

4. Regresión Lineal Múltiple

El modelo de regresión lineal múltiple es uno de los modelos más utilizados entre todos los modelos estadísticos.

En la mayoría de las situaciones prácticas en las que se quiere explicar una variable continua Y se dispone de muchas potenciales variables predictoras. Usualmente, el modelo de regresión lineal simple (es decir, con una sola variable predictora) provee una descripción inadecuada de la respuesta ya que suele suceder que son muchas las variables que ayudan a explicar la respuesta y la afectan de formas distintas e importantes. Más aún, en general estos modelos suelen ser muy imprecisos como para ser útiles (tienen mucha variabilidad). Entonces es necesario trabajar con modelos más complejos, que contengan variables predictoras adicionales, para proporcionar predicciones más precisas y colaborar en la cuantificación del vínculo entre ellas. En este sentido, el modelo de regresión múltiple es una extensión natural del modelo de regresión lineal simple, aunque presenta características propias que es de interés estudiar en detalle.

El modelo de regresión múltiple se puede utilizar tanto para datos observacionales como para estudios controlados a partir de ensayos aleatorizados o experimentales.

4.1. El modelo

La regresión múltiple es un modelo para la esperanza de una variable continua Y cuando se conocen variables explicativas o predictoras que denotaremos X_1, X_2, \dots, X_{p-1} . Antes de formularlo en general, describiremos a modo ilustrativo la situación en la que se tienen dos variables predictoras (i.e. $p = 3$). En este caso, proponemos el siguiente modelo para la esperanza condicional de Y dado X_1 y X_2

$$E(Y | X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (41)$$

donde $\beta_0, \beta_1, \beta_2$ son constantes desconocidas que se denominan *parámetros* del modelo, o *coeficientes* de la ecuación. Muchas veces, por simplicidad, escribiremos $E(Y)$ en vez de $E(Y | X_1, X_2)$. El modelo se denomina “lineal” puesto que la esperanza de Y condicional a las X ’s depende linealmente de las covariables X_1 y X_2 . Los coeficientes del modelo se estiman a partir de una muestra aleatoria de n observaciones (X_{i1}, X_{i2}, Y_i) con $1 \leq i \leq n$, donde Y_i es la variable respuesta medida en el i -ésimo individuo (o i -ésima repetición o i -ésima unidad experimental, según el caso), X_{i1} y X_{i2} son los valores de las variables predictoras en el i -ésimo individuo (o i -ésima repetición o i -ésima unidad experimental, según el caso). Una manera alternativa de escribir el modelo (41) en términos de las variables (en vez de sus valores esperados) es la siguiente

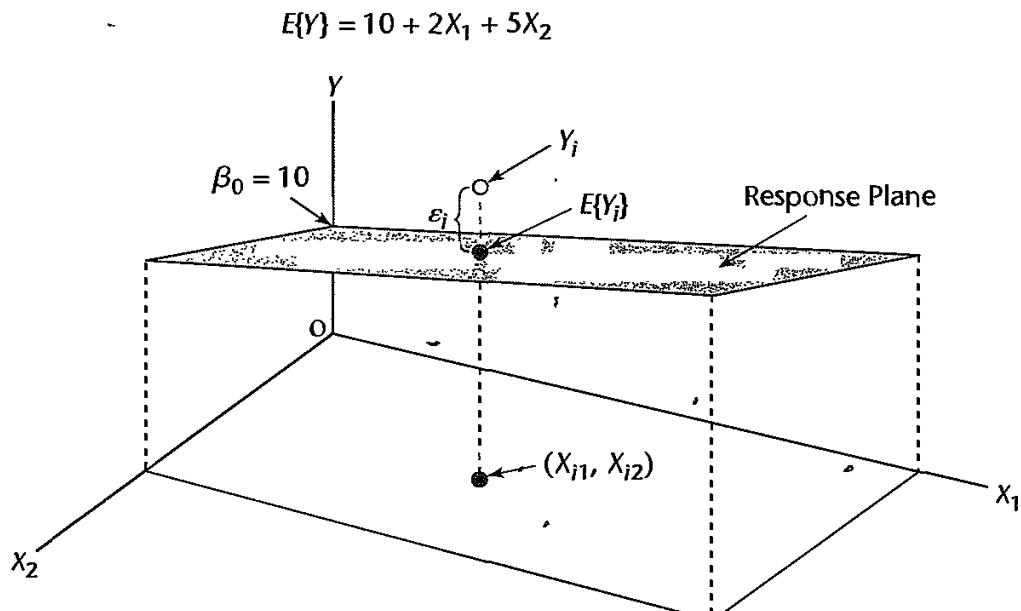
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad (42)$$

donde ε_i es el término del error para el individuo i -ésimo, que no es observable. A la ecuación (41) se la suele llamar función de respuesta. En analogía con la regresión lineal simple donde la función $E(Y | X) = \beta_0 + \beta_1 X_1$ es una recta, la función de regresión (41) es un plano. En la Figura 36 se representa una porción de la función de respuesta

$$E(Y | X_1, X_2) = 10 + 2X_1 + 5X_2. \quad (43)$$

Por supuesto, la única situación en la que podemos graficar es cuando $p \leq 3$ (dos o menos variables explicativas), es por eso que hemos comenzado con este caso.

Figura 36: En regresión lineal con dos variables explicativas la función de respuesta es un plano. Fuente Kutner et al. [2005], pág. 215.



Observemos que cualquier punto de la Figura 36 corresponde a una respuesta media $E(Y)$ para una combinación dada de X_1 y X_2 . La Figura 36 también muestra una observación Y_i correspondiente a los niveles (X_{i1}, X_{i2}) de las dos variables predictoras. El segmento vertical entre Y_i y el gráfico de la función (el plano) de respuesta representa la diferencia entre Y_i y la media $E(Y_i) = E(Y_i | X_{i1}, X_{i2})$ de la distribución de probabilidad de Y para la combinación de (X_{i1}, X_{i2}) . Por lo tanto, la distancia vertical entre Y_i y el plano de respuesta representa el término

de error $\varepsilon_i = Y_i - E(Y_i)$. En regresión lineal múltiple, a la función de respuesta también suele llamársela *superficie de regresión* o *superficie de respuesta*.

4.2. Significado de los coeficientes de regresión

Consideremos ahora el significado de los coeficientes en la función de regresión múltiple (42). El parámetro β_0 es el intercepto ordenada al origen del plano. Si dentro de los valores que estamos ajustando el modelo, se encuentra incluido el punto $X_1 = 0, X_2 = 0$, el origen de coordenadas, entonces β_0 representa la respuesta media $E(Y)$ en $X_1 = 0, X_2 = 0$. De lo contrario, β_0 no tiene ningún significado en particular como un término separado del modelo de regresión.

El parámetro β_1 indica el cambio en la respuesta media $E(Y)$ cuando aumentamos a X_1 en una unidad, manteniendo a X_2 constante (en cualquier valor). Del mismo modo, β_2 indica el cambio en la respuesta media $E(Y)$ cuando aumentamos a X_2 en una unidad, manteniendo a X_1 constante. En el ejemplo (43) graficado, supongamos que fijamos X_2 en el nivel $X_2 = 3$. La función de regresión (43) ahora es la siguiente:

$$E(Y) = 10 + 2X_1 + 5(3) = 25 + 2X_1, \quad X_2 = 3.$$

Notemos que esta función de respuesta es una línea recta con pendiente $\beta_1 = 2$. Lo mismo es cierto para **cualquier otro valor de X_2** ; sólo el intercepto de la función de respuesta será diferente. Por lo tanto, $\beta_1 = 2$ indica que la respuesta media $E(Y)$ aumenta en 2 unidades, cuando se produce un incremento unitario en X_1 , cuando X_2 se mantiene constante, sin importar el nivel de X_2 .

Del mismo modo, $\beta_1 = 5$, en la función de regresión (43) indica que la respuesta media $E(Y)$ se incrementa en 5 unidades, cuando se produce un incremento unitario en X_2 , siempre que X_1 se mantenga constante.

Cuando el efecto de X_1 en la respuesta media no depende del nivel de X_2 , y además el efecto de X_2 no depende del nivel de X_1 , se dice que las dos variables predictoras tienen *efectos aditivos* o *no interactúan*. Por lo tanto, el modelo de regresión tal como está propuesto en (41) está diseñado para las variables predictoras cuyos efectos sobre la respuesta media son aditivos.

Los parámetros β_1 y β_2 a veces se llaman *coeficientes de regresión parcial* porque reflejan el efecto parcial de una variable de predicción cuando la otra variable predictora es incluida en el modelo y se mantiene constante.

Observación 4.1 *El modelo de regresión para el que la superficie de respuesta es un plano puede ser utilizado tanto porque se crea que modela la verdadera relación entre las variables, o como una aproximación a una superficie de respuesta más compleja. Muchas superficies de respuesta complejas se pueden aproximar razonablemente bien por un plano.*

blemente bien por planos para valores limitados (o acotados) de las covariables X_1 y X_2 .

4.3. Modelo de Regresión Lineal Múltiple

El modelo de regresión lineal múltiple es un modelo para la variable aleatoria Y cuando se conocen X_1, X_2, \dots, X_{p-1} las variables regresoras. El modelo es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i, \quad (44)$$

donde $\beta_0, \beta_1, \dots, \beta_{p-1}$ son parámetros (es decir, números) desconocidos, $X_{i1}, X_{i2}, \dots, X_{ip-1}$ son los valores de las variables predictoras medidas en el i -ésimo individuo (o i -ésima repetición del experimento o i -ésima unidad experimental, según el caso) con $1 \leq i \leq n$, n es el tamaño de muestra, Y_i es la variable respuesta medida en el i -ésimo individuo (observado) y ε_i es el error para el individuo i -ésimo, que no es observable. Haremos supuestos sobre ellos:

$$\varepsilon_i \sim N(0, \sigma^2), \quad 1 \leq i \leq n, \quad \text{independientes entre sí.} \quad (45)$$

Es decir,

- los ε_i tienen media cero, $E(\varepsilon_i) = 0$.
- los ε_i tienen todos la misma varianza desconocida que llamaremos σ^2 y que es el otro parámetro del modelo, $Var(\varepsilon_i) = \sigma^2$.
- los ε_i tienen distribución normal.
- los ε_i son independientes entre sí, e independientes de las covariables $X_{i1}, X_{i2}, \dots, X_{ip-1}$.

Si definimos $X_{i0} = 1$ para todo i , podemos escribir a (44) de la siguiente forma equivalente

$$\begin{aligned} Y_i &= \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i \\ &= \sum_{j=0}^{p-1} \beta_j X_{ij} + \varepsilon_i \end{aligned}$$

Observemos que del hecho de que los ε_i son independientes y tienen distribución $N(0, \sigma^2)$ y de (44) se deduce que, condicional a X_1, \dots, X_{p-1} , $Y_i \sim$

$N\left(\sum_{j=0}^{p-1} \beta_j X_{ij}, \sigma^2\right)$ independientes entre sí. Tomando esperanza (condicional) en (44) obtenemos

$$E(Y | X_1, \dots, X_{p-1}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1},$$

que es una manera alternativa de escribir el modelo (44). Las variables predictoras pueden ser acomodadas para contemplar una serie de situaciones cuyo tratamiento iremos desarrollando a lo largo del curso. Esencialmente pueden ser

- variables continuas, y todas distintas. En la Sección 4.7 veremos un ejemplo de dos continuas.
- variables categóricas o cualitativas, en la Sección 4.13 veremos varios ejemplos donde aparecerán categóricas de dos categorías, que se suelen denominar binarias o dicotómicas o dummies, o de más de dos categorías.
- variables continuas, algunas representando potencias de otras. A esta situación se le suele llamar regresión polinomial.
- variables continuas, pero aparecen en el modelo transformaciones de las originales.
- variables modelando efectos de interacción entre dos o más variables, continuas o categóricas (ver Secciones 4.16 y 4.18).
- combinaciones de algunos o de todos los casos anteriores.

Observación 4.2 *Como ya dijimos en el caso $p = 3$, el término **lineal** en modelo lineal se refiere al hecho de que el modelo (44) es lineal tanto en los parámetros $\beta_0, \dots, \beta_{p-1}$ como en las covariables X_1, \dots, X_{p-1} que no tienen porqué ser las variables originalmente observadas para cada individuo o para cada repetición del experimento, pudiendo ser una transformación o recodificación o combinación de ellas. En este sentido, el modelo*

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

*se estudia y ajusta como un modelo de regresión lineal en dos variables: X_i y X_i^2 , (aunque matemáticamente se trate de una función cuadrática en una sola variable). Un ejemplo de modelo **no lineal** es el siguiente*

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i$$

puesto que no puede expresarse de la forma (44). Varios libros tratan el tema de regresión no lineal, por ejemplo Kutner et al. [2005], parte III.

4.4. Modelo de Regresión Lineal en notación matricial

Ahora presentaremos el modelo (44) en notación matricial. Es una notable propiedad del álgebra de matrices el hecho de que tanto la presentación del modelo como los resultados del ajuste del modelo de regresión lineal múltiple (44) escrito en forma matricial tienen el mismo aspecto (la misma forma) que los que ya vimos para regresión lineal simple. Sólo cambian algunos grados de libertad y algunas constantes.

Enfatizamos en la notación matricial puesto que éste es el tratamiento estándar del tema, y además porque refleja los conceptos esenciales en el ajuste del modelo. Nosotros no calcularemos nada, las cuentas las hace la computadora.

Para expresar el modelo (44) de forma matricial definimos las siguientes matrices

$$\begin{aligned} \mathbf{Y}_{n \times 1} &= \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} & \mathbf{X}_{n \times p} &= \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \\ \boldsymbol{\beta}_{p \times 1} &= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} & \boldsymbol{\varepsilon}_{n \times 1} &= \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \end{aligned} \quad (46)$$

Observemos que los vectores \mathbf{Y} y $\boldsymbol{\varepsilon}$ son los mismos que para la regresión lineal simple. El vector $\boldsymbol{\beta}$ contiene los parámetros de regresión adicionales. Cada fila de la matriz \mathbf{X} corresponde a las observaciones correspondientes a cada individuo (la fila i -ésima contiene las observaciones del individuo i -ésimo) y las columnas identifican a las variables.

El modelo (44) se escribe matricialmente en la siguiente forma

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

donde

\mathbf{Y} es un vector de respuestas

$\boldsymbol{\beta}$ es un vector de parámetros

\mathbf{X} es una matriz de covariables

$\boldsymbol{\varepsilon}$ es un vector de variables aleatorias normales independientes con esperanza $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ y matriz de varianzas y covarianzas

$$Var(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}.$$

Entonces tomando a las variables equis como fijas, o, lo que es lo mismo, condicional a las variables equis, la esperanza de \mathbf{Y} resulta ser

$$E(\mathbf{Y} | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

y la matriz de covarianza de las \mathbf{Y} resulta ser la misma que la de $\boldsymbol{\varepsilon}$

$$Var(\mathbf{Y} | \mathbf{X}) = \sigma^2 \mathbf{I}.$$

Al igual que hicimos con el modelo de regresión simple, muchas veces omitiremos la condicionalidad a las equis en la notación, es decir, como es bastante habitual en la literatura, escribiremos $E(\mathbf{Y})$ en vez de $E(\mathbf{Y} | \mathbf{X})$.

4.5. Estimación de los Parámetros (Ajuste del modelo)

Usamos el método de mínimos cuadrados para ajustar el modelo. O sea, definimos la siguiente función

$$g(b_0, b_1, \dots, b_{p-1}) = \sum_{i=1}^n (Y_i - b_0 X_{i0} - b_1 X_{i1} - b_2 X_{i2} - \dots - b_{p-1} X_{ip-1})^2 \quad (47)$$

y los estimadores $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$ serán aquellos valores de b_0, b_1, \dots, b_{p-1} que minimicen a g . Los denominamos estimadores de mínimos cuadrados. Denotaremos al vector de coeficientes estimados por $\hat{\boldsymbol{\beta}}$.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}_{p \times 1}$$

Las ecuaciones de mínimos cuadrados normales para el modelo de regresión lineal general son

$$\mathbf{X}^t \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^t \mathbf{Y}$$

donde \mathbf{X}^t quiere decir la matriz traspuesta. Algunos autores lo notan \mathbf{X}' (recordemos que la matriz traspuesta es aquella matriz $p \times n$ que tiene por filas a las columnas de \mathbf{X}). Los estimadores de mínimos cuadrados son

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

Observación 4.3 *Para encontrar los estimadores de $\boldsymbol{\beta}$ no se necesita que los errores sean normales.*

Observación 4.4 En el caso de la regresión lineal, los estimadores de mínimos cuadrados de los betas coinciden también con los estimadores de máxima verosimilitud para el modelo antes descripto, es decir, cuando se asume normalidad de los errores.

4.6. Valores Ajustados y Residuos

Denotemos al vector de valores ajustados (*fitted values*, en inglés) \hat{Y}_i por $\hat{\mathbf{Y}}$ y al vector de residuos $e_i = Y_i - \hat{Y}_i$ lo denotamos por \mathbf{e}

$$\hat{\mathbf{Y}}_{n \times 1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \quad \mathbf{e}_{n \times 1} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Los valores ajustados se calculan del siguiente modo

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

que son los valores que están en la *superficie de respuesta ajustada* (o sea, en el plano ajustado en el caso $p = 3$). Los residuos se escriben matricialmente como

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)\mathbf{Y} \end{aligned}$$

Llamando

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \in \mathbb{R}^{n \times n} \tag{48}$$

a la “*hat matrix*” (la matriz que “sombraea”) tenemos que

$$\hat{\mathbf{Y}} = \mathbf{HY}$$

y

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

La matriz de varianzas de los residuos es

$$Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}). \tag{49}$$

Observación 4.5 (residuos) *El modelo de regresión lineal impone que los errores ε_i sean independientes, normales y tengan todos la misma varianza. Como ya hemos dicho, los errores no son observables. Los residuos e_i , que son el correlato empírico de los errores, son observables. Sin embargo, los residuos no son independientes entre sí y sus varianzas no son iguales. Veámoslo.*

Por (49), la varianza de e_i es el elemento que ocupa el lugar ii de la matriz $\sigma^2(\mathbf{I} - \mathbf{H})$. Si la matriz \mathbf{H} fuera igual a cero (que no tendría sentido para el modelo de regresión lineal), todos los residuos tendrían la misma varianza σ^2 (igual que la varianza de los errores). Sin embargo esto no sucede. Calculemos el elemento que ocupa el lugar ii de la matriz $\sigma^2(\mathbf{I} - \mathbf{H})$.

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$$

donde h_{ii} representa el elemento que ocupa el lugar ii de la matriz \mathbf{H} . Pero sabemos que

$$\begin{aligned} h_{ij} &= \left(\mathbf{X} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \right)_{ij} = [\text{fila } i \text{ de } X] (\mathbf{X}^t \mathbf{X})^{-1} [\text{fila } j \text{ de } X]^t \\ &= \mathbf{x}_i^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_j \end{aligned}$$

donde \mathbf{x}_i^t representa la i -ésima fila de \mathbf{X} . Luego,

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii}) = \sigma^2 \left(1 - \mathbf{x}_i (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_i^t \right)$$

Como en el caso de regresión lineal simple, al elemento ii de la matriz \mathbf{H} , es decir, a h_{ii} , se lo denominará el **leverage o palanca de la observación i -ésima**. Esta cantidad servirá para detectar observaciones atípicas o potencialmente influyentes. Nos ocuparemos de esto en la Sección 5.2.

En cuanto a la independencia, los residuos no son independientes entre sí ya que la $\text{cov}(e_i, e_j)$ ocupa el lugar ij -ésimo de la matriz $\sigma^2(\mathbf{I} - \mathbf{H})$. Nuevamente, si la matriz \mathbf{H} fuera igual a cero (que no tendría sentido), entonces dichas covarianzas valdrían cero. Pero

$$\text{cov}(e_i, e_j) = \sigma^2(-h_{ij}) = \sigma^2 \left(-\mathbf{x}_i (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_j^t \right)$$

donde h_{ij} representa el elemento que ocupa el lugar ij de la matriz \mathbf{H} .

Observación 4.6 (teórica) \mathbf{H} , y por lo tanto $\mathbf{I} - \mathbf{H}$, son matrices de proyección (es decir que $\mathbf{H}^2 = \mathbf{H}$ y lo mismo ocurre con $\mathbf{I} - \mathbf{H}$). \mathbf{H} proyecta al subespacio de \mathbb{R}^n generado por las columnas de \mathbf{X} . Algunos textos la notan con la letra \mathbf{P} .

4.7. Dos predictoras continuas

Antes de seguir con las sumas de cuadrados, las estimaciones de los intervalos de confianza para los coeficientes y el test F, veamos un ejemplo numérico con $p = 3$. Consideremos los datos correspondientes a mediciones de 100 niños nacidos con bajo peso en Boston, Massachusetts presentados en el artículo de Leviton et al. [1991], tratados en el libro de Pagano et al. [2000]. Al estudiar el modelo de regresión lineal simple encontramos una relación lineal significativa entre el perímetro cefálico y la edad gestacional para la población de niños nacidos con bajo peso. La recta ajustada a esos datos era

$$\hat{Y} = 3,9143 + 0,7801X_1$$

Nos preguntamos ahora si el perímetro cefálico también dependerá del peso del niño al nacer. Veamos un scatter plot (gráfico de dispersión) del perímetro cefálico versus el peso al nacer, para los 100 niños. El scatter plot de la Figura 37 sugiere que el perímetro cefálico aumenta al aumentar el peso. Pero una vez que hayamos ajustado por la edad gestacional, ¿será que el conocimiento del peso al nacer mejorará nuestra habilidad para predecir el perímetro cefálico de un bebé?

Para responder a esta pregunta ajustamos un modelo de regresión lineal múltiple con dos variables predictoras. Sean

Y_i = perímetro cefálico del i-ésimo niño, en centímetros (**headcirc**)

X_{i1} = edad gestacional del i-ésimo niño, en semanas (**gestage**)

X_{i2} = peso al nacer del i-ésimo niño, en gramos (**birthwt**)

Proponemos el modelo (42), o lo que es lo mismo, el modelo (44) con $p = 3$, o sea dos covariables. Lo reescribimos

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

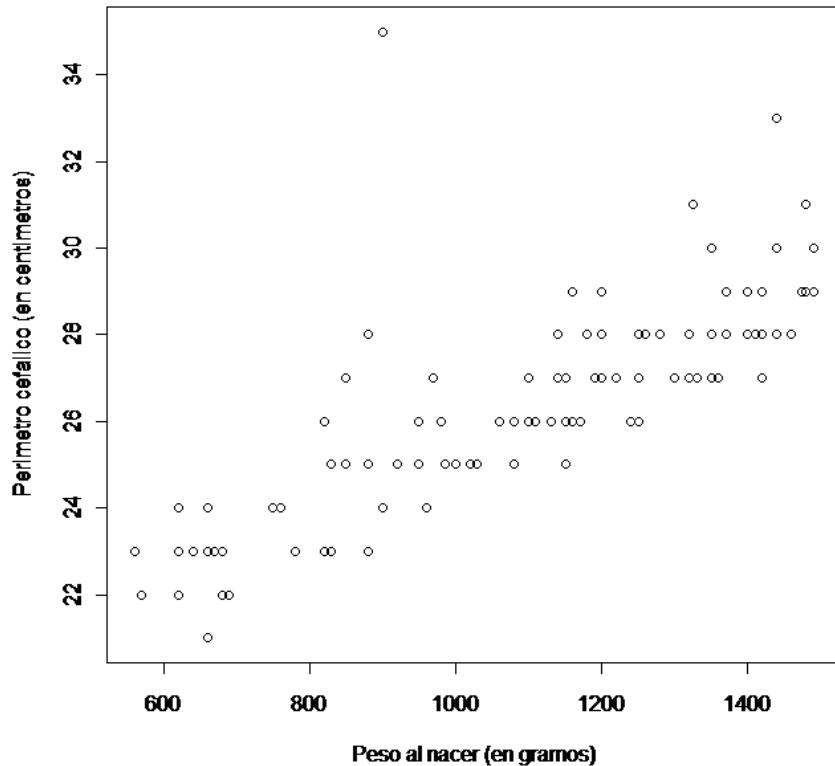
Para darnos una idea de las herramientas con las que trabaja la computadora que ajustará el modelo, listamos los primeros siete datos en la Tabla 18.

El modelo ajustado y las instrucciones para hacerlo en R, figuran en la Tabla 19. La superficie ajustada resulta ser

$$\hat{Y} = 8,3080 + 0,4487X_1 + 0,0047X_2.$$

La ordenada al origen, que es 8,3080 es, en teoría, el valor medio del perímetro cefálico para bebés de bajo peso con edad gestacional de 0 semanas y peso al nacer de 0 gramos, y por lo tanto carece de sentido. El coeficiente estimado de edad gestacional (0,4487) no es el mismo que cuando la edad gestacional era la

Figura 37: Perímetro cefálico versus peso al nacer para la muestra de 100 bebés de bajo peso.



única variable explicativa en el modelo; su valor descendió de 0,7801 a 0,4487. Esto implica que, si mantenemos el peso al nacer de un niño constante, cada incremento de una semana en la edad gestacional corresponde a un aumento de 0,4487 centímetros en su perímetro cefálico, en promedio. Una manera equivalente de decirlo es que dados dos bebés con el mismo peso al nacer pero tales que la edad gestacional del segundo de ellos es una semana más grande que la del primero, el perímetro cefálico esperado para el segundo bebé será 0,4487 centímetros mayor que el primero.

De forma similar, el coeficiente del peso al nacer indica que si la edad gestacional de un bebé no cambia, cada incremento de un gramo en el peso al nacer redundaría en un aumento de 0,0047 centímetros en el perímetro cefálico, en promedio. En este

Tabla 18: Primeros siete datos de bebés de bajo peso

Niño i	$Y_i = \text{headcirc}$	$X_{i1} = \text{gestage}$	$X_{i2} = \text{birthwt}$
1	27	29	1360
2	29	31	1490
3	30	33	1490
4	28	31	1180
5	29	30	1200
6	23	25	680
7	22	27	620

caso en el que el valor del coeficiente estimado es tan pequeño, puede tener más sentido expresar el resultado aumentando las unidades involucradas, por ejemplo decir: si la edad gestacional no cambia, cada incremento de 10 g. en el peso al nacer redundaría en un aumento de 0,047 cm. en el perímetro cefálico, en promedio.

