos variables.

Ahora tu puedes proceder a determinar si existe o no diferencias salariales entre las mujeres y los hombres para esta muestra.

8.5.2 Solución del problema removiendo variables con alto VIF

Ahora supongamos que si se deseara solucionar el problema de multicolinealidad. Nota que en este caso en específico esto no parece una buena opción, pero de todas maneras procederemos a resolver el problema solo para ejemplificar la técnica.

Creemos una función que permite eliminar de manera automática e iterativa las variables cuyos respectivos coeficientes tengan un *VIF* superior a un umbral determinado (u). La función **remueve.VIF.grande()** se presenta a continuación, sigue 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 = "")))
  # creando el nuevo modelo con nueva fórmula
  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 de clase **lm** y el segundo el umbral para el *VIF*.

No obstante no es necesario solucionar el problema de multicolinealidad en el caso del modelo estudiado. A continuación se muestra un ejemplo de cómo emplear la función **remueve.VIF.grande()**.

```

res2 <- remueve.VIF.grande(res1, 15)
summary(res2)

##
## Call:
## lm(formula = myForm, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.11988 -0.29420 -0.00166  0.30334  2.63580
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.874e+00  6.593e-02  89.090 < 2e-16 ***
## yedu                      1.307e-01  3.979e-03  32.846 < 2e-16 ***
## exp                       3.204e-02  3.918e-03   8.178 6.39e-16 ***
## I(exp^2)                  -2.616e-04  7.688e-05  -3.402 0.000687 ***
## sexomujer                 -1.539e-01  3.977e-02  -3.871 0.000114 ***
## I(exp^2):sexomujer        -2.025e-05  4.902e-05  -0.413 0.679642
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```

## Residual standard error: 0.5742 on 1409 degrees of freedom
## Multiple R-squared:  0.4542, Adjusted R-squared:  0.4522
## F-statistic: 234.5 on 5 and 1409 DF,  p-value: < 2.2e-16

vif(res2)

##          yedu         exp      I(exp^2)        sexo
## 1.248258  10.134731  10.214214  1.696079
## I(exp^2):sexo
## 2.584824

```

8.5.3 Solución del problema empleando la regresión de Ridge

Como se mencionó anteriormente, no estamos frente a un problema de multicolinealidad perfecta o casi perfecta, y no se está violando el segundo supuesto del teorema de Gauss Markov, por lo que no deberíamos emplear la regresión de Ridge, ya que ésta sólo se utiliza como un recurso de última instancia. Sin embargo, si quisieramos emplearla podríamos utilizar la función la función **linearRidge()** del paquete *ridge* (Cule y col., 2021) . Los argumentos más importantes de esta función son

linearRidge(formula,data,lambda,scaling)

donde:

- **formula:** formula del modelo de regresión.
- **data:** objeto de clase **data.frame** que contiene los datos
- **lambda:** forma en la que se va a hallar el λ . El valor por defecto es automático, es decir se encuentra de manera automática el parámetro de reducción de sesgo λ , si esto no se desea se puede colocar un vector con los valores de λ que se desean probar.
- **scaling:** el valor por defecto es **corrform** que ajusta los valores de las variables explicativas de tal forma que la matriz de correlaciones sea unitaria en la diagonal.

Un ejemplo de cómo emplear esta función se presenta a continuación.

```

##
## Call:
## linearRidge(formula = Lnih ~ yedu + exp + I(exp^2) + sexo + sexo *
##               exp + sexo * I(exp^2), data = datos)
##
##
## Coefficients:
##                               Estimate Scaled estimate
## (Intercept)           6.233e+00            NA

```

```
## yedu          1.131e-01    1.823e+01
## exp           1.431e-02    6.676e+00
## I(exp^2)      1.340e-05    3.198e-01
## sexomujer    -1.539e-01   -2.893e+00
## exp:sexomujer 1.403e-03    6.810e-01
## I(exp^2):sexomujer -4.618e-05   -8.699e-01
##
##             Std. Error (scaled) t value (scaled) Pr(>|t|)
## (Intercept)          NA          NA          NA
## yedu              5.604e-01    32.525 < 2e-16
## exp               6.188e-01    10.788 < 2e-16
## I(exp^2)          6.242e-01    0.512   0.608
## sexomujer         6.105e-01    4.739  2.15e-06
## exp:sexomujer    5.195e-01    1.311   0.190
## I(exp^2):sexomujer 6.117e-01    1.422   0.155
##
## (Intercept)
## yedu          ***
## exp           ***
## I(exp^2)
## sexomujer    ***
## exp:sexomujer
## I(exp^2):sexomujer
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge parameter: 0.118604, chosen automatically, computed using 2 PCs
##
## Degrees of freedom: model 3.89 , variance 3.152 , residual 4.627
##
## lambda 0.118604 chosen automatically using 2 PCs
##
##             Estimate (scaled) Std. Error (scaled)
## yedu          18.2284        0.5604
## exp            6.6762        0.6188
## I(exp^2)       0.3198        0.6242
## sexomujer     -2.8932        0.6105
## exp:sexomujer 0.6810        0.5195
## I(exp^2):sexomujer -0.8699        0.6117
##
##             t value (scaled) Pr(>|t|)
## yedu          32.525 < 2e-16 ***
## exp           10.788 < 2e-16 ***
## I(exp^2)       0.512   0.608
## sexomujer     4.739  2.15e-06 ***
## exp:sexomujer 1.311   0.190
## I(exp^2):sexomujer 1.422   0.155
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8.5.4 Solución del problema empleando componentes principales

Si quisieramos emplear PCA para solucionar un problema de multicolinealidad podríamos emplear la función **lm(princomp)** del paquete base de R. Ésta permite calcular los PCA empleando la matriz de varianzas y covarianzas o empleando la matriz de correlaciones. Vamos a emplear la matriz de correlaciones para el cálculo. Para un mayor discusión de la técnica de PCA se puede consultar Alonso (2020)

Los argumentos más importantes de esta función son

princomp(formula,data, cor=TRUE)

donde:

- **formula:** formula con las variables independientes numéricas, no incluir la variable dependiente
- **data:** data frame que contiene los datos
- **cor:** indica si se va a calcular los PC empleando la matriz de varianzas y covarianzas o empleando la matriz de correlaciones

Los componentes principales se calculan de manera sencilla con el siguiente código

```
res.pca <- princomp(~yedu + exp + I(exp^2), data = datos)

summary(res.pca)

## Importance of components:
##                 Comp.1        Comp.2        Comp.3
## Standard deviation   634.6930542 4.026690e+00 3.775598e+00
## Proportion of Variance 0.9999244 4.024724e-05 3.538434e-05
## Cumulative Proportion 0.9999244 9.999646e-01 1.000000e+00
```

Ahora decidamos el número de componentes. Para esto podemos emplear un gráfico de codo (*elbow graph* también conocido como *scree plot*). Para esto es necesario mirar la proporción de la varianza de todos los datos explicada por cada componente principal.

El primer paso, es calcular la desviación estándar de cada componente y la varianza.

```
std_dev <- res.pca$sdev
pr_var <- std_dev^2
pr_var

##          Comp.1        Comp.2        Comp.3
## 402835.27299    16.21424    14.25514
```

Nuestro objetivo es encontrar los componentes que explican la mayor varianza. Recuerden que queremos conservar tanta información como sea posible utilizando estos componentes. Entonces, cuanto mayor sea la varianza explicada, mayor será la información contenida en esos componentes.

Para calcular la proporción de la varianza explicada por cada componente, simplemente dividimos la varianza por la suma de la varianza total. Esto resulta en:

```
prop_varex <- pr_var/sum(pr_var)
prop_varex

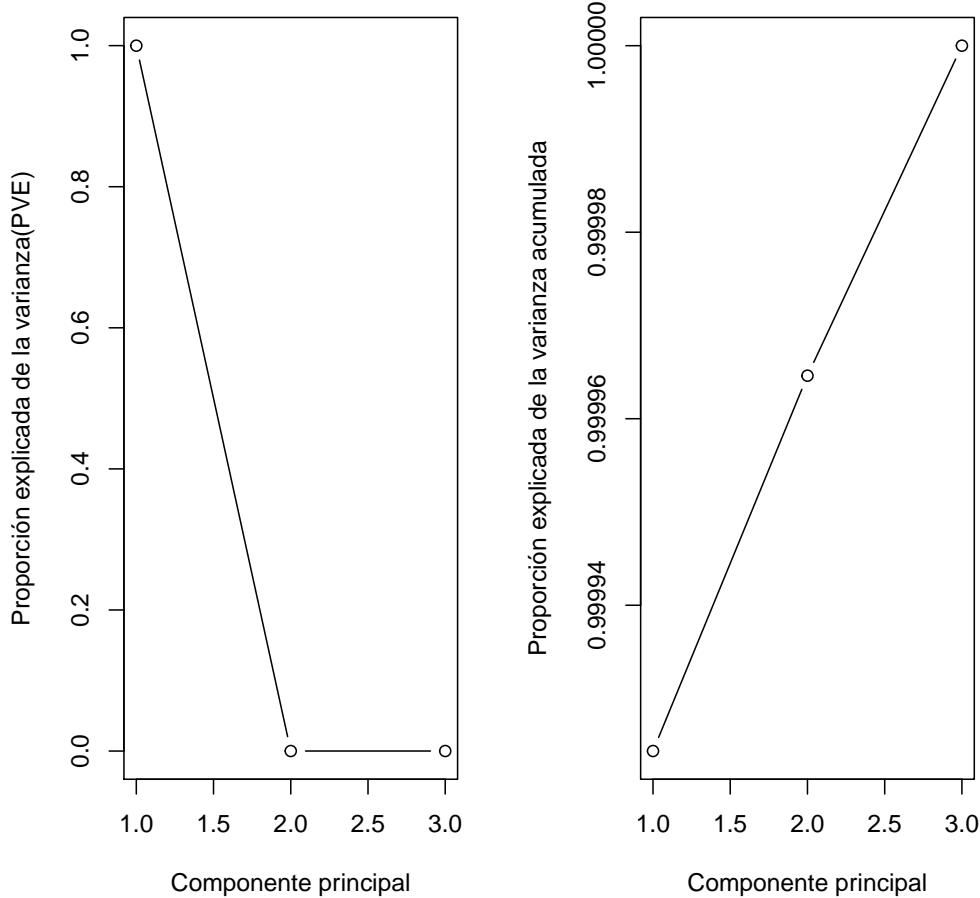
##          Comp.1          Comp.2          Comp.3
## 9.999244e-01 4.024724e-05 3.538434e-05
```

Noten que el primer componente principal explica el 99.9% de la varianza. Aunque parece evidente que debemos quedarnos con el primer componente, ¿cómo decidimos cuántos componentes deberíamos seleccionar para la etapa de modelado cuando no es tan clara la respuesta?

La respuesta a esta pregunta es proporcionada por un *elbow graph* y un gráfico acumulado de varianza.

```
par(mfrow=c(1,2))
plot(prop_varex, xlab = "Componente principal",
     ylab = "Proporción explicada de la varianza(PVE)",
     type = "b")

plot(cumsum(prop_varex), xlab = "Componente principal",
     ylab = "Proporción explicada de la varianza acumulada",
     type = "b")
```



En este tipo de gráficos se busca el “codo” (elbow). En este caso no se observa el codo, es decir la información se resume en el primer componente. Finalmente se extraen la matriz proyectada (valores de los componentes principales que remplazarán a las variables originales), los cuales vamos a emplear en estimar de nuevo el modelo inicial.

```
head(res.pca$scores)
PCA.res <- res.pca$scores[, 1]
head(PCA.res)

res2 <- lm(Lnih ~ sexo + PCA.res, datos)
summary(res2)
```

Noten que en este caso es imposible interpretar el coeficiente asociado al componente principal.

Ahora, podemos reproducir el ejercicio con la descomposición a partir de la matriz de varianzas y covarianza.

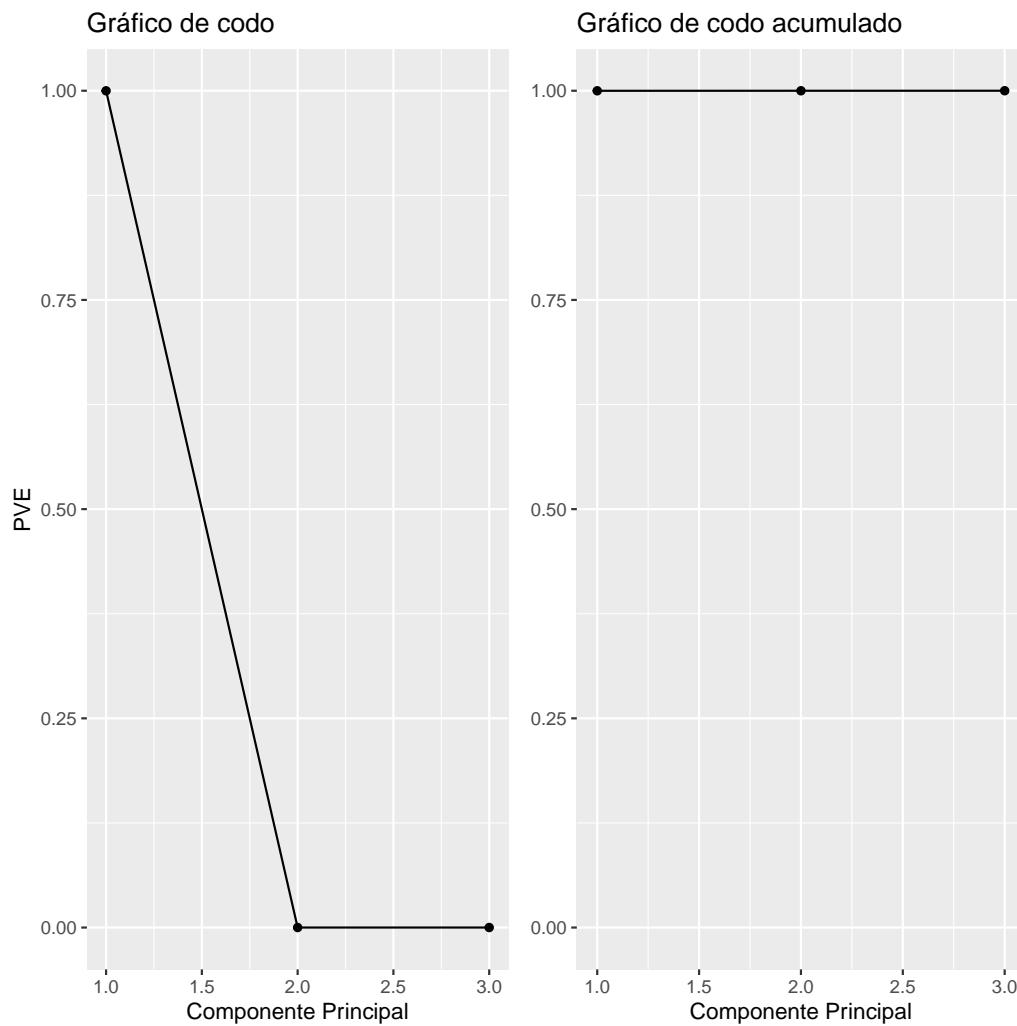
```
res.pca2 <- princomp(~yedu + exp + I(exp^2), data = datos, cor = FALSE)
std_dev2 <- res.pca2$sdev
pr_var2 <- std_dev2^2
pr_var2
prop_varex2 <- pr_var2/sum(pr_var2)
prop_varex2
```

Y para variar realicemos los gráficos de codo con el paquete *ggplot2* (Wickham, 2016) y el paquete *gridExtra* (Auguie, 2017).

```
library(gridExtra)
PVEplot <- qplot(c(1:3), prop_varex2) +
  geom_line() +
  xlab("Componente Principal") +
  ylab("PVE") +
  ggtitle("Gráfico de codo") +
  ylim(0, 1)

cumPVE <- qplot(c(1:3), cumsum(prop_varex2)) +
  geom_line() +
  xlab("Componente Principal") +
  ylab(NULL) +
  ggtitle("Gráfico de codo acumulado") +
  ylim(0, 1)

grid.arrange(PVEplot, cumPVE, ncol = 2)
```



Noten que el resultado es muy similar. Deberíamos emplear 1 componente. Los demás pasos del procedimiento los puede realizar por su cuenta.

8.6 Comentarios finales

En este Capítulo discutimos el problema de multicolinealidad. El problema de multicolinealidad perfecta no permite estimar el modelo, pero es muy fácil de solucionar. Por otro lado, vimos que el problema de multicolinealidad no perfecta es un problema que se manifiesta en t calculados bajos acompañados de $F - Global$ y R^2 altos, sensibilidad de los β 's estimados a cambios pequeños en la muestra y sensibilidad de los β 's a la inclusión o exclusión de regresores.

No obstante la multicolinealidad no perfecta provoca estos síntomas, en muchos casos este problema es ignorado por los científicos de datos. Una 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 buenas predicciones (se está haciendo analítica predictiva), 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.

Ejercicios

8.1 Un científico de datos es contratado por un fabricante de automóviles para estimar un modelo sencillo que le permita predecir cuál debe ser la posición del asiento para un conductor determinado, en otras palabras cuál es la distancia horizontal del punto medio de las caderas desde una ubicación fija en el automóvil. El fabricante desea que sus carros sean cómodos y seguros. Para elaborar la tarea se cuenta con 38 observaciones (la información se encuentra en el archivo *datosmulti.csv*). Las observaciones corresponden a características del conductor como peso, edad, su altura si emplea zapatos o no, su altura sentado, la longitud de muslo, pierna y brazo. Para explicar cuál debe ser la distancia horizontal (D_i en la base de datos *Distancia_centro*) no emplearemos por ahora todas las variables disponibles. Estimemos el siguiente modelo:

$$D_i = \alpha_0 + \alpha_1 edad_i + \alpha_2 peso_i + \alpha_3 alturadescalzo_i + \alpha_4 alturadeszapatos_i + \varepsilon_i \quad (8.9)$$

Estime el modelo y determine si existe o no multicolinealidad por medio de las pruebas estudiadas en este capítulo. Adicionalmente, corrija (de ser posible) el problema de multicolinealidad si es que existe. Explique.



8.7 Anexos

8.7.1 Demostración de la insesgadez del estimador MCO con variables explicativas aleatorias

Recordemos que el estimador MCO ($\hat{\beta}$) esta dado por

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

. Este estimador se puede reescribir de la siguiente manera:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \varepsilon) \\ \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \\ \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \\ \hat{\beta} &= \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon\end{aligned}\tag{8.10}$$

Ahora consideremos el valor esperado para un valor dado de \mathbf{X}
 \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.

8.7.2 Demostración de la eficiencia del estimador MCO con variables explicativas aleatorias

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.

$$\begin{aligned} Var[\hat{\beta} | \mathbf{X}] &= E \left[(\hat{\beta} - E[\hat{\beta}]) (\hat{\beta} - E[\hat{\beta}])^T | \mathbf{X} \right] \\ Var[\hat{\beta} | \mathbf{X}] &= E \left[(\hat{\beta} - \beta) (\hat{\beta} - \beta)^T | \mathbf{X} \right] \\ Var[\hat{\beta} | \mathbf{X}] &= E \left[(\beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon - \beta) (\hat{\beta} - \beta)^T | \mathbf{X} \right] \\ Var[\hat{\beta} | \mathbf{X}] &= E \left[((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon) ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon)^T | \mathbf{X} \right] \\ Var[\hat{\beta} | \mathbf{X}] &= E \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \varepsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} | \mathbf{X} \right] \\ Var[\hat{\beta} | \mathbf{X}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\varepsilon \varepsilon^T | \mathbf{X}] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ Var[\hat{\beta} | \mathbf{X}] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\sigma^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ Var[\hat{\beta} | \mathbf{X}] &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \end{aligned}$$

Ahora, debemos demostrar que dicha varianza es la mínima 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

$$\mathbf{D}\mathbf{y} = \tilde{\beta} - \hat{\beta}$$

y

$$\mathbf{C} = \mathbf{D} + \mathbf{A}$$

Entonces,

$$\begin{aligned} \text{Var} [\tilde{\beta} | \mathbf{X}] &= \text{Var} [(\mathbf{D} + \mathbf{A}) \mathbf{y} | \mathbf{X}] \\ \text{Var} [\tilde{\beta} | \mathbf{X}] &= \text{Var} [(\mathbf{D} + \mathbf{A})(\mathbf{X}\beta + \varepsilon) | \mathbf{X}] \end{aligned}$$

Por lo tanto, tenemos que la varianza del otro estimador es

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

Entonces, recordemos que para que $\tilde{\beta}$ 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

$$\text{Var} [\tilde{\beta} | \mathbf{X}] = \sigma^2 (\mathbf{DD}^T + \mathbf{DX} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}^T + \mathbf{AA}^T)$$

$$\text{Var} [\tilde{\beta} | \mathbf{X}] = \sigma^2 (\mathbf{DD}^T + \mathbf{AA}^T)$$

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

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

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

$$\text{Var} [\tilde{\beta} | \mathbf{X}] = \text{Var} [\hat{\beta} | \mathbf{X}] + \sigma^2 \mathbf{DD}^T$$

$$\mathbf{q}^T \mathbf{DD}^T \mathbf{q} = \mathbf{z}^T \mathbf{z} \geq 0$$

$$\text{Var} [\tilde{\beta} | \mathbf{X}] \geq \text{Var} [\hat{\beta} | \mathbf{X}]$$

Q.E.D. Por lo tanto no es posible obtener un estimador insesgado con una varianza menor, aún en presencia de regresores aleatorios.

8.7.3 Demostración del sesgo del estimador de la regresión de Ridge

Partamos de la definición del estimador de Ridge

$$\mathbf{b}(l) = (\mathbf{X}^T \mathbf{X} + l \mathbf{I})^{-1} \mathbf{X}^T y$$

Ahora calculemos el valor esperado de dicho estimador

$$E [\mathbf{b}(l) | \mathbf{X}] = E [(\mathbf{X}^T \mathbf{X} + l \mathbf{I})^{-1} \mathbf{X}^T y | \mathbf{X}]$$

$$E [\mathbf{b}(l) | \mathbf{X}] = (\mathbf{X}^T \mathbf{X} + l \mathbf{I})^{-1} \mathbf{X}^T E [y | \mathbf{X}]$$

$$E [\mathbf{b}(l) | \mathbf{X}] = (\mathbf{X}^T \mathbf{X} + l \mathbf{I})^{-1} \mathbf{X}^T E [\mathbf{X}\beta + \varepsilon | \mathbf{X}]$$

$$E [\mathbf{b}(l) | \mathbf{X}] = (\mathbf{X}^T \mathbf{X} + l \mathbf{I})^{-1} \mathbf{X}^T \mathbf{X} \beta$$

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, $\text{Var} [\hat{\beta} | \mathbf{X}] > \text{Var} [\mathbf{b}(l) | \mathbf{X}]$

9 . Heteroscedasticidad

Diseñado por Freepik

Objetivos del capítulo

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

- Explicar en sus propias palabras los efectos de la heteroscedasticidad sobre los estimadores MCO.
- 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.

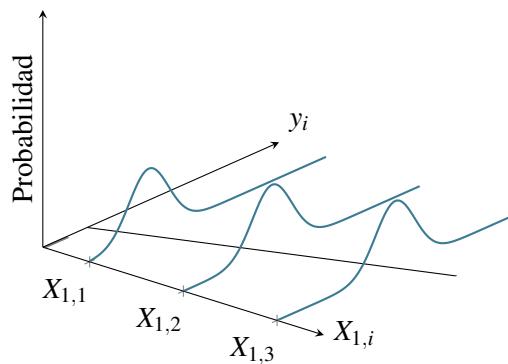
9.1 Introducción

Como lo hemos discutido en capítulos anteriores, si los supuestos del modelo de regresión se cumplen, entonces el Teorema de Gauss-Markov demuestra que los estimadores MCO son MELI (Mejor Estimador Lineal Insesgado). En el Capítulo 8 analizamos las consecuencias de la violación del supuesto de independencia lineal entre variables explicativas en un modelo de regresión múltiple; es decir la parecencia de multicolinealidad. También discutimos cómo dicho problema era un problema de los datos y que en ocasiones no será necesario resolver el problema. En este capítulo nos concentraremos en la violación del supuesto que el término de error tiene varianza constante.

Supuestos del modelo de regresión múltiple

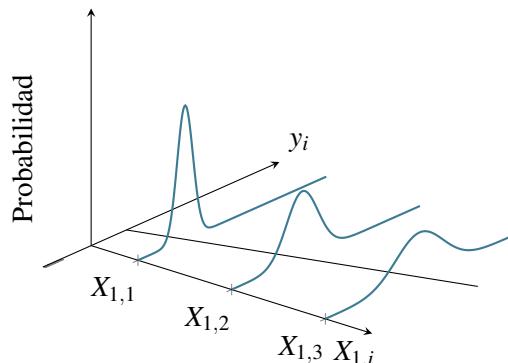
1. Relación lineal entre y y X_2, X_3, \dots, X_k
2. Las X_2, X_3, \dots, X_k son fijas y linealmente independientes (i.e. la matriz X tiene rango completo)
3. el vector de errores ε satisface:
 - Media cero ($E[\varepsilon] = 0$),
 - Varianza constante
 - No autocorrelación
 Es decir, $\varepsilon_i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

Antes de concentrarnos en la violación del supuesto, veamos un poco la intuición detrás del cumplimiento de este supuesto. En la Figura 9.1 se presenta una muestra de observaciones para una variable explicativa (X_1) y una dependiente (y) (puntos grises) en la que se cumple el supuesto de homoscedasticidad. El supuesto de homoscedasticidad implica que la dispersión de las observaciones (puntos) con respecto al modelo de regresión poblacional (línea de regresión) será siempre la misma (Ver Figura 9.1). El tercer eje mide la probabilidad de que ocurra un valor de y dado el valor de X_1 . En este caso esperamos que en promedio los valores de y se encuentren cercanos a la línea de regresión (por el supuesto de que $E[\varepsilon] = 0$) y el supuesto de homoscedasticidad implica que la dispersión con respecto a la linea de regresión siempre es la misma.

Figura 9.1. Muestras con errores homoscedasticidad

Fuente: Elaboración propia

Por otro lado, si el supuesto no se cumple entonces las observaciones de y seguirán estando en promedio al rededor de la linea de regresión (Ver Figura 9.2) gracias al supuesto de que $E [\varepsilon] = 0$. La heteroscedasticidad provoca que la dispersión sea diferente dependiendo de la observación. En este caso en la Figura 9.2 se simuló un problema de heteroscedasticidad que depende del valor de la variable explicativa. La volatilidad es mayor a medida que X_1 se hace mas grande¹.

Figura 9.2. Muestras con errores heteroscedasticidad

Fuente: Elaboración propia

En ocasiones el supuesto de homoscedasticidad o varianza constante del término de error no tiene mucho sentido. Un ejemplo de esto se presenta al analizar el consumo en función del ingreso con

¹Es importante aclarar que para las Figuras 9.1 y 9.2 se ha supuesto que los errores siguen una distribución normal para construir una gráfica más intuitiva. Pero este supuesto de normalidad no es necesario en la práctica solo se ha realizado para efectos de la gráfica

una muestra de hogares (corte transversal). Supongamos que se desea emplear un modelo en que las compras de un producto del hogar i y se cuenta con una base de datos amplia para muchos niveles de ingresos. Las compras entre muchas cosas depende de su nivel de ingresos y otras variables que caracterizan la condición socio demográfica del hogar. Los hogares con ingresos bajos típicamente presentan un comportamiento de compra 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 posible². 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:³

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

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.

²En un anexo al final del capítulo (ver sección 9.5) se presenta una demostración.

³En los anexos al final del capítulo (Ver sección 9.5 se presenta la demostración.

Antes de continuar hagamos un ejercicio similar al realizado en el Capítulo 3 para exemplificar el problema que genera la heteroscedasticidad. Realicemos el siguiente experimento de Monte Carlo. Creemos una muestra de datos de tamaño 50 del siguiente DGP:

$$Y_i = 1 + 7X2_i + \varepsilon_i$$

donde ε_i puede ser homoscedástico y heteroscedástico. Nuestro experimento tendrá como finalidad ver cuál es el efecto sobre los estimadores MCO y su desviación estándar (errores estándar) de la heteroscedasticidad. Adicionalmente, crearemos una segunda variable explicativa $X3_i$ que no es significativa y la incluiremos en nuestro análisis.

En otras palabras, generaremos una muestra aleatoria del DGP con errores homoscedásticos y otra con errores heteroscedásticos y estimaremos los correspondientes coeficientes. Y repetiremos este ejercicio 10 mil veces. Al final tendremos una muestra de 10000 para cada coeficiente estimado para cada uno de los escenarios: con y sin heteroscedasticidad. Esto nos permitirá determinar si el promedio de los coeficientes coincide o no con el valor poblacional (es insensado) y comparar las varianzas del estimador MCO en presencia o no de este problema.

Empecemos nuestro experimento fijando una semilla para los números aleatorios y creando los objetos donde guardaremos los resultados de la simulaciones. Así mismo definamos los parámetros del tamaño de la muestra y el número de repeticiones.

```
# se fija una semilla
set.seed(1234557)
# se crean objetos para guardar resultados creamos un objeto
# nulo para guardar los coeficientes de X2
X2.estcoef <- NULL
# creamos un objeto nulo para guardar los coeficientes de X3
X3.estcoef <- NULL
# se fija el tamaño de cada muestra en 50
n = 50

# se fija el número de repeticiones en 10mil
N = 10000
```

Ahora empleemos un *loop* para replicar nuestro ejercicio 100 mil veces.

```
for (i in 1:N) {
  # creación de variable explicativa
  X2 <- matrix(rnorm(n), n, 1)
  # creación de variable no significativa
  X3 <- matrix(rnorm(n), n, 1)
  # error homoscedástico
  err <- rnorm(n)
  # error heteroscedástico
  herr <- (X2^2) * err
```

```

# variable explicativa sin heteroscedasticidad
y1 <- 1 + X2 * 7 + err
# variable explicativa con heteroscedasticidad
y2 <- 1 + X2 * 7 + herr
# estimación del modelo con homoscedasticidad
res.homo <- summary(lm(y1 ~ X2 + X3))

# guardamos los resultados regresión homoscedasticidad
X2.cf.homo <- res.homo$coefficients[2, 1]
X3.cf.homo <- res.homo$coefficients[3, 1]

# estimación del modelo con heteroscedasticidad
res.hetero <- summary(lm(y2 ~ X2 + X3))
# guardamos los resultados regresión heteroscedasticidad
X2.cf.hetero <- res.hetero$coefficients[2, 1]
X3.cf.hetero <- res.hetero$coefficients[3, 1]
# guardando los resultados de la iteración i
X2.estcoef <- rbind(X2.estcoef, cbind(X2.cf.homo, X2.cf.hetero))
X3.estcoef <- rbind(X3.estcoef, cbind(X3.cf.homo, X3.cf.hetero))

}

```

Ahora, comparemos los resultados para el caso de un error homoscedástico y uno homoscedástico para la pendiente que acompaña a X_2 . Primero veamos el promedio de los 10 mil coeficientes estimados en los dos escenarios

```

round(apply(X2.estcoef, 2, mean), 3)

##   X2.cf.homo X2.cf.hetero
##       6.999      6.997

```

Se puede observar que en ambos escenarios en promedio los coeficientes estimados están muy cercanos al valor poblacional de 7. Ahora miremos que ocurre con la error estándar (desviación estándar de los coeficientes estimados)

```

round(apply(X2.estcoef, 2, sd), 2)

##   X2.cf.homo X2.cf.hetero
##       0.15      0.52

```

Noten que el error estándar es menor en el escenario de homoscedasticidad que en el de heteroscedasticidad. Estos resultados muestran que para el estimador de la pendiente de X_2 :

- Los estimadores MCO siguen siendo insesgados en presencia de heteroscedasticidad.
- El error estándar de los coeficientes es más grande en presencia de heteroscedasticidad.

Ahora, comparemos los resultados para para la pendiente que acompaña a $X3_i$.

```
round(apply(X3.estcoef, 2, mean), 2)

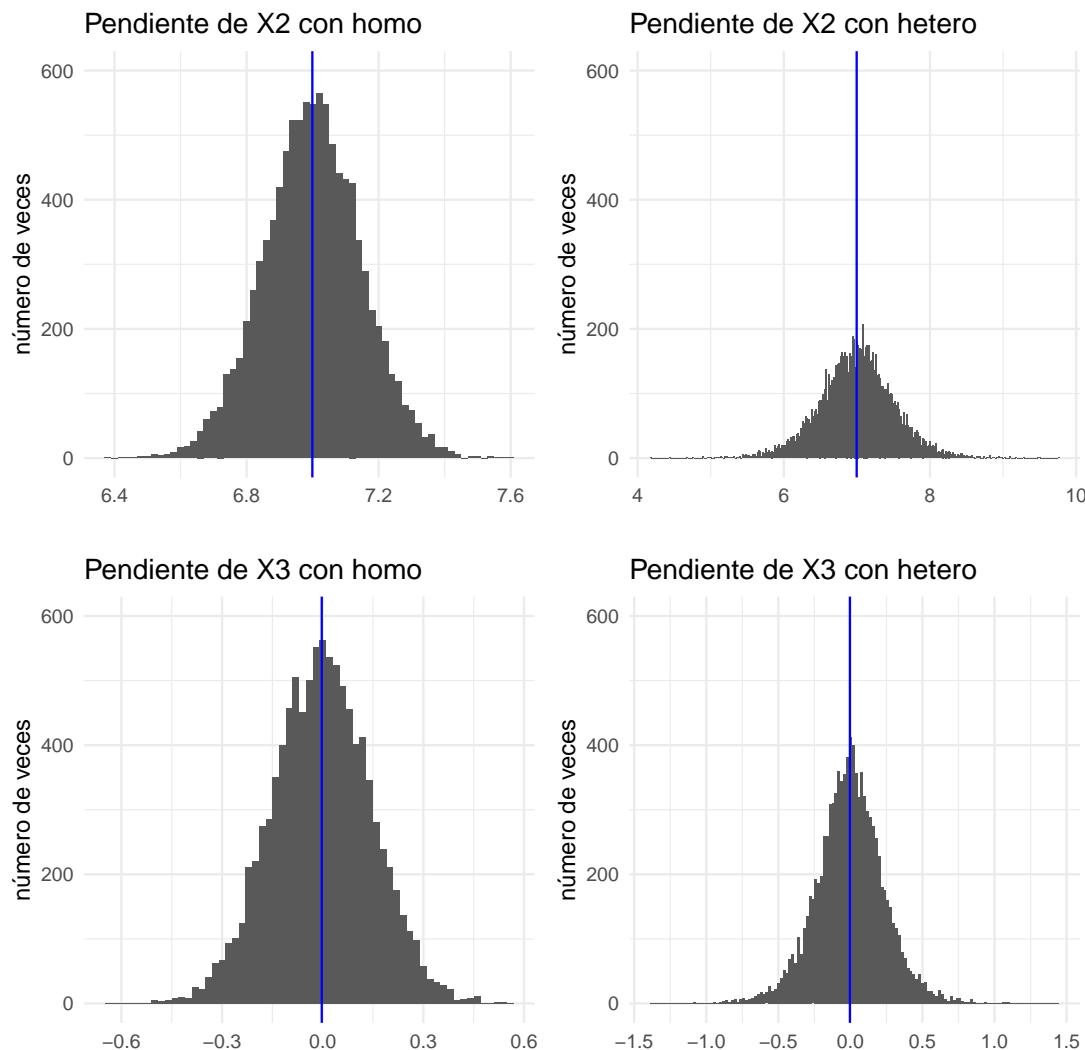
##   X3.cf.homo X3.cf.hetero
##           0          0

round(apply(X3.estcoef, 2, sd), 2)

##   X3.cf.homo X3.cf.hetero
##       0.15      0.24
```

Los resultados son similares para el estimador de la pendiente de la variable que no debería estar en el modelo. Finalmente, puedes inspeccionar los histogramas de la distribución empírica de los coeficientes estimados en los dos escenarios.

Figura 9.3. Distribución muestral de los estimadores bajo homocedasticidad y heteroscedasticidad (en azul se presenta la media de los coeficientes estimados)



Fuente: Elaboración propia

En este capítulo estudiaremos como detectar si se viola el supuesto de varianza constante del error y de encontrar el problema cómo resolverlo.

9.2 Pruebas para la detección de heteroscedasticidad

En general, una buena práctica cuando se estiman modelos económicos 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.

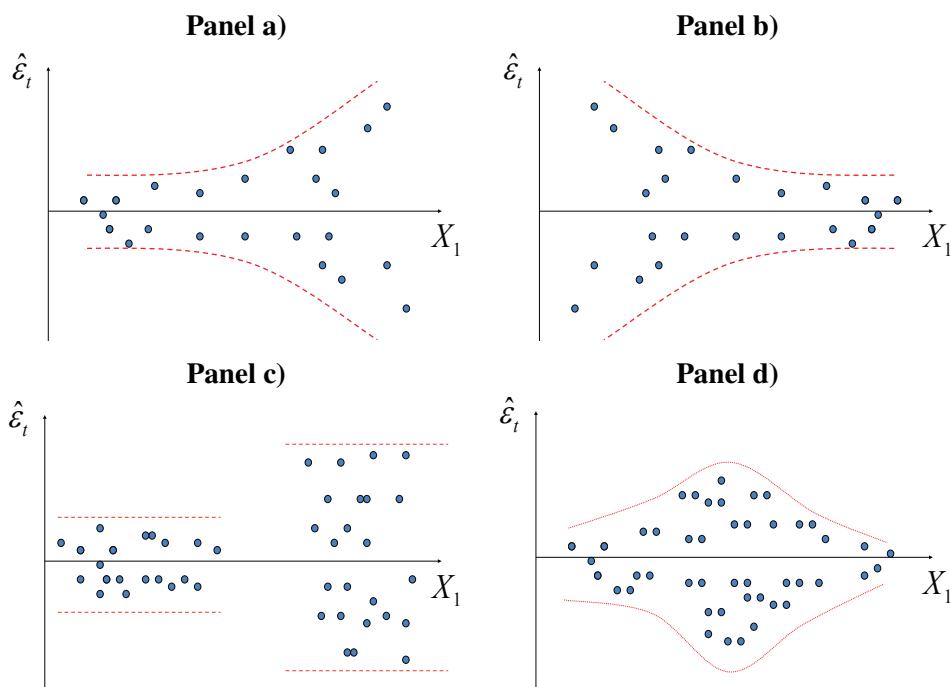
Pero, el científico de datos se puede enfrentar a la necesidad de estimar muchos modelos o tener muchas iteraciones entre modelos y tener que automatizar el proceso. En esos escenarios el análisis gráfico no es una opción. No obstante, para ganar un poco de la intuición del problema.

Regresando al análisis gráfico, dado que la heteroscedasticidad implica una variabilidad no constante del error, entonces lo más adecuado será graficar el vector de errores ($\hat{\epsilon}$). Lastimosamente, este vector no es observable, pero la mejor aproximación que tenemos para conocer ese vector es el error estimado ($\hat{\epsilon}$). Así, para intuir gráficamente la presencia de heteroscedasticidad se deben visualizar los residuos del modelo estimado. Las visualizaciones más empleadas son gráficos de dispersión de:

1. Los errores estimados versus las observaciones ($\hat{\epsilon}_i$ vs $i = 1, 2, \dots, n$)
2. Los errores estimados versus cada una de las variables explicativas ($\hat{\epsilon}_i$ vs cada una de las columnas de bX)

Estas visualizaciones de los errores estimados se emplean para explorar algún tipo de regularidad. En la Figura 9.4 se presentan tipos de patrones que se pueden observar en los residuales estimados que pueden sugerir presencia de heteroscedasticidad.

Figura 9.4. Posibles patrones de comportamiento de los residuales que sugieren heteroscedasticidad



En el panel a) de la Figura 9.4 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 tarea de identificar el problema con gráficos y con pruebas.

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 mucha intuición sobre este problema. En esas situaciones, el análisis gráfico puede ser inútil.

A continuación consideraremos dos pruebas de heteroscedasticidad, cada una diseñada para detectar diferentes formas de heteroscedasticidad. En los dos 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.

9.2.1 Prueba de Breusch-Pagan

Esta prueba, permite la posibilidad de determinar si más de una variable causa el problema de heteroscedasticidad.

Breusch y Pagan (1979) 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 científico de datos 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. Calcule⁴ $\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n}$.
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 $BP = \frac{SSR}{2}$.

Breusch y Pagan (1979) 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 Koenker (1981)). Koenker (1981) 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.

9.2.2 Prueba de White

White (1980) 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:

$$\begin{aligned} H_0 : \sigma_i^2 &= \sigma^2; \quad \forall i \\ H_A : \text{No } H_0 \end{aligned} \tag{9.1}$$

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: $\hat{\varepsilon}_i^2 = \gamma + \sum_{m=1}^k \sum_{j=1}^k \delta_s X_{mi} X_{ji} + \mu_i$. 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 será:⁵ $\hat{\varepsilon}_i^2 = \gamma + \delta_1 X_{2i} + \delta_2 X_{3i} + \delta_3 X_{2i}^2 + \delta_4 X_{3i}^2 + \delta_5 X_{2i} X_{3i} + \vartheta_i$.
3. Calcule el R^2 de la regresión auxiliar.
4. Calcule el estadístico de White $W_a = n \times R^2$.

White (1980) demostró que su estadístico W_a sigue una distribución Chi-cuadrado con un número

⁴Noten que esto corresponde al estimador de la varianza del error del Método de Máxima Verosimilitud.

⁵El último término se conoce con el nombre de término cruzado o términos de interacción. 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. Lo recomendable es incluir los términos de interacción siempre que sea posible, dado que así tiene mas poder la prueba White (1980).

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 $\chi^2_{g,\alpha}$.

9.3 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)⁶ que hace parte de la 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.

La segunda opción implica solucionar los síntomas de la heteroscedasticidad, tratando de estimar de manera consistente⁷ 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.

9.3.1 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, White (1980) ideó una forma de corregir el estimador que tiene el problema y no la muestra.

En especial, White (1980) mostró que es posible aún encontrar un estimador apropiado para la matriz de varianzas y covarianzas de los β 's obtenidos por MCO.⁸ Recordemos que en presencia de heteroscedasticidad, la varianza de los coeficientes tiene la siguiente estructura:

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

White (1980) sugiere usar el siguiente estimador consistente para la matriz de varianzas y covarianzas de los β 's obtenidos por MCO

$$\text{Est.Var} [\hat{\beta}] = n(\mathbf{X}^T \mathbf{X})^{-1} S_0 (\mathbf{X}^T \mathbf{X})^{-1}$$

⁶En los anexos (ver sección 9.5) se presenta una breve introducción a este método de solucionar el problema de heteroscedasticidad.

⁷Consistencia en estadística es la propiedad de un estimador de ser insesgado cuando la muestra es grande.

⁸Recuerden que el estimador MCO del vector β sigue siendo insesgado, el problema se presenta en el estimador de la matriz de varianzas y covarianzas.

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, White (1980) demostró que ese estimador es consistente; es decir, es sesgado en muestras pequeñas (y por tanto no funciona) pero insesgado en muestras grandes (y por tanto funciona bien en muestras grandes). Por esta razón este estimador (y otros similares) es conocido como estimador H.C. (heteroskedasticity consistent). Es importante mencionar, que esto no hace que al emplear esta solución se obtengan un estimador MELI. La varianza del estimador MCO sigue siendo más grande que por ejemplo los del estimador GLS.

Por otro lado, Davidson, Russell and MacKinnon (1993) 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, Russell and MacKinnon (1993) proponen: $\omega_i = \frac{n}{n-k} e_i^2$ (HC1).
- Davidson, Russell and 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 \mathbf{P} gorro.
- Davidson, Russell and 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, Russell and 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.

Por ejemplo, la prueba de Wald para una restricción de la forma $\mathbf{R}\beta = \mathbf{q}$ será:

$$W = (\mathbf{R}\hat{\beta})^T \left[\mathbf{R} \left(\text{Est.Asy.Var}[\hat{\beta}] \right) \mathbf{R}^T \right]^{-1} (\mathbf{R}\hat{\beta})$$

9.4 Práctica en R

Continuaremos con la pregunta de negocio que discutimos en el Capítulo anterior (ver sección 8.5). Recordemos que en este caso el gerente del hospital de nivel tres “SALUD PARA LOS COLOMBIANOS”, ubicada en la ciudad de Bogotá, tenía la siguiente pregunta de negocio: ¿existe o no una brecha salarial entre hombres y mujeres? Los datos se encuentran disponibles en el archivo *DatosMultiColinealidad.csv*⁹.

El modelo modelo que estimamos fue el siguiente:

$$\ln(ih_i) = \beta_0 + \beta_1 yedu_i + \beta_2 exp_i + \beta_3 exp_i^2 + \beta_4 D_i + \beta_5 Dexp_i + \beta_6 Dexp_i^2 + \varepsilon_i \quad (9.2)$$

donde

$$D_i = \begin{cases} 1 & \text{si el individuo } i \text{ es mujer} \\ 0 & \text{o.w.} \end{cases}$$

Además, $\ln(ih_i)$ 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 .

Adicionalmente, recuerden que habíamos encontrado un problema de multicolinealidad que decidimos era mejor no resolver. Ahora nuestro tarea será aplicar las diferentes pruebas de heteroscedasticidad y de ser el caso aplicar una solución.

Carga los datos y estima el modelo de nuevo guardando los resultados en un objeto `res1`.

9.4.1 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 tu puedes efectuar todos los otros gráficos). Recuerden que es muy probable que este análisis gráfico no sea útil con las cantidades de variables explicativas que trabajan los científicos de datos.

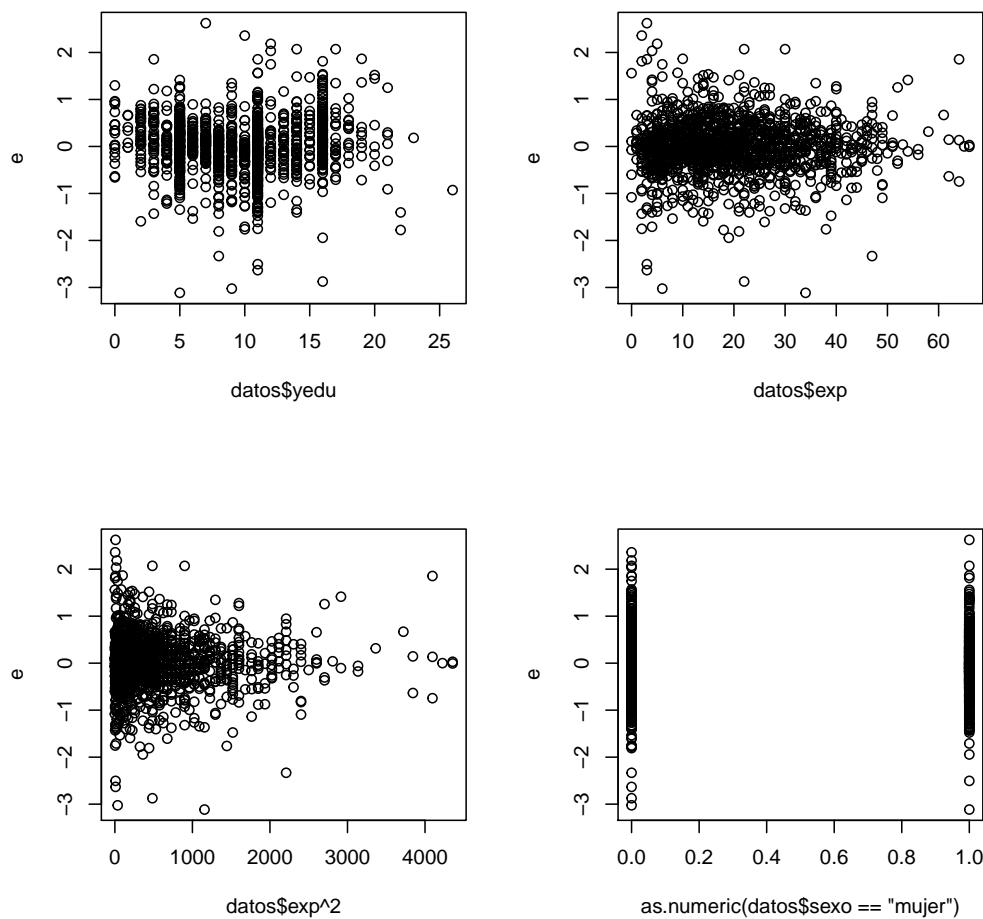
Extraigamos los residuales del objeto LM con la función `resid()` del paquete central de R. Esta función solo requiere como argumento un objeto de clase **lm**. Además, guardemos los residuales en el objeto `e`.

