Regresión simple y múltiple

Regresión: modelos y métodos

Francesc Carmona Pontaque

PID_00298353

Universitat Oberta de Catalunya

Francesc Carmona Pontaque

Cómo citar este recurso de aprendizaje con el estilo Harvard:

Carmona Pontaque, F. (2024) *Regresión simple y múltiple. Regresión: modelos y métodos*. [Recurso de aprendizaje textual]. 1.ª ed. Barcelona: Fundació Universitat Oberta de Catalunya (FUOC).

Primera edición: febrero 2024

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Francesc Carmona Pontaque
Producción: FUOC

Planteamiento y objetivos

En este módulo concretaremos la estimación de parámetros, la inferencia estadística y otros elementos particulares de la regresión lineal simple y la regresión lineal múltiple.

Para la regresión simple, en el libro de Carmona (2005) se muestran las fórmulas exactas de la pendiente y el punto de intercepción. Sin embargo, **R** proporciona estos coeficientes y otros elementos del modelo de forma automática.

En cuanto a la medida de ajuste, se define el *coeficiente de determinación* que coincide con el coeficiente de correlación al cuadrado, aunque conceptualmente son distintos. El coeficiente de determinación es una medida de ajuste y, por lo tanto, siempre tiene sentido. En cambio, el coeficiente de correlación solo tiene sentido entre variables aleatorias y su versión muestral entre muestras de variables aleatorias.

En el caso de la regresión múltiple hay que tener en cuenta el coeficiente de determinación ajustado.

Considerar como modelo la recta que pasa por el origen de coordenadas, es decir, sin intercepción, depende exclusivamente del modelo científico subyacente y no de la estadística.

Otro punto que trataremos aquí es el ejemplo de Anscombe (1973), que sirve para evidenciar que la regresión depende de un modelo que hay que estudiar y validar, más allá de los resultados numéricos.

El tema de la comparación de rectas de regresión se puede tratar como un contraste de modelos con matrices de diseño o como un modelo de análisis de la covarianza (ANCOVA). Este último modelo se presentará con más detalle al final de la asignatura.

Finalmente, un tema importante en este módulo es todo lo que concierne al cálculo de predicciones y sus diferentes intervalos de confianza: predicción media y predicción concreta. También se trata el problema de la extrapolación en regresión múltiple.

1. Regresión lineal simple

Sea Y una variable aleatoria y x una variable controlable, es decir, los valores que toma x los fija la persona que lleva a cabo el experimento. Luego calculamos Y para diferentes valores de x y obtenemos un conjunto de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Supongamos que estos valores se ajustan al modelo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

donde los errores ϵ_i verifican las condiciones de Gauss-Markov.

Este modelo es la formulación lineal del problema de hallar la recta de regresión de Y sobre x. Un ejemplo lo tenemos con los datos de los buitres leonados del Módulo 1. Los parámetros β_0 y β_1 reciben el nombre de coeficientes de regresión. El parámetro β_0 es la ordenada en el origen o intercepto, *intercept* en inglés, y β_1 es la pendiente de la recta, *slope* en inglés. Como sabemos, la expresión matricial de modelo es

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \qquad \text{rg } \mathbf{X} = 2$$

La matriz de diseño

La matriz de diseño $\mathbf{X}$ y algunas de sus propiedades han sido estudiadas en el apéndice B del Módulo 1: El modelo lineal.

En los módulos anteriores ya hemos visto cómo se aplica la teoría general de los modelos lineales al caso particular de la regresión lineal simple. Haremos ahora un resumen.

La estimación de los parámetros por el método de los mínimos cuadrados es

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x} = \frac{s_{xy}}{s_x^2}$$

Estimaciones MC

Estas estimaciones son la solución de las ecuaciones normales que hemos visto en el apéndice A del Módulo 1.

donde

$$S_{xy} = \sum x_i y_i - (1/n) \sum x_i \sum y_i = \sum (x_i - \bar{x})(y_i - \bar{y}) = n\, s_{xy}$$

$$S_x = \sum x_i^2 - (1/n)(\sum x_i)^2 = \sum (x_i - \bar{x})^2 = n\, s_x^2$$

Estas estimaciones son insesgadas y de varianza mínima entre todos los estimadores lineales (teorema de Gauss-Markov). Las varianzas y covarianza de los estimadores se calculan a partir de la matriz inversa de $\mathbf{X}'\mathbf{X}$

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{var}(\hat{\beta}_1) \end{pmatrix} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Es decir

$$E(\hat{\beta}_0) = \beta_0 \qquad \text{var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x}\right)$$

$$E(\hat{\beta}_1) = \beta_1 \qquad \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_x}$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2\frac{\bar{x}}{S_x}$$

Para calcularlas, habrá que sustituir el parámetro desconocido σ^2 por su estimación $\widehat{\sigma}^2 = \text{ECM}$. Con los datos de los buitres leonados, las varianzas y covarianza de los estimadores son

```
mod <- lm(metabol ~ heartbpm, data=vulture)
ss <- summary(mod)
ss$sigma^2 * ss$cov.unscaled
```

```
             (Intercept)      heartbpm
(Intercept)  0.719281950 -0.0099460212
heartbpm    -0.009946021  0.0001464929
```

Las raíces cuadradas de los elementos de la diagonal proporcionan las desviaciones estándar de cada estimación.

2. Medidas de ajuste

La evaluación global del ajuste de la regresión se puede hacer con la SCR o, mejor aún, con la varianza muestral de los residuos $(1/n)\sum e_i^2$. Pero los residuos no son todos independientes, sino que están ligados por dos ecuaciones ($\sum e_i = 0$ y $\sum x_i e_i = 0$), de forma que utilizaremos la llamada *varianza residual* o estimación MC de σ^2:

$$\hat{\sigma}^2 = \text{SCR}/(n-2) = \text{ECM}$$

Su raíz cuadrada $\hat{\sigma}$, que tiene las mismas unidades que Y, es el llamado *error estándar de la regresión*.

La varianza residual o el error estándar son índices de la precisión del modelo, pero dependen de las unidades de la variable respuesta y no son útiles para comparar rectas de regresión de variables diferentes. Así pues, como medida de ajuste utilizaremos el llamado *coeficiente de determinación* gracias a la siguiente descomposición de la variabilidad de la variable respuesta.

Demostración

La demostración se puede ver en el apartado 6.2. de Carmona (2005).

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$$

Esta descomposición de la suma de cuadrados de las observaciones en dos términos independientes se interpreta así: la variabilidad de la variable Y se descompone en un primer término que refleja la variabilidad no explicada por la regresión, que es debida al azar. El segundo término, en cambio, contiene la variabilidad explicada o eliminada por la regresión y puede interpretarse como la parte determinista de la variabilidad de la respuesta.