4.8. Resultados de Análisis de la Varianza (y estimación de σ^2)

4.8.1. Sumas de cuadrados y cuadrados medios (SS y MS)

Las sumas de cuadrados para el análisis de la varianza son,

SSTo = suma de cuadrados total

$$\begin{aligned} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 \end{aligned}$$

que en términos matriciales puede escribirse como

$$\text{SSTo} = \mathbf{Y}^t \mathbf{Y} - \frac{1}{n} \mathbf{Y}^t \mathbf{J} \mathbf{Y} = \mathbf{Y}^t \left[\mathbf{I} - \frac{1}{n} \mathbf{J} \right] \mathbf{Y},$$

Tabla 19: Ajuste del modelo lineal para los datos de bebés de bajo peso, `headcirc` con dos explicativas continuas: `gestage` y `birthwt`

```
> ajuste2<-lm(headcirc~gestage+birthwt)
>
> summary(ajuste2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.3080154	1.5789429	5.262	8.54e-07
gestage	0.4487328	0.0672460	6.673	1.56e-09
birthwt	0.0047123	0.0006312	7.466	3.60e-11
<hr/>				

Residual standard error: 1.274 on 97 degrees of freedom
Multiple R-squared: 0.752, Adjusted R-squared: 0.7469
F-statistic: 147.1 on 2 and 97 DF, p-value: < 2.2e-16

donde \mathbf{J} es una matriz $n \times n$ toda de unos. De igual modo,

$$\begin{aligned} \text{SSRes} &= \text{suma de cuadrados de los residuos (SSE)} \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2 \\ &= \mathbf{e}^t \mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}^t \mathbf{Y} - \hat{\boldsymbol{\beta}}^t \mathbf{X}^t \mathbf{Y} = \mathbf{Y}^t [\mathbf{I} - \mathbf{H}] \mathbf{Y} \end{aligned}$$

que en términos matriciales se escribe

$$\text{SSRes} = \mathbf{Y}^t \mathbf{Y} - \hat{\boldsymbol{\beta}}^t \mathbf{X}^t \mathbf{Y} = \mathbf{Y}^t [\mathbf{I} - \mathbf{H}] \mathbf{Y}$$

y

$\text{SSReg} = \text{suma de cuadrados de la regresión o del modelo (SSM)}$

$$= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2,$$

y vale

$$\text{SSReg} = \hat{\boldsymbol{\beta}}^t \mathbf{X}^t \mathbf{Y} - \frac{1}{n} \mathbf{Y}^t \mathbf{J} \mathbf{Y} = \mathbf{Y}^t \left[\mathbf{H} - \frac{1}{n} \mathbf{J} \right] \mathbf{Y}.$$

Más allá de las expresiones matemáticas que permiten calcularlas, en la regresión múltiple se cumple la misma propiedad que en la regresión simple en cuanto a las sumas de cuadrados. Volvamos sobre ellas. Recordemos que como los estimadores $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$ se eligen como aquellos valores de b_0, b_1, \dots, b_{p-1} que minimicen a g dada en (47), luego los parámetros elegidos hacen que la suma de los cuadrados de los residuos (SSRes) sea lo más chica posible. Pero, aunque esta superficie sea la mejor superficie disponible, todavía cabe preguntarse cuan bueno es el ajuste encontrado, es decir, cuan bien ajusta el modelo a los datos observados. Para ello, una manera es comparar el ajuste que proporciona el modelo de regresión lineal con algo, y el algo que siempre podemos elegir es el modelo más básico que podemos encontrar. Entonces usamos las sumas de cuadrados para calcular el ajuste del modelo más básico (un solo parámetro que ajuste a todas las observaciones). Es decir, elegimos el valor de μ tal que minimice

$$\sum_{i=1}^n (Y_i - \mu)^2,$$

sin tener en cuenta para nada los valores de las covariables (X_1, \dots, X_{p-1}) . Es un resultado de un curso inicial de estadística que el valor de μ que minimiza dicha suma es el promedio de las Y s es decir, $\mu = \bar{Y}$. Esencialmente, estamos tomando como medida de cuan bien ajusta un modelo, a la suma de los cuadrados; en general

$$\Delta_{\text{modelo}} = \sum (\text{observados} - \text{modelo})^2 \quad (50)$$

donde el modelo es la superficie de respuesta (44) en regresión lineal múltiple y un sólo parámetro en el modelo más básico. Para cada modelo usamos la ecuación (50) para ajustar ambos modelos, es decir, encontramos los valores de los parámetros que minimizan (50) entre todos los valores posibles y, luego, básicamente si el modelo lineal es razonablemente bueno ajustará a los datos significativamente mejor que el modelo básico. Es decir, la resta

$$\Delta_{\text{modelo básico}} - \Delta_{\text{regresión lineal}} = \text{SSTo} - \text{SSRes}$$

será pequeña comparada con lo que era la SSTo. Esto es un poco abstracto así que mejor lo miramos en un ejemplo.

Imaginemos que nos interesa predecir el perímetro cefálico de un niño al nacer (Y) a partir de la edad gestacional del bebé (X_1) y de su peso al nacer (X_2). ¿Cuánto será el perímetro cefálico de un bebé con 33 semanas de edad gestacional y que pesa 1490 gramos al nacer? Si no tuviéramos un modelo preciso de la relación entre las tres variables en niños nacidos con bajo peso, ¿cuál podría ser nuestro mejor pronóstico? Bueno, posiblemente la mejor respuesta sea dar el número promedio de perímetros cefálicos en nuestra base de datos, que resulta ser 26,45 cm. Observemos

que la respuesta sería la misma si ahora la pregunta fuera: ¿cuánto será el perímetro cefálico de un niño con 25 semanas de gestación y que pesó 680 g. al nacer? Nuevamente, en ausencia de un vínculo preciso, nuestro mejor pronóstico sería dar el promedio observado de perímetros cefálicos, o sea 26,45 cm. Claramente hay un problema: no importa cual es la edad gestacional o el peso al nacer del niño, siempre predecimos el mismo valor de perímetro cefálico. Debería ser claro que la media es poco útil como modelo de la relación entre dos variables, pero es el modelo más básico del que se dispone.

Repasemos entonces los pasos a seguir. Para ajustar el modelo más básico, predecimos el outcome Y por \bar{Y} , luego calculamos las diferencias entre los valores observados y los valores que da el modelo (\bar{Y} siempre para el modelo básico) y la ecuación (50) se convierte en la SSTo (es decir, SSTo es la cantidad total de diferencias presentes cuando aplicamos el modelo básico a los datos). La SSTo representa una medida del desajuste que surge de usar el promedio como único resumen de los datos observados. En un segundo paso ajustamos el modelo más sofisticado a los datos (el modelo de regresión lineal múltiple con dos predictores). Este modelo permite pronosticar un valor distinto para cada combinación de covariables. A este valor lo hemos llamado valor predicho y resulta ser

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}.$$

En el ejemplo, para la primer pregunta nuestra respuesta sería

$$\hat{\beta}_0 + \hat{\beta}_1 33 + \hat{\beta}_2 1490 = 8,3080 + 0,4487 \cdot 33 + 0,0047 \cdot 1490 = 30,118$$

y para la segunda pregunta tendríamos

$$\hat{\beta}_0 + \hat{\beta}_1 25 + \hat{\beta}_2 680 = 8,3080 + 0,4487 \cdot 25 + 0,0047 \cdot 680 = 22,722.$$

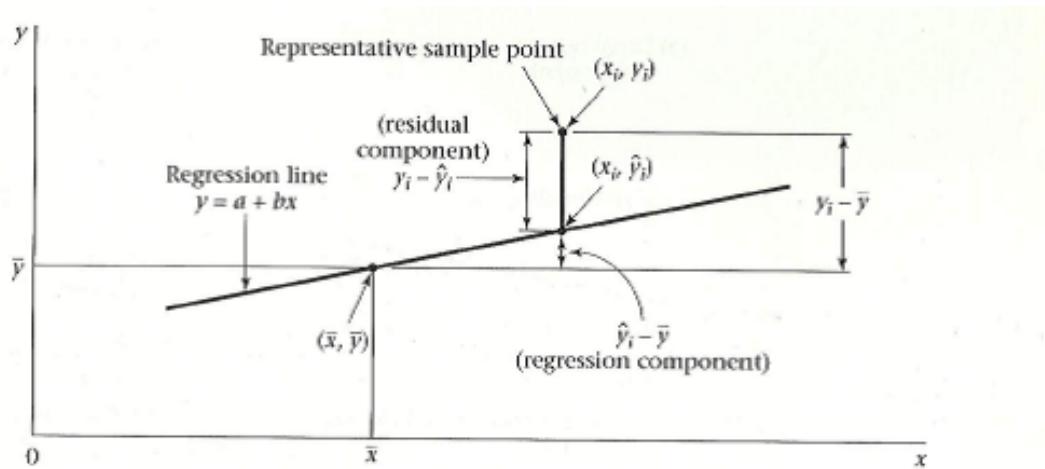
Hemos visto que el modelo de regresión lineal múltiple encuentra los valores de $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ por el método de mínimos cuadrados, es decir minimizando las diferencias entre el modelo ajustado a los datos y los propios datos. Sin embargo, aun en este modelo optimizado hay todavía imprecisiones que se representan por las diferencias entre cada valor observado (Y_i) y cada valor predicho por la regresión (\hat{Y}_i). Como antes, calculamos esas diferencias, elevamos al cuadrado cada una de ellas y las sumamos (si las sumáramos sin elevarlas al cuadrado la suma terminaría dando cero). El resultado se conoce como la suma de los cuadrados de los residuos (SSRes). Este valor representa el grado de imprecisión del modelo lineal con estas dos covariables ajustado a los datos. Podemos usar estos dos valores para calcular cuánto mejor es usar la superficie de respuesta estimada en vez de la media como modelo (es decir, ¿cuánto mejor es el mejor modelo posible comparado con el peor?) La mejora en predicción resultante al usar el mejor modelo en vez de la

media se calcula al hacer la resta entre SSTo y SSRes. Esta diferencia nos muestra la reducción en la imprecisión que se obtiene por usar un modelo de regresión lineal. Como en el caso de regresión lineal simple, puede verse que esta resta da SSReg, es decir

$$\text{SSTo} - \text{SSRes} = \text{SSReg}.$$

La Figura 38 muestra ambas distancias para una misma observación, en el caso de regresión lineal simple.

Figura 38: Distancias que intervienen en las sumas de cuadrados para una observación. Fuente: Rosner [2006], pág. 473.



Si el valor de SSReg es grande, entonces usar el modelo de regresión lineal es muy distinto a usar la media para predecir el outcome. Esto implica que el modelo de regresión ha hecho una gran mejora en la calidad de la predicción de la variable respuesta. Por otro lado, si SSReg es chico, entonces el hecho de usar el modelo de regresión es sólo un poco mejor que usar la media. (Observemos de paso que SSTo siempre será mayor que SSRes ya que tomando $b_1 = b_2 = \dots = b_{p-1} = 0$ y $b_0 = \bar{Y}$ que son valores posibles para los parámetros de la regresión lineal múltiple recuperamos al modelo básico, es decir, el modelo básico está contenido entre todos los modelos posibles bajo la regresión lineal múltiple). Pero ahora, por supuesto, aparece la natural pregunta de cuándo decimos que un valor de SSReg es “grande” o “pequeño”.

4.8.2. Coeficiente de Determinación Múltiple (R^2 y R^2 ajustado)

Una primera manera de zanjar esto es calcular la proporción de mejora debida al modelo. Esto es fácil de hacer dividiendo la suma de cuadrados de la regresión por la suma de cuadrados total. Es lo que hacíamos también en regresión lineal simple. El resultado se denomina R^2 , el **coeficiente de determinación múltiple**. Para expresar este valor como un porcentaje hay que multiplicarlo por 100. Luego, como en el caso de regresión lineal simple, R^2 representa la proporción de variabilidad de la variable respuesta que queda explicada por el modelo de regresión relativa a cuánta variabilidad había para ser explicada antes de aplicar el modelo. Luego, como porcentaje, representa el porcentaje de variación de la variable respuesta que puede ser explicada por el modelo

$$R^2 = \frac{SSReg}{SSTo} = 1 - \frac{SSRes}{SSTo}.$$

De igual modo que para el modelo de regresión lineal simple, R (la raíz cuadrada de R^2) resulta ser la correlación de Pearson entre los valores observados de (Y_i) y los valores predichos (\widehat{Y}_i) sin tener en cuenta el signo. Por lo tanto los valores grandes de R múltiple (al que se lo suele llamar *coeficiente de correlación múltiple*) representan una alta correlación entre los valores observados y predichos del outcome. Un R múltiple igual a uno representa una situación en la que el modelo predice perfectamente a los valores observados.

Observación 4.7 *El hecho de agregar variables explicativas X al modelo de regresión sólo puede aumentar el R^2 y nunca reducirlo, puesto que la suma de cuadrados de los residuos $SSReg$ nunca puede aumentar con más covariables X y la suma de cuadrados total $SSTo$ siempre vale lo mismo para un conjunto fijo de respuestas Y_i . Por este hecho, de que la inclusión de más covariables siempre aumenta el R^2 , sean estas importantes o no, se sugiere que cuando se quieran comparar modelos de regresión con distinto número de covariables en vez de usarse el R^2 se utilice una medida modificada que ajusta por el número de covariables explicativas incluidas en el modelo. El **coeficiente de determinación múltiple ajustado**, que se suele denominar R_a^2 , ajusta a R^2 dividiendo cada suma de cuadrados por sus correspondientes grados de libertad, de la siguiente forma*

$$R_a^2 = 1 - \frac{\frac{SSRes}{n-p}}{\frac{SSTo}{n-1}} = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSRes}{SSTo}$$

Este coeficiente de determinación múltiple puede, de hecho, disminuir cuando se agrega una covariable al modelo, ya que cualquier disminución de la $SSRes$ puede ser más que compensada por la pérdida de un grado de libertad en el denominador

$n - p$. Si al comparar un modelo con las covariables X_1, \dots, X_k para explicar a Y con un modelo que tiene las mismas X_1, \dots, X_k y además a X_{k+1} como covariables vemos un aumento de los R_a^2 , esto es una indicación de que la covariable X_{k+1} es importante para predecir a Y , aún cuando las covariables X_1, \dots, X_k ya están incluidas en el modelo. Si en cambio, el R_a^2 no aumenta o incluso disminuye al incorporar a X_{k+1} al modelo, esto es señal de que una vez que las variables X_1, \dots, X_k se utilizan para predecir a Y , la variable X_{k+1} no contribuye a explicarla y de debe incluirse en el modelo.

Observación 4.8 Hemos dicho que en el modelo lineal múltiple, el R^2 representa el cuadrado del coeficiente de correlación muestral de Pearson entre los valores Y_i observados y los valores \hat{Y}_i predichos. Esto también sucede en regresión lineal simple. Es decir,

$$r = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}}$$

es tal que el valor absoluto de r es la raíz de R^2 , $|r| = \sqrt{R^2}$. En este caso, el signo de r es positivo ya que los valores observados y los predichos están positivamente correlacionados. Entonces, ¿cómo juega la raíz cuadrada? Como R^2 es un número comprendido entre 0 y 1, la raíz cuadrada es en dicho intervalo una función creciente que es la inversa de la función elevar al cuadrado. Por lo tanto, como puede verse en la Figura 39, $r = \sqrt{R^2}$ será **mayor** que R^2 .

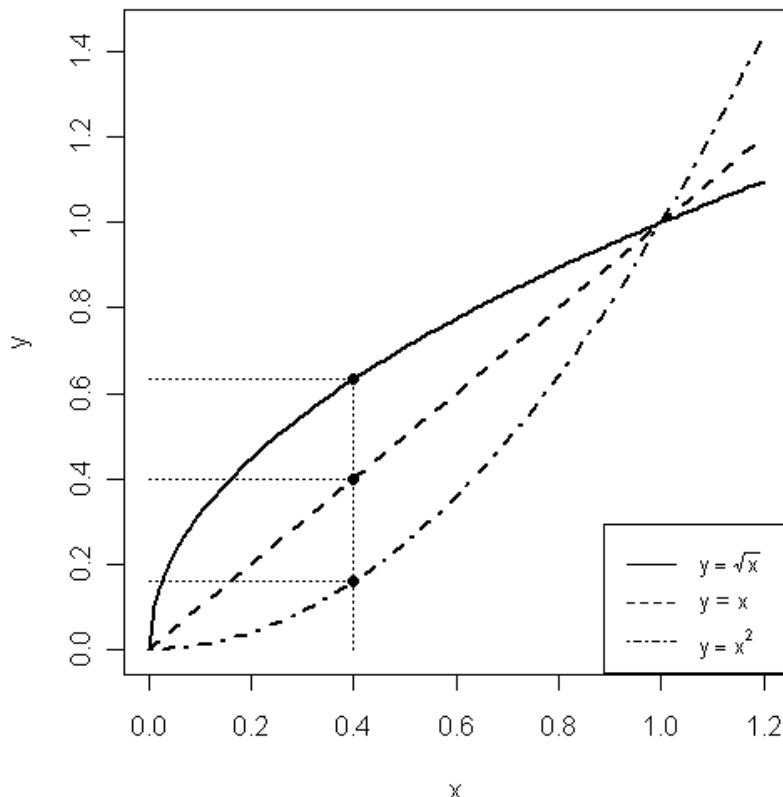
Para ver cómo funciona este vínculo entre r y R^2 inspeccionamos un par de ejemplos numéricos, que exhibimos en la Tabla 20.

Tabla 20: Algunos valores del coeficiente de determinación múltiple R^2 con el respectivo valor del coeficiente de correlación muestral de Pearson, r entre valores predichos y valores observados.

R^2	r
0,1	0,316
0,4	0,632
0,6	0,775
0,7	0,837
0,9	0,949
0,99	0,995

Desde esta óptica, otra interpretación del R^2 es pensar que un buen modelo debería producir valores predichos altamente correlacionados con los valores observados.

Figura 39: Función raíz cuadrada comparada con la función elevar al cuadrado y la identidad en el intervalo $(0, 1)$. Están graficadas las imágenes del $x = 0,4$, con tres puntos cuyas alturas son (en orden ascendente) $0,4^2 = 0,16$; $0,4$ y $\sqrt{0,4} = 0,632$.



Esta es otra manera de visualizar por qué un R^2 alto es, en general, una buena señal de ajuste.

4.8.3. Test F

Como en el modelo de regresión lineal simple, una segunda forma de usar las sumas de cuadrados para evaluar la bondad de ajuste del modelo de regresión lineal múltiple a los datos es a través de un test F . Este test se basa en el cociente de la mejora debida al modelo (SSReg) y la diferencia entre el modelo y los datos observados (SSRes). La Tabla 21 resume la información que involucra a la construcción del test F . De hecho, en vez de utilizar las sumas de cuadrados por sí mismas,

tomamos lo que se denominan los cuadrados medios (MS *mean squares* o sumas medias de cuadrados o cuadrados medios). Para trabajar con ellos, es necesario primero dividir a las sumas de cuadrados por sus respectivos grados de libertad. Para la SSReg, los grados de libertad son simplemente el número de covariables en el modelo, es decir, $p - 1$. Del mismo modo que sucedía con la regresión lineal simple, las diferencias $(\hat{Y}_i - \bar{Y})$ quedan determinadas al fijar los $p - 1$ coeficientes que acompañan a las $p - 1$ covariables, luego las diferencias $(\hat{Y}_i - \bar{Y})$ tienen $p - 1$ grados de libertad.

Tabla 21: Tabla de ANOVA para el modelo de Regresión Lineal General (44)

Fuente de variación	SS	g.l.	MS
Regresión	$SSReg = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$p - 1$	$MSReg = \frac{SSReg}{p-1}$
Residuos	$SSRes = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - p$	$MSRes = \frac{SSRes}{n-p}$
Total	$SSTo = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$	

Para la SSRes los grados de libertad son el número de observaciones menos el número de parámetros que se estiman (es decir, el número de coeficientes beta incluyendo el β_0), en este caso $n - p$. Esto proviene, al igual que en el caso de regresión lineal simple, del hecho de que los residuos satisfacen p ecuaciones normales. Las p ecuaciones que se obtienen al igualar a cero las p derivadas. Luego, si conocemos $n - p$ de ellos, podemos hallar los restantes p a partir de despejarlos de las p ecuaciones lineales.

Los resultados son, respectivamente, el cuadrado medio de regresión (que notaremos MSReg o MSM, es decir *regression mean square* o *model mean square*) y el cuadrado medio de residuos (MSRes o MSE, es decir, *residual mean square* o *mean square error*). El estadístico F es una medida de cuánto mejora el modelo la predicción de la variable respuesta comparada con el nivel de imprecisión de los datos originales. Si el modelo es bueno, esperamos que la mejora en la predicción debida al modelo sea grande (de manera que MSReg sea grande) y que la diferencia entre el modelo y los datos observados sea pequeña (o sea, MSRes pequeña). Por eso, un buen modelo debe tener un estadístico F grande (al menos mayor a 1 porque en tal caso el numerador, de decir, la mitad superior de (51) será mayor que el denominador -la mitad inferior de (51)). El estadístico F es

$$F = \frac{MSReg}{MSRes} = \frac{\frac{SSReg}{p-1}}{\frac{SSRes}{n-p}} = \frac{SSReg(n-p)}{SSRes(p-1)}. \quad (51)$$

Se construye para testear las hipótesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

$$H_1 : \text{no todos los } \beta_k \ (k = 1, 2, \dots, p-1) \text{ son iguales a 0}$$

Observemos que H_0 dice que no hay vínculo entre la variable respuesta y las regresoras. En cambio, H_1 dice que al menos una de las variables regresoras sirve para predecir a Y . La distribución de F cuando H_0 es cierta es la distribución F (de Snedecor o de Fisher) con $p-1$ grados de libertad en el numerador y $n-p$ grados de libertad en el denominador⁴. Esto es porque bajo el supuesto de normalidad de los errores, se tiene que

$$\text{SSRes} \sim \chi_{n-p}^2$$

y si además H_0 es verdadera, entonces

$$\text{SSReg} \sim \chi_{p-1}^2$$

y además SSRes y SSReg son independientes. El test rechaza H_0 cuando $F > F_{p-1, n-p, 1-\alpha}$, el $1 - \alpha$ percentil de la distribución válida cuando H_0 es verdadera. Para valores grandes de F (es decir, p-valores pequeños) el test rechaza H_0 y concluye que no todos los coeficientes que acompañan a las covariables del modelo de regresión lineal son nulos.

Observación 4.9 *Cuando $p-1 = 1$, este test se reduce al test F visto en el modelo de regresión lineal simple para testear si β_1 es 0 o no.*

Observación 4.10 *La existencia de una relación de regresión lineal, por supuesto, no asegura que puedan hacerse predicciones útiles a partir de ella.*

Usualmente, como ya hemos visto en el modelo lineal simple, estos valores aparecen en la salida de cualquier paquete estadístico en lo que se conoce como tabla de ANOVA (*Analysis of Variance table*, que presentamos en la Tabla 21).

Usualmente la tabla se completa con dos últimas columnas que se denominan F y p-valor. La columna F tiene un único casillero completo (el correspondiente a la primer fila) con el valor del estadístico, es decir

$$F_{obs} = \frac{\text{MSReg}}{\text{MSRes}}.$$

La columna p-valor tiene también un único casillero con el p-valor del test, que es la probabilidad, calculada asumiendo que H_0 es verdadera, de observar un valor del estadístico F tan alejado de lo esperado como el observado en la muestra, o más alejado aún, o sea

$$p\text{-valor} = P(F_{p-1, n-p} > F_{obs}).$$

⁴Por definición, si U es una variable aleatoria con distribución χ_k^2 y V es otra variable aleatoria independiente de U con distribución χ_m^2 , entonces la variable $W = \frac{U/k}{V/m}$ se denomina *F de Fisher con k grados de libertad en el numerador y m grados de libertad en el denominador*.

4.8.4. Estimación de σ^2

El modelo de regresión lineal dado en (44) y (45) impone que los errores $\varepsilon_1, \dots, \varepsilon_n$ sean variables aleatorias independientes con esperanza cero y $\text{Var}(\varepsilon_i) = \sigma^2$. Si tuviéramos los errores, sabemos que un estimador insesgado de σ^2 es

$$\frac{1}{n-1} \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2.$$

El problema es que en el modelo de regresión lineal múltiple, al igual que en el caso de regresión lineal simple, los errores **no son observables**. Para estimar a σ^2 los podemos reemplazar por sus correlatos empíricos, los residuos e_1, \dots, e_n . Pero, como ya vimos en la Observación 4.5 los residuos **no** son independientes. En el caso del modelo lineal simple habíamos visto que los residuos están ligados entre sí ya que satisfacen dos ecuaciones lineales (las dos ecuaciones normales):

- la suma de los residuos e_1, \dots, e_n es cero.
- la correlación muestral entre e_1, \dots, e_n y X_1, \dots, X_n es cero, o equivalentemente, el coeficiente de correlación de Pearson calculado para $(X_1, e_1), \dots, (X_n, e_n)$ es cero.

En el caso de regresión lineal múltiple con $p - 1$ variables predictoras, los residuos están ligados entre sí de una manera más estrecha, ya que satisfacen p ecuaciones lineales (linealmente independientes): como $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ y \mathbf{H} es una matriz de proyección de rango p resulta que $\mathbf{H}\mathbf{e} = \mathbf{0}$. Una de ellas es, también, que la suma de los residuos vale cero. Informalmente se dice que los residuos tienen $n - p$ grados de libertad. Esto quiere decir que conociendo $n - p$ de ellos, podemos deducir cuánto valen los p restantes despejándolos de las ecuaciones normales. Luego, el estimador de σ^2 se basará en los residuos de la siguiente forma

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-p} \sum_{i=1}^n (e_i - \bar{e})^2 = \frac{1}{n-p} \sum_{i=1}^n (e_i)^2 \\ &= \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{\text{SSRes}}{n-p} \\ &= \text{MSRes}. \end{aligned} \tag{52}$$

Es decir, el cuadrado medio de los residuos es el estimador de σ^2 dado por el modelo de regresión. En la salida de un paquete estadístico se puede encontrar en el casillero correspondiente en la tabla de ANOVA.

4.9. Inferencias sobre los parámetros de la regresión

Los estimadores de mínimos cuadrados $\hat{\beta}_k$ son insesgados, es decir,

$$E(\hat{\beta}_k) = \beta_k.$$

La matriz de covarianza de dichos estimadores $Var(\hat{\beta})$ está dada por una matriz $p \times p$ que en la coordenada jk tiene la covarianza entre $\hat{\beta}_j$ y $\hat{\beta}_k$ y que resulta ser

$$Var(\hat{\beta}) = \sigma^2 (X^t X)^{-1}.$$

Como vimos en la Sección 4.8.4, MSRes es el estimador de σ^2 , por lo que la estimación de dicha matriz está dada por

$$\widehat{Var}(\hat{\beta}) = \widehat{\sigma}^2 (X^t X)^{-1} = \text{MSRes}(X^t X)^{-1}.$$

4.9.1. Intervalos de confianza para β_k

Para el modelo de errores normales dado por (44) y (45) tenemos que

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}} \sim t_{n-p} \text{ para } k = 0, 1, \dots, p-1.$$

Recordemos que $n-p$ es el número de observaciones menos el número de covariables del modelo menos uno. Muchas veces al denominador $\sqrt{\widehat{Var}(\hat{\beta}_k)}$ se lo llama $s(\hat{\beta}_k)$. Luego, el intervalo de confianza de nivel $1-\alpha$ para cada β_k es

$$\hat{\beta}_k \pm t_{n-p, 1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\beta}_k)}. \quad (53)$$

4.9.2. Tests para β_k

Los tests para β_k se llevan a cabo de la forma usual. Para testear

$$\begin{aligned} H_0 &: \beta_k = 0 \\ H_1 &: \beta_k \neq 0 \end{aligned}$$

usamos el estadístico

$$T = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}}$$

y rechazamos H_0 cuando $|T| \geq t_{n-p,1-\frac{\alpha}{2}}$. El p-valor, a su vez, se calcula como

$$p\text{-valor} = P(|t_{n-p}| \geq |T_{obs}|).$$

Observemos que cuando realizamos este test asumimos que en el modelo aparecen todas las restantes covariables. Se puede calcular la potencia de este test.

4.9.3. Inferencias conjuntas

El objetivo de los intervalos de confianza y tests presentados en las secciones 4.9.1 y 4.9.2 es proveer conclusiones con un nivel prefijado de confianza sobre cada uno de los parámetros $\beta_0, \beta_1, \dots, \beta_{p-1}$ por separado. La dificultad es que éstos no proporcionan el 95 por ciento de confianza de que las conclusiones de los p intervalos son correctas. Si las inferencias fueran independientes, la probabilidad de que los p intervalos construidos cada uno a nivel 0,95, contengan al verdadero parámetro sería $(0,95)^p$, o sea, solamente 0,857 si p fuese 3. Sin embargo, las inferencias no son independientes, ya que son calculadas a partir de un mismo conjunto de datos de la muestra, lo que hace que la determinación de la probabilidad de que todas las inferencias sean correctas sea mucho más difícil.

En esta sección propondremos intervalos de confianza de **nivel conjunto** 0,95. Esto quiere decir que nos gustaría construir una serie de intervalos (o tests) para los cuales tengamos una garantía sobre la exactitud de todo el conjunto de intervalos de confianza (o tests). Al conjunto de intervalos de confianza (o tests) de interés lo llamaremos familias de intervalos de confianza de nivel conjunto o simultáneo (o regiones de confianza de nivel simultáneo o tests o inferencias conjuntas). En nuestro ejemplo, la familia se compone de p estimaciones, para $\beta_0, \beta_1, \dots, \beta_{p-1}$. Podríamos estar interesados en construir regiones de confianza para una cantidad g entre 1 y p de estos parámetros, con g prefijado. Distingamos entre un intervalo de confianza de nivel 0,95 para un parámetro, y una familia de intervalos de nivel simultáneo 0,95 para g parámetros. En el primer caso, 0,95 es la proporción de intervalos construido con el método en cuestión que cubren al verdadero parámetro de interés cuando se seleccionan repetidamente muestras de la población de interés y se construyen los intervalos de confianza para cada una de ellas. Por otro lado, cuando construimos una familia de regiones o intervalos de confianza de nivel simultáneo 0,95 para g parámetros: $\theta_1, \dots, \theta_g$ el valor 0,95 indica la proporción de familias de g intervalos que están enteramente correctas (cubren a los g parámetros de interés, simultáneamente) cuando se seleccionan repetidamente muestras de la población de interés y se construyen los intervalos de confianza específicos para los g parámetros en cuestión, o sea

$$P(\{\theta_1 \in I_1\} \cap \{\theta_2 \in I_2\} \cap \cdots \cap \{\theta_g \in I_g\}) = 0,95,$$

si I_1, \dots, I_g son los g intervalos construidos usando los mismos datos. Luego, el **nivel simultáneo** de una familia de regiones o intervalos de confianza corresponde a la probabilidad, calculada previa al muestreo, de que la familia entera de afirmaciones sea correcta.

Ilustremos esto en el caso del ejemplo de los 100 bebés de bajo peso. Si nos interesaría construir intervalos de confianza de nivel simultáneo 0,95 para β_1 y β_2 , una familia de intervalos de confianza simultáneos para estos datos consistiría en dos intervalos de confianza de modo tal que si tomáramos muestras de 100 bebés de bajo peso, les midiéramos la edad gestacional, el perímetro cefálico y el peso al nacer, y luego construyéramos para cada muestra los dos intervalos de confianza para β_1 y β_2 , para el 95 % de las muestras ambos intervalos construidos con este método cubrirían tanto al verdadero β_1 como al verdadero β_2 . Para el 5 % restante de las muestras, resultaría que uno o ambos intervalos de confianza sería incorrecto.

En general es sumamente deseable contar con un procedimiento que provea una familia de intervalos de confianza de nivel simultáneo cuando se estiman varios parámetros con una misma muestra de datos, ya que le permite al analista entrelazar varios resultados juntos en un conjunto integrado de conclusiones con la seguridad de que todo el conjunto de inferencias es correcto. Para obtenerlos hay básicamente dos herramientas estadísticas disponibles. Una de ellas es el estudio matemático en detalle del fenómeno en cuestión, en este caso, estudiar matemáticamente las propiedades de los estimadores $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ de manera de poder obtener la distribución exacta de alguna medida numérica que los resuma, como el $\max_{0 \leq k \leq p-1} |\hat{\beta}_k|$ o las descripciones matemáticas del elipsoide p dimensional más pequeño que los contenga, con probabilidad 0,95, para contar un par de ejemplos que son utilizados en distintas áreas de la estadística para construir regiones de confianza de nivel simultáneo. Veremos otro en la Sección 4.10.2. La otra herramienta consiste en construir intervalos de confianza con nivel simultáneo a partir de ajustar el nivel de confianza de cada intervalo individual a un valor más alto, de modo de poder asegurar el nivel simultáneo de la construcción. Esto es lo que se conoce como el método de Bonferroni para la construcción de intervalos de nivel simultáneo. Una descripción detallada de este método puede consultarse en Kutner et al. [2005], pág. 155 a 157. Este procedimiento es de aplicación bastante general en la estadística. En vez de usar el percentil de la t propuesto en la Sección 4.9.1 para cada intervalo de confianza para β_k se usa el percentil correspondiente a un nivel mayor. Cuando se quieren construir intervalos de confianza de nivel simultáneo $1 - \alpha$ para g coeficientes de la regresión, el percentil que se utiliza es el correspondiente a un nivel $1 - \frac{\alpha}{2g}$ en cada intervalo en particular. Resultan ser intervalos más anchos que los presentados en la Sección 4.9.1. Una observación importante es que el procedimiento de Bonferroni es conservativo, es decir, el nivel conjunto de los intervalos así construidos resulta ser mayor o igual a $1 - \alpha$.

Así, se pueden construir los intervalos de confianza simultáneos de Bonferroni para estimar varios coeficientes de regresión de manera simultánea. Si se desean estimar simultáneamente g parámetros (donde $g \leq p$), los intervalos de confianza con nivel simultáneo $1 - \alpha$ son los siguientes

$$\hat{\beta}_k \pm t_{n-p, 1-\frac{\alpha}{2g}} \sqrt{\widehat{Var}(\hat{\beta}_k)}.$$

Más adelante discutiremos tests que conciernen varios parámetros de regresión en forma simultánea.

4.9.4. Aplicación al ejemplo

Antes de seguir presentando teoría, veamos cómo se calculan e interpretan estas cuestiones en el ejemplo de los 100 bebés de bajo peso. Para dicho ejemplo, cuyo modelo contenía a la edad gestacional y el peso al nacer como variables explicativas, $p - 1$ resulta ser igual a 2 (luego $p = 3$). La distribución t involucrada en la construcción de intervalos de confianza o tests para los β_k tiene en este caso $n - p = 100 - 3 = 97$ grados de libertad. En la Tabla 19 que figura en la página 125 exhibimos los coeficientes estimados. Los recordamos a continuación

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.3080154	1.5789429	5.262	8.54e-07
gestage	0.4487328	0.0672460	6.673	1.56e-09
birthwt	0.0047123	0.0006312	7.466	3.60e-11

Luego,

$$\hat{\beta}_0 = 8,3080 \quad \hat{\beta}_1 = 0,4487 \quad \hat{\beta}_2 = 0,0047$$

y sus errores estándares respectivos resultan ser

$$\sqrt{\widehat{Var}(\hat{\beta}_0)} = s(\hat{\beta}_0) = 1,5789 \quad s(\hat{\beta}_1) = 0,0672 \quad s(\hat{\beta}_2) = 0,00063$$

Luego, los respectivos estadísticos t observados en cada caso son

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{0,4487}{0,0672} = 6,67$$

cuando $k = 1$ y

$$T = \frac{\hat{\beta}_2 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} = \frac{0,0047}{0,00063} = 7,46$$

cuando $k = 2$. En ambos casos, los p-valores resultan ser menores que 0,001. Observemos que en la salida de cualquier paquete estadístico figuran tanto las estimaciones de los betas, como sus desvíos estándares estimados, los valores de t observados y los p-valores respectivos. En ambos casos rechazamos las hipótesis nulas a nivel 0,05 y concluimos que β_1 es distinta de cero cuando en el modelo aparece X_2 como explicativa (en el primer test) y que β_2 es distinta de cero cuando en el modelo aparece X_1 como explicativa (en el segundo test). Como además ambos estimadores son positivos, concluimos que el perímetro cefálico aumenta cuando aumenta tanto la edad gestacional como cuando aumenta el peso al nacer. Debemos tener presente, sin embargo, que varios tests de hipótesis basados en los mismos datos no son independientes; si cada test se realiza a nivel de significación α , la probabilidad global de cometer un error de tipo I –o rechazar la hipótesis nula cuando es verdadera– es, de hecho, mayor que α . Para eso se pueden realizar los tests simultáneos presentados, como los de Bonferroni.

Los intervalos de confianza para ambos parámetros de la regresión resultan ser

$$\begin{aligned}\hat{\beta}_1 &\pm t_{97,0,975} \sqrt{\widehat{Var}(\hat{\beta}_1)} \\ &= [0,4487 - 1,9847 \cdot 0,06724; \quad 0,4487 + 1,9847 \cdot 0,06724] \\ &= [0,31525; \quad 0,58215]\end{aligned}$$

y

$$\begin{aligned}\hat{\beta}_2 &\pm t_{97,0,975} \sqrt{\widehat{Var}(\hat{\beta}_2)} \\ &= [0,004712 - 1,9847 \cdot 0,00063; \quad 0,004712 + 1,9847 \cdot 0,00063] \\ &= [0,00346; \quad 0,00596]\end{aligned}$$

o, calculados con el R, como figuran en la Tabla 22.

Si usáramos el procedimiento de Bonferroni para construir los intervalos, tendríamos que usar el percentil

$$1 - \frac{\alpha}{2g} = 1 - \frac{0,05}{2 \cdot 3} = 0,99167$$

de una t_{97} , es decir, $t_{97,0,99167} = 2,43636$ en vez de $t_{97,0,975} = 1,9847$, que nos dará intervalos más anchos, como puede observarse comparando los intervalos de confianza de las Tablas 22 y 23, la primera contiene a los intervalos de confianza de nivel 0,95 cada uno, y la segunda contiene los intervalos de confianza de nivel simultáneo 0,95.

Si calculamos el R^2 para este modelo (que figura en la Tabla 19) vemos que es $R^2 = 0,752$, luego el modelo que contiene a la edad gestacional y el peso al nacer

Tabla 22: Intervalos de confianza de nivel 0,95 para β_0 , β_1 y β_2 para los datos de niños de bajo peso al nacer

