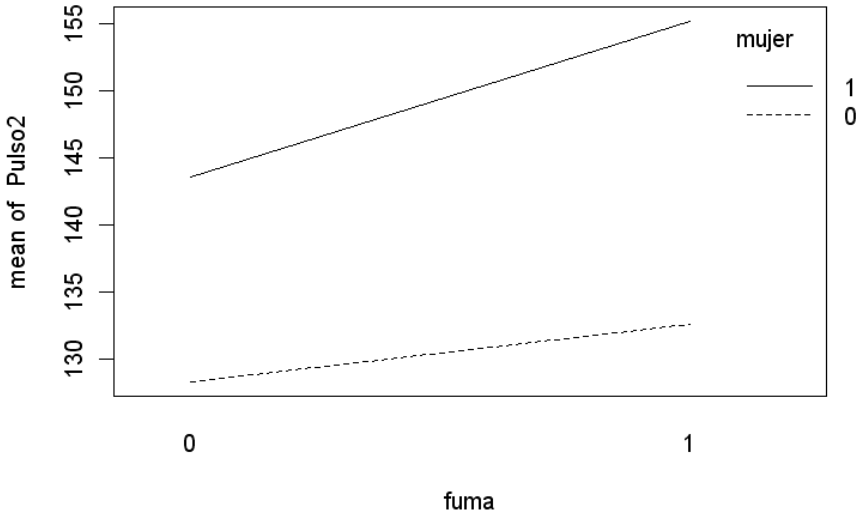
Figura 57: Gráfico de las medias muestrales de los cuatro grupos, de los datos de pulso-post ejercicio.

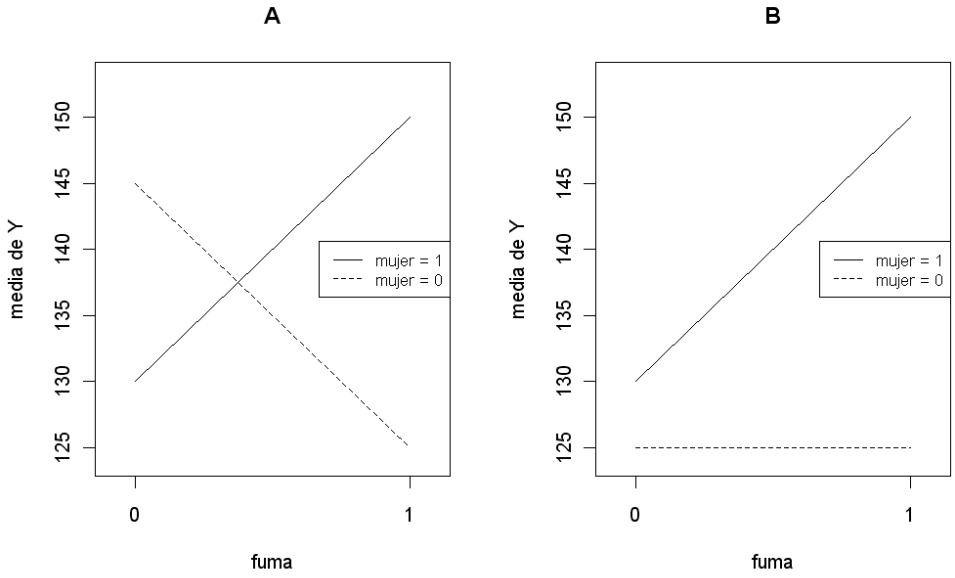


5. Diagnóstico del modelo

5.1. Diagnóstico del modelo: definiciones y gráficos

Los métodos de inferencia anteriormentes descriptos (cálculo de p-valores, intervalos de confianza y de predicción, por citar algunos) requieren que se satisfagan los cuatro supuestos (45) que subyacen a nuestro modelo. El diagnóstico del modelo consiste en la validación de estos supuestos para los datos en cuestión. Esta validación puede hacerse a través de una serie de gráficos, muchos de los cuales ya describimos en la regresión lineal simple, o bien a través de diversos cálculos. El diagnóstico desempeña un papel importante en el desarrollo y la evaluación de los modelos regresión múltiple. La mayoría de los procedimientos de diagnóstico para la regresión lineal simple que hemos descrito anteriormente se trasladan directamente a la regresión múltiple. A continuación revisaremos dichos procedimientos de diagnóstico.
Por otro lado, también se han desarrollado herramientas de diagnóstico y procedimientos especializados para la regresión múltiple. Algunas de las más importantes se discuten en la Sección 5.2.

Figura 58: Gráficos de las medias de una variable respuesta Y para dos ejemplos ficticios, en las figuras A y B.

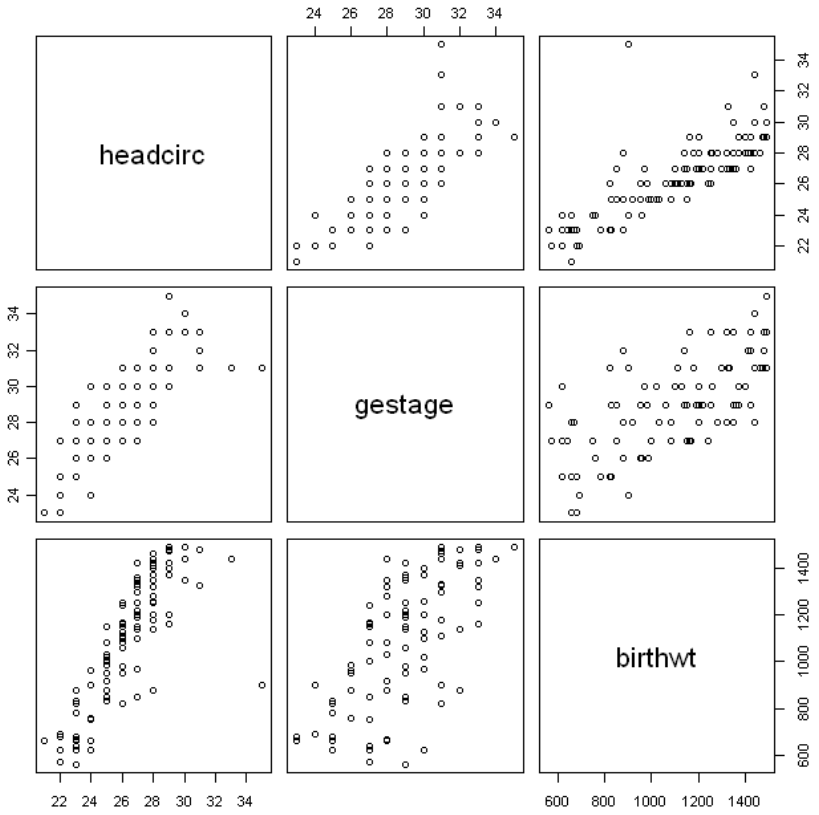


5.1.1. Matriz de scatter plots o gráficos de dispersión

Los boxplots, histogramas, diagramas de tallo y hojas, y gráficos de puntos para cada una de las variables predictoras y para la variable de respuesta pueden proporcionar información univariada preliminar y útil sobre estas variables. Los diagramas de dispersión (scatter plots) de la variable de respuesta versus cada variable predictora pueden ayudar a determinar la naturaleza y la fuerza de las relaciones bivariadas entre cada una de las variables de predicción y la variable de respuesta así como pueden permitir la identificación de lagunas en las regiones de datos. También pueden permitir identificar outliers u observaciones atípicas o alejadas del patrón del resto de los datos. Los diagramas de dispersión de cada variable predictora versus cada una de las otras variables de predicción son útiles para el estudio de las relaciones bivariadas entre las distintas variables predictoras y también para buscar espacios con ausencia de datos y detectar valores atípicos. El análisis resulta más fácil si los gráficos de dispersión se ensamblan en una matriz diagrama de dispersión (*scatter plot matrix*), como vemos en la Figura 59. En esta figura, la variable graficada en el eje vertical para cualquier gráfico de dispersión es aquella cuyo nombre se encuentra en su fila, y la variable graficada en el eje