Podemos definir:

$$\text{Variación total} = \text{VT} = \sum(y_i - \bar{y})^2 = S_y$$

$$\text{Variación no explicada} = \text{VNE} = \sum(y_i - \hat{y}_i)^2 = \text{SCR}$$

$$\text{Variación explicada} = \text{VE} = \sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 S_x$$

de forma que $\text{VT} = \text{VNE} + \text{VE}$. Finalmente, como medida del ajuste de la recta de regresión a los datos, definimos la proporción de variabilidad explicada con el nombre de *coeficiente de determinación* así:

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{SCR}}{S_y}$$

Por otra parte, también es posible calcular el *coeficiente de correlación muestral*, cuyo significado es convencional:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{S_{xy}}{(S_x S_y)^{1/2}}$$

De modo que $\text{SCR} = (1 - r^2)S_y$ y entonces

$$R^2 = 1 - \frac{(1 - r^2)S_y}{S_y} = r^2$$

Coeficiente de correlación lineal

que es el cuadrado del coeficiente de correlación lineal entre las dos variables.

¡Alerta!

El coeficiente de correlación muestral r es un estimador del coeficiente de correlación poblacional $\rho(X, Y)$, que únicamente tiene sentido entre variables aleatorias. Es decir, si la variable regresora está definida por la persona que experimenta, no es aleatoria, luego ρ no existe, r no es un estimador y no debería darse.

El coeficiente de determinación R^2 se puede utilizar en cualquier tipo de regresión, puesto que es una medida de la bondad del ajuste, $0 \leq R^2 \leq 1$, mientras que el coeficiente de correlación es una medida de la dependencia lineal entre dos variables aleatorias.

Con los datos de los buitres leonados, el coeficiente de determinación es

```
ss$r.squared
```

```
[1] 0.9688697
```

Más adelante definiremos el coeficiente de determinación "ajustado" que únicamente tiene sentido en la regresión múltiple.

Advertencia:

Un R^2 elevado no es una garantía de que el modelo sea válido. Ver los ejemplos en el gráfico 2.3. de Faraway (2014).
Para comprobar que el modelo verifica las condiciones de Gauss-Markov debemos realizar una completa diagnosis con los residuos.

3. Intervalo para la respuesta media

Nota

Si el valor de x_0 está fuera del recorrido de x, entonces diremos que se trata de una extrapolación para señalar que la predicción no es fiable.

Uno de los usos principales de los modelos de regresión es la estimación de la respuesta media $E[Y|x_0]$ para un valor particular x_0 de la variable regresora. Asumiremos que x_0 es un valor dentro del recorrido de los datos originales de x. Un estimador puntual insesgado de $E[Y|x_0]$ se obtiene con la predicción

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x})$$

Podemos interpretar $\beta_0 + \beta_1 x_0$ como una función paramétrica estimable

$$\beta_0 + \beta_1 x_0 = (1, x_0)\boldsymbol{\beta} = \mathbf{x}_0'\boldsymbol{\beta}$$

cuyo estimador es $\hat{y}_0 = \mathbf{x}_0'\widehat{\boldsymbol{\beta}}$, de manera que

$$\text{var}(\mathbf{x}_0'\widehat{\boldsymbol{\beta}}) = \sigma^2\, \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

y el error estándar de $\mathbf{x}_0'\widehat{\boldsymbol{\beta}}$ es

$$\text{ee}(\mathbf{x}_0'\widehat{\boldsymbol{\beta}}) = [\hat{\sigma}^2(1/n + (x_0 - \bar{x})^2/S_x)]^{1/2}$$

Entonces, el intervalo de confianza para la respuesta media $E[Y|x_0]$ es

$$\hat{y}_0 \pm t_{n-2}(\alpha) \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x}}$$ error estándar

Destacaremos el hecho de que evidentemente el ancho del intervalo depende de x_0, es mínimo para $x_0 = \bar{x}$ y crece cuando $|x_0 - \bar{x}|$ crece. Esto es intuitivamente razonable.

La predicción de `metabol` cuando `heartbpm` es 70 se hace así:

```
as.numeric(coef(mod)[1] + coef(mod)[2]*70)
```

```
[1] 11.69152
```

o mejor así

```
predict(mod, newdata = data.frame(heartbpm=70))
```

Si lo que queremos es el intervalo de confianza para la respuesta media, entonces

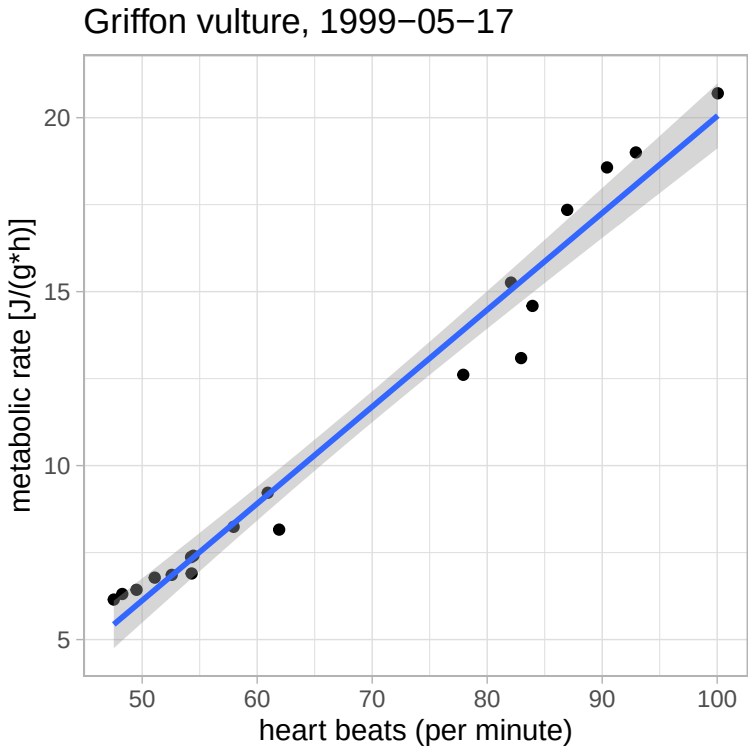
```
predict(mod, newdata = data.frame(heartbpm=70),
              interval = "confidence")


       fit      lwr      upr
1 11.69152 11.24568 12.13736
```

Con el paquete `ggplot2` disponemos de la función `geom_smooth()`, que tiene un parámetro para añadir a la recta de regresión unas curvas que representan los intervalos de confianza para la respuesta media de todos los valores de x. Así, a la figura 1 del primer módulo le podemos añadir estos intervalos de confianza para la respuesta media.