```
> confint(ajuste2)
              2.5 %      97.5 %
(Intercept) 5.174250734 11.441780042
gestage     0.315268189  0.582197507
birthwt     0.003459568  0.005964999
```

Tabla 23: Intervalos de confianza de nivel simultáneo 0,95 para β_0 , β_1 y β_2 para los datos de niños de bajo peso al nacer, construidos con el método de Bonferroni

```
> confint(ajuste2, level=(1-(0.05/3)))
              0.833 %      99.167 %
(Intercept) 4.461384677 12.154646098
gestage     0.284907765  0.612557932
birthwt     0.003174601  0.006249966
> 0.05/(2*3)
[1] 0.008333333
```

como variables explicativas explica el 75,20 % de la variabilidad en los datos observados de perímetro cefálico; el modelo que tenía solamente a la edad gestacional explicaba el 60,95 %. Este aumento en el R^2 sugiere que agregar la variable peso al modelo mejora nuestra habilidad para predecir el perímetro cefálico para la población de bebés nacidos con bajo peso. Pero, como ya vimos, debemos ser muy cuidadosos al comparar coeficientes de determinación de dos modelos diferentes. Ya dijimos que la inclusión de una nueva covariante al modelo nunca puede hacer que el R^2 decrezca; el conocimiento de la edad gestacional y el peso al nacer, por ejemplo, nunca puede explicar menos de la variabilidad observada en los perímetros cefálicos que el conocimiento de la edad gestacional sola (aun si la variable peso no contribuyera en la explicación). Para sortear este problema podemos usar una segunda medida (cuando el interés sea comparar el ajuste que producen dos o más modelos entre sí), el R^2 ajustado (que notaremos R_a^2), que compensa por la complejidad extra que se le agrega al modelo. El R^2 ajustado aumenta cuando la inclusión de una variable mejora nuestra habilidad para predecir la variable y disminuye cuando no lo hace. Consecuentemente, el R^2 ajustado nos permite hacer

una comparación más justa entre modelos que contienen diferente número de covariables. Como el coeficiente de determinación, el R^2 ajustado es una estimación del coeficiente de correlación poblacional ρ ; a diferencia del R^2 , sin embargo, no puede ser directamente interpretado como la proporción de la variabilidad de los valores Y que queda explicada por el modelo de regresión. En este ejemplo, el R^2 ajustado resulta ser 0,7469 (ver nuevamente la Tabla 19) que al ser mayor que el R^2 ajustado del modelo con sólo una variable explicativa, la edad gestacional (era $R^2_a = 0,6055$) indica que la inclusión del peso al nacer en el modelo, mejora nuestra capacidad para predecir el perímetro cefálico del niño.

Finalmente, la tabla ANOVA para estos datos aparece en la Figura 40 con el SPSS y en la Tabla 24 con el R.

Figura 40: Tabla de ANOVA para los datos de niños de bajo peso al nacer

ANOVA^b					
Modelo	Suma de cuadrados	gl	Media cuadrática	F	Sig.
1 Regresión	477,327	2	238,663	147,058	,000 ^a
Residual	157,423	97	1,623		
Total	634,750	99			

a. Variables predictoras: (Constante), birthwt, Edad gestacional (semanas)

b. Variable dependiente: Perímetro cefálico (centímetros)

Tabla 24: Tabla de Anova para los datos de bebés de bajo peso, en R.

```
> ajuste2<-lm(headcirc~gestage+birthwt)
> ajuste1<-lm(headcirc~1)
> anova(ajuste1,ajuste2)
Analysis of Variance Table

Model 1: headcirc ~ 1
Model 2: headcirc ~ gestage + birthwt
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1      99 634.75
2      97 157.42  2     477.33 147.06 < 2.2e-16 ***
---

```

Observemos que el estimador de σ^2 que surge del modelo de regresión es

$$\text{MSRes} = \frac{\text{SSRes}}{n-p} = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 1,62, \quad (54)$$

por la Sección 4.8.4. Si comparamos el valor observado del estimador $\hat{\sigma}^2 = 1,62$ para este modelo con el estimador de la varianza no explicada por el modelo de regresión lineal simple que sólo tiene a la edad gestacional como explicativa, que era 2,529 (ver Tabla 2.8) observamos que con la inclusión del peso hemos reducido la variabilidad no explicada por el modelo, mejorando la calidad del ajuste obtenido (y de las predicciones que pueden hacerse con él).

4.10. Estimación de la Respuesta Media

4.10.1. Intervalo de confianza para $E(Y_h)$

Nos interesa estimar la respuesta media o esperada cuando (X_1, \dots, X_{p-1}) toma el valor dado $(X_{h1}, \dots, X_{h,p-1})$. Notamos a esta respuesta media por $E(Y_h)$ o bien $E(Y_h | (X_{h1}, \dots, X_{h,p-1}))$. Como en regresión lineal simple estos valores $(X_{h1}, \dots, X_{h,p-1})$ pueden ser valores que hayan ocurrido en la muestra considerada o pueden ser algunos otros valores de las variables predictoras dentro del alcance (*scope*) del modelo. Definimos el vector

$$\mathbf{X}_h = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

de modo que la respuesta a ser estimada es

$$E(Y_h) = E(Y_h | \mathbf{X}_h) = \mathbf{X}_h^t \boldsymbol{\beta}.$$

La respuesta media estimada correspondiente a \mathbf{X}_h , que denotamos por \hat{Y}_h es la variable aleatoria que se calcula del siguiente modo

$$\hat{Y}_h = \mathbf{X}_h^t \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 X_{h1} + \hat{\beta}_2 X_{h2} + \cdots + \hat{\beta}_{p-1} X_{h,p-1}.$$

Para el modelo de errores normales (45) la distribución de \hat{Y}_h será normal, con media

$$E(\hat{Y}_h) = \mathbf{X}_h^t \boldsymbol{\beta} = E(Y_h) \quad (55)$$

y varianza

$$\text{Var}(\hat{Y}_h) = \sigma^2 \mathbf{X}_h^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_h = \mathbf{X}_h^t \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{X}_h.$$

Como la esperanza del predicho es igual a lo que queremos estimar, es decir, $E(\hat{Y}_h) = E(Y_h)$, el estimador resulta ser insesgado. La varianza estimada resulta ser

$$\widehat{\text{Var}}(\hat{Y}_h) = \text{MSRes} \cdot \mathbf{X}_h^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_h = \mathbf{X}_h^t \widehat{\text{Var}}(\hat{\beta}) \mathbf{X}_h. \quad (56)$$

A partir de (55) y (56) puede obtenerse un intervalo de confianza de nivel $1 - \alpha$ para $E(Y_h)$, la respuesta media esperada cuando las covariables son \mathbf{X}_h , que viene dado por

$$\hat{Y}_h \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}(\hat{Y}_h)}. \quad (57)$$

En general, estos intervalos serán calculados usando un paquete estadístico.

4.10.2. Región de Confianza para la Superficie de Regresión

La región de confianza para toda la superficie de regresión es una extensión de la banda de confianza de Hotelling o Working-Hotelling para una recta de regresión (cuando hay una sola variable predictora). Los puntos de la frontera de la región de confianza en \mathbf{X}_h , se obtienen a partir de

$$\hat{Y}_h \pm W \cdot \sqrt{\widehat{\text{Var}}(\hat{Y}_h)}$$

donde

$$W^2 = pF_{p,n-p;1-\alpha}. \quad (58)$$

Puede probarse que eligiendo este percentil, la región resultante cubrirá a la superficie de regresión **para todas las combinaciones posibles de las variables \mathbf{X}** (dentro de los límites observados), con nivel $1 - \alpha$. Es por eso que esta región de confianza tiene nivel simultáneo o global $1 - \alpha$, como discutimos en la Sección 4.9.3.

4.10.3. Intervalos de Confianza Simultáneos para Varias Respuestas Medias

Para estimar un número de respuestas medias $E(Y_h)$ correspondientes a distintos vectores \mathbf{X}_h con coeficiente de confianza global $1 - \alpha$ podemos emplear dos enfoques diferentes:

1. Usar las regiones de confianza para la superficie de regresión basadas en la distribución de Hotelling (58) para varios vectores \mathbf{X}_h de interés

$$\hat{Y}_h \pm W \cdot \sqrt{\widehat{\text{Var}}(\hat{Y}_h)}.$$

donde \widehat{Y}_h , W y $\widehat{Var}(\widehat{Y}_h)$ están definidos respectivamente en (55), (58) y (56). Como la región de confianza para la superficie de regresión basada en la distribución de Hotelling cubre la respuesta media para todos los vectores \mathbf{X}_h posibles con nivel conjunto $1 - \alpha$, los valores de frontera seleccionados cubrirán las respuestas medias para los vectores \mathbf{X}_h de interés con nivel de confianza global mayor a $1 - \alpha$.

2. Usar intervalos de confianza simultáneos de Bonferroni. Cuando se quieren hallar g intervalos de confianza simultáneos, los límites serán

$$\widehat{Y}_h \pm B \cdot \sqrt{\widehat{Var}(\widehat{Y}_h)}.$$

donde

$$B = t_{n-p, 1-\frac{\alpha}{2g}}.$$

Para una aplicación en particular, podemos comparar los valores de W y B para ver cuál procedimiento conduce a tener los intervalos de confianza más angostos. Si los niveles \mathbf{X}_h no son conocidos antes de aplicar el modelo, sino que surgen del análisis, es mejor usar los intervalos basados en la distribución de Hotelling, puesto que la familia de estos intervalos incluye a todos los posibles valores de \mathbf{X}_h .

4.11. Intervalos de Predicción para una Nueva Observación $Y_{h(\text{nueva})}$

Como en el caso de regresión lineal simple, estamos interesados ahora en predecir una nueva observación Y correspondiente a un nivel dado de las covariables \mathbf{X}_h . La nueva observación Y a ser predicha se puede ver como el resultado de una nueva repetición del experimento u observación, independiente de los resultados anteriores en los que se basa el análisis de regresión. Denotamos el nivel de \mathbf{X} para la nueva observación por \mathbf{X}_h y a la nueva observación de Y como $Y_{h(\text{nueva})}$. Por supuesto, asumimos que el modelo de regresión subyacente aplicable a los datos con los que contamos sigue siendo apropiado para la nueva observación.

La diferencia entre la estimación de la respuesta media $E(Y_h)$, tratado en la sección anterior, y la predicción de una nueva respuesta $Y_{h(\text{nueva})}$, que discutimos en esta, es básica. En el primer caso, se estima la media de la distribución de Y . En el segundo caso, queremos predecir un *resultado individual* surgido a partir de la distribución de Y . Por supuesto, la gran mayoría de los resultados individuales se desvían de la respuesta media, y esto debe ser tenido en cuenta por el procedimiento para la predicción de la $Y_{h(\text{nueva})}$.

4.11.1. Intervalo de predicción para $Y_{h(\text{nueva})}$ cuando los parámetros son conocidos

Para ilustrar la naturaleza de un intervalo de predicción para una nueva observación de la $Y_{h(\text{nueva})}$ de la manera más simple posible, en primer lugar supondremos que todos los parámetros de regresión son conocidos. Más adelante abandonaremos este supuesto para tener el enfoque realista y haremos las modificaciones pertinentes.

Consideremos el ejemplo de los niños con bajo peso al nacer. Supongamos que supiéramos que los parámetros del modelo son

$$\begin{aligned}\beta_0 &= 8 \quad \beta_1 = 0,5 \quad \beta_2 = 0,004 \quad \sigma = 1,25 \\ E(Y) &= 8 + 0,5X_1 + 0,004X_2\end{aligned}$$

El analista considera ahora un bebé de 30 semanas de edad gestacional y que pesó 1360g. al nacer. El perímetro cefálico medio para $X_{h1} = 30$ y $X_{h2} = 1360$ es

$$E(Y) = 8 + 0,5 \cdot 30 + 0,004 \cdot 1360 = 28,44$$

En la Figura 41 se muestra la distribución para Y_h para $\mathbf{X}_h^t = (1, 30, 1360)$. Su media es $E(Y_h) = 28,44$ y su desvío estándar es $\sigma = 1,25$. La distribución es normal debido al modelo de regresión (44) y (45).

Supongamos que fuéramos a predecir el perímetro cefálico de un bebé con estos valores de las covariables, diríamos que está entre

$$\begin{aligned}E(Y_h) &\pm 3\sigma \\ 28,44 &\pm 3 \cdot 1,25\end{aligned}$$

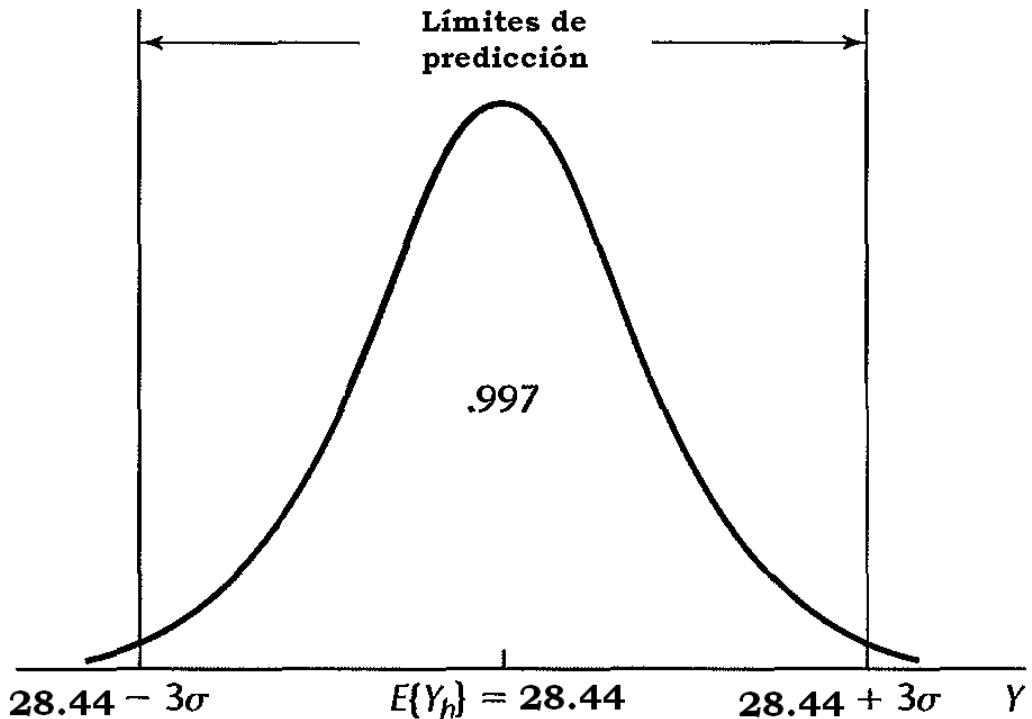
de modo que el intervalo de predicción sería

$$24,69 \leq Y_{h(\text{nueva})} \leq 32,19$$

Como el 99,7 por ciento del área en una distribución de probabilidad normal cae dentro de los tres desvíos estándares de la media, hay una probabilidad de 0,997 de que este intervalo de predicción dé una predicción correcta para el perímetro cefálico del bebé en cuestión, con 30 semanas de gestación y que pesó 1360g. al nacer. Los límites de predicción en este caso son bastante amplios, por lo que la predicción no es muy precisa, sin embargo, el intervalo de predicción indica que el bebé tendrá un perímetro cefálico mayor a 24 cm., por ejemplo.

La idea básica de un intervalo de predicción es, pues, elegir un rango en la distribución de Y en donde la mayoría de las observaciones caerá, y luego, declarar que la observación siguiente caerá en este rango. La utilidad del intervalo de predicción

Figura 41: Distribución de Y_h cuando $\mathbf{X}_h^t = (1, 30, 1360)$. Fuente: Kutner et al. [2005], pág. 57.



depende, como siempre, del ancho del intervalo y de la necesidad de precisión por parte del usuario.

En general, cuando los parámetros del modelo de regresión con errores normales son conocidos, los límites de la predicción de la $Y_{h(\text{nueva})}$ son

$$E(Y_h) \pm z_{1-\frac{\alpha}{2}} \cdot \sigma \quad (59)$$

4.11.2. Intervalo de predicción para $Y_{h(\text{nueva})}$ cuando los parámetros son desconocidos

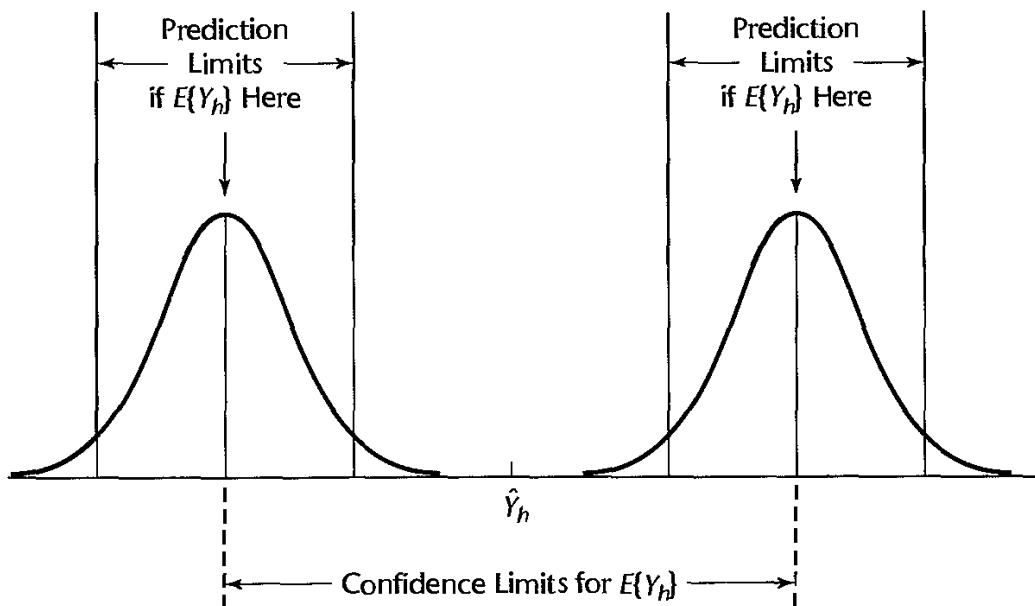
Cuando los parámetros de regresión son desconocidos, deben ser estimados. La media de la distribución de Y se estima por \hat{Y}_h , como de costumbre, y la varianza de la distribución de Y se estima por la MSRes. No podemos, sin embargo sólo utilizar los límites de la predicción de (59) con los parámetros reemplazados por los estimadores puntuales correspondientes. La razón de ello es ilustrada de manera

intuitiva en la Figura 42. En ella se muestran dos distribuciones de probabilidad de Y , que corresponde a los límites superior e inferior de un intervalo de confianza para $E(Y_h)$. En otras palabras, la distribución de Y puede ser ubicada tan a la izquierda como la distribución que se exhibe a la extrema izquierda, o tan a la derecha como la distribución que se exhibe a la extrema derecha, o en cualquier lugar en el medio. Dado que no sabemos la media $E(Y_h)$ y sólo la podemos estimar por un intervalo de confianza, no podemos estar seguros de la localización de la distribución de Y .

La Figura 42 también muestra los límites de predicción para cada una de las dos distribuciones de probabilidad de Y allí presentadas. Ya que no podemos estar seguros de la localización del centro de la distribución de Y , los límites de la predicción de $Y_{h(\text{nueva})}$ claramente deben tener en cuenta dos elementos, como se muestra en la Figura 42:

1. La variación en la posible ubicación de la (esperanza o centro de la) distribución de Y .
2. La variación dentro de la distribución de probabilidad de Y .

Figura 42: Predicción de $Y_{h(\text{nueva})}$ cuando los parámetros son desconocidos. Fuente: Kutner et al. [2005], pág 58.



Los límites de predicción para una nueva observación $Y_{h(\text{nueva})}$ en un determinado nivel \mathbf{X}_h se obtienen por medio del siguiente resultado

$$\frac{Y_{h(\text{nueva})} - \widehat{Y}_h}{s(\text{pred})} \sim t_{n-p} \quad (60)$$

Observemos que en el estadístico de Student utilizamos el estimador puntual \widehat{Y}_h en el numerador y no la verdadera media $E(Y_h)$ porque la media real se desconoce y no puede ser utilizada al hacer la predicción. El desvío estándar estimado de la predicción, $s(\text{pred})$, en el denominador se define por

$$\begin{aligned} s^2(\text{pred}) &= \text{MSRes} + \widehat{\text{Var}}(\widehat{Y}_h) \\ &= \text{MSRes} \cdot \left(1 + \mathbf{X}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h\right), \end{aligned}$$

de manera análoga a lo que habíamos calculado para el modelo de regresión lineal simple. A partir de dicho resultado, el intervalo de predicción de la $Y_{h(\text{nueva})}$ correspondiente a \mathbf{X}_h de nivel $1 - \alpha$ es

$$\begin{aligned} \widehat{Y}_h \pm t_{n-p, 1-\alpha/2} \cdot s(\text{pred}) \\ \widehat{Y}_h \pm t_{n-p, 1-\alpha/2} \cdot \sqrt{\text{MSRes} \cdot \left(1 + \mathbf{X}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h\right)} \end{aligned}$$

Observemos que el numerador del estadístico de Student (60) representa cuán lejos se desviará la nueva observación $Y_{h(\text{nueva})}$ de la media estimada \widehat{Y}_h basada en los n casos originales en el estudio. Esta diferencia puede ser vista como el error de predicción, con \widehat{Y}_h jugando el papel de la mejor estimación puntual del valor de la nueva observación $Y_{h(\text{nueva})}$. La varianza de este error de predicción puede ser fácilmente obtenida mediante la utilización de la independencia de la nueva observación, $Y_{h(\text{nueva})}$ y los n casos originales de la muestra en la que se basa \widehat{Y}_h .

$$\begin{aligned} \text{Var}(\text{pred}) &= \text{Var}(Y_{h(\text{nueva})} - \widehat{Y}_h) \\ &= \text{Var}(Y_{h(\text{nueva})}) + \text{Var}(\widehat{Y}_h) \\ &= \sigma^2 + \text{Var}(\widehat{Y}_h) \end{aligned}$$

Luego, la varianza del error de predicción $\text{Var}(\text{pred})$ tiene dos componentes:

1. La varianza de la distribución de Y en $\mathbf{X} = \mathbf{X}_h$, es decir, σ^2 .
2. La varianza de la distribución muestral de \widehat{Y}_h , es decir, $\text{Var}(\widehat{Y}_h)$.

Un estimador insesgado de $\text{Var}(\text{pred})$ es

$$s^2(\text{pred}) = \text{MSRes} + \widehat{\text{Var}}\left(\widehat{Y}_h\right).$$

Por supuesto, como este estimador es siempre mayor que $\widehat{\text{Var}}\left(\widehat{Y}_h\right)$, que aparece en el intervalo de confianza (57), el intervalo de predicción de la $Y_{h(\text{nueva})}$ correspondiente a \mathbf{X}_h de nivel $1 - \alpha$ siempre será más largo que el intervalo de confianza de nivel $1 - \alpha$ para $E(Y_h)$, la respuesta media esperada cuando las covariables son \mathbf{X}_h .

4.11.3. Ejemplo de cálculo de Intervalo de Confianza para $E(Y_h)$ y de un Intervalo de Predicción para $Y_{h(\text{nueva})}$

Aplicaremos estos dos resultados (cálculo de intervalo de confianza e intervalo de predicción) a un caso particular, usando los datos de bebés de bajo peso. Buscamos un intervalo de confianza para la media del perímetro cefálico de un bebé con 30 semanas de gestación y que pesó 1360g. al nacer, de nivel 0,95. El intervalo de confianza resulta ser

Tabla 25: Intervalos de confianza y predicción de nivel 0,95 para los datos de niños de bajo peso al nacer, para edad gestacional de 30 semanas y peso al nacer de 1360g.