horizontal es aquella cuyo nombre se encuentra en su columna. Por lo tanto, la matriz de gráfico de dispersión en la Figura 59 muestra en la primera fila los gráficos de Y (perímetro cefálico: `headcirc`) versus X_1, (edad gestacional: `gestage`) y de Y versus X_2 (peso: `birthwt`). En la segunda fila tenemos los gráficos de X_1 versus Y y de X_1 versus X_2. Finalmente, en la tercer fila tenemos los gráficos de X_2 versus Y y de X_2 versus X_1. Una matriz de diagramas de dispersión facilita el estudio de las relaciones entre las variables mediante la comparación de los diagramas de dispersión dentro de una fila o una columna. Esta matriz muestra, por supuesto, información repetida de los datos. Bastaría con dar los scatter plots que quedan por encima (o bien, por debajo) de la diagonal.
Si el dataframe que contiene a los datos se denomina `low`, la matriz de scatterplots se realiza con `pairs(low)`, en `R`.

Figura 59: Matriz de scatter plots para los datos de bebés con bajo peso, con las covariables edad gestacional y peso



Un complemento a la matriz de diagramas de dispersión que puede ser útil a veces es la matriz de correlaciones. Esta matriz contiene los coeficientes de correlación simple $r_{YX_1}, r_{YX_2}, ..., r_{YX_{p-1}}$ entre Y y cada una de las variables predictoras, así como todos los coeficientes de correlación simple entre las distintas variables predictoras entre sí. El formato de la matriz de correlación sigue el de la matriz de scatter plots

$$\begin{bmatrix} 1 & r_{YX_1} & r_{YX_2} & \cdots & r_{YX_{p-1}} \\ r_{YX_1} & 1 & r_{X_1X_2} & \cdots & r_{X_1X_{p-1}} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{YX_{p-1}} & r_{X_1X_{p-1}} & r_{X_2X_{p-1}} & \cdots & 1 \end{bmatrix}$$

y en el caso de los datos de bebés de bajo peso es

```
> cor(low)
          headcirc   gestage   birthwt
headcirc 1.0000000 0.7806919 0.7988372
gestage  0.7806919 1.0000000 0.6599376
birthwt  0.7988372 0.6599376 1.0000000
```

Observemos que la matriz de correlación es simétrica y en la diagonal contiene unos pues el coeficiente de correlación de una variable consigo misma es 1.

5.1.2. Gráficos de dispersión en tres dimensiones

Algunos paquetes estadísticos proporcionan gráficos de dispersión en tres dimensiones y permiten girar estos gráficos para permitir al usuario ver la nube de puntos desde diferentes perspectivas. Esto puede ser muy útil para identificar los patrones que sólo se desprenden de la observación desde ciertas perspectivas. En `R` la instrucción `plot3d` de la librería `rgl` permite hacerlo. Incluso se puede rotar el gráfico y guardar la película de la rotación con la instrucción `movie3d`.

5.1.3. Gráficos de residuos

Es importante recalcar que aunque las observaciones $(X_{i1}, X_{i2}, \ldots, X_{i(p-1)}, Y)$ no puedan graficarse en el caso de tener más de dos covariables, siempre tanto el valor predicho o ajustado $\widehat{Y}_i$ como el residuo e_i están en $\mathbb{R}$ y pueden ser graficados. De modo que un gráfico de los residuos contra los valores ajustados es útil para evaluar la idoneidad de la regresión múltiple para modelar los datos observados, y la homoscedasticidad (constancia de la varianza) de los términos de error. También

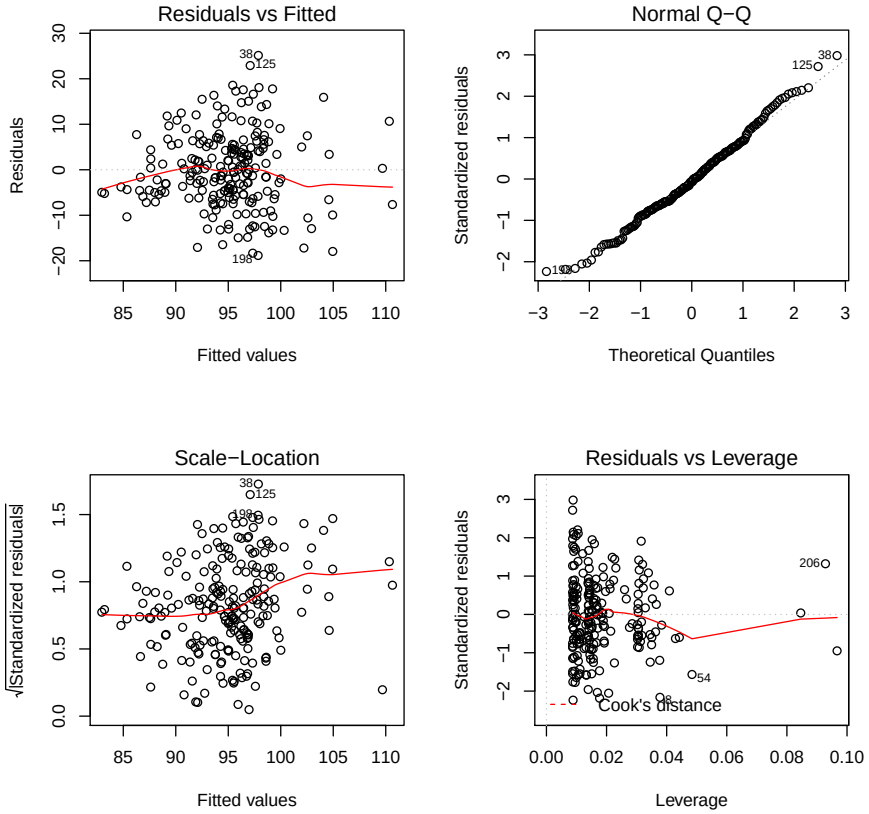
permite proporcionar información acerca de los valores extremos como lo vimos en la regresión lineal simple. Del mismo modo, un gráfico de los residuos contra el tiempo (u orden en el que fueron recopilados los datos, si este fuera relevante) o en contra de otra secuencia puede proporcionar información acerca de las posibles correlaciones entre los términos de error en la regresión múltiple. Boxplots y gráficos de probabilidad normal de los residuos son útiles para examinar si el supuesto de distribución normal sobre los términos de error se satisface razonablemente para los valores observados.
Además, los residuos deben ser graficados versus cada una de las variables predictivas. Cada uno de estos gráficos pueden proporcionar más información sobre la idoneidad de la función de regresión con respecto a la variable de predicción (por ejemplo, si un efecto que dé cuenta de la curvatura es necesario para dicha variable) y también puede proporcionar información sobre la posible variación de la varianza del error en relación con dicha variable predictora.
Los residuos también deben ser graficados versus cada una de las variables de predicción importantes que se omitieron del modelo, para ver si las variables omitidas tienen importantes efectos adicionales sobre la variable de respuesta que aún no han sido reconocidos en el modelo de regresión. Además, los residuos deben graficarse versus términos de interacción para los posibles efectos no incluidos en el modelo de regresión (trabajamos con interacción en la Sección 4.16 y subsiguientes), para ver si es necesario incluir algún término de interacción en el modelo.
Un gráfico de los residuos o los residuos al cuadrado contra los valores ajustados es útil para examinar si la varianza de los términos de error es constante. Si se detecta que la varianza no es constante, suele ser apropiado realizar gráficos del valor absoluto de los residuos, o de sus cuadrados, versus cada una de las variables predictoras. Estos gráficos pueden permitir la identificación de una o más variables predictoras con las que se relaciona la magnitud de la variabilidad del error. Se puede incluir una transformación de esta variable en el modelo para tratar de eliminar la estructura observada en los residuos.
Por supuesto, cualquier paquete estadístico realizará estos gráficos de manera más o menos automática. R produce automáticamente cuatro gráficos de diagnóstico cuando uno aplica la función `plot()` directamente a una salida del `lm()`. En general, este comando produce un gráfico por vez, apretando enter en el teclado producirá un gráfico por vez. Para ver los cuatro juntos, una opción es dividir la ventana gráfica en cuatro con la instrucción `mfrow`:

```
par(mfrow=c(2,2))
plot(ajuste4)
```