¡Atención!

No son bandas de confianza. Las curvas son los límites inferior y superior del intervalo para cada valor de x en vertical.

```
p + geom_smooth(method=lm, se=TRUE)
```



4. Predicción de nuevas observaciones

Otra de las importantes aplicaciones de los modelos de regresión es la predicción de nuevas observaciones para un valor x_0 de la variable regresora. El intervalo definido en el apartado anterior es adecuado para el valor esperado de la respuesta, pero ahora queremos un intervalo de predicción para una respuesta individual concreta. Estos intervalos reciben el nombre de intervalos de predicción en lugar de intervalos de confianza, ya que se reserva el nombre de intervalo de confianza para los que se construyen como estimación de un parámetro. Los intervalos de predicción tienen en cuenta la variabilidad en la predicción del valor medio y la variabilidad al exigir una respuesta individual.

Si x_0 es el valor de nuestro interés, entonces

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

es el estimador puntual de un nuevo valor de la respuesta $Y_0 = Y|x_0$.

Si consideramos la obtención de un intervalo de confianza para esta futura observación Y_0, el intervalo de confianza para la respuesta media en $x = x_0$ es inapropiado, ya que es un intervalo sobre la *media* de Y_0 (un parámetro), no sobre futuras observaciones de la distribución.

Se puede hallar un intervalo de predicción para una respuesta concreta de Y_0 del siguiente modo:

Consideremos la variable aleatoria $Y_0 - \hat{y}_0 \sim N(0, \text{var}(Y_0 - \hat{y}_0))$ donde

$$\text{var}(Y_0 - \hat{y}_0) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x} \right)$$

ya que Y_0, una futura observación, es independiente de $\hat{y}_0$.

Si utilizamos el valor muestral de $\hat{y}_0$ para predecir Y_0, obtenemos un intervalo de predicción al $100(1-\alpha)$ % para Y_0, tal como se muestra a continuación:

$$\hat{y}_0 \pm t_{n-2}(\alpha) \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x}}$$

El intervalo para una predicción concreta de `metabol` cuando `heartbpm` es 70 se hace así:

```
predict(mod, newdata = data.frame(heartbpm=70),
              interval = "prediction")


       fit      lwr      upr
1 11.69152 9.711502 13.67154
```

Al gráfico del apartado anterior con los intervalos de confianza para la respuesta media, también le podemos añadir los intervalos de predicción.

```
pred_var <- predict(mod, interval="prediction")
pred_var <- pred_var[order(pred_var[,"fit"]),]
new_df <- cbind(vulture, pred_var)
ggplot(new_df, aes(x=heartbpm, y=metabol)) +
    geom_point() + labs(x = "heart beats (per minute)",
                        y = "metabolic rate [J/(g*h)]",
                        title = "Griffon vulture, 1999-05-17") +
    geom_line(aes(y=lwr), color = "blue", linetype = "dashed") +
    geom_line(aes(y=upr), color = "blue", linetype = "dashed") +
    geom_smooth(method=lm, se=TRUE) +
    theme_light()
```

5. Recta que pasa por el origen

Cuidado con este modelo --> contraste de hipótesis

Supongamos que, por alguna razón justificada, la persona que experimenta decide proponer el modelo de regresión simple sin el término β_0.

$$y_i = \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

El estimador MC del parámetro β_1 es $\hat{\beta}_1 = (\sum x_i y_i)/(\sum x_i^2)$ y su varianza es

$$\text{var}(\hat{\beta}_1) = \frac{1}{(\sum x_i^2)^2} \sum x_i^2 \text{var}(y_i) = \sigma^2 \frac{1}{\sum x_i^2}$$

El estimador de σ^2 es $\hat{\sigma}^2 = \text{SCR}/(n-1)$.

Con la hipótesis de normalidad se pueden construir intervalos de confianza al $100(1-\alpha)$ % para β_1

$$\hat{\beta}_1 \pm t_{n-1}(\alpha) \cdot \hat{\sigma} \sqrt{\frac{1}{\sum x_i^2}}$$

para $E[Y|x_0]$

$$\hat{y}_0 \pm t_{n-1}(\alpha) \cdot \hat{\sigma} \sqrt{\frac{x_0^2}{\sum x_i^2}}$$

y para predecir una futura observación

$$\hat{y}_0 \pm t_{n-1}(\alpha) \cdot \hat{\sigma} \sqrt{1 + \frac{x_0^2}{\sum x_i^2}}$$

Es preciso tener mucha seguridad para utilizar este modelo. Frecuentemente la relación entre la variable respuesta Y y la variable regresora x varía cerca del origen. Hay ejemplos en química y en otras ciencias. El diagrama de dispersión nos puede ayudar a decidir el mejor modelo. Si no tenemos plena confianza, es mejor utilizar el modelo completo y contrastar la hipótesis $H_0 : \beta_0 = 0$.

Una medida del ajuste del modelo a los datos es el error cuadrático medio ECM_0, que se puede comparar con el del modelo completo ECM. El coefi-

ciente de determinación R^2 no es un buen índice para comparar los dos tipos de modelos. Para el modelo sin β_0, la descomposición

$$\sum y_i^2 = \sum (y_i - \hat{y}_i)^2 + \sum \hat{y}_i^2$$

justifica que la definición del coeficiente de determinación sea

$$R_0^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$$

Nota

Faraway (2014, p. 24) sugiere utilizar $\text{cor}^2(\hat{y}, y)$ para recuperar el R^2 original.

que no es comparable con el R^2 del modelo completo (con β_0). De hecho, puede ocurrir que $R_0^2 > R^2$, aunque $\text{ECM}_0 < \text{ECM}$.

6. Carácter lineal de la regresión simple

En algunos experimentos es posible contrastar si la regresión de Y sobre x es realmente lineal o hay una falta de ajuste (en inglés, *lack of fit*). Consideremos las hipótesis

$$H_0 : Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$H_1 : Y_i = g(x_i) + \epsilon_i$$

donde $g(x)$ es una función no lineal desconocida de x. Sin embargo, podemos reconducir el contraste a la situación prevista por el test F para la elección entre dos modelos lineales.

Lecturas complementarias

Más detalles y ejemplos se pueden leer en el apartado 8.3. del libro de Faraway (2014) y en el apartado 6.6. de Carmona (2005).

Necesitamos n_i valores de Y para cada x_i. Si indicamos $\delta_i = g(x_i), i = 1, \ldots, k$ convertimos la hipótesis H_1 en una hipótesis lineal con k parámetros. Cuando H_1 es cierta, la estimación de δ_i es la media de cada grupo de respuestas para el mismo x_i, $\hat{\delta}_i = \bar{y}_i$.