```
> new<-data.frame(gestage=30, birthwt= 1360)
> predict.lm(ajuste2,new,interval="confidence")
      fit      lwr      upr
1 28.17871 27.81963 28.53778
> predict.lm(ajuste2,new,interval="prediction")
      fit      lwr      upr
1 28.17871 25.62492 30.73249
```

O, bien, operando a mano, la matriz de varianzas de los coeficientes beta da

```
> vcov(sal2)
            (Intercept)      gestage      birthwt
(Intercept) 2.4930607944 -9.986181e-02 3.714576e-04
gestage     -0.0998618122  4.522022e-03 -2.801056e-05
birthwt      0.0003714576 -2.801056e-05  3.983870e-07
```

Recordemos que $\widehat{Var}(\widehat{Y}_h)$ está definida en (56), luego

$$\begin{aligned}\widehat{Var}(\widehat{Y}_h) &= \mathbf{X}_h^t \widehat{Var}(\widehat{\beta}) \mathbf{X}_h \\ &= [1 \ 30 \ 1360] \begin{bmatrix} 2,4930607944 & -9,986181 \times 10^{-2} & 3,714576 \times 10^{-4} \\ -0,0998618122 & 4,522022 \times 10^{-3} & -2,801056 \times 10^{-5} \\ 0,0003714576 & -2,801056 \times 10^{-5} & 3,983870 \times 10^{-7} \end{bmatrix} \begin{bmatrix} 1 \\ 30 \\ 1360 \end{bmatrix} \\ &= 0,032731\end{aligned}$$

Como

$$\begin{aligned}t_{n-p,1-\alpha/2} &= t_{97,0,975} = 1,984723 \\ \widehat{Y}_h &= 8,3080 + 0,4487 \cdot 30 + 0,0047122 \cdot 1360 = 28,178\end{aligned}$$

resulta que el intervalo de confianza de nivel $1-\alpha = 0,95$ para $E(Y_h)$, la respuesta media esperada cuando las covariables son \mathbf{X}_h , es

$$\begin{aligned}\widehat{Y}_h &\pm t_{n-p,1-\alpha/2} \cdot \sqrt{\widehat{Var}(\widehat{Y}_h)} \\ 28,178 &\pm 1,984723 \cdot \sqrt{0,032731} \\ 28,178 &\pm 0,35907\end{aligned}$$

es decir

$$[27,819; \quad 28,537]$$

Por otro lado, el intervalo de predicción de la $Y_{h(\text{nueva})}$ correspondiente a \mathbf{X}_h de nivel $1-\alpha = 0,95$ es

$$\begin{aligned}\widehat{Y}_h &\pm t_{n-p,1-\alpha/2} \cdot s(\text{pred}) \\ \widehat{Y}_h &\pm t_{n-p,1-\alpha/2} \cdot \sqrt{\text{MSRes} + \widehat{Var}(\widehat{Y}_h)}\end{aligned}$$

Como

$$\text{MSRes} = 1,62,$$

el intervalo de predicción de la $Y_{h(\text{nueva})}$ resulta ser

$$\begin{aligned}28,178 &\pm 1,984723 \cdot \sqrt{1,62 + 0,032731} \\ 28,178 &\pm 2,5515\end{aligned}$$

es decir,

$$[25,62; \quad 30,730].$$

4.11.4. Precaución Respecto de Extrapolaciones Ocultas

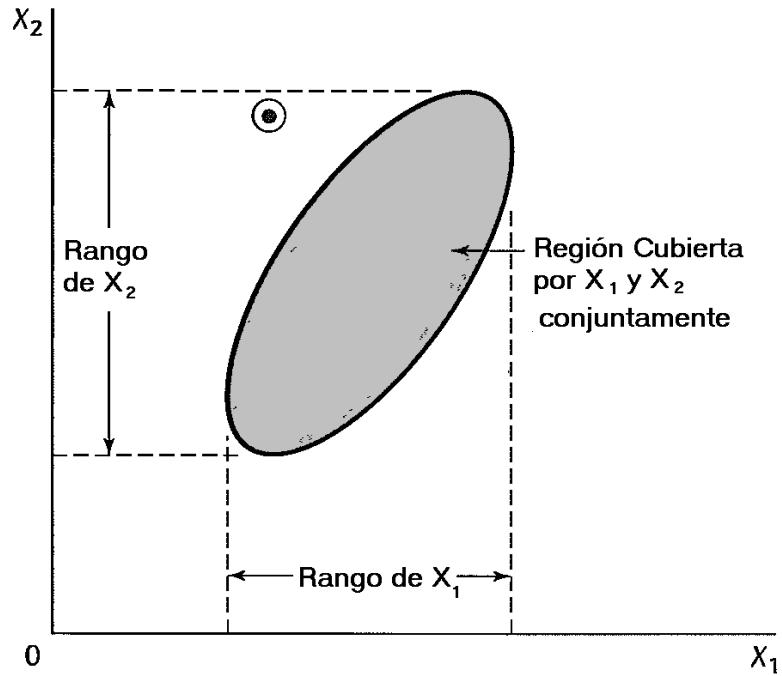
Al estimar una respuesta media o al predecir una nueva observación en la regresión múltiple, hay que tener especial cuidado de que la estimación o predicción esté comprendida dentro del alcance del modelo. El peligro, por supuesto, es que el modelo puede no ser apropiado cuando se lo extiende fuera de la región de las observaciones. En regresión múltiple, es particularmente fácil perder la noción de esta región ya que los niveles de X_1, \dots, X_{p-1} definen a la región en *forma conjunta*. Por lo tanto, uno no puede simplemente mirar los rangos de cada variable predictora de forma individual. Para visualizar el problema, consideremos la Figura 43, donde la región sombreada es la región de las observaciones para una regresión múltiple con dos variables de predicción y el punto con un círculo alrededor representa los valores (X_{h1}, X_{h2}) para los que se desea predecir la $Y_{h(\text{nueva})}$. Dicho punto está dentro de los rangos de las variables predictoras X_1 y X_2 en forma individual, sin embargo, está bien fuera de la región conjunta de las observaciones. Cuando sólo hay dos variables de predicción es fácil descubrir que se está frente a esta extrapolación, a través de un scatterplot (o gráfico de dispersión) pero esta detección se hace mucho más difícil cuando el número de variables predictivas es muy grande. Se discute en la Sección 5.2 un procedimiento para identificar las extrapolaciones ocultas cuando hay más de dos variables predictoras.

4.12. Ejercicios (primera parte)

Ejercicio 4.1 *Medidas del cuerpo V. Base de datos `bdim5` del paquete `openintro`.*

- (a) *En el ejercicio 2.1 explicamos el peso de las personas registradas en esta base de datos, por el contorno de la cadera y en el ejercicio 2.2 la explicamos con un modelo con la altura como covariable. Proponga un modelo de regresión múltiple que explique el peso medido en kilogramos (`wgt`) utilizando el contorno de la cadera medida en centímetros (`hip.gi`) y la altura media en centímetros (`hgt`) como covariables. Escriba el modelo que está ajustando. Realice el ajuste con el R.*
- (b) *Interprete los coeficientes estimados. ¿Resultan significativos? Cambian sus valores respecto de los que tenían los coeficientes que acompañaban a estas variables en los modelos de regresión lineal simple?*
- (c) *Evalúe la bondad del ajuste realizado, a través del R^2 . Indique cuánto vale y qué significa. Se quiere comparar este ajuste con el que dan los dos modelos lineales simples propuestos en los ejercicios 2.1 y 2.2. ¿Es correcto comparar los R^2 de los tres ajustes? ¿Qué valores puedo comparar? ¿Es mejor este ajuste múltiple?*

Figura 43: Región de observaciones en X_1 y X_2 conjuntamente, comparada con los rangos de X_1 y X_2 por separado.



- (d) Estime la varianza de los errores. Compare este estimador con los obtenidos en los dos ajustes simples.
- (e) Estime el peso esperado para la población de adultos cuyo contorno de cadera mide 100 cm y su altura es de 174cm. Dé un intervalo de confianza de nivel 0.95 para este valor esperado.
- (f) Prediga el peso de un adulto cuyo contorno de cadera mide 100 cm y su altura es de 174cm. Dé un intervalo de predicción de nivel 0.95 para este valor. Compare las longitudes de los tres intervalos de predicción que se obtienen usando el modelo que solamente tiene al contorno de cadera como explicativa, al que solamente usa la altura y al modelo múltiple que contiene a ambas.

4.13. Predictores Categóricos

Hasta ahora hemos visto el modelo de regresión lineal simple o múltiple con uno o varios predictores continuos. Sin embargo, tanto en regresión lineal simple como múltiple los predictores pueden ser variables binarias, categóricas, numéricas discretas o bien numéricas continuas.

4.13.1. Predictores Binarios

Comencemos con un ejemplo.

Los niveles de glucosa por encima de 125 mg/dL son diagnóstico de diabetes, mientras que los niveles en el rango de 100 a 125 mg/dL señalan un aumento en el riesgo de progresar a esta condición grave. Por lo tanto, es de interés determinar si la actividad física, una característica del estilo de vida que es modificable, podría ayudar a las personas a reducir sus niveles de glucosa y, por ende, evitar la diabetes. Responder a esta pregunta de manera concluyente requeriría un ensayo clínico aleatorizado, lo cual es a la vez difícil y costoso. Por ello, preguntas como estas son con frecuencia, inicialmente respondidas utilizando datos observacionales. Pero esto es complicado por el hecho de que las personas que hacen ejercicio físico difieren en muchos aspectos de las que no lo hacen, y algunas de las otras diferencias podrían explicar cualquier asociación (no ajustada) entre el ejercicio físico y el nivel de glucosa.

Usaremos un modelo lineal simple para predecir el nivel de glucosa usando una medida de la cantidad y frecuencia de ejercicio físico que realizan. La base de datos está en el archivo `azucar.txt`, se compone de los datos de $n = 220$ personas. Corresponde a datos artificialmente creados. La pregunta que queremos responder es si el hecho de hacer actividad física puede contribuir a bajar el nivel de glucosa y ayudar a prevenir la progresión a la diabetes entre las personas en riesgo.

Hay muchas maneras de codificar numéricamente las clases de una variable cualitativa. Usaremos variables indicadoras que valen 0 ó 1. Estas variables indicadoras son fáciles de usar y son ampliamente utilizadas, pero de ninguna manera son la única forma de cuantificar una variable cualitativa. En la Observación 4.12 comentamos una propuesta alternativa de codificación. Para el ejemplo, definimos la variable indicadora (o binaria, o dummy) por

$$X_{i1} = \begin{cases} 1 & \text{si el } i\text{ésimo paciente hace actividad física} \\ & \text{(al menos 3 veces por semana)} \\ 0 & \text{si no} \end{cases} \quad (61)$$

El modelo de regresión lineal para este caso es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

La función de respuesta para este modelo de regresión es

$$E(Y | X_1) = \beta_0 + \beta_1 X_1. \quad (62)$$

Para entender el significado de los coeficientes de regresión en este modelo, consideremos primero el caso de una persona que no hace ejercicio. Para tal persona, $X_1 = 0$, y la función de respuesta (62) se reduce a

$$E(Y) = \beta_0 + \beta_1 0 = \beta_0 \quad \text{no ejercita}$$

Para una persona que sí hace ejercicio, $X_1 = 1$, y la función de respuesta (62) se convierte en

$$E(Y) = \beta_0 + \beta_1 1 = \beta_0 + \beta_1 \quad \text{ejercita}$$

Luego, el modelo de regresión lineal en este caso consiste simplemente en expresar la media del nivel de glucosa en cada población mediante dos coeficientes distintos, donde β_0 es la media de la glucosa para las personas que no ejercitan y $\beta_0 + \beta_1$ es la media de la glucosa para las personas que ejercitan; por lo tanto, β_1 es la diferencia (positiva o negativa, dependiendo del signo) en niveles medios de glucosa para las personas que ejercitan respecto de las que no. Observemos que esto es consistente con nuestra interpretación más general de β_j como el cambio en $E[Y|X_j]$ por un aumento de una unidad de X_j . En este caso, si el ejercicio estuviera asociado con menores niveles de glucosa (como se presume) β_1 debería ser negativo.

En la Tabla 26 presentamos el resultado de ajustar el modelo propuesto a los datos. Los datos están en el archivo `azucar.txt`. Las variables se denominan `glucosa` (la respuesta) y `ejercicio` la explicativa.

El coeficiente estimado para la actividad física (`ejercicio`) muestra que los niveles basales de glucosa fueron alrededor de 7,4 mg/dL más bajos para personas que hacían ejercicios al menos tres veces por semana que para las personas que ejercitaban menos. Esta diferencia es estadísticamente significativa ($t = -6,309$, $p - \text{valor} = 1,5410^{-9} < 0,05$).

Sin embargo, en la base de datos considerada, las personas que hacen ejercicio resultaron ser en promedio, un poco más jóvenes, un poco más propensas a consumir alcohol, y, en particular, como puede verse en la Figura 44 tienen en promedio un menor índice de masa corporal (BMI), todos factores asociados con el nivel de glucosa. Esto implica que el promedio más bajo de la glucosa que observamos entre las personas que hacen ejercicio puede deberse al menos en parte, a diferencias en estos otros predictores. En estas condiciones, es importante que nuestra estimación de la diferencia en los niveles promedio de glucosa asociados con el ejercicio se “ajuste” a los efectos de estos factores de confusión potenciales de la asociación sin ajustar. Idealmente, el ajuste de un modelo de regresión múltiple (o sea, de múltiples predictores) proporciona una estimación del efecto de ejercitarse en el nivel medio de glucosa, manteniendo las demás variables constantes.

Tabla 26: Ajuste de la regresión para la variable glucosa con ejercicio como explicativa.

```
> ajuste1<-lm(glucosa ~ ejercicio, data = azucar)
> summary(ajuste1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	98.9143	0.8512	116.212	< 2e-16 ***
ejercicio	-7.4273	1.1773	-6.309	1.54e-09 ***

Residual standard error: 8.722 on 218 degrees of freedom
 Multiple R-squared: 0.1544, Adjusted R-squared: 0.1505
 F-statistic: 39.8 on 1 and 218 DF, p-value: 1.545e-09

Observación 4.11 ¿Qué pasa si ponemos dos variables binarias para modelar ejercicio? O sea, si definimos X_1 como antes,

$$X_{i1} = \begin{cases} 1 & \text{si la } i\text{ésima persona ejercita} \\ 0 & \text{si no} \end{cases}$$

y

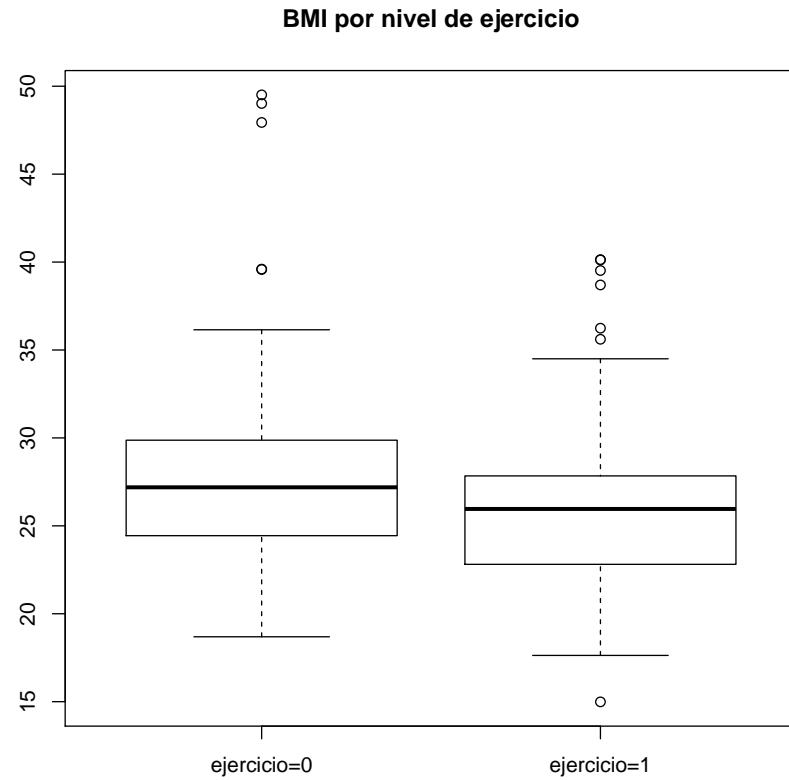
$$X_{i2} = \begin{cases} 1 & \text{si la } i\text{ésima persona no ejercita} \\ 0 & \text{si no} \end{cases}$$

Acá decimos que ejercita si hace actividad física más de tres veces por semana. Entonces el modelo sería

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (63)$$

Esta manera intuitiva de incorporar una variable indicadora para cada clase de la predictoras cualitativas, desafortunadamente, conduce a problemas tanto estadísticos (de identificación de parámetros) como computacionales. Para verlo, supongamos que tuviéramos $n = 4$ observaciones, las primeras dos compuestas por personas que ejercitan ($X_1 = 1, X_2 = 0$) y las dos segundas que no lo hacen

Figura 44: Boxplot del `bmi`, separados por niveles de la variable `ejercicio`, para los datos del archivo `azucar`.



$(X_1 = 0, X_2 = 1)$. Entonces la matriz X sería

$$X = \begin{bmatrix} X_1 & X_2 \\ \hline 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Observemos que la suma de las columnas X_1 y X_2 da la primera columna, de modo que las columnas de esta matriz son linealmente dependientes. Esto tiene un efecto

serio en la matriz $X^t X$.

$$X^t X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}$$

Vemos que la primer columna de la matriz $X^t X$ es igual a la suma de las últimas dos, de modo que las columnas son linealmente dependientes. Luego, la matriz $X^t X$ no tiene inversa, y por lo tanto, no se pueden hallar únicos estimadores de los coeficientes de regresión. De hecho, no hay unicidad tampoco en los parámetros del modelo (lo que en estadística se conoce como identificabilidad de los parámetros) puesto que la función de respuesta para el modelo (63) es

$$E(Y | X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} \beta_0 + \beta_1 & \text{si ejercita} \\ \beta_0 + \beta_2 & \text{si no ejercita} \end{cases}$$

En particular, tomando

$$\begin{aligned} \beta_0 &= a \\ \beta_1 &= b \\ \beta_2 &= c \end{aligned}$$

o bien

$$\begin{aligned} \beta_0 &= a - b \\ \beta_1 &= 2b \\ \beta_2 &= c \end{aligned}$$

resulta, en ambos casos

$$E(Y | X_1, X_2) = \begin{cases} a + b & \text{si ejercita} \\ a + c & \text{si no ejercita} \end{cases}$$

para cualesquiera números reales a, b, c . Una salida simple a este problema es desprenderte de una de las variables indicadoras. En nuestro ejemplo nos deshacemos de X_2 . Esta forma de resolver el problema de identificabilidad no es la única pero, como hemos visto, permite una interpretación sencilla de los parámetros. Otra posibilidad en este caso consiste en eliminar β_0 y proponer el modelo

$$E(Y | X_1, X_2) = \beta_1 X_1 + \beta_2 X_2 = \begin{cases} \beta_1 & \text{si ejercita} \\ \beta_2 & \text{si no ejercita} \end{cases}$$

Sin embargo, no la exploraremos ya que nuestra propuesta anterior es, no sólo satisfactoria sino también la más utilizada en el área.

Comparemos este modelo lineal con una sola regresora dicotómica con el test t para comparar las medias de dos poblaciones, a través de dos muestras independientes. Sean W_1, \dots, W_{n_1} variables aleatorias independientes idénticamente distribuidas con $E(W_i) = \mu_0$ e independientes de Z_1, \dots, Z_{n_2} que a su vez son variables aleatorias independientes entre sí e idénticamente distribuidas con $E(Z_i) = \mu_1$. El test t permite decidir entre las hipótesis

$$\begin{aligned} H_0 &: \mu_0 = \mu_1 \\ H_1 &: \mu_0 \neq \mu_1 \end{aligned}$$

donde $\mu_0 = E(Y | X_1 = 0)$ es decir, la esperanza de la glucosa para las personas que no ejercitan y $\mu_1 = E(Y | X_1 = 1)$ la esperanza de la glucosa para las personas que sí lo hacen. Recordemos que este test presupone que las observaciones de cada población tienen distribución normal con las medias μ_0 y μ_1 respectivamente, y la misma varianza (aunque desconocida). Para el conjunto de datos `azucar`, la salida de correr el test t figura en la Tabla 27.

Tabla 27: Test t para dos muestras normales independientes, datos `azucar`.

```
> t.test(glucosa ~ ejercicio, var.equal = TRUE, data = azucar)
```

```
Two Sample t-test

data: glucosa by ejercicio
t = 6.309, df = 218, p-value = 1.545e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.107071 9.747588
sample estimates:
mean in group 0 mean in group 1
 98.91429      91.48696
```

Recordemos que el estadístico del test es

$$\sqrt{n_1 + n_2} \frac{(\bar{W}_{n_1} - \bar{Z}_{n_2})}{S_p} \underset{\text{Bajo } H_0}{\sim} t_{n_1+n_2-2}$$

donde n_1 y n_2 son los tamaños de las muestras respectivas, y

$$S_p^2 = \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} (W_i - \bar{W}_{n_1})^2 + \sum_{j=1}^{n_2} (Z_j - \bar{Z}_{n_2})^2 \right]$$

es la varianza *poolada* o combinada de ambas muestras. Por otra parte, para el modelo (26), el test de $H_0 : \beta_1 = 0$ es también un test t , observemos que tanto el estadístico calculado como el p-valor son los mismos.

Finalmente podemos concluir que en el caso en el que el modelo lineal tiene una sola variable explicativa categórica, realizar el test de si el coeficiente que la acompaña es estadísticamente significativo es equivalente a utilizar un test t de comparación de medias entre dos poblaciones normales independientes, con igual varianza.

Dos observaciones con respecto a la codificación de la variable binaria dada en (61):

- Comparemos el valor de β_0 estimado en la Tabla 26 (que es $\hat{\beta}_0 = 98,9143$) con el promedio de la glucosa de las personas que no ejercitan (el grupo correspondiente a `ejercicio = 0`) calculado en la Tabla 27, que es 98,914, como anticipáramos. De igual modo, recuperamos el promedio de glucosa de las personas que ejercitan (91,487 en la Tabla 27) a partir de sumar $\hat{\beta}_0 + \hat{\beta}_1$ de la Tabla 26

$$\hat{\beta}_0 + \hat{\beta}_1 = 98,9143 - 7,4273 = 91,487.$$

- Codificando de esta forma, el promedio de la variable `ejercicio` da la proporción de personas que hacen ejercicio en la muestra, que son el 52,27% de la muestra, como puede comprobarse en la Tabla 28 que tiene los estadísticos descriptivos de la variable `ejercicio`.

Tabla 28: Estadísticos descriptivos de la variable `ejercicio`.

```
> summary(ejercicio)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.
0.0000 0.0000 1.0000 0.5227 1.0000 1.0000
```

Observación 4.12 Una alternativa comúnmente utilizada para la codificación de las variables binarias es ($1 = \text{sí}$, $2 = \text{no}$). Si definimos la variable X_3 con este código, el modelo es

$$E(Y | X_3) = \beta_0 + \beta_3 X_3.$$

Luego la función de respuesta para las personas que ejercitan ($X_3 = 1$) es

$$E(Y) = \beta_0 + \beta_3,$$

y para las que no ejercitan ($X_3 = 2$) es

$$E(Y) = \beta_0 + 2\beta_3.$$

Nuevamente, la diferencia entre ambas medias es el coeficiente β correspondiente, en este caso β_3 . Luego el coeficiente β_3 conserva su interpretación como la diferencia en el nivel medio de glucosa entre grupos, pero ahora entre las personas que no hacen ejercicio, comparadas con aquellas que sí lo hacen, una manera menos intuitiva de pensarlo. De hecho, β_0 sólo no tiene una interpretación directa, y el valor promedio de la variable binaria no es igual a la proporción de observaciones de la muestra que caen en ninguno de los dos grupos. Observar que, sin embargo, en general el ajuste del modelo, es decir, los valores ajustados, los errores estándares, y los p -valores para evaluar la diferencia de la glucosa en ambos grupos serán iguales con cualquier codificación.

4.13.2. Un predictor binario y otro cuantitativo

Incorporemos al modelo una variable cuantitativa. Tomaremos el índice de masa corporal que se denomina **bmi** (*body mass index*, medido en kg/m^2) en la base de datos,

$$X_{i2} = \text{BMI de la persona } i\text{ésima.}$$

El índice de masa corporal (BMI) es una medida de asociación entre el peso y la talla de un individuo ideada por el estadístico belga L. A. J. Quetelet, por lo que también se conoce como índice de Quetelet. Se calcula según la expresión matemática

$$BMI = \frac{\text{peso}}{\text{estatura}^2}$$

donde la masa o peso se expresa en kilogramos y la estatura en metros, luego la unidad de medida del BMI es kg/m^2 . En el caso de los adultos se ha utilizado como uno de los recursos para evaluar su estado nutricional, de acuerdo con los valores propuestos por la Organización Mundial de la Salud: a grandes rasgos se divide en tres categorías: delgadez (si $BMI < 18,5$), peso normal (cuando $18,5 \leq BMI < 25$) y sobrepeso (si $BMI \geq 25$), con subclasificaciones que contemplan los casos de infrapeso u obesidad.

Luego el modelo de regresión lineal múltiple que proponemos es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

O, si escribimos la función de respuesta (o sea, el modelo para la esperanza de Y) obtenemos

$$E(Y | X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \quad (64)$$

Interpretemos los parámetros. Para las personas que no hacen ejercicio ($X_1 = 0$) la función de respuesta es

$$E(Y) = \beta_0 + \beta_1 0 + \beta_2 X_2 = \beta_0 + \beta_2 X_2 \quad \text{no ejercita} \quad (65)$$

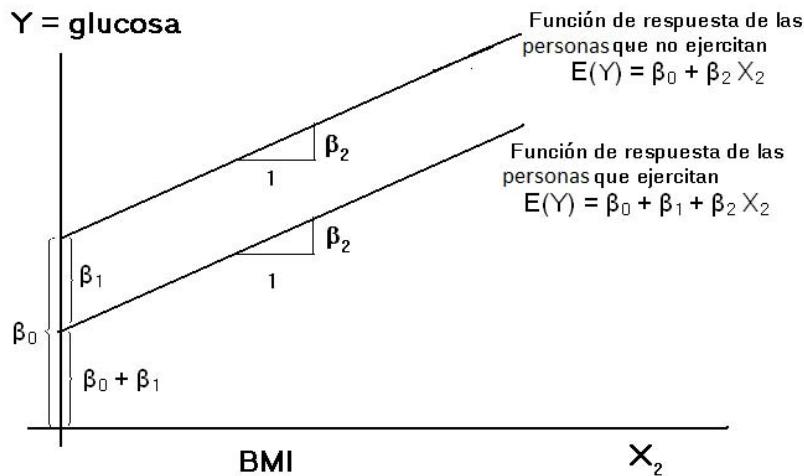
O sea, la función de respuesta para la glucosa media de las personas que no ejercitan es una línea recta con ordenada al origen β_0 y pendiente β_2 .

Para las que sí hacen ejercicio ($X_1 = 1$) la función de respuesta (64) se convierte en

$$E(Y) = \beta_0 + \beta_1 1 + \beta_2 X_2 = (\beta_0 + \beta_1) + \beta_2 X_2 \quad \text{ejercita} \quad (66)$$

Esta función también es una línea recta, con la misma pendiente β_2 pero con ordenada al origen $(\beta_0 + \beta_1)$. En la Figura 45 se grafican ambas funciones.

Figura 45: Significado de los coeficientes del modelo de regresión (64) con una variable indicadora X_1 de ejercicio y una variable continua $X_2 = \text{bmi}$ (datos azucar).



Enfoquémosnos en el significado de los coeficientes en la expresión (64) en el caso de las mediciones del nivel de glucosa. Vemos que el nivel medio de glucosa, $E(Y)$, es una función lineal del BMI (X_2) de la persona, con la misma pendiente β_2 para ambos tipos de personas. β_1 indica cuánto más baja (o más alta) es la función de respuesta para las personas que hacen ejercicio respecto de las que no lo hacen,

fijado el BMI. Luego β_1 mide el efecto diferencial por ejercitarse. Como el ejercicio debiera reducir el nivel de glucosa, esperamos que β_1 sea menor que cero y que la recta de valores de glucosa esperados para personas que ejercitan (66) esté por debajo de las que no lo hacen (65). En general, β_1 muestra cuánto más baja (o más alta) se ubica la recta de respuesta media para la clase codificada por 1 respecto de la recta de la clase codificada por 0, para cualquier nivel fijo de X_2 .