```
e <- resid(res1)
```

⁹Los datos y la aproximación de esta sección corresponden a Bernat (2004).

El correspondiente modelo estimado se reportó en el capítulo anterior (Ver Cuadro 8.1). Pero noten que esos resultados no son interesantes hasta que estemos seguros que no existe un problema de heteroscedasticidad. Ahora grafiquemos los residuos versus las variables explicativas del modelo.

```
par(mfrow = c(2, 2))
plot(datos$yedu, e)
plot(datos$exp, e)
plot(datos$exp^2, e)
plot(as.numeric(datos$sexo == "mujer"), 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 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.

9.4.2 Pruebas de heteroscedasticidad

Prueba de Breusch-Pagan

Hay muchas funciones que permiten calcular esta prueba en R. Por ejemplo, existe la función **bptest()** del paquete *lmtest* (Zeileis y Hothorn, 2002) 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* (Hebbali, 2020). 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

bptest(formula, varformula = NULL, studentize = TRUE, data = list())

donde:

- **formula:** fórmula del modelo de regresión o un objeto de clase **lm**.
- **varformula:** una fórmula que incluya las posibles variables explicativas de la varianza. Por defecto se toman las mismas variables explicativas que en el modelo de regresión principal.
- **studentize:** permite determinar si se empleará la versión studentizada de la prueba o no. El valor por defecto es **studentize = "TRUE"**. Es decir, si no se especifica este argumento se calculará la versión studentizada.
- **data:** Si se emplea una fórmula para la varianza, este argumento es obligatorio y tendrá que contener el **data.frame** donde se encuentran las variables que se emplean para explicar la varianza.

De esta manera la prueba de Breush-Pagan cuya hipótesis alterna es que todas las variables del modelo causan el problema de heteroscedasticidad se puede calcular con el siguiente código:

```
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. Es decir, la varianza es función de todas las variables. En otras palabras, existe un problema de heteroscedasticidad.

Si queremos ver si es la variable *exp* la que causa el problema de heteroscedasticidad, podemos hacerlo fácilmente, ,

```
bptest(res1, ~exp, studentize = FALSE, data = datos)

##
## 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 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.

ols_test_breusch_pagan(model,rhs,multiple,p.adju)

donde:

- **model:** El objeto de clase **lm** al que se le realizará la prueba
- **rhs:** es un valor lógico, 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).
- **multiple:** es igual a un valor lógico y permite realizar todas las pruebas en las que la varianza depende de cada una de las variables explicativas y de todas ella (**multiple = TRUE**).
- **p.adj:** permite especificar qué tipo de corrección estadística aplicarle al valor p para tener en cuenta que se están realizando múltiples comparaciones al tiempo. Las opciones para este argumento son “none” para ninguna corrección, “bonferroni” para la corrección de Bonferroni, “sidak” para la de Sidak y “holm” para la de Holm.

Por ejemplo, el siguiente código realiza pruebas de Breusch-Pagan 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 : Lnih
## 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 homoscedasticidad 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. Las pruebas mas comunes de normalidad son Shapiro-Wilk (Shapiro y Francia, 1972), Kolmogorov-Smirnov (Kolmogorov, 1933), Cramer-von Mises (Cramér, 1928) y Anderson-Darling(Anderson y Darling, 1952). Una forma rápida de realizar estas pruebas de normalidad es emplear la función **ols_test_normality()** del paquete **olsrr**. La hipótesis nula de estas pruebas de normalidad es que la variable fue generada por una distribución normal, versus la alterna que es otra distribución.

```

ols_test_normality(res1)

## -----
##      Test        Statistic     pvalue
## -----
##      Shapiro-Wilk  0.9583     0.0000
##      Kolmogorov-Smirnov 0.0636     0.0000
##      Cramer-von Mises 174.9537    0.0000

```

```
## 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 la versión studentizada de la prueba propuesta por Koenker (1981) para el caso en que los residuos no sigan una distribución normal.

```
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.

Prueba de White

Por último, realicemos la prueba de White¹⁰ para contrastar la hipótesis nula de no heteroscedasticidad versus la alterna de heteroscedasticidad se puede calcular empleando la función **white_lm()** del paquete *skedastic* (Farrar, 2020). Esta función típicamente incluye los siguientes argumentos:

```
white_lm(mainlm, interactions = FALSE)
```

donde:

- **mainlm**: es un objeto de clase *lm* que contiene la regresión sobre la cual se realizará la prueba.
- **interactions**: es un valor lógico, si es igual a TRUE se incluyen los términos de interacción; si interactions = FALSE entonces no se incluyen los términos de interacción. Esta última es la opción por defecto. En general es recomendable incluir los términos de interacción (esto comúnmente no tiene mucho sentido).

```
library(skedastic)

white_lm(res1, interactions = TRUE)
```

¹⁰Waldman (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, en principio se puede emplear la función **bptest()** para esta prueba. Pero en la mayoría de las esta aproximación puede ser engorrosa cuando el modelo incluye muchas variables explicativas y por tanto los productos cruzados pueden ser muchos para incluirlos en la fórmula.

```
## # A tibble: 1 x 5
##   statistic p.value parameter method      alternative
##       <dbl>     <dbl>    <dbl> <chr>      <chr>
## 1      37.9    0.0793     27 White's Test greater
```

En este caso con un 95 % de confianza no se puede rechazar la hipótesis nula de homoscedasticidad.

9.4.3 Solución al problema de heteroscedasticidad con HC

Sumando los resultados de las pruebas anteriores, se puede concluir que existe evidencia de la presencia de heteroscedasticidad. Por tanto, es pertinente solucionar el problema para hacer inferencia.

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* (Zeileis, 2004) y la función **vcovHC()**. Esta función requiere dos argumentos: el objeto de clase **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, Russell and MacKinnon (1993) y sugerida por Cribari-Neto (2004)).

Para este caso, tenemos que la corrección HC3 se puede estimar de la siguiente manera:

```
library(sandwich)
# HC3 Davidson y MacKinnon (1993)
vcovHC(res1)

##          (Intercept)      yedu      exp
## (Intercept) 8.317882e-03 -2.271974e-04 -4.947456e-04
## yedu        -2.271974e-04  2.059261e-05  1.825652e-06
## exp         -4.947456e-04  1.825652e-06  4.626380e-05
## I(exp^2)     7.713180e-06  1.795725e-08 -8.473137e-07
## sexomujer   -5.423997e-03 -3.514125e-05  4.723621e-04
## exp:sexomujer 4.554557e-04  1.742774e-06 -4.609217e-05
## I(exp^2):sexomujer -7.815056e-06 -8.851575e-09  8.506639e-07
##                      I(exp^2)  sexomujer exp:sexomujer
## (Intercept) 7.713180e-06 -5.423997e-03  4.554557e-04
## yedu        1.795725e-08 -3.514125e-05  1.742774e-06
## exp         -8.473137e-07  4.723621e-04 -4.609217e-05
## I(exp^2)     1.718700e-08 -7.955854e-06  8.527277e-07
## sexomujer   -7.955854e-06  9.727628e-03 -8.059566e-04
## exp:sexomujer 8.527277e-07 -8.059566e-04  8.079147e-05
## I(exp^2):sexomujer -1.721951e-08  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 más bien los respectivos t individuales y sus correspondientes valores p, podemos emplear la función **coeftest()** del paquete *lmtest* para realizar las pruebas individuales. La función **coeftest()** necesita dos argumentos:

$$\text{coeftest}(x, \text{vcov.} = \text{NULL})$$

donde:

- **x**: es el objeto al que se le realizará la prueba, generalmente de clase *lm*.
- **vcov.**: la matriz de varianzas y covarianzas que se quiera emplear. Si no se especifica una matriz de varianzas y covarianzas, entonces se empleará la de los MCO.

Por ejemplo, para este caso podemos ver como individualmente, los coeficientes de las variables asociadas a la dummy de sexo no son significativos individualmente.

```
# HC3 Davidson y MacKinnon (1993)
coeftest(res1, vcov = (vcovHC(res1)))

##
## t test of coefficients:
##
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.8557e+00  9.1202e-02 64.2056 < 2.2e-16 ***
## yedu            1.3064e-01  4.5379e-03 28.7882 < 2.2e-16 ***
## exp             3.3948e-02  6.8017e-03  4.9911 6.754e-07 ***
## I(exp^2)       -2.9550e-04  1.3110e-04 -2.2540   0.02435 *
## sexomujer      -1.1666e-01  9.8629e-02 -1.1828   0.23707
## exp:sexomujer -3.9387e-03  8.9884e-03 -0.4382   0.66131
## I(exp^2):sexomujer 5.3015e-05  1.7510e-04  0.3028   0.76211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estos resultados comparados con la estimación MCO se reportan en el Cuadro 9.1.

Cuadro 9.1: Modelo estimado por MCO y corrección HC

	<i>Dependent variable:</i>	
	Lnih	
	MCO	HC3
	(1)	(2)
yedu	0.131*** (0.004)	0.131*** (0.005)
exp	0.034*** (0.005)	0.034*** (0.007)
I(exp^2)	-0.0003*** (0.0001)	-0.0003** (0.0001)
sexomujer	-0.117 (0.084)	-0.117 (0.099)
exp:sexomujer	-0.004 (0.008)	-0.004 (0.009)
I(exp^2):sexomujer	0.0001 (0.0002)	0.0001 (0.0002)
Constant	5.856*** (0.075)	5.856*** (0.091)
F Statistic (df = 6; 1408)	195.348***	147.36***
Observations	1,415	1,415
R ²	0.454	0.454
Adjusted R ²	0.452	0.452
Residual Std. Error (df = 1408)	0.574	0.574

Note:

*p<0.1; **p<0.05; ***p<0.01

A manera de ejemplo se muestra las otras posibles correcciones.

```
# White (1980)
coefest(res1, vcov = (vcovHC(res1, "HCO")))
##
```

```
## t test of coefficients:  
##  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.8557e+00 9.0230e-02 64.8974 < 2.2e-16 ***  
## yedu        1.3064e-01 4.5104e-03 28.9636 < 2.2e-16 ***  
## exp         3.3948e-02 6.6308e-03  5.1197 3.483e-07 ***  
## I(exp^2)    -2.9550e-04 1.2605e-04 -2.3443   0.0192 *  
## sexomujer   -1.1666e-01 9.7408e-02 -1.1977   0.2312  
## exp:sexomujer -3.9387e-03 8.7934e-03 -0.4479   0.6543  
## I(exp^2):sexomujer 5.3015e-05 1.6940e-04  0.3130   0.7544  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
# Davidson y MacKinnon (1993)  
coeftest(res1, vcov = (vcovHC(res1, "HC1")))  
  
##  
## t test of coefficients:  
##  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.8557e+00 9.0454e-02 64.7367 < 2.2e-16 ***  
## yedu        1.3064e-01 4.5216e-03 28.8919 < 2.2e-16 ***  
## exp         3.3948e-02 6.6473e-03  5.1070 3.721e-07 ***  
## I(exp^2)    -2.9550e-04 1.2637e-04 -2.3385   0.0195 *  
## sexomujer   -1.1666e-01 9.7650e-02 -1.1947   0.2324  
## exp:sexomujer -3.9387e-03 8.8152e-03 -0.4468   0.6551  
## I(exp^2):sexomujer 5.3015e-05 1.6982e-04  0.3122   0.7549  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
# Davidson y MacKinnon (1993)  
coeftest(res1, vcov = (vcovHC(res1, "HC2")))  
  
##  
## t test of coefficients:  
##  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.8557e+00 9.0708e-02 64.5556 < 2.2e-16 ***  
## yedu        1.3064e-01 4.5241e-03 28.8759 < 2.2e-16 ***  
## exp         3.3948e-02 6.7141e-03  5.0562 4.839e-07 ***  
## I(exp^2)    -2.9550e-04 1.2851e-04 -2.2994   0.02163 *  
## sexomujer   -1.1666e-01 9.8006e-02 -1.1904   0.23411  
## exp:sexomujer -3.9387e-03 8.8884e-03 -0.4431   0.65774  
## I(exp^2):sexomujer 5.3015e-05 1.7217e-04  0.3079   0.75819  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Cribari-Neto (2004)
coeftest(res1, vcov = (vcovHC(res1, "HC4")))

##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.8557e+00 9.1595e-02 63.9306 < 2.2e-16 ***
## yedu        1.3064e-01 4.5321e-03 28.8250 < 2.2e-16 ***
## exp         3.3948e-02 6.9554e-03  4.8808 1.177e-06 ***
## I(exp^2)    -2.9550e-04 1.3620e-04 -2.1697    0.0302 *
## sexomujer   -1.1666e-01 9.9245e-02 -1.1755    0.2400
## exp:sexomujer -3.9387e-03 9.1539e-03 -0.4303    0.6671
## I(exp^2):sexomujer 5.3015e-05 1.8070e-04  0.2934    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 Monte Carlo 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), datos)

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 que 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, sí existe diferencia en como se trata a los hombres y a las mujeres.

Ejercicios

9.1 Realice un experimento de Monte Carlo similar al que se realiza en la Introducción de este Capítulo. En este caso nos interesa conocer cual es la proporción de veces que se rechaza la hipótesis nula de las pruebas individuales para las dos pendientes con los estimadores MCO sin heteroscedasticidad, los estimadores MCO con heteroscedasticidad y la corrección HC3 en presencia de heteroscedasticidad. Realice el experimento para muestras de tamaño 20, 50, 200 y 1000. ¿Qué puedes concluir?

9.5 Anexos

9.5.1 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,

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

Donde $E[\boldsymbol{\varepsilon}_t] = 0$, $Var[\boldsymbol{\varepsilon}_t] = \sigma^2_t$ y $E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j] = 0, \forall i \neq j$. Ahora determinemos si $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ sigue siendo insesgado o no. Así,

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

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

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

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

9.5.2 Demostración del 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 ($\widehat{Var}[\hat{\beta}] = s^2(\mathbf{X}^T \mathbf{X})$) 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 Anexo anterior, en este caso tenemos que:

$$Var[\varepsilon] = E[\varepsilon^T \varepsilon] = \Omega = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Ahora podemos calcular la varianza de los estimadores MCO. Es decir,

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

Por tanto tendremos que

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

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

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

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 ($\widehat{Var}[\hat{\beta}] = s^2 (\mathbf{X}^T \mathbf{X})^{-1}$) en presencia de heteroscedasticidad se obtendrá un estimador cuyo valor esperado no es igual a la varianza real; es decir, será insesgado.

9.5.3 Solución por mínimos cuadrados ponderados

Si conocemos la varianza de cada uno de los errores es posible utilizar el método de Mínimos Cuadrados Ponderados (MCP) que hace parte de la familia de los Mínimos Cuadrados Generalizados (MCG). La idea de esta solución es muy sencilla, e implica transformar los datos de la muestra de tal forma que el problema desaparezca de la muestra. Por ejemplo, partamos del siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (9.3)$$

Con un vector de errores con las siguientes características: $E[\varepsilon_i] = 0$, $E[\varepsilon_i \varepsilon_j] = 0$, y con una varianza de los errores conocida $Var[\varepsilon_i] = W_i^2 \sigma_i^2$. Si dividimos el modelo por W_i obtenemos:¹¹

$$\frac{Y_i}{W_i} = \beta_1 \frac{1}{W_i} + \beta_2 \frac{X_{2i}}{W_i} + \beta_3 \frac{X_{3i}}{W_i} + \dots + \beta_k \frac{X_{ki}}{W_i} + \frac{\varepsilon_i}{W_i} \quad (9.4)$$

¹¹Los W_i son conocidos como los pesos de ponderación, de ahí el nombre del método

Note que:

$$\text{Var} \left[\frac{\varepsilon_i}{W_i} \right] = \frac{1}{W_i^2} \text{Var} [\varepsilon_i] = \frac{1}{W_i^2} W_i^2 \sigma^2 = \sigma^2$$

Es decir, ahora cada observación tiene varianza constante y por lo tanto el problema de heteroscedasticidad ha sido resuelto, de tal manera que los estimadores MCO para el modelo 9.4 ya son MELI.

El problema práctico de esta aproximación es conocer exactamente la naturaleza de la heteroscedasticidad, por eso en la práctica es aconsejable para los científicos de datos emplear una solución H.C. dados los grandes volúmenes de observaciones que se emplean.

10 . Primer caso de negocio (actualizado)

Diseñado por Freepik

Objetivos del capítulo

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

- Emplear las herramientas estudiadas en los capítulos anteriores para responder una pregunta de negocio que implique analítica diagnóstica
- Presentar los resultados de una regresión de manera gráfica empleando R con solución H.C..
- Determinar cuál variable tiene mas efecto sobre la variable explicativa empleando R en presencia de heteroscedasticidad.

10.1 Introducción

En el Capítulo 7 trabajamos en responder una pregunta de negocio del portal de noticias en Internet Mashable (<https://mashable.com>). La pregunta de negocio del editor era: ¿De qué depende el número de *shares* de un artículo? A esta pregunta venía anexa otra: ¿Existe alguna variable accionable que pueda ser modificada para generar una recomendación a los escritores?

Responder esta pregunta de negocio implicaba realizar analítica diagnóstica y realizamos una búsqueda automática para encontrar modelos candidatos a ser el mejor modelo. Posteriormente los comparamos empleando pruebas de Y finalmente analizamos el mejor modelo para encontrar el mejor modelo. Este análisis lo hicimos sin tener en cuenta que podíamos estar violando supuestos del Teorema de Gauss-Markov. En este capítulo realizaremos las modificaciones necesarias a nuestro análisis para brindar una respuesta estadísticamente correcta a la pregunta de negocio.

Recordemos que para realizar nuestro análisis contamos con una base de datos que contenía los artículos publicados en un periodo de dos años suministrada por Fernandes y col. (2015). La base de datos se encuentra en el archivo *DatosCaso1.csv*. Estos datos son reales y fueron descargados de la siguiente página <https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity>. La base de datos contiene 39644 observaciones y 61 variables que se describen en el Cuadro 7.1.

De esta manera, la base de datos con que trabajaremos corresponde a datos de corte transversal. De inmediato debemos sospechar de un posible problema de heteroscedasticidad. A continuación recrearemos nuevamente todos los pasos seguidos en el Capítulo 7 y se aclararán cuáles pasos deben ser modificados en presencia de multicolinealidad y heteroscedasticidad (¡si es que encontramos que existe alguno o ambos problemas!)

10.2 El plan

Recordemos que nuestra primera tarea fue trazar una ruta analítica para responder la pregunta de negocio.

Ahora, con esta aproximación completa debemos ajustar nuestro plan (en negrilla se resaltan los nuevos pasos) que será el siguiente:

1. Encontrar diferentes modelos candidatos a ser el mejor modelo
2. **Determinar si existe un problema de heteroscedasticidad en los modelos candidatos.**
3. **Limpiar los modelos candidatos de variables no significativas.**
4. Comparar los modelos candidatos para seleccionar un único modelo
5. **Determinar si existe un problema de multicolinealidad.**
6. Identificar la variable más importante para explicar la variable dependiente
7. Generar las recomendaciones
8. Generar visualizaciones de los resultados

Empecemos a ejecutar esa ruta analítica para resolver la pregunta de negocio

10.3 Detección de posibles modelos

Recuerda que la primera variables en la base de datos no es relevante (url), ésta corresponde al enlace del artículo, y por tanto la debemos eliminar. Lo cual implica que tendremos $5,7646075 \times 10^{17}$ posibles modelos.

Lee los datos empleando la función `read.csv()`, guárdalos en el objeto `datos.caso1` y elimina la primera variable que no es relevante.

```
datos.caso1 <- read.csv("../Data/DatosCaso1.csv", sep = ",")  
datos.caso1 <- datos.caso1[, -1]
```

El siguiente paso es encontrar los mejores modelos empleando las estrategias de regresión paso a paso forward, backward y combinada con el AIC, con el valor p y el R^2 ajustado. No obstante, en este caso que es posible que exista heteroscedasticidad (mas adelante lo demostraremos) es mejor no emplear el criterio del valor p, pues sabemos que este criterio depende del estadístico t, el cual a su vez depende del estimador del error estándar. Y sabemos que en presencia de heteroscedasticidad, estos errores estándar no son los mínimos posibles. Es decir, si existe heteroscedasticidad, los errores estándar no son los mas pequeños posibles, y por tanto los t calculados estarán errados y por tanto los valores p.

Por otro lado, el problema de multicolinealidad puede hacer que el R^2 esté inflado y por tanto el R^2 ajustado puede también estar inflado. Pero por ahora no tenemos por que intuir este problema. Trabajaremos con esta métrica, pero teniendo en mente que nuestro resultado podría estar afectado por un problema de multicolinealidad cuando empleemos el R^2 ajustado para seleccionar un candidato a mejor modelo.

Esto nos deja con 6 opciones de algoritmos y criterios de selección que se presentan en el Cuadro 10.1. Mantendremos los mismos nombres de modelos que se emplearon en el Capítulo 7 para evitar confusiones. Los modelos que obtenemos los guardaremos con los nombres que se presentan el Cuadro 7.2.

Cuadro 10.1: Modelos a estimar con los diferentes algoritmos

Nombre del objeto	Algoritmo	Criterio
modelo1	Forward	R^2 ajustado
modelo3	Forward	AIC
modelo4	Backward	R^2 ajustado
modelo6	Backward	AIC
modelo7	Both	R^2 ajustado
modelo9	Both	AIC

Partamos de estimar los modelos lineales con todas las variables potenciales (`max.model`).

```
# modelo con todas las variables
max.modelo <- lm(shares ~ ., data = datos.caso1)
```

Ahora procedamos a encontrar modelos candidatos para los mejores modelos. Empecemos con la estrategia *stepwise Forward*.

10.3.1 Stepwise forward

Empecemos por el modelo 1: algoritmo Forward y criterio de R^2 ajustado.

```
# cargando la libreria library(leaps)
modelo1 <- regsubsets(x = datos.caso1[, 1:59], y = datos.caso1[, 60], nvmax = 60, method = "forward")

## Reordering variables and trying again:

m1 <- summary(modelo1)

index <- 1:length(modelo1$xnames)
n.index <- index[m1$which[which.max(m1$adjr2), ] == 1]
vars.modelo1 <- modelo1$xnames[index[m1$which[which.max(m1$adjr2), ] == 1]]
vars.modelo1 <- vars.modelo1[-1]
# vars.modelo1 <- vars.modelo1[-30]
formula.modelo1 <- as.formula(paste("shares ~ ", paste(vars.modelo1,
collapse = " + "), sep = ""))

modelo1 <- lm(formula.modelo1, data = datos.caso1)
```

Ahora tenemos que constatar si el modelo 1 tiene o no heteroscedasticidad.

Para realizar la prueba de Breusch-Pagan, debemos constatar si el supuesto de normalidad se cumple. Para esto extraigamos los residuales con la función **residuals()** y efectuemos las pruebas de normalidad. En esta ocasión contamos con una muestra muy grande y la función **ols_test_normality()** del paquete *olsrr* no trabaja con muestras mayores a 5mil observaciones. Tenemos otras opciones como la función **ks.test()** del paquete base que permite también realizar las pruebas de Kolmogorov-Smirnov (Kolmogorov, 1933). Otra prueba de normalidad que se puede emplear es la de Jarque y Bera (1987). La función **jarque.bera.test()** del paquete *tseries* (Trapletti y Hornik, 2019).

```
# se extraen los residuales
res.modelo1 <- residuals(modelo1)
# pruebas de normalidad Kolmogorov-Smirnov
ks.test(res.modelo1, y = pnorm)

##
```

```

## One-sample Kolmogorov-Smirnov test
##
## data: res.modelo1
## D = 0.76329, p-value < 2.2e-16
## alternative hypothesis: two-sided

# pruebas de normalidad Kolmogorov-Smirnov
library(tseries)
jarque.bera.test(res.modelo1)

##
## Jarque Bera Test
##
## data: res.modelo1
## X-squared = 5856780436, df = 2, p-value < 2.2e-16

```

Los residuales de este modelo no siguen una distribución normal. Por eso no son confiables los resultados de la prueba de Breusch-Pagan tradicional. Deberíamos entonces emplear la versión studentizada de la prueba propuesta por Koenker (1981)

```

# se carga la librería
library(lmtest)
# prueba de Breusch-Pagan studentizada
bptest(modelo1, studentize = TRUE)

##
## studentized Breusch-Pagan test
##
## data: modelo1
## BP = 77.535, df = 29, p-value = 2.643e-06

```

Con un 99% de confianza podemos rechazar la hipótesis de homoscedasticidad. Y para la prueba de White obtendremos el mismo resultado (¡Inténtalo!).

Esto implica que no es posible hacer inferencia sobre el modelo obtenido anteriormente empleando los estimadores MCO y por tanto eliminar las variables no significativas como lo hicimos en el Capítulo 10. Corrijamos el problema empleando los estimadores HC3 (la propuesta por Davidson, Russell and MacKinnon (1993) y sugerida por Cribari-Neto (2004)).

```

# se carga la librería
library(sandwich)
# HC3 Davidson y MacKinnon (1993)
coeftest(modelo1, vcov = (vcovHC(modelo1)))

```

```

## t test of coefficients:
##
##                               Estimate Std. Error t value
## (Intercept)           -2.9618e+03 6.7808e+02 -4.3679
## timedelta              1.8912e+00 3.9325e-01  4.8092
## n_tokens_title          1.1101e+02 2.8435e+01  3.9041
## n_tokens_content         2.6766e-01 2.7823e-01  0.9620
## num_hrefs               2.8980e+01 7.2799e+00  3.9809
## num_self_hrefs          -5.6136e+01 1.7265e+01 -3.2515
## num_imgs                 1.3931e+01 9.7521e+00  1.4285
## average_token_length    -4.6962e+02 1.1079e+02 -4.2387
## num_keywords              5.7890e+01 3.3770e+01  1.7142
## data_channel_is_lifestyle -5.3930e+02 2.1705e+02 -2.4847
## data_channel_is_entertainment -6.5722e+02 1.4188e+02 -4.6323
## kw_min_min                2.2183e+00 1.1497e+00  1.9294
## kw_max_min                 1.3328e-01 7.4593e-02  1.7868
## kw_avg_min                 -6.7975e-01 5.7434e-01 -1.1835
## kw_min_max                 -2.5086e-03 6.8535e-04 -3.6603
## kw_min_avg                 -4.2869e-01 7.1805e-02 -5.9702
## kw_max_avg                 -2.4084e-01 2.3767e-02 -10.1334
## kw_avg_avg                  1.9493e+00 1.3451e-01 14.4913
## self_reference_min_shares   2.2720e-02 1.3491e-02  1.6841
## self_reference_max_shares   3.4592e-03 1.9480e-03  1.7758
## weekday_is_monday            4.9749e+02 1.8912e+02  2.6305
## weekday_is_saturday           5.9379e+02 2.9796e+02  1.9928
## LDA_00                         2.5922e+02 2.7314e+02  0.9490
## LDA_04                         1.1381e+02 2.0600e+02  0.5525
## global_subjectivity            2.6104e+03 6.6176e+02  3.9447
## global_rate_positive_words    -8.8872e+03 4.1561e+03 -2.1384
## rate_positive_words             3.2315e+02 4.2289e+02  0.7641
## avg_negative_polarity          -1.5861e+03 7.8656e+02 -2.0165
## max_negative_polarity           -4.0605e+02 8.1437e+02 -0.4986
## weekday_is_sunday                2.4024e+02 1.4086e+02  1.7056
## Pr(>|t|)
## (Intercept)                   1.258e-05 ***
## timedelta                      1.521e-06 ***
## n_tokens_title                  9.472e-05 ***
## n_tokens_content                  0.3360459
## num_hrefs                       6.879e-05 ***
## num_self_hrefs                   0.0011488 **
## num_imgs                          0.1531625
## average_token_length              2.253e-05 ***
## num_keywords                      0.0864956 .
## data_channel_is_lifestyle        0.0129687 *
## data_channel_is_entertainment     3.628e-06 ***
## kw_min_min                        0.0536866 .
## kw_max_min                        0.0739811 .
## kw_avg_min                        0.2365999

```

```

## kw_min_max          0.0002522 ***
## kw_min_avg          2.390e-09 ***
## kw_max_avg          < 2.2e-16 ***
## kw_avg_avg          < 2.2e-16 ***
## self_reference_min_shares 0.0921767 .
## self_reference_max_shares 0.0757804 .
## weekday_is_monday   0.0085284 **
## weekday_is_saturday 0.0462877 *
## LDA_00               0.3426063
## LDA_04               0.5806201
## global_subjectivity 8.004e-05 ***
## global_rate_positive_words 0.0324942 *
## rate_positive_words 0.4447881
## avg_negative_polarity 0.0437544 *
## max_negative_polarity 0.6180640
## weekday_is_sunday    0.0880925 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Noten que con la corrección, los resultados de la inferencia cambian si empleamos un nivel de confianza superior al 95 %. En este caso no podemos emplear la función **remueve.no.sinifica()** que habíamos empleado en el Capítulo 10. Podemos modificar esta función para que tenga en cuenta la corrección HC3. . Creemos la función **remueve.no.sinifica.HC3()**.

```

remueve.no.sinifica.HC3 <- 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 <- coeftest(modelo, vcov = (vcovHC(modelo)))
  # extrae los valores p
  pvals <- summ[, 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 <- coeftest(modelo, vcov = (vcovHC(modelo)))
pvals <- summ[, 4]
not_signif <- character()
not_signif <- names(which(pvals > p))
not_signif <- not_signif[!not_signif %in% "(Intercept)"]
}
modelo.limpio <- modelo
return(modelo.limpio)
}
```

El modelo 1 después de la eliminación de las variables explicativas no significativas (con la corrección HC3) se reporta en el Cuadro 10.2.

```
# remueve las variables no significativas HC3
modelo1 <- remueve.no.sinifica.HC3(modelo1, 0.05)
```

Cuadro 10.2: Modelo 1 estimado por MCO y corrección HC

	Dependent variable:	
	shares	
	MCO (1)	HC3 (2)
timedelta	2.012*** (0.289)	2.012*** (0.313)
n_tokens_title	113.101*** (28.523)	113.101*** (28.417)
num_hrefs	31.979*** (6.007)	31.979*** (7.436)
num_self_hrefs	-62.929*** (16.793)	-62.929*** (16.991)
num_imgs	19.292** (7.591)	19.292** (7.855)
average_token_length	-461.225*** (91.192)	-461.225*** (93.976)
data_channel_is_lifestyle	-439.375* (262.568)	-439.375** (209.846)
data_channel_is_entertainment	-775.435*** (156.015)	-775.435*** (130.636)
kw_min_max	-0.003** (0.001)	-0.003*** (0.001)
kw_min_avg	-0.419*** (0.070)	-0.419*** (0.069)
kw_max_avg	-0.213*** (0.020)	-0.213*** (0.024)
kw_avg_avg	1.845*** (0.106)	1.845*** (0.129)
self_reference_max_shares	0.009*** (0.001)	0.009*** (0.003)
weekday_is_monday	504.489*** (157.131)	504.489** (191.284)
global_subjectivity	2,420.731*** (675.566)	2,420.731*** (608.437)
avg_negative_polarity	-1,777.658*** (514.280)	-1,777.658** (693.150)
is_weekend	415.816** (174.886)	415.816** (166.009)
Constant	-2,299.812*** (542.508)	-2,299.812*** (627.582)
Observations	39,644	39,644
R ²	0.021	0.021
Adjusted R ²	0.021	0.021
Residual Std. Error (df = 39626)	11,506.430	11,506.430

Note:

*p<0.1; **p<0.05; ***p<0.01

Ahora sigamos con el modelo 3: algoritmo Forward y criterio AIC. En este caso tenemos:

```
model3 <- ols_step_forward_aic(max.modelo)
# se extraen las variables del mejor modelo según el
# algoritmo
vars.modelo3 <- model3$predictors
# se construye la fórmula
```

```
formula.modelo3 <- as.formula(paste("shares ~ ", paste(vars.modelo3,
  collapse = " + "), sep = ""))

# se estima el modelo con la fórmula construida
modelo3 <- lm(formula.modelo3, data = datos.caso1)
```

Ahora procedamos de nuevo a realizar las pruebas de heteroscedasticidad.

```
# se extraen los residuales
res.modelo3 <- residuals(modelo3)
# pruebas de normalidad Kolmogorov-Smirnov
ks.test(res.modelo3, y = pnorm)

## 
## One-sample Kolmogorov-Smirnov test
##
## data: res.modelo3
## D = 0.76089, p-value < 2.2e-16
## alternative hypothesis: two-sided

# pruebas de normalidad Kolmogorov-Smirnov
jarque.bera.test(res.modelo3)

## 
## Jarque Bera Test
##
## data: res.modelo3
## X-squared = 5853868793, df = 2, p-value < 2.2e-16

# prueba de Breusch-Pagan studentizada
bptest(modelo3, studentize = TRUE)

## 
## studentized Breusch-Pagan test
##
## data: modelo3
## BP = 79.676, df = 28, p-value = 7.458e-07
```

La versión studentizada de la prueba de Breusch-Pagan permite concluir con un 99 % de confianza que hay heteroscedasticidad. Ahora realicemos la corrección HC3 y eliminemos las variables no significativas. El resultado se reporta en el Cuadro 10.3. Si comparamos estos resultados con los obtenidos en el Capítulo 7 (Cuadro 7.3) podemos ver diferencias entre las conclusiones para cada modelo. Es este caso los modelo 1 y 3 no se encuentra anidados.

Cuadro 10.3: Modelo seleccionado el algoritmo stepwise forward con corrección HC3

	<i>Dependent variable:</i>	
	shares	
	Modelo 1(HC3) (R^2 adj)	Modelo 3 (HC3) (AIC)
	(1)	(2)
timedelta	2.012*** (0.313)	1.950*** (0.316)
n_tokens_title	113.101*** (28.417)	118.105*** (28.905)
num_hrefs	31.979*** (7.456)	33.158*** (7.453)
num_self_hrefs	−62.929*** (16.991)	−62.426*** (16.690)
num_imgs	19.292** (7.855)	
LDA_02		−787.203*** (192.924)
global_rate_positive_words		−10.992.610*** (3.488.953)
min_positive_polarity		−2.150.639*** (733.476)
average_token_length	−461.225*** (93.976)	−336.557*** (97.749)
data_channel_is_lifestyle	−439.375** (209.846)	−556.753*** (215.703)
num_keywords		66.533** (33.090)
abs_title_sentiment_polarity		661.541** (262.492)
abs_title_subjectivity		658.216** (329.711)
data_channel_is_entertainment	−775.435*** (130.636)	−818.212*** (143.022)
kw_min_max	−0.003*** (0.001)	−0.002*** (0.001)
kw_min_avg	−0.419*** (0.069)	−0.365*** (0.076)
kw_max_avg	−0.213*** (0.024)	−0.204*** (0.024)
kw_avg_avg	1.845*** (0.129)	1.766*** (0.140)
self_reference_max_shares	0.009*** (0.003)	0.009*** (0.003)
weekday_is_monday	504.489*** (191.284)	434.380** (190.030)
global_subjectivity	2.420.731*** (608.437)	2.709.752*** (625.581)
avg_negative_polarity	−1.777.658** (693.150)	−1.675.733** (701.510)
is_weekend	415.816** (166.009)	
Constant	−2,299.812*** (627.582)	−2,710.776*** (664.314)
Observations	39.644	39.644
R ²	0.021	0.021
Adjusted R ²	0.021	0.021
Residual Std. Error	11,506.430 (df = 39626)	11,504.740 (df = 39622)

Note: *p<0.1; **p<0.05; ***p<0.01

10.3.2 Stepwise backward

De manera similar en el Cuadro 10.4 se presentan los resultados de emplear el algoritmo *stepwise backward* y tras limpiar las variables no significativas con la corrección HC3 (con un 95 % de confianza). Previamente puedes mostrar que estos dos modelos tienen un problema de heteroscedasticidad. Estos resultados implican que el modelo 6 está anidado en el 4 (pero no en el 1 o 3).

```
## Reordering variables and trying again:
```

Cuadro 10.4: Modelos seleccionados con el algoritmo stepwise backward con corrección HC3

	<i>Dependent variable:</i>	
	shares	
	Modelo 4 (HC3) (R^2 aj)	Modelo 6 (HC3) (AIC)
	(1)	(2)
timedelta	2.012*** (0.295)	
n_tokens_title	111.215*** (28.360)	
num_hrefs	30.457*** (7.301)	
num_self_hrefs	−56.059*** (16.488)	
num_imgs	17.013** (7.561)	
average_token_length	−252.397*** (88.422)	
data_channel_is_entertainment	−826.411*** (142.573)	
kw_min_max	−0.003*** (0.001)	
kw_min_avg	−0.392*** (0.074)	
kw_max_avg	−0.202*** (0.024)	
kw_avg_avg	1.754*** (0.148)	
self_reference_max_shares	0.009*** (0.003)	
weekday_is_monday	500.657*** (191.424)	
weekday_is_saturday		717.793** (294.231)
weekday_is_sunday		320.349** (134.332)
LDA_03	693.450** (320.933)	3,962.645*** (242.651)
avg_negative_polarity	−2,196.253*** (699.360)	
LDA_04		1,580.746*** (166.055)
LDA_01		1,037.607*** (194.758)
data_channel_is_bus		−703.496** (296.249)
is_weekend	408.830** (165.954)	
kw_avg_max		0.002*** (0.0005)
LDA_00		2,321.693*** (340.952)
min_negative_polarity		−741.708*** (218.448)
self_reference_avg_shares		0.024*** (0.009)
Constant	−2,237.443*** (627.939)	626.971*** (195.051)
Observations	39,644	39,644
R ²	0.021	0.012
Adjusted R ²	0.020	0.012
Residual Std. Error	11,507.320 (df = 39627)	11,557.930 (df = 39633)

Note:

*p<0.1; **p<0.05; ***p<0.01

10.3.3 Combinando forward y backward

Y finalmente, el Cuadro 10.5 se presentan los resultados de emplear el algoritmo combinado y tras limpiar las variables no significativas y la corrección HC3(con un 95 % de confianza). (Recuerda hacer las pruebas de heteroscedasticidad para cada modelo) Los modelos 7 y 9 no están anidados.

```
## Reordering variables and trying again:
```

Cuadro 10.5: Modelo seleccionado el algoritmo stepwise forward y backward con corrección HC3

	Dependent variable:	
	shares	
	Modelo 7 (HC3) (R^2 aj)	Modelo 9 (HC3) (AIC)
timedelta	1.858*** (0.310)	1.950*** (0.316)
n_tokens_title	111.485*** (28.296)	118.105*** (28.905)
num_hrefs	30.030*** (7.406)	33.158*** (7.453)
num_self_hrefs	-55.725*** (16.107)	-62.426*** (16.690)
data_channel_is_entertainment	-899.253*** (150.990)	-818.212*** (143.022)
kw_min_max	-0.003*** (0.001)	-0.002*** (0.001)
kw_min_avg	-0.378*** (0.075)	-0.365*** (0.076)
kw_max_avg	-0.194*** (0.024)	-0.204*** (0.024)
kw_avg_avg	1.700*** (0.155)	1.766*** (0.140)
self_reference_max_shares	0.009*** (0.003)	0.009*** (0.003)
data_channel_is_lifestyle		-556.753*** (215.703)
num_keywords		66.533** (33.090)
weekday_is_monday	490.851** (191.719)	434.380** (190.030)
LDA_02	-826.886*** (189.820)	-787.203*** (192.924)
LDA_03	749.927** (328.032)	
global_rate_positive_words	-8,507.198*** (3,236.867)	-10,992.610*** (3,488.953)
min_positive_polarity	-2,189.555*** (751.224)	-2,150.639*** (733.476)
avg_negative_polarity	-1,869.996*** (645.946)	-1,675.733** (701.510)
average_token_length		-336.557*** (97.749)
global_subjectivity		2,709.752*** (625.581)
abs_title_sentiment_polarity	491.221** (244.026)	661.541** (262.492)
is_weekend	427.803** (167.261)	
abs_title_subjectivity		658.216** (329.711)
Constant	-2,409.251*** (542.894)	-2,710.776*** (664.314)
Observations	39.644	39.644
R ²	0.021	0.021
Adjusted R ²	0.021	0.021
Residual Std. Error	11,506.550 (df = 39625)	11,504.740 (df = 39622)

Note:

*p<0.1; **p<0.05; ***p<0.01

10.4 Comparación de modelos

En resumen, contamos con 6 modelos con las variables explicativas que se representan con una X en el Cuadro 10.6. El modelo 6 está anidado en el modelo 4 y en el 7. Puedes probar rápidamente que el modelo 4 es mejor que el 6 y el 7 es mejor que el 4. Por otro lado, el modelo 9 y el 3 son los mismos. Así, descartaremos el modelo 6 y el 9 del análisis. Los modelos otros modelos no se encuentran anidados.

Cuadro 10.6: Variables explicativas incluidas en cada uno de los modelos calculados (con corrección HC3)

	Forward		Backward		Both	
	modelo 1	modelo 3	modelo 4	modelo 6	modelo 7	modelo 9
timedelta	X	X	X		X	X
n_tokens_title	X	X	X		X	X
n_tokens_content						
num_href	X	X	X		X	X
num_self_href	X	X	X		X	X
num_keywords						X
average_token_length	X	X	X			X
num_imgs			X			
data_channel_is_lifestyle	X	X				X
data_channel_is_entertainment	X	X	X		X	X
kw_min_max	X	X	X		X	X
kw_min_avg	X	X	X		X	X
kw_max_avg	X	X	X		X	X
kw_avg_avg	X	X	X		X	X
self_reference_max_shares	X	X	X		X	X
data_channel_is_lifestyle	X	X				X
weekday_is_monday	X	X	X		X	X
is_weekend			X		X	
LDA_02		X			X	X
LDA_03			X	X	X	
LDA_04						
global_rate_positive_words		X			X	X
global_subjectivity	X	X				X
min_positive_polarity		X			X	X
avg_negative_polarity	X	X	X		X	X
abs_title_sentiment_polarity		X			X	X
abs_title_subjectivity	X					X

De esta manera tendremos que comparar estos modelos con pruebas de modelos no anidados empleando la prueba J con corrección HC3. La prueba de Cox no es posible implementarla con H.C.. A continuación se presenta el código ajustado de la Prueba J para la corrección H.C. empleando la función `jtest()` del *AER* (Kleiber y Zeileis, 2008).

```
library(AER)
# comparación de modelos no anidados

J.res1.3 <- jtest(modelo1, modelo3, vcov. = vcovHC)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    NA TRUE FALSE  TRUE
## [2,]   TRUE    NA  TRUE FALSE
## [3,]  TRUE  TRUE    NA  TRUE
## [4,]  TRUE  TRUE   TRUE   NA
```

Empecemos comparando todos los modelos con la prueba J. En el Cuadro 10.7 se reportan los valores p de las pruebas J que permiten probar la hipótesis nula de que el modelo de la fila es mejor que el de la columna.

Cuadro 10.7: Valores p de las pruebas J (H_0 : modelo de la fila es mejor que el de la columna (con corrección HC3))

	Modelo 1	Modelo 3	Modelo 4	Modelo 7
Modelo 1		0.000	0.072	0.000
Modelo 3	0.002		0.006	0.014
Modelo 4	0.000	0.000		0.000
Modelo 7	0.000	0.000	0.004	

Si miramos la primera fila, con un 99 % de confianza, podemos concluir que el modelo 1 no es mejor que los otros modelos¹. Si miramos la primera columna, podemos ver como la nula de que el modelo 3 es mejor que el 1 no se puede rechazar. Es decir, con un 99 % de confianza, podemos concluir que los modelos 3 es mejor que el 1. Para los otros modelos no se puede afirmar algo similar, y por tanto la prueba no es concluyente. Al comparar el modelo 3 con el 4, se encuentra que se puede rechazar la nula de que el modelo 3 es mejor que el 4, y lo contrario es cierto. Es decir, la prueba no es concluyente. La prueba si puede concluir en favor que el modelo 3 es mejor que el 7.

Para las otras comparaciones que no se mencionan la prueba no es concluyente. Es decir, poniendo todo junto el modelo 3 es el mejor.

Ahora empleemos las métricas AIC y BIC. Estas medidas no se ven afectadas por la presencia de heteroscedasticidad ni la multicolinealidad. No obstante, el R^2 ajustado al depender del R^2 podría verse afectado en presencia de una multicolinealidad fuerte. Los resultados se reportan en el Cuadro 10.8 y BIC()

Cuadro 10.8: Medidas de bondad de ajuste para los 4 modelos comprados

	AIC	BIC
Modelo 1	853919.997	854083.163
Modelo 3	853912.391	854109.908
Modelo 4	853925.168	854079.747
Modelo 7	853921.852	854093.606

El AIC sugieren que el mejor modelo es el 3, mientras que el BIC selecciona el 1. Poniendo todo junto, el mejor modelo será el modelo 3 (que es igual al 9) el cuál se reporta en el Cuadro 10.9.

¹Nota que la hipótesis nula asociados a los valores p reportados en la primera final del Cuadro 10.7 corresponde a que el modelo 1 es mejor al modelo de la respectiva columna. Y esa hipótesis nula se puede rechazar en todos los casos.

Cuadro 10.9: Modelo estimado por MCO y corrección HC

<i>Dependent variable:</i>	
	shares HC3
kw_avg_avg	1.766*** (0.140)
kw_max_avg	−0.204*** (0.024)
kw_min_avg	−0.365*** (0.076)
timedelta	1.950*** (0.316)
num_hrefs	33.158*** (7.453)
n_tokens_title	118.105*** (28.905)
data_channel_is_entertainment	−818.212*** (143.022)
avg_negative_polarity	−1,675.733** (701.510)
average_token_length	−336.557*** (97.749)
global_subjectivity	2,709.752*** (625.581)
weekday_is_monday	434.380** (190.030)
kw_min_max	−0.002*** (0.001)
num_self_hrefs	−62.426*** (16.690)
LDA_02	−787.203*** (192.924)
global_rate_positive_words	−10,992.610***

Antes de pasar a sacar conclusiones, procedamos a determinar si el mejor modelo tiene o no multicolinealidad. Nota que el *VIF* depende de la matriz de varianzas y covarianzas de los estimadores MCO y en presencia de heteroscedasticidad, éstos no son confiables. Así no podemos emplearlos en esta situación. Así que nos concentraremos en la prueba de Belsley y col. (1980) también conocida como la prueba Kappa.