En R

En **R** este cambio es muy sencillo, ya que basta con transformar el vector de datos respuesta a `factor`.

En el siguiente ejemplo tenemos datos disponibles sobre el efecto de un suplemento dietético en las tasas de crecimiento de unas ratas. Aquí la variable regresora es la dosis de suplemento dietético y la respuesta es la tasa de crecimiento.

```
supplement <- c(10,10,15,15,20,20,25,25,25,30,35,35)
rate <- c(73,78,85,88,90,91,87,86,91,75,65,63)
g0 <- lm(rate ~ supplement)
g <- lm(rate ~ factor(supplement))
anova(g0,g)


Analysis of Variance Table

Model 1: rate ~ supplement
Model 2: rate ~ factor(supplement)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     10 891.73
2      6  33.50  4    858.23 38.428 0.0002061
```

En este caso hay una falta de ajuste, ya que rechazamos la hipótesis nula que corresponde a la regresión lineal simple.

7. Comparación de rectas de regresión

El ejemplo que explicamos a continuación se basa en un estudio realizado por Alan Pearson, veterinario del Animal Health Laboratory, Lincoln, Nueva Zelanda y se puede hallar en el libro de Saville y Wood (1997).

El experimento tenía como objetivo determinar si el programa estándar de desparasitado por vía oral en 6 granjas de cabras era adecuado. Para ello se seleccionaron 40 cabras en cada granja. Veinte de ellas, elegidas completamente al azar, se desparasitaron con el programa estándar, mientras que las veinte restantes se desparasitaron con más frecuencia. Las cabras se pesaron al principio y al final del estudio, el cual duró un año. Para nuestro ejemplo hemos tomado los datos de una única granja y se pueden cargar en **R** con las instrucciones del apéndice A. Así pues, las variables consideradas son:

- Aumento de peso en vivo (kg.)
- Peso al inicio (kg.)
- Tratamiento: estándar o intensivo

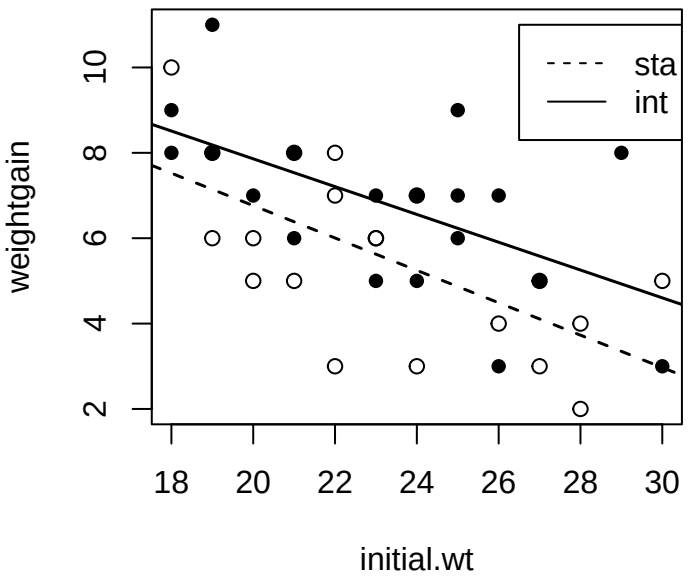


Figura 1

Gráfico de dispersión de los datos de las cabras con las dos rectas de regresión por separado. Las instrucciones se hallan en el apéndice A.

Aunque se puede comparar el aumento de peso sin tener en cuenta el peso inicial, es mucho mejor comparar los dos tratamientos a través de las rectas de regresión que relacionan las dos variables cuantitativas. El gráfico de dispersión de la figura 1 nos permite ver la situación.

Así pues, necesitamos un método para comparar dos rectas de regresión asociadas a dos poblaciones o grupos distintos.

Supongamos, como es el caso, que disponemos de dos muestras independientes de tamaños n_1 y n_2

$$(x_{11}, y_{11}), (x_{12}, y_{12}), \ldots, (x_{1n_1}, y_{1n_1})$$

$$(x_{21}, y_{21}), (x_{22}, y_{22}), \ldots, (x_{2n_1}, y_{2n_1})$$

sobre la misma variable regresora x y la misma variable respuesta Y con distribución normal, pero para dos poblaciones distintas.

Los dos modelos de regresión simple para las dos poblaciones por separado son

$$y_{1i} = \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \qquad i = 1, \ldots, n_1$$

$$y_{2i} = \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \qquad i = 1, \ldots, n_2$$

Si queremos estudiar contrastes sobre todos los parámetros, debemos construir un único modelo lineal conjunto. Para ello hacemos

$$\mathbf{Y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2})'$$

y

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & x_{11} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{1n_1} & 0 \\ 0 & 1 & 0 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

donde $\mathbf{X}$ es $(n_1 + n_2) \times 4$ de $\text{rg}(\mathbf{X}) = 4$.

Así pues, el modelo que presenta las dos regresiones simples conjuntamente $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ es un modelo lineal siempre que los errores verifiquen las condiciones de Gauss-Markov. En particular, es necesario suponer que las varianzas de los errores para las dos poblaciones son iguales $\sigma_1^2 = \sigma_2^2$. Para este modelo lineal, las estimaciones MC de los parámetros $\alpha_1, \alpha_2, \beta_1, \beta_2$ coinciden con las estimaciones MC de las rectas por separado $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2$.

Para contrastar la hipótesis de homogeneidad de varianzas $H_0 : \sigma_2^2 = \sigma_2^2$ podemos utilizar el estadístico

$$F = \frac{\text{SCR}_1/(n_1 - 2)}{\text{SCR}_2/(n_2 - 2)} \sim F_{n_1-2, n_2-2}$$

y la estimación de la varianza común es

$$\text{ECM} = \text{SCR}/(n_1 + n_2 - 4)$$

Nota

Este test F de comparación de varianzas supone que hay normalidad. En caso de duda, también se puede aplicar un test de Levene sobre los residuos de cada recta (Faraway, 2014, p. 228).

Con los datos de las cabras, el test F de homogeneidad de varianzas no es significativo (ver apéndice A) y podemos considerar el modelo común.

Para comparar las dos rectas podemos empezar por el llamado test de paralelismo, es decir, el contraste sobre la igualdad de las pendientes $H_0 : \beta_1 = \beta_2$. En **R** este contraste es muy sencillo. En primer lugar, definimos el modelo conjunto. Observemos que este no tiene intercepto.