Tabla 29: Ajuste de la regresión para la variable `glucosa` con `ejercicio` y `bmi` como explicativas

```
> ajuste2<-lm(glucosa ~ ejercicio + bmi, data = azucar)
> summary(ajuste2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	84.4141	3.2336	26.105	< 2e-16 ***
ejercicio	-6.4879	1.1437	-5.673	4.46e-08 ***
bmi	0.5227	0.1128	4.633	6.20e-06 ***
<hr/>				

Residual standard error: 8.339 on 217 degrees of freedom

Multiple R-squared: 0.2305, Adjusted R-squared: 0.2234

F-statistic: 32.5 on 2 and 217 DF, p-value: 4.491e-13

En la Tabla 29 figura el ajuste del modelo propuesto. La función de respuesta ajustada es

$$\hat{Y} = 84,414 - 6,488X_1 + 0,523X_2.$$

Nos interesa medir el efecto de ejercitarse (X_1) en el nivel de glucosa en sangre. Para ello buscamos un intervalo de confianza del 95% para β_1 . Necesitamos el percentil 0,975 de la t de Student con $n - 3 = 217$ grados de libertad. Como $t(0,975, 217) = 1,970956 \simeq 1,959964 = z_{0,975}$, los límites para el intervalo de confianza resultan ser

$$-6,488 \pm 1,971 \cdot 1,1437$$

o sea,

$$\begin{aligned} -6,488 - 1,971 \cdot 1,1437 &\leq \beta_1 \leq -6,488 + 1,971 \cdot 1,1437 \\ -8,742 &\leq \beta_1 \leq -4.234 \end{aligned}$$

Podemos obtener este resultado directamente con R.

```
> confint(ajuste2)
              2.5 %    97.5 %
(Intercept) 78.0408659 90.7873793
ejercicio   -8.7421142 -4.2336953
bmi         0.3003302  0.7449921
```

Luego, con el 95 por ciento de confianza concluimos que las personas que ejercitan tienen un nivel de glucosa entre 4,23 y 8,74 mg/dL, **más bajo** que las que no lo hacen, en promedio, para un cada nivel de **bmi** fijo. Un test formal de

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

con nivel de significatividad de 0,05 nos conduciría a rechazar H_0 y aceptar H_1 , es decir, que el ejercicio tiene efecto cuando en el modelo incluimos el **bmi**, pues el intervalo de confianza del 95 % para β_1 no contiene al cero. Eso lo vemos también en la tabla de salida del paquete estadístico, en el p-valor de dicho coeficiente, que es $4,46 \cdot 10^{-8} < 0,05$.

Observación 4.13 ¿Por qué no ajustar dos regresiones lineales separadas (una para las personas que ejercitan y otra para las que no) en vez de hacer un ajuste con el total de datos? O sea, ajustar

$$E(Y | X_2) = \beta_0^{(0)} + \beta_2^{(0)} X_2 \quad \text{no ejercitan} \quad (67)$$

para las que no ejercitan y

$$E(Y | X_2) = \beta_0^{(1)} + \beta_2^{(1)} X_2 \quad \text{ejercitan} \quad (68)$$

para las que ejercitan. Hay dos razones para esto.

- El modelo (64) asume pendientes iguales en (67) y (68) y la misma varianza del error para cada tipo de persona. En consecuencia, la pendiente común β_2 se puede estimar mejor usando la información en la muestra conjunta. Ojo, este modelo **no** debería usarse si no se cree que este supuesto sea correcto para los datos a analizar.
- Usando el modelo (64) otras inferencias, como por ejemplo las realizadas sobre β_0 y β_1 resultarán más precisas pues se dispone de más observaciones para estimarlos y estimar a σ^2 (lo que se traduce en más grados de libertad en el MSRes). De todos modos, en este ejemplo donde hay doscientas observaciones, tenemos grados de libertad suficientes para proponer dos modelos si creyéramos que el modelo (64) no describe bien a los datos.

Observación 4.14 Los modelos de regresión múltiple en los que todas las variables explicativas son cualitativas se suelen denominar **modelos de análisis de la varianza (ANOVA)**. Los modelos que contienen algunas variables explicativas cuantitativas y otras variables explicativas cualitativas, para los que la variable explicativa de interés principal es cualitativa (por ejemplo, tipo de tratamiento que recibe el paciente) y las variables cuantitativas se introducen primariamente para reducir la varianza de los términos del error, se suelen denominar **modelos de análisis de la covarianza (ANCOVA)**.

4.14. Predictores Cualitativos con más de dos clases

4.14.1. Una sola predictora cualitativa con más de dos clases

Otra de las acciones que suelen recomendarse para evitar que las personas con niveles de glucosa alto entren en el diagnóstico de diabetes es bajar el 7% de su peso, o más. Para los datos del archivo `azucar` se registró la variable `peso.evo` que es una variable categórica que mide la evolución del peso del paciente en el último año. Tiene tres categorías: “*bajó de peso*”, si su peso disminuyó un 7% o más respecto del control médico anual anterior, “*su peso se mantuvo igual*”, cuando la diferencia entre ambos pesos difiere en menos de un 7% respecto del peso anterior, o bien “*aumentó de peso*”, si su peso actual aumentó un 7% o más respecto del registrado en el control médico anual anterior. La evolución del peso fue codificada en orden de 1 a 3, según las categorías recién definidas. Este es un ejemplo de una variable *ordinal* (con valores o categorías cuyo orden relativo es relevante, pero separados por incrementos que pueden no estar reflejados en forma precisa en la codificación numérica asignada). Las categorías de la variable `peso.evo` figuran en la Tabla 30.

Tabla 30: Niveles de la variable `peso.evo`, que codifica la evolución del peso en el último año.

Categorías de peso.evo	codificación original
Disminuyó de peso en el último año (7% o más)	1
Pesa igual que hace un año (menos de 7% de diferencia)	2
Aumentó de peso en el último año (7% o más)	3

Las variables categóricas de más de dos niveles también puede ser *nominales*, en el sentido que no haya un orden intrínseco en las categorías. Etnia, estado civil, ocupación y región geográfica son ejemplos de variables nominales. Con las variables nominales es aún más claro que la codificación numérica usada habitualmente

para representar a la variable en la base de datos no puede ser tratada como los valores de una variable numérica como nivel de glucosa en sangre.

Las categorías se suelen crear para ser mutuamente excluyentes y exhaustivas, por lo que cada miembro de la población se encuentra en una y sólo una categoría. En este sentido, tanto las categorías ordinales como las nominales definen subgrupos de la población.

Es sencillo acomodar ambos tipos de variables tanto en la regresión lineal múltiple como en otros modelos de regresión, usando variables indicadoras o *dummies*. Como en las variables binarias, donde dos categorías se representan en el modelo con una sola variable indicadora, las variables categóricas con $K \geq 2$ niveles se representan por $K - 1$ indicadoras, una para cada nivel de la variable, excepto el nivel de referencia o basal. Supongamos que elegimos el nivel 1 como nivel de referencia. Entonces para $k = 2, 3, \dots, K$, la k -ésima variable indicadora toma el valor 1 para las observaciones que pertenecen a la categoría k , y 0 para las observaciones que pertenecen a cualquier otra categoría. Observemos que para $K = 2$ esto también describe el caso binario, en el cual la respuesta “*no*” define el nivel basal o de referencia y la variable indicadora toma el valor 1 sólo para el grupo “*sí*”.

Traduzcamos todo al ejemplo. Como la variable ordinal `peso.evo` tiene 3 categorías, necesitamos definir 2 variables dummies. Las llamamos `Ievo2` e `Ievo3`. En la Tabla 31, observamos los valores para las dos variables indicadoras correspondientes a la variable categórica `peso.evo`. Cada nivel de `peso.evo` queda definido por una combinación única de las dos variables indicadoras.

Tabla 31: Codificación de las variables indicadoras para una variable categórica multinivel

peso.evo	Variables indicadoras		
	Ievo2	Ievo3	Categoría
1	0	0	bajó de peso
2	1	0	mantuvo su peso
3	0	1	aumentó de peso

Por el momento consideremos un modelo simple en el cual los tres niveles de `peso.evo` sean los únicos predictores. Entonces

$$E(Y | X) = \beta_0 + \beta_2 Ievo2 + \beta_3 Ievo3 \quad (69)$$

donde X representa las dos variables dummies recién definidas, es decir,

$$X = (Ievo2, Ievo3).$$

Para tener mayor claridad, en (69) hemos indexado a los β' s en concordancia con los niveles de `peso.evo`, de modo que β_1 no aparece en el modelo. Si dejamos que las dos indicadoras tomen el valor 0 ó 1 de manera de definir los tres (¿por qué no cuatro?) niveles de `peso.evo`, obtenemos

$$E(Y | X) = \begin{cases} \beta_0 & \text{si } \text{peso.evo} = 1, \text{ o sea } \text{Ievo2} = 0 \text{ e } \text{Ievo3} = 0 \\ \beta_0 + \beta_2 & \text{si } \text{peso.evo} = 2, \text{ o sea } \text{Ievo2} = 1 \text{ e } \text{Ievo3} = 0 \\ \beta_0 + \beta_3 & \text{si } \text{peso.evo} = 3, \text{ o sea } \text{Ievo2} = 0 \text{ e } \text{Ievo3} = 1 \end{cases} \quad (70)$$

De (70) es claro que β_0 , la ordenada al origen, da el valor de $E(Y | X)$ en el grupo de referencia, el grupo “bajó de peso”, o `peso.evo` = 1. Entonces es sólo cuestión de restarle a la segunda línea la primera línea para ver que β_2 da la diferencia en el promedio de glucosa en el grupo “mantuvo su peso” (`peso.evo` = 2) comparado con el grupo “bajó de peso”. De acuerdo con esto, el test de $H_0 : \beta_2 = 0$ es un test para chequear si los niveles medios de glucosa son los mismos en los dos grupos “bajó de peso” y “mantuvo su peso” (`peso.evo` = 1 y 2). Y de manera similar para β_3 .

Podemos hacer unas cuantas observaciones a partir de (70).

- Sin otros predictores, o covariables, el modelo es equivalente a un ANOVA de un factor (one-way ANOVA). También se dice que el modelo está *saturado* (es decir, no impone estructura alguna a las medias poblacionales) y las medias de cada grupo de la población se estimarán bajo el modelo (70) por el promedio de las muestras correspondientes. Con covariables, las medias estimadas para cada grupo se ajustarán a las diferencias entre grupos en las covariables incluidas en el modelo.
- Los parámetros del modelo (y por lo tanto las dummies que los acompañan) pueden ser definidos para que sean iguales a la media poblacional de cada grupo o, sino, para que sean las diferencias entre las medias poblacionales de dos grupos distintos, como en (70). Por ejemplo, la diferencia en los niveles medios de la variable Y entre los grupos “aumentó de peso” (`peso.evo` = 3) y “mantuvo su peso” (`peso.evo` = 2) está dada por $\beta_3 - \beta_2$ (chequearlo). Todos los paquetes estadísticos permiten calcular de manera directa estimadores y tests de hipótesis acerca de estos *contrastos lineales*. Esto implica que la elección del grupo de referencia es, en algún sentido, arbitraria. Mientras que alguna elección en particular puede ser la mejor para facilitar la presentación, posiblemente porque los contrastes con el grupo de referencia seleccionado sean los de mayor interés, cuando se toman grupos de referencia alternativos, esencialmente se está definiendo el mismo modelo.

Tabla 32: Ajuste de regresión lineal múltiple para explicar a la variable glucosa con la evolución del peso como categórica, Ievo (datos de la base azucar). El R produce las dos binarias de forma automática (Ievo2 e Ievo3).

```
> Ievo<-factor(peso.evo) #convierte a la variable en factor
> contrasts(Ievo)       #da la codificacion
  2 3
1 0 0
2 1 0
3 0 1
> ajuste3<-lm(glucosa ~ Ievo)
> summary(ajuste3)
```

Coefficients:

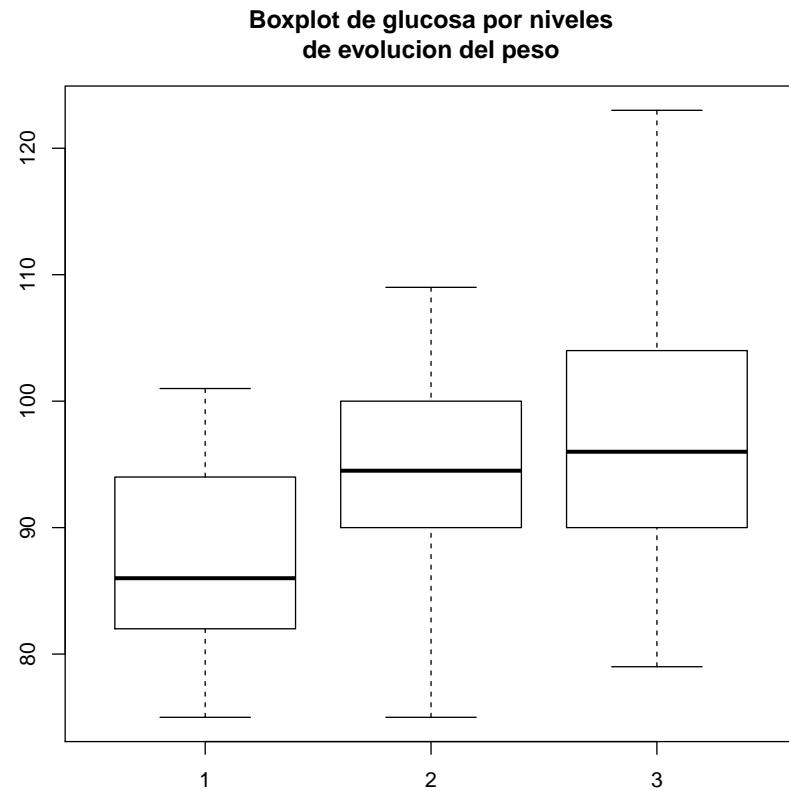
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	88.273	1.564	56.449	< 2e-16 ***
Ievo2	6.283	1.888	3.327	0.00103 **
Ievo3	8.997	1.774	5.072	8.45e-07 ***

Residual standard error:	8.983	on 217 degrees of freedom		
Multiple R-squared:	0.1071,		Adjusted R-squared:	0.09884
F-statistic:	13.01	on 2 and 217 DF,	p-value:	4.607e-06

La Tabla 32 muestra los resultados para el modelo con `peso.evo` tratada como una variable categórica, utilizando de nuevo los datos del archivo `azucar`. La estimación de $\hat{\beta}_0$ es 88,273 mg / dL, esta es la estimación del nivel de glucosa medio para el grupo que bajó de peso (grupo de referencia). Las diferencias entre los niveles de glucosa del grupo de referencia y los otros dos grupos (de distinta evolución del peso) resultan ser estadísticamente significativas; como todas dan positivas indican que la glucosa estaría relacionada con la evolución del peso. Por ejemplo, el nivel promedio de glucosa en el grupo “subió de peso” (`Ievo3`) es 8,997 mg / dL mayor que la del grupo “bajó de peso” (`peso.evo = 1`) ($t = 5,072$, p -valor = $8,45 \cdot 10^{-7}$). En la Figura 46 vemos un boxplot de los datos de glucosa separados según sus niveles de `peso.evo`, donde se aprecia esta diferencia.

Es de interés testear si la variable `peso.evo` sirve para explicar al nivel de glucosa. Para evaluarla en su conjunto se utiliza el test F que describiremos en la Sección 4.14.4. Antes de hacerlo discutamos otra manera de introducir a la variable `peso.evo` en el modelo.

Figura 46: Boxplot de los datos de glucosa, según sus niveles de peso.evo.



4.14.2. Variables indicadoras versus variables numéricas

Una alternativa al uso de variables indicadoras de una variable de predicción cualitativa es pensarla como numérica. En el ejemplo de la glucosa, podríamos utilizar una única variable predictora Z y asignar valores 1,2 y 3 a las clases, como se describe en la Tabla 33.

Los valores numéricos son, por supuesto, arbitrarios y podrían ser cualquier otro conjunto de números. El modelo en este caso sería

$$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i \quad (71)$$

La principal dificultad en tomar a las variables categóricas como numéricas es que la numeración otorgada a las categorías distintas define una métrica (una distancia) entre las clases de la variable cualitativa que puede no resultar razonable para

Tabla 33: Variable categórica mirada como numérica

peso.evo	Z
Bajó de peso	1
Mantuvo su peso	2
Aumentó de peso	3

modelizar. Veámoslo en el ejemplo. Escribimos la función de respuesta media con el modelo (71) para las tres clases de la variable cualitativa

$$E(Y | Z) = \begin{cases} \beta_0 + \beta_1 & \text{si } \text{peso.evo} = 1 \\ \beta_0 + 2\beta_1 & \text{si } \text{peso.evo} = 2 \\ \beta_0 + 3\beta_1 & \text{si } \text{peso.evo} = 3 \end{cases}$$

Notemos la implicación clave de este modelo:

$$\begin{aligned} E(Y | \text{peso.evo} = 2) - E(Y | \text{peso.evo} = 1) \\ = E(Y | \text{peso.evo} = 3) - E(Y | \text{peso.evo} = 2) \\ = \beta_1 \end{aligned}$$

Luego, la codificación 1 a 5 implica que pensamos que la respuesta media cambia **en la misma cantidad** cuando pasamos de `peso.evo=1` a `peso.evo=2` o de `peso.evo=2` a `peso.evo=3`. Esto puede no estar en coincidencia con la realidad y resulta de la codificación 1 a 3 que asigna igual distancia entre los 3 tipos de evolución del peso. Por supuesto, con distintas codificaciones podemos imponer espaciamientos diferentes entre las clases de la variable cualitativa pero esto sería siempre arbitrario.

En contraposición, el uso de variables indicadoras no hace supuestos sobre el espacioamiento de las clases y descansa en los datos para mostrar los efectos diferentes que ocurren. En el caso del modelo (70) no se impone ningún patrón o vínculo entre sí a las cinco medias de los grupos definidos por la variable categórica, tanto en el modelo sin covariables como si las tuviera. Aquí β_2 da la diferencia en el promedio de glucosa en el grupo `peso.evo=2` comparado con el grupo `peso.evo=1`, y β_3 da la diferencia en el promedio de glucosa en el grupo `peso.evo=3` comparado con el grupo `peso.evo=1` y $\beta_3 - \beta_2$ da la diferencia en el promedio de glucosa en el grupo `peso.evo=3` comparado con el grupo `peso.evo=2`. Observemos que no hay restricciones arbitrarias que deban cumplir estos tres efectos. En cambio, si la variable `peso.evo` fuera tratada como una variable numérica que toma valores de 1 a 3, las esperanzas poblacionales de cada grupo se verían obligadas a yacer en una línea recta. En síntesis: para variables categóricas es preferible usar la codificación que proporcionan las variables dummies.

4.14.3. Variables numéricas como categóricas

Algunas veces, aún cuando las variables son originalmente cuantitativas se las puede incluir en un modelo como categóricas. Por ejemplo, la variable cuantitativa edad se puede transformar agrupando las edades en las categorías: menor de 21, 21 a 34, 35 a 49, etc. En este caso, se usan variables indicadoras o dummies para las clases de este nuevo predictor. A primera vista, este enfoque parece cuestionable, ya que la información sobre la edad real se pierde. Además, se ponen parámetros adicionales en el modelo, lo que conduce a una reducción de los grados de libertad asociados con el MSRes.

Sin embargo, hay ocasiones en las que la sustitución de una variable cuantitativa por indicadoras puede ser apropiado. Por ejemplo, cuando se piensa que la relación entre la respuesta y la explicativa puede no ser lineal (en el caso en que la glucosa aumentara tanto para personas muy jóvenes o muy grandes) o en una encuesta a gran escala, donde la pérdida de 10 ó 20 grados de libertad es irrelevante. En una primera etapa exploratoria, donde se está muy en duda acerca de la forma de la función de regresión, puede ajustarse un modelo como (70) en una primera etapa exploratoria, y luego, en virtud de lo observado, incluir a la variable (o una transformación de ella) como numérica.

Esto es de hecho lo que se hizo con la variable evolución del peso: a partir de dos variables numéricas medidas sobre el mismo paciente (el peso en un año y el peso en el siguiente) se construyó una variable categórica.

4.14.4. El test F

A pesar de que todos los contrastes entre los niveles de una variable explicativa categórica están disponibles para ser estimados y comparados luego de ajustar un modelo de regresión, los test t para estas comparaciones múltiples en general no proporcionan una evaluación conjunta de la importancia de la variable categórica para predecir a la variable respuesta, o más precisamente no permiten realizar un único test de la hipótesis nula de que el nivel medio de la variable respuesta es el mismo para todos los niveles de este predictor. En el ejemplo, esto es equivalente a un test de si alguno de los dos coeficientes correspondientes a I_{evo2} o I_{evo2} difieren de cero. El resultado que aparece en la Tabla 32 ($F_{obs} = 13,01$, con 2 grados de libertad en el numerador y 217 en el denominador, $p\text{-valor} = 4,6 \cdot 10^{-6} < 0,05$) muestra que los niveles medios de glucosa son claramente diferentes entre los grupos definidos por peso.evo . Las hipótesis que chequea este test en este caso son

$$H_0 : \beta_2 = \beta_3 = 0 \tag{72}$$

$$H_1 : \text{al menos uno de los } \beta_i \text{ con } i \text{ entre } 2 \text{ y } 3 \text{ es tal que } \beta_i \neq 0$$

En este caso se rechaza la hipótesis nula ($p\text{-valor} = 4,6 \cdot 10^{-6} < 0,05$) y se concluye que no todos los β_i con i entre 2 y 3 son simultáneamente iguales a cero. Luego la evolución del peso es útil para predecir el nivel de glucosa. En general este resultado puede leerse en la tabla de ANOVA del ajuste.

Es por este motivo que conviene ingresar en la base de datos a la variable `peso.evo` con sus tres niveles y pedirle al software que compute las dos variables dicotómicas, en vez de ponerlas a mano en el archivo, pues en tal caso no hay cómo decirle al paquete que las dos variables están vinculadas de esta forma.

4.14.5. Comparaciones Múltiples

Una vez que, a través del test F , logramos concluir que la variable categórica es significativa para explicar a la respuesta, aparece el interés de comparar la media de la variable respuesta para los grupos definidos por los distintos niveles de la variable categórica. Es decir, interesa comparar las medias de dos grupos, digamos μ_j y μ_k . Esto suele hacerse estudiando si la diferencia entre ellos $\mu_j - \mu_k$ es distinta de cero. Tal diferencia entre los niveles medios de una variable categórica (o factor) se denomina una comparación de a pares. Cuando una variable categórica toma K niveles, el número de comparaciones de a pares que pueden hacerse es $\binom{K}{2} = \frac{K(K-1)}{2}$.

La salida que da el `summary` del `lm` en R que analizamos antes, por ejemplo en la Tabla 32, o la salida estándar que proporciona cualquier paquete estadístico a un ajuste lineal, tiene, en este sentido dos limitaciones importantes.

1. Los p-valores que aparecen en la columna de la derecha son válidos para cada comparación individual.
2. Cuando la variable categórica tiene más de dos niveles, dicha tabla no nos da información de todas las comparaciones de a pares de forma directa. En el ejemplo de `azucar`, vemos en la Tabla 32 que nos falta la comparación entre los niveles 2 y 3 de la variable evolución de peso.

Con la primera limitación veníamos trabajando desde el modelo lineal simple, pero en el caso de regresión con covariables categóricas se hace particularmente seria por la gran cantidad de comparaciones que tienen interés para el experimentador. Cuando se realizan varios tests con los mismos datos, tanto el nivel de significatividad como la potencia de las conclusiones acerca de la familia de tests se ve afectada. Consideremos por ejemplo, la realización de tres tests de t , cada uno a nivel $\alpha = 0,05$, para testear las hipótesis

$$\begin{aligned} H_0^{(1)} &: \mu_2 - \mu_1 = 0 \text{ versus } H_1^{(1)} : \mu_2 - \mu_1 \neq 0 \\ H_0^{(2)} &: \mu_3 - \mu_1 = 0 \text{ versus } H_1^{(2)} : \mu_3 - \mu_1 \neq 0 \\ H_0^{(3)} &: \mu_3 - \mu_2 = 0 \text{ versus } H_1^{(3)} : \mu_3 - \mu_2 \neq 0 \end{aligned}$$

La probabilidad de que los tres tests concluyan que las tres hipótesis nulas H_0 son verdaderas cuando en realidad las tres H_0 son verdaderas, asumiendo independencia de los tests, será

$$0,95^3 = 0,857.$$

Luego, la probabilidad de concluir H_1 para al menos una de las tres comparaciones es $1 - 0,857 = 0,143$ en vez de 0,05. Vemos que el nivel de significatividad de una familia de tests no es el mismo que para un test individual. Lo mismo pasa para los intervalo de confianza.

El objetivo de hacer estas comparaciones múltiples de manera justa es mantener el error de tipo I acotado, sin inflarlo por sacar muchas conclusiones con el mismo conjunto de datos. Es decir, queremos un test de nivel 0,05 para las hipótesis

$$H_0 : \begin{cases} \mu_2 - \mu_1 = 0 \\ \mu_3 - \mu_1 = 0 \\ \mu_3 - \mu_2 = 0 \end{cases} \text{ versus } H_1 : \text{alguna de las 3 igualdades no vale.}$$

Con la notación de regresión lineal múltiple, podemos reescribir a la hipótesis nula del siguiente modo

$$H_0 : \begin{cases} E(Y | \text{peso.evo} = 2) - E(Y | \text{peso.evo} = 1) = 0 \\ E(Y | \text{peso.evo} = 3) - E(Y | \text{peso.evo} = 1) = 0 \\ E(Y | \text{peso.evo} = 3) - E(Y | \text{peso.evo} = 2) = 0 \end{cases}$$

Para eso, primero hay que mirar el resultado del test conjunto F que evalúa la significatividad conjunta de la variable categórica para explicar a la respuesta. Si este test no resulta significativo, suele descartarse la variable categórica de entre las covariables de interés, y se la excluye del modelo. Si este test resulta estadísticamente significativo, entonces suelen mirarse con más detalle cuáles de las comparaciones entre grupos son estadísticamente significativas, para proporcionar un mejor análisis de los datos en consideración. Hay diversas propuestas para llevar estas comparaciones a cabo, de acuerdo esencialmente a cuáles son las comparaciones que resultan más interesantes al experimentador.

¿Qué pasa si alguna de las comparaciones llevadas a cabo resulta no significativa? ¿Conviene redefinir las categorías? La recomendación general es que si esto pasa, de todos modos conviene mantener las categorías originales puesto que de esta forma se estimará mejor a la varianza σ^2 de los errores (resultará menor), y por lo tanto las conclusiones que se obtendrán serán más potentes. Además, cuando el interés esté puesto en la conclusión respecto de una comparación entre dos categorías en particular, el hecho de mantener las grupos originales permitirá clasificar bien a cada observación permitiendo mantener clara la diferencia que se está buscando establecer. No conviene recodificar a posteriori del análisis, es mejor dejar

todos los niveles de la variable categórica en el modelo, aunque algunos resulten no significativos.

Hay distintos métodos disponibles para controlar la tasa de error de tipo I. La comparación múltiple propuesta por Tukey, la *diferencia honestamente significativa* de Tukey (HSD: *honestly significant difference*) es un procedimiento que permite hacer todas las comparaciones con nivel conjunto prefijado. El método de Scheffé también provee un procedimiento que asegura el nivel de todas las comparaciones posibles, incluso asegura el nivel de cualquier combinación lineal de los parámetros (no necesariamente una resta) que pueda interesar. Si sólo interesan comparaciones de a pares, Tukey proporciona intervalos más angostos, y su método es preferible. Los intervalos sólo difieren en el percentil utilizado. En las Tablas 34 y 35 aparecen las salidas de las comparaciones de Tukey para los datos de `azucar` realizado en R utilizando dos comandos diferentes para hacerlo: `TukeyHSD` y `glht`, este último del paquete `multcomp`. En ella vemos que el nivel medio de glucosa del grupo “bajó de peso” difiere significativamente tanto del grupo “mantuvo su peso igual” como del grupo “aumentó de peso”, así como también vemos que el nivel medio de glucosa de estos dos últimos grupos no difiere (significativamente) entre sí.

En la Sección 4.9.3 presentamos el procedimiento de comparaciones múltiples de Bonferroni. También es aplicable para el contexto de ANOVA, ya sea que interesen las comparaciones de a pares, o combinaciones lineales de los coeficientes mientras estos se hayan fijado **con anterioridad** a hacer el análisis de datos. Otros métodos que pueden aplicarse cuando las comparaciones que interesan tienen características específicas son la *mínima diferencia significativa* de Fisher (LSD: *least significant difference*), el método de Sidak o el procedimiento de Dunnett. Puede verse Seber y Lee [1977] o Kutner et al. [2005] para más detalle sobre las comparaciones múltiples.

4.15. Una predictora cualitativa y una numérica

Ajustemos ahora un modelo de regresión lineal múltiple con una covariable numérica y una categórica. Siguiendo con los datos de la glucosa, proponemos ajustar un modelo donde aparezcan `peso.evo` y `bmi` como variables explicativas, donde la primera es categórica (como ya vimos, la incluimos en el modelo como las 2 dummies definidas por `Ievo`) y la segunda es continua. Proponemos ajustar el siguiente modelo

$$E(Y | X) = \beta_0 + \beta_2 I_{evo2} + \beta_3 I_{evo3} + \beta_{BMI} bmi \quad (73)$$

En este caso, $X = (I_{evo2}, I_{evo3}, bmi)$. Para entender este modelo, nuevamente dejamos que las indicadoras tomen el valor 0 ó 1 de manera de definir los tres

Tabla 34: Intervalos de confianza de nivel simultáneo para la variable `glucosa` para los distintos niveles de evolución del peso, `Ievo` (datos de la base `azucar`), usando el comando `TukeyHSD` del R.