Alternativamente, uno puede calcular los residuos a partir de un ajuste lineal con las instrucciones `residuals()`, `rstudent()` para los residuos estudentizados,

`rstandard()` para los estandarizados, `fitted.values()` para los valores predichos. Y luego graficarlos usando la función `plot`. En la Figura 60 vemos el `plot` de los residuos para el `ajuste4` propuesto en el modelo (73), página 173.

Figura 60: Gráficos de diagnóstico producidos por la instrucción plot aplicada a la salida lm. Se grafican, los residuos versus los valores ajustados, el Q-Q plot normal de los residuos estandarizados, la raíz cuadrada de los residuos versus los valores ajustados, y finalmente los residuos estandarizados versus los leverage.



5.2. Identificación de outliers y puntos de alto leverage

Como en el caso de regresión lineal simple, aquí también tiene sentido identificar las observaciones que no siguen el patrón de los demás datos. Medidas para identificar estas observaciones son, nuevamente, el leverage, los residuos estudentizados, y algunas que no estudiaremos aquí como los DFFITS y los DFBETAS.

5.2.1. Ajuste robusto: permite ignorar a los outliers automáticamente

Como vimos para regresión lineal simple, el problema de las observaciones atípicas en un conjunto de datos es que su presencia puede influir de manera dramática sobre el ajuste del modelo propuesto, incluso llegando a tergiversar por completo las conclusiones que se pueden extraer de él cuando el ajuste se lleva a cabo usando estimadores de mínimos cuadrados. Esta distorsión de los valores ajustados además tiene la potencia de enmascarar las observaciones atípicas y muchas veces disimularlas entre los datos, dificultando el buen funcionamiento de las herramientas de detección de atipicidad más difundidas en el área: leverage, distancias de Cook, dfits, etc. Como ya dijimos en la Sección 3.2.4, una forma automática de evitar estos problemas consiste en cambiar el método de estimación de mínimos cuadrados por un ajuste robusto de los coeficientes. Los MM-estimadores de regresión son una buena alternativa. El ajuste que presentamos en la Sección 3.2.4 a traves de la rutina `lmrob` de la librería `robustbase` se extiende trivialmente para el caso de regresión lineal múltiple. A modo de ejemplo, en la Tabla 51 vemos el ajuste de dicha rutina a los datos del archivo `azucar` considerados previamente, el mismo ajuste de mínimos cuadrados figura en la Tabla 36.
Si el ajuste robusto y el clásico (o sea el de mínimos cuadrados) no difieren, esta es una señal de que no hay observaciones atípicas en el conjunto de datos con el que se trabaja. Como el ajuste por mínimos cuadrados es el más difundido en estadística, cuando el interés del análisis incluya la comunicación de los resultados a otros especialistas, es recomendable reportar la salida obtenida con el ajuste clásico. Esto es lo que sucede para el ajuste de los datos de `azucar`.
En cambio, cuando el ajuste robusto y el clásico difieren entre sí, esto se deberá a la presencia de datos atípicos. Con el ajuste robusto estos se podrán detectar claramente: corresponderán a aquellas observaciones cuyos pesos (robustos) asignados por el ajuste del `lmrob` sean muy chicos (pesos cero o muy cercanos a él). Estos se calculan como `ajusterob$rweights` para el ajuste presentado en la Tabla 51. Como los pesos, que van entre 0 y 1, se calculan en función de los residuos, un criterio equivalente será investigar aquellas observaciones con residuos grandes del ajuste robusto. En el caso de los datos de `azucar`, vemos que el menor peso corresponde a 0,24. La instrucción `plot` aplicada al ajuste robusto dado por `lmrob` proporciona 5 gráficos que permite visualizar las observaciones extremas, basadas en el cómputo del leverage robusto (*robust distances*) y que pueden verse en la Figura 61 para el ajuste robusto de los datos de `azucar`. Una descripción más detallada de los estimadores robustos disponibles puede consultarse en los Capítulos 4 y 5 de Maronna et al. [2006].

Tabla 51: Ajuste de un MM-estimador de regresión para el modelo con `glucosa` como variables respuesta y `peso.evo` (categórica) y `bmi` (numérica), archivo de datos `azucar`. El ajuste clásico puede verse en la Tabla 36, página 177.

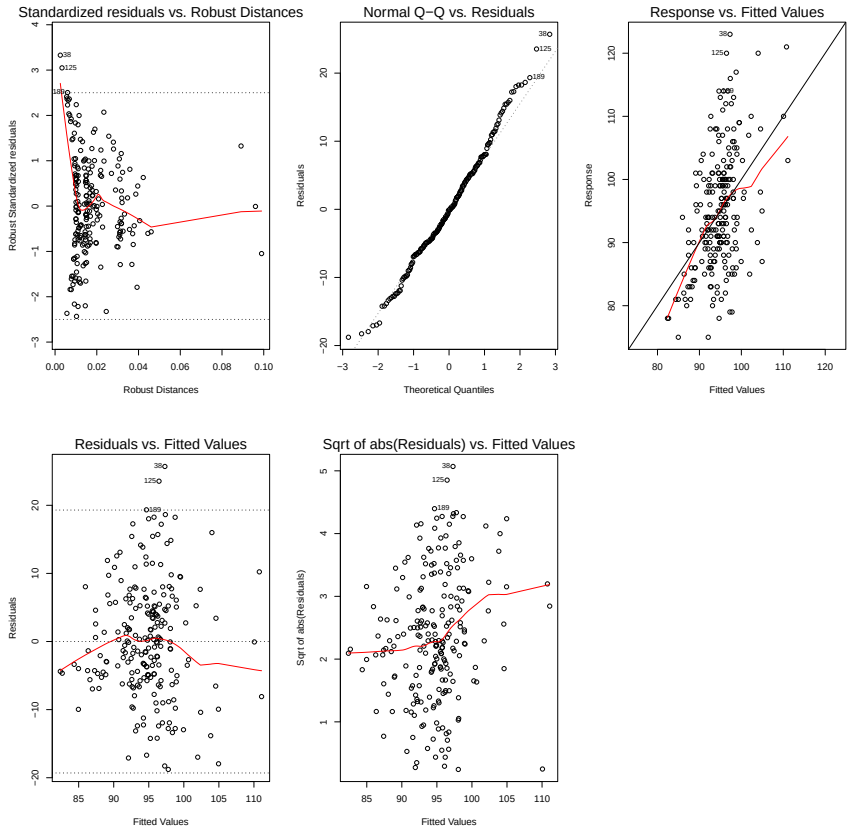
```
> library(robustbase)
> Ievo<-factor(peso.evo)
> ajusterob <- lmrob(glucosa ~ bmi + Ievo, data = azucar)
> summary(ajusterob)
 \--> method = "MM"
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.6589     3.7731  18.727  < 2e-16 ***
bmi           0.6552     0.1356   4.832 2.56e-06 ***
Ievo2         7.2255     1.3691   5.278 3.18e-07 ***
Ievo3         7.9919     1.4531   5.500 1.07e-07 ***
---
Robust residual standard error: 7.721
Multiple R-squared:  0.2222,        Adjusted R-squared:  0.2114
Convergence in 12 IRWLS iterations

Robustness weights:
 18 weights are ~= 1. The remaining 202 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2452  0.8557  0.9502  0.8921  0.9815  0.9990
```