Lectura complementaria

En el apartado 6.7. de Carmona (2005) se explicitan las fórmulas exactas para los contrastes de paralelismo y de coincidencia. Sin embargo, en la práctica resultan más sencillos con el contraste de modelos explicado aquí.

```
n1 <- n2 <- 20
x1 <- c(rep(1,n1),rep(0,n2))
x2 <- c(rep(0,n1),rep(1,n2))
x3 <- c(initial.wt[treatment=="standard"],rep(0,n2))
x4 <- c(rep(0,n1),initial.wt[treatment=="intensive"])
g1 <- lm(weightgain ~ 0 + x1 + x2 + x3 + x4)
```

Ahora el modelo bajo la hipótesis de paralelismo es

```
g2 <- lm(weightgain ~ 0 + x1 + x2 + I(x3 + x4))
```

A continuación, hacemos el contraste

```
anova(g2,g1)


Analysis of Variance Table

Model 1: weightgain ~ 0 + x1 + x2 + I(x3 + x4)
```

```
Model 2: weightgain ~ 0 + x1 + x2 + x3 + x4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     37 96.857
2     36 96.514  1   0.34225 0.1277  0.723
```

Una vez aceptamos el paralelismo, el siguiente contraste es la igualdad total, es decir, solo tenemos una recta.

```
g3 <- lm(weightgain ~ initial.wt)
anova(g3,g2)


Analysis of Variance Table

Model 1: weightgain ~ initial.wt
Model 2: weightgain ~ 0 + x1 + x2 + I(x3 + x4)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     38 112.852
2     37  96.857  1    15.995 6.1104 0.01816
```

En consecuencia, rechazamos la hipótesis $H_0 : \alpha_1 = \alpha_2$ y los tratamientos dan una diferencia significativa en la ganancia de peso a lo largo de todos los pesos iniciales por igual.

Otra forma de resolver los mismos contrastes es mediante el análisis de la covarianza o ANCOVA. La idea es considerar un modelo de regresión con un indicador del tratamiento y la interacción entre la variable regresora cuantitativa y ese indicador. En realidad, se trata de una reparametrización del modelo general que hemos visto con $\alpha_1 = \mu$, $\alpha_2 = \mu + \delta$, $\beta_1 = \beta$ y $\beta_2 = \beta + \gamma$. Así, el test de paralelismo es $H_0 : \gamma = 0$. La matriz de diseño para el ANCOVA es

Lectura complementaria

En el documento de Carmona (2018) se trata el ANCOVA con más detalle. También se expone el caso multinivel o de varias rectas.

$$\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 1 & x_{21} & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n_2} & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} \mu \\ \beta \\ \delta \\ \gamma \end{pmatrix}$$

Las dos primeras columnas de $\widetilde{\mathbf{X}}$ son la regresión conjunta, la tercera es el indicador y la cuarta la interacción o producto entre la segunda y la tercera. En **R** este planteamiento es muy sencillo.

```
g <- lm(weightgain ~ initial.wt * treatment, data=goats)
anova(g)


Analysis of Variance Table

Response: weightgain
                     Df Sum Sq Mean Sq F value  Pr(>F)
initial.wt            1 59.548  59.548 22.2115 3.6e-05
treatment             1 15.995  15.995  5.9663 0.01962
initial.wt:treatment  1  0.342   0.342  0.1277 0.72296
Residuals            36 96.514   2.681
```

La ventaja es que los contrastes de paralelismo y de coincidencia son consecutivos de abajo hacia arriba y en la misma tabla. Salvo una pequeña diferencia en los grados de libertad, los resultados coinciden con los tests anteriores.

Las estimaciones de los parámetros β y $\widetilde{\beta}$ y otros detalles se pueden ver en el documento de Carmona (2018).

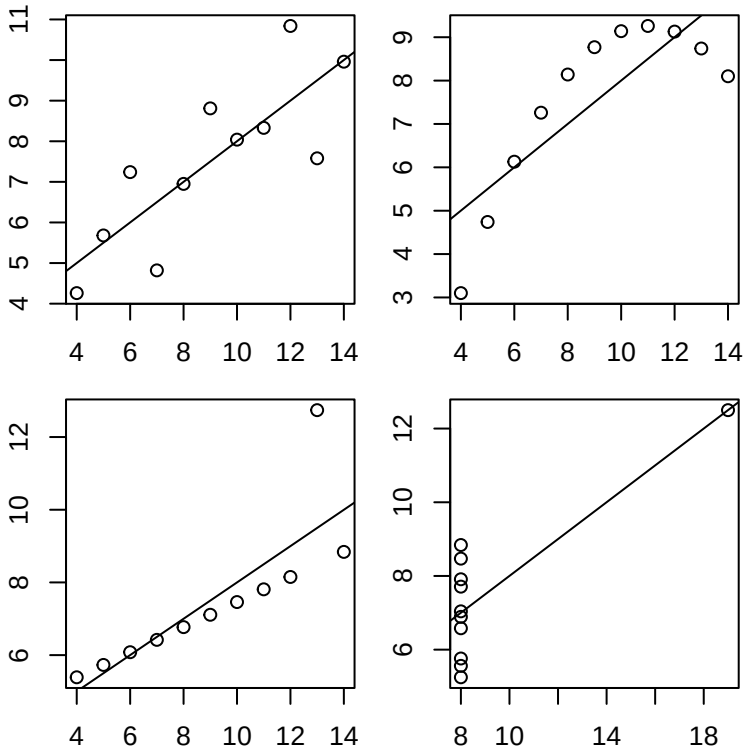
8. Un ejemplo para la reflexión

Con los datos de Anscombe (1973) se pueden calcular 4 regresiones completamente distintas que, sin embargo, tienen prácticamente la misma recta de regresión.

¡Atención!

Para evitar la confusión entre vectores ya utilizados con el mismo nombre, es conveniente ejecutar la instrucción `rm(list = ls())` o iniciar una nueva sesión de **R**.

```
data(anscombe)
attach(anscombe)
g1 <- lm(y1 ~ x1)
g2 <- lm(y2 ~ x2)
g3 <- lm(y3 ~ x3)
g4 <- lm(y4 ~ x4)
par(mfrow=c(2,2), mar=c(2,2,1,1))
plot(x1,y1); abline(g1)
plot(x2,y2); abline(g2)
plot(x3,y3); abline(g3)
plot(x4,y4); abline(g4)
```



```
cbind(coef(g1), coef(g2), coef(g3), coef(g4))
```

```
                [,1]     [,2]      [,3]      [,4]
(Intercept) 3.0000909 3.000909 3.0024545 3.0017273
x1          0.5000909 0.500000 0.4997273 0.4999091
```

```
detach(anscombe)
```

El mensaje que se deduce de este ejemplo es que no podemos fiarnos del procedimiento de cálculo de la regresión sin un análisis más profundo de la validación del modelo. Esto implica un análisis de los residuos y otras técnicas específicas para comprobar que se verifican las condiciones ideales del modelo lineal.