```
> tu<-TukeyHSD(aov(glucosa ~ Ievo),"Ievo")
> tu
  Tukey multiple comparisons of means
    95% family-wise confidence level

$Ievo
      diff      lwr      upr      p adj
2-1 6.282828 1.8263489 10.739308 0.0029505
3-1 8.996838 4.8103925 13.183283 0.0000025
3-2 2.714010 -0.4718461 5.899865 0.1121165
```

niveles de `peso.evo`, y obtenemos

$$E(Y | X) = \begin{cases} \beta_0 + \beta_{BMI} \text{bmi} & \text{si } \text{peso.evo} = 1 \\ \beta_0 + \beta_2 + \beta_{BMI} \text{bmi} & \text{si } \text{peso.evo} = 2 \\ \beta_0 + \beta_3 + \beta_{BMI} \text{bmi} & \text{si } \text{peso.evo} = 3 \end{cases}$$

es decir, que este modelo propone ajustar una recta distinta para la glucosa media de cada grupo, **todas con igual pendiente** que en este caso hemos denominado β_{BMI} , y tres ordenadas al origen diferentes, una por cada grupo. Como vemos, estamos ajustando tres rectas paralelas. Acá β_2 indica cuánto aumenta (o disminuye, dependiendo del signo) el valor medio de glucosa para las personas cuyo nivel de evolución del peso es 2 (las personas “que mantuvieron su peso”) respecto de aquellas cuyo nivel de evolución del peso es 1 (las personas que “bajaron de peso”). En la Figura 47 puede verse el gráfico que proponemos para el valor esperado de la glucosa en función de la evolución del peso y del BMI. Como esperamos que a medida que la evolución del peso aumente (o sea, a medida que el paciente aumente de peso) el nivel de glucosa aumente, hemos acomodado las rectas de manera que vayan aumentando al aumentar la variable que codifica esta evolución. Así mismo, es de esperar que a mayor BMI aumente el nivel de glucosa, por eso en el dibujo proponemos una pendiente (común a todos los grupos) positiva, como ya vimos que pasaba en el ajuste anterior.

La Tabla 36 exhibe el modelo ajustado.

En este caso vemos que cuando incorporamos la variable BMI al modelo, todos los coeficientes asociados a la variable `peso.evo` siguen siendo significativos. El test de, por ejemplo, $H_0 : \beta_2 = 0$ da significativo ($t = 3,82$, p -valor = 0,000177)

Tabla 35: Tests e intervalos de confianza de nivel simultáneo para la variable glucosa para los distintos niveles de evolución del peso, Ievo (datos de la base azucar), usando el comando `glht` de la librería `multcomp` de R.

```

> ajuste3<-lm(glucosa ~ Ievo)
> library(multcomp) #para testear combinaciones de los parametros
> tu.otro <- glht(ajuste3, linfct = mcp(Ievo = "Tukey"))
> summary(tu.otro)

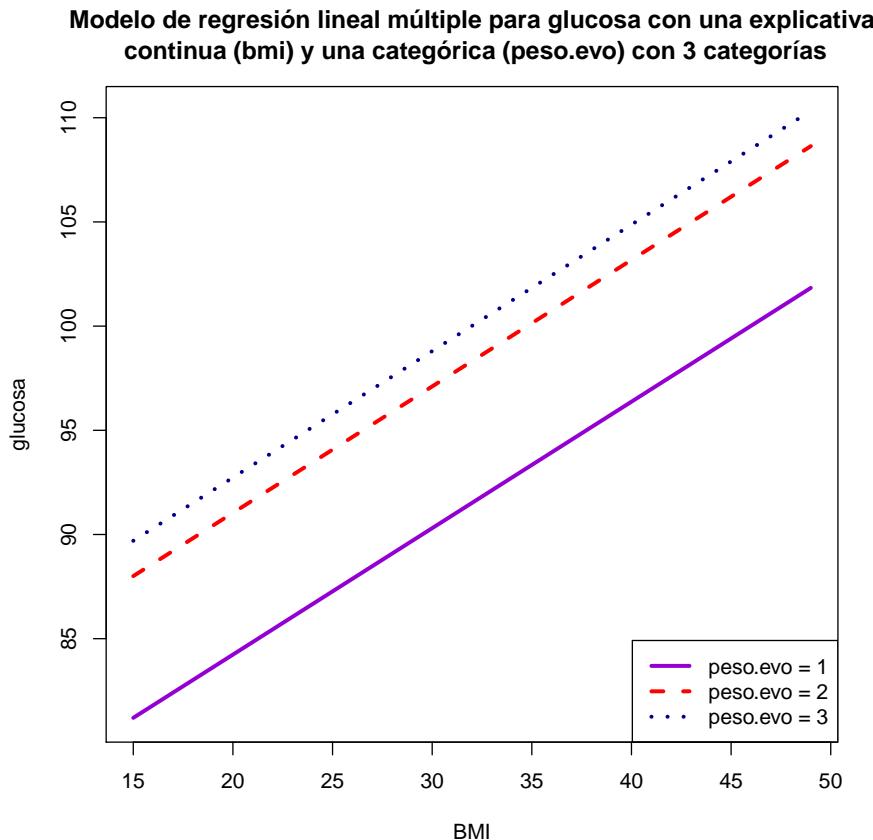
   Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = glucosa ~ Ievo)

Linear Hypotheses:
Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0     6.283     1.888    3.327  0.00286 **
3 - 1 == 0     8.997     1.774    5.072  < 1e-04 ***
3 - 2 == 0     2.714     1.350    2.010  0.10977
---
> confint(tu.otro)

   Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Quantile = 2.35
95% family-wise confidence level
Linear Hypotheses:
Estimate lwr      upr
2 - 1 == 0  6.2828  1.8450 10.7206
3 - 1 == 0  8.9968  4.8280 13.1657
3 - 2 == 0  2.7140 -0.4585  5.8865

```

Figura 47: Modelo propuesto para explicar la glucosa con una covariante explicativa categórica (`peso.evo`) con tres niveles y otra continua (`bmi`).



indicando que hay diferencia significativa en los niveles medios de glucosa para personas que no bajaron de peso con respecto a las que sí bajaron (grupo basal). Lo mismo sucede al testear la comparación entre la glucosa esperada del grupo que aumentó de peso y el que bajó de peso, cuando en el modelo se ajusta por BMI ($t = 5,074$, p -valor $= 8,38 \cdot 10^{-7}$). Es decir que los niveles medios de glucosa en los distintos grupos definidos por la evolución del peso difieren del basal. Además, como sus coeficientes estimados crecen al aumentar el peso, vemos que los valores estimados son consistentes con lo que bosquejamos a priori en la Figura 47. Antes de comparar los niveles medios de los distintos grupos entre sí observemos que si queremos evaluar a la variable `peso.evo` en su conjunto, debemos recurrir a un test F que evalue las hipótesis (72), cuando además en el modelo aparece BMI

Tabla 36: Regresión de glucosa en las regresoras: peso.evo (categórica) y bmi (numérica).

```
> ajuste4<-lm(glucosa ~ Ievo + bmi)
> summary(ajuste4)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	72.0993	3.3764	21.354	< 2e-16 ***
Ievo2	6.8023	1.7823	3.817	0.000177 ***
Ievo3	8.4963	1.6744	5.074	8.38e-07 ***
bmi	0.6068	0.1140	5.324	2.54e-07 ***

```
Residual standard error: 8.466 on 216 degrees of freedom
Multiple R-squared:  0.2107,          Adjusted R-squared:  0.1997
F-statistic: 19.21 on 3 and 216 DF,  p-value: 4.411e-11
```

como explicativa. A presentarlo nos abocamos en la siguiente sección.

4.15.1. Test F para testear si varios parámetros son cero, y tabla de ANOVA para comparar modelos

En forma análoga a la descripta en la Sección 4.8.1, pueden usarse las sumas de cuadrados para comparar el ajuste proporcionado por dos modelos lineales distintos. Esto puede hacerse de manera general, para diversos modelos. Lo describiremos con cierto detalle para la situación que nos interesa ahora. En el caso de los datos de azúcar queremos testear si la variable categórica que describe la actividad física es significativa para explicar el nivel de glucosa **cuando en el modelo tenemos a BMI como explicativa**. Es decir, para el modelo (64)

$$E(Y | X) = \beta_0 + \beta_2 Ievo2 + \beta_3 Ievo3 + \beta_{BMI} bmi$$

queremos testear las hipótesis

$$H_0 : \beta_2 = \beta_3 = 0 \tag{74}$$

$$H_1 : \text{al menos uno de los } \beta_i \text{ con } i \text{ entre 2 y 3 es tal que } \beta_i \neq 0$$

Para ello, ajustamos dos modelos lineales a los datos y usaremos la suma de cuadrados propuesta en (50) como medida de cuan bueno es cada ajuste, es decir, calcularemos y compararemos las

$$\Delta_{\text{modelo}} = \sum (\text{observados} - \text{modelo})^2$$

para cada uno de dos modelos. En este caso el modelo básico será el que vale si H_0 es verdadera, el modelo lineal simple que tiene a BMI como única explicativa del nivel medio de glucosa:

$$Y_i = \beta_0^{\text{básico}} + \beta_{BMI}^{\text{básico}} \mathbf{bmi}_i + \varepsilon_i.$$

Para este modelo se calculan las estimaciones de los parámetros $\hat{\beta}_0^{\text{básico}}$ y $\hat{\beta}_{BMI}^{\text{básico}}$, y con ellos los predichos

$$\hat{Y}_i^{\text{básico}} = \hat{\beta}_0^{\text{básico}} + \hat{\beta}_{BMI}^{\text{básico}} \mathbf{bmi}_i$$

y la suma de cuadrados que mide el desajuste

$$\Delta_{\text{modelo básico}} = \sum_{i=1}^n (Y_i - \hat{Y}_i^{\text{básico}})^2.$$

El modelo más complejo será el que figura en (64), es decir

$$Y_i = \beta_0^{\text{comp}} + \beta_2^{\text{comp}} \mathbf{Ievo2}_i + \beta_3^{\text{comp}} \mathbf{Ievo3}_i + \beta_{BMI}^{\text{comp}} \mathbf{bmi}_i + \varepsilon_i.$$

Nuevamente se estiman los parámetros bajo este modelo obteniéndose $\hat{\beta}_0^{\text{comp}}$, $\hat{\beta}_2^{\text{comp}}$, $\hat{\beta}_3^{\text{comp}}$ y $\hat{\beta}_{BMI}^{\text{comp}}$, con ellos se calculan los predichos para este modelo

$$\hat{Y}_i^{\text{comp}} = \hat{\beta}_0^{\text{comp}} + \hat{\beta}_2^{\text{comp}} \mathbf{Ievo2}_i + \hat{\beta}_3^{\text{comp}} \mathbf{Ievo3}_i + \hat{\beta}_{BMI}^{\text{comp}} \mathbf{bmi}_i$$

y la suma de cuadrados que mide el desajuste que tienen los datos a este modelo complejo

$$\Delta_{\text{modelo complejo}} = \sum_{i=1}^n (Y_i - \hat{Y}_i^{\text{comp}})^2.$$

Por supuesto, como el modelo complejo tiene al modelo básico como caso particular, resulta que el ajuste del modelo complejo a los datos será siempre tan satisfactorio como el del modelo básico o más satisfactorio aún, de modo que $\Delta_{\text{modelo complejo}} \leq \Delta_{\text{modelo básico}}$. Es de interés observar que la estimación del coeficiente que acompaña al BMI depende de qué covariables hay en el modelo, excepto cuando todas las covariables presentes en el modelo sean **no correlacionadas** con BMI, lo cual ocurrirá las menos de las veces: en general las variables explicativas

Tabla 37: Tabla de ANOVA para comparar dos modelos de regresión

Modelo	SS	g.l.	Diferencia	g.l.	F
Básico	$\Delta_{\text{mod bás}}$	$n - 2$			
Complejo	$\Delta_{\text{mod comp}}$	$n - 4$	$\Delta_{\text{mod bás}} - \Delta_{\text{mod comp}}$	2	$\frac{(\Delta_{\text{mod bás}} - \Delta_{\text{mod comp}})/2}{\Delta_{\text{mod comp}}/(n-4)}$

están vinculadas entre sí de manera más o menos estrecha, eso significa que en general estarán (linealmente) correlacionadas.

Nuevamente se puede construir una tabla de ANOVA para resumir la información descripta hasta ahora. En la Tabla 37 describimos la forma en la que se presenta la información.

La resta $\Delta_{\text{modelo básico}} - \Delta_{\text{modelo complejo}}$ mide la mejora en el ajuste debido al modelo más complejo respecto del más sencillo. Los grados de libertad de esta resta será la resta de los grados de libertad de los dos ajustes, en el ejemplo $(n - 4) - (n - 2) = 2$. Esta cuenta da siempre la diferencia entre el número de coeficientes del modelo más complejo respecto del más básico. El test F se basa en la comparación de la mejora en el ajuste debido al modelo más complejo respecto del simple relativa al ajuste proporcionado por el modelo complejo (el mejor ajuste disponible), ambos divididos por sus grados de libertad. El test F para las hipótesis (74) rechaza H_0 cuando $F > F_{2,n-4,\alpha}$ (el percentil $1 - \alpha$ de la distribución F con 2 grados de libertad en el numerador y $n - 4$ grados de libertad en el denominador) o, equivalentemente, cuando el p -valor calculado como $P(F_{2,n-4} > F_{\text{obs}})$ es menor que α . En general, cuando se comparan

$$\begin{aligned} \text{Modelo complejo: } Y_i &= \beta_0^c + \sum_{k=1}^{p-1} \beta_k^c X_{ik} + \varepsilon_i \\ \text{Modelo simple: } Y_i &= \beta_0^s + \sum_{k=1}^{q-1} \beta_k^s X_{ik} + \varepsilon_i \end{aligned} \quad (75)$$

Es decir, cuando se testea

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$

$$H_1 : \text{al menos uno de los } \beta_k \text{ con } k \text{ entre } q \text{ y } p - 1 \text{ es tal que } \beta_k \neq 0$$

en el modelo (75), los grados de libertad del estadístico F serán $p - q$ en el numerador y $n - p$ en el denominador. Para los datos del archivo `azucar`, la tabla de ANOVA para chequear las hipótesis (74) es la que figura en la Tabla 38. Como el p -valor es menor a 0,05 resulta que cuando controlamos a la glucosa por el BMI,

el nivel de evolución del peso de cada paciente resulta significativo. Luego la evolución del peso es útil para predecir el nivel de glucosa, aún cuando controlamos por el BMI.

Tabla 38: Comparación de sumas de cuadrados para evaluar la significatividad de physact (categórica) una vez que se tiene a BMI (numérica) como regresora de glucosa

```
> uno<-lm(glucosa ~ bmi)
> dos<-lm(glucosa ~ Ievo + bmi)
> anova(uno,dos)
Analysis of Variance Table

Model 1: glucosa ~ bmi
Model 2: glucosa ~ Ievo + bmi
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     218 17328
2     216 15480  2      1848.1 12.894 5.128e-06 ***
---

```

4.15.2. Comparaciones múltiples

Cuando usamos un modelo de regresión (73), podemos querer estimar los efectos diferenciales entre los dos niveles de `peso.evo` que no involucren al basal. Cuando la variable categórica tenga más de tres categorías, habrá todavía más comparaciones que considerar, en este caso sólo resta considerar una. Esto puede hacerse estimando diferencias entre coeficientes de regresión. En el ejemplo, $\beta_3 - \beta_2$ indica cuánto más alta (o baja) es la función de respuesta para “subió de peso” (`peso.evo=3`) comparada con “mantuvo su peso” (`peso.evo=1`) para cualquier nivel de BMI pues

$$\begin{aligned} & E(Y | \text{bmi}, \text{peso.evo} = 3) - E(Y | \text{bmi}, \text{peso.evo} = 2) \\ &= \beta_0 + \beta_3 + \beta_{BMI} \text{BMI} - \beta_0 - \beta_2 - \beta_{BMI} \text{BMI} \\ &= \beta_3 - \beta_2. \end{aligned}$$

El estimador puntual de esta cantidad es, por supuesto, $\widehat{\beta}_3 - \widehat{\beta}_2$, y la varianza estimada de este estimador es

$$\widehat{Var}(\widehat{\beta}_3 - \widehat{\beta}_2) = \widehat{Var}(\widehat{\beta}_3) + \widehat{Var}(\widehat{\beta}_2) + 2\widehat{Cov}(\widehat{\beta}_3, \widehat{\beta}_2).$$

Las varianzas y covarianzas necesarias se pueden obtener a partir de la matriz de covarianza de los coeficientes de regresión.

```
>mm <- summary(ajuste4)$cov.unscaled * (summary(ajuste4)$sigma)^2
      (Intercept)      Ievo2      Ievo3       bmi
(Intercept) 11.4002401 -2.46807671 -1.88607917 -0.34625635
Ievo2        -2.4680767  3.17653751  2.16249632  0.01112123
Ievo3        -1.8860792  2.16249632  2.80368157 -0.01071534
bmi         -0.3462563  0.01112123 -0.01071534  0.01299155
> sqrt(diag(mm)) #coincide con la columna de Std. Error del lm
(Intercept)      Ievo2      Ievo3       bmi
3.3764242    1.7822844   1.6744198   0.1139805
```

De todos modos, esta cuenta la realiza, en general, el software estadístico. A continuación vemos las comparaciones de a pares realizadas con el método de la diferencia honestamente significativa de Tukey (HSD: *honestly significant difference*), realizada con nivel conjunto del 95 %. Mostramos los intervalos de confianza calculados con la instrucción TukeyHSD. Es muy importante en esta instrucción incorporar primero la variable continua y luego la categórica para obtener los resultados que queremos. También presentamos la salida (intervalos de confianza y tests) del comando glht, del paquete multcomp. Ahí vemos que exceptuando la diferencia entre los subgrupos dados por los niveles 2 y 3 de actividad física, las restantes diferencias son estadísticamente significativas a nivel conjunto 95 %.

```
#cambiamos el orden, covariable cont primero
> TukeyHSD(aov(glucosa ~ bmi + Ievo),"Ievo")
  Tukey multiple comparisons of means
  95% family-wise confidence level
Fit: aov(formula = glucosa ~ bmi + Ievo)

$Ievo
  diff      lwr      upr      p adj
2-1 6.827351  2.627495 11.027207 0.0004795
3-1 8.472188  4.526816 12.417560 0.0000026
3-2 1.644837 -1.357564  4.647237 0.4007100

> ajuste4<-lm(glucosa ~ bmi + Ievo)
> library(multcomp) #para testear combinaciones de los parametros
> posthoc <- glht(ajuste4, linfct = mcp(Ievo = "Tukey"),test = Ftest())
> summary(posthoc)
```

```

Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = glucosa ~ bmi + Ievo)
Linear Hypotheses:
Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0     6.802      1.782   3.817 0.000513 ***
3 - 1 == 0     8.496      1.674   5.074 < 1e-04 ***
3 - 2 == 0     1.694      1.287   1.317 0.383040
---
> confint(posthoc)
Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = glucosa ~ bmi + Ievo)
Quantile = 2.3509
95% family-wise confidence level

Linear Hypotheses:
Estimate lwr      upr
2 - 1 == 0  6.8023  2.6123 10.9923
3 - 1 == 0  8.4963  4.5599 12.4327
3 - 2 == 0  1.6940 -1.3305  4.7186

```

En la Figura 48 se ve un gráfico de estos intervalos de confianza de nivel simultáneo 95 %. Sólo el intervalo para la diferencia de medias entre los grupos 2 y 3 contiene al cero. Los restantes quedan ubicados a la izquierda del cero. En el gráfico esto se ve más fácilmente que leyendo la tabla. Para el modelo con las covariables Ievo pero sin bmi también podríamos haber exhibido un gráfico como éste (de hecho, no lo hicimos puesto que ambos dan muy parecidos).

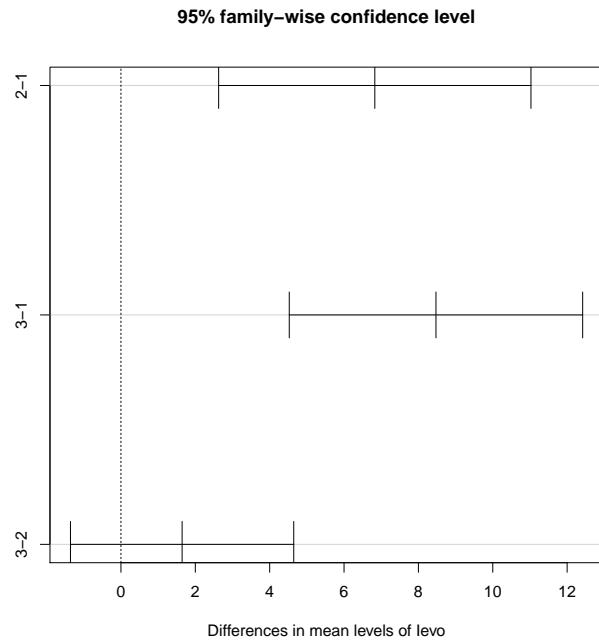
4.16. Modelos con interacción entre variables cuantitativas y cualitativas

Como ya dijimos, cuando proponemos un modelo de regresión lineal múltiple del estilo de

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad (76)$$

estamos asumiendo que los efectos de las variables X_1 y X_2 sobre la respuesta Y no interactúan entre sí: es decir, que el efecto de X_1 en Y no depende del valor que tome X_2 (y al revés, cambiando X_1 por X_2 , el efecto de X_2 en Y no depende del valor que tome X_1). Cuando esto no sucede, es inadecuado proponer el modelo (76), y es necesario agregarle a dicho modelo un témino que intente dar cuenta

Figura 48: Intervalos de confianza de nivel simultáneo para las diferencias de los niveles medios de glucosa de cada grupo, controlados por el BMI.



de la *interacción* entre X_1 y X_2 en su relación con Y , es decir, del hecho de que el efecto de un predictor sobre la respuesta difiere de acuerdo al nivel de otro predictor. La manera estándar de hacerlo es agregarle al modelo (76) un término de interacción, es decir

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} \cdot X_{i2} + \varepsilon_i. \quad (77)$$

El modelo (77) es un caso particular del modelo de regresión lineal múltiple. Sea $X_{i3} = X_{i1} \cdot X_{i2}$ el producto entre las variables X_1 y X_2 medidas en el iésimo individuo, entonces el modelo (77) puede escribirse de la forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i,$$

que es un caso particular del modelo de regresión lineal múltiple presentado en (44). Algunas veces al coeficiente de la interacción se lo nota con los subíndices 1 : 2, es decir $\beta_{1:2} = \beta_3$ para explicitar que es el coeficiente asociado a la interacción. Veamos un ejemplo.

Ejemplo 4.1 Consideremos datos sobre la frecuencia cardíaca o pulso medido a 40 personas antes y después de ejercitarse. Estos datos aparecen publicados en el manual del paquete BMDP, sin citar las fuentes. Figuran en la carpeta de datos de este apunte en el archivo *pulso.txt*. Se les pidió que registraran su pulso, luego que corrieran una milla, y luego volvieran a registrar su pulso. Además se registró su sexo, edad y si eran o no fumadores. De este modo, para cada individuo, se midieron las siguientes variables

$$\begin{aligned} Y &= \text{pulso luego de correr una milla (Pulso2)} \\ X_1 &= \text{pulso en reposo (Pulso1)} \\ X_2 &= \begin{cases} 1 & \text{si la persona es mujer} \\ 0 & \text{en caso contrario} \end{cases} \\ X_3 &= \begin{cases} 1 & \text{si la persona fuma} \\ 0 & \text{en caso contrario} \end{cases} \\ X_4 &= \text{edad} \end{aligned}$$

Interesa explicar el pulso post-ejercicio, en función de algunas de las demás covariables. Es de interés saber si la edad, o el hábito de fumar inciden en él. La frecuencia cardíaca es el número de contracciones del corazón o pulsaciones por unidad de tiempo. Su medida se realiza en unas condiciones determinadas (reposo o actividad) y se expresa en latidos por minuto.

Tanto el sexo como la condición de fumador son variables dummies o binarias. En la base de datos se las denomina $X_2 = \text{mujer}$ y $X_3 = \text{fuma}$. Las restantes son variables continuas. En la Figura 49 hacemos un scatter plot de Y versus X_1 . En él se puede ver que a medida que X_1 crece también lo hace Y , y que una relación lineal es una buena descripción (inicial) de la relación entre ellas.

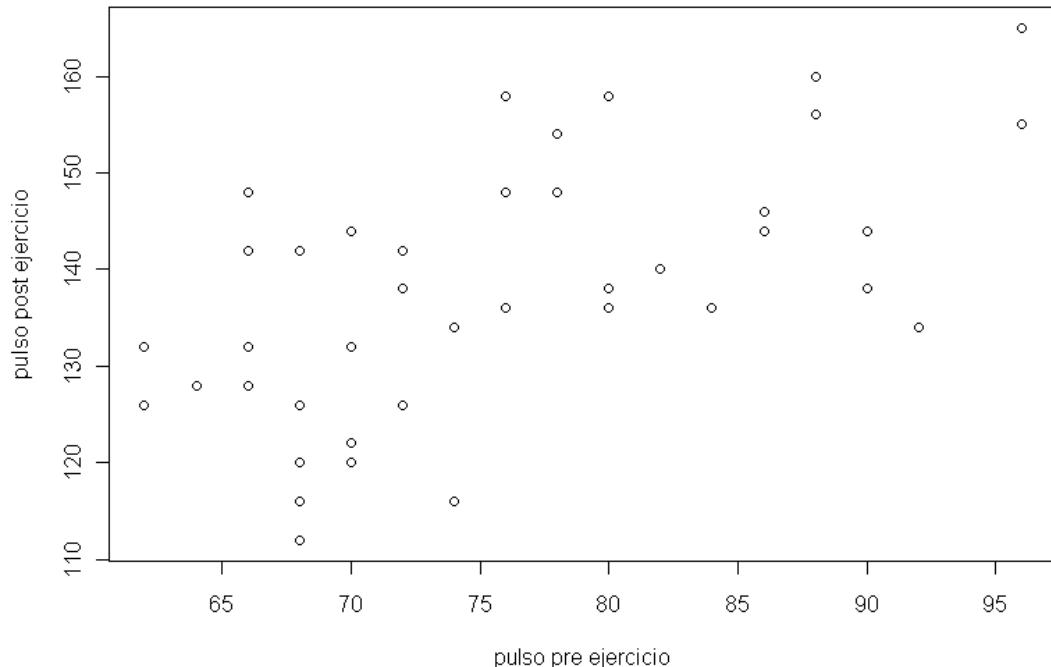
Si identificamos en ese gráfico a las observaciones según su sexo, obtenemos el gráfico de dispersión que aparece en la Figura 50. En él observamos que el género de la persona parece influir en la relación entre ambas variables.

Querríamos cuantificar el efecto del género en el pulso medio post ejercicio. Para ello vamos a ajustar un modelo de regresión lineal múltiple con el pulso post ejercicio como variable dependiente. Proponemos un modelo lineal múltiple para estos datos. El modelo múltiple sería en este caso

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad (78)$$

Como ya vimos en la Sección 4.13.2, este modelo sin interacción propone que el pulso medio post-ejercicio es una función lineal del pulso pre-ejercicio, con dos

Figura 49: Gráfico de dispersión del pulso post-ejercicio versus el pulso pre-ejercicio, para 40 adultos. Archivo: `pulso.txt`



rectas diferentes para las mujeres y los hombres, pero estas rectas tienen la misma pendiente. O sea, la ecuación (78) propone que para las mujeres, (o sea, cuando $X_2 = 1$)

$$\begin{aligned} E(Y | X_1, X_2 = 1) &= \beta_0 + \beta_1 X_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 X_1 \end{aligned}$$

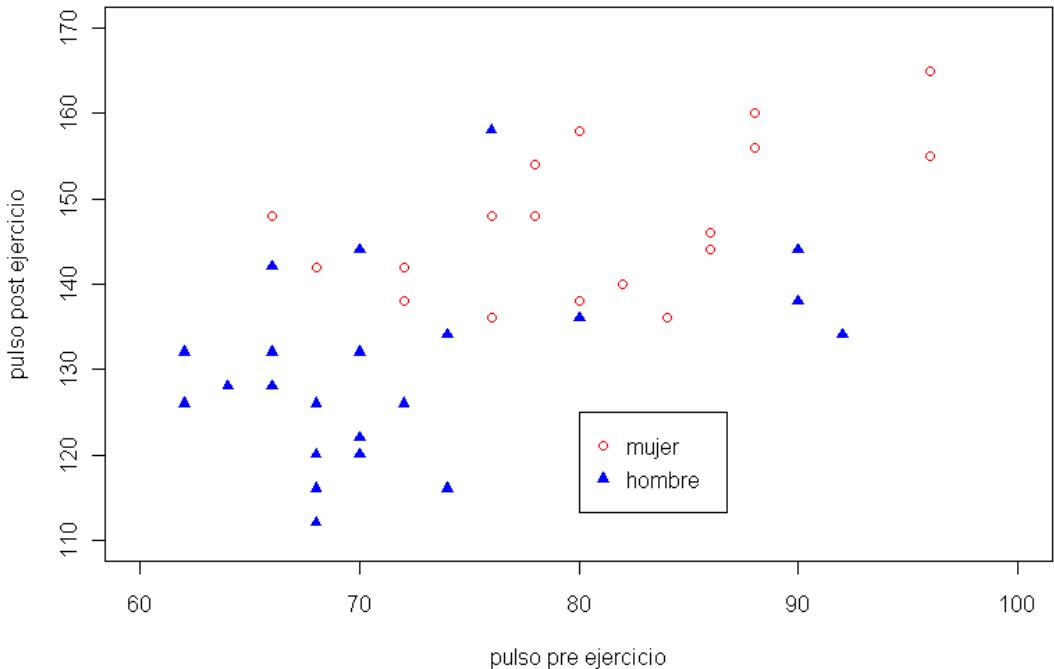
mientras que para los hombres (cuando $X_2 = 0$) se tiene

$$E(Y | X_1, X_2 = 0) = \beta_0 + \beta_1 X_1.$$

La salida del ajuste del modelo está en la Tabla 39. De acuerdo a ella, la recta ajustada es

$$\hat{Y} = 93,0970 + 0,5157 \cdot X_1 + 12,7494 \cdot X_2$$

Figura 50: Gráfico de dispersión del pulso post-ejercicio versus el pulso pre-ejercicio, identificando el sexo de cada observación.