5.2.2. Leverage

Vimos en la Observación 4.5, en la Sección 4.6 que los residuos no son homoscedásticos. Y además vimos que la varianza dependía del leverage de una observación, que también definimos en esa sección a partir de la matriz de proyección o "*hat matrix*" H. El leverage de la i-ésima observación será el elemento h_{ii} de la matriz de proyección, y en general será calculado por el software. En el caso de regresión múltiple, sin embargo, es mucho más importante asegurarse que no haya observaciones potencialmente influyentes, o si uno sospecha de algunas, estudiar cómo cambia el ajuste cuando esa observación es eliminada de la base de datos. Para la detección de observaciones potencialmente influyentes en regresión lineal simple, muchas veces basta mirar con cuidado el scatter plot de los datos. El problema que aparece aquí es que no podemos, en general, dibujar el scatter plot de los datos, por lo que tendremos que calcular el leverage de cada observación. El criterio para

Figura 61: Salida de `plot(ajusterob)` para los datos de `azucar`, cuyo ajuste figura en la Tabla 51.



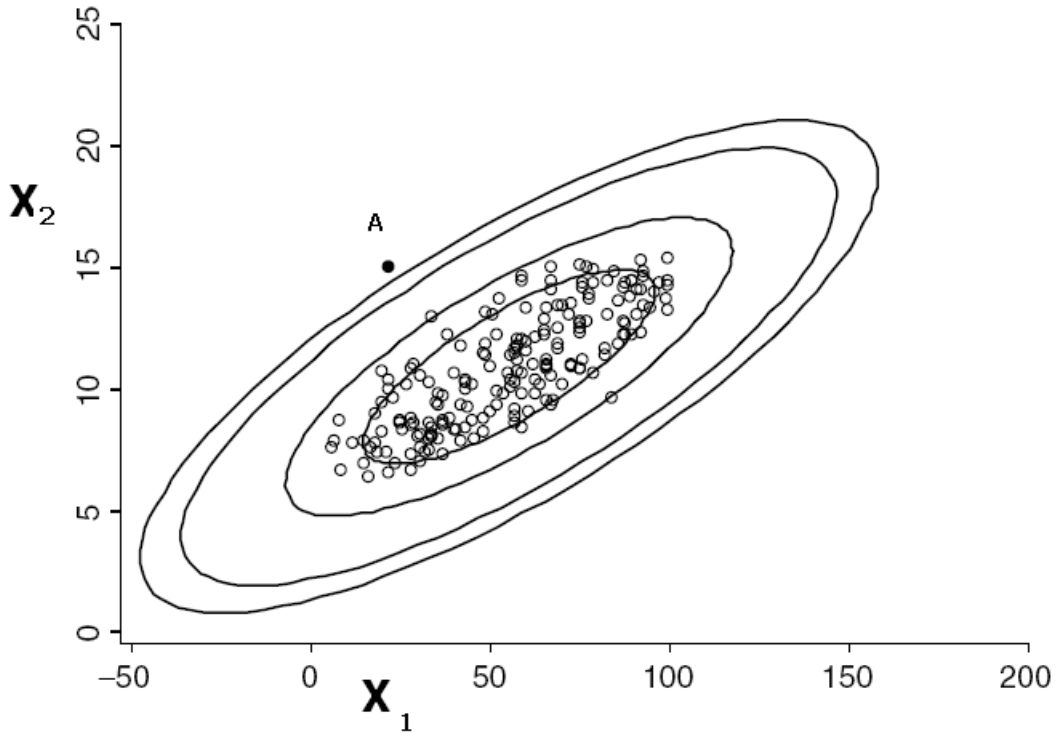
encontrar observaciones potencialmente influyentes será la extensión del visto anteriormente. El leverage alto indica que una observación no sigue el patrón de las demás covariables X. Nuevamente se tiene

$$0 \leq h_{ii} \leq 1 \qquad \sum_{i=1}^{n} h_{ii} = p$$

donde p es el número de parámetros de regresión (betas) que hay en la función de regresión, incluyendo el término de intercept. Puede mostrarse que h_{ii} es una medida de la distancia entre los valores de las covariables X de la i-ésima observación respecto del valor del promedio de todas las X observadas en los n casos. Es lo que se conoce como **distancia de Mahalanobis** de la i-ésima observación $\left(X_{i1}, X_{i2}, \ldots, X_{i(p-1)}\right)$ cuando se tiene una muestra de ellas. Este concepto se es-

tudia en detalle en los cursos de análisis multivariado. La idea subyacente es que las distancia usual no expresa bien las distancias entre observaciones cuando hay dependencia entre las covariables, entonces esta correlación o dependencia se toma en cuenta para definir una nueva noción de distancia entre puntos. En la Figura 62 se ve un gráfico de dispersión para un conjunto de observaciones, con curvas superpuestas. Estas curvas representan los puntos que tienen el mismo leverage. Vemos que son elipses. En el gráfico hay una observación alejada, indicada con A.

Figura 62: Contornos de leverage constante en dos dimensiones. Las elipses más pequeñas representan un menor leverage. Vemos una observación identificada con el nombre A que tiene alto leverage y no sigue el patrón de las restantes. Fuente: Weisberg [2005], pág. 170



Los dos criterios para evaluar si una observación tiene alta palanca presentados en el caso de regresión lineal simple se extienden sin grandes modificaciones al caso múltiple. Ellos son

1. *(Ajustado por la cantidad de covariables)* Declarar a la observación i-ésima con alto leverage si $h_{ii} > 2\overline{h} = \frac{2}{n}\sum_{j=1}^{n} h_{jj} = \frac{2p}{n}$.

2. *(Sin ajustar por la cantidad de covariables)* Declarar a la observación i-ésima con muy alto leverage si $h_{ii} > 0{,}5$ y con leverage moderado si $0{,}2 < h_{ii} \leq 0{,}5$.

Una evidencia adicional para declarar que una cierta observación tiene un leverage notoriamente alto, consiste en graficar un histograma de los h_{ii} y ver si existe una brecha destacable que separa al mayor leverage o a un pequeño conjunto de mayores leverages del resto de las observaciones.

5.2.3. Uso de la matriz de proyección para identificar extrapolaciones

La matriz H de proyección también es útil para determinar si una inferencia respecto de la respuesta media o de la predicción para una nueva observación X_{nueva} de valores de las predictoras involucra una extrapolación sustancial respecto del rango de los valores observados. Cuando sólo tenemos dos predictoras X_1 y X_2 esto puede resolverse con un scatter plot como muestra la Figura 62. Este sencillo análisis gráfico no se encuentra disponible si $p \geq 3$, donde las extrapolaciones pueden ocultarse.
Para detectarlas podemos utilizar los cálculos de leverage presentados anteriormente. Para una nueva combinación de variables

$$X_{\text{nue}} = (X_{1\text{nue}}, \ldots, X_{p-1\ \text{nue}})$$

para la que interesa hacer predicción se puede calcular

$$h_{\text{nue}} = X_{\text{nue}}^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} X_{\text{nue}}$$

donde la matriz $\mathbf{X}$ tiene dimensión $n \times p$ y se armó en base a la muestra con la que se calculó el modelo ajustado, ver (46), en la página 118. Si h_{nue} está bien incluida dentro del rango de leverages observados en el conjunto de datos disponibles, estamos seguros de que no hay extrapolación involucrada. Si, por el contrario, h_{nue} es mucho mayor que los leverages observados, entonces no debería llevarse a cabo la estimación o predicción para esta combinación X_{nue} de covariables.