El siguiente código permite hacer la prueba.

```
XTX <- model.matrix(modelo3)
e <- eigen(t(XTX) %*% XTX)

lambda.1 <- max(e$val)
lambda.k <- min(e$val)
kappa <- sqrt(lambda.1/lambda.k)
kappa

## [1] 4304419
```

Este estadístico es muy grande ($\kappa = 4,3044185 \times 10^6$). Esta prueba sugiere la existencia de un problema serio de multicolinealidad. Nota que los síntomas del problema de multicolinealidad no estaban presentes. No obstante la prueba si detecta un problema serio. Solucionar este problema no será fácil, pues no podemos emplear el *VIF* para eliminar variables. Así que la interpretación que realicemos de los coeficientes puede tener problemas.

Así continuaremos con este problema, teniendo precaución con las conclusiones que saquemos.

10.5 Identificación de la variable más importante

Como lo discutimos una pregunta habitual cuando estamos haciendo analítica diagnóstica es ¿cuál variable es la mas importante para explicar la variable dependiente? en el Capítulo 7 discutimos dos formas de hacer esto empleando el aumento en el R^2 que se obtiene al adicionar el respectivo regresor dado que ya están en el modelo las otras $k - 2$ variables explicativas y con los coeficientes estandarizados. Nota que en este caso no es posible emplear el R^2 , pues el problema de multicolinealidad podría afectar esta métrica. Así que es mejor descartar esta aproximación.

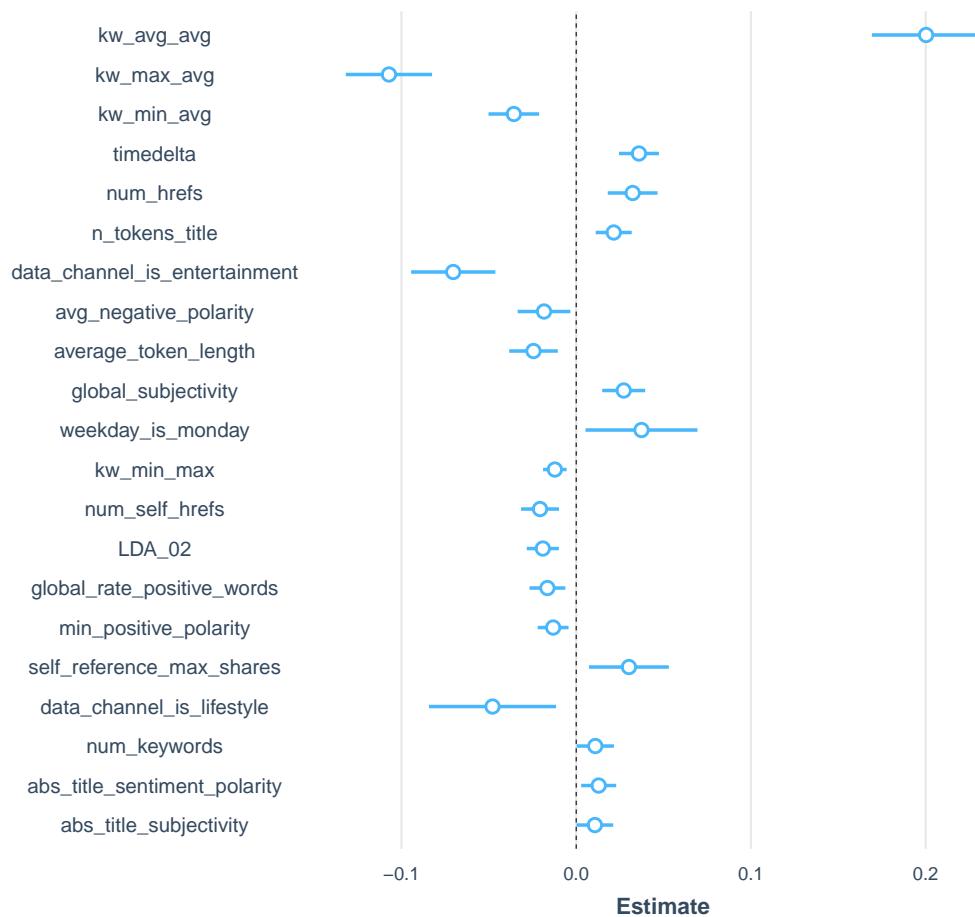
Por otro lado, también hay que reconocer que en presencia de heteroscedasticidad los coeficientes estandarizados como los discutimos en el Capítulo 7 se ven afectados. Es decir, al tener un error heteroscedástico, tendremos que la variable dependiente también tiene una varianza no constante. Y por tanto emplear s_y no sería adecuado. Pero podemos emplear el estimador H.C. para tener un estimador de la desviación estándar de la variable dependiente que no sea constante para toda la muestra. Esto no lo podemos hacer con el paquete *relaimpo* (Grömping, 2006). Pero si lo podemos hacer con el paquete *jtools* (Long, 2020) como lo veremos a continuación.

10.6 Generación de visualizaciones de los resultados

Ahora veamos cómo se pueden presentar los resultados de una manera más amigable y en presencia de heteroscedasticidad. Las visualizaciones que generamos en el Capítulo 7 emplean las estimaciones del modelo por MCO, pero podemos modificarlas rápidamente para incluir la corrección H.C..

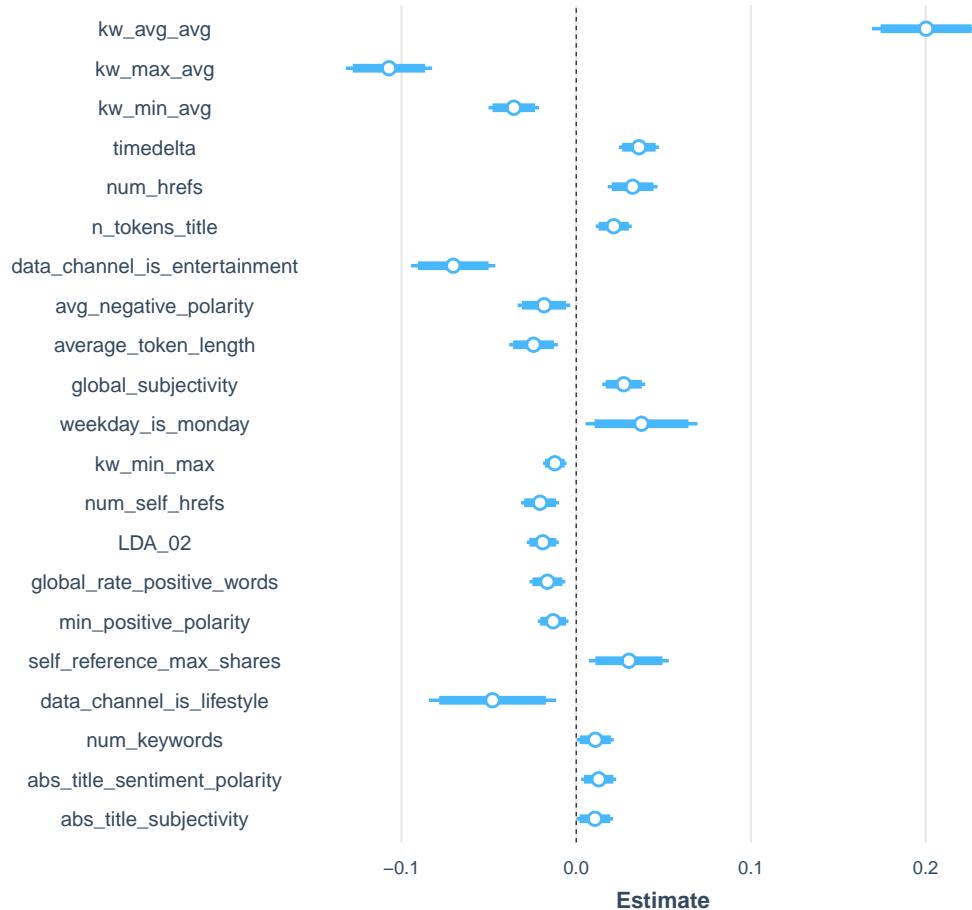
Seguiremos empleando el paquete *jtools* (Long, 2020) para visualizar los resultados del objeto `lm` pero con la corrección de heteroscedasticidad. La función `plot_summs()` tiene el argumento `robust` que permite incluir la corrección H.C. o H.A.C.. En este caso.

```
library(jtools)
plot_summs(modelo3, robust = "HC3", scale = TRUE, transform.response = TRUE)
```



Este gráfico permite ver los coeficientes y sus respectivos intervalos de confianza con la corrección HC3. También podemos graficar los coeficientes estandarizados teniendo en cuenta la corrección HC3, de la siguiente manera:

```
plot_summs(modelo3, robust = "HC3", scale = TRUE, transform.response = TRUE,  
inner_ci_level = 0.9)
```



Ahora ya estás listos para proceder a generar las recomendaciones al editor. ¿Qué recomiendas?

10.6.1 Comentarios Finales

En este capítulo hemos seguido el proceso paso a paso para responder una pregunta de negocio empleando un modelo de regresión para hacer analítica diagnóstica. A diferencia de lo realizado en el Capítulo 7, esta vez realizamos el chequeo del cumplimiento de los supuestos del Teorema de Gauss-Markov y ahora si podemos estar seguros de sacar conclusiones con nuestro modelo.



11 : Autocorrelación

Diseñado por Freepik

Objetivos del capítulo

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

- Explicar en sus propias palabras los efectos de la autocorrelación sobre los estimadores MCO.
- Realizar en R diferentes tipos de análisis gráficos que revelen la posibilidad de autocorrelación en los residuos.
- Efectuar las pruebas estadísticas necesarias para detectar la violación del supuesto de no autocorrelación en los residuos empleado R. 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.

11.1 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 (Capítulo 8), posteriormente nos enfocamos en las consecuencias de un error homoscedástico (Capítulo 9).

Finalmente, y para concluir nuestra discusión de la violación de los supuestos que garantizan el cumplimiento del Teorema de Gauss-Markov 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.

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, $\varepsilon_t \sim i.i.d (0, \sigma^2)$ o en forma matricial $\varepsilon_{n \times 1} \sim (0_{n \times 1}, \sigma^2 I_n)$

Si existe autocorrelación entre los errores, entonces los estimadores MCO siguen siendo insesgados pero no son eficientes¹ 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:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \quad (11.1)$$

donde el término del error tiene media cero y se encuentra correlacionado con el error del periodo anterior²: $E[\varepsilon_t] = 0$; $\varepsilon_t = \rho \varepsilon_{t-1} + v_t \quad \forall t$ 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 que:

$$\sigma_\varepsilon^2 = \frac{\sigma_v^2}{(1 - \rho^2)}$$

¹Una demostración de esta afirmación se presenta en el anexo al final de capítulo en la sección 11.5

²A la observación del periodo anterior se le conoce como la variable rezagada. Es decir, Y_{t-1} es la variable rezagada de Y_t ; rezagada un periodo. En este orden de ideas Z_{t-2} está rezagada dos periodos. En general es común emplear la letra p para representar el número de rezagos de una variable.

Por otro lado, dado que el valor esperado del error es cero se puede mostrar fácilmente que:

$$\text{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \rho \sigma_\varepsilon^2$$

Entonces, la correlación entre los errores adyacentes³ 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:⁴

$$\rho(s) = \rho^s \sigma_\varepsilon^2$$

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 los ejemplos 11.1 y 11.1. El Ejemplo 11.1 muestra la situación cuando no existiese un problema de autocorrelación, la autocorrelación para los diferentes rezagos será cero.

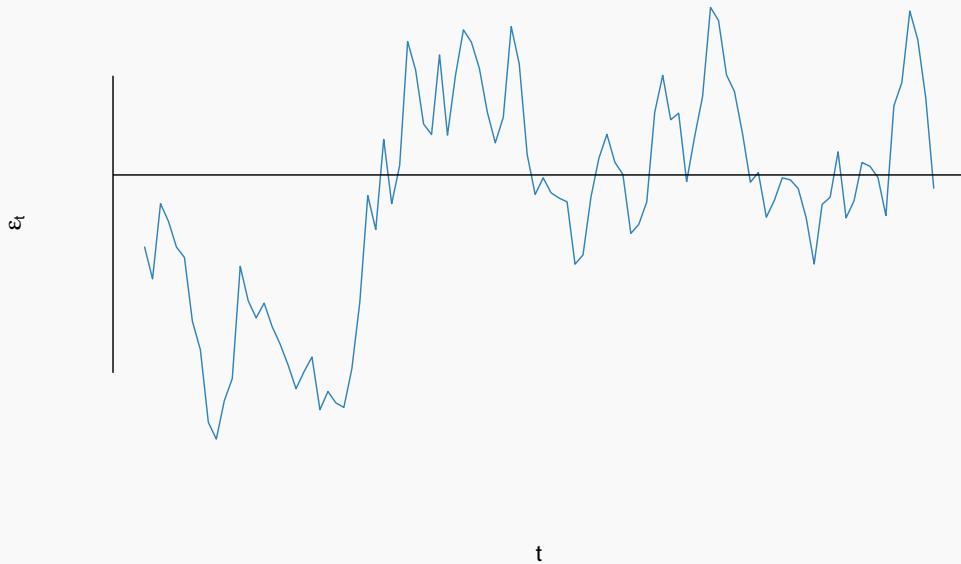
Ejemplo 11.1 Errores con un proceso AR(1) y autocorrelación positiva

En la Figura 11.1 se presenta un error cuya autocorrelación es de 0.8 ($\rho = 0.8$). Noten que para un periodo t los errores tienden a mantener el signo del error del periodo anterior ($t - 1$). Esto se puede ver de mejor manera si se visualizan los errores del periodo t (ε_t) en función de los del periodo anterior (ε_{t-1}) como se presenta en la Figura 11.2. Nota que se observa una relación lineal muy fuerte entre los errores.

³La correlación entre errores inmediatamente adyacentes también se domina la autocorrelación a un rezago. Por ejemplo, la correlación entre ε_2 y ε_1 o la correlación entre ε_3 y ε_2 . Y en general será la correlación entre ε_t y ε_{t-1} . Si se considera la relación entre errores separados por dos periodos se denomina autocorrelación para a dos rezagos (ε_t y ε_{t-2}), etc.

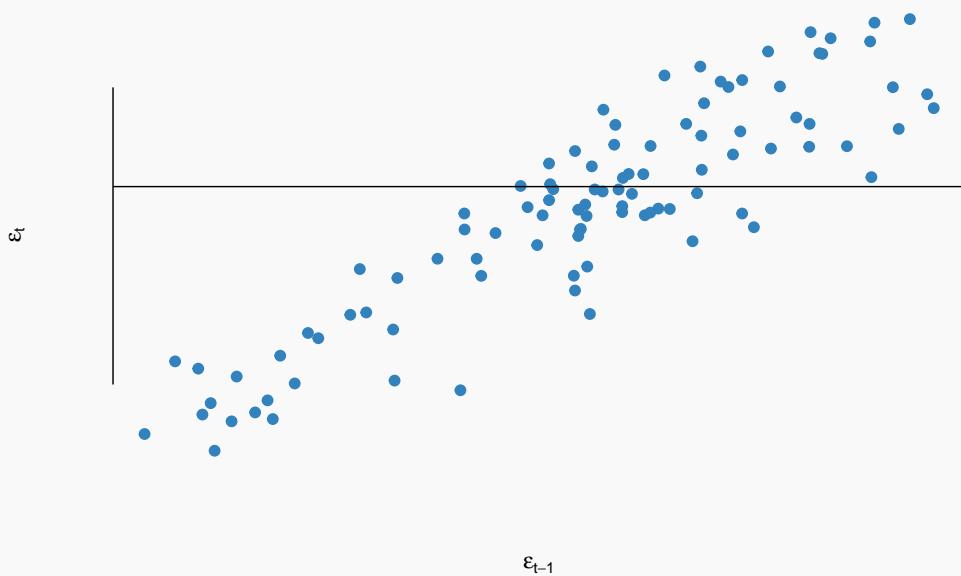
⁴A la función que muestra la correlación para diferentes rezagos de un proceso se le denomina función de autocorrelación.

Figura 11.1. Comportamiento en el tiempo del error simulado de un proceso AR(1) con $\rho = 0,8$



Fuente: Elaboración propia

Figura 11.2. Error simulado para el periodo t versus el mismo error en el periodo anterior (error simulado de un proceso AR(1) con $\rho = 0,8$)

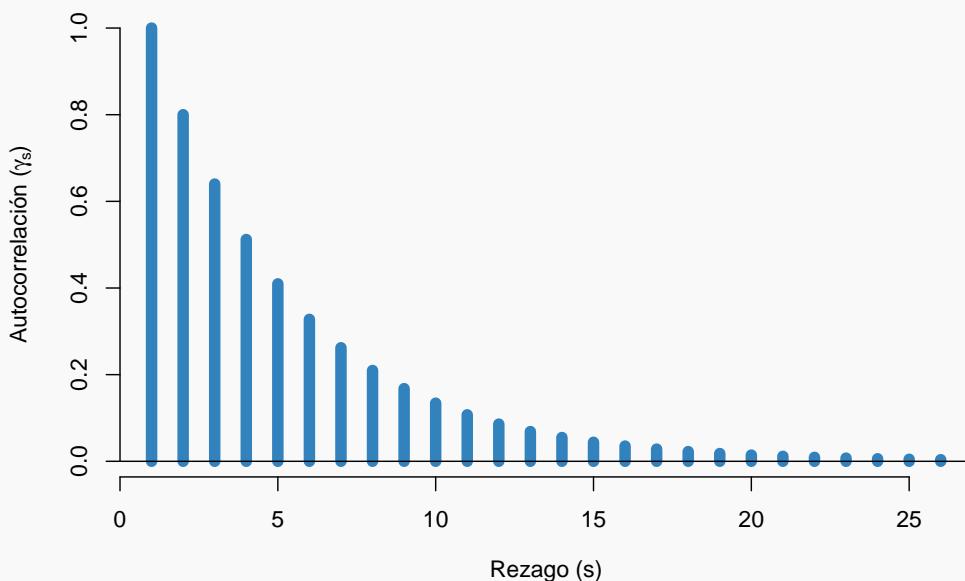


Fuente: Elaboración propia

Adicionalmente, podemos observar en la Figura 11.3 la autocorrelación para estos errores para diferentes rezagos. Este gráfico muestra como para errores autocorrelacionados con un

proceso AR(1) y con un $\rho = 0,8$) la correlación entre las observaciones se hace menos fuerte entre mas alejados estén.

Figura 11.3. Comportamiento de la autocorrelación para diferentes rezagos del error simulado de un proceos AR(1) con $\rho = 0,8$



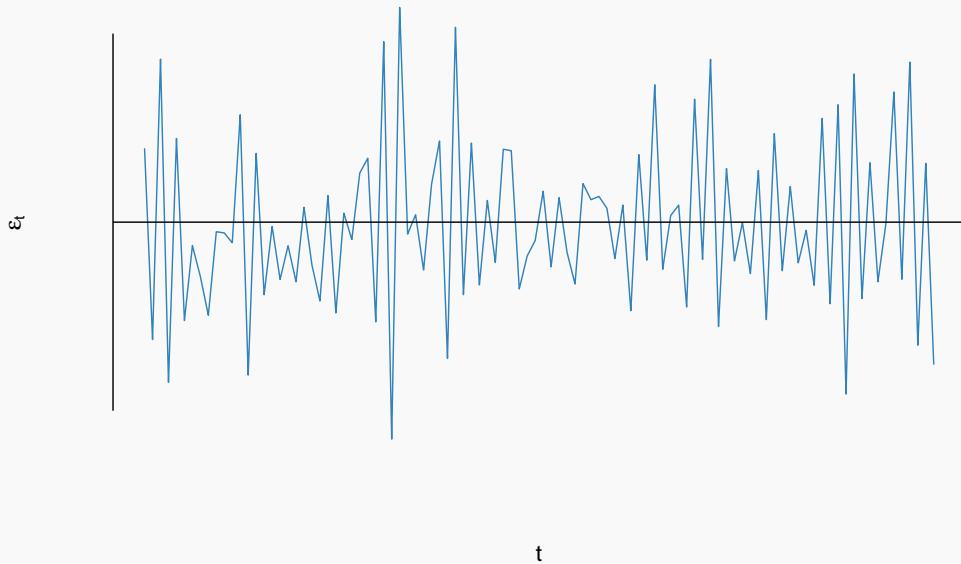
Fuente: Elaboración propia

En general, **cuando un error de un modelo sigue un proceso AR(1) y $0 < \rho < 1$ diremos que el modelo tiene Autocorrelación positiva**. En este caso, los errores de períodos adyacentes tienden a tener el mismo signo y esto implicará obtener gráficos similares a los obervados en las Figuras 11.1 y 11.2.

Ejemplo 11.2 Errores con un proceso AR(1) y autocorrelación negativa

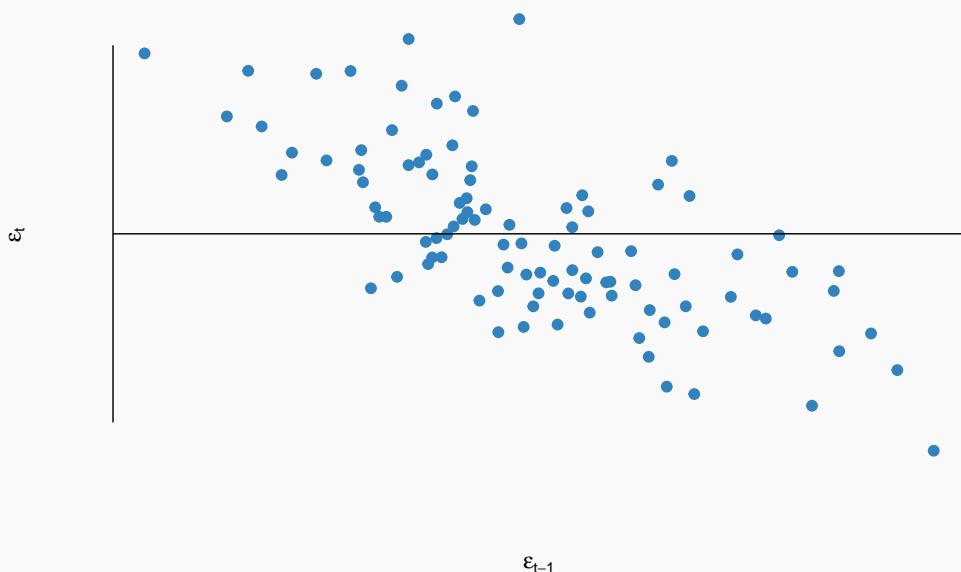
En la Figura 11.4 se presenta un error cuya autocorrelación es de -0.8 ($\rho = -0,8$). Noten que a diferencia del ejemplo anterior, para un periodo t los errores tienden a cambiar el signo observado en el periodo anterior ($t - 1$). Esto se puede ver de mejor manera si se visualizan los errores del periodo t (ε_t) en función de los del periodo anterior (ε_{t-1}) como se presnta en la Figura 11.5. Nota que se observa una realción lineal muy fuerte entre los errores, pero con pendiente negativa.

Figura 11.4. Comportamiento en el tiempo del error simulado de un proceso AR(1) con $\rho = 0,8$



Fuente: Elaboración propia

Figura 11.5. Error simulado para el periodo t versus el mismo error en el periodo anterior (error simulado de un proceso AR(1) con $\rho = -0,8$)

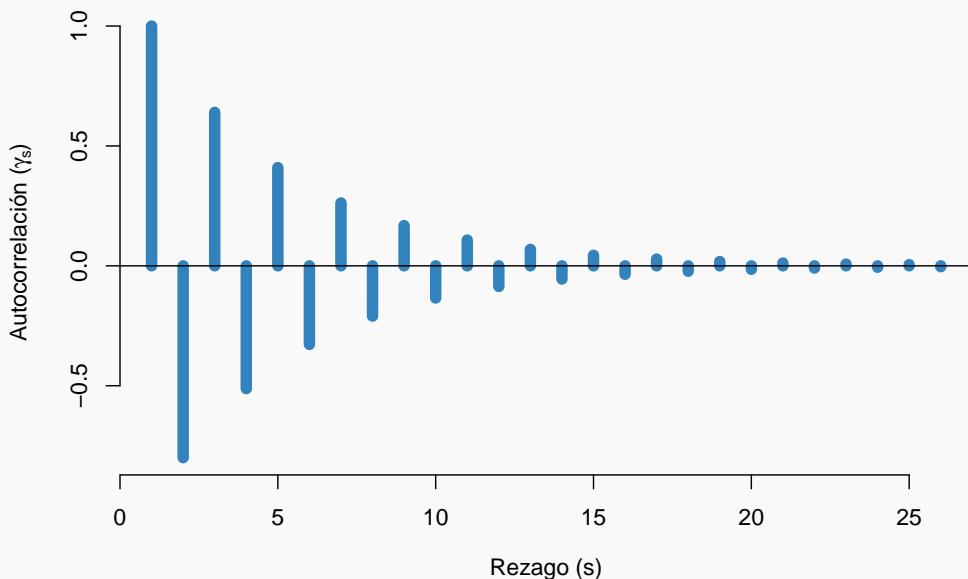


Fuente: Elaboración propia

Adicionamente, podemos observar en la Figura 11.6 la autorrelación para estos errores para diferentes rezagos. Este gráfico muestra como para errores autocorrelacionados con un proceso

AR(1) y con un $\rho = -0,8$) la correlación entre las observaciones se hacee menos fuerte entre mas alejados estén, pero la correlación tiene signo contrario.

Figura 11.6. Comportamiento de la autocorrelación para diferentes rezagos del error simulado de un proceos AR(1) con $\rho = -0,8$



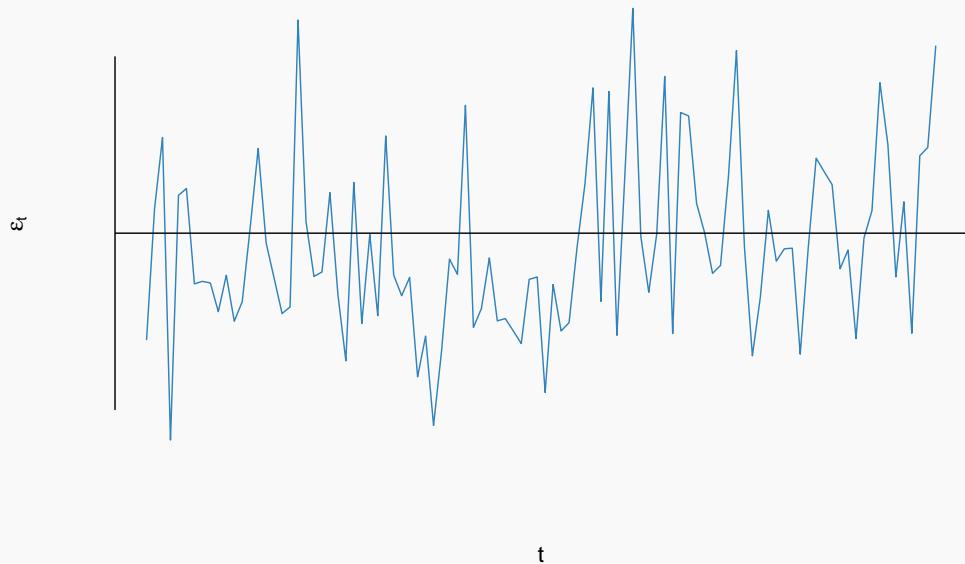
Fuente: Elaboración propia

En general, **cuando un error de un modelo sigue un proceso AR(1) y $-1 < \rho < 0$ diremos que el modelo tiene Autocorrelación negativa**. En este caso, los errores de períodos adyacentes tienden a tener signo contrario y esto implicará obtener gráficos similares a los observados en las Figuras 11.4 y 11.5.

Ejemplo 11.3 Errores sin autocorrelación $\rho = 0$

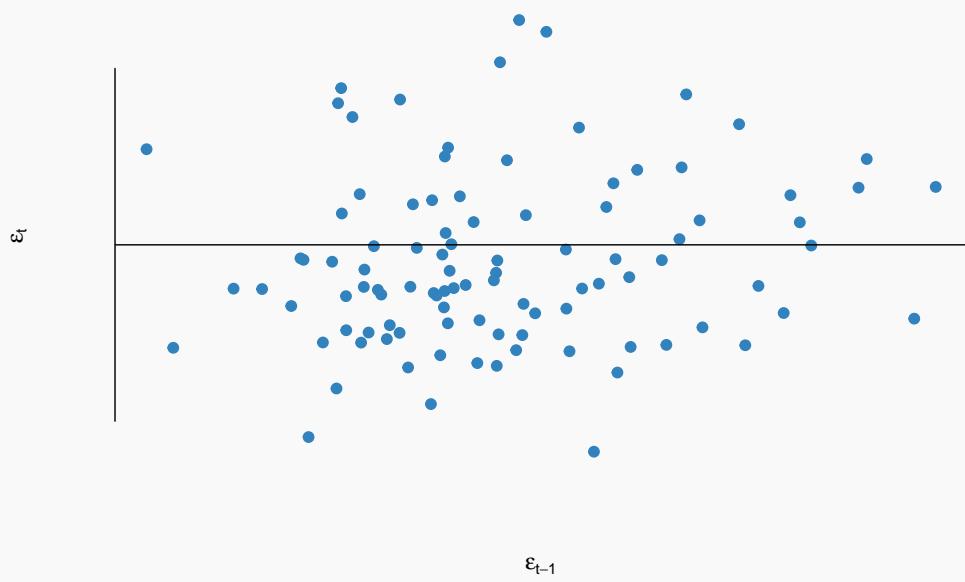
En la Figura 11.7 se presenta un error sin autocorrelación ($\rho = 0$). En este caso el patrón de los residuales es totalmente aleatorio. Al visualizar los errores del periodo t (ε_t) en función de los del periodo anterior (ε_{t-1}) no se observa ningún patrón (ver Figura 11.8. Nota que tampoco existirán autocorrelaciones y por eso no las graficaremos.

Figura 11.7. Comportamiento en el tiempo del error simulado sin autocorrelación



Fuente: Elaboración propia

Figura 11.8. Error simulado no autocorrelacionado para el periodo t versus el mismo error en el periodo anterior ($t - 1$)



Fuente: Elaboración propia

Regresando al problema de autocorrelación, este problema 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. 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 individuos tendrán que esperar unos períodos para ajustar sus decisiones. En general, la autocorrelación es un problema muy común en modelos que emplean series de tiempo.

Es poco probable que exista autocorrelación en datos de corte trascversal, pero de existirlo es muy fácil de eliminarlo. Si implementen podríamos reorganizar las observaciones para eliminar la relación entre las observaciones adyacentes. Por otro lado esto es imposible en el caso de una muestra de serie de tiempo, pues el orden de las observaciones es importante en las series de tiempo y no se puede modificar.

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 $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + \cdots + \rho_p \varepsilon_{t-p} + v_t$ 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 (dado que hay homoscedasticidad) 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:

$$E [\varepsilon^T \varepsilon] = \Omega = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho & 1 \end{bmatrix} \neq \sigma^2 I_n$$

Como se mencionó anteriormente, en presencia de autocorrelación los estimadores MCO continúan siendo insesgados pero no tienen la mínima varianza posible.⁵ Es más, en presencia de autocorrelación, el estimador MCO de la matriz de varianzas y covarianzas del vector β será sesgado.⁶ 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.

11.2 Pruebas para la detección de autocorrelación

En el análisis tradicional el primer paso, antes de realizar pruebas estadísticas formales, es verificar gráficamente si los residuales presentan síntomas de la presencia del problema. No obstante, así

⁵En el anexo al final del capítulo (ver sección 11.5) se presenta una demostración de esta afirmación.

⁶En el anexo al final del capítulo (ver sección 11.5) se presenta una demostración de esta afirmación.

como en el caso de los dos problemas anteriormente estudiados, es posible que el científico de datos se enfrente a tener que estimar muchos modelos de manera automática y por tanto no tenga sentido realizar un análisis preliminar gráfico.

Los síntomas de autocorrelación en una regresión se pueden observar en el vector de errores. Pero dicho vector no es observable, por tanto la mejor aproximación es examinar los errores estimados. Las gráficas más empleadas son:

1. Los errores contra el tiempo $\hat{\epsilon}$ vs $t = 1, 2, \dots, n$. Como por ejemplo las Figuras 11.1, 11.4 y 11.7.
2. Los errores contra los errores del periodo anterior $\hat{\epsilon}_t$ vs $\hat{\epsilon}_{t-1}$. Como por ejemplo las Figuras 11.2, 11.5 y 11.8.

Estos gráficos permiten identificar algún tipo de regularidad como los que se presentan en el ejemplo 11.1. Por un lado, 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 (ver ejemplo 11.1), ya que el signo de los residuos cambia de un periodo a otro. Si no existe autocorrelación, en estos gráficos (ver ejemplo 11.1) no podremos observar patrón claro de comportamiento en el signo de los residuales y en su relación con valore pasados.

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. El científico de datos tendrá que decidir si este tipo de análisis gráfico agrega o no valor al análisis y la presentación de resultados que esta realizando para cada pregunta de negocio en particular.

11.2.1 Prueba de Rachas (Runs test)

La prueba de rachas (o *Runs test* en inglés), propuesta por Wald y Wolfowitz (1940), es una prueba de independencia lineal no paramétrica cuya idea es relativamente sencilla. Si no hay autocorrelación, entonces no deberían observarse 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 anexo al final del capítulo (ver sección 11.5)
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:

$$\left[E[k] \pm z_{\frac{\alpha}{2}} \sqrt{Var[k]} \right] \quad (11.2)$$

donde, el valor esperado y la varianza de k (las rachas) son:

$$E(k) = \frac{2N_+N_-}{N_+ + N_-} + 1 \quad (11.3)$$

$$Var(k) = \frac{2N_+N_- (2N_+N_- - N_+ - N_-)}{(N_+ + N_-)^2 (N_+ + N_- - 1)} \quad (11.4)$$

La hipótesis nula se puede rechazar si el número de rachas observadas no están contenidas en el intervalo de confianza.⁷

11.2.2 Prueba de Durbin-Watson

Durbin y Watson (1951) 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:

$$DW = \frac{\sum_{t=2}^n (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\hat{\epsilon}^T \hat{\epsilon}}$$

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 Cuadro 11.1, 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.

Cuadro 11.1: Relación del estadístico DW y los casos de autocorrelación

Tipo de problema	ρ	$\hat{\rho}$	valor aproximado del DW
No correlación	$\rho = 0$	$\hat{\rho} \approx 0$	$DW \approx 2$
Correlación positiva	$1 > \rho > 0$	$1 > \hat{\rho} > 0$	$DW < 2$
Correlación negativa	$-1 < \rho < 0$	$-1 < \hat{\rho} < 0$	$DW > 2$

Fuente: Elaboración propia

⁷Una 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 $|RA| > z_{\frac{\alpha}{2}}$.

Naturalmente la regla que se presenta en el Cuadro 11.1 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 Cuadro 11.2.

Cuadro 11.2: Estadístico DW en casos de Autocorrelación

Hipótesis nula (H_0)	H_0 en términos de ρ	Hipótesis alterna (H_A)
No autocorrelación	$\rho = 0$	$\rho \neq 0$
No autocorrelación Positiva	$0 < \rho < 1$	$\rho > 0$
No autocorrelación Negativa	$-1 < \rho < 0$	$\rho < 0$

Fuente: Elaboración propia

Durbin y Watson (1951) 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 como lo discutiremos mas adelante.

Sobre esta prueba es importante destacar varios aspectos:

- El DW no tiene sentido si no hay intercepto (ver Durbin y Watson, 1951).
- El DW depende del supuesto que las X 's sean no estocásticas (ver Durbin y Watson, 1951).
- 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.

11.2.3 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 Durbin (1970) 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:

$$h = \left(1 - \frac{DW}{2}\right) \sqrt{\frac{n}{1 - n \widehat{Var}(\hat{\alpha})}} \quad (11.5)$$

Durbin (1970) 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 11.5, esta prueba no es válida en los casos en que $n \left(\widehat{\text{Var}}(\hat{\alpha}) \right) \geq 1$.

11.2.4 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 cero⁸. Box y Pierce (1970) 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:

$$\gamma_j = \frac{\text{Cov}(\varepsilon_t, \varepsilon_{t-j})}{\sqrt{\text{Var}(\varepsilon_t) \text{Var}(\varepsilon_{t-j})}} = \frac{\text{Cov}(\varepsilon_t, \varepsilon_{t-j})}{\sigma_\varepsilon^2}$$

Y la correspondiente autocorrelación muestral es:

$$\hat{\gamma}_j = \frac{\sum_{t=k+1}^n (\hat{\varepsilon}_t - \bar{\varepsilon})(\hat{\varepsilon}_{t-j} - \bar{\varepsilon})}{\sum_{t=1}^n (\hat{\varepsilon}_t - \bar{\varepsilon})^2} = \frac{\sum_{t=j+1}^n \hat{\varepsilon}_t \hat{\varepsilon}_{t-j}}{\sum_{t=1}^n (\hat{\varepsilon}_t)^2}$$

Box y Pierce (1970) 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 en los primeros 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, Box y Pierce (1970) 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 Ljung y Box (1979) proponen una modificación del estadístico anterior para que presente un mejor comportamiento en muestras pequeñas. El estadístico de Ljung-Box corresponde a:

$$Q' = n(n+2) \sum_{k=1}^s \frac{r_k^2}{n+j}$$

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

⁸Esto explota las características observadas en las Figuras 11.3 y 11.6.

rezagos; por ejemplo, se calculan los correspondientes estadísticos para comprobar las siguientes hipótesis alternas:

$$\begin{aligned} H_0 : \gamma_1 &= 0 \\ H_0 : \gamma_1 = \gamma_2 &= 0 \\ &\vdots \\ H_0 : \gamma_1 = \gamma_2 = \dots \gamma_{n/3} &= 0 \end{aligned}$$

La decisión de si los errores están o no autocorrelacionados se toma teniendo en cuenta las decisiones de cada una de estas pruebas.

11.2.5 Prueba de Breusch-Godfrey

Breusch (1978) y Godfrey (1978) diseñaron una prueba que permite comprobar la hipótesis nula de no autocorrelación versus la alterna de que el error sigue un proceso auto-regresivo 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 estos son explicados por sus valores pasados, pero si no hay autocorrelación 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:⁹

$$LM = (n - s) \times R^2$$

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 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 $(n - s)$.

11.3 Solución a la autocorrelación

Así como en el caso de la heteroscedasticidad (Ver Capítulo 9), 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. Una breve introducción a este método se presenta en el anexo al final del capítulo (Ver sección 11.5).

Otra opción para resolver el problema Incluir variables rezagadas...

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 serán diferentes. Una aproximación muy similar a la de White para solucionar el problema de heteroscedasticidad.

11.3.1 Estimación Consistente en presencia de Autocorrelación de los errores estándar.

De manera similar a la solución de White (1980) para la heteroscedasticidad, Newey y West (1987) sugieren una estimador para la matriz de varianzas y covarianzas.

Recordemos, que en presencia de perturbaciones no esféricas¹⁰ tendremos que:

$$\Psi = \text{Var} [\hat{\beta} | \mathbf{X}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Omega \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$$

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:

$$\hat{\Phi} = \frac{1}{n} \sum_{i,j=1}^n w_{|i-j|} \hat{\mathbf{V}}_i \hat{\mathbf{V}}_j^T$$

donde $w = [w_0 \dots w_{n-1}]^T$ corresponde a un vector de pesos.

Newey-West (1987) sugieren

$$w_\ell = 1 - \frac{\ell}{L+1}$$

¹⁰Esta es otra forma de decir que no se cumplen los supuestos del Teorema de Gauss-Markov.

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 esto 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 MCO sigue siendo más grande que por ejemplo los del estimador GLS.

De manera similar al caso de la heteroscedasticidad, el estimador de Newey-West genera varianzas de los estimadores que son relativamente más pequeñas y por tanto los t calculados son 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:

$$W = (\mathbf{R}\hat{\beta})^T \left[\mathbf{R} \left(\text{Est.Asy.Var} [\hat{\beta}] \right) \mathbf{R}^T \right]^{-1} (\mathbf{R}\hat{\beta})$$

Por otro lado, así como en el caso de la corrección H.A.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 $z = 6\pi/5 \cdot \ell/B$. Esta aproximación es conocida como H.A.C de kernel.
- Lumley and Heagerty (1999) sugieren otra forma de pesar los datos que es conocida como la aproximación H.A.C de **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.

Finalmente, es importante anotar que otra forma común para solucionar el problema de autocorrelación en una regresión es emplear la variable dependiente como variable explicativa, pero rezagada unos periodos. Es decir, de encontrar autocorrelación se acostumbra incluir la variable y rezagada y_{t-1} como variable explicativa¹¹. Esto implica probar nuevamente si existe o no autocorrelación en los nuevos residuales. Si el problema de autocorrelación persiste, se sigue incluyendo la variable dependiente rezagada mas periodos hasta que el problema de autocorrelación desaparezca. Esta aproximación implica diferentes iteraciones y estar probando la autocorrelación en cada paso.

En la actualidad los científicos de datos prefieren la aproximación de emplear H.A.C para hacer analítica diagnóstica . Si se de desea hacer analítica predictiva para hacer proyecciones es más común emplear técnicas de series de tiempo que emplean esta lógica de incluir las autocorrelaciones como los modelos ARIMA o ARIMAX. Estas técnicas no las cubriremos en este libro pero puedes consultar Alonso y Hoyos (2021) para una introducción a los métodos de proyección.

¹¹Esto se puede hacer en R empleando la función **tslm()** del paquete *forecast* (Hyndman y Khandakar, 2008)

11.4 Práctica en R: Explicando los rendimientos de una acción (continuación)

En el Capítulo 3 construimos un modelo para responder una pregunta de negocio que tenían en el área financiera de una organización. La pregunta era ¿Cómo está relacionado el rendimiento de la acción grupo SURA con el rendimiento de aquellas acciones en la que ya tenemos inversiones? Las acciones que ya se tenían en el portafolio eran: ECOPETROL, NUTRESA, EXITO, ISA, GRUPOAVAL, CONCONCRETO, VALOREM y OCCIDENTE. Esta pregunta nació de la necesidad de diversificar el riesgo del portafolio. Ahora, tras unas semanas, contamos con una variable mas que puede ser útil en el análisis la tasa de depósitos a término fijo a 90 días (DTF). Esta variable es común que se incluya en este tipo de modelos como una variable que captura el rendimiento de un activo libre de riesgo, poner el dinero en un CDT. Esta variable está disponible en el archivo *DataCDTs.xlsx* y la adicionaremos a nuestro modelo. Entonces el modelo que estimaremos será:

$$\begin{aligned} \text{GRUPOSURA}_t = & \beta_1 + \beta_2 \text{ECOPETROL}_t + \beta_3 \text{NUTRESA}_t \\ & + \beta_4 \text{EXITO}_t + \beta_5 \text{ISA}_t + \beta_7 \text{GRUPOAVAL}_t \\ & + \beta_8 \text{CONCONCRETO}_t + \beta_9 \text{VALOREM}_t \\ & + \beta_{10} \text{OCCIDENTE}_t + \beta_{11} \text{DTF}_t + \varepsilon_t \end{aligned}$$

Nuestra misión en esta ocasión es terminar el análisis y responder la pregunta de negocio incluyendo la nueva variable, pero antes debemos estar seguros que el modelo estimado está libre de autocorrelación.

11.4.1 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 objeto). Los datos de los rendimientos se encuentran en un archivo *RetornosDiarios.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.

Los datos de la *DTF* la podemos cargar con la función **read_excel()** del paquete *readxl* (Wickham y Bryan, 2019). 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. Esto se puede hacer mediante el argumento **col_types**. El otro argumento que requiere esta función es el nombre del archivo que contiene los datos y su correspondiente ruta. En nuestro caso, 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)
# carga los datos
DTF <- read_excel("../Data/DataCDTs.xlsx", col_types = c("date",
  "numeric"))
head(DTF, 2)

## # A tibble: 2 x 2
##   `Fecha(dd/mm/aaaa)` DTF90dias
##   <dttm>              <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"

# renombrando la primera variable
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**. Ésta es una clase que permite manipular fácilmente objetos de series de tiempo. Procedamos a cambiar dicha clase para hacer más fácil la unión de las dos bases de datos.

Carguemos el paquete **xts** (Ryan y Ulrich, 2020) . En dicho paquete tenemos una función con el mismo nombre del paquete (**xts()**) que permite crear objetos de la clase **xts**. Esta función requiere dos argumentos. El primero es el objeto que se quiere transformar y el segundo (**order.by**) es un vector que contiene las fechas que se le asignará al objeto. Estas fechas deben ser de un formato fecha. Por ahora, tenemos el objeto para ser convertido en objeto **xts** y nos falta el vector con las correspondientes fechas. Esto lo podemos crear con la función **as.Date()** del paquete base de R.

Creemos las fechas de la variable *Fecha* del objeto DTF empleando la función **as.Date()** y luego procedamos a crear un objeto con el nombre DTF90dias que sea de clase **xts**.

```

# Carga librería
library(xts)
# se crea objeto con fechas
DTF$Fecha <- as.Date(DTF$Fecha)
# se verifica la clase de la fecha
class(DTF$Fecha)

## [1] "Date"

```

```
# se crea el nuevo objeto de clase xts
DTF90dias <- xts(DTF$DTF90dias, order.by = DTF$Fecha)
# se verifica la clase del nuevo objeto
class(DTF90dias)

## [1] "xts" "zoo"
```

Antes de unir los dos objetos podemos constatar que ambos objetos tienen la misma periodicidad y cubren el mismo periodo. La función **periodicity()** del paquete *xts* nos muestra la periodicidad de un objeto de serie de tiempo.

```
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 la función **merge()** del paquete *xts* (es decir, **merge.xts()**). Esta función permite pegar dos objetos de clase **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”¹². Si **join= “inner”** se construirá una nueva base de datos únicamente con las filas (fechas) que están 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 llenará esa información con un “NA”. De manera similar, Si **join= “right”**, el nuevo objeto tendrá las fechas del segundo objeto.

En nuestro caso, puedes 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 sí se recoge información para la DTF. Así, que dado que 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,

```
datos.ejeauto <- merge(retornos.diarios, DTF90dias, join = "left")
head(datos.ejeauto, 3)

##           GRUPOSURA ECOPETROL      NUTRESA      EXITO       ISA
## 2012-01-02  1.2779727 -0.3565066 -0.9216655  2.02184181 -1.983834
```

¹²La función permite cambiar como se rellena los datos faltantes empleando el argumento **fill**. Por defecto **fill = NA**.