El coeficiente estimado de **mujer** es positivo, indicando que cuando la variable X_2 aumenta de 0 a 1 (**mujer** = 0 quiere decir que se trata de un hombre), el pulso medio post ejercicio crece, es decir, el pulso medio de las mujeres es mayor que el de los hombres si uno controla por pulso en reposo. ¿Será estadísticamente significativa esta observación? Para evaluarlo, hacemos un test de

$$H_0 : \beta_2 = 0 \text{ versus } H_0 : \beta_2 \neq 0$$

asumiendo que el modelo contiene al pulso en reposo. El estadístico observado resulta ser $t_{obs} = 3,927$ y p -valor = 0,000361. Entonces, rechazamos la hipótesis nula y concluimos que $\beta_2 \neq 0$. Si construyéramos un intervalo de confianza para β_2 , éste resultaría contenido enteramente en $(-\infty, 0)$. Por eso concluimos que el verdadero valor poblacional de β_2 es menor a cero. Es decir, para las dos poblaciones de personas (hombres y mujeres) con el mismo pulso en reposo, en promedio los pulsos medios luego de ejercitarse serán mayores en las mujeres que en los hombres.

Tabla 39: Ajuste del modelo lineal múltiple $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$, donde X_1 = pulso pre ejercicio (Pulso1), X_2 = indicador de mujer (mujer), Y = pulso post ejercicio (Pulso2).

```
> ajuste1<-lm(Pulso2~ Pulso1+mujer)
> summary(ajuste1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	93.0970	12.5157	7.438	7.44e-09
Pulso1	0.5157	0.1715	3.007	0.004725
mujer	12.7494	3.2468	3.927	0.000361

Residual standard error: 9.107 on 37 degrees of freedom
Multiple R-squared: 0.5445, Adjusted R-squared: 0.5199
F-statistic: 22.12 on 2 and 37 DF, p-value: 4.803e-07

Para entender mejor este modelo escribimos las dos rectas ajustadas en cada caso. El modelo ajustado para las mujeres, ($X_2 = 1$) es

$$\begin{aligned}\hat{Y} &= (93,0970 + 12,7494) + 0,5157 \cdot X_1 \\ &= 105,85 + 0,5157 \cdot X_1\end{aligned}$$

mientras que para los hombres ($X_2 = 0$)

$$\hat{Y} = 93,0970 + 0,5157 \cdot X_1.$$

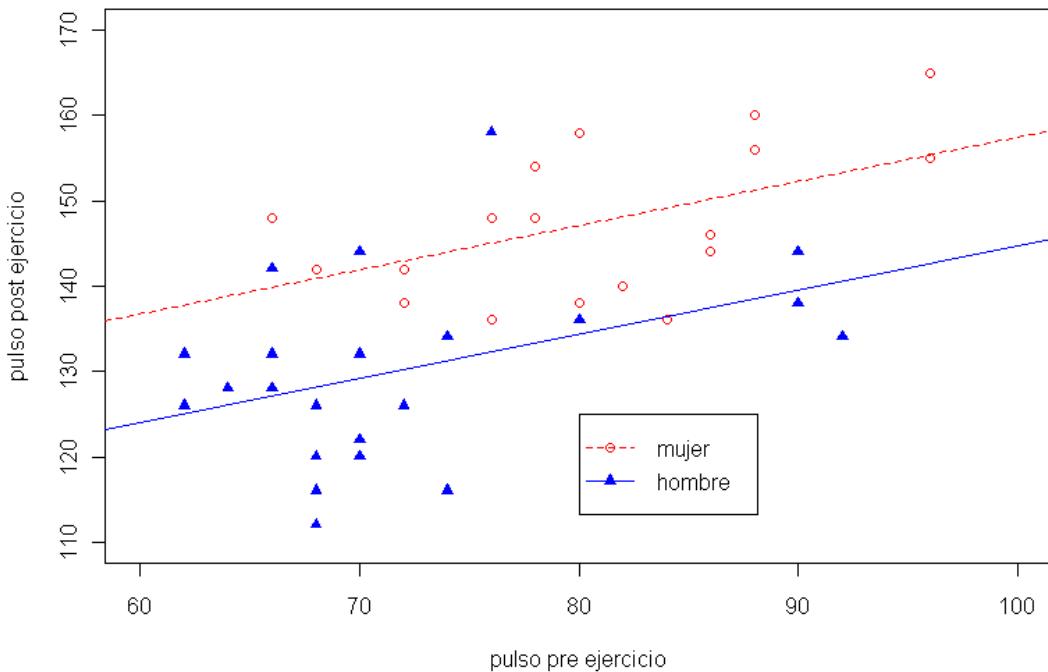
Las dos rectas están graficadas en la Figura 51, junto con las observaciones identificadas por sexo. Observemos que ambas rectas son paralelas: en ambos grupos una unidad (un latido por minuto) de aumento en el pulso en reposo está asociado con un incremento en 0,5157 latidos por minuto de la frecuencia cardíaca post ejercicio, en promedio. Esto es consecuencia del modelo propuesto.

Ahora queremos proponer un modelo con interacción para estos datos. Es decir proponemos el modelo

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{1:2} X_{i1} \cdot X_{i2} + \varepsilon_i \quad (79)$$

Como la variable X_2 asume solamente valores 0 y 1, el término de la interacción $X_{i1} \cdot X_{i2}$ valdrá 0 siempre que $X_2 = 0$ (o sea para los hombres), y será igual a

Figura 51: Rectas ajustadas para los dos géneros (modelo sin interacción).



X_1 siempre que $X_2 = 1$ (o sea para las mujeres). En la población de personas ejercitando, esta nueva variable tendrá coeficiente $\beta_{1:2}$. Llaremos $X = (X_1, X_2)$. Si escribimos el modelo propuesto para los dos grupos de observaciones, tendremos que cuando **mujer** = 1,

$$\begin{aligned} E(Y | X) &= \beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_{1:2} X_1 \cdot 1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{1:2}) X_1 \quad \text{mujeres} \end{aligned}$$

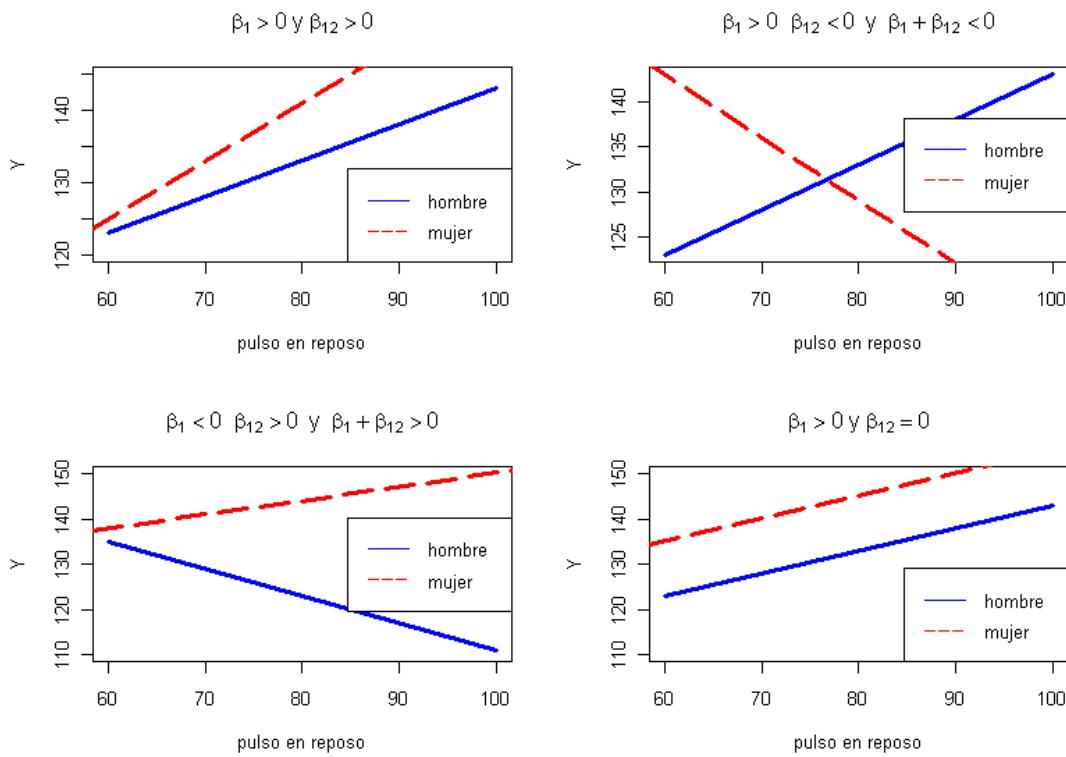
Mientras que cuando **mujer** = 0, proponemos

$$\begin{aligned} E(Y | X) &= \beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_{1:2} X_1 \cdot 0 \\ &= \beta_0 + \beta_1 X_1 \quad \text{hombres} \end{aligned}$$

Es decir que para cada grupo estamos proponiendo ajustar dos rectas distintas. Observemos que estas rectas no están atadas (como sí lo estaban en el modelo aditivo con una explicativa binaria y una continua, en el que ajustábamos **dos**

rectas paralelas). Por otro lado, la interpretación de los coeficientes del modelo cambia. Analicemos cada uno. El coeficiente de X_1 (β_1) es la pendiente del pulso en el grupo de hombres. Indica que por cada aumento en una unidad en el pulso en reposo entre los hombres, el pulso medio post ejercicio aumenta (o disminuye, según el signo) β_1 unidades. El coeficiente de la interacción ($\beta_{1:2}$) representa el aumento (o la disminución) de la pendiente en el grupo de las mujeres con respecto al de los hombres. Si $\beta_{1:2} = 0$ esto significaría que ambas rectas son paralelas. Los distintos valores que pueden tomar β_1 y $\beta_{1:2}$ dan lugar a distintos posibles tipos de interacción entre las variables, según se ve en la Figura 52.

Figura 52: Gráfico de posibles combinaciones de valores de β_1 y $\beta_{1:2}$ para el modelo (79).



Esas posibilidades de modelo no están limitadas a los tratamientos significativos de la Tabla 10, el test de

$$H_0 : \beta_{1:2} = 0 \text{ versus } H_0 : \beta_{1:2} \neq 0$$

asumiendo que el modelo contiene al pulso en reposo y a la indicadora de mujer, tiene por estadístico $t_{obs} = 0,211$ y p -valor = 0,834. Esto nos dice que esta muestra

Tabla 40: Ajuste del modelo lineal con interacción entre $X_1 =$ pulso pre ejercicio (`Pulso1`), $X_2 =$ indicador de mujer (`mujer`), $Y =$ pulso post ejercicio (`Pulso2`).

```
> ajuste2<-lm(Pulso2~ Pulso1 * mujer)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	95.42838	16.80929	5.677	1.88e-06
Pulso1	0.48334	0.23157	2.087	0.044
mujer	7.05575	27.14749	0.260	0.796
Pulso1:mujer	0.07402	0.35033	0.211	0.834

Residual standard error: 9.227 on 36 degrees of freedom

Multiple R-squared: 0.5451, Adjusted R-squared: 0.5072

F-statistic: 14.38 on 3 and 36 DF, p-value: 2.565e-06

no provee evidencia suficiente de que el pulso en reposo tenga un efecto diferente en el pulso post ejercicio dependiendo del sexo de la persona.

Como la interacción no es estadísticamente significativa, no la retendremos en el modelo de regresión. Sin embargo, veamos cuánto dan las dos rectas ajustadas en este caso. Cuando `mujer` = 1,

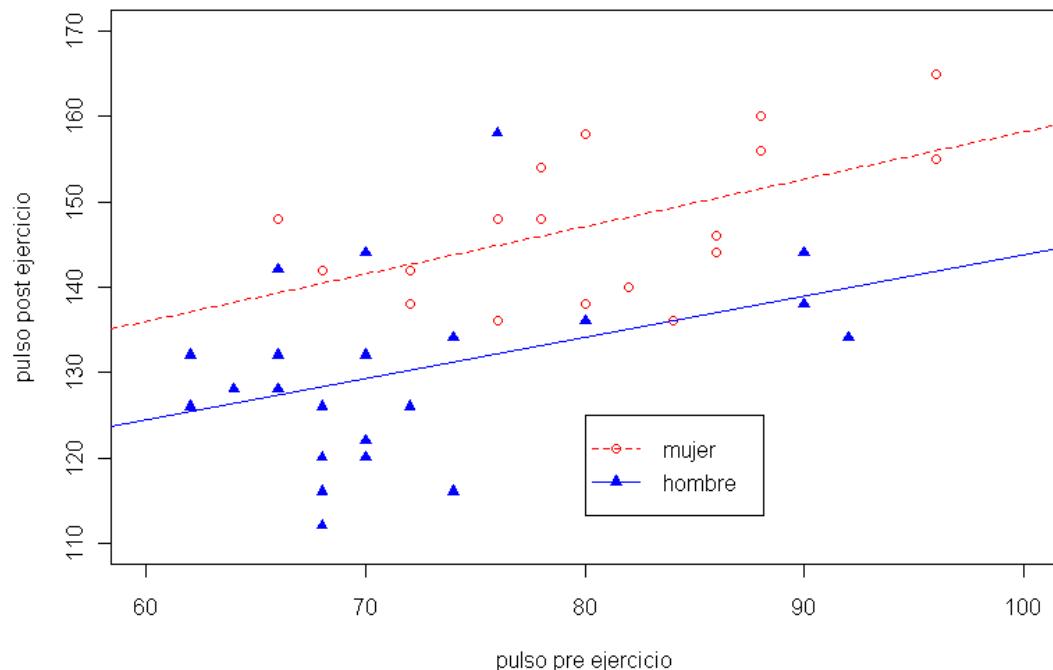
$$\begin{aligned}\hat{Y} &= 95,429 + 0,483 \cdot X_1 + 7,056 \cdot 1 + 0,074X_1 \cdot 1 \\ &= 102,485 + 0,557 \cdot X_1\end{aligned}$$

Mientras que cuando `mujer` = 0, resulta

$$\begin{aligned}\hat{Y} &= 95,429 + 0,483 \cdot X_1 + 7,056 \cdot 0 + 0,074X_1 \cdot 0 \\ &= 95,429 + 0,483 \cdot X_1\end{aligned}$$

El gráfico de ambas rectas puede verse en la Figura 53. Estas dos rectas no tienen la misma pendiente, ni la misma ordenada al origen. En el rango de interés, sin embargo, la recta que describe el pulso medio post-ejercicio para las mujeres está completamente sobre la de los hombres. Esto implica que a lo largo de todos los valores relevantes del pulso en reposo, predeciremos valores de pulso post-ejercicio mayores para las mujeres que para los hombres. Si comparamos los ajustes obtenidos para los modelos que explican a Y con las variables `Pulso1` y `mujer` sin interacción (78) y con interacción (79), que aparecen en las Tablas 39 y 40, respectivamente, vemos que son muy diferentes.

Figura 53: Rectas ajustadas por mínimos cuadrados para distintos niveles de sexo, con el término de interacción incluido.



En la Tabla 41 resumimos lo observado. Cuando el término de la interacción se incluye en el modelo, el coeficiente de **mujer** se reduce en magnitud, casi a la mitad. Además, su error estándar aumenta multiplicándose por un factor de 8. En el modelo sin término de interacción, el coeficiente de **mujer** es significativamente distinto de cero, a nivel 0,05; esto no ocurre cuando incluimos el término de interacción en el modelo, en ese caso la variable **mujer** deja de ser significativa. El coeficiente de determinación (R^2) no cambia al incluir la interacción, sigue valiendo 0,545. Más aún, el coeficiente de determinación ajustado decrece ligeramente con la incorporación de una covariante más al modelo. Al tomar en cuenta simultáneamente todos estos hechos, concluimos que la inclusión del término de interacción de $\text{Pulso1} \cdot \text{mujer}$ en el modelo no explica ninguna variabilidad adicional en los valores observados del pulso post-ejercicio, más allá de lo que es explicado por las variables **mujer** y **Pulso1** en forma aditiva. La información proporcionada por este término es redundante.

¿Por qué sucede esto? Muchas veces sucede que al incorporar una nueva variable al modelo ajustado, se pierde la significatividad de alguna o varias variables ya incluidas previamente. Si además de suceder esto aparece una inestabilidad de los coeficientes estimados, difiriendo sustancialmente los valores estimados de algunos coeficientes en los dos modelos, y en particular, se observa un aumento grosero de los errores estándares: esto suele ser un síntoma de *colinealidad* o *multicolinealidad* entre los predictores. La colinealidad ocurre cuando dos o más variables explicativas están altamente correlacionadas, a tal punto que, esencialmente, guardan la misma información acerca de la variancia observada de Y . En la Sección 5.3.1 presentaremos algunas maneras de detectar y resolver la multicolinealidad.

En este caso, la variable artificial `Pulso1 · mujer` está fuertemente correlacionada con `mujer` ya que el coeficiente de correlación de Pearson es $r_{mujer, Pulso1 \cdot mujer} = 0,99$, como aparece en la Tabla 42. Como la correlación entre las variables es tan grande, la capacidad explicativa de `Pulso1 · mujer` cuando `mujer` está en el modelo es pequeña.

Tabla 41: Tabla comparativa de los ajustes con y sin interacción para las covariables `Pulso1` y `mujer`.

	Sin interacción	Con interacción
Coeficiente $\hat{\beta}_2$	12,749	7,056
Error estándar de $\hat{\beta}_2$	3,247	27,147
Valor del estadístico t	3,927	0,26
p–valor	0,000361	0,796
R^2	0,5445	0,5451
R^2 ajustado	0,5199	0,5072

Tabla 42: Correlaciones de Pearson entre $X_1 =$ pulso pre ejercicio (`Pulso1`), $X_2 =$ indicador de mujer (`mujer`) e $Y =$ pulso post ejercicio (`Pulso2`).

	Pulso1	mujer	Pulso1·mujer
Pulso1	1	0,453	0,53
mujer	0,453	1	0,99
Pulso1·mujer	0,53	0,99	1

Un modo de resolver el problema de la multicolinealidad es trabajar con los datos centrados para la o las variables predictoras que aparecen en más de un término del modelo. Esto es, usar no la variable X tal como fue medida, sino la diferencia entre el valor observado y el valor medio en la muestra.

4.17. Interacción entre dos variables cuantitativas

En la sección anterior presentamos la interacción entre dos variables cuando una es cualitativa y la otra cuantitativa. Ahora nos ocuparemos de estudiar la situación en la que las dos variables que interesan son cuantitativas. Vimos que el modelo aditivo propone que cuando la covariable X_j aumenta una unidad, la media de Y aumenta en β_j unidades independientemente de cuáles sean los valores de las otras variables. Esto implica paralelismo de las rectas que relacionan a Y y X_j , cualesquiera sean los valores que toman las demás variables.

En nuestro ejemplo de los bebés de bajo peso, analizado en la Sección 4.7, propusimos un modelo de regresión lineal múltiple con dos variables predictoras. Recorremos que habíamos definido

$$Y_i = \text{perímetro cefálico del iésimo niño, en centímetros (headcirc)}$$

$$X_{i1} = \text{edad gestacional del iésimo niño, en semanas (gestage)}$$

$$X_{i2} = \text{peso al nacer del iésimo niño, en gramos (birthwt)}$$

Propusimos el siguiente modelo,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \quad (80)$$

El modelo ajustado figura en la Tabla 19, página 125. La superficie ajustada resultó ser

$$\hat{Y} = 8,3080 + 0,4487X_1 + 0,0047X_2.$$

Cuando controlamos por X_2 (peso al nacer), la ecuación (parcial) ajustada que relaciona el perímetro cefálico y la edad gestacional es

$$X_2 = 600, \quad \hat{Y} = 8,3080 + 0,4487X_1 + 0,0047 \cdot 600 = 11,128 + 0,4487X_1$$

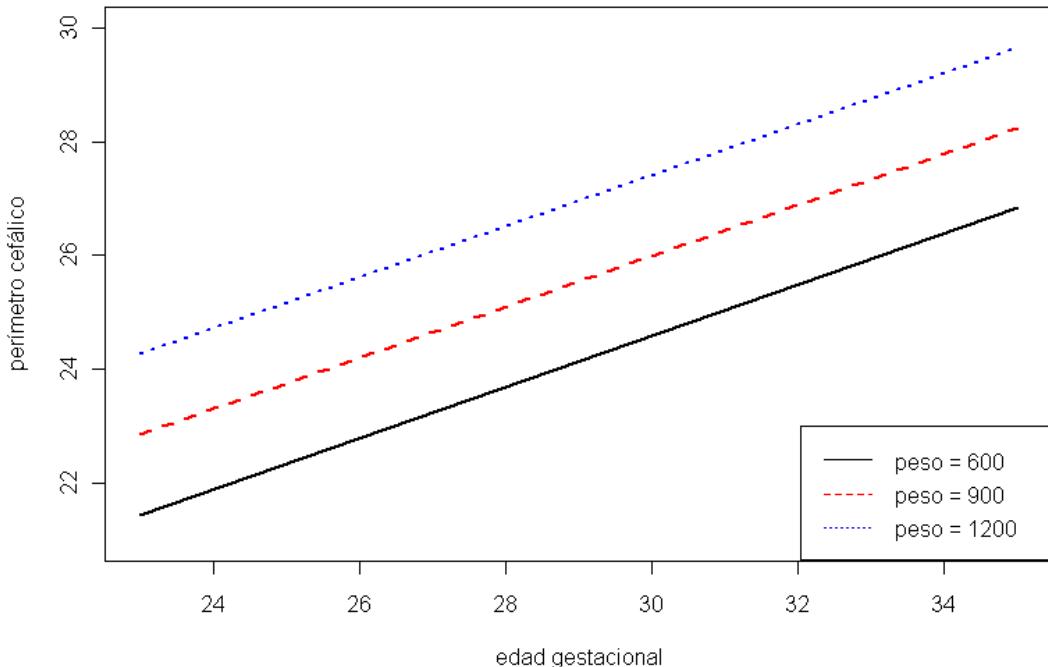
$$X_2 = 900, \quad \hat{Y} = 8,3080 + 0,4487X_1 + 0,0047 \cdot 900 = 12,538 + 0,4487X_1$$

$$X_2 = 1200, \quad \hat{Y} = 8,3080 + 0,4487X_1 + 0,0047 \cdot 1200 = 13,948 + 0,4487X_1$$

Para cada nivel posible de peso al nacer, por cada unidad de aumento en la edad gestacional se espera un aumento de 0,448 unidades (cm.) en el perímetro cefálico al nacer. Gráficamente, esto se ve representado en la Figura 54. Lo mismo sucedería si controláramos por X_1 en vez de X_2 : tendríamos rectas paralelas, de pendiente 0,0047.

Este modelo asume que no existe interacción entre las variables. El modelo (80) fuerza a que los efectos de las covariables en la variable dependiente sean aditivos, es decir, el efecto de la edad gestacional será el mismo para todos los valores del peso al nacer, y viceversa, porque el modelo no le permitirá ser de ninguna otra forma. A menudo este modelo es demasiado simple para ser adecuado, aunque en

Figura 54: Perímetro cefálico esperado en función de la edad gestacional, controlando por peso al nacer, para tres posibles valores de peso al nacer (600, 900 y 1200g.) en el modelo sin interacción.



muchos conjuntos de datos proporciona una descripción satisfactoria del vínculo entre las variables.

Cuando esto no suceda, es decir, cuando pensemos que tal vez la forma en que el perímetro cefálico varíe con la edad gestacional dependa del peso al nacer del bebé, será necesario descartar (o validar) esta conjectura. Una manera de investigar esta posibilidad es incluir un término de interacción en el modelo. Para ello, creamos la variable artificial que resulta de hacer el producto de las otras dos: $X_3 = X_1 \cdot X_2 = \text{gestage} \cdot \text{birthwt}$, y proponemos el modelo

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \\ Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 \cdot X_2 + \varepsilon \end{aligned} \quad (81)$$

Este es un caso especial de un modelo de regresión con tres variables regresoras. ¿Cómo se interpreta este modelo para dos variables cuantitativas? En este caso

decimos que existe interacción estadística cuando la pendiente de la relación entre la variable respuesta y una variable explicativa cambia para distintos niveles de las otras variables. Para entenderlo, escribamos el modelo propuesto cuando controlamos el valor de X_2 .

$$\begin{aligned} E(Y | X_1, X_2 = 600) &= \beta_0 + \beta_1 X_1 + \beta_2 600 + \beta_{1:2} X_1 \cdot 600 \\ &= \underbrace{\beta_0 + \beta_2 600}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 600)}_{\text{pendiente}} X_1 \end{aligned}$$

$$\begin{aligned} E(Y | X_1, X_2 = 900) &= \beta_0 + \beta_1 X_1 + \beta_2 900 + \beta_{1:2} X_1 900 \\ &= \underbrace{\beta_0 + \beta_2 900}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 900)}_{\text{pendiente}} X_1 \end{aligned}$$

$$\begin{aligned} E(Y | X_1, X_2 = 1200) &= \beta_0 + \beta_1 X_1 + \beta_2 1200 + \beta_{1:2} X_1 1200 \\ &= \underbrace{\beta_0 + \beta_2 1200}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 1200)}_{\text{pendiente}} X_1 \end{aligned}$$

En general

$$\begin{aligned} E(Y | X_1, X_2) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2 \\ &= \underbrace{\beta_0 + \beta_2 X_2}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} X_2)}_{\text{pendiente}} X_1 \end{aligned} \tag{82}$$

Luego, en el modelo (81), la pendiente de la relación entre X_1 e Y depende de X_2 , decimos entonces que existe interacción entre las variables.

Entonces, cuando X_2 aumenta en una unidad, la pendiente de la recta que relaciona Y con X_1 aumenta en $\beta_{1:2}$. En este modelo, al fijar X_2 , $E(Y | X_1, X_2)$ es una función lineal de X_1 , pero la pendiente de la recta depende del valor de X_2 . Del mismo modo, $E(Y | X_1, X_2)$ es una función lineal de X_2 , pero la pendiente de la relación varía de acuerdo al valor de X_1 .

Si $\beta_{1:2}$ no fuera estadísticamente significativa, entonces los datos no avalarían la hipótesis de que el cambio en la respuesta con un predictor dependa del valor del otro predictor, y podríamos ajustar directamente un modelo aditivo, que es mucho más fácil de interpretar.