5.2.4. Residuos estudentizados y distancias de Cook

Ambos estadísticos se definen y calculan del mismo modo que en regresión lineal simple. Las distancia de Cook para la i-ésima observación se define por

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{Y}_j - \widehat{Y}_{j(i)}\right)^2}{pMS\text{Res}}$$

donde $\widehat{Y}_j$ es el valor ajustado para la j-ésima observación, cuando se usaron las n observaciones en el ajuste del modelo, y $\widehat{Y}_{j(i)}$ es el valor ajustado para la j-ésima observación, cuando se usaron $n-1$ observaciones en el ajuste del modelo,

todas menos la i-ésima. Esto se repite para cada observación, para poder calcular todas las Distancias de Cook. Afortunadamente, las D_i pueden ser calculadas sin necesidad de ajustar una nueva función de regresión cada vez, en la que se deja una observación distinta afuera del conjunto de datos. Esto es porque puede probarse la siguiente igualdad que permite calcular las distancias de Cook

$$D_i = \frac{e_i^2}{pMS\text{Res}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right].$$

Observemos que las Distancias de Cook dependen de dos factores:

1. el tamaño del residuo i-ésimo, e_i
2. el leverage i-ésimo, h_{ii}.

Cuanto más grande sean e_i o h_{ii}, mayor será D_i. Luego el i-ésimo caso puede ser influyente por

1. tener un alto residuo e_i y sólo un moderado valor de leverage h_{ii},
2. o bien por tener un alto valor de leverage h_{ii} con sólo un moderado valor de residuo e_i,
3. o bien por tener tanto un alto valor de leverage h_{ii} como un alto valor de residuo e_i.

Los puntos de corte sugeridos para detectar una observación influyente con la Distancia de Cook suelen ser percentiles de la distribución F de Fisher con p grados de libertad en el numerador y $n-p$ en el denominador. Si la $D_i \geq F\,(p, n-p, 0{,}50)$ la observación i-ésima es considerada influyente.
El residuo estudentizado (o estudentizado eliminado) se define por

$$restud_i = \frac{Y_i - \widehat{Y}_{i(i)}}{\frac{MS\text{Res}_{(i)}}{1-h_{ii}}},$$

donde $\widehat{Y}_{i(i)}$ es el valor ajustado para la i-ésima observación, cuando se usaron $n-1$ observaciones en el ajuste del modelo, todas menos la i-ésima y $MS\text{Res}_{(i)}$ es el cuadrado medio de los residuos cuando el caso i-ésimo es omitido en el ajuste de la regresión lineal. Nuevamente, no necesitamos ajustar las regresiones excluyendo los casos de a uno por vez, pues una expresión alternativa para el residuo estudentizado es

$$restud_i = e_i \left[\frac{n-p-1}{SS\text{Res}\,(1-h_{ii}) - e_i^2} \right]^{1/2}$$

Los puntos de corte sugeridos para detectar una observación influyente con el residuo estudentizado están dados por el criterio de Bonferroni y consiste en declarar influyente a una observación si

$$|restud_i| > t_{n-p-1,\ 1-\frac{\alpha}{2n}}.$$

5.3. Colinealidad de los predictores

5.3.1. Diagnóstico de multicolinealidad

Cuando las variables predictoras incluidas en el modelo están correlacionadas entre ellas, decimos que existe intercorrelación o multicolinealidad. Algunos de los problemas típicos que aparecen cuando las variables regresoras están fuertemente correlacionadas son:

1. Los coeficientes de regresión estimados se modifican sustancialmente cuando se agregan o se quitan variables del modelo.

2. Los errores estándares de los estimadores de los coeficientes aumentan espúreamente cuando se incluyen covariables muy correlacionadas en el modelo. Esto se denomina *inflar la varianza estimada de los estimadores.*

3. Los coeficientes pueden ser no significativos aún cuando exista una asociación verdadera entre la variable de respuesta y el conjunto de variables regresoras.

5.3.2. Diagnóstico informal

Las siguientes situaciones son indicativas de multicolinealidad severa:

1. Cambios importantes en los coeficientes estimados al agregar o quitar variables o al modificar levemente las observaciones.

2. Tests no significativos para los coeficientes asociados a variables que teóricamente son importantes predictores, aún cuando observamos que existe una relación estadística entre las predictoras y la respuesta. El modelo puede tener R^2 cercano a 1 y el test F para el modelo ser fuertemente significativo y los tests para los coeficientes pueden ser no significativos. Recordemos que el efecto de la multicolinealidad es inflar la varianza estimada y en consecuencia el estadístico t asociado a cada coeficiente beta será pequeño. Por lo tanto, cuando existe multicolinealidad es difícil evaluar los efectos parciales.

3. Coeficientes estimados con signo contrario al que se espera según consideraciones teóricas.

4. Coeficientes de correlación grandes para las predictoras tomadas de a pares.

Aunque este último diagnóstico parece ser el modo más simple de detectar multicolinealidad, adolece de un problema: al calcular los coeficientes de correlación de Pearson de todas las variables regresoras tomadas de a pares sólo estamos mirando los vínculos lineales entre dos covariables. El problema es que podría haber un vínculo lineal muy estrecho entre una colección de variables y otra variable en particular. Un enfoque más apropiado es hacer una regresión de cada variable regresora sobre las demás variables regresoras. Cuando el R^2 de alguna de estas regresiones sea cercano a 1, deberíamos preocuparnos por el efecto de la multicolinealidad.
Finalmente diremos que la interpretación de los coeficientes se vuelve dudosa cuando existe multicolinealidad. Recordemos que en regresión múltiple (aditiva) cada coeficiente representa el efecto de la variable regresora cuando todas las demás variables se mantienen constantes. Pero si dos variables regresoras, por ejemplo X_1 y X_2, están fuertemente correlacionadas tiene poco sentido pensar en el efecto de X_1 sobre Y cuando X_2 se mantiene constante.

5.3.3. Diagnóstico formal

Un método formal para detectar la presencia de multicolinealidad que está ampliamente difundido es el uso de los factores de inflación de la varianza, más conocidos como **Variance Inflation Factor (VIF).** Es un número que se calcula para cada covariable. El VIF de la $k-$ésima covariable se calcula del siguiente modo

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad 1 \leq k \leq p-1,$$

donde R_k^2 es el coeficiente de determinación múltiple cuando X_k es regresado en las $p-2$ restantes covariables X en el modelo.
El VIF_k es igual a uno si $R_k^2 = 0$, es decir si la $k-$ésima covariable no está correlacionada con las restantes covariables. Cuando $R_k^2 \neq 0$, el VIF_k es mayor a uno. Cuando R_k^2 está muy cerca de uno, el VIF_k se vuelve un número enorme.
Para un conjunto de datos, el mayor VIF observado se usa como medida de diagnóstico. Si el máximo VIF es mayor a 10, eso es señal de multicolinealidad. Otro criterio es que cuando el promedio de los VIF es considerablemente mayor a uno se está frente a problemas de multicolinealidad.

5.3.4. ¿Cómo tratar el problema de multicolinealidad?