Lectura complementaria

En el apartado 6.8. de Carmona (2005) se puede ver un ejemplo parecido con más detalles.

9. Medidas de ajuste en regresión múltiple

Como en la regresión simple, la evaluación del ajuste del hiperplano de regresión a los datos se puede hacer con la *varianza residual* o estimación MC de σ^2:

$$\hat{\sigma}^2 = \text{SCR}/(n-m) = \text{ECM}$$

El estimador $\hat{\sigma}^2$ recibe el nombre de *error cuadrático medio*. (ECM)

Su raíz cuadrada $\hat{\sigma}$, que tiene las mismas unidades que Y, es el *error estándar de la regresión múltiple*. Como sabemos, la varianza residual y el error estándar dependen de las unidades de la variable respuesta y no son útiles para comparar diversas regresiones.

También en regresión múltiple se verifica la misma descomposición en sumas de cuadrados que hemos visto para la regresión simple. Así pues, el coeficiente de determinación se define de la misma forma y tiene la misma relación con el coeficiente de correlación múltiple al cuadrado.

$$R^2 = 1 - \frac{\text{SCR}}{S_y} = 1 - \frac{(1-r^2_{y\mathbf{x}})S_y}{S_y} = r^2_{y\mathbf{x}}$$

Donde $r_{y\mathbf{x}} = \text{cor}(Y, \hat{Y})$.

Como R^2 es la proporción de variabilidad explicada por las variables regresoras, resulta que si $R^2 \approx 1$, entonces la mayor parte de la variabilidad es explicada por dichas variables. Pero R^2 es la proporción de la variabilidad total explicada por el modelo con todas las variables frente al modelo $y = \beta_0 + \epsilon$, de manera que un R^2 alto muestra que el modelo mejora el modelo nulo y, por tanto, solo tiene sentido comparar coeficientes de determinación entre modelos anidados (casos particulares).

Además, un valor grande de R^2 no necesariamente implica que el modelo lineal es bueno. El coeficiente R^2 no mide si el modelo lineal es apropiado. Es posible que un modelo con un valor alto de R^2 proporcione estimaciones y predicciones pobres, poco precisas. El análisis de los residuos es imprescindible.

Un valor "grande"

Tampoco está claro lo que significa un valor "grande", ya que problemas en diversas ciencias (física, ingeniería, sociología...) tienen razonablemente diferentes criterios.

Por otra parte, cuando se añaden variables regresoras R^2 crece, pero eso no significa que el nuevo modelo sea superior:

$$R^2_{\text{nuevo}} = 1 - \frac{\text{SCR}_{\text{nuevo}}}{S_y} \geq R^2 = 1 - \frac{\text{SCR}}{S_y} \qquad \Rightarrow \qquad \text{SCR}_{\text{nuevo}} \leq \text{SCR}$$

pero es posible que

$$\text{ECM}_{\text{nuevo}} = \frac{\text{SCR}_{\text{nuevo}}}{n-(m+p)} \geq \text{ECM} = \frac{\text{SCR}}{n-m}$$

Luego, en esta situación, el nuevo modelo será peor. Así, como R^2 crece al añadir nuevas variables regresoras, se corre el peligro de sobreajustar el modelo añadiendo términos innecesarios. El coeficiente de determinación ajustado penaliza esto.

Para corregir el peligro de sobreajuste se define el *coeficiente de determinación ajustado* como

$$\bar{R}^2 = 1 - \frac{\text{SCR}/(n-m)}{S_y/(n-1)} = 1 - \frac{n-1}{n-m}(1-R^2)$$

Cuando $\bar{R}^2$ y R^2 son muy distintos, el modelo ha sido sobreajustado y debemos eliminar variables o términos.

Con los datos de las islas Galápagos tenemos los siguientes resultados.

```
ss <- summary(lmod)
ss$sigma


[1] 60.97519


ss$r.squared


[1] 0.7658469


ss$adj.r.squared


[1] 0.7170651
```

El ajuste no es excelente, pero parece razonablemente bueno para el tipo de datos. Por otra parte, la diferencia entre $\bar{R}^2$ y R^2 no es muy grande.

10. Inferencia sobre los coeficientes de regresión

Cuando asumimos la hipótesis de normalidad sobre la distribución de los errores $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, se deduce la normalidad de la variable respuesta $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, lo que nos permite utilizar las distribuciones asociadas a los estimadores de los parámetros y los contrastes de hipótesis que hemos estudiado en los módulos anteriores. Además, se supone que todas las funciones paramétricas son estimables, ya que el rango de la matriz de diseño es máximo.

La hipótesis de mayor interés, que es imprescindible rechazar, es la afirmación de que Y es independiente de las variables $x_1, \ldots, x_k$, es decir $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$. Se trata del llamado **contraste de significación de la regresión**, que se resuelve con el test F que hemos visto en el módulo anterior. Esta hipótesis equivale a afirmar que el coeficiente de correlación múltiple poblacional es cero y se resuelve con el mismo contraste y la tabla ANOVA asociada.

Un test F equivalente

El contraste de significación de la regresión se puede decidir con una expresión del estadístico F, basada en el coeficiente de correlación múltiple

$$F = \frac{r^2_{y\mathbf{x}}}{1 - r^2_{y\mathbf{x}}} \cdot \frac{n-p-1}{p}$$

El contraste de significación de un coeficiente de regresión particular $H_0 : \beta_j = 0$, para un j fijo, se resuelve con el estadístico t y la región crítica asociada. Aceptar esta hipótesis significa que la variable regresora x_j se puede eliminar del modelo. Sin embargo, es preciso actuar con cuidado, ya que se trata de un contraste *parcial* porque el coeficiente $\hat{\beta}_j$ depende de todas las otras variables regresoras x_i $(i \neq j)$. Es un contraste de la contribución de x_j dada la presencia de las otras variables regresoras en el modelo. De modo que la eliminación de variables del modelo se debe realizar con algún algoritmo o criterio, como veremos más adelante.

En cuanto a los intervalos de confianza para la respuesta media o los intervalos de predicción para una respuesta concreta, su deducción es similar al caso de la regresión simple.

Si $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0p})'$ recoge una observación particular del conjunto de variables regresoras, el intervalo de confianza con nivel $100(1-\alpha)\,\%$ para la respuesta media $E[Y|\mathbf{x}_0]$ está centrado en su estimación $\hat{y}_0 = \mathbf{x}_0'\widehat{\boldsymbol{\beta}}$

$$\hat{y}_0 \pm t_{n-m}(\alpha) \cdot (\mathrm{ECM}\, \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}$$

ya que $E(\hat{y}_0) = \mathbf{x}_0'\boldsymbol{\beta} = E[Y|\mathbf{x}_0]$ y $\mathrm{var}(\hat{y}_0) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.