Ejemplo 4.2 Consideremos un ejemplo de datos generados. Para $n = 40$ pacientes se miden tres variables:

X_1 = cantidad de droga A consumida

X_2 = cantidad de droga B consumida

$Y = \text{variable respuesta}$

Proponemos un modelo con interacción para los datos, que figuran en el archivo `ejemploint.txt`. Antes de ajustar un modelo, veamos los estadísticos descriptivos de las dos variables, en la Tabla 43. Ajustamos el modelo (81). En la Tabla 44 aparece la salida. Vemos que el coeficiente asociado al término de interacción

Tabla 43: Estadísticos descriptivos de las variables $X_1 = \text{drogaA}$ y $X_2 = \text{drogaB}$.

```
> summary(drogaA)
  Min. 1st Qu. Median     Mean 3rd Qu.    Max.
 3.207   4.449   7.744   8.107  11.100  13.590

> summary(drogaB)
  Min. 1st Qu. Median     Mean 3rd Qu.    Max.
10.18   38.44   63.02   59.58   82.61   93.76
```

es 2,771 y el test t rechaza la hipótesis $H_0 : \beta_{1:2} = 0$ (p -valor < $2 \cdot 10^{-16}$). Concluimos que la interacción resulta estadísticamente significativa, así como lo son los restantes coeficientes asociados a las dos drogas. Luego, hay variación en la pendiente de la relación entre la respuesta y la cantidad de droga A ingerida, al variar la cantidad de droga B consumida. Esto puede verse más fácilmente en el gráfico de la Figura 55. En este caso vemos que las dos drogas potencian su efecto en la variable respuesta, ya que a medida que la cantidad de droga A crece (en el gráfico pasa de 4 a 7 y luego a 11) la variable respuesta crece al crecer la droga B, con pendientes cada vez mayores. Tienen interacción positiva. Las rectas graficadas en dicha figura son

$$\begin{aligned} \text{drogaA} &= 4 \quad \hat{Y} = -53,92 + 16,59 \cdot 4 + 6,22X_2 + 2,77 \cdot 4 \cdot X_2 \\ \hat{Y} &= 12,44 + 17,3X_2 \end{aligned}$$

$$\begin{aligned} \text{drogaA} &= 7 \quad \hat{Y} = -53,92 + 16,59 \cdot 7 + 6,22X_2 + 2,77 \cdot 7 \cdot X_2 \\ \hat{Y} &= 62,21 + 25,61X_2 \end{aligned}$$

$$\begin{aligned} \text{drogaA} &= 11 \quad \hat{Y} = -53,92 + 16,59 \cdot 11 + 6,22X_2 + 2,77 \cdot 11 \cdot X_2 \\ \hat{Y} &= 128,57 + 36,69X_2 \end{aligned}$$

Debería resultar claro en este caso, que necesitamos conocer el valor de la droga A para poder decir cuánto aumenta la respuesta media al aumentar en una unidad la

Tabla 44: Ajuste del modelo lineal múltiple (81) $E(Y | X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2$, donde $X_1 = \text{drogaA}$, $X_2 = \text{drogaB}$, $Y = \text{respuesta}$.

```
> summary( ajuste5)

Call:
lm(formula = YY ~ drogaA * drogaB)

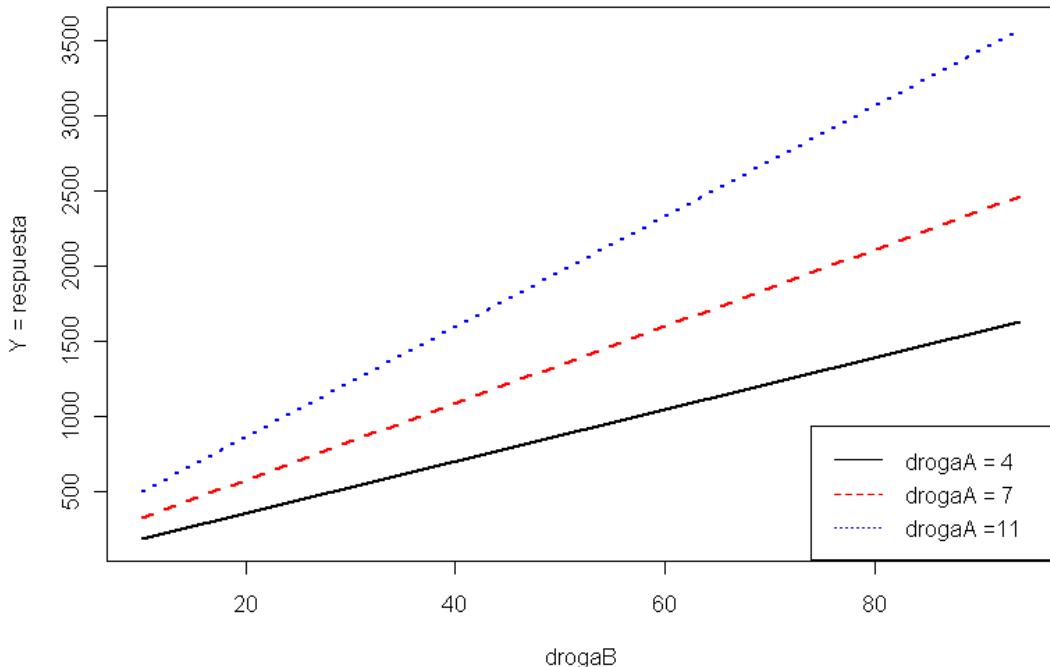
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.92176   42.27242  -1.276  0.21027
drogaA        16.59288   4.92500   3.369  0.00181
drogaB        6.22153   0.63436   9.808 1.04e-11
drogaA:drogaB  2.77152   0.07774  35.651 < 2e-16
---
Residual standard error: 44.04 on 36 degrees of freedom
Multiple R-squared:  0.9979 ,    Adjusted R-squared:  0.9977
F-statistic:  5650 on 3 and 36 DF,  p-value: < 2.2e-16
```

cantidad de droga B consumida. Para los tres valores graficados, tendríamos tres respuestas distintas: 17,3, 25,61 y 36,69 (para $\text{drogaA} = 4, 7$ y 11 , respectivamente).

Hay que tener mucho cuidado en la interpretación de los coeficientes de cada covariable cuando el modelo contiene interacciones. Este modelo es mucho más complicado que el aditivo. Por esta razón, cuando se ajusta un modelo con interacción y no se rechaza la hipótesis de que la interacción sea cero, es mejor eliminar el término de interacción del modelo antes de interpretar los efectos parciales de cada variable. Sin embargo, cuando existe clara evidencia de interacción (se rechaza $H_0 : \beta_{1:2} = 0$), hay que conservar los términos asociados a las variables originales en el modelo lineal, aún cuando no resulten ser significativos, ya que el efecto de cada variable cambia según el nivel de las otras variables, ver (82). Es decir, si en el ajuste presentado en la Tabla 44 la interacción hubiera resultado significativa y el efecto de la droga A no hubiera resultado significativo, de todos modos, debería conservarse la droga A como covariable en el modelo, puesto que se conservará la interacción.

Veamos un ejemplo donde el efecto de la interacción es más fuerte aún.

Figura 55: Variable respuesta Y ajustada en función de la `drogaB`, controlando por `drogaA`, para tres posibles valores de `drogaA` (4, 7 y 11) en el modelo con interacción.



Ejemplo 4.3 El conjunto de datos está en el archivo `ejemploint3.txt`. Nuevamente se trata de datos generados para los que se midieron las tres variables descriptas en el principio de esta sección, es decir, niveles de droga A (X_1), droga B (X_2) y la respuesta (Y). El modelo ajustado figura en la Tabla 45.

Ahí vemos que tanto el coeficiente de la interacción, como los otros dos coeficientes que acompañan a las covariables son significativamente distintos de cero. En la Figura 56 vemos las rectas ajustadas para tres valores fijos de `drogaB`. En ella vemos que el efecto de la interacción cambia de sentido al vínculo entre la respuesta Y y la `drogaA` al aumentar la cantidad de `drogaB`, ya que pasa de ser un potenciador de la variable respuesta, aumentándola considerablemente al aumentar la cantidad de `drogaA`, cuando la cantidad de `drogaB` es 10, a tener un vínculo inverso con Y cuando la cantidad de `drogaB` es 90, en el sentido que a mayor cantidad de `drogaA` la variable respuesta disminuye en este caso. En el caso de `drogaB` = 50,

Tabla 45: Modelo ajustado para los datos del archivo `ejemploint3.txt`, con las variables explicativas $X_1 = \text{drogaA}$ y $X_2 = \text{drogaB}$ y la interacción entre ellas, para explicar a Y .

```
> summary(ajuste7)
Call:
lm(formula = Y7 ~ drogaA * drogaB)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2488.19403   31.27861   79.55 < 2e-16
drogaA       151.87124    3.64415   41.67 < 2e-16
drogaB        4.92268    0.46938   10.49 1.71e-12
drogaA:drogaB -3.00872    0.05752  -52.30 < 2e-16
---
Residual standard error: 32.59 on 36 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9962
F-statistic: 3427 on 3 and 36 DF,  p-value: < 2.2e-16
```

vemos que el vínculo entre `drogaA` y la respuesta desaparece, ya que la recta parece horizontal (la pendiente estimada es exactamente cero cuando $\text{drogaB} = 50,47703$). Las tres rectas graficadas son

$$\begin{aligned} \text{drogaB} &= 10 & \hat{Y} &= 2488,194 + 151,871X_1 + 4,923 \cdot 10 - 3,0087 \cdot X_1 \cdot 10 \\ \hat{Y} &= 2537,4 + 121,78X_1 \end{aligned}$$

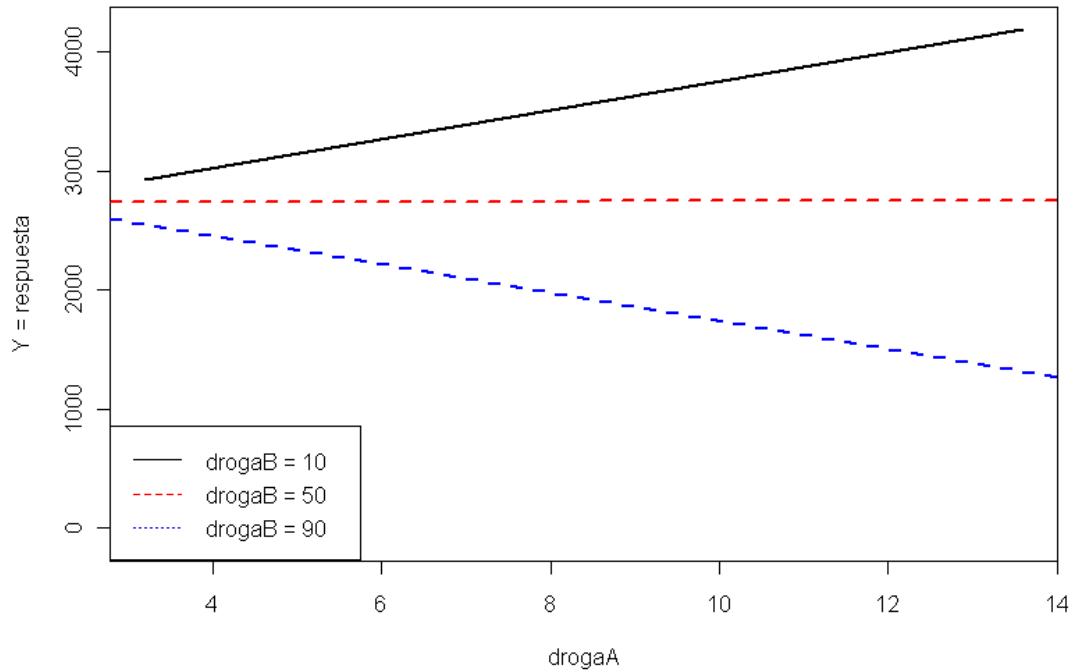
$$\begin{aligned} \text{drogaB} &= 50 & \hat{Y} &= 2488,194 + 151,871X_1 + 4,923 \cdot 50 - 3,0087 \cdot X_1 \cdot 50 \\ \hat{Y} &= 2734,2 + 1,436X_1 \end{aligned}$$

$$\begin{aligned} \text{drogaB} &= 90 & \hat{Y} &= 2488,194 + 151,871X_1 + 4,923 \cdot 90 - 3,0087 \cdot X_1 \cdot 90 \\ \hat{Y} &= 2931,3 - 118,91X_1 \end{aligned}$$

En este caso observamos que para hablar del efecto que tiene en la media de Y el aumento de una unidad de la `drogaA` debemos saber cuál es el valor de la `drogaB` (ya que Y podría crecer, quedar constante o incluso disminuir) con un aumento de una unidad de la `drogaA`. En el modelo aditivo (sin interacción) uno podía siempre cuantificar la variación de la respuesta ante un aumento de

una unidad de una covariante sin necesidad de conocer siquiera el valor de la otra covariante, mientras se mantuviera constante. Decíamos, en el ejemplo de los bebés de bajo peso, que manteniendo el peso constante, el aumento de una semana en la edad gestacional de un bebé repercutía en un aumento de 0,45 cm. del perímetrocefálico esperado del bebé al nacer. Esto vale tanto para bebés que pesan 600 g., 900 g. o 1200 g. al nacer. Cuando hay interacción, esta interpretación se dificulta.

Figura 56: Variable respuesta Y ajustada en función de la drogaA, controlando por drogaB, para tres posibles valores de drogaB (10, 50 y 90) en el modelo con interacción, para los datos de ejemplointer3.txt.



Ejercicio 4.2 Hacer el ejercicio 3 del Taller 3.

Ejercicio 4.3 Hacer el ejercicio 4 del Taller 3.

4.18. Interacción entre dos variables cualitativas

Finalmente restaría presentar un modelo de regresión lineal con interacción entre dos variables cualitativas. Retomemos el ejemplo del pulso post ejercicio.

Ejemplo 4.4 A cuarenta personas se les miden el pulso antes y después de ejercitarse, junto con otras covariables. Estos datos fueron presentados en el Ejemplo 4.1. Para cada individuo, se midieron las siguientes variables

$$Y = \text{pulso luego de correr una milla (Pulso2)}$$

$$X_2 = \begin{cases} 1 & \text{si la persona es mujer} \\ 0 & \text{en caso contrario} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{si la persona fuma} \\ 0 & \text{en caso contrario} \end{cases}$$

Antes de presentar el modelo con interacción, proponemos un modelo aditivo para explicar el pulso post-ejercicio, en función de las covariables X_2 y X_3 . Tanto el sexo como la condición de fumador son variables dummies o binarias. En la base de datos se las denomina $X_2 = \text{mujer}$ y $X_3 = \text{fuma}$.

El modelo (aditivo) es

$$E(Y | X_2, X_3) = \beta_0 + \beta_M \text{mujer} + \beta_F \text{fuma}. \quad (83)$$

Hemos puesto el subíndice de los beta de acuerdo a la variable explicativa que acompañan. En la Tabla 46 escribimos el significado del modelo para las cuatro combinaciones posibles de los valores de X_2 y X_3 .

Tabla 46: Modelo de regresión lineal múltiple aditivo para el pulso post-ejercicio con covariables $X_2 = \text{mujer}$ y $X_3 = \text{fuma}$.

Grupo	$X_2 = \text{mujer}$	$X_3 = \text{fuma}$	$E(Y X_2, X_3)$
1	0	0	β_0
2	0	1	$\beta_0 + \beta_F$
3	1	0	$\beta_0 + \beta_M$
4	1	1	$\beta_0 + \beta_F + \beta_M$

En la Tabla 46 vemos que β_F representa el aumento (o disminución, según el signo) en el pulso medio post ejercicio al comparar el grupo de hombres fumadores con

el grupo de hombres no fumadores (grupo 2 menos grupo 1), pues

$$\begin{aligned} & E(Y \mid \text{mujer} = 0, \text{fuma} = 1) - E(Y \mid \text{mujer} = 0, \text{fuma} = 0) \\ &= (\beta_0 + \beta_F) - \beta_0 \\ &= \beta_F \end{aligned}$$

y también representa el cambio en el pulso medio post ejercicio al comparar el grupo de mujeres fumadoras con el de las mujeres no fumadoras (grupo 4 menos grupo 3), pues

$$\begin{aligned} & E(Y \mid \text{mujer} = 1, \text{fuma} = 1) - E(Y \mid \text{mujer} = 1, \text{fuma} = 0) \\ &= (\beta_0 + \beta_F + \beta_M) - (\beta_0 + \beta_M) \\ &= \beta_F. \end{aligned}$$

Como ambas diferencias dan el mismo número, decimos que β_F representa el cambio en el valor esperado del pulso post-ejercicio por efecto de fumar, cuando se controla (o estratifica) por la variable sexo, o sea, cuando mantenemos la otra variable fija sin importar su valor. Observemos que esta es la misma interpretación que hemos hecho de los coeficientes en los modelos de regresión lineal aditivos. Del mismo modo, β_M representa la diferencia en el pulso medio post-ejercicio entre mujeres y varones, al controlar por la variable **fuma**.

Observemos que este modelo dispone de tres coeficientes β_0 , β_M y β_F para reflejar las medias de cuatro grupos distintos.

¿Cómo es el modelo que explica a Y con X_2 , X_3 y la interacción entre ambas? El modelo es el siguiente

$$E(Y \mid X_2, X_3) = \beta_0 + \beta_M \text{mujer} + \beta_F \text{fuma} + \beta_{M:F} \text{mujer} \cdot \text{fuma}. \quad (84)$$

Como tanto $X_2 = \text{mujer}$ y $X_3 = \text{fuma}$ son variables dicotómicas, el término producto $X_2 \cdot X_3 = \text{mujer} \cdot \text{fuma}$ también resulta ser una variable indicadora o dicotómica, en este caso

$$X_2 \cdot X_3 = \begin{cases} 1 & \text{si la persona es mujer y fuma} \\ 0 & \text{en caso contrario.} \end{cases}$$

Nuevamente, en la Tabla 47, escribimos el significado del modelo para las cuatro combinaciones posibles de los valores de $X_2 = \text{mujer}$ y $X_3 = \text{fuma}$.

Hagamos las mismas comparaciones que hicimos en el modelo aditivo. Comparamos el valor medio de la variable respuesta del grupo 2 con el del grupo 1:

$$\begin{aligned} & E(Y \mid \text{mujer} = 0, \text{fuma} = 1) - E(Y \mid \text{mujer} = 0, \text{fuma} = 0) \\ &= (\beta_0 + \beta_F) - \beta_0 \\ &= \beta_F \end{aligned}$$

Tabla 47: Modelo de regresión lineal múltiple con interacción, para el pulso post-ejercicio con covariables $X_2 = \text{mujer}$ y $X_3 = \text{fuma}$.

Grupo	$X_2 = \text{mujer}$	$X_3 = \text{fuma}$	$X_2 \cdot X_3$	$E(Y X_2, X_3)$
1	0	0	0	β_0
2	0	1	0	$\beta_0 + \beta_F$
3	1	0	0	$\beta_0 + \beta_M$
4	1	1	1	$\beta_0 + \beta_F + \beta_M + \beta_{M:F}$

Ahora comparemos los valores medios de la respuesta en los grupos 4 y 3:

$$\begin{aligned} & E(Y | \text{mujer} = 1, \text{fuma} = 1) - E(Y | \text{mujer} = 1, \text{fuma} = 0) \\ &= (\beta_0 + \beta_M + \beta_F + \beta_{M:F}) - (\beta_0 + \beta_M) \\ &= \beta_F + \beta_{M:F}. \end{aligned}$$

Por lo tanto, β_F mide el efecto de fumar en los hombres, y $\beta_F + \beta_{M:F}$ mide el efecto de fumar en las mujeres. De modo que el término de la interacción $\beta_{M:F}$ da la diferencia del pulso medio post-ejercicio por efecto de fumar de las mujeres respecto de los hombres. Si $\beta_{M:F} > 0$, el hecho de fumar en las mujeres redunda en un aumento de la respuesta media respecto de la de los hombres. Un test de $H_0 : \beta_{M:F} = 0$ versus $H_1 : \beta_{M:F} \neq 0$ para el modelo (84) es una prueba para la igualdad del efecto de fumar en el pulso medio post-ejercicio de hombres y mujeres. Observemos que si no se rechaza H_0 , tenemos un modelo aditivo: el efecto de fumar en el pulso post-ejercicio resulta ser el mismo para hombres y mujeres.

También se podrían tomar diferencias análogas entre los grupos 1 y 3 (no fumadores) y entre los grupos 4 y 2 (fumadores) y llegar a la misma interpretación de la interacción. En este ejemplo, esta aproximación parece menos intuitiva, ya que interesa evaluar el efecto de fumar (controlando por el sexo) en la respuesta.

Antes de pasar a los ajustes de los modelos, propongamos un modelo de comparación de las medias de cuatro muestras aleatorias normales, todas con la misma varianza (o sea, una generalización del test de t para de dos muestras). Tal modelo, propondría que se tienen 4 muestras de pulso post-ejercicio tomadas en 4 grupos diferentes (en este caso, los definidos en la primer columna de la Tabla 47) y para cada uno de ellos proponemos

$$\begin{aligned} Y_{i1} &\sim N(\mu_1, \sigma^2) \quad (1 \leq i \leq n_1) && \text{grupo 1 (hombres no fumadores)} \quad (85) \\ Y_{i2} &\sim N(\mu_2, \sigma^2) \quad (1 \leq i \leq n_2) && \text{grupo 2 (hombres fumadores)} \\ Y_{i3} &\sim N(\mu_3, \sigma^2) \quad (1 \leq i \leq n_3) && \text{grupo 3 (mujeres fumadoras)} \\ Y_{i4} &\sim N(\mu_4, \sigma^2) \quad (1 \leq i \leq n_4) && \text{grupo 4 (mujeres no fumadoras)}. \end{aligned}$$

Todas las observaciones son independientes entre sí. Este modelo propone ajustar 4 parámetros que dan cuenta de la media (uno para cada grupo, que hemos denominado μ_k que se estimarán con las observaciones del respectivo grupo k -ésimo) y un parámetro que da cuenta de la varianza de cada observación en el modelo homoscedástico (σ^2 que se estimará de forma conjunta con todas las $n_1 + n_2 + n_3 + n_4$ observaciones). Si comparamos este modelo con el propuesto en (84), vemos que ambos tienen 4 parámetros para las medias. Más aún, resultará que se vinculan de la siguiente forma, por lo desarrollado en la Tabla 47.

$$\begin{aligned}\mu_1 &= \beta_0 \\ \mu_2 &= \beta_0 + \beta_F \\ \mu_3 &= \beta_0 + \beta_M \\ \mu_4 &= \beta_0 + \beta_F + \beta_M + \beta_{F:M}.\end{aligned}\tag{86}$$

Otra forma de escribir el modelo (85) es la siguiente

$$Y_{i1} = \mu_1 + \epsilon_{i1} \quad (1 \leq i \leq n_1) \quad \text{grupo 1 (hombres no fumadores)}\tag{87}$$

$$Y_{i2} = \mu_2 + \epsilon_{i2} \quad (1 \leq i \leq n_2) \quad \text{grupo 2 (hombres fumadores)}$$

$$Y_{i3} = \mu_3 + \epsilon_{i3} \quad (1 \leq i \leq n_3) \quad \text{grupo 3 (mujeres fumadoras)}$$

$$Y_{i4} = \mu_4 + \epsilon_{i4} \quad (1 \leq i \leq n_4) \quad \text{grupo 4 (mujeres no fumadoras)},$$

donde los $\epsilon_{ik} \sim N(0, \sigma^2)$ y son todos independientes

Vemos pues que ambos modelos (84) y (85) son equivalentes, ya que conociendo los parámetros de uno de ellos (los μ_k por ejemplo) podemos despejar los valores del otro (los β_h por ejemplo) por medio de las ecuaciones (86). O al revés, obtener los μ_k a partir de los β_h . La varianza del error se estimará en forma conjunta en ambos modelos. La diferencia está en el significado de los parámetros. En el modelo (85), μ_k representa el valor esperado de la variable respuesta en el grupo k -ésimo, mientras que en el modelo (84) los β_h representan (algunas de) las diferencias entre los valores de las respuestas medias entre los distintos grupos.

En las Tablas 48 y 49 se muestran los valores ajustados de los modelos aditivos (83) y con interacción (84).

Analicemos primero el modelo con interacción. En la salida vemos que el coeficiente de la interacción no resulta significativo (el p-valor es 0,245 que no es menor a 0,05), por lo tanto concluimos que el efecto de fumar en el pulso medio post-ejercicio de mujeres y varones es el mismo. Luego, para los datos del pulso el modelo apropiado es el aditivo (83). En dicho ajuste vemos que todos los coeficientes son significativos, y que el hecho de fumar aumenta el pulso post-ejercicio en 7,36 pulsaciones por minuto, cuando uno controla por sexo. Es interesante graficar

Tabla 48: Ajuste del modelo lineal múltiple aditivo $Y_i = \beta_0 + \beta_M X_{i2} + \beta_F X_{i3} + \varepsilon_i$, donde X_2 = indicador de mujer (`mujer`), X_3 = indicador de fumar (`fuma`), e Y = pulso post ejercicio (`Pulso2`).

```
> ajusteA<-lm(Pulso2 ~ mujer + fuma)
> summary(ajusteA)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	126.926	2.452	51.754	< 2e-16
mujer	18.064	3.027	5.967	6.96e-07
fuma	7.362	3.074	2.395	0.0218
<hr/>				

```
Residual standard error: 9.453 on 37 degrees of freedom
Multiple R-squared: 0.5093, Adjusted R-squared: 0.4828
F-statistic: 19.2 on 2 and 37 DF, p-value: 1.906e-06
```

las cuatro medias muestrales y los cuatro valores esperados bajo el modelo. Esos valores figuran en la Tabla 50.

Mirando la Tabla 50 podemos corroborar que los estimadores obtenidos con el modelo con interacción son los mismos que obtendríamos si estimáramos las medias de cada grupo por separado. En este caso además, vemos que el ajuste obtenido por el modelo sin interacción no difiere demasiado del con interacción, en sus valores ajustados, es por eso que la interacción no resulta significativa en este modelo. El Gráfico 57 permite visualizar más claramente la situación. En él vemos que al pasar del grupo de no fumadores al grupo de fumadores, aumenta el pulso medio post-ejercicio, tanto en hombres como en mujeres, siempre en una cantidad parecida (tan parecida, que la diferencia entre ambos no es estadísticamente significativa). Este gráfico suele llamarse gráfico de interacción. Sirve para evaluar si tiene sentido ajustar un modelo con interacción a nuestros datos. Si dicho gráfico resultara como se muestra en alguno de los dos de la Figura 58, entonces se justificaría agregar el término de interacción al modelo con dos covariables categóricas. En el gráfico A vemos un ejemplo donde al pasar del grupo no fumador al grupo fumador, para las mujeres se produce un aumento de la respuesta media, y para los hombres una disminución de la respuesta media. Para este ejemplo, tiene sentido incluir el término de la interacción, ya que la respuesta cambia de sentido para distintas combinaciones de las dos explicativas. En el gráfico B sucede algo parecido: cuando controlamos por el sexo de la persona, el efecto de fumar es diferente en los dos

Tabla 49: Ajuste del modelo lineal múltiple con interacción $Y_i = \beta_0 + \beta_M X_{i2} + \beta_F X_{i3} + \beta_{M:F} X_{i2} \cdot X_{i3} + \varepsilon_i$, donde X_2 = indicador de mujer (**mujer**), X_3 = indicador de fumar (**fuma**), Y = pulso post ejercicio (Pulso2).

```
> ajusteB <- lm(Pulso2 ~ mujer * fuma)
> summary(ajusteB)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 128.333     2.714   47.280 < 2e-16
mujer        15.250     3.839   3.973 0.000326
fuma         4.267     4.026   1.060 0.296306
mujer:fuma   7.317     6.190   1.182 0.244922
---
Residual standard error: 9.403 on 36 degrees of freedom
Multiple R-squared:  0.5276,    Adjusted R-squared:  0.4883
F-statistic: 13.4 on 3 and 36 DF,  p-value: 4.978e-06
```

grupos, para las mujeres aumenta la media de la respuesta, para los hombres la deja igual.

4.19. Generalización a más de dos variables.

Cuando el número de variables regresoras es mayor que dos, se pueden incluir términos de interacción para cada par de covariables. Por ejemplo, en un modelo con tres variables regresoras X_1 , X_2 y X_3 , podemos tener:

$$\begin{aligned} E(Y | X_1, X_2, X_3) = & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ & + \beta_{1:2} X_1 \cdot X_2 + \beta_{1:3} X_1 \cdot X_3 + \beta_{2:3} X_2 \cdot X_3 \end{aligned}$$

En este modelo hemos considerado las interacciones de las variables tomadas de a pares, a las que se denomina interacciones de segundo orden. Pero podríamos haber considerado además la interacción de tercer orden incorporando un término $\beta_{1:2:3} X_1 \cdot X_2 \cdot X_3$.

A partir de los tests de significación podremos evaluar si alguna(s) de estas interacciones son necesarias en el modelo.

Tabla 50: Medias muestrales calculadas por grupos, comparadas con el ajuste de los modelos sin y con interacción, para el pulso post-ejercicio con covariables $X_2 = \text{mujer}$ y $X_3 = \text{fuma}$.

Grupo	X_2	X_3	Media muestral	$E(Y X_2, X_3)$ sin interacción
1	0	0	128,3333	$\hat{\beta}_0 = 126,926$
2	0	1	132,6	$\hat{\beta}_0 + \hat{\beta}_F = 126,926 + 7,362 = 134,29$
3	1	0	143,5833	$\hat{\beta}_0 + \hat{\beta}_M = 126,926 + 18,064 = 144,99$
4	1	1	155,1667	$\hat{\beta}_0 + \hat{\beta}_F + \hat{\beta}_M = 126,926 + 7,362 + 18,064 = 152,35$

Grupo	X_2	X_3	Media muestral	$E(Y X_2, X_3)$ con interacción
1	0	0	128,3333	$\hat{\beta}_0 = 128,333$
2	0	1	132,6	$\hat{\beta}_0 + \hat{\beta}_F = 128,333 + 4,267 = 132,6$
3	1	0	143,5833	$\hat{\beta}_0 + \hat{\beta}_M = 128,333 + 15,25 = 143,58$
4	1	1	155,1667	$\hat{\beta}_0 + \hat{\beta}_F + \hat{\beta}_M + \hat{\beta}_{F:M} = 128,333 + 4,267 + 15,25 + 7,317 = 155,1667$

Cuando alguna interacción es significativa y el modelo debe incluir estos términos, es más compleja la presentación de los resultados. Una aproximación posible es graficar una colección de rectas como en las figuras anteriores, para describir gráficamente cómo cambia la relación con los valores de las demás variables.