El recurso más simple es elegir un subconjunto de las variables regresoras poco correlacionadas. Si detectamos dos variables muy correlacionadas ¿cómo decidir cuál omitir? En general, conviene omitir aquella que tenga:

- mayor número de datos faltantes,
- mayor error de medición o
- que sea menos satisfactoria en algún sentido.

Otra posibilidad es eliminar variables a través de procedimientos de selección automáticos (se presentarán más adelante).
Cuando varios predictores están altamente correlacionados y son indicadores de una característica común, uno puede construir un índice combinando estas covariables. Los índices de bienestar, como el IDH (índice de desarrollo humano), o el índice de inflación, construidos como promedios ponderados de variables que miden el bienestar en una cierta región o bien los precios asociados a una determinada canasta, son ejemplos clásicos de esta construcción. En aplicaciones en las que se miden varias covariables muy correlacionadas esta puede resultar una buena solución.
En modelos polinómicos o que contienen interacciones, una solución al problema de multicolinealidad es trabajar con los datos centrados para la o las variables predictoras que aparecen en más de un término del modelo. Esto es, no usar la variable X tal como fue medida, sino la diferencia entre el valor observado y el valor medio de X en la muestra.
Existen otros procedimientos para tratar multicolinealidad, tanto clásicos como modernos, que complementan el ajuste de regresión lineal múltiple que hemos descrito, potenciando su alcance. Podemos agruparlos en tres clases importantes de métodos.

- Selección de modelos (o selección de variables). Este enfoque implica la identificación de un subconjunto de predictores que creemos están relacionados con la respuesta. Una vez seleccionado el subconjunto de variables relevantes, ajustamos con mínimos cuadrados un modelo que explica la respuesta con el conjunto reducido de variables. Describimos con detalle este procedimiento en la Sección 5.4.

- Regularización o penalización. Este enfoque involucra ajustar un modelo que contiene a todos los predictores. Sin embargo, el método de ajuste fuerza a encojer a los estimadores, achicándolos en valor absoluto, en relación con los estimadores que se obtienen usando mínimos cuadrados. Este encojimiento (también conocido en la literatura matemática como regularización) tiene el efecto de reducir la varianza estimada. Dependiendo de qué regularización se realice, la estimación de algunos coeficientes terminará siendo exactamente cero. Por eso, los métodos de regularización o penalización también pueden llevar a cabo la selección de variables. Los estimadores de regularización

más utilizados son los estimadores ridge o los lasso, y una combinación de ambos que se denominan elastic net. El libro de James, Witten, Hastie, y Tibshirani [2013] constituye una fuente muy accesible y actualizada, que además comenta los comandos de `R` para apropiados para implementarlos. Un enfoque más técnico por los mismos autores es Friedman, Hastie, y Tibshirani [2008]. No nos ocuparemos de estos temas en el curso.

- Reducción de la dimensión. Este enfoque propone proyectar las $p-1$ variables predictoras en un espacio de dimensión M, con $M < p$. Esto se realiza calculando M combinaciones lineales, o proyecciones, diferentes de los predictores. Luego, estas M proyecciones se utilizan como nuevas covariables para ajustar un modelo de regresión lineal por mínimos cuadrados. Puede verse el libro de James et al. [2013] para una introducción al tema, que tampoco trataremos en estas notas. Los métodos comprendidos en esta categoría son *Principal Components Regression*, *Partial Least Squares*, *Fitted Principal Components*, *Projection Pursuit Regression*, entre otros.

5.4. Selección de modelos

Ya hemos observado que cuando tenemos K covariables disponibles y una variable a explicar Y, pueden, en principio, ajustarse 2^K modelos distintos. Decimos en principio, pues este total de modelos no incluye aquellos que tienen interacciones. En esta sección estamos pensando que si uno quiere evaluar ciertas interacciones, las debe incluir en esas K covariables iniciales. Lo mismo si uno quisiera evaluar algunas potencias de las covariables originales, o algunas transformaciones más complicadas de ellas. De este modo, cuando K es un número grande, la cantidad de modelos posibles crece exponencialmente, y evaluarlos uno por uno puede ser inmanejable. Por ejemplo, para $K = 8$, hay $2^8 = 256$ modelos posibles: hay un modelo sin covariables, 8 modelos de regresión lineal simple, cada uno con una sola covariable, $\binom{8}{2} = 28$ modelos con dos covariables $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_1, X_4\}$, $\{X_2, X_3\}$, etc. , $\binom{8}{3} = 56$ modelos con tres covariables, etcétera.
Lo que se denomina selección de modelos corresponde a la tarea de elegir el mejor modelo para nuestros datos.

5.4.1. Criterios para comparar modelos

Una vez que se tienen todas las variables, es de interés contar con un criterio numérico para resumir la bondad del ajuste que un modelo lineal con un cierto conjunto de covariables da a la variable dependiente observada. A partir de este criterio se podrán ranquear los modelos y elegir un conjunto de unos pocos buenos candidatos para estudiar luego en detalle.

A continuación presentamos algunos de los criterios más frecuentemente utilizados en regresión lineal para la selección de modelos. No son los únicos, pero sí los más difundidos. Cuando ajustamos un modelo con $p-1$ covariables, es decir, con p coeficientes $\beta's$ podemos tomar como criterio para evaluar el ajuste a:

- R_p^2 o $SS\text{Res}_p$: Un primer criterio para comparar modelos es mirar el R^2 obtenido con cada uno de ellos y elegir aquél con mayor R^2. Usamos el subíndice p para indicar la cantidad de parámetros $\beta's$ hay en el modelo (es decir, $p-1$ covariables). Como tenemos que

$$R_p^2 = 1 - \frac{SS\text{Res}_p}{SSTotal},$$

resulta que comparar modelos usando el criterio de elegir aquél cuyo R_p^2 sea lo más grande posible equivale a elegir aquel que tenga la menor suma de cuadrados de residuos $SS\text{Res}_p$ (ya que la suma de cuadrados total $SSTotal = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ no depende de las covariables del modelo ajustado y por eso permanece constante). Pero como ya observamos, el R^2 aumenta al aumentar $p-1$, el número de covariables, sean estas apropiadas para ajustar los datos o no. Es por eso que el criterio no es identificar el modelo con mayor R^2 (ese será siempre el modelo con todas las covariables disponibles) sino encontrar el punto a partir del cual no tiene sentido agregar más variables ya que estas no inciden en un aumento importante del R^2. Muchas veces esto sucede cuando se han incorporado unas pocas variables al modelo de regresión. Por supuesto, encontrar el punto donde este aumento se empieza a estancar es un asunto de criterio individual. Suele ser bastante informativo graficar el mejor R_p^2 en función de p y evaluar gráficamente cuándo el crecimiento en el R^2 es tan poco que no justifica la inclusión de la covariable adicional.

- $R_{a,p}^2$ o MSE_p : Como el R_p^2 no toma en cuenta el número de parámetros en el modelo de regresión, un criterio de decisión mucho más objetivo y automatizable es calcular y comparar modelos por medio del R_a^2. Lo subindicaremos como $R_{a,p}^2$ para indicar la cantidad de coeficientes $\beta's$ presentes en el modelo. Recordemos que

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SS\text{Res}_p}{SSTotal} = 1 - \frac{MS\text{Res}_p}{\frac{SSTotal}{n-1}}.$$

Como $\frac{SSTotal}{n-1}$ está fijo en un conjunto de datos dado (sólo depende de las Y observadas), el $R_{a,p}^2$ aumenta si y sólo si el $MS\text{Res}_p$ disminuye. Luego, el coeficiente de determinación múltiple ajustado $R_{a,p}^2$ y el cuadrado medio del error $MS\text{Res}_p$, proveen información equivalente acerca del ajuste obtenido.