```
## 2012-01-03 2.5079684 2.0036027 -0.2781643 0.07695268 1.626052
## 2012-01-04 0.4324999 1.0446990 -0.5586607 1.07116556 2.478003
## GRUPOAVAL CONCONCRET VALOREM OCCIDENTE DTF90dias
## 2012-01-02 0.0000000 0 0.000000 0 0.05131078
## 2012-01-03 -2.4292693 0 12.583905 0 0.05576149
## 2012-01-04 -0.4106782 0 1.342302 0 0.04769692
```

Ahora ya tenemos un objeto con la base de datos.

11.4.2 Residuales del modelo y análisis gráfico de los residuales

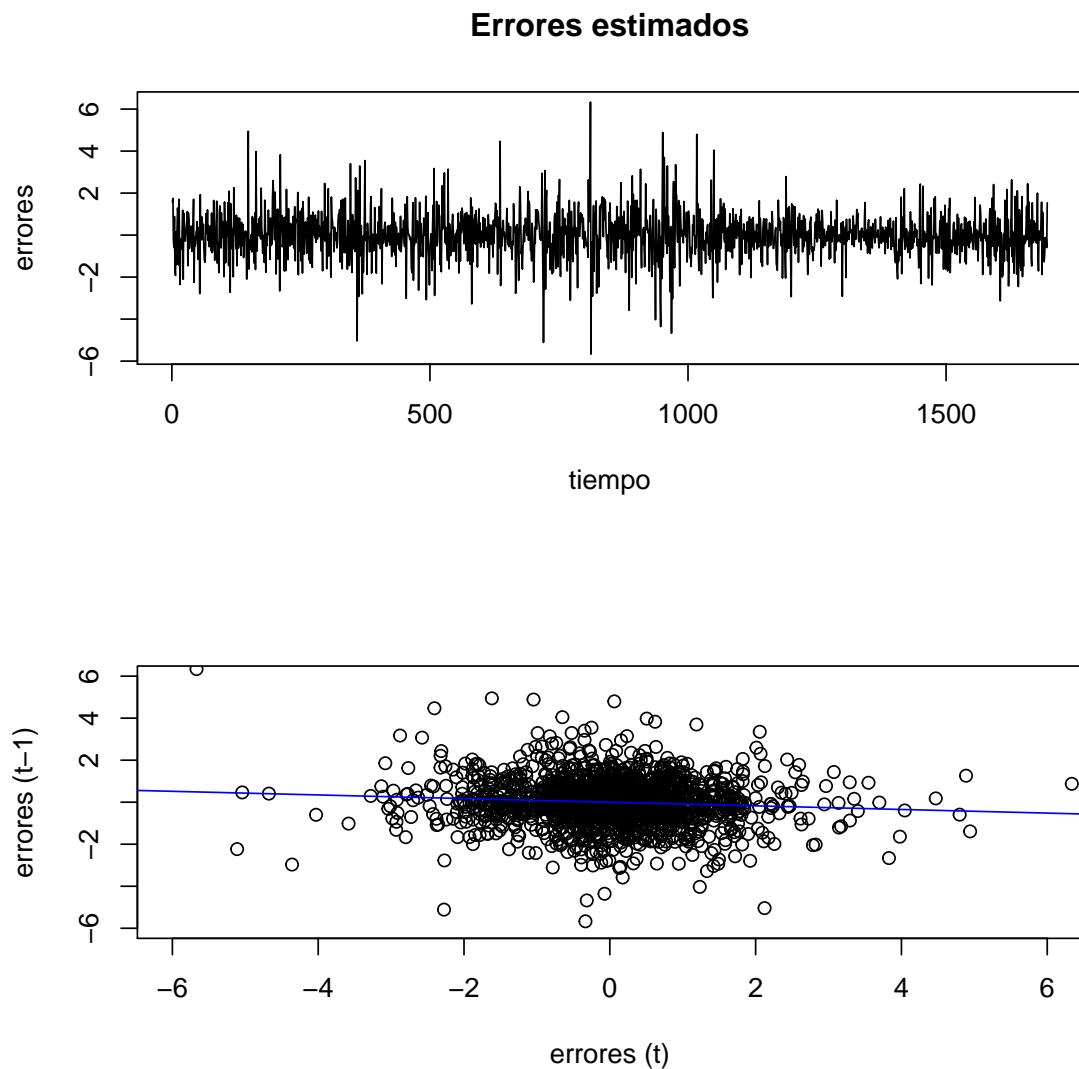
Ya podemos correr el modelo y examinar los respectivos residuos¹³. Esto lo podemos hacer con un gráfico de líneas de los residuos en función del tiempo y uno de dispersión de los residuos en el periodo actual versus los residuos rezagados¹⁴.

```
# estimación del modelo
modelo1 <- lm(GRUPOSURA ~ ., datos.ejeauto)
# se extraen los residuales
e <- residuals(modelo1)

# creamos los gráficos
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")
```

¹³En el Capítulo 9 vimos que podemos extraer los residuales de un objeto LM con la función **resid()** del paquete central de R.

¹⁴Recuerda que el término 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()** del paquete *xts*



El primer gráfico muestra unos 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 sí nos permiten tener una intuición de lo que está ocurriendo con los residuos. A continuación se presentan las pruebas de autocorrelación discutidas anteriormente.

11.4.3 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.

11.4.4 Prueba de Rachas

Para realizar la prueba de rachas, podemos emplear la función **runs.test()** del paquete *tseries* (Trapletti y Hornik, 2019). Los argumentos más importantes de esta función son

$$\text{runs.test}(x, \text{alternative})$$

donde:

- **x**: objeto que contiene una variable de clase **factor** que muestre si el error es positivo o negativo
- **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 (más cambios de signos que los esperados).

Así, nuestro primer paso para efectuar esta prueba es convertir los residuos estimados en una variable de clase **factor** que muestre si el signo del residual es positivo o no para cada periodo.

```
# se crea la variable tipo factor
signo.error <- factor(e > 0)
# chequeo de la variable creada
head(signo.error, 3)

## 2012-01-02 2012-01-03 2012-01-04
##      TRUE      TRUE     FALSE
## Levels: FALSE TRUE

# se carga la librería
library(tseries)
# prueba de rachas dos colas
runs.test(signo.error)

##
## Runs Test
##
## data: signo.error
## Standard Normal = 2.7324, p-value = 0.006288
## alternative hypothesis: two.sided

# la alterna es que existe autocorrelación positiva
runs.test(signo.error, alternative = "less")
```

```

## 
## Runs Test
##
## data: signo.error
## Standard Normal = 2.7324, p-value = 0.9969
## alternative hypothesis: less

# la alterna es que existe autocorrelación negativa
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 negativa (en favor de la alterna de autocorrelación positiva) 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).

11.4.5 Prueba de Durbin-Watson

Esta prueba se puede implementar empleando la función **dwtest()** del paquete *AER* (Kleiber y Zeileis, 2008). Los argumentos más importantes de esta función son

```
dwtest(formula, alternative = c("greater", "two.sided", "less"))
```

donde:

- **formula:** objeto de clase *lm* que contenga el modelo al cual se le quiere efectuar la prueba, este argumento es obligatorio.
- **alternative:** 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:

```
# se carga la librería
library(AER)

# prueba de Durbin-Watson dos colas
dwtest(modelo1, alternative = "two.sided")

## 
## Durbin-Watson test
##
## data: modelo1
## DW = 2.1703, p-value = 0.0004734
## alternative hypothesis: true autocorrelation is not 0

# la alterna es que existe autocorrelación positiva
dwtest(modelo1, alternative = "greater")

## 
## Durbin-Watson test
##
## data: modelo1
## DW = 2.1703, p-value = 0.9998
## alternative hypothesis: true autocorrelation is greater than 0

# la alterna es que existe autocorrelación negativa
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 en este modelo con la variable dependiente rezagada 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*(Bansal, 2021).

11.4.6 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 los siguientes tres argumentos importantes,

$$\text{Box.test}(x, \text{lag} = 1, \text{type} = c(\text{"Box-Pierce"}, \text{"Ljung-Box"}))$$

donde:

- **x**: el vector al que se le quiere hacer la prueba.
- **lag**: el número de rezagos (*lag*) para incluir en la hipótesis nula y alterna. Por defecto `lag = 1`.
- **type**: el tipo de prueba. Las opciones son `type = "Box-Pierce"` para realizar la prueba de Box-Pierce (la opción por defecto de la función) y `type = "Ljung-Box"` para el de la prueba de Ljung-Box.

Entonces para probar la hipótesis $H_0 : \gamma_1 = 0$ con la prueba de Box-Pierce podemos emplear el siguiente código

```
# Prueba de Box-Pierce para un rezago
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 pruebas podemos crear una función para construir una tabla con todas las pruebas que se deseen. A continuación se presenta la función **tabla.Box.Pierce()**.

```
# se crea función
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 obtener los resultados que se presentan el Cuadro 11.3.

Cuadro 11.3: Prueba de Box-Pierce de los errores para los primeros 20 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 concluir que las autocorrelaciones de los errores no son cero. Así podemos concluir que existe autocorrelación.

Puede generar fácilmente el Cuadro 11.4 que contienen los resultados de la prueba de Ljung-Box

aplicada a los mismos residuos.

Cuadro 11.4: Prueba de Ljung-Box de los errores para los primeros 20 rezagos

Rezagos	Ljung-Box	p-valor
1.00	12.66	0.00
2.00	17.57	0.00
3.00	17.66	0.00
4.00	18.56	0.00
5.00	22.82	0.00
6.00	25.25	0.00
7.00	26.42	0.00
8.00	28.92	0.00
9.00	28.99	0.00
10.00	30.01	0.00
11.00	30.17	0.00
12.00	31.20	0.00
13.00	31.20	0.00
14.00	37.99	0.00
15.00	37.99	0.00
16.00	39.49	0.00
17.00	42.62	0.00
18.00	44.13	0.00
19.00	44.15	0.00
20.00	44.20	0.00

No es sorprendente que los resultados de esta prueba sean los mismos 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.

11.4.7 Prueba de Breusch-Godfrey

Esta prueba se puede realizar empleando la función **bgtest()** del paquete *lmtest*(Zeileis y Hothorn, 2002). Similar a las anteriores pruebas, esta función tiene dos argumentos. El primero es el objeto de clase **Im** 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

```
# se carga la libreria
library(lmtest)
# prueba reusch-Godfrey para AR(1)
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 que 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.

```
# se crea función
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 Cuadro 11.5.

Cuadro 11.5: 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.

11.4.8 Solución al problema de autocorrelación 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* (Zeileis, 2004) y las siguientes funciones:

- *NeweyWest()*: para obtener la corrección de Newey y West (1987).
- *kernHAC()*: para obtener la corrección de Andrews (1991).
- *weave()*: para obtener la corrección de Lumley y Heagerty (1999).

Las tres funciones tienen como argumento el objeto de clase **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 y West (1987) podemos emplear la siguiente linea de código.

```
# se carga la libreria
library(sandwich)
# Newey - West
NeweyWest(modelo1)

##              (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
## CONCONCRET 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
## CONCONCRET  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 más bien los respectivos t individuales y sus correspondientes valores p, podemos emplear la función del paquete *lmttest* para realizar las pruebas individuales. Esta función ya la habíamos estudiado en el Capítulo 9. Por ejemplo:

```

# pruebas individuales
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 ***
## CONCONCRET 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 CONCONCRETO, OCCIDENTE y la DTF a 90 días no son significativos individualmente.

Ahora hagamos lo mismo para las correcciones de Andrews (1991) y Lumley y Heagerty (1999).

```
coefest(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 ***
## CONCONCRET  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
```

```
coefest(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 ***
## CONCONCRET  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: CONCONCRETO, 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, datos.ejeauto)

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 + CONCONCRET +
##           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 + CONCONCRET +
##           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 + CONCONCRET +
##           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 el sin restringir. Así, podemos afirmar que la DTF a 90 días, los rendimientos de CONCONCRETO y OCCIDENTE no afectan el rendimiento de la acción de Suramericana. Los resultados finales se presentan en el Cuadro 11.6.

Cuadro 11.6: Modelo estimado por MCO y correcciones H.A.C.

	<i>Dependent variable:</i>		
	GRUPOSURA		
	MCO	NW	Kenel
	(1)	(2)	(3)
ECOPETROL	0.121*** (0.014)	0.121*** (0.018)	0.121*** (0.018)
NUTRESA	0.141*** (0.028)	0.141*** (0.035)	0.141*** (0.035)
EXITO	0.123*** (0.018)	0.123*** (0.028)	0.123*** (0.028)
ISA	0.188*** (0.019)	0.188*** (0.027)	0.188*** (0.025)
GRUPOAVAL	0.084*** (0.020)	0.084*** (0.017)	0.084*** (0.020)
CONCONCRET	0.021 (0.015)	0.021	0.021
VALOREM	0.036** (0.014)	0.036** (0.015)	0.036** (0.016)
OCCIDENTE	0.031 (0.027)	0.031	0.031
DTF90dias	-1.696 (2.725)	-1.696	-1.696
Constant	0.092 (0.141)	0.092*** (0.022)	0.092*** (0.025)
F Statistic (df = 6; 1695)	99.025***	49.099***	62.358***
Observations	1,696	1,696	1,696
R ²	0.262	0.262	0.262
Adjusted R ²	0.258	0.258	0.258
Residual Std. Error (df = 1686)	1.110	1.110	1.110

Note:

*p<0.1; **p<0.05; ***p<0.01

Ejercicios

11.1 Diseñe un experimento de Monte Carlo similar al que se realizó en el Capítulo 9 en la Introducción. Suponga que existe una autocorrelación en el error que sigue un proceso AR(1) con $\rho = 0,7$.

11.2 Ahora continuando con el mismo *DGP* diseñe un experimento de Monte Carlo que nos permita estudiar cuál es la proporción de veces que se rechaza la hipótesis nula de las pruebas individuales para las dos pendientes con los estimadores MCO sin autocorrelación, los estimadores MCO con autocorrelación y la corrección NeweyWest en presencia de autocorrelación. Realice el experimento para una muestra de tamaño 500, 100, 50 y 20. ¿Qué puedes concluir?

11.3 Un científico de datos es contratado por el área de gestión humana de una de las principales empresas que venden automóviles del fabricante WMB en todo el mundo. El interés del gerente del área es entender si la cantidad de vendedores depende de las unidades fabricadas (en miles). Para esto cuenta con datos mensuales de 5 años que se encuentran en el archivo *autoEmployee.txt*. Estime el modelo y reporte sus resultados en una tabla. Efectúe el análisis gráfico de los errores estimados. ¿Qué tipo de problema puede intuir a partir de este análisis? Explique.

Adicionalmente, realice las pruebas que considere necesarias para determinar la existencia o no de un problema de autocorrelación en el modelo. De ser necesario corrija el problema y saque sus conclusiones.

■

11.5 Anexos

11.5.1 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,

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

Donde $E[\boldsymbol{\varepsilon}_t] = 0$, $Var[\boldsymbol{\varepsilon}_t] = \sigma^2_{\varepsilon}$ y $\boldsymbol{\varepsilon}_t = \rho \boldsymbol{\varepsilon}_{t-1} + v_t$, $\forall t$. Ahora determinemos si $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ sigue siendo insesgado o no. Así,

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

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

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

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

11.5.2 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:

$$Var[\varepsilon] = E[\varepsilon^T \varepsilon] = \Omega = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho & 1 \end{bmatrix} \quad (11.6)$$

Ahora podemos calcular la varianza de los estimadores MCO. Es decir,

$$Var[\hat{\beta}] = Var[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}] \quad (11.7)$$

Por tanto tendremos que

$$Var[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Var[\mathbf{y}] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (11.8)$$

$$Var[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Var[\varepsilon] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (11.9)$$

$$Var[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Omega \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (11.10)$$

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.

11.5.3 Límites de la prueba de Rachas para muestras pequeñas

Cuadro 11.7: Límite inferior de la prueba de Rachas. Nivel de confianza del 95 %

		Número de errores negativos (N_-)													
		2	2	2	2	2	2	2	2	2	2	2	2	2	2
		3	2	2	2	2	2	2	2	2	3	3	3	3	3
		4	2	2	2	3	3	3	3	3	3	4	4	4	4
		5	2	2	3	3	3	3	4	4	4	4	4	5	5
		6	2	2	3	3	3	4	4	4	5	5	5	5	6
		7	2	2	3	3	3	4	4	5	5	5	6	6	6
		8	2	3	3	3	4	4	5	5	6	6	6	7	7
		9	2	3	3	4	4	5	5	6	6	7	7	8	8
N_+	10	2	3	3	4	5	5	5	6	6	7	7	7	8	8
	11	2	3	4	4	5	5	6	6	7	7	7	8	8	9
	12	2	2	3	4	4	5	6	6	7	7	8	8	9	10
	13	2	2	3	4	5	5	6	6	7	7	8	9	10	10
	14	2	2	3	4	5	5	6	7	7	8	8	9	10	11
	15	2	3	3	4	5	6	6	7	7	8	9	9	10	11
	16	2	3	4	4	5	6	6	7	8	8	9	10	11	12
	17	2	3	4	4	5	6	7	7	8	9	9	10	11	12
	18	2	3	4	5	5	6	7	8	8	9	9	10	11	12
	19	2	3	4	5	6	6	7	8	8	9	10	10	11	13
	20	2	3	4	5	6	6	7	8	9	9	10	10	11	13

Fuente: Swed y Eisenhart (1943)

Cuadro 11.8: Límite superior de la prueba de Rachas. Nivel de confianza del 95 %

		Número de errores negativos (N_-)																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
N_+	2																			
	3																			
	4			9	9															
	5		9	10	10	11	11													
	6		9	10	11	12	12	13	13	13	13	13								
	7			11	12	13	13	14	14	14	14	14	15	15	15	15				
	8				11	12	13	14	14	15	15	15	16	16	16	16	17	17	17	
	9					13	14	14	15	16	16	16	17	17	18	18	18	18	18	
	10						13	14	15	16	16	17	17	18	18	18	19	19	20	
	11							13	14	15	16	17	17	18	19	19	20	20	21	
	12								13	14	15	16	17	18	19	20	20	21	22	
	13									15	16	17	18	19	19	20	20	21	23	
	14										15	16	17	18	19	20	21	22	23	
	15											15	16	18	18	19	20	21	24	
	16												17	18	19	20	21	22	25	
	17													17	18	19	20	21	26	
	18														17	18	19	20	27	
	19															17	18	20	27	
	20																17	18	28	

Fuente: Swed y Eisenhart, 1943

11.5.4 Solución por el método de diferencias generalizadas

Si conocemos la naturaleza de la autocorrelación entonces podemos usar una transformación de la muestra del modelo original para construir una muestra sin este problema. Este método se conoce como transformación de diferencias generalizadas.

Es decir, partiendo del modelo 11.5 con errores $AR(1)$, se le puede restar el mismo modelo 11.5 rezagado un período y multiplicado por la correlación. De tal manera que se obtiene:

$$\begin{aligned} Y_t - \rho Y_{t-1} = & \beta_1 (1 - \rho) + \beta_2 (X_{2t} - \rho X_{2t-1}) + \beta_3 (X_{3t} - \rho X_{3t-1}) + \dots \\ & + \beta_k (X_{kt} - \rho X_{kt-1}) + \varepsilon_t - \rho \varepsilon_{t-1} \end{aligned}$$

Reparametrizando tenemos:

$$Y_t^* = \beta_1 (1 - \rho) + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \dots + \beta_k X_{kt}^* + v_t$$

donde

$$Y_t^* = Y_t - \rho Y_{t-1}, \quad X_{2t}^* = X_{2t} - \rho X_{2t-1} \quad \text{y} \quad X_{3t}^* = X_{3t} - \rho X_{3t-1} \quad (11.11)$$

Así, el modelo transformado (11.11) ya no tiene problemas de autocorrelación. Esto es emplear el método de Mínimos Cuadrados Generalizados en su versión especial denominado Método de diferencias generalizadas .

Sin embargo, en la práctica es imposible que conozcamos la naturaleza del problema de autocorrelación con certeza y mucho menos el valor de ρ . Durbin, 1960 plantea la solución a este problema,

método que tomó el nombre de Método de Durbin. El método de Durbin, 1960 permite implementar el método de diferencias generalizadas al estimar en un primer paso el valor de ρ . Por eso se le conoce como un método de Mínimos Cuadrados Factibles . Este método implica los siguientes tres pasos:

1. Corra la regresión de la variable dependiente en función de las variables explicativas del modelo original, además incluya las mismas variables independientes rezagadas un período y la variable dependiente rezagada un período. Es decir:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \dots + \beta_k X_{kt} + \beta_{k+1} X_{2t-1} \\ + \beta_{k+2} X_{3t-1} + \dots + \beta_{k+(k-1)} X_{kt-1} + \rho Y_{t-1} + \varepsilon_t$$

2. De la anterior estimación se obtiene el valor estimado de ρ . A partir de ese valor estimado, realice las siguientes transformaciones:

$$Y_t^* = Y_t - \hat{\rho} Y_{t-1}, \quad X_{2t}^* = X_{2t} - \hat{\rho} X_{2t-1} \quad \text{y} \quad X_{3t}^* = X_{3t} - \hat{\rho} X_{3t-1}$$

3. Finalmente estime el siguiente modelo:

$$Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \dots + \beta_k X_{kt}^* + v_t,$$

$$\text{donde } \beta_1^* = \beta_1 (1 - \hat{\rho})$$

Una vez se estima el modelo con el método de Durbin, a los nuevos residuales se les debe hacer las pruebas para asegurarnos que el problema fue solucionado. Si la autocorrelación en los residuales no es de orden uno, es muy probable que esta solución no funcione. En la práctica es aconsejable para los científicos de datos emplear una solución H.A.C. dados los grandes volúmenes de observaciones que se emplean esta solución provee errores estándar insesgados (robustos) y se garantiza una solución aceptable al problema.

Parte IV

Modelo de regresión y la analítica predictiva.

12 . Predicciones con datos de corte transversal

Diseñado por Freepik

Objetivos del capítulo

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

- Explicar en sus propias palabras la diferencia entre un pronóstico y una predicción.
- Explicar en sus propias palabras qué es la validación cruzada, para que sirve y cuándo se debe emplear.
- Efectuar en R una validación cruzada por los métodos de retención, LOOCV y k-folds para seleccionar un modelo de regresión para hacer analítica predictiva (en muestras de corte transversal).
- Construir intervalos de confianza para predicciones.

12.1 Introducción

Los modelos de regresión múltiple pueden emplearse para hacer analítica predictiva. La analítica predictiva busca responder la pregunta: ¿qué es posible que ocurra? Con datos ya existentes se estima un modelo que permita predecir datos que aún no se tienen o no han ocurrido. Es importante resaltar que la creación de predicciones no necesariamente implica predecir lo que ocurrirá en el **futuro**. De hecho, la creación de pronósticos (forecasting en inglés) es un área de la predicción en la que se realizan predicciones sobre el futuro, basándonos en datos de series de tiempo. La única diferencia entre la predicción y los pronósticos es que en esta última se considera la dimensión temporal. La analítica predictiva tiene como intención generar predicciones de variables cuantitativas.

El modelo de regresión múltiple puede emplearse tanto para hacer predicciones como pronósticos. Para la creación de pronósticos es necesario la estimación de un modelo empleando datos de serie de tiempo. Cuando empleamos muestras de corte trasversal, podemos realizar predicciones pero no pronósticos.

Por ejemplo, si se emplea una muestra de muchos individuos en el mismo periodo (muestra de corte trasversal) para encontrar las variables asociadas a la cantidad de unidades que compra un cliente, el modelo podría ser empleado para responder la pregunta ¿cuánto compraría un nuevo cliente con determinadas características?

Si el modelo de regresión múltiple se estima con una serie de tiempo (se observa un objeto de estudio periodo tras periodo), éste podrá ser empleado para hacer pronósticos. Por ejemplo, si se cuenta con las ventas mensuales para muchos periodos y el modelo encuentra que variables están asociadas a estas ventas mes tras mes, el modelo podría responder la pregunta que ocurrirá en el futuro con las ventas. Con este tipo de modelos podemos responder preguntas como: ¿cuánto es lo más probable que venda un producto el próximo año?

En este capítulo nos concentraremos en la construcción de predicciones y no de pronósticos. Es decir, nos concentraremos en la analítica predictiva que emplea modelos de corte trasversal.

Hasta ahora hemos discutido la estimación de modelos de regresión múltiple y el diagnóstico y solución de los problemas que aparecen tras la violación de los supuestos del Teorema de Gauss-Markov . Hemos discutido la interpretación de los coeficientes que permiten determinar el efecto de una variable independiente sobre el comportamiento de la variable dependiente, aislando el efecto de las otras variables. Esto corresponde a un análisis de analítica diagnóstica. . Por eso, hasta ahora no se había discutido el poder predictivo de los modelos estimados. Los modelos se han estimado con el total de la muestra y esto es lo común si se desea hacer analítica diagnóstica. Es decir, hemos empleado las herramientas de la estadística para responder la pregunta ¿qué tan bueno es el modelo para explicar la muestra que tenemos?

Para responder dicha pregunta, hemos empleado herramientas como pruebas de hipótesis de modelos anidados y no anidados, el R^2 , el R^2 ajustado, SBC y AIC. Cómo lo vimos, no existe una única forma de determinar la bondad de ajuste de un modelo a una determinada muestra. No obstante las herramientas estudiadas si tienen en común que tienen como objetivo medir que tan bien se ajusta el

modelo estimado a la muestra que empleamos para su estimación (entrenamiento).

Emplear estas herramientas nos llevan a encontrar modelos que explican lo mejor posible la muestra bajo estudio, pero no necesariamente nuevas muestras que aparezcan. Es decir, es posible que tengamos un modelo muy bueno para explicar la muestra (analítica diagnóstica) pero no necesariamente para hacer analítica predictiva. Esto se conoce como el problema de *overfitting* (sobreajuste).

Cuando los científicos de datos están interesados en responder una pregunta de negocio que involucra emplear analítica predictiva, será necesario evaluar el poder predictivo del modelo. En esta tarea, queremos evitar el *overfitting* porque podemos estar dando demasiado poder predictivo a peculiaridades específicas de la muestra que se empleó para estimar el modelo. Pero al mismo tiempo, queremos evitar tener un modelo con bajo ajuste a la muestra (*underfitting* (infraajuste)) porque estaríamos ignorando patrones en la muestra útiles para determinar el comportamiento de un nuevo individuo.

De esta manera, necesitaremos otras herramientas para responder la pregunta ¿qué tan bueno es el modelo para predecir? Al proceso de encontrar un modelo que responda esta pregunta empleando parte de la muestra original, se le conoce como **validación cruzada del modelo** (*cross-validation* en inglés). En general, para realizar la valoración del poder predictivo de varios modelos candidatos empleando la muestra disponible será necesario contar con:

1. una muestra de evaluación que sea diferente a la muestra de estimación o también conocida como la muestra de entrenamiento y
2. una métrica que permita determinar que tan cerca se encuentran las predicciones del valor realmente observado en la muestra de evaluación.

Este capítulo discutiremos estos elementos y cómo realizar la validación de modelos empleando diferentes aproximaciones.

12.2 Estrategias para la validación cruzada de modelos

Para evitar *overfitting* (sobreajuste) una buena práctica es emplear muestras diferentes para estimar y evaluar la capacidad predictiva de este. En general, cualquier técnica de validación cruzada de modelos implicará dividir la muestra en una muestra de evaluación que sea diferente a la muestra de estimación o también conocida como la muestra de entrenamiento . A esta práctica se le conoce como **validación cruzada** o en inglés *cross-validation*¹. En las siguientes tres subsecciones se discuten las técnicas de validación cruzada más empleadas.

¹Noten que esta técnica solo tiene sentido en datos de corte transversal, donde el orden de los datos no es importante. Para muestras de series de tiempo el orden es importante y seleccionar una muestra de manera aleatoria estaría negando una de las características mas importantes en las series de tiempo. Es por esto que cuando realizamos pronósticos emplearemos otras técnicas para validar los modelos. Para mayor detalle ver Alonso y Hoyos (2021)

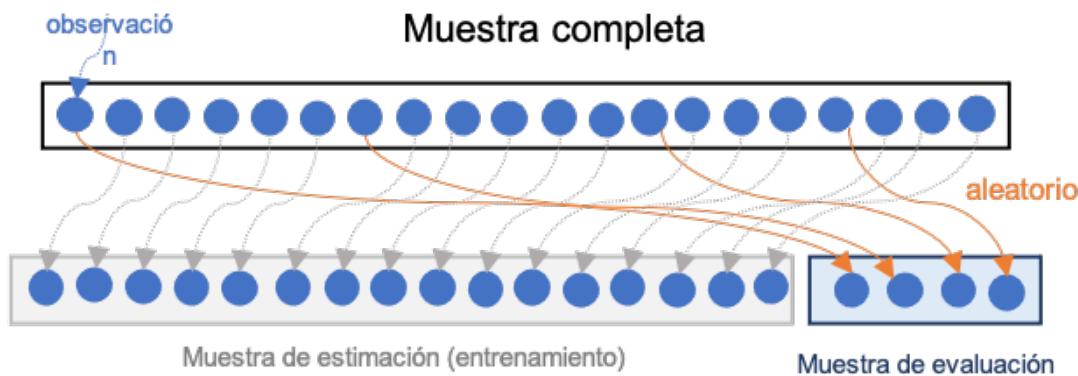
12.2.1 Método de retención

La técnica de validación cruzada más sencilla implica seleccionar aleatoriamente unas observaciones de la muestra para hacer la estimación (muestra de estimación o muestra de entrenamiento) y otra para evaluar el comportamiento del modelo para predecir (muestra de evaluación). Esta técnica es conocida como el **método de retención** o *holdout method*.

En este caso, una parte relativamente pequeña de la muestra es seleccionada al azar² como muestra de evaluación y la muestra restante es la de evaluación. Es común que se emplee el 80 % de la muestra para la estimación y el 20 % restante para realizar la evaluación. En la Figura 12.1 se presenta un diagrama de esta aproximación. Con la muestra de evaluación se comparan los diferentes modelos de regresión candidatos a ser el mejor modelo para predecir las observaciones. La comparación implica calcular diferentes métricas que discutiremos en la siguiente sección de este capítulo.

Una desventaja de esta aproximación es que la selección del mejor modelo para predecir podría estar determinada por el azar, pues la muestra seleccionada para la evaluación es totalmente aleatoria. Así, si se replicara el ejercicio con otra muestra, el mejor modelo seleccionado podría ser diferente.

Figura 12.1. Diagrama del Método de retención para la evaluación cruzada de modelos



Fuente: Elaboración propia

Antes de continuar con otras aproximaciones de validación cruzada de modelos, definamos unos términos importantes en esta literatura. El error del modelo en la muestra de estimación se conoce como el **error de entrenamiento**. El error del modelo en la muestra de evaluación se conoce como el **Error de prueba**.

12.2.2 Método LOOCV

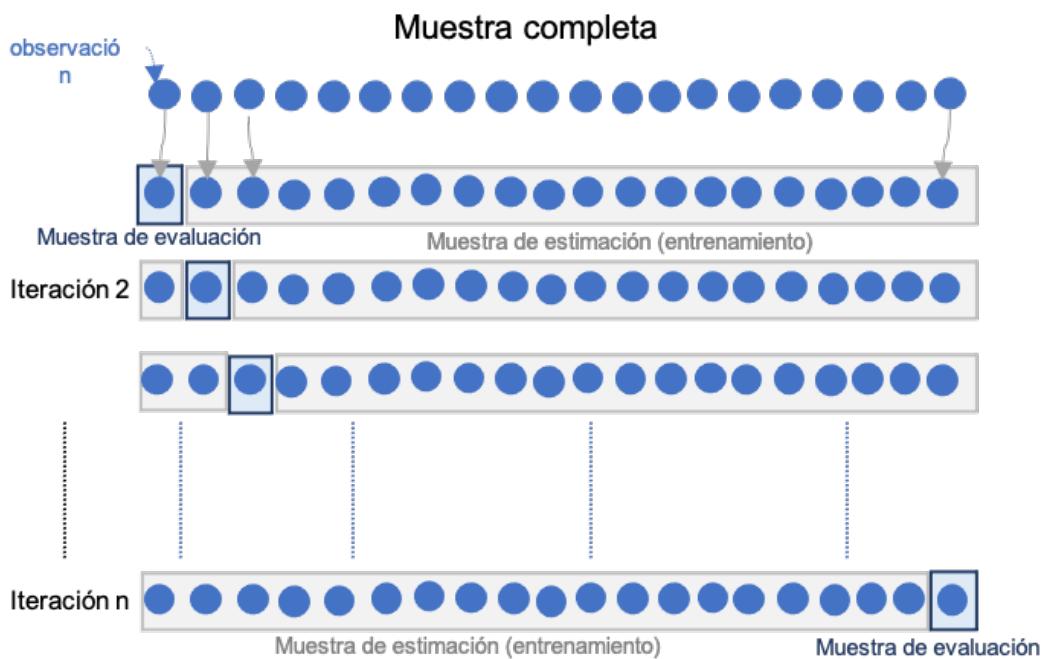
Otro método de valoración cruzada comúnmente empleado es el denominado LOOCV (por la sigla en inglés del término *Leave one out Cross-validation*). En este método, a diferencia del método

²por muestreo aleatorio sin reposición.

de retención, no se emplea una única muestra de evaluación elegida al azar. Este método implica realizar la estimación y la valoración del modelo en diferentes muestras.

En este caso, se fija una observación en la muestra de evaluación y se estima el modelo con el resto de observaciones ($n - 1$). Este proceso se repite hasta que todas las observaciones de la muestra han sido empleadas en una muestra de evaluación. En la Figura 12.2 se presenta un diagrama de esta aproximación. Los modelos candidatos se comparan empleando el promedio de las métricas deseadas para evaluar las predicciones para cada una de las n muestras de evaluación.

Figura 12.2. Diagrama del Método de validación cruzada de k iteraciones para la evaluación de modelos



Fuente: Elaboración propia

A priori el método LOOCV parecería ser el ideal al emplear todas las observaciones como muestra de evaluación. Esto claramente quitaría el problema de la aleatoriedad que genera el método de retención al momento de seleccionar el mejor modelo para predecir. No obstante, este método puede ser costoso computacionalmente pues requiere estimar el modelo n veces.

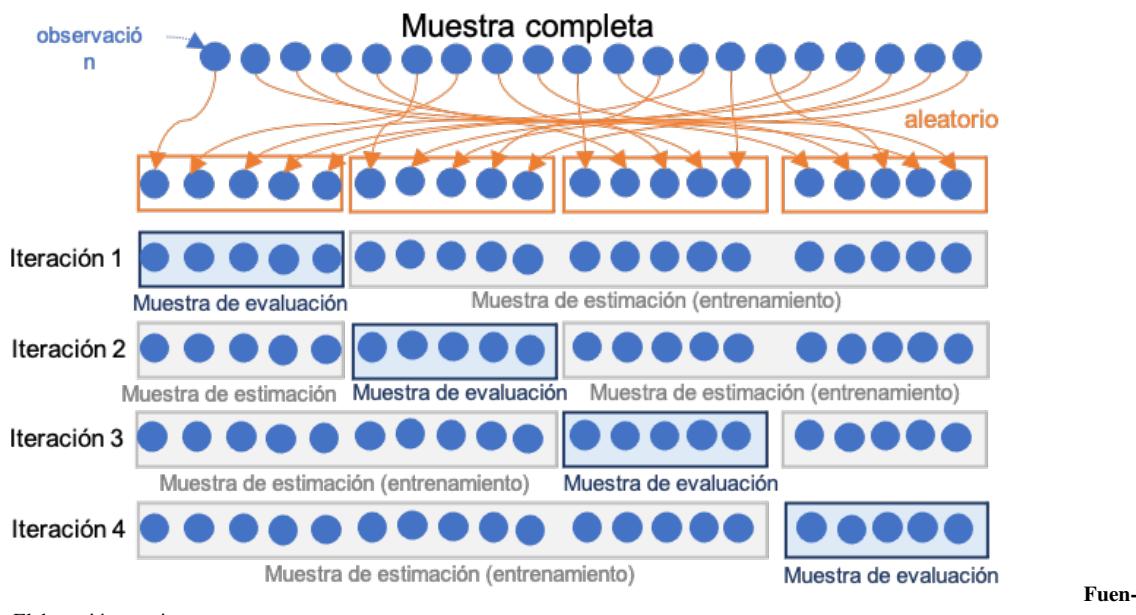
12.2.3 Método de k iteraciones

A parte del costo computacional del método LOOCV, puede existir potencialmente otro problema con esa aproximación. El LOOCV puede generar una mayor variación en el error de predicción si algunas observaciones son atípicas. Para evitar este problema, lo ideal sería utilizar una buena proporción de observaciones como muestra de prueba para evitar el peso de las observaciones

atípicas.

El método de **validación cruzada de k iteraciones** o *k -fold Cross-validation* intenta ser un intermedio entre el método de retención y el LOOCV. Este método implica dividir de manera aleatoria la muestra completa en k grupos de aproximadamente el mismo tamaño. Para cada uno de los k grupos (o iteraciones) se emplean los restantes $k - 1$ grupos como muestra de estimación y el grupo k de observaciones se emplea como muestra de evaluación para la cual se calculan las respectivas métricas deseadas para los modelos a comparar. Y finalmente, para obtener la métrica para todo el ejercicio se calcula el promedio de las k métricas calculadas para cada modelo. En la Figura 12.3 se presenta un diagrama de esta aproximación. Lo más común en la práctica es emplear un valor de k de 5 o 10.

Figura 12.3. Diagrama del Método de validación cruzada de k iteraciones para la evaluación de modelos



En la práctica, la validación cruzada de k iteraciones se recomienda generalmente sobre los otros dos métodos debido a su equilibrio entre la variabilidad que puede aparecer por los datos atípicos (método LOOCV), el sesgo fruto de emplear sólo una muestra de evaluación (método de retención) y el tiempo de ejecución computacional.

En la siguiente sección discutiremos las métricas para valorar la precisión de las predicciones.

12.3 Métricas para medir la precisión de las predicciones

Independientemente del método empleado para generar la muestra o muestras de evaluación, necesitaremos métricas que permitan evaluar que tan cerca está cada predicción (\hat{y}_i) del valor real

observado (y_i) para las n_e observaciones que conforman la muestra de evaluación. Las predicciones del modelo se puede construir fácilmente para los valores determinados de las variables explicativas para la i -esima observación de la muestra de evaluación de la siguiente manera:

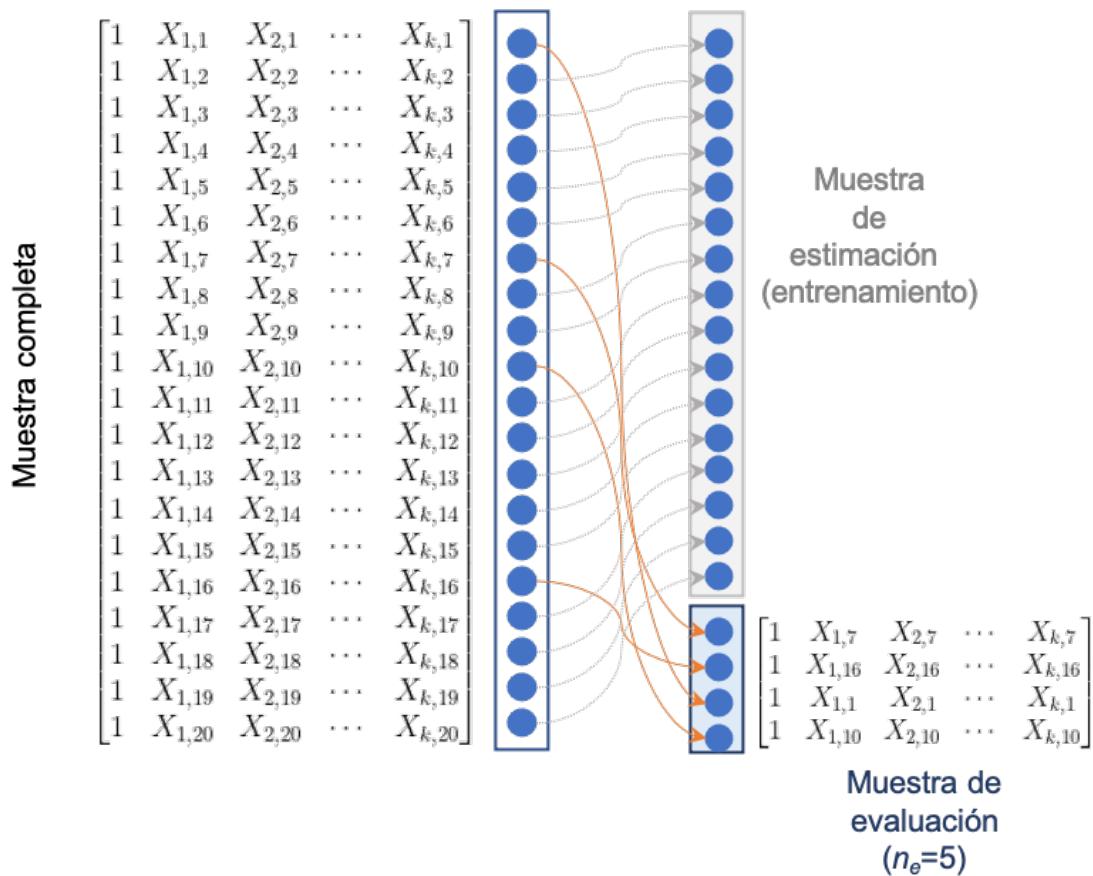
$$\hat{y}_i = \mathbf{x}_i^T \hat{\beta}, \quad (12.1)$$

donde \mathbf{x}_i^T corresponde al vector de valores que toman las variables explicativas para el individuo i en la muestra de evaluación. Es decir,

$$\mathbf{x}_i^T = \mathbf{x}_i^T = (1 \ X_{1,i} \ X_{2,i} \ \cdots \ X_{k,i}). \quad (12.2)$$

En otras palabras, \mathbf{x}_i^T es la fila de la matriz X de la muestra completa que corresponde al individuo i que fue asignado a la muestra de evaluación. En el ejemplo de la Figura 12.4, la muestra de evaluación es de tamaño 5 ($n_e = 5$) y la primera observación de la muestra de evaluación es la séptima de la muestra completa.

Figura 12.4. De la muestra completa a la muestra de evaluación (Ejemplo del Método de retención)



Fuente: Elaboración propia

Las dos métricas más empleadas para evaluar el comportamiento de las predicciones son:

- *RMSE* (Raíz media cuadrada del error):

$$RMSE = \sqrt{\frac{1}{n_e} \sum_{i=1}^{n_e} (\hat{y}_i - y_i)^2}$$

- *MAE* (Error absoluto promedio):

$$MAE = \frac{1}{n_e} \sum_{i=1}^{n_e} |\hat{y}_i - y_i|$$

La *RMSE* y el *MAE* intenta que los errores negativos y positivos no se cancelen al ser sumados. El *MAE* emplea el valor absoluto de los errores de prueba para eliminar los signos, mientras que la *RMSE* eleva los errores de prueba al cuadrado³. Esto implica que la *RMSE* al elevar al cuadrado penalice mas errores de prueba grandes que errores de prueba pequeños, mientras que el *MAE* pondera igual a todos los errores de prueba. Estas dos métricas tienen en común que una menor métrica es deseable.

Cada métrica pone énfasis en un aspecto diferente y por eso es una buena práctica emplear la mayor cantidad de métricas posibles para la selección del modelo que se comporta mejor fuera de la muestra de estimación. La selección de la o las métricas adecuadas para la evaluación de los modelos dependerá de la pregunta de negocio y de lo que es deseable para cada organización. En la siguiente sección se presenta un ejemplo práctico.

12.4 Validación cruzada en R

Para mostrar como responder a la pregunta cuál modelo tiene un mejor poder predictivo, emplearemos los mismos datos del Capítulo 6. En dicho capítulo empleamos una muestra de 150 observaciones que contenía una variable dependiente (y_i) y 25 posibles variables explicativas X_j, i donde $j = 1, 2, \dots, 25$. Los datos se encuentran disponibles en el archivo **DATOSautoSel.txt**.

En Capítulo 6 (sección 6.4.2) evaluamos 3 posibles modelos teniendo en cuenta que nuestra tarea era encontrar el mejor modelo para explicar la variable y_i . Ahora en este caso supongamos que queremos encontrar un modelo para hacer analítica predictiva y no diagnóstica. Es decir, ahora queremos comparar los tres modelos candidatos con la óptica de encontrar el mejor modelo predictivo.

Empecemos cargando los datos y estimando los siguientes tres modelos

$$y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + \beta_6 x_{5i} + \beta_7 x_{20i} + \varepsilon_i \quad (12.3)$$

$$y_i = \beta_1 + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + \beta_6 x_{5i} + \beta_7 x_{20i} + \varepsilon_i \quad (12.4)$$

³Después de sumarlos les saca la raíz cuadrada para retornarlos a su escala inicial

$$y_i = \beta_1 + \beta_3 x_2 i + \beta_4 x_3 i + \beta_5 x_4 i + \beta_6 x_5 i + \varepsilon_i \quad (12.5)$$

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

modeloA <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x20, datos)
modeloB <- lm(y ~ x2 + x3 + x4 + x5 + x20, datos)
modeloC <- lm(y ~ x2 + x3 + x4 + x5, datos)
```

Recordemos que estos tres modelos tenían todas sus variables significativas individual y conjuntamente (Ver 12.1). Además podrás constatar que los supuestos del teorema de Gauss-Markov se cumplen en los tres casos.

Cuadro 12.1: Modelos a valorar su capacidad predictiva.

	<i>Dependent variable:</i>		
	Modelo A (1)	y Modelo B (2)	ModeloC (3)
x1	0.663** (0.257)		
x2		1.780*** (0.266)	1.701*** (0.268)
x3		0.946*** (0.263)	0.857*** (0.263)
x4		1.212*** (0.254)	1.034*** (0.245)
x5		1.156*** (0.253)	0.997*** (0.247)
x20	-0.715*** (0.266)	-0.608** (0.268)	
Constant	11.464*** (1.343)	12.444*** (1.314)	11.924*** (1.312)
Observations	150	150	150
R ²	0.720	0.707	0.696
Adjusted R ²	0.708	0.697	0.688
Residual Std. Error	2.447 (df = 143)	2.494 (df = 144)	2.530 (df = 145)
F Statistic	61.274*** (df = 6; 143)	69.459*** (df = 5; 144)	83.157*** (df = 4; 145)

Note: *p<0.1; **p<0.05; ***p<0.01

Para realizar la validación cruzada podemos emplear diferentes paquetes, pero tal vez de los más empleados es el paquete *caret*(Kuhn, 2020) . Ahora procedamos a realizar la validación cruzada de estos tres modelos con cada uno de los tres métodos estudiados.

12.4.1 Método de retención

Para seleccionar aleatoriamente las observaciones que harán parte de la muestra de estimación (entrenamiento) y la de evaluación se puede emplear la función `createDataPartition()`. Esta función típicamente incluye los siguientes argumentos que serán útiles para este caso:

```
createDataPartition(y, times = 1, p = 0.5, list = TRUE))
```

donde:

- **y**: es un vector columna que corresponde a la variable dependiente.
- **times**: el número de muestras (particiones) que se desean crear. Por defecto `times = 1`.
- **p**: el porcentaje de observaciones de la muestra original que estarán en la muestra de entrenamiento. Por defecto `p = 0.5`.
- **list**: un valor lógico que le indica a la función en qué formato presentar los resultados. Si `list = TRUE` entonces el objeto será de clase `list` y en caso contrario será un objeto de clase `matrix`. Por defecto `list = TRUE`.

En nuestro caso,

```
# se carga el paquete
library(caret)
# se fija una semilla para los números aleatorios
set.seed(123)
# muestra de estimación
est.index <- createDataPartition(y = datos$y, p = 0.8, list = FALSE)
head(est.index, 3)

##      Resample1
## [1,]      2
## [2,]      6
## [3,]      7
```

Esto genera el objeto `est.index` de clase `list` que contiene las filas seleccionadas al azar para conformar la muestra de estimación. Entonces, podemos proceder a crear las muestras de estimación y evaluación.

```
# muestra de estimación
datos.est <- datos[est.index, ]

# muestra de evaluación
datos.eval <- datos[-est.index, ]
```

Ahora tendremos que reestimar los modelos con la muestra de entrenamiento (muestra de estimación).