11. Extrapolación oculta

En la estimación de la respuesta media o la predicción de nuevas respuestas en un punto $(x_{01}, \ldots, x_{0p})$ debemos tener mucho cuidado con la extrapolación. Si únicamente tenemos en cuenta el producto cartesiano de los recorridos de las variables regresoras, es fácil considerar la predicción para un punto que puede estar fuera de la nube de puntos con la que hemos calculado la regresión. Para evitar este problema deberemos ceñirnos al menor conjunto convexo que contiene los n puntos originales y que recibe el nombre de casco (*hull*) de las variables regresoras (ver figura 3).

La función `chull()`

La función `chull()` de **R** calcula el casco convexo de un conjunto de puntos de dos dimensiones.

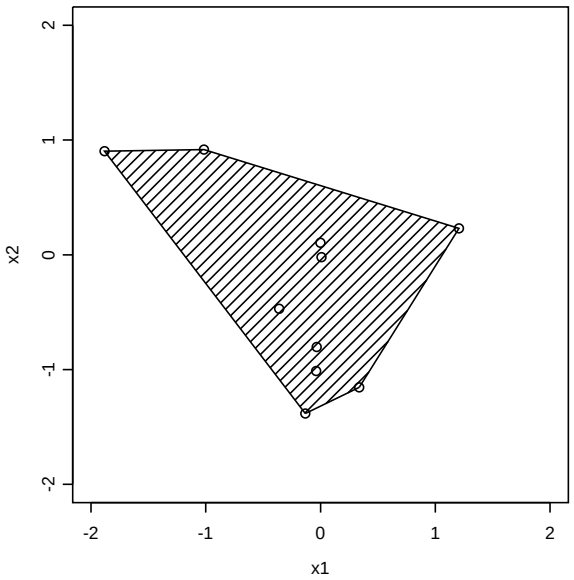


Figura 3: Conjunto convexo para los puntos de dos variables regresoras

Si consideramos los elementos h_{ii} de la diagonal de la matriz proyección $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, podemos definir $h_{\text{máx}} = \text{máx}\{h_{11}, \ldots, h_{nn}\}$ y se puede comprobar que

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \leq h_{\text{máx}}$$

es un elipsoide que contiene al casco. No es el menor elipsoide, pero es el más fácil de calcular.

Así pues, para evitar en lo posible la extrapolación, podemos comprobar en el punto $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})'$ si

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 < h_{\text{máx}}$$

12. Cambios de escala

En muchos casos es necesario un cambio de escala de las variables. Esto puede pasar cuando nos interesa, por ejemplo, cambiar las unidades de medida. También puede ocurrir que los valores observados sean excesivamente grandes o muy muy pequeños y sea imprescindible utilizar menos dígitos o menos decimales. En casos extremos, un cambio de escala permite evitar problemas de estabilidad numérica.

Si cambiamos x_i por $(x_i + a)/b$, entonces los contrastes t y F y los valores de $\hat{\sigma}^2$ y R^2 quedan igual, pero los coeficientes son $b\hat{\beta}_i$. Si se reescala la respuesta Y del mismo modo, entonces los contrastes t y F y el coeficiente R^2 no cambian, pero $\hat{\sigma}^2$ y $\hat{\beta}_i$ se reescalan por b.

Veamos qué ocurre con la variable `Area` dividida por 100 en el ejemplo con los datos de las islas Galápagos.

```
lmod2 <- lm(Species ~ I(Area/100) + Elevation + Nearest + Scruz +
                      Adjacent, data = gala)
rbind(summary(lmod)$coef[2,], summary(lmod2)$coef[2,])
```

```
        Estimate Std. Error   t value Pr(>|t|)
[1,] -0.02393834 0.02242235 -1.067611 0.296318
[2,] -2.39383383 2.24223507 -1.067611 0.296318
```

```
c(summary(lmod)$sigma, summary(lmod2)$sigma)
```

```
[1] 60.97519 60.97519
```

```
c(summary(lmod)$r.squared, summary(lmod2)$r.squared)
```

```
[1] 0.7658469 0.7658469
```

Sin embargo, si la modificación la hacemos en la variable respuesta, tenemos:

```
lmod3 <- lm(Species/100 ~ Area + Elevation + Nearest + Scruz +
                          Adjacent, data = gala)
```

```
rbind(summary(lmod)$coef[2,], summary(lmod3)$coef[2,])
```

```
          Estimate   Std. Error   t value Pr(>|t|)
[1,] -0.0239383383 0.0224223507 -1.067611 0.296318
[2,] -0.0002393834 0.0002242235 -1.067611 0.296318
```

```
c(summary(lmod)$sigma, summary(lmod3)$sigma)
```

```
[1] 60.9751884  0.6097519
```

```
c(summary(lmod)$r.squared, summary(lmod3)$r.squared)
```

```
[1] 0.7658469 0.7658469
```

Por otra parte, la magnitud de los coeficientes de regresión $\hat{\beta}_j$ refleja las unidades de medida de la variable regresora. En concreto, las unidades de los coeficientes de regresión son

$$\text{unidades } \hat{\beta}_j = \frac{\text{unidades } Y}{\text{unidades } x_j}$$

Si queremos comparar los coeficientes de diferentes variables, es necesario hacerlo con los **coeficientes de regresión estandarizados**, un caso especial de cambio de escala. La estandarización más habitual es

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\hat{s}_j} \qquad i = 1, \ldots, n;\ j = 1, \ldots, p$$
$$y_i^* = \frac{y_i - \bar{y}}{\hat{s}_y} \qquad i = 1, \ldots, n$$

Lectura complementaria

En el apartado 8.4. de Carmona (2005) se puede ver otra estandarización y más detalles de este punto.

donde

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad \hat{s}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \qquad \hat{s}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