Al usar este criterio buscamos el subconjunto de $p-1$ covariables que maximicen el $R^2_{a,p}$, o un subconjunto de muchas menos covariables para las cuales $R^2_{a,p}$ esté muy cerca del máx $R^2_{a,p}$, en el sentido que el aumento en el R^2_a sea tan pequeño que no justifique la inclusión de la o las covariables extra.

- *C_p de Mallows*: Para utilizar esta medida hay que asumir que en el modelo con el total de las K covariables (el más grande posible) están todas las covariables importantes de modo que en ese modelo completo, la estimación de la varianza del error, σ^2, es insesgada. El valor del C_p se define por

$$C_p = \frac{SS\text{Res}_p}{MS\text{Res}\,(X_1, \ldots, X_K)} - (n - 2p)$$

donde $SS\text{Res}_p$ es la suma de los cuadrados de los errores del modelo con p parámetros (es decir, con $p-1$ covariables) y $MS\text{Res}\,(X_1, \ldots, X_K)$ es el estimador de la varianza del error σ^2, calculado bajo el modelo con todas las posibles covariables $X_1, \ldots, X_K$. Cuando se usa el C_p como criterio, se busca aquel subconjunto de p covariables X que tengan un C_p pequeño, lo más cercano a p posible. Es fácil ver que para el modelo completo, $C_K = K$.

- *AIC_p*, o el *Criterio de Akaike* y *SBC_p* o el *Criterio Bayesiano de Schwartz*, son otros dos criterios que, al igual que el C_p de Mallows, penalizan a los modelos con muchas covariables. Se buscan los modelos que tienen valores pequeños de AIC_p o SBC_p, donde estas cantidades están dadas por

$$\begin{aligned} AIC_p &= n \ln (SS\text{Res}_p) - n \ln (n) + 2p \\ SBC_p &= n \ln (SS\text{Res}_p) - n \ln (n) + p \ln (n) \end{aligned}$$

Observemos que para ambas medidas, el primer sumando decrece al aumentar p. El segundo sumando está fijo (puesto que n lo está, para un conjunto de datos) y el tercer sumando crece al crecer p, es decir, el número de covariables. Ambas medidas representan una buena ponderación entre ajuste apropiado (es decir, $SS\text{Res}_p$ pequeña) y parsimonia del modelo (es decir, pocos parámetros a ajustar, o sea, p pequeño). El Criterio SBC_p también se llama *Criterio Bayesiano de Información* (*BIC*, por sus siglas en inglés).

5.4.2. ¿Cuál de estos criterios utilizar?

Todos estos criterios miden cualidades deseables en un modelo de regresión. Ocasionalmente, una única ecuación de regresión produce valores óptimos de los cuatro criterios simultáneamente, con lo que uno puede confiar que éste es el mejor modelo en términos de estos criterios.

Desafortunadamente esto raramente ocurre y diferentes instrumentos identifican diferentes modelos. Sin embargo, tomados en conjunto estos criterios permiten identificar un conjunto pequeño de modelos de regresión que pueden ser construidos a partir de las variables independientes relevadas. Conviene entonces estudiar estos pocos modelos más detalladamente, teniendo en cuenta los objetivos del estudio, nuestro conocimiento del área del que provienen los datos y la evaluación de los supuestos del análisis de regresión para realizar una selección criteriosa de cual es el "mejor" modelo.

5.4.3. Selección automática de modelos

Al inicio de la Sección 5.4 hemos visto que en el proceso de selección de modelos es necesario comparar un número muy grande de modelos entre sí. Para simplificar esta tarea, existen una variedad de procedimientos automáticos de selección de modelos, programados en los paquetes estadísticos.
Un gran problema de estas búsquedas automáticas es que en general, están programadas para trabajar con la base completa de $n \times K$ observaciones. Si hubiera una observación faltante (*missing data*) (es decir, un caso para el cual no se registró **una** de las variables) estos algoritmos remueven el caso **completo** y hacen la selección de modelos basados en $n - 1$ observaciones. Esto puede volverse un problema si n es pequeño y hay varias variables con observaciones faltantes.
Los métodos más populares de selección de variables son:

1. Todos los subconjuntos posibles (*Best subset*).
2. Eliminación *backward* (hacia atrás).
3. Selección *forward* (incorporando variables).
4. *Stepwise regression* (regresión de a pasos).

A continuación los describimos. Asumimos que $n > K$ (o sea, que tenemos más observaciones que covariables).

5.4.4. Todos los subconjuntos posibles (*Best subset*)

Estos algoritmos ajustan **todos** los submodelos posibles (los 2^K) y luego los ranquean de acuerdo a algún criterio de bondad de ajuste. Por supuesto, esto involucra hacer 2^K regresiones. Siempre que sea posible es aconsejable usar este procedimiento ya que es el único método que garantiza que se obtendrá el modelo final que realmente optimice la búsqueda con el criterio elegido: por ejemplo mayor R^2_a, o mejor C_p, etc. Es decir, garantiza que el modelo final es el "mejor" para el presente conjunto de datos y para los criterios utilizados.

Una vez que todos los modelos han sido ajustados, en general el paquete exhibe los 10 (o una cantidad prefijable) mejores modelos de acuerdo al criterio elegido, entre todos los que tienen el mismo número de variables.
Cuando la cantidad original de potenciales covariables es muy grande, K mayor a 40, por ejemplo, no es posible ajustar todos los modelos posibles ya que $2^{40} = 1\,099\,511\,627\,776$. Se vuelve necesario usar otro tipo de procedimientos, computacionalmente más realizables, que buscan elegir un modelo luego de una búsqueda que explora una sucesión de modelos de regresión que en cada paso agrega o quita una covariable X. El criterio para agregar o quitar una covariable, en el caso secuencial, puede escribirse equivalentemente en términos de la suma de los cuadrados de los residuos, los estadísticos F parciales, el estadístico t asociado a un coeficiente, o el R_a^2. Son los tres procedimientos que describimos a continuación.

5.4.5. Eliminación *backward* (hacia atrás).

El procedimiento comienza construyendo el modelo con todas las predictoras y en cada paso se elimina una variable. La secuencia del procedimiento es la siguiente. Se define un nivel de significación fijo α.

1. El modelo inicial contiene todos los potenciales predictores (que hemos denominado K).

2. Si todas las variables producen una contribución parcial significativa (es decir, un estadístico t con p−valor $< \alpha$) entonces el modelo completo es el modelo final.

3. De otro modo, se elimina la variable que tenga la menor contribución parcial (es decir, el mayor p−valor de su estadístico t) cuando todas las demás están en el modelo.

4. Se ajusta el nuevo modelo con $(K-1)$ predictores y se repiten los pasos 2 y 3 hasta que todas las variables en el modelo tengan un coeficiente estimado cuyo p−valor asociado al estadístico t sea menor a α.

Si hay una alta multicolinealidad en el conjunto de los K predictores, este procedimiento no es muy recomendable.