```
# estimación de modelos con muestra de estimación
modeloA.ret <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x20, datos.est)
modeloB.ret <- lm(y ~ x2 + x3 + x4 + x5 + x20, datos.est)
modeloC.ret <- lm(y ~ x2 + x3 + x4 + x5, datos.est)
```

Ahora, procedamos a generar las predicciones con la función **predict()** de R⁴. Esta función realiza la operación que se presenta en la ecuación (12.1). Es decir, encuentra los valores estimados para la variable dependiente (\hat{y}_i) (predicciones) dado unos coeficientes estimados para un modelo. Los argumentos necesarios, para este caso, en esta función son el modelo y los nuevos datos para los cuales se hará la predicción.

```
# predicciones para el modelo A para la muestra de evaluación
pred.A <- predict(modeloA.ret, datos.eval)
head(pred.A, 2)
```

```
##      1      3
## 39.38654 36.63466
```

```
# predicciones para el modelo A para la muestra de evaluación
pred.B <- predict(modeloB.ret, datos.eval)
```

```
# predicciones para el modelo A para la muestra de evaluación
pred.C <- predict(modeloC.ret, datos.eval)
```

Ahora calculemos la *RMSE* y el *MAE* para cada uno de los tres modelos. Esto lo podemos hacer con las funciones **RMSE()** y **MAE()** del paquete *caret*. Las dos funciones emplean como primer argumento las predicciones y como segundo argumento los valores observados. En nuestro caso podemos encontrar estas métricas con el siguiente código.

```
# Cálculo del RMSE
RMSE.modeloA <- RMSE(pred.A, datos.eval$y)
RMSE.modeloB <- RMSE(pred.B, datos.eval$y)
RMSE.modeloC <- RMSE(pred.C, datos.eval$y)

# Cálculo del MAE
MAE.modeloA <- MAE(pred.A, datos.eval$y)
MAE.modeloB <- MAE(pred.B, datos.eval$y)
MAE.modeloC <- MAE(pred.C, datos.eval$y)
```

Ahora podemos comparar los resultados. En el Cuadro 12.2 se presenta un resumen de estos resultados. El modelo que minimiza la *RMSE* es Modelo C y el que minimiza el *MAE* es Modelo A. Veamos si estos resultados se mantienen con los otros métodos de validación cruzada.

⁴Para ser mas preciso se empleará la función **predict.lm()** . La función **predict()** es una función genérica para predicciones que de acuerdo con la clase de objeto que se emplee como argumento redirecciona a la función respectiva. Como empleamos objetos de clase **lm** la función **predict()** redirecciona a la función **predict.lm()**.

Cuadro 12.2: Métricas de precisión de los 3 modelos en muestra de evaluación

	RMSE	MAE
Modelo A	2.09	1.62
Modelo B	2.18	1.72
Modelo C	2.05	1.64

12.4.2 Método LOOCV

El método LOOCV puede ser implementado fácilmente empleando el paquete *caret*. En este caso debemos emplear dos funciones. La primera es función **trainControl()** . Esta función establece el método (**method**) que se empleará para la validación cruzada. Definamos el método a LOOCV de la siguiente manera.

```
train.control <- trainControl(method = "LOOCV")
```

Ahora podemos emplear la función principal para hacer la valoración: **train()** . Para emplear cualquier método recursivo de valoración cruzada, esta función necesita los siguientes argumentos:

```
train(formula, data, method = "lm", trControl = trainControl())
```

donde:

- **formula:** es la fórmula correspondiente al modelo que será evaluado .
- **data:** **data.frame** con los datos (la muestra completa).
- **method:** en nuestro caso deberá fijarse en **method = "lm"** porque estamos evaluando modelos lineales. Es importante anotar que este no es el valor por defecto de este argumento, así que es importante especificar este argumento de esta manera. Las otras opciones de este argumento no son relevantes para nuestro caso.
- **trControl:** este argumento define que tipo de valoración realizar.

Para implementar el método LOOCV para el primer método se puede emplear el siguiente código

```
LOOCV.modeloA <- train(y ~ x1 + x2 + x3 + x4 + x5 + x20, data = datos,
  method = "lm", trControl = train.control)

# resultados
LOOCV.modeloA

## Linear Regression
##
## 150 samples
##   6 predictor
##
```

```

## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 149, 149, 149, 149, 149, 149, ...
## Resampling results:
##
##    RMSE     Rsquared     MAE
##    2.505238  0.6922227  1.99463
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

```

Cuadro 12.3: Métricas de precisión de los 3 modelos con LOOCV

	RMSE	MAE
Modelo A	2.51	1.99
Modelo B	2.55	2.04
Modelo C	2.57	2.02

Ahora realice la misma valoración para los otros dos modelos. Los resultados se presentan en el Cuadro 12.3. Nuevamente, el modelo que minimiza la *RMSE* y *MAE* es Modelo A. Cómo se discutió anteriormente este resultado no necesariamente debe coincidir entre el método de retención y el LOOCV.

12.4.3 Método de k iteraciones

El método de k iteraciones puede ser implementado de manera similar al método a LOOCV. Lo único que debemos modificar es la definición de la función **trainControl()**. Empleemos 5 iteraciones ($k = 5$).

```
train.control <- trainControl(method = "cv", number = 5)
```

Y finalmente, podemos implementar el método de k iteraciones empleando el mismo código anteriormente discutido. Es decir,

```

fold.5.modeloA <- train(y ~ x1 + x2 + x3 + x4 + x5 + x20, data = datos,
  method = "lm", trControl = train.control)

# resultados
fold.5.modeloA

## Linear Regression
##
## 150 samples
##   6 predictor
##
```

```

## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 119, 121, 120, 120, 120
## Resampling results:
##
##   RMSE     Rsquared    MAE
##   2.494089  0.7040192  1.974375
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

```

Cuadro 12.4: Métricas de precisión de los 3 modelos con 5 iteraciones

	RMSE	MAE
Modelo A	2.49	1.97
Modelo B	2.59	2.10
Modelo C	2.56	2.02

Ahora realice la misma valoración para los otros dos modelos. Los resultados se presentan en el Cuadro 12.4. Nuevamente, el modelo que minimiza la *RMSE* y *MAE* es Modelo A. Esto implica que con los tres métodos de validación cruzada podemos concluir que el mejor modelo para hacer predicciones es el Modelo A.

En este caso esta conclusión coincide con el mejor modelo detectado cuando estábamos interesados encontrar el mejor modelo para entender que ocurría con la muestra (analítica diagnóstica). Este resultado no es común y por esto debemos realizar el análisis presentado si se desea realizar analítica predictiva.

12.5 Intervalo de confianza para la predicción

En este capítulo hemos discutido cómo seleccionar entre modelos candidatos cuando el objetivo es realizar analítica predictiva. Una vez seleccionado el mejor modelo para realizar las predicciones podemos proceder a realizar predicciones con dicho modelo, empleando la muestra completa para estimar el modelo y aplicar los coeficientes estimados al nuevo individuo para el cuál se quiere predecir el valor de la variable explicativa.

Antes de finalizar es importante anotar que emplear la expresión (12.1) producirá un estimador puntual para el valor esperado de la variable dependiente. En algunas situaciones podría ser interesante producir un intervalo de confianza para la predicción porque de esta manera estaremos generando una región con una certidumbre del $(1 - \alpha) \times 100\%$.

Para construir un intervalo de confianza para la predicción para un nuevo individuo⁵ será necesario suponer una distribución para los errores del modelo. Si se supone una distribución normal, entonces

⁵Técnicamente hay que hacer una distinción entre un intervalo de confianza para la predicción individual y un intervalo para el valor esperado (valor promedio). Si bien en ambos casos el centro del intervalo de confianza es el mismo los

el intervalo de confianza para la predicción de la variable dependiente para un individuo i cuyas valores de las variables explicativas toman el valor de $\mathbf{x}_{p,i}^T$ será:

$$\hat{y}_{p,i} \pm t_{\frac{\alpha}{2}, n-k} \cdot \sqrt{s^2 \cdot \mathbf{x}_{p,i}^T (X^T X)^{-1} \mathbf{x}_{p,i}} \quad (12.6)$$

donde s^2 es la varianza estimada del error y $t_{\frac{\alpha}{2}, n-k}$ es el valor de la distribución t con $n - k$ grados de libertad⁶ y

$$\hat{y}_{p,i} = \mathbf{x}_{p,i}^T \hat{\beta} \quad (12.7)$$

Para construir el intervalo de confianza para la predicción individual se puede realizar en R con la función **predict()** de la base de R. Esta función realiza la operación que se presenta en la ecuación (12.6). El único argumento adicional que se debe tener en cuenta al discutido en las secciones anteriores es **interval** que permite establecer el cálculo del intervalo de confianza. Y para que se calcule el intervalo de confianza para la predicción bajo el supuesto de errores con una distribución normal, este argumento debe ser **interval = "prediction"**. Por defecto la función calcula un intervalo de confianza del 95 %, si se desea modificar el nivel de confianza, se puede emplear el argumento **level**.

Por ejemplo, supongamos que empleamos el Modelo A con todos los datos disponibles y queremos hacer una predicción para un individuo que tenga los siguientes valores para las variables explicativas: $x1_p = 4, x2_p = 3, x3_p = 5, x4_p = 2, x5_p = 4$ y $x20_p = 1$. Esto se puede realizar de la siguiente manera.

```
# nuevos datos
nuevos.datos <- data.frame(x1 = 4, x2 = 3, x3 = 5, x4 = 2, x5 = 4,
                             x20 = 1)

prediccion <- predict(modeloA, nuevos.datos, interval = "prediction",
                      level = 0.95)

prediccion

##          fit      lwr      upr
## 1 29.44943 24.22057 34.6783
```

intervalos tienen amplitudes diferentes. En general es mucho más fácil estimar en promedio cuál sería el valor de la variable dependiente si aparecen muchos individuos con unos valores determinados para las variables explicativas que predecir el valor de la variable dependiente para un solo individuo con unos valores determinados para las variables explicativas. En el caso del intervalo de confianza para el valor esperado, se puede invocar el Teorema del Límite Central y por tanto no se requiere ningún supuesto. En la introducción del Capítulo 3 se presentó una discusión del Teorema del Límite Central. Caso muy diferente para el intervalo de confianza de una predicción individual en la cual es necesaria suponer la distribución del término de error.

⁶En la sección 3.2 se discutió cómo si un estimador sigue una distribución normal y se tiene que estimar la varianza (σ^2) del error, entonces se tendrá que emplear una distribución t con $n - k$ grados de libertad.

En este caso lo pronóstico para el individuo cuyas las variables explicativas son $x1_p = 4$, $x2_p = 3$, $x3_p = 5$, $x4_p = 2$, $x5_p = 4$ y $x20_p = 1$ es de 29.45. El límite inferior del intervalo del 95% de confianza es 24.22 y el límite superior es 34.68. Si el modelo tiene problemas heteroscedasticidad⁷ y dicho problema se resolvió con una matriz de varianzas y covarianzas H.C., entonces la predicción debería tener en cuenta dicha corrección. Esto se puede realizar con la función **Predcit()** del paquete *car*(Fox y Weisberg, 2019). Esta función tiene argumentos similares a la función **predcit.lm()** del paquete central de R. Una gran diferencia de la función **Predcit()** es el argumento **vcov**, que permite especificar la corrección H.C. que se desea emplear. Por ejemplo, mantengamos el mismo ejemplo anterior y supongamos que existe un problema de heteroscedasticidad que se soluciona con HC3.

```
# se cargan las librerías
library(car)
library(sandwich)

Predict(modeloA, nuevos.datos, interval = "prediction", level = 0.95,
        vcov. = vcovHC(modeloA, "HC3"))

##          fit      lwr      upr
## 1 29.44943 24.20053 34.69834
```

Si el supuesto de normalidad de los errores no se cumple, una opción es simular la distribución de los errores empleando el método de *bootstrapping*. En el apéndice se describe el método de bootstrapping y el código que se puede emplear. Recuerden que ya habíamos discutido como comprobar el supuesto de normalidad de los residuales en la sección 9.4 cuatro pruebas de normalidad: Shapiro-Wilk (Shapiro y Francia, 1972), Kolmogorov-Smirnov (Kolmogorov, 1933), Cramer-von Mises (Cramér, 1928) y Anderson-Darling(Anderson y Darling, 1952).

12.6 Comentarios finales

Las técnicas de validación cruzada que hemos estudiado en este capítulo permiten evaluar el desempeño predictivo de diferentes modelos candidatos a ser el mejor modelo predictivo. Los modelos candidatos fueron construidos con los algoritmos de selección automática estudiados en el Capítulo 6.

Una práctica diferente para seleccionar modelos para hacer analítica predictiva es seleccionar el mejor modelo empleando directamente la validación cruzada. Es decir, empleando los métodos de LOOCV o k iteraciones para comparar los posibles modelos empleando el comportamiento predictivo y no el ajuste del modelo a la muestra como lo hicimos en el Capítulo 6. Esto se puede realizar empleando la librería *caret*, si te interesa podrás encontrar en la documentación del paquete información de cómo desarrollar esta tarea.

⁷El problema de autocorrelación no se considera, pues los datos para esta tarea de predicción (y no de proyección) deberán ser de corte transversal y no una serie de tiempo.

Para terminar es importante recordar que en la práctica, la validación cruzada de k iteraciones se recomienda generalmente sobre los otros dos métodos estudiados en este capítulo. La razón de preferir este método es que presenta un equilibrio entre la variabilidad que puede aparecer por los datos atípicos (método LOOCV), el sesgo fruto de emplear sólo una muestra de evaluación (método de retención) y el tiempo de ejecución computacional.

12.7 Anexo: Método de Bootstrapping para construcción de intervalos de confianza para las predicciones

Existen diferentes métodos para la construcción de intervalos de confianza para las predicciones empleando bootstrapping, tal vez el mas sencillo es el propuesto por Davison y Hinkley, 1997 (sección 6.3.3).

Partamos de reconocer que el error de predicción está dado por:

$$e_{p,i} = y_{p,i} - \hat{y}_{p,i} \quad (12.8)$$

Por lo tanto,

$$y_{p,i} = \hat{y}_{p,i} + e_{p,i} \quad (12.9)$$

Por lo tanto, un intervalo de confianza del 90 % de confianza implica encontrar un estimador del percentil quinto (e_p^5) y 95 (e_p^{95}) del error de predicción ($e_{p,i}$). En este caso el intervalo estará definido como

$$[\hat{y}_{p,i} + e_p^5, \hat{y}_{p,i} + e_p^{95}] \quad (12.10)$$

Ahora el problema es encontrar los percentiles 5 y 95 del error de predicción. Reescribamos el error de predicción de la siguiente manera:

$$\begin{aligned} e_{p,i} &= y_{p,i} - \hat{y}_{p,i} \\ &= \mathbf{x}_{p,i}^T \boldsymbol{\beta} + \varepsilon_{p,i} - \mathbf{x}_{p,i}^T \hat{\boldsymbol{\beta}} \\ &= \mathbf{x}_{p,i}^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon_{p,i} \end{aligned} \quad (12.11)$$

La estrategia consistirá en muestrear (a esto se le denomina bootstrap) muchas veces de $e_{p,i}$ y luego calcular los percentiles de la manera habitual. Así, por ejemplo podemos sacar 10 mil muestras de $e_{p,i}$, y luego estimemos los percentiles 5 y 95. Esto en otras palabras será encontrar el error de predicción simulado tal que 500 de ellos sean menores que el para encontrar el percentil 5. Y para el percentil 95 la observación simulada tal que 9500 sean menores que ella.

Para muestrear $\mathbf{x}_{p,i}^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$, podemos hacer un bootstrap de los errores. Así, en cada réplica bootstrap (cada muestra), se extrae n veces con reemplazo de los residuales estandarizados (residuales divididos por la varianza) para obtener ε^* (vector de errores simulados para todas las observaciones

originales), y luego se reconstruye un nuevo vector de observaciones para la variable dependiente: $y^* = \mathbf{X}\hat{\beta} + \varepsilon^*$. Esto permite calcular con el nuevo conjunto de datos (y^* y la matriz \mathbf{X} original) un nuevo vector de coeficientes ($\hat{\beta}^*$) para cada muestra (iteración). Por último, $\mathbf{x}_{p,i}^T (\beta - \hat{\beta})$ puede ser aproximado por $\mathbf{x}_{p,i}^T (\hat{\beta} - \hat{\beta}^*)$.

Ahora queda faltando como encontrar $\varepsilon_{p,i}$. Dado los supuestos sobre el error del Teorema de Gauss-Markov, la manera natural de muestrear $\varepsilon_{p,i}$ es utilizar los residuos que tenemos de la regresión ($\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \dots, \hat{\varepsilon}_n^*$) en cada iteración. Los residuos tienen varianzas diferentes y, por lo general, demasiado pequeñas, por lo que querremos muestrear de los residuos corregidos por la varianza.

Resumiendo, el algoritmo para hacer un intervalo de confianza con un nivel de significancia de α será:

1. Realizar la predicción $\hat{y}_{p,i} = \mathbf{x}_{p,i}^T \hat{\beta}$
2. Construir el vector de residuales ajustados por la varianza $[s_1 - \bar{s}, s_2 - \bar{s}, \dots, s_n - \bar{s}]$ donde $s_i = \hat{\varepsilon}_i / \sqrt{1 - h_i}$ y h_i es el leverage de la observación i^8 .
3. Seleccionar una muestra de tamaño n de los residuales ajustados. Es decir, los errores de bootstrap ($\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \dots, \hat{\varepsilon}_n^*$)
4. Construir un nuevo vector de variables explicativas simuladas. El y de bootstrap $y^* = \mathbf{X}\hat{\beta} + \varepsilon^*$
5. Estimar por MCO los coeficientes de bootstrap. Es decir, $\hat{\beta}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y^*$
6. Obtener los residuales de bootstrap. En otras palabras, $\hat{\varepsilon}^* = y^* - \mathbf{X}\hat{\beta}^*$
7. Calcular los residuales de bootstrap ajustados por varianza ($s^* - \bar{s}$).
8. Seleccionar al azar uno de los residuales de bootstrap ajustados por varianza ($\varepsilon_{p,i}^*$).
9. Calcular $e_{p,i}$ como $e_{p,i}^* = \mathbf{x}_{p,i}^T (\hat{\beta} - \hat{\beta}^*) + \varepsilon_{p,i}^*$
10. Repetir 3 a 9 R veces⁹.
11. Encontrar los percentiles $\alpha/2 \times 100 (e_p^{(\alpha/2) \times 100})$ y $(1 - \alpha/2) \times 100 (e_p^{(1-\alpha/2) \times 100})$ de los $e_{p,i}^*$ simulados (son R simulaciones disponibles).
12. Construir el intervalo de bootstrapping como $[\hat{y}_{p,i} + e_p^5, \hat{y}_{p,i} + e_p^{95}]$

Para construir una función que realice este algoritmo construiremos dos funciones auxiliares. La primera realiza el paso 2 que corresponde a la estimación de los residuales ajustados por la varianza sobre un objeto de clase `lm`. Para esto construiremos la función `errores.ajustados.varianza()`.

```
# función para crear los errores ajustados por varianza (paso
# 2) argumento objeto lm

errores.ajustados.varianza <- function(modelo) {
  require(MASS)
  require(Hmisc)
  leverage <- influence(modelo)$hat
  s.resid <- residuals(modelo)/sqrt(1 - leverage)
```

⁸El leverage de una observación i corresponde al elemento i -ésimo de la diagonal de la matriz \mathbf{H} que está definida como $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Es decir, $h_i = [\mathbf{H}]_{i,i}$

⁹ R debe ser un número relativamente grande como 10 mil

```

s.resid <- s.resid - mean(s.resid)

return(s.resid)
}

```

La segunda función auxiliar realizará los pasos 3 a 9 del algoritmo empleando un objeto de clase `lm` (modelo), un vector de residuales ajustados por la varianza (`s`) y unos nuevos datos para la predicción (`nuevos.datos`). A esta nueva función la llamaremos `ic.boot.predic.iteration()`.

```

# función para correr pasos 3 a 9 argumentos modelo = objeto
# lm s = vector de residuales ajustados por la varianza
# nuevos datos = data.frame con datos para la predicción

ic.boot.predic.iteration <- function(modelo, s, nuevos.datos) {
  # pasos 3 a 9
  require(MASS)
  require(Hmisc)
  # residuales de bootstrap
  ep.star <- sample(s, size = length(modelo$residuals), replace = TRUE)

  # crea y de bootstrap
  y.star <- fitted(modelo) + ep.star

  # coeficientes de bootstrap
  x <- model.frame(modelo)[, -1]
  bs.data <- cbind(y.star, x)
  bs.modelo <- lm(y.star ~ ., bs.data)

  # residuales de bootstrap ajustados
  bs.lev <- influence(bs.modelo)$hat
  bs.s <- residuals(bs.modelo)/sqrt(1 - bs.lev)
  bs.s <- bs.s - mean(bs.s)

  # Selección del error de predicción
  xb.xb <- coef(modelo)[["(Intercept)"]] - coef(bs.modelo)[["(Intercept)"]]
  xb.xb <- xb.xb + (coef(modelo)[-1] - coef(bs.modelo)[-1]) *
    nuevos.datos
  return(unname(xb.xb + sample(bs.s, size = 1)))
}

```

Ahora construyamos la función `ic.boot.predic()` que ponga todo junto, replicando los pasos 3 a 9 R veces.

```

# función para crear IC para predicción con bootstrapping
# argumentos modelo = objeto lm s = vector de residuales
# ajustados por la varianza nuevos datos = data.frame con
# datos para la predicción R = número de iteraciones para el

```

```
# bootstrapping, por defecto R = 1000 alpha = nivel de
# significancia, por defecto alpha = 0.05

ic.boot.predic <- function(modelo, nuevos.datos, R = 1000, alpha = 0.05) {
  # paso 1
  y.p <- predict.lm(modelo, nuevos.datos)

  # paso 2
  s <- errores.ajustados.varianza(modelo)

  # paso 10 (repite pasos 3 a 9 R veces)

  ep.draws <- replicate(R, ic.boot.predic.iteration(modelo,
    s, nuevos.datos))

  # paso 11 y 12 encontrar los percentiles y construcción
  # intervalo

  res <- y.p + quantile(as.numeric(ep.draws), probs = c(alpha/2,
    1 - alpha/2))

  return(c(fit = y.p, lwr = res[1], upr = res[2]))
}
```

Apliquemos esta función al ejemplo que trabajamos en este capítulo y comparemos el resultado con la aproximación tradicional con el supuesto de normalidad.

```
set.seed(123445)
ic.boot.predic(modeloA, nuevos.datos, R = 10000, alpha = 0.05)

##      fit.lwr.2.5% upr.97.5%
## 29.44943 23.73147 34.98249

prediccion

##      fit      lwr      upr
## 1 29.44943 24.22057 34.6783
```

13 . Segundo caso de negocio

Diseñado por Freepik

Objetivos del capítulo

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

- Emplear las herramientas estudiadas en los capítulos anteriores para responder una pregunta de negocio que implique analítica diagnóstica
- Presentar los resultados de una regresión de manera gráfica empleando R con solución H.A.C..
- Determinar cuál variable tiene mas efecto sobre la variable explicativa empleando R en presencia de heteroscedasticidad.

13.1 Introducción

En los capítulos anteriores hemos estudiado las bases del modelo clásico de regresión múltiple, cómo encontrar el mejor modelo para hacer analítica diagnóstica o analítica predictiva. . En este capítulo pondremos todos los elementos juntos para resolver un caso de negocio que implica analítica predictiva.

13.2 La pregunta de negocio

Una empresa de consultoría en seguridad en los Estados Unidos quiere contar con un modelo que le permita, al visitar a un alcalde de cualquier ciudad, poder tener un estimado de la tasa de crímenes violentos por cada 100000 habitantes y como esta puede cambiar cuando se “intervengan” algunas variables. Es decir, esta empresa quiere tener una “fórmula” que le permita determinar cuál sería la tasa de crímenes violentos por cada 100000 habitantes bajo diferentes escenarios. La empresa quiere ofrecer una nueva linea de servicios que le permita asesorar a los alcaldes en que variables intervenir para disminuir la tasa de crímenes violentos.

Para realizar esta tarea, contamos con una base de datos con información sociodemográfica y datos de crimen del FBI para diferentes comunidades de Estados Unidos, los datos se encuentran en el archivo *DatosCaso2.csv*. Los datos los datos son reales y fueron suministrada por Redmond y Baveja (2002) y tomados de la siguiente página <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime>.

La base de datos contiene 1994 observaciones y las siguientes 101 variables:

1. population: población para la comunidad: (numérico - decimal)
2. householdsize: media de personas por hogar (numérico - decimal)
3. racepctblack: porcentaje de población afroamericana (numérico - decimal)
4. racePctWhite: porcentaje de población caucásica (numérico - decimal)
5. racePctAsian: porcentaje de población de origen asiático (numérico - decimal)
6. racePctHisp: porcentaje de población de origen hispano (numérico - decimal)
7. agePct12t21: porcentaje de población de 12 a 21 años (numérico - decimal)
8. agePct12t29: porcentaje de población de 12-29 años (numérico - decimal)
9. agePct16t24: porcentaje de población de 16-24 años (numérico - decimal)
10. agePct65up: porcentaje de población de 65 años o más (numérico - decimal)
11. numbUrban: número de personas que viven en zonas clasificadas como urbanas (numérico - decimal)
12. pctUrban: porcentaje de personas que viven en zonas clasificadas como urbanas (numérico - decimal)
13. medIncome: ingreso familiar medio (numérico - decimal)
14. pctWWage: porcentaje de hogares con ingresos salariales en 1989 (numérico - decimal)
15. pctWFarmSelf: porcentaje de hogares con ingresos agrícolas o por cuenta propia en 1989 (numérico - decimal)
16. pctWInvInc: porcentaje de hogares con ingresos por inversiones/alquilares en 1989 (numérico

- decimal)
- 17. pctWSocSec: porcentaje de hogares con ingresos de la seguridad social en 1989 (numérico - decimal)
 - 18. pctWPubAsst: porcentaje de hogares con ingresos de la asistencia pública en 1989 (numérico - decimal)
 - 19. pctWRetire: porcentaje de hogares con ingresos por jubilación en 1989 (numérico - decimal)
 - 20. medFamInc: renta familiar mediana (difiere de la renta de los hogares no familiares) (numérico - decimal)
 - 21. perCapInc: renta per cápita (numérico - decimal)
 - 22. whitePerCap: renta per cápita de los caucásicos (numérico - decimal)
 - 23. blackPerCap: renta per cápita de los afroamericanos (numérico - decimal)
 - 24. IndianPerCap: renta per cápita de los nativos americanos (numérico - decimal)
 - 25. AsianPerCap: renta per cápita de las personas de origen asiático (numérico - decimal)
 - 26. OtherPerCap: renta per cápita de las personas con “otra” herencia (numérico - decimal)
 - 27. HispPerCap: renta per cápita de las personas de origen hispano (numérico - decimal)
 - 28. NumUnderPov: número de personas por debajo del nivel de pobreza (numérico - decimal)
 - 29. PctPopUnderPov: porcentaje de personas bajo el nivel de pobreza (numérico - decimal)
 - 30. PctLess9thGrade: porcentaje de personas de 25 años o más con menos de 9º grado de educación (numérico - decimal)
 - 31. PctNotHSGrad: porcentaje de personas de 25 años o más que no tienen estudios secundarios (numérico - decimal)
 - 32. PctBSorMore: porcentaje de personas de 25 años o más con una licenciatura o educación superior (numérico - decimal)
 - 33. PctUnemployed: porcentaje de personas de 16 años o más, que forman parte de la población activa y están desempleadas (numérico - decimal)
 - 34. PctEmploy: porcentaje de personas de 16 años o más que están empleadas (numérico - decimal)
 - 35. PctEmplManu: porcentaje de personas de 16 años o más que trabajan en la industria manufacturera (numérico - decimal)
 - 36. PctEmplProfServ: porcentaje de personas de 16 años o más que trabajan en servicios profesionales (numérico - decimal)
 - 37. PctOccupManu: porcentaje de personas de 16 años o más que trabajan en la industria manufacturera (numérico - decimal)
 - 38. PctOccupMgmtProf: porcentaje de personas de 16 años o más empleadas en ocupaciones de gestión o profesionales (numérico - decimal)
 - 39. MalePctDivorce: porcentaje de hombres divorciados (numérico - decimal)
 - 40. MalePctNevMarr: porcentaje de hombres que nunca se han casado (numérico - decimal)
 - 41. FemalePctDiv: porcentaje de mujeres divorciadas (numérico - decimal)
 - 42. TotalPctDiv: porcentaje de población divorciada (numérico - decimal)
 - 43. PersPerFam: número medio de personas por familia (numérico - decimal)
 - 44. PctFam2Par: porcentaje de familias (con hijos) encabezadas por dos padres (numérico - decimal)
 - 45. PctKids2Par: porcentaje de niños en viviendas familiares con dos padres (numérico - decimal)
 - 46. PctYoungKids2Par: porcentaje de niños de 4 años o menos en hogares con dos padres (numérico - decimal)

- rico - decimal)
- 47. PctTeen2Par: porcentaje de niños de 12 a 17 años en hogares con dos padres (numérico - decimal)
 - 48. PctWorkMomYoungKids: porcentaje de madres de niños de 6 años o menos que trabajan (numérico - decimal)
 - 49. PctWorkMom: porcentaje de madres de niños menores de 18 años que trabajan (numérico - decimal)
 - 50. NumIlleg: número de niños nacidos de madres nunca casadas (numérico - decimal)
 - 51. PctIlleg: porcentaje de hijos nacidos de personas nunca casadas (numérico - decimal)
 - 52. NumImmig: número total de personas que se sabe que han nacido en el extranjero (numérico - decimal)
 - 53. PctImmigRecent: porcentaje de inmigrantes que han inmigrado en los últimos 3 años (numérico - decimal)
 - 54. PctImmigRec5: porcentaje de inmigrantes que han inmigrado en los últimos 5 años (numérico - decimal)
 - 55. PctImmigRec8: porcentaje de inmigrantes que han inmigrado en los últimos 8 años (numérico - decimal)
 - 56. PctImmigRec10: porcentaje de inmigrantes que han inmigrado en los últimos 10 años (numérico - decimal)
 - 57. PctRecentImmig: porcentaje de la población que ha inmigrado en los últimos 3 años (numérico - decimal)
 - 58. PctRecImmig5: porcentaje de población que ha inmigrado en los últimos 5 años (numérico - decimal)
 - 59. PctRecImmig8: porcentaje de población que ha inmigrado en los últimos 8 años (numérico - decimal)
 - 60. PctRecImmig10: porcentaje de población que ha inmigrado en los últimos 10 años (numérico - decimal)
 - 61. PctSpeakEnglOnly: porcentaje de personas que sólo hablan inglés (numérico - decimal)
 - 62. PctNotSpeakEnglWell: porcentaje de personas que no hablan bien el inglés (numérico - decimal)
 - 63. PctLargHouseFam: porcentaje de hogares familiares que son grandes (6 o más) (numérico - decimal)
 - 64. PctLargHouseOccup: porcentaje de hogares ocupados que son grandes (6 o más personas) (numérico - decimal)
 - 65. PersPerOccupHous: media de personas por hogar (numérico - decimal)
 - 66. PersPerOwnOccHous: media de personas por hogar ocupado por el propietario (numérico - decimal)
 - 67. PersPerRentOccHous: media de personas por hogar de alquiler (numérico - decimal)
 - 68. PctPersOwnOccup: porcentaje de personas en hogares ocupados por el propietario (numérico - decimal)
 - 69. PctPersDenseHous: porcentaje de personas en viviendas densas (más de 1 persona por habitación) (numérico - decimal)
 - 70. PctHousLess3BR: porcentaje de viviendas con menos de 3 dormitorios (numérico - decimal)

71. MedNumBR: número medio de dormitorios (numérico - decimal)
72. HousVacant: número de viviendas vacías (numérico - decimal)
73. PctHousOccup: porcentaje de viviendas ocupadas (numérico - decimal)
74. PctHousOwnOcc: porcentaje de hogares ocupados por el propietario (numérico - decimal)
75. PctVacantBoarded: porcentaje de viviendas vacías que están tapiadas (numérico - decimal)
76. PctVacMore6Mos: porcentaje de viviendas desocupadas que han estado desocupadas más de 6 meses (numérico - decimal)
77. MedYrHousBuilt: año medio de construcción de las viviendas (numérico - decimal)
78. PctHousNoPhone: porcentaje de viviendas ocupadas sin teléfono (en 1990, esto era raro) (numérico - decimal)
79. PctWOFullPlumb: porcentaje de viviendas sin instalaciones completas de acueducto (numérico - decimal)
80. OwnOccLowQuart: viviendas ocupadas por sus propietarios - valor del cuartil inferior (numérico - decimal)
81. OwnOccMedVal: viviendas ocupadas por sus propietarios - valor medio (numérico - decimal)
82. OwnOccHiQuart: vivienda ocupada por el propietario - valor del cuartil superior (numérico - decimal)
83. RentLowQ: vivienda de alquiler - renta del cuartil inferior (numérico - decimal)
84. RentMedian: vivienda de alquiler - renta mediana (variable censal H32B del fichero STF1A) (numérico - decimal)
85. RentHighQ: vivienda de alquiler - renta del cuartil superior (numérico - decimal)
86. MedRent: alquiler bruto medio (variable censal H43A del fichero STF3A - incluye los servicios públicos) (numérico - decimal)
87. MedRentPctHousInc: alquiler bruto medio como porcentaje de los ingresos del hogar (numérico - decimal)
88. MedOwnCostPctInc: costo medio de los propietarios como porcentaje de los ingresos del hogar - para propietarios con hipoteca (numérico - decimal)
89. MedOwnCostPctIncNoMtg: costo medio de los propietarios como porcentaje de los ingresos del hogar - para los propietarios sin hipoteca (numérico - decimal)
90. NumInShelters: número de personas en refugios para personas sin hogar (numérico - decimal)
91. NumStreet: número de personas sin hogar contabilizadas en la calle (numérico - decimal)
92. PctForeignBorn: porcentaje de personas nacidas en el extranjero (numérico - decimal)
93. PctBornSameState: porcentaje de personas nacidas en el mismo estado en el que viven actualmente (numérico - decimal)
94. PctSameHouse85: porcentaje de personas que viven en la misma casa que en 1985 (5 años antes) (numérico - decimal)
95. PctSameCity85: porcentaje de personas que viven en la misma ciudad que en 1985 (5 años antes) (numérico - decimal)
96. PctSameState85: porcentaje de personas que viven en el mismo estado que en 1985 (5 años antes) (numérico - decimal)
97. LandArea: superficie terrestre en millas cuadradas (numérico - decimal)
98. PopDens: densidad de población en personas por milla cuadrada (numérico - decimal)
99. PctUsePubTrans: porcentaje de personas que utilizan el transporte público para desplazarse

(numérico - decimal)

100. LemasPctOfficDrugUn: porcentaje de agentes asignados a unidades de drogas (numérico - decimal)
101. ViolentCrimesPerPop: número total de delitos violentos por cada 100000 habitantes (numérico - decimal) (Variable dependiente) Traducido con la versión gratuita del Traductor de DeepL.

Responder esta pregunta de negocio implicaba realizar analítica predictiva al tener que encontrar una “formula” que pueda que le permita determinar cuál sería la tasa de crímenes violentos por cada 100000 habitantes para diferentes valores de las variables explicativas (escenarios).

13.3 El plan

Recordemos que nuestra primera tarea siempre es trazar una ruta analítica para responder la pregunta de negocio. Para este momento, ya debes estar intuyendo que es muy probable que exista un problema de heteroscedasticidad dado que la base de datos con que trabajaremos corresponde a datos de corte transversal. Esto implicará que nuestra ruta tendrá los siguientes pasos:

1. Encontrar diferentes modelos candidatos a ser el mejor modelo
2. Determinar si existe un problema de heteroscedasticidad en los modelos candidatos.
3. Limpiar los modelos candidatos de variables no significativas.
4. Comparar los modelos candidatos para seleccionar el mejor modelo para hacer analítica descriptiva empleando Validación cruzada!de k iteraciones
5. Determinar si existe un problema de multicolinealidad.
6. Generar escenarios para mostrar como funcionaría la herramienta

Empecemos a ejecutar esa ruta analítica para resolver la pregunta de negocio

13.4 Detección de posibles modelos

Lee los datos empleando la función **read.csv()** , guárdalos en el objeto **datos.caso2** y elimina la primera variable que no es relevante (se carga una variable “X” con el número de la observación. Esto puede cambiar de computador a computador.)

```
datos.caso2 <- read.csv("../Data/DatosCaso2.csv", sep = ",")  
datos.caso2 <- datos.caso2[, -1]
```

El siguiente paso es encontrar los mejores modelos empleando las estrategias de regresión paso a paso forward, backward y combinada con el AIC, con el valor p y el R^2 ajustado. Como lo discutimos en el Capítulo 10, dado que es posible que exista heteroscedasticidad (mas adelante lo demostraremos) es mejor no emplear el criterio del valor p.

Entonces tendremos 6 opciones de algoritmos y criterios de selección que se presentan en el Cuadro 13.1. Los modelos que obtenemos los guardaremos con los nombres que se presentan el Cuadro 13.1.

Cuadro 13.1: Modelos a estimar con los diferentes algoritmos

Nombre del objeto	Algoritmo	Criterio
modelo1.C2	Forward	R^2 ajustado
modelo2.C2	Forward	AIC
modelo3.C2	Backward	R^2 ajustado
modelo4.C2	Backward	AIC
modelo5.C2	Both	R^2 ajustado
modelo6.C2	Both	AIC

Partamos de estimar los modelos lineales con todas las variables potenciales (`max.model`).

```
# modelo con todas las variables
max.model <- lm(ViolentCrimesPerPop ~ ., data = datos.caso2)
```

Ahora procedamos a encontrar modelos candidatos para ser los mejores modelos. Empecemos con la estrategia *stepwise Forward*.

13.4.1 Stepwise forward

Empecemos por el modelo 1: algoritmo Forward y criterio de R^2 ajustado. Estímalo, guarda lo en el objeto `modelo1.C2` y ahora constatemos si el modelo 1 tiene o no heteroscedasticidad.

Para realizar la prueba de Breusch-Pagan, debemos constatar si el supuesto de normalidad se cumple. Para esto extraigamos los residuales con la función `residuals()` y efectuemos las pruebas de normalidad con la función `ks.test()` del paquete base (Prueba de Kolmogorov-Smirnov) y la función `jarque.bera.test()` del paquete `tseries` (Trapletti y Hornik, 2019) (Prueba de Jarque-Bera).

```
# se extraen los residuales
res.modelo1.C2 <- residuals(modelo1.C2)
# pruebas de normalidad Kolmogorov-Smirnov
ks.test(res.modelo1.C2, y = pnorm)

##
## One-sample Kolmogorov-Smirnov test
##
## data: res.modelo1.C2
## D = 0.38847, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# pruebas de normalidad Kolmogorov-Smirnov
library(tseries)
jarque.bera.test(res.modelo1.C2)

##
##  Jarque Bera Test
##
## data: res.modelo1.C2
## X-squared = 1185.3, df = 2, p-value < 2.2e-16
```

Los residuales de este modelo no siguen una distribución normal. Por eso no son confiables los resultados de la prueba de Breusch-Pagan tradicional. Deberíamos entonces emplear la versión studentizada de la prueba propuesta por Koenker (1981)

```
# se carga la librería
library(lmtest)
# prueba de Breusch-Pagan studentizada
bptest(modelo1.C2, studentize = TRUE)

##
## studentized Breusch-Pagan test
##
## data: modelo1.C2
## BP = 307.37, df = 65, p-value < 2.2e-16
```

Con un 99 % de confianza podemos rechazar la hipótesis de homoscedasticidad. Y para la prueba de White obtendremos el mismo resultado (¡Inténtalo!).

Por tanto eliminar las variables no significativas podemos emplear la función construída en el Capítulo10: **remueve.no.sinifica.HC3()**.

El modelo 1 después de la eliminación de las variables explicativas no significativas (con la corrección HC3) se reporta en el Cuadro 13.2.