El modelo es $y_i^* = b_0 + b_1 z_{i1} + b_2 z_{i2} + \cdots + b_p z_{ip} + \eta_i, \quad i = 1, \ldots, n$, donde las variables regresoras y la variable respuesta tienen media cero y varianza muestral uno. La estimación del modelo es $\widehat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_p)' = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}^*$ y $\hat{b}_0 = \bar{y}^* = 0$.

```
gala.scaled <- data.frame(scale(gala))
lmod.scaled <- lm(Species ~ Area + Elevation + Nearest + Scruz +
                    Adjacent, data = gala.scaled)
summary(lmod.scaled)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala.scaled)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97423 -0.30443 -0.06859  0.29188  1.59277

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.187e-16  9.711e-02   0.000 1.000000
Area        -1.804e-01  1.690e-01  -1.068 0.296318
Elevation    1.175e+00  1.974e-01   5.953 3.82e-06
Nearest      1.139e-03  1.313e-01   0.009 0.993151
Scruz       -1.427e-01  1.278e-01  -1.117 0.275208
Adjacent    -5.641e-01  1.335e-01  -4.226 0.000297

Residual standard error: 0.5319 on 24 degrees of freedom
Multiple R-squared:  0.7658,Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Ahora bien, tal como se demuestra en el apartado 8.4. de Carmona (2005), esta estimación es la que se obtiene con el sistema de ecuaciones

$$\begin{aligned} b_1 &+ r_{12}b_2 + \cdots + r_{1p}b_p = r_{1y} \\ r_{21}b_1 + &\ b_2 + \cdots + r_{2p}b_p = r_{2y} \\ \vdots &\quad \vdots \quad\quad\quad \vdots \quad \vdots \\ r_{k1}b_1 + r_{k2}b_2 &+ \cdots + \ b_p = r_{py} \end{aligned}$$

es decir, $\mathbf{R_{xx}b} = \mathbf{R}_{\mathbf{x}y}$, donde $\mathbf{R_{xx}}$ es la matriz de coeficientes de correlación entre las variables regresoras y $\mathbf{R}_{\mathbf{x}y} = (r_{1y}, \ldots, r_{py})'$ el vector columna con los coeficientes de correlación entre las variables regresoras y la respuesta.

Los coeficientes de regresión ordinarios se deducen de las ecuaciones

$$\hat{\beta}_j = \hat{b}_j \left(\frac{S_y}{S_j}\right)^{1/2} = \hat{b}_j \frac{s_y}{s_j} \qquad j = 1, \ldots, p$$

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j$$

Nota

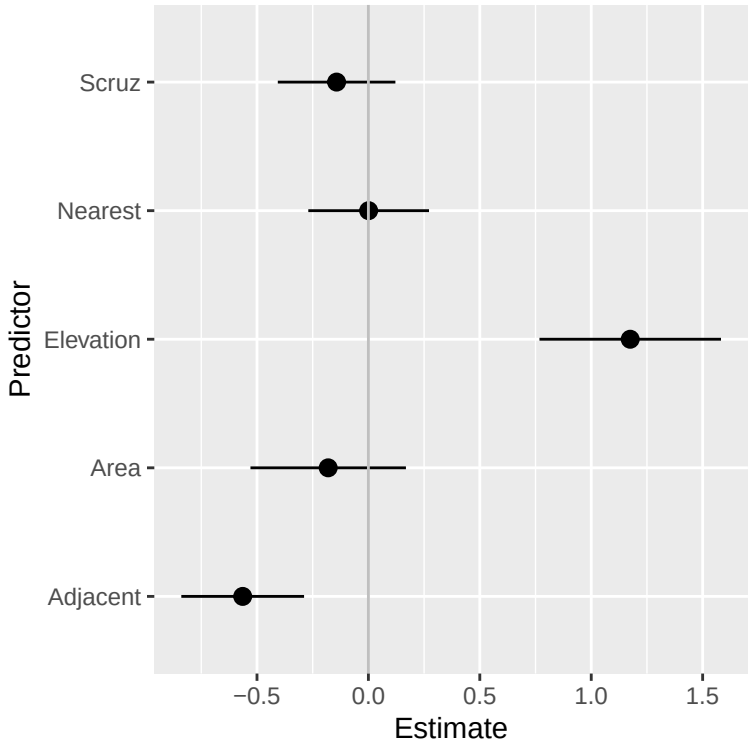
Algunos paquetes estadísticos calculan ambos conjuntos de coeficientes de regresión. En algún caso, a los coeficientes de regresión estandarizados les llaman *beta coeficientes*, lo que puede resultar confuso.

Además, el coeficiente de determinación es

$$R^2 = r^2_{y\mathbf{x}} = \hat{b}_1 r_{1y} + \hat{b}_2 r_{2y} + \cdots + \hat{b}_p r_{py}$$

Cuando las variables predictoras están en la misma escala, puede ser de ayuda dibujar el siguiente gráfico comparando los intervalos de confianza.

```
edf <- data.frame(coef(lmod.scaled),confint(lmod.scaled))[-1,]
names(edf) <- c("Estimate","lb","ub")
require(ggplot2)
p <- ggplot(aes(y=Estimate,ymin=lb,ymax=ub,x=row.names(edf)),
                data=edf) + geom_pointrange()
p + coord_flip() + xlab("Predictor") +
    geom_hline(yintercept=0, col=gray(0.75))
```



Finalmente señalaremos que debemos cuidar las interpretaciones, puesto que los coeficientes estandarizados todavía son parciales, es decir, miden el efecto de x_j dada la presencia de las otras variables regresoras. También $\hat{b}_j$ está afectado por el recorrido de los valores de las variables regresoras, de modo que es peligroso utilizar $\hat{b}_j$ para medir la importancia relativa de la variable regresora x_j.

Bibliografía

Anscombe, F.J. (1973). *Graphs in statistical analysis*. The American Statistician, 27(1). Disponible en: https://doi.org/10.2307/2682899.

Carmona, F. (2005) *Modelos lineales*. e-UMAB, Universitat de Barcelona.

Carmona, F. (2018) *Análisis de la Covarianza con **R***. Universitat de Barcelona. Disponible en: https://www.ub.edu/cursosR/files/ancova.pdf

Faraway, J.J. (2014) *Linear Models with R*. 2.ª ed. Chapman and Hall/CRC.

Saville, D.J. y Wood, G.R. (1991) *Statistical Methods: The Geometric Approach*. Springer.

Apéndice A: Datos de Alan Pearson

Con las siguientes instrucciones de **R** tenemos los datos estudiados por Alan Pearson.

```
treatment <- c(rep(1,20),rep(2,20))
weightgain <- c(5,3,8,7,6,4,8,6,7,5,5,2,4,5,10,3,8,6,6,3,
                9,8,8,8,11,8,7,6,5,7,6,9,3,7,5,7,7,5,3,8)
initial.wt <- c(21,24,21,22,23,26,22,23,24,20,
                27,28,28,30,18,27,19,20,19,22,
                18,18,19,19,19,21,20,21,23,23,
                25,25,26,24,24,25,26,27,30,29)
goats <- data.frame(treatment, weightgain, initial.wt)
goats$treatment <- factor(goats$treatment,
                          labels = c("standard","intensive"))
rm(treatment, weightgain, initial.wt)
attach(goats)
```