5.4.6. Selección *forward* (incorporando variables)

En este caso, comenzamos con el modelo sin variables y vamos agregando las variables de a una por vez. Ingresa la variable que más contribuye a explicar a Y cuando las otras ya están en el modelo. Se elige un nivel de significación fijo α. La secuencia de pasos es la siguiente:

1. Primero se ajustan todos los modelos de regresión lineal simple con Y como respuesta y una sola covariable explicativa. Se elige la que tiene el mayor valor del estadístico F o, equivalentemente, el menor p−valor del estadístico t asociado al coeficiente, siempre que dicho p−valor sea inferior a α, sino el procedimiento termina y se elige el modelo sin covariables

2. En el segundo paso, se busca elegir entre todos los modelos de dos covariables que tienen a la que fue seleccionada en el primer paso aquél para el cuál el test F parcial dé mas significativo. El test F parcial es el que compara el ajuste del modelo con dos variables con el ajuste del modelo con una variable elegido en el primer paso. Es decir, es el test que mide la significatividad de la segunda variable a ser incorporada en el modelo cuando la primera ya está en él. Para aquel modelo que tenga el F parcial más significativo o, equivalentemente, el test t asociado al coeficiente de la variable a ser incorporada más significativo, o sea, el menor p−valor, se compara a dicho p−valor con el valor crítico α. Si el p−valor es menor que α se elige dicho modelo, si el p−valor supera el valor crítico, el procedimiento se detiene, y el output del proceso es el modelo que tiene una única covariable significativa, que fue seleccionada en el paso 1.

3. Ahora se calculan los estadísticos F parciales de todos los modelos con tres covariables, que tienen a las dos covariables ya elegidas e incorporan una tercera. Se continua de esta manera (como en el paso 2) hasta que ninguna variable produce un F parcial (o t) significativo.

Si se usa un punto de corte muy exigente (digamos $\alpha < 0{,}01$) serán incluídas menos variables y existe la posibilidad de perder covariables importantes. Si se usa un punto de corte menos exigente ($\alpha < 0{,}20$) es menos probable que se pierdan covariables explicativas importantes pero el modelo contendrá más variables.
Una vez que el procedimiento finaliza, no todas las variables en el modelo necesariamente tendrán coeficientes parciales significativos.

5.4.7. Selección stepwise

Es una modificación del procedimiento forward que elimina una variable en el modelo si ésta pierde significación cuando se agregan otras variables. La aproximación es la misma que la selección forward excepto que a cada paso, después de incorporar una variable, el procedimiento elimina del modelo las variables que ya no tienen contribución parcial significativa. Una variable que entró en el modelo en una etapa, puede eventualmente, ser eliminada en un paso posterior.
En este caso será necesario definir un punto de corte para que ingrese una variable α_I y otro para eliminarla del modelo α_E. Uno puede desear ser menos exigente

(mayor p−valor) en el punto de corte para que una variable salga del modelo una vez que ingresó, o usar el mismo valor para ambos.
Este procedimiento, en general produce modelos con menos variables que la selección forward.

5.4.8. Limitaciones y abusos de los procedimientos automáticos de selección de variables

Cualquier método automático de selección de variables debe ser usado con precaución y no debería ser sustituto de un investigador que piensa, ya que no hay garantías que el modelo final elegido sea "óptimo". Conviene tener en cuenta las siguientes observaciones.

- Cuando se proponen términos de interacción entre las variables regresoras, el modelo debe contener las interacciones significativas y los efectos principales de las variables involucradas en estas interacciones, sean éstas significativas o no. De otro modo el modelo carece de interpretación. La mayoría de los procedimientos automáticos no tienen este cuidado. Lo mismo sucede cuando uno incorpora una variable categórica codificada con dummies: o entran todas las dummies, o ninguna, pero no es correcto poner algunas de ellas (las significativas) y otras no, porque sino el modelo carece de interpretación. Con las categóricas, otra posibilidad consiste en recategorizarlas, agrupando algunas categorías, y luego ajustar el modelo nuevamente, esperando que se obtenga un mejor ajuste (no siempre ocurre).

- El hecho de que un modelo sea el mejor en términos de algún criterio (C_p o R^2_a, por ejemplo) no significa que sea el mejor desde el punto de vista práctico. Ni tampoco que para este modelo valgan los supuestos.

- El procedimiento de selección automática puede excluir del modelo variables que realmente deberían estar en el modelo de acuerdo a otros criterios teóricos.Una posibilidad es forzar a que ciertas variables aparezcan en el modelo, independientemente del hecho de que tengan coeficientes significativos. Por ejemplo, podemos hacer una regresión backward sujeta a la restricción de que el modelo incluya ciertos términos especificados de antemano. Esto asegura que el modelo final contiene las variables de interés primario y toda otra variable o interacción que sea útil a los efectos de predicción. Algunos paquetes permiten esta alternativa.

- Una vez que hemos seleccionado un modelo final usando cualquier procedimiento de selección, la inferencia realizada sobre ese modelo es **sólo aproximada**. En particular, los p−valores serán menores y los intervalos de confianza más angostos que lo que deberían ser, puesto que el modelo seleccionado

es aquél que más fuertemente refleja los datos. (Hemos hecho uso y abuso de nuestros datos para obtener un modelo, es de esperar que otra muestra aleatoria de observaciones del mismo tipo a la que se le ajuste este modelo tenga menor capacidad predictiva).

- Existe una diferencia sustancial entre selección de modelos explicativos y exploratorios. En investigación explicativa, uno tiene un modelo teórico y pretende testearlo a través de un análisis de regresión. Uno podría querer testear si una relación que se propone como espúrea desaparece al incorporar una nueva variable en el modelo. En este enfoque, los procedimientos de selección automática en general no son apropiados, ya que es la teoría la que determina cuáles son las variables que deben estar en el modelo.

- En investigación exploratoria el objetivo es encontrar un buen conjunto de predictores. Uno intenta maximizar R^2 independientemente de explicaciones teóricas.

- ¿Por qué podría dejarse en el modelo final una variable que no resulta estadísticamente significativa? Muchas veces pueden aparecer variables en el modelo seleccionado para las cuáles el p−valor del test t no es menor que 0,05. Esto puede deberse a que haya motivos teóricos que indican que la respuesta depende de dicha covariable y que tal vez el tamaño de muestra no haya sido lo suficientemente grande como para comprobarse la significatividad estadística. Se deja para que el modelo no resulte sesgado. Los estimadores de los coeficientes son insesgados si el modelo es correcto (es decir, contiene todas las covariables apropiadas en la forma correcta, dejar covariables con sustento teórico para que estén permite que los estimadores de los efectos de otras covariables sean insesgados). Otro motivo para dejarla puede ser porque su presencia ayuda a reducir la varianza estimada del error, permitiendo que otros coeficientes resulten significativos. Y también pueden dejarse covariables aunque no sean significativas pero que permitan comparar el modelo presentado con otros modelos publicados con antelación.

En resumen, los procedimientos de selección automática de modelos no son sustitutos de una cuidadosa construcción teórica que guíe la formulación de los modelos.

5.4.9. Validación de modelos

El paso final en el proceso de construcción o selección de modelos lo constituye el proceso de validación de los modelos. Esta etapa de validación involucra, usualmente, chequear el modelo candidato con datos independientes a los utilizados para proponer el modelo. Hay cuatro formas básicas de validar un modelo de regresión:

1. Recolectar un nuevo conjunto de datos que permita chequear el modelo y su habilidad predictiva.

2. Comparar los resultados con las expectativas teóricas, resultados empíricos previos y resultados de simulaciones.

3. Cuando fuera posible, usar otras técnicas experimentales para confirmar el modelo. Esto, por supuesto, dependerá de las herramientas propias de cada disciplina.

4. Cuando el tamaño de muestra lo permitiera, otra posibilidad es dividir al conjunto de observaciones disponible en dos grupos disjuntos. Con uno de ellos se selecciona el modelo más apropiado. Este grupo se denomina *muestra de entrenamiento* (*training sample*). Con el segundo grupo, que se llama *muestra de validación* (*validation set*) se evalúa la razonabilidad y la capacidad predictiva del modelo seleccionado. A este proceso de validación se lo denomina a veces, *cross-validation*, es decir, validación cruzada.