```
# remueve las variables no significativas HC3
modelo1.C2 <- remueve.no.sinifica.HC3(modelo1.C2, 0.05)
```

Cuadro 13.2: Modelo 1 estimado por MCO y corrección HC

	<i>Dependent variable:</i>	
	ViolentCrimesPerPop	
	MCO	HC3
	(1)	(2)
racepctblack	0.196*** (0.030)	0.196*** (0.036)
agePct12t29	-0.263*** (0.067)	-0.263*** (0.077)
pctUrban	0.040*** (0.009)	0.040*** (0.009)
pctWWage	-0.261*** (0.060)	-0.261*** (0.056)
pctWFarmSelf	0.050*** (0.019)	0.050*** (0.017)
pctWInvInc	-0.156*** (0.057)	-0.156*** (0.054)
pctWRetire	-0.063** (0.031)	-0.063** (0.028)
medFamInc	0.185** (0.090)	0.185** (0.085)
whitePerCap	-0.190*** (0.070)	-0.190*** (0.071)
indianPerCap	-0.036* (0.019)	-0.036** (0.016)
OtherPerCap	0.050*** (0.017)	0.050*** (0.017)
PctPopUnderPov	-0.095** (0.043)	-0.095** (0.045)
PctEmploy	0.203*** (0.059)	0.203*** (0.058)
MalePctNevMarr	0.226*** (0.054)	0.226*** (0.061)
FemalePctDiv	-0.286*** (0.112)	-0.286*** (0.108)
TotalPctDiv	0.336*** (0.114)	0.336*** (0.109)
PctKids2Par	-0.330*** (0.071)	-0.330*** (0.083)
PctWorkMom	-0.129*** (0.027)	-0.129*** (0.028)
PctIlleg	0.137*** (0.041)	0.137*** (0.052)
NumImmig	-0.174*** (0.059)	-0.174** (0.069)
PctNotSpeakEnglWell	-0.153*** (0.048)	-0.153*** (0.053)
PctLargHouseFam	-0.097* (0.051)	-0.097** (0.039)
PersPerOccupHous	0.270*** (0.087)	0.270*** (0.075)
PersPerRentOccHous	-0.219*** (0.070)	-0.219*** (0.054)
PctPersOwnOccup	-0.634*** (0.206)	-0.634*** (0.196)
PctPersDenseHous	0.280*** (0.059)	0.280*** (0.067)
HousVacant	0.175*** (0.033)	0.175*** (0.041)
PctHousOccup	-0.056** (0.023)	-0.056** (0.023)
PctHousOwnOcc	0.574*** (0.203)	0.574*** (0.189)
PctVacMore6Mos	-0.046** (0.022)	-0.046** (0.020)
OwnOccLowQuart	-0.435*** (0.159)	-0.435*** (0.149)
OwnOccMedVal	0.323** (0.154)	0.323** (0.146)
RentLowQ	-0.227*** (0.054)	-0.227*** (0.043)
MedRent	0.267*** (0.068)	0.267*** (0.053)
MedOwnCostPctIncNoMtg	-0.100*** (0.021)	-0.100*** (0.021)
NumStreet	0.202*** (0.043)	0.202*** (0.063)
PctForeignBorn	0.136*** (0.041)	0.136*** (0.044)
Constant	0.645*** (0.093)	0.645*** (0.107)
Observations	1,993	1,993
R ²	0.685	0.685
Adjusted R ²	0.679	0.679
Residual Std. Error (df = 1955)	0.132	0.132

Note:

*p<0.1; **p<0.05; ***p<0.01

Ahora sigamos con el modelo 2: algoritmo Forward y criterio AIC. Tu puedes encontrar que este modelo también tiene heteroscedasticidad (*¿Hazlo!*). Ahora realicemos la corrección HC3 y eliminemos las variables no significativas. El resultado se reporta en el Cuadro 13.3

Cuadro 13.3: Modelo seleccionado el algoritmo stepwise forward con corrección HC3

	<i>Dependent variable:</i>	
	ViolentCrimesPerPop	
	Modelo 1 (HC3) (R^2 aj)	Modelo 2 (HC3) (AIC)
	(1)	(2)
racepctblack	0.196*** (0.036)	0.220*** (0.033)
agePct12t29	-0.263*** (0.077)	-0.199*** (0.067)
pctUrban	0.040*** (0.009)	0.047*** (0.009)
pctWWage	-0.261*** (0.056)	-0.186*** (0.047)
pctWFarmSelf	0.050*** (0.017)	0.038** (0.017)
pctWInvInc	-0.156*** (0.054)	-0.156*** (0.055)
pctWRetire	-0.063** (0.028)	-0.079*** (0.028)
medFamInc	0.185** (0.085)	
whitePerCap	-0.190*** (0.071)	-0.137*** (0.048)
indianPerCap	-0.036** (0.016)	-0.036** (0.016)
PctBSorMore		0.099*** (0.035)
OtherPerCap	0.050*** (0.017)	0.051*** (0.017)
PctPopUnderPov	-0.095** (0.045)	-0.170*** (0.048)
PctEmploy	0.203*** (0.058)	0.155*** (0.054)
MalePctNevMarr	0.226*** (0.061)	0.134** (0.056)
FemalePctDiv	-0.286*** (0.108)	
TotalPctDiv	0.336*** (0.109)	-0.267** (0.108)
MedRentPctHousInc		0.061** (0.024)
PctKids2Par	-0.330*** (0.083)	-0.356*** (0.081)
PctWorkMom	-0.129*** (0.028)	-0.144*** (0.026)
PctIlleg	0.137*** (0.052)	0.134*** (0.050)
NumImmig	-0.174** (0.069)	
PctNotSpeakEnglWell	-0.153*** (0.053)	
PctLargHouseFam	-0.097** (0.039)	
PersPerOccupHous	0.270*** (0.075)	
PersPerRentOccHous	-0.219*** (0.054)	
PctPersOwnOccup	-0.634*** (0.196)	
PctPersDenseHous	0.280*** (0.067)	0.202*** (0.028)
HousVacant	0.175*** (0.041)	0.278*** (0.063)
PctHousOccup	-0.056** (0.023)	
PctHousOwnOcc	0.574*** (0.189)	
PctVacMore6Mos	-0.046** (0.020)	
OwnOccLowQuart	-0.435*** (0.149)	
OwnOccMedVal	0.323** (0.146)	
RentLowQ	-0.227*** (0.043)	-0.218*** (0.040)
MedRent	0.267*** (0.053)	0.221*** (0.048)
MedOwnCostPctIncNoMtg	-0.100*** (0.021)	-0.105*** (0.020)
NumStreet	0.202*** (0.063)	0.178*** (0.057)
PctForeignBorn	0.136*** (0.044)	
numbUrban		-0.201** (0.082)
MalePctDivorce		0.347*** (0.097)
Constant	0.645*** (0.107)	0.583*** (0.092)
Observations	1,993	1,993
R ²	0.685	0.680
Adjusted R ²	0.679	0.676
Residual Std. Error	0.132 (df = 1955)	0.133 (df = 1965)

Note:

*p<0.1; **p<0.05; ***p<0.01

13.4.2 Stepwise backward

De manera similar en el Cuadro 13.4 se presentan los resultados de emplear el algoritmo *stepwise backward* y tras limpiar las variables no significativas con la corrección HC3 (con un 95 % de confianza). Previamente puedes mostrar que estos dos modelos tienen un problema de heteroscedasticidad.

Cuadro 13.4: Modelos seleccionados con el algoritmo stepwise backward con corrección HC3

	<i>Dependent variable:</i>	
	ViolentCrimesPerPop	
	Modelo 3 (HC3) (R^2 aj)	Modelo 4 (HC3) (AIC)
	(1)	(2)
racePctBlack	0.181*** (0.036)	
agePct12t29	-0.224*** (0.075)	
pctUrban	0.040*** (0.009)	
pctWWage	-0.262*** (0.057)	
pctWFarmSelf	0.050*** (0.017)	
pctWInvInc	-0.167*** (0.057)	
pctWRetire	-0.075*** (0.028)	
whitePerCap	-0.108** (0.052)	
indianPerCap	-0.033** (0.016)	
OtherPerCap	0.049*** (0.017)	
PctPopUnderPov	-0.137*** (0.047)	
PctEmploy	0.162*** (0.057)	
PctOccupMgmtProf	0.100** (0.039)	
MalePctDivorce	0.320*** (0.096)	
MalePctNevMarr	0.203*** (0.059)	
TotalPctDiv	-0.241** (0.111)	
PctKids2Par	-0.358*** (0.083)	
PctWorkMom	-0.124*** (0.028)	
PctIlleg	0.138*** (0.052)	
NumImmig	-0.171** (0.067)	
PctNotSpeakEnglWell	-0.135** (0.053)	
PctLargHouseOccup	-0.140*** (0.044)	
PersPerOccupHous	0.358*** (0.073)	
PersPerRentOccHous	-0.222*** (0.053)	
PctPersOwnOccup	-0.592*** (0.199)	
PctPersDenseHous	0.268*** (0.067)	
HousVacant	0.161*** (0.040)	
PctHousOccup	-0.050** (0.023)	
PctHousOwnOcc	0.547*** (0.191)	
PctVacMore6Mos	-0.044** (0.020)	
OwnOccLowQuart	-0.089** (0.037)	
RentLowQ	-0.219*** (0.042)	
MedRent	0.270*** (0.054)	
MedOwnCostPctIncNoMtg	-0.094*** (0.021)	
NumStreet	0.204*** (0.062)	
PctForeignBorn	0.130*** (0.044)	
PctFam2Par		-0.580*** (0.043)
racePctAsian		-0.072*** (0.022)
PctSameState85		-0.050** (0.024)
agePct12t21		-0.102*** (0.026)
MedYrHousBuilt		0.070*** (0.018)
racePctWhite		-0.300*** (0.031)
numUrban		0.242*** (0.036)
PersPerFam		0.104*** (0.032)
blackPerCap		-0.047** (0.019)
PctSameCity85		0.073** (0.029)
RentHighQ		0.078*** (0.018)
PctWorkMomYoungKids		-0.078*** (0.020)
PctHousLess3BR		0.062** (0.030)
Constant	0.636*** (0.105)	0.748*** (0.046)
Observations	1,993	1,993
R ²	0.685	0.633
Adjusted R ²	0.679	0.631
Residual Std. Error	0.132 (df = 1956)	0.142 (df = 1979)

Note:

*p<0.1; **p<0.05; ***p<0.01

13.4.3 Combinando forward y backward

Y finalmente, el Cuadro 13.5 se presentan los resultados de emplear el algoritmo combinado y tras limpiar las variables no significativas y la corrección HC3 (con un 95 % de confianza). (Recuerda hacer las pruebas de heteroscedasticidad para cada modelo)

Cuadro 13.5: Modelo seleccionado el algoritmo stepwise forward y backward con corrección HC3

	<i>Dependent variable:</i>	
	ViolentCrimesPerPop	
	Modelo 5 (HC3) (R^2 aj)	Modelo 6 (HC3) (AIC)
	(1)	(2)
racePctBlack	0.173*** (0.033)	0.193*** (0.031)
agePct12t29	-0.202*** (0.075)	-0.204*** (0.068)
pctUrban	0.041*** (0.009)	0.038*** (0.009)
pctWWage	-0.247*** (0.054)	-0.261*** (0.053)
pctWFarmSelf	0.047*** (0.017)	0.045*** (0.017)
pctWInvInc	-0.141*** (0.048)	-0.181*** (0.051)
pctWRetire	-0.068** (0.027)	-0.085*** (0.027)
indianPerCap	-0.032** (0.016)	-0.034** (0.016)
TotalPctDiv		-0.320*** (0.104)
PctVacMore6Mos		-0.039** (0.019)
OtherPerCap	0.047*** (0.017)	0.052*** (0.017)
PctPopUnderPov	-0.109** (0.045)	-0.099** (0.044)
PctEmploy	0.141** (0.055)	0.144*** (0.055)
MalePctDivorce	0.118*** (0.044)	0.373*** (0.091)
MalePctNevMarr	0.213*** (0.060)	0.194*** (0.057)
PctKids2Par	-0.290*** (0.079)	-0.376*** (0.082)
PctWorkMom	-0.111*** (0.025)	-0.121*** (0.024)
PctIlleg	0.152*** (0.052)	0.159*** (0.051)
NumImmig	-0.160** (0.067)	
PctNotSpeakEnglWell	-0.150*** (0.052)	
PctLargHouseFam	-0.107*** (0.039)	-0.145*** (0.039)
PctLess9thGrade		-0.092*** (0.029)
PersPerOccupHous	0.342*** (0.069)	0.151*** (0.050)
PersPerRentOccHous	-0.242*** (0.053)	
PctPersOwnOccup	-0.611*** (0.197)	
NumIlleg		-0.179** (0.073)
PctPersDenseHous	0.262*** (0.066)	0.265*** (0.042)
HousVacant	0.155*** (0.040)	0.225*** (0.047)
PctHousOccup	-0.049** (0.022)	
PctHousOwnOcc	0.552*** (0.190)	
OwnOccLowQuart	-0.121*** (0.033)	
RentLowQ	-0.215*** (0.042)	-0.222*** (0.041)
MedRent	0.280*** (0.053)	0.202*** (0.045)
MedOwnCostPctIncNoMtg	-0.102*** (0.020)	-0.078*** (0.020)
NumStreet	0.205*** (0.062)	0.177*** (0.055)
PctForeignBorn	0.140*** (0.043)	
Constant	0.531*** (0.095)	0.631*** (0.103)
Observations	1,993	1,993
R ²	0.683	0.680
Adjusted R ²	0.677	0.676
Residual Std. Error	0.132 (df = 1960)	0.133 (df = 1964)

Note:

*p<0.1; **p<0.05; ***p<0.01

13.5 Selección del mejor modelo para predecir

En resumen, contamos con 6 modelos. Todos los modelos se encuentran anidados en el modelo 1 y cada modelo es diferente. En el Cuadro 13.6 se presentan los 6 modelos.

Cuadro 13.6: Modelos seleccionados por los diferentes algoritmos con corrección HC3

	Dependent variable:					
	ViolentCrimesPerPop					
	Modelo 1 (1)	Modelo 2 (2)	Modelo 3 (3)	Modelo 4 (4)	Modelo 5 (5)	Modelo 6 (6)
raceptblack	0.196*** (0.036)	0.220*** (0.033)	0.181*** (0.036)		0.173*** (0.033)	0.193*** (0.031)
agePct1229	-0.263*** (0.077)	-0.199*** (0.067)	-0.224*** (0.075)		-0.202*** (0.075)	-0.204*** (0.068)
pctUrban	0.040*** (0.009)	0.047*** (0.009)	0.040*** (0.009)		0.041*** (0.009)	0.038*** (0.009)
pctWWage	-0.261*** (0.056)	-0.186*** (0.047)	-0.262*** (0.057)		-0.247*** (0.054)	-0.261*** (0.053)
pctWFarmSelf	0.050*** (0.017)	0.038** (0.017)	0.050*** (0.017)		0.047*** (0.017)	0.045*** (0.017)
pctWInvInc	-0.156*** (0.054)	-0.156*** (0.055)	-0.167*** (0.057)		-0.141*** (0.048)	-0.181*** (0.051)
pctWRetire	-0.063** (0.028)	-0.079*** (0.028)	-0.075*** (0.028)		-0.068** (0.027)	-0.085*** (0.027)
medFamInc	0.185** (0.085)					
whitePerCap	-0.190*** (0.071)	-0.137*** (0.048)	-0.108** (0.052)			
indianPerCap	-0.036** (0.016)	-0.036** (0.016)	-0.033** (0.016)		-0.032** (0.016)	-0.034** (0.016)
PctBSorMore		0.099*** (0.035)				
OtherPerCap	0.050*** (0.017)	0.051*** (0.017)	0.049*** (0.017)		0.047*** (0.017)	0.052*** (0.017)
PctPopUnderPov	-0.095** (0.045)	-0.170*** (0.048)	-0.137*** (0.047)		-0.109** (0.045)	-0.099** (0.044)
PctEmploy	0.203*** (0.058)	0.155*** (0.054)	0.162*** (0.057)		0.141** (0.055)	0.144*** (0.055)
PctOccupMgmtProf			0.100** (0.039)			
MalePctNevMarr	0.226*** (0.061)	0.134** (0.056)	0.203*** (0.059)		0.213*** (0.060)	0.194*** (0.057)
FemalePctDiv	-0.286*** (0.108)					
TotalPctDiv	0.336*** (0.109)	-0.267** (0.108)	-0.241** (0.111)			-0.320*** (0.104)
MedRentPctHousInc		0.061** (0.024)				
PctKids2Par	-0.330*** (0.083)	-0.356*** (0.081)	-0.358*** (0.083)		-0.290*** (0.079)	-0.376*** (0.082)
PctWorkMom	-0.129*** (0.028)	-0.144*** (0.026)	-0.124*** (0.028)		-0.111*** (0.025)	-0.121*** (0.024)
PctIlleg	0.137*** (0.052)	0.134*** (0.050)	0.138*** (0.052)		0.152*** (0.052)	0.159*** (0.051)
NumImmig	-0.174** (0.069)		-0.171** (0.067)		-0.160** (0.067)	
PctNotSpeakEnglWell	-0.153*** (0.053)		-0.135** (0.053)		-0.150*** (0.052)	
PctLargHouseFam	-0.097** (0.039)				-0.107*** (0.039)	-0.145*** (0.039)
PctLargHouseOccup			-0.140*** (0.044)			
PctLess9thGrade						-0.092*** (0.029)
PersPerOccupHous	0.270*** (0.075)		0.358*** (0.073)		0.342*** (0.069)	0.151*** (0.050)
PersPerRentOccHous	-0.219*** (0.054)		-0.222*** (0.053)		-0.242*** (0.053)	

Ahora comparemos el comportamiento de los modelos para predecir fuera de muestra. Emplearemos el método de k iteraciones para hacer la validación cruzada. Empleemos, el paquete *caret* (Kuhn, 2020) y las funciones **trainControl()** y **train()** (tal como lo discutimos en el Capítulo 12). Además empleemos 5 iteraciones ($k = 5$).

```
# se fija el método de k iteraciones
train.control <- trainControl(method = "cv", number = 5)

# se realiza la evaluación fuera de muestra
fold.modelo1 = train(formula(modelo1.C2), data = datos.caso2,
                     method = "lm", trControl = train.control)
fold.modelo2 = train(formula(modelo2.C2), data = datos.caso2,
                     method = "lm", trControl = train.control)
fold.modelo3 = train(formula(modelo3.C2), data = datos.caso2,
                     method = "lm", trControl = train.control)
fold.modelo4 = train(formula(modelo4.C2), data = datos.caso2,
                     method = "lm", trControl = train.control)
fold.modelo5 = train(formula(modelo5.C2), data = datos.caso2,
                     method = "lm", trControl = train.control)
fold.modelo6 = train(formula(modelo6.C2), data = datos.caso2,
                     method = "lm", trControl = train.control)
```

Resumamos los resultados usando el RMSE y el MAE de cada modelo como se presenta en el Cuadro 13.7.

Cuadro 13.7: Metricas de precisión de los 6 modelos con 5 iteraciones

	RMSE	MAE
Modelo 1	0.1332	0.09348
Modelo 2	0.1337	0.09333
Modelo 3	0.1333	0.09350
Modelo 4	0.1419	0.09954
Modelo 5	0.1339	0.09378
Modelo 6	0.1340	0.09378

Aunque las métricas no varían mucho entre modelos, de acuerdo con la RMSE el mejor modelo es el 1 y de acuerdo al MAE el mejor modelo es el 3. Noten que en este caso puede tener sentido penalizar errores de predicción grandes y favorecer los errores pequeños. Así, emplearemos el criterio del RMSe que sugiere el modelo 1. Este modelo se construyó a partir del algoritmo Forward y empleando el criterio de R^2 ajustado.

Ahora, antes de pasar a emplear el modelo, procedamos a determinar si el mejor modelo tiene o no multicolinealidad. Nota que el *VIF* depende de la matriz de varianzas y covarianzas de los estimadores MCO y en presencia de heteroscedasticidad, éstos no son confiables. Así no podemos emplearlos en esta situación. Así que nos concentraremos en la prueba de Belsley y col. (1980) también conocida como la prueba Kappa.

El siguiente código permite hacer la prueba.

```
XTX <- model.matrix(modelo1.C2)
e <- eigen(t(XTX) %*% XTX)

lambda.1 <- max(e$val)
lambda.k <- min(e$val)
kappa <- sqrt(lambda.1/lambda.k)
kappa

## [1] 270.928
```

Este estadístico es muy grande ($\kappa = 270.9280298$). Esta prueba sugiere la existencia de un problema serio de multicolinealidad. Solucionar este problema no será necesario, pues la multicolinealidad alta nos está ayudando a mejorar el R^2 del modelo y cómo nuestro objetivo es hacer predicciones, no será necesario solucionar el problema.

13.6 Predicciones (escenarios)

El modelo construido nos permite generar los escenarios que estaba buscando la empresa consultora. Recuerda que la pregunta de negocio implicaba tener una “fórmula” que le permita determinar cuál sería la tasa de crímenes violentos por cada 100000 habitantes bajo diferentes escenarios.

Por ejemplo, supongamos que queremos tener un escenario en el que todas las variables se encuentran en sus medias.

Recuerden que tenemos heteroscedasticidad y el modelo no cumple el supuesto de normalidad. Por eso deberemos emplear la técnica de bootstrapping para crear los intervalos de confianza para las predicciones. Esto lo podemos hacer empleando la función **ic.boot.predic()** que construimos en la sección 12.7.

```
data.temp = as.data.frame(t(apply(datos.caso2, 2, mean)))

set.seed(123445)
ic_boot_predic(modelo1.C2, data.temp, R = 1000, alpha = 0.05)

##      fit.1    lwr.2.5%   upr.97.5%
##  0.23798294 -0.04723444  0.59553341
```

Ahora la empresa de consultoría puede jugar con este modelo mostrando diferentes escenarios. Por ejemplo, supongamos que visitamos el municipio de la fila 120 de la base de datos. Esta municipio tiene las siguientes características (\mathbf{x}_{120})

```
modelo1.C2$model[120, ]

##      ViolentCrimesPerPop racepctblack agePct12t29 pctUrban
## 120          0.68          0.09          0.34          1
##      pctWWage pctWFarmSelf pctWInvInc pctWRetire medFamInc
## 120          0.49          0.2          0.56          0.26          0.55
##      whitePerCap indianPerCap OtherPerCap PctPopUnderPov
## 120          0.88          0.24          0.36          0.24
##      PctEmploy MalePctNevMarr FemalePctDiv TotalPctDiv
## 120          0.59          0.73          0.81          0.82
##      PctKids2Par PctWorkMom PctIlleg NumImmig PctNotSpeakEnglWell
## 120          0.59          0.43          0.29          0.15          0.3
##      PctLargHouseFam PersPerOccupHous PersPerRentOccHous
## 120          0.2          0.05          0.07
##      PctPersOwnOccup PctPersDenseHous HousVacant PctHousOccup
## 120          0.12          0.25          0.15          0.77
##      PctHousOwnOcc PctVacMore6Mos OwnOccLowQuart OwnOccMedVal
## 120          0.08          0.5          1          1
##      RentLowQ MedRent MedOwnCostPctIncNoMtg NumStreet
## 120          0.43          0.42          0.17          0.84
##      PctForeignBorn
## 120          0.75
```

Y la tasa de crímenes violentos por cada 100000 habitantes de ese municipio es de:

```
datos.caso2[120, 101]

## [1] 0.68
```

Ahora, se le podría proponer al alcalde una política pública para aumentar el porcentaje de personas de 16 años o más que están empleadas (variable *PctEmploy*) en 5 punto porcentuales de tal manera que pase de 59 % a 64 %. Entonces en ese caso se esperaría que la tasa de crímenes violentos por cada 100000 habitantes de ese municipio cambiaría a:

```
# se extraen los datos para el municipio 10
data.escenario <- modelo1.C2$model[120, ]
# se modifica los datos del municipio 10 para ajustarse al
# escenario
data.escenario["PctEmploy"] <- data.escenario["PctEmploy"] +
  0.05

# predicción

ic_boot_predic(modelo1.C2, data.escenario, R = 1000, alpha = 0.05)

##   fit.120 lwr.2.5% upr.97.5%
```

```
## 0.5044719 0.1684555 0.8725683
```

Nota que esto implicaría que el intervalo de confianza de la predicción sigue conteniendo el valor original de la tasa de crímenes violentos por cada 100000 habitantes, por tanto ese escenario no cambiaría la tasa esperada. Ahora los consultores pueden jugar con esta fórmula para proponer diferentes escenarios a cada alcalde para disminuir la tasa de crímenes violentos o si es del caso saber cuáles municipios no tiene sentido visitar, pues no se puede hacer mucho para disminuir dicha tasa. De aquí en adelante, la imaginación es le límite.

@

Parte V

Apéndices: Conceptos básicos álgebra matricial y estadística



14 . Elementos de álgebra matricial

Diseñado por Freepik

Objetivos del capítulo

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

- Explicar en sus propias palabras los conceptos de: matriz, suma y multiplicación de matrices, matriz identidad, transpuesta de una matriz e inversa de una matriz
- Calcula en R la suma y multiplicación de matrices, la matriz identidad, la transpuesta de una matriz y la inversa de una matriz

14.1 Introducción

Este libro supone el conocimiento básico de álgebra matricial que le permiten al científico de datos trabajar con grandes volúmenes de datos organizados en forma matricial. Este Apéndice presenta dichos elementos necesarios para seguir algunas demostraciones y operaciones descritas en el libro. Este Apéndice no pretende ser un tratado autocontenido de álgebra matricial, sino por el contrario un breve resumen que permitirá al lector ya familiarizado con el álgebra matricial recordar los concepto¹.

El concepto de matrices es una noción que inicialmente desarrollado en el siglo XVII, asociado a la manipulación de gráficos y soluciones de ecuaciones lineales simultáneas. Hoy en día la aplicación de las matrices y sus operaciones están ligadas a áreas tan diversas como la física, gráficos de computador, la economía, los métodos estadísticos, la teoría de juegos, redes, criptología y la ciencia de datos.

Una matriz es una forma eficiente de ordenar información en columnas y filas. Por ejemplo, consideremos una base de datos que contiene información de ventas, costos y utilidades mensuales para dos empresas (empresa 1 y 2) en millones de dólares. La información de un mes se puede ordenar fácilmente en una matriz de la siguiente forma:

$$M = \begin{bmatrix} \text{Empresa} & \text{Ventas} & \text{Costos} & \text{Utilidades} \\ \text{Empresa1} & 4 & 2 & 2 \\ \text{Empresa2} & 2 & 1,5 & 0,5 \end{bmatrix} \quad (14.1)$$

Generalmente, se omiten los nombres que toman cada una de las columnas y filas de las matrices; es decir, (14.1) se puede reescribir como

$$M = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 1,5 & 0,5 \end{bmatrix}$$

El tamaño de una matriz está determinado por el número de filas y de columnas; así, se dice que la matriz M es una matriz de dimensiones 2×3 (2 filas por 3 columnas). Una forma rápida de escribir esto es $M_{2 \times 3}$. En general una matriz puede tener n filas y m columnas, y se representa así:

$$A_{n \times m} = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} = [a_{ij}]_{i=1,\dots,n; j=1,\dots,m}$$

En el caso especial cuando $m = 1$, la matriz A se conoce como un vector columna. Si $n = 1$, la matriz A se conoce como un vector fila². Cuando el número de filas y de columnas es el mismo ($n = m$), la matriz A es llamada matriz cuadrada de orden n . Si $n = m = 1$, entonces A es conocido como un escalar (lo que coloquialmente se conoce como un número).

¹Este apéndice corresponde a una versión adaptada de Alonso (2006b).

²En este Apéndice y en general en el libro cuando empleemos el término vector nos referiremos a un vector columna, a menos que se especifique lo contrario.

En el caso de las matrices, se dice que $A = B$ si y solamente si todos los elementos de la matriz A son iguales a los de B , es decir $a_{ij} = b_{ij}$ para todo i y j , donde $i = 1, \dots, n$ y $j = 1, \dots, m$.

A continuación repasaremos rápidamente las operaciones matriciales, definiciones y resultados más importantes que serán útiles a lo largo del libro.

14.2 Matriz triangular y diagonal

Antes de definir algunas matrices especiales que serán útiles, es importante definir el concepto de diagonal principal de una matriz. La diagonal principal de una matriz cuadrada A , es el conjunto de elementos cuya posición corresponden a la misma fila y columna. En otras palabras, son los elementos que se encuentran formando una “diagonal” entre la esquina superior izquierda y la esquina inferior derecha de la matriz. Por ejemplo, sea la matriz

$$C = \begin{bmatrix} 1 & 2 & 4 \\ 6 & 3 & 5 \\ 11 & 10 & 8 \end{bmatrix}$$

. La diagonal principal de la matriz C está dada por el conjunto $1, 3, 8$.

En general, sea A una matriz cuadrada representada por:

$$A_{n \times m} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = [a_{ij}]_{i=1,\dots,n; j=1,\dots,n} \quad (14.2)$$

Entonces, la **diagonal principal** está dada por el conjunto de elementos $\{a_{ii}\}_{i=1,\dots,n}$.

Por otro lado, una matriz triangular superior (inferior) es una matriz cuadrada con todos los elementos por debajo (encima) de la diagonal principal iguales a cero. Por ejemplo, la siguiente matriz:

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 5 \\ 0 & 0 & 8 \end{bmatrix}$$

es una matriz triangular superior.

Si una matriz es al mismo tiempo una matriz triangular superior y triangular inferior, entonces se conoce como una **matriz diagonal**. En otras palabras, una matriz diagonal es una matriz cuyos elementos por fuera de la diagonal principal son cero. Es decir, A es una matriz diagonal si tiene la siguiente forma

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & a_{nn} \end{bmatrix}$$

Una matriz diagonal se puede escribir de forma corta de la siguiente manera $A = \text{diag}(a_{11}, a_{22}, \dots, a_{nn})$.

14.3 Adición, multiplicación por un escalar y multiplicación de matrices

Sean dos matrices A y B , cada una de dimensiones $n \times m$. Es decir, las dos matrices tienen el mismo número de filas y columnas, cuando ocurre esto se le denomina a las dos matrices conformes para la suma. La suma de estas dos matrices corresponde a una matriz con dimensiones $n \times m$, cuyos elementos son iguales a la suma de los elementos correspondientes de las matrices A y B . En otras palabras,

$$\begin{aligned} A + B &= \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nm} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1m} + b_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & \dots & a_{nm} + b_{nm} \end{bmatrix} \\ &= [a_{ij} + b_{ij}]_{i=1,\dots,n; j=1,\dots,m} \end{aligned}$$

Si las matrices a sumar no tienen las mismas dimensiones, entonces la operación no se puede efectuar y se dice que las matrices no cumplen la condición de conformidad.

Ejemplo 14.1 Suma de matrices

Suponga que contamos con 3 matrices que corresponden a la información para los tres primeros meses del año de las ventas, costos y utilidades mensuales (columnas) para dos empresas (filas).

$$E = \begin{bmatrix} 600 & 250 & 350 \\ 550 & 180 & 400 \end{bmatrix} \quad F = \begin{bmatrix} 650 & 330 & 250 \\ 600 & 270 & 400 \end{bmatrix} \quad M = \begin{bmatrix} 580 & 270 & 350 \\ 6250 & 350 & 410 \end{bmatrix}$$

Encuentre el valor de las ventas, costos y utilidades para el primer trimestre.

Respuesta: Las ventas, costos y utilidades trimestrales para las dos empresas son:

$$E + F + M = \begin{bmatrix} 1830 & 850 & 950 \\ 1775 & 800 & 1210 \end{bmatrix}$$

La suma de matrices posee varias propiedades similares a las de la suma de escalares. A continuación se exponen estas propiedades:

- Propiedad Comutativa: $A + B = B + A$
- Propiedad Asociativa: $A + B + C = (A + B) + C = A + (B + C)$

Antes de avanzar un poco más, note que si se suma dos veces la matriz A , se obtiene:

$$\begin{aligned}
 A + A &= \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} + \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \\
 &= \begin{bmatrix} 2a_{11} & \dots & 2a_{1n} \\ \vdots & \ddots & \vdots \\ 2a_{n1} & \dots & 2a_{nn} \end{bmatrix} = 2 \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \\
 &= 2 \cdot [a_{ij}]_{i=1,\dots,n;j=1,\dots,n} = 2 \cdot A
 \end{aligned}$$

Es decir, el resultado de sumar dos veces la matriz A es igual a cada uno de los elementos de la matriz A multiplicado por dos. Así es fácil mostrar que en general:

$$\lambda \cdot A = \begin{bmatrix} \lambda a_{11} & \dots & \lambda a_{1n} \\ \vdots & \ddots & \vdots \\ \lambda a_{n1} & \dots & \lambda a_{nn} \end{bmatrix}$$

donde λ es un escalar.

Ejemplo 14.2 Continuación

Suponga que quiere calcular $E - F + 2M$.

Respuesta: Tenemos que:

$$E - F + 2M = E + (-1)F + 2M$$

Por lo tanto,

$$E - F + 2M = \begin{bmatrix} 600 & 250 & 350 \\ 550 & 180 & 400 \end{bmatrix} + \begin{bmatrix} -650 & -330 & -250 \\ -600 & -270 & -400 \end{bmatrix} + \begin{bmatrix} 2 \cdot 580 & 2 \cdot 270 & 2 \cdot 350 \\ 2 \cdot 6250 & 2 \cdot 350 & 2 \cdot 410 \end{bmatrix}$$

Y por lo tanto

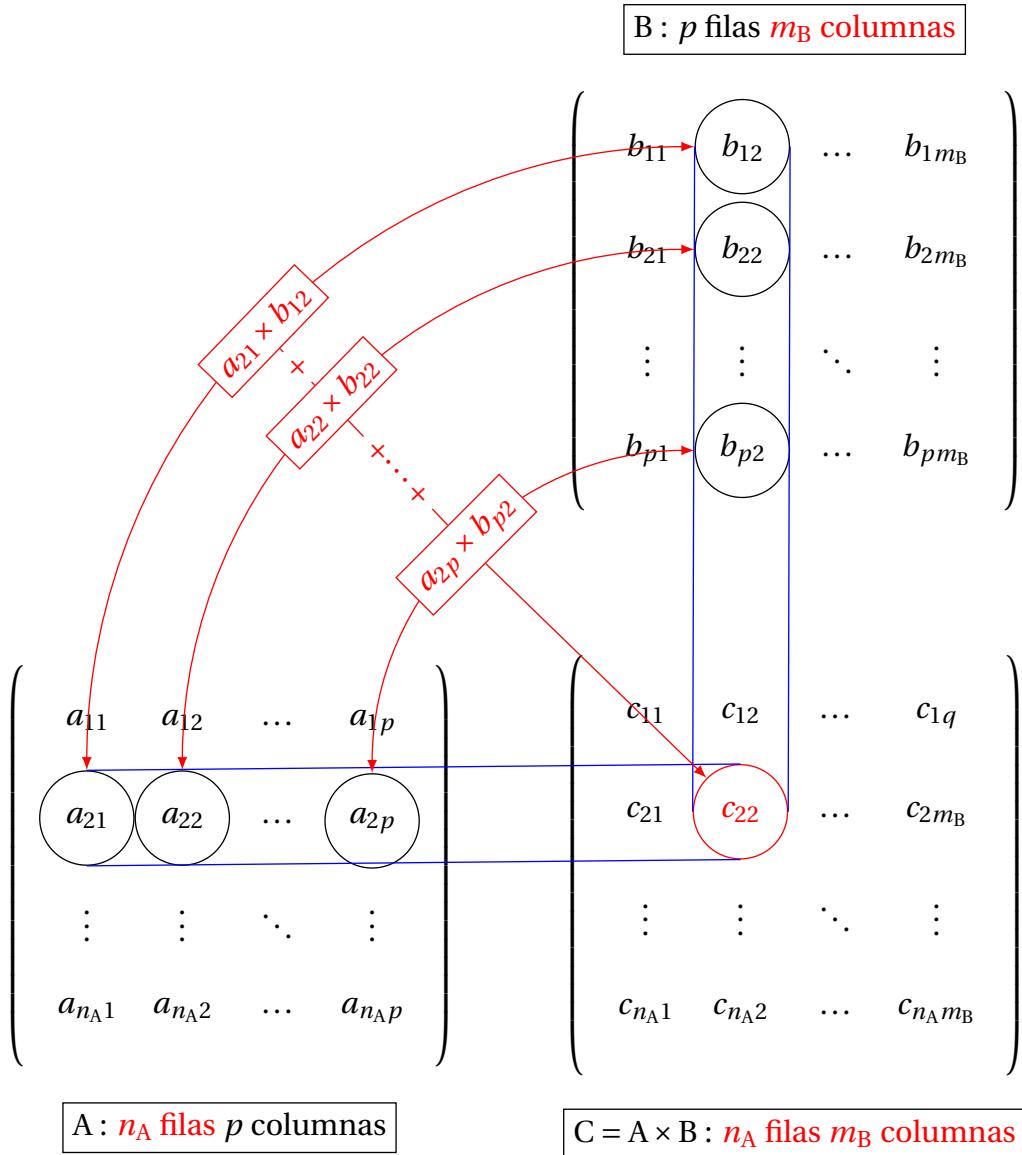
$$E - F + 2M = \begin{bmatrix} 1110 & 460 & 800 \\ 1200 & 610 & 820 \end{bmatrix}$$

Finalmente, consideremos la multiplicación de dos matrices. En este caso, a diferencia de la multiplicación entre escalares, las matrices a multiplicar deben cumplir una condición de conformidad para que el producto exista. Esta condición es que el número de columnas de la primera matriz debe ser igual al número de filas de la segunda matriz. Es decir, dadas dos matrices A y B con dimensiones $n_A \times m_A$ y $n_B \times m_B$, respectivamente; el producto $A \cdot B$ cumple la condición de conformidad si y solamente si $m_A = n_B$. En caso de que esta condición se cumpla, el producto estará dado por

$$A \cdot B = [c_{ij}]_{i=1,\dots,n;j=1,\dots,m} \tag{14.3}$$

donde $c_{ij} = \sum_{l=1}^m a_{il} \cdot b_{il}$. La Figura 14.1 ilustra el uso de esta fórmula.

Figura 14.1. Diagrama de multiplicación de matrices



Fuente: Este diagrama es una modificación del encontrado en <https://texexample.net/tikz/examples/matrix-multiplication/>
Nota: Esta representación se conoce como el esquema de Falk. Cada una de las parejas unidas por las líneas rojas se multiplican entre sí. Posteriormente, se suman dichos productos para obtener el correspondiente elemento de la matriz final C .

La multiplicación de matrices presenta varias propiedades, pero antes es importante recalcar que la propiedad conmutativa de la multiplicación para los escalares no se cumple para las matrices (aún si este producto cumple la condición de conformidad). Es decir, $A \cdot B \neq B \cdot A$ en caso de que ambos productos estén definidos. Por esto, es importante tener en cuenta el orden en que se multiplican las matrices, y hablaremos de pre-multiplicar o post-multiplicar por una matriz, cuando se multiplica una matriz por la izquierda o por la derecha, respectivamente.

Las siguientes son las propiedades que se cumplen para la multiplicación de matrices:

- Propiedad Asociativa: $A \cdot B \cdot C = (A \cdot B) \cdot C = A \cdot (B \cdot C)$
- Propiedad Distributiva: $C(A + B) = (C \cdot A) + (C \cdot B) = C \cdot A + C \cdot B$ y $(A + B)C = A \cdot C + B \cdot C$.

Un caso especial se presenta cuando consideramos una matriz diagonal. Sea A una matriz diagonal de orden n , entonces tenemos que $A \cdot A = A^2 = [a_{ij}^2]_{i=1,\dots,n; j=1,\dots,n}$. Por ejemplo, si

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 6 \end{bmatrix},$$

tendremos que

$$A \cdot A = A^2 = \begin{bmatrix} 16 & 0 \\ 0 & 36 \end{bmatrix}.$$

En general, si A es una matriz diagonal de orden n , tendremos que $A^\alpha = [a_{ij}^\alpha]_{i=1,\dots,n; j=1,\dots,n}$. Es decir, cuando se eleva una matriz diagonal a la α , el resultado es una matriz diagonal, cuyos elementos en la diagonal principal son iguales a los correspondientes elementos de la diagonal principal de la matriz original elevados cada uno a la α .

Ejemplo 14.3 Multiplicación de matrices

Dadas las siguientes dos matrices $A \cdot A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \\ 1 & 6 \end{bmatrix}$ y $B = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 1 \end{bmatrix}$, encuentre $A \cdot B$.

Respuesta: Inicialmente, es importante chequear que se cumpla la condición de conformidad. En este caso el número de columnas de la matriz A es igual al número de filas de la matriz B . Así, el producto $A \cdot B$ está definido: las matrices son conformables. Procedamos a encontrar dicho producto.

$$A \cdot B = \begin{bmatrix} 1 & 2 \\ 4 & 3 \\ 1 & 6 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 1 \end{bmatrix}.$$

Y por tanto,

$$A \cdot B = \begin{bmatrix} (1 \cdot 1) + (2 \cdot 2) & (1 \cdot 2) + (2 \cdot 4) & (1 \cdot 1) + (2 \cdot 1) \\ (4 \cdot 1) + (3 \cdot 2) & (4 \cdot 2) + (3 \cdot 4) & (4 \cdot 1) + (3 \cdot 1) \\ (1 \cdot 1) + (6 \cdot 2) & (1 \cdot 2) + (6 \cdot 4) & (1 \cdot 1) + (6 \cdot 1) \end{bmatrix}.$$

Esto implica que

$$A \cdot B = \begin{bmatrix} 5 & 10 & 3 \\ 10 & 20 & 7 \\ 13 & 26 & 7 \end{bmatrix}.$$

Finalmente, noten que $B \cdot$ también está definido. Encuentre a que es igual dicha multiplicación.

Es importante anotar que las matrices, así como los escalares, pueden ser sumadas, restadas o multiplicadas (siempre y cuando el producto esté definido). Pero la división para las matrices no es posible.

Si consideramos dos números (escalares) a y b , entonces el cociente $\frac{a}{b}$ (siempre y cuando $b \neq 0$) se puede expresar como $ab^{-1} = b^{-1}a$, donde b^{-1} se conoce como el recíproco o inverso de b . Debido a la propiedad conmutativa de la multiplicación de escalares, la expresión $\frac{a}{b}$ se puede emplear sin ningún problema para expresar ab^{-1} o $b^{-1}a$.

En el caso de las matrices esto es diferente. Debe ser claro que aunque los productos AB^{-1} y $B^{-1}A$ estén definidos³, estos dos productos usualmente son diferentes. De manera que, la expresión $\frac{A}{B}$ no puede emplearse porque es ambigua. Es decir, no es claro si esta expresión se refiere a AB^{-1} o $B^{-1}A$. Y peor aún, es posible que alguno de estos productos no exista. Así, cuando manipulamos matrices, es mejor evitar el uso de la expresión $\frac{A}{B}$ (matriz A dividida por la matriz B) (Chiang, 1996).

14.4 La matriz identidad y la matriz de ceros.

La matriz identidad es una matriz diagonal especial, cuyos elementos de la diagonal principal son iguales a uno. Es decir, la matriz identidad es una matriz con unos en la diagonal principal y ceros en las otras posiciones. La matriz identidad se denota por I_n , donde n corresponde al número de columnas y filas de la matriz (orden de la matriz). Formalmente,

$$I_n = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \quad (14.4)$$

La matriz identidad tiene la propiedad de ser el módulo de la multiplicación de matrices. Es decir, $I_k \cdot A_{k \times g} = A$ y $A_{k \times g} \cdot I_g = A$.

Un caso especial se produce cuando $A_{k \times g} \cdot I_g = I_g$, donde tiene que ser cierto que $A_{k \times g} = I_g$. ¿Por qué? (asegurese que puede encontrar el por qué de esta afirmación)

Otra matriz importante es la matriz de ceros, cuyos elementos son iguales a cero y se denota por $0_{n \times m}$. Esta matriz no necesariamente debe ser cuadrada y tiene la siguiente propiedad: $A_{n \times m} + 0_{n \times m} = 0_{n \times m} + A_{n \times m} = A$. Es decir, $0_{n \times m}$ es el módulo de la suma (siempre que la suma sea comnformable).

14.5 La transpuesta de una matriz y la matriz simétrica.

La transpuesta de una matriz cualquiera A es una matriz cuyas filas corresponden a las columnas de la matriz original; o lo que es lo mismo, es una matriz cuyas columnas corresponden a las filas de la matriz A . La transpuesta de una matriz A se denota por A^T o A' .

³ B^{-1} denota la inversa de la matriz B en caso de que exista. Este concepto será repasado más adelante.

Por ejemplo, la transpuesta de

$$C = \begin{bmatrix} 1 & 2 & 4 \\ 6 & 3 & 5 \\ 11 & 10 & 8 \end{bmatrix}$$

es

$$C^T = \begin{bmatrix} 1 & 6 & 11 \\ 2 & 3 & 10 \\ 4 & 5 & 8 \end{bmatrix}.$$

Las propiedades de la operación de transposición son las siguientes:

- Transpuesta de la transpuesta: $(A^T)^T = A$
- Transpuesta de una suma: $(A + B)^T = A^T + B^T$
- Transpuesta de un producto: $(A \cdot B)^T = B^T \cdot A^T$

Cuando $A^T = A$, se dice que A es una matriz simétrica. En otras palabras, una matriz simétrica es una matriz cuya transpuesta es igual a ella misma. Por ejemplo, la siguiente matriz es simétrica:

$$D = \begin{bmatrix} 1 & 2 & 4 \\ 6 & 3 & 5 \\ 4 & 5 & 8 \end{bmatrix}.$$

Dado que

$$D^T = \begin{bmatrix} 1 & 2 & 4 \\ 6 & 3 & 5 \\ 4 & 5 & 8 \end{bmatrix} = D.$$

Generalmente, por convención, las matrices simétricas son escritas omitiendo los elementos por debajo de la diagonal principal. Para nuestro ejemplo tendremos que la matriz D se puede reescribir de la siguiente forma

$$D = \begin{bmatrix} 1 & 2 & 4 \\ & 3 & 5 \\ & & 8 \end{bmatrix}.$$

Cuando se emplea esta notación se da por entendido que se trata de una matriz simétrica.

14.6 Matriz idempotente y matrices ortogonales

Una matriz cuadrada A se denomina idempotente si y solamente si $A \cdot A = A$. Un ejemplo de matriz idempotente es la matriz identidad. Por otro lado, dos matrices A y B son matrices ortogonales si y solamente si $A \cdot B = 0$.

14.7 Combinaciones lineales de vectores e independencia lineal

Dado un conjunto de vectores (fila o columna) v_1, v_2, \dots, v_k y un conjunto de escalares $\alpha_1, \alpha_2, \dots, \alpha_k$ para $k = 1, 2, \dots, K$, una combinación lineal de los vectores está dada por

$$c = \sum_{i=1}^K \alpha_i v_i. \quad (14.5)$$

Ejemplo 14.4 Combinaciones lineales de vectores

Dados los siguientes 3 vectores y 3 escalares, encuentre 4 diferentes combinaciones lineales de los vectores.

$$v = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \end{bmatrix}, w = \begin{bmatrix} 4 \\ 6 \\ 1 \\ 0 \end{bmatrix}, z = \begin{bmatrix} 5 \\ 10 \\ 0 \\ 1 \end{bmatrix}, \alpha = 2, \beta = 1, \gamma = -1.$$

Respuesta: A partir de estos vectores y escalares podemos encontrar las siguientes combinaciones lineales:

$$c_1 = \alpha v + \beta w + \gamma z = 2 \begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \end{bmatrix} + 1 \begin{bmatrix} 4 \\ 6 \\ 1 \\ 0 \end{bmatrix} + (-1) \begin{bmatrix} 5 \\ 10 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 11 \\ 11 \end{bmatrix}$$

$$c_2 = \beta v + \alpha w + \gamma z = 1 \begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \end{bmatrix} + 2 \begin{bmatrix} 4 \\ 6 \\ 1 \\ 0 \end{bmatrix} + (-1) \begin{bmatrix} 5 \\ 10 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \\ 7 \\ 5 \end{bmatrix}$$

$$c_3 = \beta v + \gamma w + \alpha z = 1 \begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \end{bmatrix} + (-1) \begin{bmatrix} 4 \\ 6 \\ 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 5 \\ 10 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 11 \\ 19 \\ -3 \\ -5 \end{bmatrix}$$

$$c_4 = \gamma v + \beta w + \alpha z = (-1) \begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \end{bmatrix} + 1 \begin{bmatrix} 4 \\ 6 \\ 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 5 \\ 10 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 12 \\ 23 \\ -4 \\ -4 \end{bmatrix}$$

Ahora tenemos todos los elementos para definir el concepto de dependencia lineal. Un conjunto de vectores será linealmente dependiente si al menos uno de los vectores puede expresarse como una combinación lineal de los otros. Así, por ejemplo, el conjunto de vectores c, v_1, v_2, \dots, v_k (definidos anteriormente), será dependiente linealmente, pues por definición el vector c es una combinación lineal de los demás vectores ($c = \sum_{i=1}^K \alpha_i v_i$).

Ahora bien, un conjunto de vectores v_1, v_2, \dots, v_k se considera linealmente independiente si y solamente si los únicos valores de los escalares $\alpha_1, \alpha_2, \dots, \alpha_k$ que cumplen la condición

$$\sum_{i=1}^K \alpha_i v_i = 0 \quad (14.6)$$

son $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$. En otras palabras, ningún vector del conjunto se puede expresar como combinación lineal de otro u otros vectores que pertenecen al mismo conjunto.

14.8 La Traza y el rango de una matriz

La traza de una matriz cuadrada A es la suma de los elementos de la diagonal principal y se denota por $tr(A)$. Es decir:

$$tr(A_{n \times n}) = \sum_{i=1}^n a_{ii}. \quad (14.7)$$

Una de las principales propiedades de la traza es: $tr(A \cdot B \cdot C) = tr(C \cdot B \cdot A) = tr(B \cdot C \cdot A) = \dots$

Por otro lado, el rango de una matriz A es el número de filas o columnas linealmente independientes y se denota por $ran(A)$. Si se trata de una matriz no simétrica de dimensiones $n \times m$, entonces tendremos que $ran(A) \leq \min(n, m)$ ⁴.

Si el rango de una matriz A es igual al número de sus columnas, entonces se dice que la matriz A tiene rango columna completo. En caso de que el rango de la matriz A sea igual al número de filas, se dice que la matriz A tiene rango fila completo. Para el caso de una matriz cuadrada, si el rango de la matriz es igual al número de filas y por tanto al número de columnas, se dice que la matriz tiene rango completo.

Ejemplo 14.5 Rango de una matriz

Encuentre el rango de la siguiente matriz

$$D = \begin{bmatrix} 2 & 4 & 5 & 12 \\ 3 & 6 & 10 & 23 \\ 5 & 1 & 0 & -4 \\ 6 & 0 & 1 & -4 \end{bmatrix}.$$

Respuesta: Noten que esta matriz se construye con los vectores columna v, w, z y c_4 del ejemplo anterior. Es más recuerden que $c_4 = \gamma v + \beta w + \alpha z$. Es decir, la última columna (vector columna) se puede construir como una combinación lineal de las primeras tres columnas. Además se puede comprobar rápidamente que las tres primeras filas son linealmente independientes entre sí. Así, $ran(D) = 3$.

⁴Siendo un poco más rigurosos se debería hablar del rango columna (número de columnas linealmente independientes) y del rango fila (número de filas linealmente independientes). Pero es fácil demostrar que el rango columna y el rango fila de una matriz es el mismo; por tanto, podemos hablar sin riesgo a ambigüedades del rango de una matriz.

Algunos resultados útiles del rango son:

- $\text{ran}(A \cdot B) = \min(\text{ran}(A), \text{ran}(B))$
- Si A tiene dimensiones $n \times m$ y B es una matriz cuadrada de rango m , entonces $\text{ran}(AB) = \text{ran}(A)$
- $\text{ran}(A) = \text{ran}(A^T A) = \text{ran}(AA^T)$

14.9 Determinante de una matriz

El determinante de una matriz cuadrada A , representado por $\det(A)$ o $|A|$, es un escalar asociado de manera única con esta matriz. Para una matriz 2×2 , el determinante está dado por

$$\det(A) = |A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12} \quad (14.8)$$

. Para una matriz de un orden superior, el cálculo del determinante es un poco más complejo.

Antes de entrar en el detalle del cálculo, definamos dos conceptos importantes. La matriz menor, asociada al elemento a_{ij} de una matriz cuadrada A de orden n , denotada por M_{ij} , es la matriz cuadrada de orden $n - 1$ obtenida al eliminar la i -ésima fila y la j -ésima columna de A . (Ver Figura 14.2). El menor del elemento a_{ij} es el determinante de la matriz menor M_{ij} ; es decir, $|M_{ij}|$. Y el cofactor del elemento a_{ij} , expresado por C_{ij} , corresponde a $C_{ij} = (-1)^{i+j} |M_{ij}|$.

Figura 14.2. Diagrama de la matriz menor asociada al elemento a_{ij} de la matriz A

$$M_{ij} = \begin{pmatrix} a_{11} & \cdots & \boxed{a_{1j}} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \cdots & \boxed{a_{ij}} & \cdots & a_{in} \\ \vdots & \ddots & \cdots & \ddots & \vdots \\ a_{n1} & \cdots & \boxed{a_{nj}} & \cdots & a_{nn} \end{pmatrix}$$

Nota: La matriz menor asociada al elemento a_{ij} de la matriz A , se forma eliminando la i -ésima fila y la j -ésima columna (columna y fila sombreadas en rojo).

Ahora bien, el determinante de una matriz de orden n puede ser calculado, a partir de cualquier columna o fila, de la siguiente manera:

- empleando la i -ésima fila: $\det(A) = |A| = \sum_{j=1}^n a_{ij}C_{ij}$
- empleando la j -ésima columna: $\det(A) = |A| = \sum_{i=1}^n a_{ij}C_{ij}$

Algunas propiedades útiles del determinante son:

- $\det(A) = \det(A^T)$.
- $|A^T| = |A|$.

- $|A \cdot B| = |A| \cdot |B|$.
- Si el producto de una fila de A por un escalar se suma a una fila de A , entonces el determinante de la matriz resultante es igual al $|A|$.
- El intercambio de dos filas o dos columnas, sin importar cuales sean, alterará el signo, pero no el valor numérico, del determinante.
- $\det(\lambda A_n) = \lambda^n \det(A)$, donde λ es un escalar.
- Si y solamente si A es cuadrada. Es decir, si A es cuadrada, entonces tiene que ser cierto que existe una columna (o fila) que es combinación lineal de una, o más de una columna (fila), de A .
- Si $\det(A) = 0$, entonces A se conoce como una matriz singular. En caso contrario (i.e. $\det(A) \neq 0$) se dice que A es una matriz no singular.
- Si A es una matriz diagonal denotada por $A = \text{diag}(a_{11}, a_{22}, \dots, a_{nn})$, entonces $\det(A) = a_{11} \cdot a_{22} \cdots \cdots a_{nn}$.

Ejemplo 14.6 Cálculo del determinante de una matriz

Encuentre el determinante de la siguiente matriz

$$D = \begin{bmatrix} 2 & -1 & 3 & 5 \\ 2 & 0 & 1 & 0 \\ 6 & 1 & 3 & 4 \\ -7 & 3 & -2 & 8 \end{bmatrix}.$$

Respuesta: El determinante puede ser calculado empleando cualquier fila o columna, en este caso será más fácil emplear la segunda fila, pues tiene dos elementos iguales a cero. Así, tendremos:

$$|D| = \left| \begin{bmatrix} 2 & -1 & 3 & 5 \\ 2 & 0 & 1 & 0 \\ 6 & 1 & 3 & 4 \\ -7 & 3 & -2 & 8 \end{bmatrix} \right| = 2 \cdot (-1)^{2+1} \left| \begin{bmatrix} -1 & 3 & 5 \\ 1 & 3 & 4 \\ 3 & -2 & 8 \end{bmatrix} \right| + 1 \cdot (-1)^{2+3} \left| \begin{bmatrix} 2 & -1 & 5 \\ 6 & 1 & 4 \\ -7 & 3 & 8 \end{bmatrix} \right|.$$

Note que esto implicará el cálculo de dos determinantes de orden 3. Esta operación se puede simplificar aún más si “creamos” más ceros en la segunda fila de la matriz D . Por ejemplo, si multiplicamos la columna 3 por -2 y sumamos este producto a la columna 1 tendremos que:

$$|D| = \left| \begin{bmatrix} 2 & -1 & 3 & 5 \\ 2 & 0 & 1 & 0 \\ 6 & 1 & 3 & 4 \\ -7 & 3 & -2 & 8 \end{bmatrix} \right| = \left| \begin{bmatrix} -4 & -1 & 3 & 5 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 3 & 4 \\ -3 & 3 & -2 & 8 \end{bmatrix} \right| = 1 \cdot (-1)^{2+3} \left| \begin{bmatrix} -4 & -1 & 5 \\ 0 & 1 & 4 \\ -3 & 3 & 8 \end{bmatrix} \right|.$$

Ahora, multiplicando la segunda columna por -4 y sumándola a la tercera columna se obtendrá:

$$\begin{aligned} |D| &= -1 \left| \begin{bmatrix} -4 & -1 & 5 \\ 0 & 1 & 4 \\ -3 & 3 & 8 \end{bmatrix} \right| = -1 \left| \begin{bmatrix} -4 & -1 & 9 \\ 0 & 1 & 0 \\ -3 & 3 & -4 \end{bmatrix} \right| = -1(-1)^{2+2} \left| \begin{bmatrix} -4 & 9 \\ -3 & -4 \end{bmatrix} \right| \\ &= -1(1)[(-4)(-4) - (-3)(9)] \\ &= -[16 + 27] = -43. \end{aligned}$$

14.10 Valores propios de una matriz

Los valores propios (*eigen values* en inglés), también conocidos como raíces características de una matriz cuadrada A_n son los escalares λ que satisfacen

$$|A - \lambda In| = 0. \quad (14.9)$$

Es decir, para encontrar los valores propios se debe resolver la anterior ecuación.

Los valores propios son importantes porque en muchos casos facilitan ciertos cálculos. En especial tenemos los siguientes resultados:

- $\det(A_n) = \prod_{i=1}^n \lambda_i$.
- El rango de cualquier matriz A es igual al número de valores propios diferentes de cero.

Ejemplo 14.7 Valores propios de una matriz

Encuentre los valores propios, rango y determinante de la siguiente matriz

$$D = \begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix}.$$

Respuesta: Para encontrar los valores propios de la matriz D , se requiere solucionar la siguiente ecuación $|D - \lambda In| = 0$. Es decir,

$$\left| \begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix} - \lambda I_2 \right| = 0.$$

$$\left| \begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0.$$

$$\left| \begin{bmatrix} 5-\lambda & 1 \\ 2 & 4-\lambda \end{bmatrix} \right| = 0.$$

$$(5-\lambda)(4-\lambda) - 2(1) = 0.$$

$$\lambda^2 - 9\lambda + 18 = 0.$$

Las dos soluciones son $\lambda_1 = 6$ ó $\lambda_2 = 3$. Así, los valores propios de la matriz D son 6 y 3. Por tanto, $\text{ran}(D) = 2$ y $\det(D) = 6 \cdot 3 = 18$.

14.11 La Matriz inversa

La matriz inversa de la matriz cuadrada A_n es una matriz cuadrada de igual orden tal que

$$B \cdot A_n = I_n. \quad (14.10)$$

Y se denota como A^{-1} , es decir $B = A^{-1}$. Además, si se post-multiplica la matriz A_n por su inversa, también se obtendrá la matriz identidad. En otras palabras, la matriz inversa además cumple que $A_n A_n^{-1} = I_n$.

Pero, ¿cómo encontrar la matriz inversa de una matriz A_n ? Primero es importante recordar que no todas las matrices cuadradas poseen inversa. De hecho, sólo aquellas matrices cuyas columnas y filas son linealmente independientes entre sí tendrán inversa. En otras palabras, una matriz singular ($\det(A) = 0$) no tendrá matriz inversa.

Existen diferentes métodos para encontrar la matriz inversa de una matriz no singular, pero aquí sólo repasaremos dos métodos. El primer método será emplear la transpuesta de la matriz de cofactores conocida como la matriz adjunta ($\text{Adj}(A)$). La matriz inversa de una matriz A_n está dada por la siguiente fórmula

$$A_n^{-1} = \frac{1}{|A_n|} \text{Adj}(A_n). \quad (14.11)$$

Figura 14.3. Ejemplo de una matriz adjunta de orden tres

$$\text{adj} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} & + \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \\ - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} & + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \\ + \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} & + \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{pmatrix}$$

Fuente: Elaboración propia

Ejemplo 14.8 Cálculo de la inversa de una matriz: Método de la matriz adjunta

Encuentre la inversa de la siguiente matriz

$$D = \begin{bmatrix} 3 & 2 \\ 1 & 0 \end{bmatrix}.$$

Respuesta: Para emplear la fórmula que se presenta en (14.11), primero debemos calcular tanto el determinante como la matriz adjunta de D . Así, tenemos que $\det(D) = -2$ y $\text{Adj}(D) = \begin{bmatrix} 0 & -2 \\ -1 & 3 \end{bmatrix}$.

Aplicando la fórmula, la inversa de D está dada por

$$D^{-1} = -\frac{1}{2} \begin{bmatrix} 0 & -2 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ \frac{1}{2} & -\frac{3}{2} \end{bmatrix}.$$

Ahora, constate que en efecto esta si es la inversa de D .

El segundo método que consideraremos es el método de reducción de Gauss-Jordan. En algunas ocasiones, este método resulta menos tedioso que aplicar la fórmula dada en (14.11). Dada una matriz cuadrada A_n , este método parte de considerar la siguiente matriz aumentada

$$\left[\begin{array}{ccc|cc} a_{11} & \dots & a_{1n} & 1 & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} & 0 & 1 \end{array} \right] = [A_n | I_n]. \quad (14.12)$$

Posteriormente, por medio de operaciones de filas podemos reducir la matriz A a la matriz identidad. Así tendremos que:

$$\left[\begin{array}{cc|ccc} 1 & 0 & d_{11} & \dots & d_{1n} \\ \ddots & \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & d_{n1} & \dots & d_{nn} \end{array} \right] = [I_n | A_n]. \quad (14.13)$$

Al final de las transformaciones, obtendremos la inversa de la matriz original a la derecha.

Ejemplo 14.9 Calculo de la inversa de una matriz: Método de reducción de Gauss-Jordan

Encuentre la inversa de la siguiente matriz empleando el método de reducción de Gauss-Jordan

$$E = \begin{bmatrix} 2 & 9 \\ 1 & 4 \end{bmatrix}.$$

Respuesta: Primero necesitamos armar la matriz aumentada. Es decir

$$\left[\begin{array}{cc|cc} 2 & 9 & 1 & 0 \\ 1 & 4 & 0 & 1 \end{array} \right].$$

Ahora necesitamos manipular esta matriz de tal forma que a la izquierda tengamos la matriz inversa. Para esto, intercambiemos la primera fila con la segunda.

$$\left[\begin{array}{cc|cc} 1 & 4 & 0 & 1 \\ 2 & 9 & 1 & 0 \end{array} \right].$$

Posteriormente, sumemos la fila uno multiplicada por (-2) a la fila 2.

$$\left[\begin{array}{cc|cc} 1 & 4 & 0 & 1 \\ 0 & 1 & 1 & -2 \end{array} \right].$$

Finalmente, sumemos la fila 2 multiplicada por (-4) a la fila 1.

$$\left[\begin{array}{cc|cc} 1 & 0 & -4 & 9 \\ 0 & 1 & 1 & -2 \end{array} \right].$$

Por tanto,

$$D^{-1} = \begin{bmatrix} -4 & 9 \\ 1 & -2 \end{bmatrix}.$$

Un resultado especial, que es muy útil, es la inversa de una matriz diagonal. Ésta es una de las inversas más fáciles de calcular. En general tenemos que:

$$\begin{bmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{nn}} \end{bmatrix} \quad (14.14)$$

Finalmente, repasemos rápidamente algunas propiedades de las matrices inversas.

- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^{-1} = A$
- $(A^T)^{-1} = (A^{-1})^T$
- Si A es una matriz simétrica, entonces A^{-1} también será simétrica

14.12 Elementos de cálculo matricial

Consideremos la siguiente función cuyo dominio es en los reales ($\mathbb{R}X$) y su rango pertenece a $\mathbb{R}^n : y = f(x_1, x_2, \dots, x_n) = f(\mathbf{x})$. El vector de derivadas parciales de $f(\mathbf{x})$ es conocido como el gradiente o vector gradiente y está definido de la siguiente manera:

$$g = g(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}. \quad (14.15)$$

Es importante tener en cuenta que el gradiente es un vector columna y no un vector fila.

El elemento i -ésimo del gradiente se interpreta como la pendiente de $f(\mathbf{x})$ con respecto al plano formado por x_i y y ; en otras palabras, es el cambio en $f(\mathbf{x})$ dado un cambio en x_i teniendo los otros elementos del vector \mathbf{x} constantes.

La segunda derivada de $f(\mathbf{x})$ está dada por una matriz denominada la matriz Hessiana que es calculada de la siguiente manera:

$$\begin{aligned} H &= \begin{bmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \frac{\partial^2 y}{\partial x_2 \partial x_1} & \frac{\partial^2 y}{\partial x_2 \partial x_2} & \dots & \frac{\partial^2 y}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \frac{\partial^2 y}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{bmatrix} \\ &= [f_{ij}] = \frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^T} \end{aligned} \quad (14.16)$$

La matriz Hessiana es cuadrada y simétrica, gracias al Teorema de Young.

Algunas derivadas especiales útiles para la construcción de modelos de regresión múltiple son:

- $\frac{\partial(\mathbf{x}^T \mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}$, donde \mathbf{a} es un vector columna
- $\frac{\partial(\mathbf{x}^T A \mathbf{x})}{\partial \mathbf{x}} = (A + A^T) \mathbf{x}$, donde A es cualquier matriz cuadrada
- $\frac{\partial(\mathbf{x}^T A \mathbf{x})}{\partial \mathbf{x}} = 2A\mathbf{x}$, donde A es una matriz cuadrada simétrica.

Ahora si consideramos funciones cuyo rango está en los \mathbb{R}^n , tenemos los siguientes resultados que son de gran utilidad:

- $\frac{\partial(A\mathbf{x})}{\partial \mathbf{x}} = A^T$, donde A es cualquier matriz tal que el producto $A\mathbf{x}$ está definido.
- $\frac{\partial(\mathbf{x}^T A \mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^T \mathbf{x}$, donde A es cualquier matriz cuadrada.
- $\frac{\ln(A)}{\partial A} = (A^{-1})^T$, donde A es cualquier matriz cuadrada no-singular.

14.13 Resultados especiales para el modelo de regresión múltiple

Antes de finalizar, es importante resaltar varios resultados importantes que serán empleados en este libro. Los datos de una muestra se organizarán en un vector columna $y_{n \times 1}$ que contendrá los datos de las n observaciones para la variable dependiente

$$y_{n \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (14.17)$$

Y una matriz $X_{n \times k}$ que contiene las variables explicativas y una columna de unos para el intercepto

$$X_{n \times k} = \begin{bmatrix} 1 & X_{21} & X_{31} & \dots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & X_{3n} & \dots & X_{kn} \end{bmatrix} \quad (14.18)$$

Noten que en este contexto hemos cambiado un poco la notación de las filas y las columnas. El primer número corresponde al número de la variable y el segundo a la observación i (fila). Es decir, X_{32} corresponde a la segunda observación de la variable X_2 .

Usando estas definiciones, tendremos los siguientes resultados:

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^n X_{2i} & \sum_{i=1}^n X_{3i} & \dots & \sum_{i=1}^n X_{ki} \\ \sum_{i=1}^n X_{2i}^2 & \sum_{i=1}^n X_{2i}X_{3i} & \dots & \sum_{i=1}^n X_{2i}X_{ki} \\ & \sum_{i=1}^n X_{3i}^2 & \ddots & \sum_{i=1}^n X_{3i}X_{ki} \\ & & \ddots & \vdots \\ & & & \sum_{i=1}^n X_{ki}^2 \end{bmatrix}_{k \times k} \quad (14.19)$$

Adicionalmente,

$$y^T y = \sum_{i=1}^n y_i^2. \quad (14.20)$$

Finalmente, tenemos que

$$X^T y = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i X_{2i}^2 \\ \sum_{i=1}^n y_i X_{3i}^2 \\ \vdots \\ \sum_{i=1}^n y_i X_{ki}^2 \end{bmatrix}_{k \times 1}. \quad (14.21)$$

Ejemplo 14.10 Matrices en la criptología

Como se mencionó al inicio de esta capítulo, una de las aplicaciones de las matrices es la criptología; es decir, el proceso de cifrar mensajes. A continuación veremos una breve aplicación de las matrices a la codificación de mensajes.

Considere una matriz fija A invertible (no singular). Entonces, podemos convertir el mensaje que se desea codificar en una matriz B , tal que $A \times B$ satisface la condición de conformidad. Así, se puede enviar el mensaje generado por el producto $A \times B$. El receptor del mensaje necesita conocer solamente A^{-1} para decodificar el mensaje. Esto gracias a que $A^{-1}AB = B$. Para entender cómo funciona esto, consideremos la siguiente matriz

$$A = \begin{bmatrix} -1 & 5 & -1 \\ -2 & 11 & 7 \\ 1 & -5 & 2 \end{bmatrix}.$$

La correspondiente inversa es

$$A^{-1} = \begin{bmatrix} 57 & -5 & 46 \\ 11 & -1 & 9 \\ -1 & 0 & -1 \end{bmatrix}.$$

Ahora supongamos que queremos transmitir el mensaje “Estamos en Cali”. Para esto necesitamos crear una equivalencia entre las letras y un número. Por ejemplo supongamos que se crea la siguiente equivalencia: $e = 3, s = 4, t = 5, a = 1, m = 6, o = 8, n = 7, c = -1, l = 9, i = 0$ y para los espacios en blanco usaremos el cero.

Esto se puede expresar de la siguiente manera (note que el receptor del mensaje también debe conocer la codificación adecuada de las letras). Así, podemos crear la siguiente matriz

$$B = \begin{bmatrix} 3 & 4 & 5 & 1 & 6 \\ 8 & 4 & 0 & 3 & 7 \\ 0 & -1 & 1 & 9 & -2 \end{bmatrix}.$$

Entonces tenemos que

$$AB = \begin{bmatrix} 37 & 17 & -6 & 5 & 31 \\ 82 & 29 & -3 & 94 & 51 \\ -37 & -18 & 7 & 4 & -33 \end{bmatrix}.$$

Por tanto, el mensaje encriptado a enviar sería: 37,17,37, 17 , -6 , 5, 31, 82, El receptor puede saber cuántas filas tendrá la matriz que le es enviada al conocer A . Y podrá fácilmente reconstruir la matriz AB . De tal forma que pre-multiplicando por A^{-1} el mensaje recibido se obtendrá el mensaje deseado.

Ejercicio: Encripte el mensaje “Tengo que repasar álgebra matricial”

14.14 Empleando R para hacer operaciones matriciales

R Core Team (2018) es un lenguaje de programación que emplea vectorización. Es decir, la forma natural de hacer cálculos en R es emplear vectores y matrices y no escalares. Esto hace el lenguaje matricial sea un lenguaje natural al momento de emplear R.

Empecemos por crear la siguiente matriz

$$D = \begin{bmatrix} 2 & -1 & 3 & 5 \\ 2 & 0 & 1 & 0 \\ 6 & 1 & 3 & 4 \\ -7 & 3 & -2 & 8 \end{bmatrix}.$$

Es se puede hacer con la función **matrix()** del paquete base de R. Esta función requiere los siguientes argumentos:

```
matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE)
```

donde:

- **data**: los datos que estarán en la matriz.
- **nrow**: número de filas.
- **ncol**: número de columnas.
- **byrow** : Valor lógico. Si **byrow** = FALSE (el valor por defecto) la matriz se llena por columnas, de lo contrario la matriz se llena por filas.

La matriz D se puede construir rápidamente empleando el código que se observa a continuación.

```
D <- matrix(c(2, -1, 3, 5, 2, 0, 1, 0, 6, 1, 3, 4, -7, 3, -2,
             8), nrow = 4, ncol = 4, byrow = TRUE)
D

##      [,1] [,2] [,3] [,4]
## [1,]    2   -1    3    5
## [2,]    2    0    1    0
## [3,]    6    1    3    4
## [4,]   -7    3   -2    8
```

```
class(D)

## [1] "matrix" "array"
```

Por otro lado, una matriz identidad puede ser construida rápidamente con la función **diag()**. Para crear una matriz identidad de orden n el único atributo que requiere la función es el tamaño de la matriz identidad (n). Es decir,

```
diag(4)

##     [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1
```

La misma función **diag()** permite extraer los valores de la diagonal principal de una matriz, si el argumento de esta es una función.

```
diag(D)

## [1] 2 0 3 8
```

Así, la traza de la matriz D se puede calcular de la siguiente manera:

```
sum(diag(D))

## [1] 13
```

Por otro lado, si queremos sumar matrices, esto se puede hacer rápidamente empleando el operador “+”. Por ejemplo, supongamos que queremos sumarle a la matriz D la siguiente matriz

$$E = \begin{bmatrix} 3 & 3 & 7 & 51 \\ 4 & 9 & 1 & 3 \\ 5 & 10 & 3 & 12 \\ 8 & 21 & -2 & 4 \end{bmatrix}.$$

```
E <- matrix(c(3, 3, 7, 51, 4, 9, 1, 3, 5, 10, 3, 12, 8, 21, -2,
             4), nrow = 4, ncol = 4, byrow = TRUE)
E
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    3    3    7   51
## [2,]    4    9    1    3
## [3,]    5   10    3   12
## [4,]    8   21   -2    4
```

D + E

```
##      [,1] [,2] [,3] [,4]
## [1,]    5    2   10   56
## [2,]    6    9    2    3
## [3,]   11   11    6   16
## [4,]    1   24   -4   12
```

Para la multiplicación de matrices se debe tener un poco de cuidado. El operador “`**`” realiza una multiplicación elemento por elemento, no realiza la multiplicación de las matrices. Por otro lado el operador “`%*%`” si realiza la multiplicación de matrices. Es decir,

```
# multiplicación elemento por elemento
D * E
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    6   -3   21  255
## [2,]    8    0    1    0
## [3,]   30   10    9   48
## [4,]  -56   63    4   32
```

```
# multiplicación de matrices
D %*% E
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   57  132   12  155
## [2,]   11   16   17  114
## [3,]   69  141   44  361
## [4,]   45  154  -68 -340
```

El determinante de una matriz, y sus valores propios se pueden calcular empleando las funciones `det()` y `eigen()`, respectivamente. Por ejemplo

```
# determinante
det(D)

## [1] -43
```

```
# valores propios
eigen(D)

## eigen() decomposition
## $values
## [1] 6.851893+5.934594i 6.851893-5.934594i -1.156348+0.000000i
## [4] 0.452562+0.000000i
##
## $vectors
## [,1] [,2] [,3]
## [1,] 0.0919841-0.4433032i 0.0919841+0.4433032i -0.47221621+0i
## [2,] -0.0983345-0.1285593i -0.0983345+0.1285593i 0.07723466+0i
## [3,] -0.0947980-0.5778436i -0.0947980+0.5778436i 0.85512230+0i
## [4,] 0.6526363+0.0000000i 0.6526363+0.0000000i -0.19953074+0i
## [,4]
## [1,] -0.2400156+0i
## [2,] 0.5743704+0i
## [3,] 0.7399695+0i
## [4,] -0.2548258+0i

eigen(D)$values

## [1] 6.851893+5.934594i 6.851893-5.934594i -1.156348+0.000000i
## [4] 0.452562+0.000000i
```

Finalmente, la matriz inversa se puede encontrar con la función **solve()** .

```
# dinversa
solve(D)

## [,1] [,2] [,3] [,4]
## [1,] -0.09302326 -1.744186 0.5348837 -0.20930233
## [2,] -0.27906977 2.767442 -0.3953488 0.37209302
## [3,] 0.18604651 4.488372 -1.0697674 0.41860465
## [4,] 0.06976744 -1.441860 0.3488372 -0.09302326

# chequeo
round(solve(D) %*% D, 3)

## [,1] [,2] [,3] [,4]
## [1,] 1 0 0 0
## [2,] 0 1 0 0
## [3,] 0 0 1 0
## [4,] 0 0 0 1
```




15 . Elementos de Estadística

Diseñado por Freepik

Objetivos del capítulo

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

- Explicar en sus propias palabras la diferencia entre una variable aleatoria y una no aleatoria.
- Explicar en sus propias palabras los conceptos de: valor esperado, varianza y covarianza

15.1 Introducción

Este libro supone el conocimiento básico de estadística que le permiten al científico de datos trabajar con grandes volúmenes de datos y modelos estadísticos. Este Apéndice presenta unos conceptos básicos de estadística necesarios para seguir algunas demostraciones y operaciones descritas en el libro. Este Apéndice no pretende ser un tratado autocontenido de estadística, sino por el contrario un breve resumen que permitirá al lector ya familiarizado con los conceptos recordar los concepto.

En general, la Estadística es definida como “la ciencia de estimar la distribución de probabilidad de una variable aleatoria basada en repetidas observaciones de variables aleatorias de la misma variable aleatoria”^{footnote}Este apéndice corresponde a una versión adaptada de Alonso (2007). (Amemiya, 1994).

Así, la estadística es una ciencia que emplea conjuntos de datos para obtener a partir de ellos inferencias (proyección, adivinanza) sobre una población (valor real). De manera que el problema estadístico consiste en encontrar la mejor predicción para un valor real desconocido para el investigador, a partir de datos recolectados (muestra) de una población.

En este capítulo repasaremos los conceptos básicos de estadística y probabilidad que son las bases para este libro.

15.2 Variables, vectores y matrices aleatorias

Una variable se define como una magnitud que puede tener un valor cualquiera de los comprendidos en un conjunto. En otras palabras, es una “letra” que puede tomar uno o diferentes valores. Por ejemplo, si la variable x cumple la condición de que $3x = 2$, entonces la variable necesariamente tomará el valor de $\frac{2}{3}$ ($x = \frac{2}{3}$). Otro ejemplo, si la variable cumple la condición $x^2 = 1$, entonces x puede tomar los valores de 1 o -1.

Ahora, consideremos la definición de una variable aleatoria, también conocida como variable estocástica. Una variable aleatoria es una “letra” que toma diferentes valores, cada uno con una probabilidad previamente definida. Por ejemplo, tiremos una moneda justa¹ al aire, y sea X la variable aleatoria que toma el valor de uno si la cara superior de la moneda es sello, en caso contrario la variable toma el valor de cero. Es decir

$$X = \begin{cases} 1 & \text{si sello} \\ 0 & \text{si cara} \end{cases} .$$

Entonces, en este caso, diremos que la variable aleatoria X tiene dos posibles realizaciones. Ahora bien, si la moneda es una moneda normal, existirá igual probabilidad que la variable aleatoria tome el valor de uno o cero. En otras palabras, tendremos que la probabilidad de que la variable aleatoria sea igual a uno es 0.5, al igual que la probabilidad que la variable aleatoria sea cero. Esto se puede abreviar de la siguiente forma: $P(X = 1) = 0.5$ y $P(X = 0) = 0.5$.

¹Por una moneda justa, se entiende una moneda que tiene una probabilidad igual de obtener cualquiera de las dos caras.

Si el conjunto de valores que toma la variable aleatoria es un conjunto finito o infinito contable, entonces la variable estocástica se denomina una variable aleatoria discreta. Por otro lado, si las posibles realizaciones de la variable aleatoria son un conjunto de realizaciones infinitamente divisible y, por tanto, imposible de contar, entonces la variable estocástica se conoce como una variable aleatoria continua. En general, si las posibles realizaciones toman valores discretos entonces estamos hablando de una variable estocástica discreta; por el contrario, si los posibles valores son parte de un rango continuo de valores, entonces estamos hablando de una variable estocástica continua.

Un vector aleatorio es un vector cuyos elementos son variables aleatorias ya sean continuas o discretas, es decir,

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad (15.1)$$

donde X_i para $i = 1, 2, \dots, n$ representan diferentes variables aleatorias. Análogamente, una matriz aleatoria es una matriz cuyos elementos son variables aleatorias.

Es importante anotar que los científicos de datos interpretan la mayoría de los aspectos de la realidad como resultados de un proceso estocástico. En la práctica observamos un único valor de una variable como las ventas mensuales o los rendimientos de un activo. Los valores observados en la realidad para esas variables aleatorias (muestra), se interpretan como las realizaciones de una variable aleatoria después de que los “dados” de la economía o el negocio ya han sido tirados. Es decir, lo que observa el científico de datos es la realización de un evento aleatorio.

15.3 Distribución de probabilidad

Una distribución de probabilidad de una variable aleatoria discreta, también conocida como la función de densidad discreta, $f(x)$, es una lista de las probabilidades asociadas a las diferentes realizaciones x que puede tomar una variable aleatoria discreta X . Para una variable aleatoria discreta tenemos que

$$f(x) = P(X = x) \quad (15.2)$$

donde debe cumplir que:

- $0 \leq f(x) \leq 1$
- $\sum_{\forall i} f(x_i) = 1$

Dado que en el caso de una variable aleatoria continua, ésta puede tomar cualquier valor dentro de un número infinito de valores, será imposible asignar una probabilidad para cada uno de los valores que puede tomar la variable aleatoria continua. Por tanto, en el caso de variables aleatorias continuas es necesario un enfoque diferente al seguido con las variables aleatorias discretas. En este caso definiremos una función que nos permita conocer la probabilidad de ocurrencia de un intervalo (conjunto continuo de puntos) y no un punto como lo hicimos para las variables aleatorias discretas.

Una distribución de probabilidad de una variable aleatoria continua, también conocida como la función de densidad continua, $f(x)$, es una función asociada a la variable aleatoria continua X , tal que

$$\int_a^b f(x)dx = P(a \leq x \leq b) \quad (15.3)$$

donde debe cumplir que:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

15.4 Valor esperado de una variable aleatoria

Eventualmente, las distribuciones de probabilidad se pueden describir con sus momentos². El primer momento de una distribución se conoce como el valor esperado o esperanza matemática.

El valor esperado de una variable aleatoria corresponde a su media poblacional y se interpreta como el valor promedio que se espera de la variable aleatoria cuando se obtiene cualquier muestra de ésta.

El valor esperado de una variable aleatoria discreta denotado por $E[X]$ se define como

$$E[X] = \sum_{\forall i} x_i P(X = x_i) = \sum_{\forall i} x_i f(x_i) \quad (15.4)$$

El valor esperado de una variable aleatoria continua, también denotado por $E[X]$ se define como:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \quad (15.5)$$

El valor esperado también es conocido como el operador de esperanza matemática, es un operador lineal cuyas principales características son:

- $E[c] = c$, donde c es una constante, o una variable no estocástica.
- $E[aX + b] = aE[X] + b$, donde a y b son constantes y X es una variable aleatoria.
- En general $E[g(X)] \neq g(E(X))$, donde $g(\cdot)$ es cualquier función. La única excepción de esto es cuando $g(\cdot)$ es una función lineal.
- $E[a_1X_1 + a_2X_2 + \dots + a_nX_n] = a_1E[X_1] + a_2E[X_2] + \dots + a_nE[X_n]$ donde cada uno de los a_i y X_i ($i=1, 2, \dots, n$) son constantes y variables aleatorias, respectivamente.
- $$X = \begin{cases} \sum_{\forall i} g(x)f(x_i) & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{si } X \text{ es cont} \end{cases}$$
.

Como se mencionó anteriormente, $E[x]$ se conoce como el primer momento de una variable aleatoria y también se denota como μ_x . Es decir, la media poblacional de X . El i -ésimo momento (alrededor del origen) de una variable aleatoria X está definido por $\mu'_i = E[X^i]$.

²Los momentos de una distribución son representados por parámetros poblacionales que se representarán de aquí en adelante con letras griegas. Los momentos de una distribución describen las características de la distribución poblacional de la variable aleatoria.

15.5 Independencia (estadística) lineal

Dos variables aleatorias, X y Y , se consideran estadísticamente independientes, u ortogonales, si y solamente si

$$E[XY] = E[X]E[Y]. \quad (15.6)$$

Es importante notar que independencia estadística entre dos variables no implica que no exista relación alguna entre las variables, como se verá más adelante, independencia estadística sólo implica que no existe una relación lineal entre las dos variables.

15.6 Varianza y momentos alrededor de la media de una variable aleatoria

La varianza de una variable aleatoria, denotada por σ^2 o $Var[X]$, se define como

$$Var[X] = E[(x - \mu)^2] \quad (15.7)$$

Así, en el caso de una variable aleatoria discreta tendremos que

$$Var[X] = \sum_{\forall i} (x_i - \mu)^2 f(x_i),$$

y para una variable estocástica continua la varianza será calculada de la siguiente manera:

$$Var[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

La varianza es una medida de la dispersión de una distribución. Generalmente se emplea la raíz cuadrada de la varianza, la desviación estándar ($\sigma = \sqrt{Var[X]}$), para describir una distribución. La ventaja de la desviación estándar es que ésta está medida en las mismas unidades de X y μ .

Un ejemplo de cómo la desviación estándar puede ser empleada para describir la dispersión de una distribución está dado por la desigualdad de Chebychev; para cualquier variable aleatoria y para cualquier constante se tiene que:

$$P(\mu - k\sigma \leq x \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2} \quad (15.8)$$

Antes de continuar, es importante anotar que el cálculo directo de la varianza es relativamente engorroso, afortunadamente es fácil mostrar que

$$Var[X] = E[X^2] - (E[X])^2 \quad (15.9)$$

este resultado permite en la práctica agilizar el cálculo de la varianza de cualquier variable aleatoria.

Las principales propiedades de la varianza son:

- $Var[c] = 0$, donde c es una constante, o una variable no estocástica.

- $\text{Var}[aX + b] = a^2\text{Var}[X]$, donde a y b son constantes y X es una variable aleatoria.
- $\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y]$ donde a y b son constantes y X y Y son variables aleatorias, respectivamente.(en la próxima sección repasaremos el concepto de covarianza ($\text{Cov}[X, Y]$)).

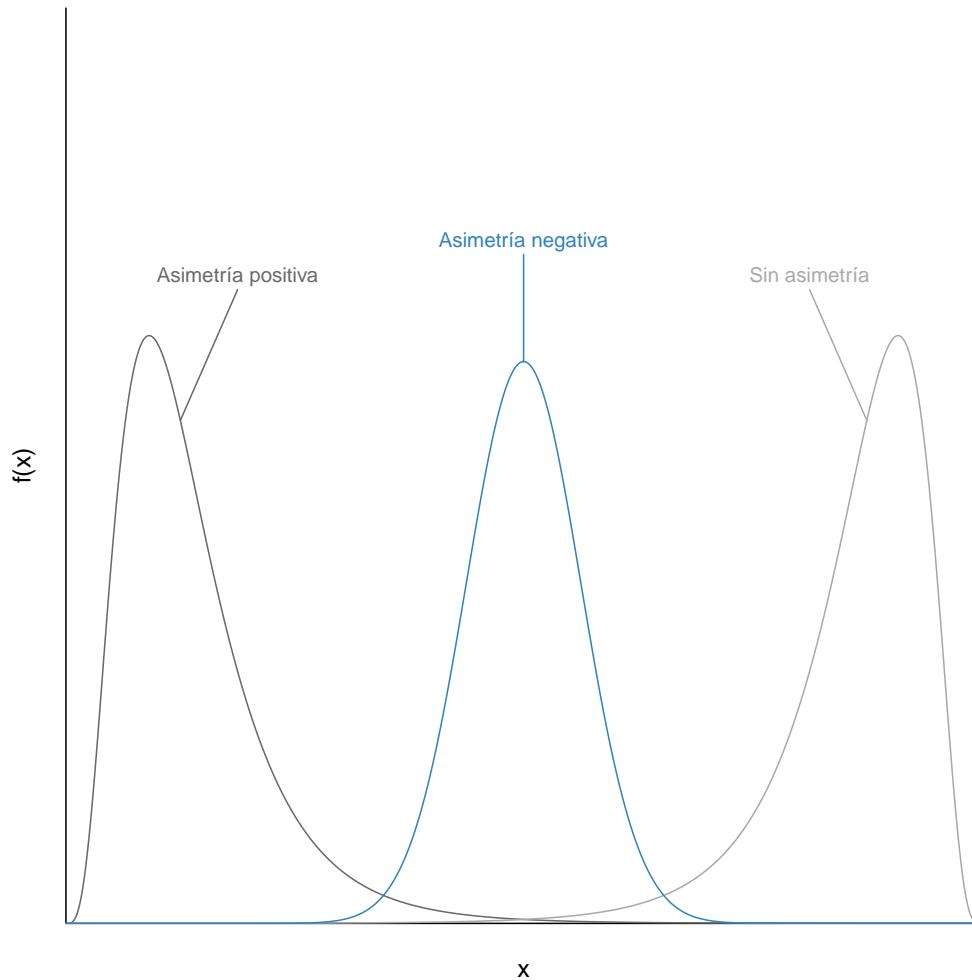
La varianza de una variable aleatoria también es conocida como el segundo momento alrededor de la media. En general, el i -ésimo momento alrededor de la media de una variable aleatoria se define como:

$$\mu_i = E[(x - \mu)^i] \quad (15.10)$$

El tercero y cuarto momento alrededor de la media se conocen como la asimetría (*skewness* en inglés) y curtosis, respectivamente. Una medida de asimetría comúnmente empleada es el coeficiente de asimetría definido como:

$$A = \frac{\mu_3}{\sigma^3} \quad (15.11)$$

En la Figura 15.1 se presentan los posibles casos extremos para interpretar el coeficiente de asimetría. En general, cuando ambas colas de la distribución tienen igual longitud, diremos que la distribución es simétrica o no posee asimetría. Por otro lado, si la cola izquierda (derecha) es más “corta” (larga) que la derecha, entonces la distribución se dirá que la distribución tiene asimetría positiva (negativa) (Ver Figura 15.1). En algunos casos la asimetría positiva también se conoce como asimetría a la derecha, mientras que la asimetría negativa se denomina asimetría a la izquierda.

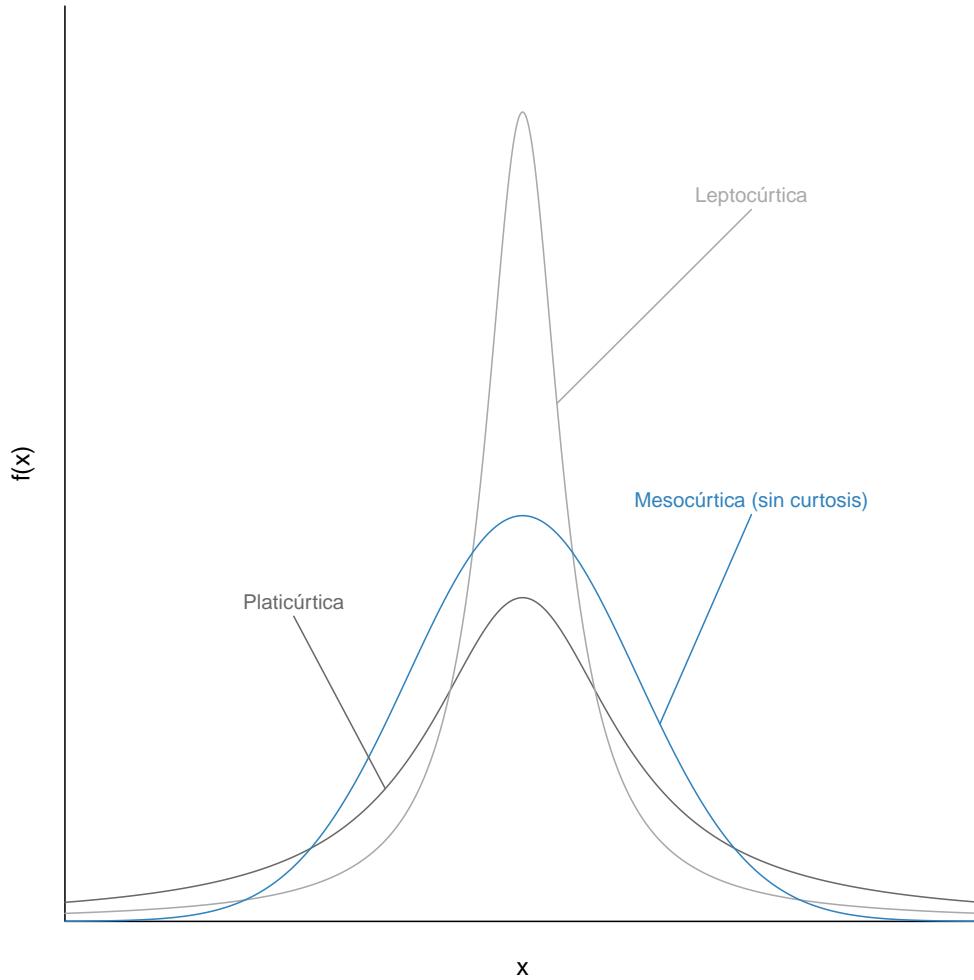
Figura 15.1. Tipos de Asimetría de diferentes distribuciones

Fuente: Elaboración propia

Otro estadístico comúnmente empleado para describir que tan “aplanada” o “picuda” es una distribución, es el coeficiente de curtosis que se define como:

$$C = \frac{\mu_4}{\sigma^4} \quad (15.12)$$

Para efectos de comparación, se emplea como distribución referente la distribución normal. Una distribución que es relativamente más picuda que una distribución normal se le denomina leptocúrtica. Contrariamente, aquella distribución más plana que la distribución normal se le denomina platicúrtica (Ver Figura 15.2).

Figura 15.2. Tipos de Curtosis de diferentes distribuciones

Fuente: Elaboración propia

15.7 Covarianza y Correlación entre dos variables aleatorias

Ahora consideremos la covarianza entre dos variables aleatorias X y Y denotada por $Cov[X, Y]$ ó $\sigma_{x,y}$ y definida como

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])] \quad (15.13)$$

Al igual que lo que ocurre con la varianza de una variable aleatoria, el cálculo directo de una covarianza es muy engorroso. Afortunadamente, es fácil mostrar que

$$Cov[X, Y] = E[XY] - E[X]E[Y] \quad (15.14)$$

La expresión 15.14 ayuda a entender la utilidad de la covarianza entre dos variables aleatorias. Noten

que en caso de que las variables estocásticas X y Y sean independientes, se tendrá por definición que $E[XY] = E[X]E[Y]$. Y por tanto, $Cov[X, Y] = 0$.

De esta manera, la covarianza entre dos variables aleatorias será cero si no existe relación lineal (hay independencia) entre ellas; y será diferente de cero si no hay independencia estadística entre ellas. Por otro lado, en el caso de que al mismo tiempo que una realización de la variable aleatoria X está por encima de su media, la realización de la variable estocástica Y también está por encima de su media, entonces la covarianza de estas dos variables será positiva. Si cuando la realización de una variable aleatoria está por encima de su media la realización de la otra variable está por debajo de la media, entonces la covarianza será negativa.

Una importante propiedad de la covarianza es:

$$Cov[a + bX, c + dY] = bdCov[X, Y],$$

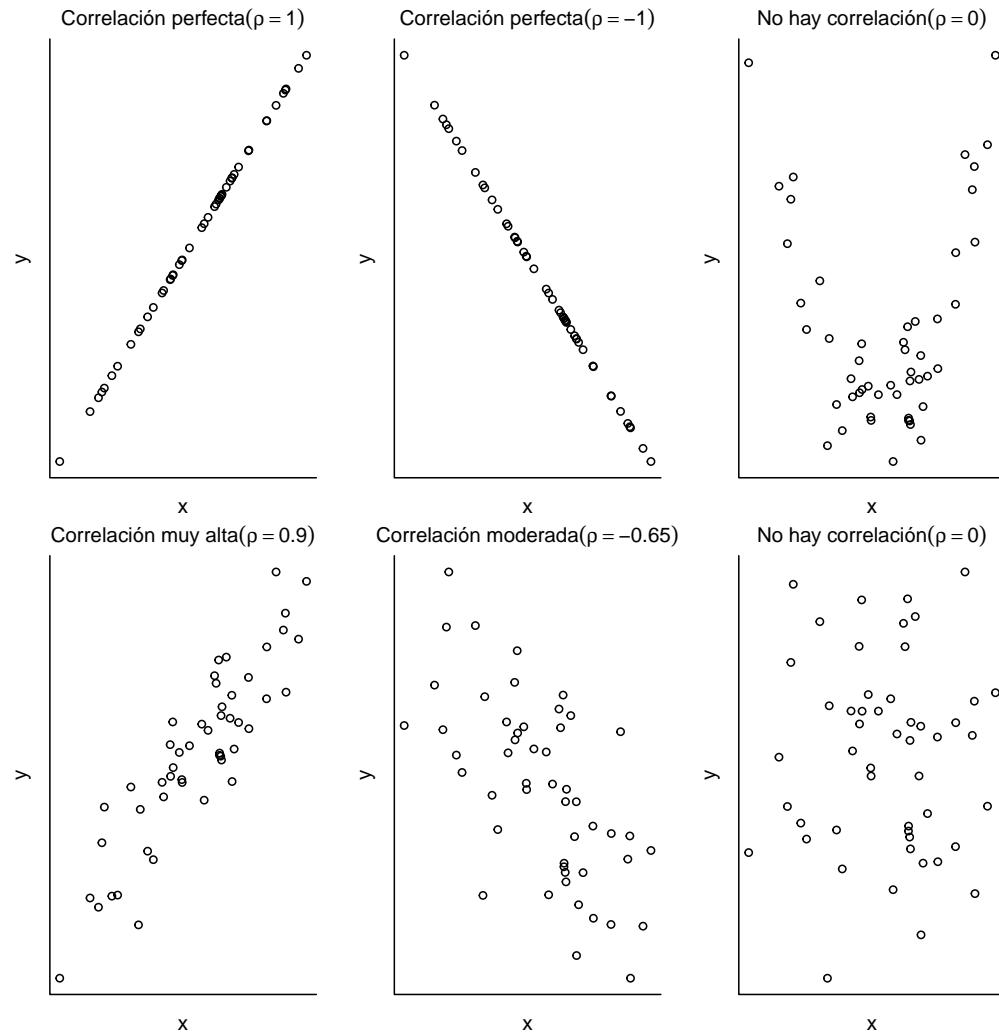
donde a, b, c y d son constantes, y X y Y son variables aleatorias.

Como se mencionó anteriormente, la covarianza entre dos variables estocásticas mide la relación lineal entre las variables, pero ésta depende de las unidades en que están medidas X y Y . Para tener una medida del grado de dependencia lineal entre dos variables aleatorias, que no dependa de las unidades, se emplea el coeficiente de correlación.

La correlación entre dos variables aleatorias, denotado por ρ , está definida por:

$$\rho = \frac{Cov[X, Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}} \quad (15.15)$$

Es muy fácil mostrar que $-1 \leq \rho \leq 1$. La correlación entre dos variables aleatorias tiene una interpretación muy sencilla; por ejemplo, una correlación de 1/1 entre las variables aleatorias X y Y implica una relación lineal positiva/negativa y perfecta entre ellas. Mientras que una correlación de cero implica que no existe relación lineal entre las variables. En la Figura 15.3

Figura 15.3. Ejemplos de diferentes valores de la correlación

Fuente: Elaboración propia

15.8 Esperanza y Varianza de vectores aleatorios.

Como se mencionó anteriormente, un vector aleatorio es un vector cuyos elementos son todas variables aleatorias. Así, el valor esperado de un vector aleatorio corresponde a un vector cuyos elementos son los valores esperados de los correspondientes elementos del vector estocástico. En otras palabras, sea \mathbf{X} un vector aleatorio, entonces:

$$E[\mathbf{X}] = E \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \mu_{n \times 1}. \quad (15.16)$$

Es muy fácil extender esta idea para encontrar el valor esperado de una matriz aleatoria. Sea $\mathbf{X}_{n \times m}$ una matriz aleatoria de dimensiones $n \times m$, entonces

$$E[\mathbf{X}_{n \times m}] = \begin{bmatrix} E[X_{11}] & E[X_{12}] & \dots & E[X_{1m}] \\ E[X_{21}] & E[X_{22}] & \dots & E[X_{2m}] \\ \vdots & \vdots & \vdots & \vdots \\ E[X_{n1}] & E[X_{n2}] & \dots & E[X_{nm}] \end{bmatrix} \quad (15.17)$$

Análogamente al caso de una variable aleatoria, la varianza de un vector aleatorio $\mathbf{X}_{n \times 1}$ se define como:

$$Var[\mathbf{X}_{n \times 1}] = E[(\mathbf{X}_{n \times 1} - \mu)(\mathbf{X}_{n \times 1} - \mu)^T] = E[\mathbf{X}_{n \times 1} \mathbf{X}_{n \times 1}^T] - \mu \mu^T. \quad (15.18)$$

En este caso tenemos que

$$Var[\mathbf{X}_{n \times 1}] = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \dots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \dots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \vdots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \dots & E[(X_n - \mu_n)(X_n - \mu_n)]. \end{bmatrix} \quad (15.19)$$

La matriz de varianzas de un vector aleatorio $\mathbf{X}_{n \times 1}$, conocida como la matriz de covarianzas o la matriz de varianzas y covarianzas, está dada por:

$$Var[\mathbf{X}_{n \times 1}] = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] & \dots & Cov[X_1, X_n] \\ Cov[X_2, X_1] & Var[X_2] & \dots & Cov[X_2, X_n] \\ \vdots & \vdots & \vdots & \vdots \\ Cov[X_n, X_1] & Cov[X_n, X_2] & \dots & Var[X_n]. \end{bmatrix} \quad (15.20)$$

En algunas ocasiones esta matriz se escribe de la siguiente manera:

$$Var[\mathbf{X}_{n \times 1}] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix} = \Sigma. \quad (15.21)$$

Nota que esta matriz es simétrica pues en general $\sigma_{ij} = \sigma_{ji}$.

Dividiendo cada uno de los σ_{ij} por las respectivas σ_i y σ_j obtendremos la matriz de correlaciones:

$$\begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{bmatrix}. \quad (15.22)$$

Antes de continuar, consideraremos las siguientes propiedades. Sean \mathbf{a} un vector de constantes, \mathbf{A} una matriz de constantes y $\mathbf{X}_{n \times 1}$ un vector aleatorio, entonces:

- $E [\mathbf{a}^T \mathbf{X}] = \mathbf{a}^T \boldsymbol{\mu}$
- $Var [\mathbf{a}^T \mathbf{X}] = \mathbf{a}^T Var [\mathbf{X}] \mathbf{a} = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}$
- $E [\mathbf{A} \mathbf{X}] = \mathbf{A} \boldsymbol{\mu}$
- $Var [\mathbf{A} \mathbf{X}] = \mathbf{A} \boldsymbol{\Sigma} \mathbf{a} \mathbf{A}^T$
- $E [tr(\mathbf{X}_{n \times n})] = tr(E[\mathbf{X}_{n \times n}])$

15.9 Estimadores puntuales y sus propiedades deseadas

Intuitivamente, un estimador se puede entender como una “fórmula” que permite pronosticar un valor poblacional (parámetro) desconocido a partir de una muestra. Por ejemplo, supongamos que deseamos conocer la media de una población. Regularmente no conocemos este valor y por tanto se recolectan observaciones de parte de la población total (muestra), y a partir de estas observaciones evaluamos una fórmula para conocer nuestro pronóstico del valor poblacional real.

Formalmente un estimador, también conocido como estimador puntual, de un parámetro poblacional es una función que indica cómo calcular una matriz, vector o escalar a partir de una muestra. El valor arrojado por esta función una vez los valores muestrales son reemplazados en el estimador se denomina estimación.

Así, un estimador $\hat{\theta}$ para pronosticar un parámetro θ a partir de una muestra aleatoria de tamaño n se define como:

$$\hat{\theta} = h(X_1, X_2, \dots, X_n) \quad (15.23)$$

donde $h(\cdot)$ es una función cualquiera y X_1, X_2, \dots, X_n corresponden a cada uno de los puntos muestrales (elementos de la muestra). Los estimadores son variables aleatorias, pues son función de variables aleatorias.

Claramente cualquier función de los puntos muestrales por definición es un estimador. Pero, ¿cómo escoger cuál función de la muestra es un buen estimador para el parámetro deseado? Existen varias propiedades deseadas en los estimadores que discutiremos a continuación.

Una propiedad muy deseable es que el valor esperado de la distribución del estimador esté lo más cercano o coincida con el valor población del parámetro. De esta forma, cada vez que se analice información nueva se estará seguro que en promedio el estimador estará correcto. En general, diremos que un estimador es insesgado si $E[\hat{\theta}] = \theta$. Así definiremos el sesgo de un estimador como:

$$Sesgo[\hat{\theta}] = E[\hat{\theta}] - \theta. \quad (15.24)$$

La insesgadez es una propiedad deseable en un estimador, pero la ausencia de sesgo no dice nada sobre la dispersión que tiene el estimador alrededor de su media. En general, se preferirá un estimador que tenga una menor dispersión alrededor de la media (varianza) a uno con mayor dispersión. Un estimador $\hat{\theta}_1$ es considerado un estimador insesgado más eficiente que si

$$Var[\hat{\theta}_1] < Var[\hat{\theta}_2] \quad (15.25)$$

Ahora consideremos el caso en que estamos comparando un estimador sesgado con una varianza relativamente pequeña con un estimador insesgado con una varianza relativamente grande. La pregunta es: ¿cuál de los dos estimadores deberá ser preferido? Un criterio para escoger un estimador entre otros, es considerar el estimador con el Mínimo Error Medio al Cuadrado, denotado MSE por su nombre en inglés (*Mean Square Error*), éste se define como:

$$MSE [\hat{\theta}] = (E [\hat{\theta} - \theta])^2 \quad (15.26)$$

Es fácil mostrar que:

$$MSE [\hat{\theta}] = (\text{Sesgo} [\hat{\theta}])^2 + \text{Var} [\hat{\theta}] \quad (15.27)$$

Así al minimizar el MSE, se está teniendo en cuenta tanto el sesgo como la dispersión del estimador.

Finalmente, otra propiedad deseada en un estimador es la consistencia. Intuitivamente, un estimador es consistente si cuando la muestra se hace grande y más cercana a la población total, entonces la probabilidad de que el estimador $\hat{\theta}$ sea diferente del valor poblacional θ es cero. Formalmente, $\hat{\theta}$ es un estimador consistente si

$$\lim_{n \rightarrow \infty} P(|\theta - \hat{\theta}| < \delta) = 1 \quad (15.28)$$

donde δ es una constante positiva arbitrariamente pequeña.

Solución a ejercicio

Diseñado por Freepik

Soluciones del Capítulo 2

2.1 Partiendo de 2.10, y sin perder generalidad encontremos la derivada con respecto X_2 .

$$\frac{\partial Y_i}{\partial X_{2i}} = \alpha_2 \left(\alpha_0 X_{1i}^{\alpha_1} X_{2i}^{\alpha_2-1} X_{3i}^{\alpha_3} \varepsilon_i \right)$$

Ahora, reconciendo que

$$Y_i = \alpha_0 X_{1i}^{\alpha_1} X_{2i}^{\alpha_2} X_{3i}^{\alpha_3} \varepsilon_i$$

podemos reexpresar la deribada de la siguiente manera:

$$\frac{\partial Y_i}{\partial X_{2i}} = \alpha_2 \frac{Y_i}{X_{2i}}.$$

Ahora multiplicando a ambos lados por X_{2i} y dividiendo a ambos lados por Y_i obtenemos:

$$\frac{\frac{\partial Y_i}{Y_i}}{\frac{\partial X_{2i}}{X_{2i}}} = \alpha_2.$$

Ahora multiplicando por 100 el numerador y denominador de la mano derecha obtendremos:

$$\frac{\frac{\partial Y_i}{Y_i} 100}{\frac{\partial X_{2i}}{X_{2i}} 100} = \alpha_2.$$

Es fácil reconocer que el cambio porcentual de Y_i ($\Delta \% Y_i$) corresponde a $\frac{\partial Y_i}{Y_i} 100$. Así, tenemos que:

$$\frac{\Delta \% Y_i}{\Delta \% X_{2i}} = \alpha_2.$$

Es decir, α_2 representa el cambio porcentual en Y_i dado un cambio del 1% en X_{2i} . Esto representa la elasticidad.

2.2 El primer paso es cargar los datos.

```
dejercicio <- read.csv("../DataFinalCap/regmult.csv", sep = ";")
str(dejercicio)

## 'data.frame': 34 obs. of 7 variables:
## $ Año      : int 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 ...
## $ CE..T.    : int 1450 1477 1504 1531 1558 1585 1578 1614 1686 1670 ...
## $ CD..T.    : int 193 188 183 177 172 166 160 148 150 147 ...
## $ I..T.     : int 71135 76495 81855 87215 92575 97935 102072 107371 113521 116662 ...
## $ Ldies..T.: int 931 905 879 853 827 802 775 741 733 691 ...
## $ LEI..T.   : int 712 699 687 675 663 650 623 621 621 606 ...
## $ V..T.     : num 36127 61829 74066 183746 220029 ...
```

Los datos han sido leídos correctamente, pero los nombres de las variables no parecen los correctos.

```
names(dejercicio) <- c("Año", "CE", "CD", "I", "Ldies", "LEI",
                         "V")
str(dejercicio)

## 'data.frame': 34 obs. of 7 variables:
## $ Año : int 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 ...
## $ CE  : int 1450 1477 1504 1531 1558 1585 1578 1614 1686 1670 ...
## $ CD  : int 193 188 183 177 172 166 160 148 150 147 ...
## $ I   : int 71135 76495 81855 87215 92575 97935 102072 107371 113521 116662 ...
## $ Ldies: int 931 905 879 853 827 802 775 741 733 691 ...
## $ LEI  : int 712 699 687 675 663 650 623 621 621 606 ...
## $ V   : num 36127 61829 74066 183746 220029 ...
```

De esta manera, podemos proceder a estimar el modelo. Los resultados se reportan en el Cuadro 15.1.

Cuadro 15.1: Modelo del ejercicio estimado por MCO

<i>Dependent variable:</i>	
	I
CE	85.760*** (8.391)
CD	40.905 (77.222)
Ldies	−55.183** (20.472)
LEI	−148.368*** (23.928)
V	0.001 (0.001)
Constant	95,755.200*** (22,438.880)
<hr/>	
Observations	34
R ²	1.000
Adjusted R ²	1.000
Residual Std. Error	969.562 (df = 28)
F Statistic	20,032.750*** (df = 5; 28)

Note: *p<0.1; **p<0.05; ***p<0.01

Ahora podemos interpretar los coeficientes. En este caso tendremos que:

- Intercepto: Si todas las otras variables del modelo fuesen cero, entonces el sector tendría egresos de $9,57552 \times 10^4$ millones de dólares. Noten que este es un ejemplo en el que el intercepto no tiene sentido.
- Por cada millón adicional Kilovatios/hora de consumo de electricidad el ingreso del sector aumentará en 85.76 millones de dólares.
- Como se discutirá en el siguiente capítulo, el consumo de diesel no tiene efecto sobre el ingreso del sector (no es significativo).
- Por cada locomotora diesel adicional el ingreso del sector caerá en 55.18 millones de dólares.
- Por cada locomotora eléctrica adicional el ingreso del sector caerá en 148.37 millones de dólares.

- Como se discutirá en el siguiente capítulo, el número de viajeros no tiene efecto sobre el ingreso del sector (no es significativo).

Soluciones del Capítulo 3

3.1 Falta por hacer

Soluciones del Capítulo 4

4.1 El primer paso es cargar los datos y hacer las modificaciones que ya habíamos efectuado.

```
dejercicio <- read.csv("../DataFinalCap/regmult.csv", sep = ";")
str(dejercicio)
names(dejercicio) <- c("Año", "CE", "CD", "I", "Ldies", "LEI",
"V")
```

De esta manera, podemos proceder a estimar los modelos. Los resultados se reportan en el Cuadro 15.2.

Cuadro 15.2: Modelos del ejercicio estimados por MCO

	Dependent variable:				
	I				
	(1)	(2)	(3)	(4)	(5)
CE	85.760*** (8.391)		101.624*** (6.799)	121.950*** (8.456)	87.625*** (7.763)
CD	40.905 (77.222)			-382.801*** (39.674)	32.506 (75.251)
Ldies	-55.183** (20.472)	-123.944*** (27.362)			-52.426** (19.784)
LEI	-148.368*** (23.928)	-176.151*** (57.659)	-214.080*** (14.991)		-148.839*** (23.664)
V	0.001 (0.001)			0.0001 (0.002)	
Constant	95.755.200*** (22,438.880)	311.892.500*** (15,639.650)	76.169.700*** (20,492.610)	-31.642.690 (19,588.990)	92.505.850*** (21,602.510)
Observations	34	34	34	34	34
R ²	1.000	0.998	1.000	0.999	1.000
Adjusted R ²	1.000	0.998	1.000	0.999	1.000
Residual Std. Error	969.562 (df = 28)	2,543.147 (df = 31)	1,144.403 (df = 31)	1,581.001 (df = 30)	959.363 (df = 29)
F Statistic	20,032.750*** (df = 5; 28)	7,265.827*** (df = 2; 31)	35,942.490*** (df = 2; 31)	12,550.250*** (df = 3; 30)	25,576.090*** (df = 4; 29)

Note: *p<0.1; **p<0.05; ***p<0.01

Ahora calculemos las métricas de bondad de ajuste. Estas se resumen en el Cuadro 15.3.

Cuadro 15.3: Medidas de bondad de ajuste para los cinco modelos estimados

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
R2.ajustado	1.000	0.998	1.000	0.999	1.000
AIC	571.512	634.546	580.247	603.108	569.986
BIC	582.196	640.651	586.352	610.739	579.144

Así el modelo 5 es el que tiene las mejores métricas de bondad de ajuste.

Dado que todos los modelos están anidados, procedamos a compararlos empleando pruebas de hipótesis. Puedes realizar todas las comparaciones posibles. En todos los casos podemos encontrar que el mejor modelo es el modelo 4.11.

```
anova(R5, R1)

## Analysis of Variance Table
##
## Model 1: I ~ CE + CD + Ldies + LEI
## Model 2: I ~ CE + CD + Ldies + LEI + V
##   Res.Df     RSS Df Sum of Sq    F Pr(>F)
## 1     29 26690966
## 2     28 26321440  1    369526 0.3931 0.5358
```

Soluciones del Capítulo 5

5.1 Tras estimar los modelos se obtienen los resultados reportados en el Cuadro 15.4.

Cuadro 15.4: Modelos del ejercicio estimados por MCO

<i>Dependent variable:</i>			
	q		
	(1)	(2)	(3)
D	−205,487.400 (385,894.200)	−463,504.900*** (72,900.210)	
inversión	5.125*** (0.570)	4.836*** (0.380)	5.347*** (0.386)
D:inversión	−0.521 (0.765)		
Dinversión		−0.921*** (0.144)	
Constant	72,169.520 (288,169.600)	215,805.500 (196,090.500)	−42,420.020 (191,355.000)
Observations	228	228	228
R ²	0.479	0.478	0.479
Adjusted R ²	0.472	0.474	0.474
Residual Std. Error	550,142.900 (df = 224)	549,486.800 (df = 225)	549,266.300 (df = 225)
F Statistic	68.749*** (df = 3; 224)	103.138*** (df = 2; 225)	103.311*** (df = 2; 225)

Note: *p<0.1; **p<0.05; ***p<0.01

Es importante anotar que se requiere de un truco para estimar el modelo solo con interacción y sin intercepto. Para esto es necesario crear la variable de interacción aparte, pues la función **lm()** automáticamente se incluye el elemento de interacción, incluye el cambio en el intercepto. Por eso este modelo se estimó de la siguiente manera:

```
datos.dummy$Dinversión <- datos.dummy$D * datos.dummy$inversión
res1b <- lm(q ~ inversión + Dinversión, datos.dummy)
```

Ahora comparemos estos modelos.

```
## Analysis of Variance Table
##
## Model 1: q ~ D + inversión
## Model 2: q ~ D + inversión + D * inversión
```

```

##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1    225 6.7936e+13
## 2    224 6.7795e+13  1 1.4032e+11 0.4636 0.4966
## Analysis of Variance Table
##
## Model 1: q ~ inversión + Dinversión
## Model 2: q ~ D + inversión + D * inversión
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1    225 6.7881e+13
## 2    224 6.7795e+13  1 8.5819e+10 0.2836 0.5949
## Analysis of Variance Table
##
## Model 1: q ~ inversión
## Model 2: q ~ D + inversión
##   Res.Df      RSS Df  Sum of Sq      F     Pr(>F)
## 1    226 8.0141e+13
## 2    225 6.7936e+13  1 1.2206e+13 40.425 1.125e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Model 1: q ~ inversión
## Model 2: q ~ inversión + Dinversión
##   Res.Df      RSS Df  Sum of Sq      F     Pr(>F)
## 1    226 8.0141e+13
## 2    225 6.7881e+13  1 1.226e+13 40.638 1.025e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Model 1: q ~ inversión
## Model 2: q ~ D + inversión + D * inversión
##   Res.Df      RSS Df  Sum of Sq      F     Pr(>F)
## 1    226 8.0141e+13
## 2    224 6.7795e+13  2 1.2346e+13 20.396 7.284e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Es decir, el modelo con solo cambio en el intercepto o solo cambio en pendiente es mejor que el modelo sin ningún cambio. Al igual que lo es el que tiene el cambio en ambos. No obstante el modelo con cambio tanto en intercepto como en pendiente muestra coeficientes individualmente no significativos para los parámetros asociados con las variables dummy. Esto hace que esta decisión no sea fácil, pero claramente el modelo con cambio en pendiente e intercepto es mejor que los otros tres.

Soluciones del Capítulo 6

Soluciones del Capítulo 7

Soluciones del Capítulo 8

8.1 El primer paso es cargar los datos.

```
dejercicio <- read.csv("../DataFinalCap/datosmulti.csv", sep = ",")  
str(dejercicio)  
  
## 'data.frame': 38 obs. of  9 variables:  
## $ Edad           : int  46 31 23 19 23 47 30 28 23 29 ...  
## $ Peso           : int  180 175 100 185 159 170 137 192 150 120 ...  
## $ Altura_desde_zapato    : num  187 168 154 190 178 ...  
## $ Altura.desde.el.pie.descalzo: num  185 166 152 187 174 ...  
## $ Altura.sentado      : num  95.2 83.8 82.9 97.3 93.9 92.4 87.7 96.9 91.4 85.2 ...  
## $ Longitud_brazo     : num  36.1 32.9 26 37.4 29.5 36 32.5 35.8 29.4 26.6 ...  
## $ Longitud_muslo     : num  45.3 36.5 36.6 44.1 40.1 43.2 35.6 39.9 35.5 31 ...  
## $ Longitud_piernainferior: num  41.3 35.9 31 41 36.9 37.4 36.2 43.1 33.4 32.8 ...  
## $ Distancia_centro   : num  -206.3 -178.2 -71.7 -257.7 -173.2 ...
```

Y procedamos a estimar el modelo. Los resultados se reportan en el Cuadro 15.5.

Cuadro 15.5: Modelo del ejercicio estimado por MCO

<i>Dependent variable:</i>	
Distancia_centro	
Edad	0.526 (0.420)
Peso	0.004 (0.316)
Altura.desde.el.pie.descalzo	-3.349 (9.159)
Altura_desde_zapato	-0.865 (9.119)
Constant	530.369*** (139.055)
<hr/>	
Observations	38
R ²	0.656
Adjusted R ²	0.615
Residual Std. Error	37.030 (df = 33)
F Statistic	15.751*** (df = 4; 33)

Note:

*p<0.1; **p<0.05; ***p<0.01

Ahora miremos si existe multicolinealidad

```
vif(R1)

##                      Edad                  Peso
## 1.126843            3.458461
## Altura.desde.el.pie.descalzo Altura_desde_zapato
## 282.611282          278.854402
```

```
# matriz X
XTX <- model.matrix(R1)
# se calculan los valores propios
e <- eigen(t(XTX) %*% XTX)
# se muestran los valores propios
e$val
```

```

## [1] 3.206402e+06 2.027935e+04 8.958858e+03 8.322439e+00
## [5] 7.090495e-02

# se crea el valor propio mas grande
lambda.1 <- max(e$val)
# se crea el valor propio mas pequeño
lambda.k <- min(e$val)
# se calcula kappa
kappa <- sqrt(lambda.1/lambda.k)
kappa

## [1] 6724.666

```

Claramente las dos pruebas muestran un problema de multicolinealidad. En este caso la mejor opción para corregir el problema es eliminar variables con *VIF* grande. En este caso.

```

R2 <- remueve.VIF.grande(R1, 4)
summary(R2)

## 
## Call:
## lm(formula = myForm, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -91.029  -25.249    0.294  25.200   54.717 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           532.877125 137.104715  3.887 0.000448 ***
## Edad                  0.557597  0.406456  1.372 0.179094    
## Peso                 -0.008688  0.310512 -0.028 0.977842    
## Altura_desde_zapato -4.178042  0.996501 -4.193 0.000186 ***  
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 36.55 on 34 degrees of freedom
## Multiple R-squared:  0.6549, Adjusted R-squared:  0.6244 
## F-statistic: 21.5 on 3 and 34 DF,  p-value: 5.46e-08

vif(R2)

##                   Edad                  Peso Altura_desde_zapato
## 1.080473          3.418028          3.417264

```

Nota que para un análisis de los resultados, sería necesario eliminar las variables que no son significativas.

Soluciones del Capítulo 9

9.1 El siguiente código permite realizar el experimento de Monte Carlo para una muestra de 20.

```
# se fija una semilla
set.seed(1234557)

# creamos un objeto nulo para guardar si se rechaza h0 o no
# para el coeficiente de X2
X2.rechaza <- matrix("NA", N, 3)
# h0 o no para el coeficiente de X3
X3.rechaza <- matrix("NA", N, 3)
# se fija el tamaño de cada muestra
n = 20

# se fija el número de repeticiones en 10mil
N = 10000

for (i in 1:N) {
  # creación de variable explicativa
  X2 <- matrix(rnorm(n), n, 1)
  # creación de variable no significativa
  X3 <- matrix(rnorm(n), n, 1)
  # error homoscedástico
  err <- rnorm(n)
  # error heteroscedástico
  herr <- (X2^2) * err
  # variable explicativa sin heteroscedasticidad
  y1 <- 1 + X2 * 7 + err
  # variable explicativa con heteroscedasticidad
  y2 <- 1 + X2 * 7 + herr
  # estimación del modelo con homoscedasticidad
  res.homo <- summary(lm(y1 ~ X2 + X3))

  # guardamos los resultados regresión homoscedasticidad
  X2.rechaza[i, 1] <- res.homo$coefficients[2, 4] < 0.01
  X3.rechaza[i, 1] <- res.homo$coefficients[3, 4] < 0.01

  # estimación del modelo con heteroscedasticidad
  res.hetero.1 <- lm(y2 ~ X2 + X3)
  res.hetero <- summary(res.hetero.1)
  # guardamos los resultados regresión heteroscedasticidad con
  # MCO
  X2.rechaza[i, 2] <- res.hetero$coefficients[2, 4] < 0.01
  X3.rechaza[i, 2] <- res.hetero$coefficients[3, 4] < 0.01
```

```

# guardamos los resultados regresión heteroscedasticidad con
# HC4
coef.test.HC4 <- coeftest(res.hetero.1, vcov = (vcovHC(res.hetero.1,
  "HC4")))

X2.rechaza[i, 3] <- coef.test.HC4[2, 4] < 0.01
X3.rechaza[i, 3] <- coef.test.HC4[3, 4] < 0.01
}

# poniendo el nombre de las columnas
X2.rechaza <- as.data.frame(X2.rechaza)
names(X2.rechaza) <- c("MC0sinHetero", "MC0conHetero", "HC3")

X3.rechaza <- as.data.frame(X3.rechaza)
names(X3.rechaza) <- c("MC0sinHetero", "MC0conHetero", "HC3")

```

Ahora, miremos la proporción de rechazos de la hipótesis nula de no significancia para el caso de la pendiente que acompaña a x_2 . Recuerden que en este caso la hipótesis nula es incorrecta.

```

X2.rechaza[, 1] <- as.logical(X2.rechaza[, 1])
X2.rechaza[, 2] <- as.logical(X2.rechaza[, 2])
X2.rechaza[, 3] <- as.logical(X2.rechaza[, 3])
apply(X2.rechaza, 2, mean)

## MC0sinHetero MC0conHetero          HC3
##      1.0000      0.9997      0.9706

```

Ahora podemos replicar este resultado para los diferentes tamaños de muestra. ¿Qué concluyes?

Soluciones del Capítulo 11

11.1 Partamos del siguiente modelo

$$y_t = 1 + 7x_t + \varepsilon_t$$

donde ε_t estará autocorrelacionado (AR1) o no. Adicionalmente, estimemos un modelo con dos variables explicativas, x_i y z_i , para entender el comportamiento de las pruebas de significancia individual. Noten que la segunda variable no es significativa por construcción.

Consideremos el siguiente código que materializa este experimento.

```

set.seed(123)
# creamos un objeto nulo para guardar los coeficientes de x
x.estcoef <- NULL
# creamos un objeto nulo para guardar los coeficientes de z
z.estcoef <- NULL

```

```

n = 50
N = 10000
for (i in 1:N) {
  # Una variable explicativa
  x <- matrix(rnorm(n), n, 1)
  # Una variable explicativa no significativa
  z <- matrix(rnorm(n), n, 1)
  # error no.autocorrelacionado
  err <- arima.sim(model = list(order = c(0, 0, 0)), n = n)
  # error autocorrelacionado
  Auto.err <- arima.sim(model = list(ar = 0.7), n = n)

  y1 <- 1 + x * 7 + err # variable explicativa sin auto
  y2 <- 1 + x * 7 + Auto.err # variable explicativa con auto
  res.no.auto <- summary(lm(y1 ~ x + z))
  x.cf.no.auto <- res.no.auto$coefficients[2, 1]
  z.cf.no.auto <- res.no.auto$coefficients[3, 1]

  res.auto <- summary(lm(y2 ~ x + z))
  x.cf.auto <- res.auto$coefficients[2, 1]
  z.cf.auto <- res.auto$coefficients[3, 1]

  # guardando los resultados de la iteración
  x.estcoef <- rbind(x.estcoef, cbind(x.cf.no.auto, x.cf.auto))
  z.estcoef <- rbind(z.estcoef, cbind(z.cf.no.auto, z.cf.auto))

}

```

Ahora, comparemos los resultados para el caso de un error no autocorrelacionado y uno si autocorrelacionado para la pendiente que acompaña a x_i .

```

round(apply(x.estcoef, 2, mean), 2)

## x.cf.no.auto      x.cf.auto
##             7              7

round(apply(x.estcoef, 2, sd), 2)

## x.cf.no.auto      x.cf.auto
##          0.14         0.20

```

Los resultados muestran que para el estimador de la pendiente de x_i :

- Los estimadores MCO siguen siendo insesgados en presencia de autocorrelación.
- El error estándar de los coeficientes es más grande en presencia de autocorrelación.

Ahora, hagamos lo mismo para la pendiente que acompaña a z_i .

```
round(apply(z.estcoef, 2, mean), 2)

## z.cf.no.auto    z.cf.auto
##              0             0

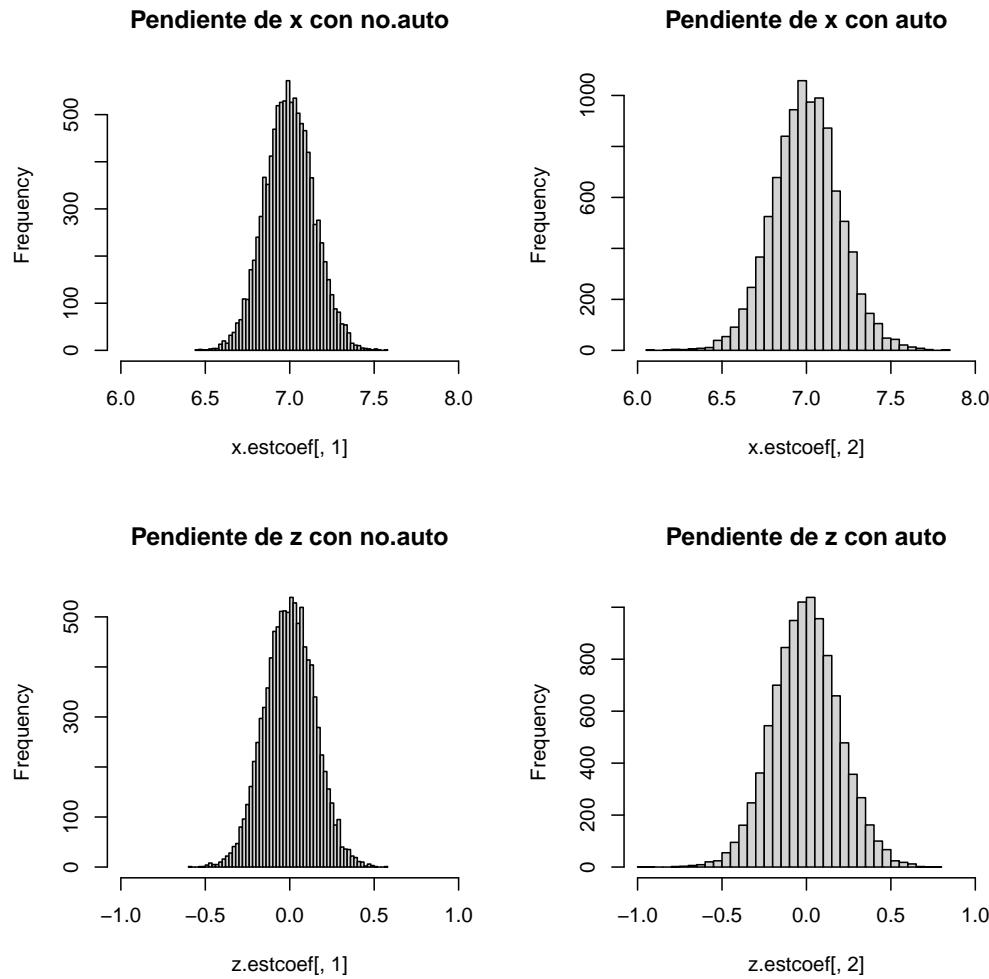
round(apply(z.estcoef, 2, sd), 2)

## z.cf.no.auto    z.cf.auto
##          0.15        0.20
```

Los resultados son similares al caso de la variable que no debería estar en el modelo.

Por otro lado, miremos por un momento los histogramas de la distribución de los coeficientes estimados.

```
par(mfrow = c(2, 2))
hist(x.estcoef[, 1], 50, main = "Pendiente de x con no.auto",
      xlim = c(6, 8))
hist(x.estcoef[, 2], 50, main = "Pendiente de x con auto", xlim = c(6,
      8))
hist(z.estcoef[, 1], 50, main = "Pendiente de z con no.auto",
      xlim = c(-1, 1))
hist(z.estcoef[, 2], 50, main = "Pendiente de z con auto", xlim = c(-1,
      1))
```



Es interesante que la distribución sigue teniendo forma acampana. ¿Qué concluyes?

11.2 Miremos lo que ocurre con la prueba de hipótesis individuales. Es decir, miremos que proporción de veces se rechaza la hipótesis nula de que cada una de las pendientes es igual cero con y sin autocorrelación. Y finalmente el efecto de emplear una corrección H.A.C para el caso de errores con autocorrelación.

Usemos el mismo experimento del ejercicio anterior para ver esto, pero usemos una muestra de 500. (tu podras reproducir este ejercicio para otros tamaños de muestra).

```
set.seed(123)

n = 500
N = 10000
# creamos un objeto nulo para guardar si se rechaza h0 o no
# para el coeficiente de z
```

```

z.rechaza <- matrix("NA", N, 3)

# creamos un objeto nulo para guardar si se rechaza h0 o no
# para el coeficiente de x. Las columnas serán los
# diferentes métodos
x.rechaza <- matrix("NA", N, 3)

library(sandwich)
library(AER)
library(lmtest)
for (i in 1:N) {
  # Una variable explicativa
  x <- matrix(rnorm(n), n, 1)
  # Una variable explicativa no significativa
  z <- matrix(rnorm(n), n, 1)
  # error no.autocorrelacionado
  err <- arima.sim(model = list(order = c(0, 0, 0)), n = n)
  # error autocorrelacionado
  Auto.err <- arima.sim(model = list(ar = 0.7), n = n)
  # variable explicativa sin auto
  y1 <- 1 + x * 7 + err
  # variable explicativa con auto
  y2 <- 1 + x * 7 + Auto.err
  modelo.no.auto <- lm(y1 ~ x + z)
  res.no.auto <- summary(modelo.no.auto)
  x.rechaza[i, 1] <- res.no.auto$coefficients[2, 4] < 0.01
  z.rechaza[i, 1] <- res.no.auto$coefficients[3, 4] < 0.01

  modelo.auto <- lm(y2 ~ x + z)
  res.auto <- summary(lm(y2 ~ x + z))
  x.rechaza[i, 2] <- res.auto$coefficients[2, 4] < 0.01
  z.rechaza[i, 2] <- res.auto$coefficients[3, 3] < 0.01

  # con corrección HCA NeweyWest
  coef.test.NeweyWesy <- coeftest(modelo.auto, vcov = NeweyWest(modelo.auto))
  x.rechaza[i, 3] <- coef.test.NeweyWesy[2, 4] < 0.01
  z.rechaza[i, 3] <- coef.test.NeweyWesy[3, 4] < 0.01
}

# poniendo el nombre de las columnas
x.rechaza <- as.data.frame(x.rechaza)
names(x.rechaza) <- c("OLS_sin_auto", "OLS_con_auto", "HCA_NeweyWest")

z.rechaza <- as.data.frame(z.rechaza)
names(z.rechaza) <- c("OLS_sin_auto", "OLS_con_auto", "HCA_NeweyWest")

```

Ahora, miremos la proporción de rechazos de la hipótesis nula de no significancia para el caso de la pendiente que acompaña a x . Recuerden que en este caso la hipótesis nula es incorrecta.

```
str(x.rechaza)

## 'data.frame': 10000 obs. of 3 variables:
## $ OLS_sin_auto : chr "TRUE" "TRUE" "TRUE" "TRUE" ...
## $ OLS_con_auto : chr "TRUE" "TRUE" "TRUE" "TRUE" ...
## $ HCA_NeweyWest: chr "TRUE" "TRUE" "TRUE" "TRUE" ...

x.rechaza[, 1] <- as.logical(x.rechaza[, 1])
x.rechaza[, 2] <- as.logical(x.rechaza[, 2])
x.rechaza[, 3] <- as.logical(x.rechaza[, 3])
str(x.rechaza)

## 'data.frame': 10000 obs. of 3 variables:
## $ OLS_sin_auto : logi TRUE TRUE TRUE TRUE TRUE ...
## $ OLS_con_auto : logi TRUE TRUE TRUE TRUE TRUE ...
## $ HCA_NeweyWest: logi TRUE TRUE TRUE TRUE TRUE ...

apply(x.rechaza, 2, mean)

##   OLS_sin_auto   OLS_con_auto HCA_NeweyWest
##                 1                  1                  1
```

Los resultado muestran que independiente del método que se emplee en todos los casos se rechaza la hipótesis nula correctamente.

Finalmente, concentrémos en el efecto sobre la prueba de significancia para el coeficiente asociado a z . Recuerden que en este caso la hipótesis nula es correcta.

```
str(z.rechaza)

## 'data.frame': 10000 obs. of 3 variables:
## $ OLS_sin_auto : chr "FALSE" "FALSE" "FALSE" "FALSE" ...
## $ OLS_con_auto : chr "TRUE" "FALSE" "TRUE" "TRUE" ...
## $ HCA_NeweyWest: chr "FALSE" "FALSE" "FALSE" "FALSE" ...

z.rechaza[, 1] <- as.logical(z.rechaza[, 1])
z.rechaza[, 2] <- as.logical(z.rechaza[, 2])
```

```
z.rechaza[, 3] <- as.logical(z.rechaza[, 3])
str(z.rechaza)

## 'data.frame': 10000 obs. of 3 variables:
## $ OLS_sin_auto : logi FALSE FALSE FALSE FALSE TRUE ...
## $ OLS_con_auto : logi TRUE FALSE TRUE TRUE FALSE TRUE ...
## $ HCA_NeweyWest: logi FALSE FALSE FALSE FALSE FALSE ...

apply(z.rechaza, 2, mean)

##   OLS_sin_auto  OLS_con_auto HCA_NeweyWest
##       0.0107        0.4995      0.0116
```

Ahora, noten que sin autocorrelación y empleando MCO, la proporción de veces que se rechaza incorrectamente la nula de no significancia es aproximadamente 1%. El nivel de significancia seleccionado.

Si se emplea erroneamente MCO en presencia de autocorrelación, se rechaza erroneamente el 50% de las veces la hipótesis nula. ¡Como tirar una moneda! Pero si se emplea la corrección H.A.C. de Newey West, la proporción de rechazos erróneos se aproxima al nivel de significancia empleado.

Noten que esto muestra que la solución de HCA mejora la inferencia en presencia de autocorrelación.

Ahora podemos replicar este resultado para los diferentes tamaños de muestra. ¿Qué concluyes?

11.3

Bibliografía

Diseñado por Freepik

- Frisch, R. (1933). Editor's Note. *Econometrica*, 1, 1.
- Spanos, A. (1986). *Statistical Foundations of Econometric Modelling*. Cambridge University Press.
- Gujarati, D. N. & Porter, D. C. (2011). *Econometria básica-5*. Amgh Editora.
- R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A. & Crowley, J. (2021). *GGally: Extension to 'ggplot2'* [R package version 2.1.2]. <https://CRAN.R-project.org/package=GGally>
- Hlavac, M. (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables* [R package version 5.2.2]. Central European Labour Studies Institute (CELSI). Bratislava, Slovakia. <https://CRAN.R-project.org/package=stargazer>
- Alonso, J. C. (2021). Una introducción a los Loops en R (y algunas alternativas). *Economics Lecture Notes*, (14), 1-18.
- Ryan, J. A. & Ulrich, J. M. (2020). *xts: eXtensible Time Series* [R package version 0.12-0]. <https://CRAN.R-project.org/package=xts>
- Alonso, J. C. (2006a). 4 Hechos Estilizados de las series de rendimientos: Una ilustración para Colombia. *Estudios Gerenciales*.
- Alonso, J. C. & Torres, G. (2014). Características estadísticas del índice general de la Bolsa de Valores de Colombia (IGBC) en sus primeros 10 años. *Journal of Economics Finance and Administrative Science*, 19(36), 45-54. <http://www.sciencedirect.com/science/article/pii/S2077188614000031>
- Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R* [ISBN 978-0-387-77316-2]. Springer-Verlag. <https://CRAN.R-project.org/package=AER>
- Alonso, J. C. & Berggrun, L. (2011). *Introducción al análisis de riesgos financiero*. Ecoe Ediciones.

- Alonso, J. C. & Gallo, B. E. (2013). The Day-of-the-Week Effect: The CIVETS Stock Markets Case. *Journal of Applied Business and Economics*, 15(3), 102-116.
- Alonso, J. C. & Hoyos, C. C. (2021). *Introducción a los pronósticos con modelos estadístico de series de tiempo para científico de datos (en R)* (Universidad Icesi). Universidad Icesi.
- Brown, C. (2012). *dummies: Create dummy/indicator variables flexibly and efficiently* [R package version 1.5.6]. <https://CRAN.R-project.org/package=dummies>
- Kaplan, J. (2020). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables* [R package version 1.6.3]. <https://CRAN.R-project.org/package=fastDummies>
- Hebbali, A. (2020). *olsrr: Tools for Building OLS Regression Models* [R package version 0.5.3]. <https://CRAN.R-project.org/package=olsrr>
- Lumley, T. (2020). *leaps: Regression Subset Selection* [R package version 3.1]. <https://CRAN.R-project.org/package=leaps>
- Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W. & Wells, M. T. (2013). AIC, Cp and estimators of loss for elliptically symmetric distributions. *arXiv preprint arXiv:1308.2766*.
- Orestes Cerdeira, J., Duarte Silva, P., Cadima, J. & Minhoto, M. (2020). *subselect: Selecting Variable Subsets* [R package version 0.15.2]. <https://CRAN.R-project.org/package=subselect>
- Lê, S., Josse, J. & Husson, F. (2008). FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1-18. <https://doi.org/10.18637/jss.v025.i01>
- Fernandes, K., Vinagre, P. & Cortez, P. A proactive intelligent decision support system for predicting the popularity of online news. En: *Portuguese Conference on Artificial Intelligence*. Springer. 2015, 535-546.
- Alonso Cifuentes, J. C. (2020). Una introducción a la construcción de WordClouds (para economistas) en R. *Economics Lecture Notes*, (9), 1-28. <https://www.researchgate.net/publication/341829699>
- Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software*, 17(1), 1-27.
- Long, J. A. (2020). *jtools: Analysis and Presentation of Social Scientific Data* [R package version 2.1.0]. <https://cran.r-project.org/package=jtools>
- Hair, J. F. J., Black, W. C., Babin, B. J. & Anderson, R. E. (2014). *Multivariate Data Analysis* (7.^a ed.). Pearson.
- Sheather, S. (2009). *A modern approach to regression with R*. Springer Science & Business Media.
- Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). *Applied Linear Regression Models* (4 th). McGraw-Hill Irwin.
- Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, Inc. <https://doi.org/10.1002/0471725153>
- Cule, E. & De Iorio, M. (2012). A semi-automatic method to guide the choice of ridge parameter in ridge regression. *arXiv preprint arXiv:1205.0686*, 1-32. <http://arxiv.org/abs/1205.0686>
- Alonso, J. C. (2020). Herramientas del Business Analytics en R : Análisis de Componentes Principales para resumir variables. *Economics Lecture Notes*, (10), 1-32. <https://www.researchgate>.

- net/publication/341829708{_}Herramientas{_}del{_}Business{_}Analitycs{_}en{_}R{_}Analisis{_}de{_}Componentes{_}Principales{_}para{_}resumir{_}variables
- Bernat, L. F. (2004). Diferencias Salariales por Género en las siete principales áreas metropolitanas Colombianas. ¿Evidencia de Discriminación? *Investigaciones sobre género y desarrollo en Colombia*, (1).
- Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14(3), 693-709.
- Fox, J. & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third). Sage. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Cule, E., Moritz, S. & Frankowski, D. (2021). *ridge: Ridge Regression with Automatic Selection of the Penalty Parameter* [R package version 2.9]. <https://CRAN.R-project.org/package=ridge>
- Auguie, B. (2017). *gridExtra: Miscellaneous Functions for “Grid” Graphics* [R package version 2.3]. <https://CRAN.R-project.org/package=gridExtra>
- Breusch, T. S. & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47(5), 1287-94.
- Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17(1), 107-112.
- White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4), 817-38.
- Davidson, Russell and MacKinnon, J. (1993). *Estimation and inference in econometrics*. Oxford University Press.
- Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, 45(2), 215-233.
- Long, J. & Ervin, L. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, 54(3), 217-224. <http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2000.10474549>
- Zeileis, A. & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, 2(3), 7-10. <https://CRAN.R-project.org/doc/Rnews/>
- Alonso, J. C. & Montenegro, S. (2015). Estudio de Monte Carlo para comparar 8 pruebas de normalidad sobre residuos de mínimos cuadrados ordinarios en presencia de procesos autorregresivos de primer orden. *Estudios Gerenciales*, 31(In press), 253-265. <https://doi.org/10.1016/j.estger.2014.12.003>
- Shapiro, S. S. & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337), 215-216. <http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1972.10481232>
- Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Attuari*, 89-91.
- Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1, 13-74.
- Anderson, T. & Darling, D. (1952). Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *The annals of mathematical statistics*, 23, 193-212. <http://www.jstor.org/stable/2236446>

- Waldman, D. M. (1983). A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity. *Economics Letters*, 13(2-3), 197-200.
- Farrar, T. J. (2020). *skedastic: Heteroskedasticity Diagnostics for Linear Regression Models* [R Package Version 1.0.0]. Bellville, South Africa. <https://github.com/tjfarrar/skedastic>
- Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software*, 11(10), 1-17. <https://doi.org/10.18637/jss.v011.i10>
- Jarque, C. & Bera, A. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55(2), 163-172. <http://www.jstor.org/stable/1403192>
- Trapletti, A. & Hornik, K. (2019). *tseries: Time Series Analysis and Computational Finance* [R package version 0.10-47.]. <https://CRAN.R-project.org/package=tseries>
- Wald, A. & Wolfowitz, J. (1940). On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics*, 11(2), 147-162.
- Durbin, J. & Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, 38(1-2), 159-178.
- Durbin, J. (1970). Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables. *Econometrica*, 38(3), 410-21.
- Box, G. E. P. & Pierce, D. A. (1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, 65(332), 1509-1526.
- Ljung, G. M. & Box, G. E. P. (1979). The Likelihood Function of Stationary Autoregressive-Moving Average Models. *Biometrika*, 66(2), 265-270.
- Breusch, T. S. (1978). Testing for Autocorrelation in Dynamic Linear Models. *Australian Economic Papers*, 17(31), 334-55.
- Godfrey, L. G. (1978). Testing against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables. *Econometrica*, 46(6), 1293-1301.
- Newey, W. K. & West, K. D. (1987). A Simple , Positive Semi-Definite , Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703-708.
- Hyndman, R. J. & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3), 1-22. <http://www.jstatsoft.org/article/view/v027i03>
- Wickham, H. & Bryan, J. (2019). *readxl: Read Excel Files* [R package version 1.3.1]. <https://CRAN.R-project.org/package=readxl>
- Bansal, G. (2021). *ecm: Build Error Correction Models* [R package version 6.3.0]. <https://CRAN.R-project.org/package=ecm>
- Andrews, D. (1991). Heteroskedasticity and autocorrelation consistent covariant matrix estimation. *Econometrica*, 59(3), 817-858.
- Lumley, T. & Heagerty, P. (1999). Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2), 459-477.
- Durbin, J. & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression. I. *Biometrika*, 37(3-4), 409-428.
- Durbin, J. & Watson, G. S. (1971). Testing for serial correlation in least squares regression.III. *Biometrika*, 58(1), 1-19.

- Swed, F. S. & Eisenhart, C. (1943). Tables for Testing Randomness of Grouping in a Sequence of Alternatives. *The Annals of Mathematical Statistics*, 14(1), 66-87.
- Durbin, J. (1960). Estimation of Parameters in Time-Series Regression Models. *Journal of the Royal Statistical Society*, 22(1), 139-153.
- Kuhn, M. (2020). *caret: Classification and Regression Training* [R package version 6.0-86]. <https://CRAN.R-project.org/package=caret>
- Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge university press.
- Redmond, M. & Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3), 660-678.
- Alonso, J. C. (2006b). Apuntes de álgebra matricial para un curso introductorio de econometría. *Apuntes de economía; No. 11-2006*.
- Chiang, A. (1996). Métodos fundamentales de economía.
- Alonso, J. C. (2007). APUNTES DE ESTADÍSTICA PARA UN CURSO INTRODUCTORIO DE ECONOMETRÍA. *Apuntes de Economía*, (12).
- Amemiya, T. (1994). *Introduction to statistics and econometrics*. Harvard University Press.

Índice alfabético

Diseñado por Freepik

- Analítica
 - descriptiva, 16
 - diagnóstica, 16, 24, 34, 46
 - predictiva, 16, 24
 - prescriptiva, 17
- Aporte relativo, 37
- Aprendizaje
 - no supervisado, 12, 13
 - supervisado, 12, 14
- Business analytics, 10
- Coeficientes
 - estandarizados, 34
- Comparación de modelos
 - no anidados, 30, 31
- Criterio de Información
 - Bayesiano, 32
 - de Akaike, 32
- Data Generating Process (DGP), 19
- Econometría, 15, 17
- Función
 - AIC(), 32
 - BIC(), 32
 - calc.relimp(), 35, 37
 - jtest(), 31
- plot_summs(), 40
- read.csv(), 26
- LDA, 25
- Modelo
 - econométrico, 18
- Métricas de bondad de ajuste, 32
- Palabra vacía, 25
- Paquete
 - jtools, 40
 - relaimpo, 35
- Predicción, 17
- Pronóstico, 17
- Prueba
 - J, 31
- Regresión
 - stepwise backward, 26, 28
 - stepwise forward, 26, 27
 - stepwise forward y backward, 26, 29
- Ruta analítica, 26
- Stop word, 25
- Tarea de
 - clasificación, 12
 - clustering, 11

detección de excepciones, 13
encontrar asociaciones, 13
estimar regresiones, 13
pronosticar, 14
resumir, 11
visualización, 11
Tareas en la analítica, 11
Trampa de las variables dummy, 27

Variable

accionable, 24, 34, 40
más importante, 34
no accionable, 40