6

Regresión lineal simple

Sea Y una variable aleatoria y x una variable controlable, es decir, los valores que toma x son fijados por el experimentador. Supongamos que calculamos Y para diferentes valores de x de acuerdo con el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{6.1}$$

donde $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2 \quad i = 1, \ldots, n$.

Este modelo es la formulación lineal del problema de hallar la recta de regresión de Y sobre x. Los parámetros β_0, β_1 reciben el nombre de coeficientes de regresión. El parámetro β_0 es la ordenada en el origen, *intercept* en inglés, y β_1 es la pendiente de la recta, *slope* en inglés. La expresión matricial de 6.1 es

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

Ahora podemos aplicar toda la teoría general desarrollada en los capítulos anteriores para un modelo lineal cualquiera, al caso particular de la regresión lineal simple.

6.1. Estimación de los coeficientes de regresión

Con los datos observados se pueden calcular los siguientes estadísticos

$$\begin{aligned} \bar{x} &= (1/n) \textstyle\sum x_i \qquad & s_x^2 &= (1/n) \textstyle\sum (x_i - \bar{x})^2 \\ \bar{y} &= (1/n) \textstyle\sum y_i \qquad & s_y^2 &= (1/n) \textstyle\sum (y_i - \bar{y})^2 \end{aligned}$$

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

donde $\bar{x}, \bar{y}, s_x^2, s_y^2, s_{xy}$ son las medias, varianzas y covarianzas muestrales, aunque el significado de s_x^2 y s_{xy} es *convencional* pues x no es variable aleatoria. Con esta notación las ecuaciones normales son:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \qquad \Leftrightarrow \qquad \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}$$

y como

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n s_x^2} \begin{pmatrix} (1/n)\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

la solución es

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_x} = \frac{s_{xy}}{s_x^2} \end{aligned}$$

donde

$$S_{xy} = \sum x_i y_i - (1/n)\sum x_i \sum y_i = \sum (x_i - \overline{x})(y_i - \overline{y}) = n\, s_{xy}$$
$$S_x = \sum x_i^2 - (1/n)(\sum x_i)^2 = \sum (x_i - \overline{x})^2 = n\, s_x^2$$

En el ejercicio 6.2 se ven otras formas de expresar $\hat{\beta}_1$.

La recta de regresión es

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

que se expresa también en la forma

$$y - \overline{y} = \hat{\beta}_1 (x - \overline{x})$$

lo que deja claro que la recta pasa por el punto $(\overline{x}, \overline{y})$ y que el modelo es válido en el rango de las x_i, centrado en $\overline{x}$. Ésta es también la recta que se obtiene a partir del modelo equivalente con los datos x_i centrados (ver ejemplo 5.6.2 y ejercicio 6.3).

Recordemos que por lo que hemos estudiado, estas estimaciones son insesgadas y de varianza mínima entre todos los estimadores lineales (teorema de Gauss-Markov). Las varianzas y covarianza de los estimadores son

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{var}(\hat{\beta}_1) \end{pmatrix} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \tag{6.2}$$

Es decir

$$E(\hat{\beta}_0) = \beta_0 \qquad \text{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_x} \right) \tag{6.3}$$

$$E(\hat{\beta}_1) = \beta_1 \qquad \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_x} \tag{6.4}$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\overline{x}}{S_x} \tag{6.5}$$

Ejemplo 6.1.1

Vamos a ilustrar el cálculo "manual" de las estimaciones de los parámetros con un ejemplo muy sencillo de muy pocos datos.

Supongamos que una empresa de compra-venta de automóviles organiza exposiciones los fines de semana i contrata un número variable de vendedores que oscila entre 3 y 8. El gerente de esta empresa quiere estudiar la relación entre el número de vendedores y el número de coches vendidos ya que, si es posible, podría prever las ventas a partir del número de vendedores que contrata. Para aclararlo, el gerente examina el registro de ventas de los últimos cuatro meses y localiza un período de 10 semanas durante las cuales no hubo ningún incentivo especial ni a la venta ni a la compra. El número de coches vendidos durante este período y el número de vendedores empleados en cada caso se muestra en la tabla adjunta.

Para examinar esta relación es muy útil empezar por dibujar un diagrama de dispersión. Este gráfico muestra una relación bastante evidente entre el número de vendedores y las ventas, relación que se podía esperar. Vamos a cuantificarla con la ayuda de la recta de regresión MC.

En la siguiente tabla tenemos los cálculos necesarios para obtener los coeficientes de regresión, las predicciones, los residuos y la suma de cuadrados de los errores para los datos de las 10 semanas. Esta tabla se ha calculado con una hoja de cálculo, lo que permite una mayor precisión en los cálculos sucesivos.

Con estos cálculos, las estimaciones de los coeficientes de regresión son

$$\hat{\beta}_1 = \frac{855 - \frac{1}{10}\, 53 \cdot 150}{301 - \frac{1}{10}\,(53)^2} = 2.9850746$$
$$\hat{\beta}_0 = 15 - \hat{\beta}_1 \cdot 5.3 = -0.820896$$

Semana	Vendedores	Coches
1	5	10
2	6	20
3	5	18
4	4	10
5	3	7
6	4	14
7	7	21
8	6	15
9	5	13
10	8	22

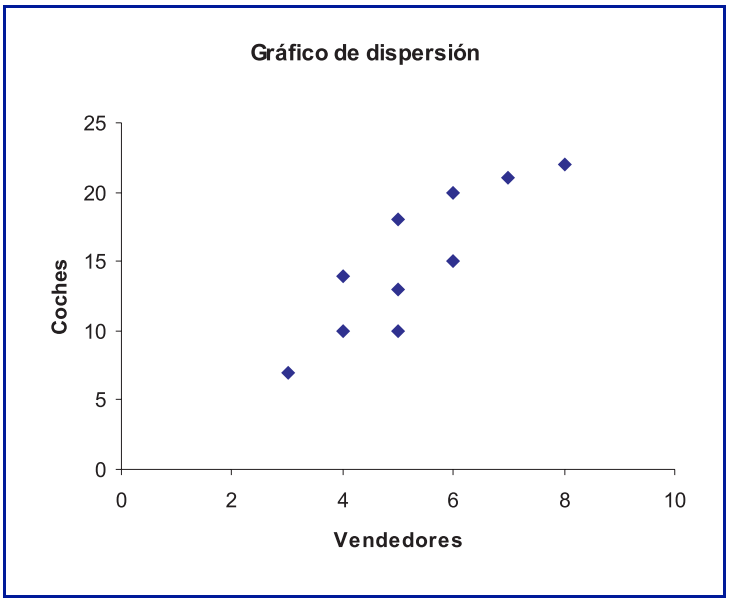

Tabla 6.1: Datos de las ventas en 10 semanas y gráfico de dispersión

i	x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i$	e_i	e_i^2
1	5	10	25	50	14.10	-4.10	16.85
2	6	20	36	120	17.09	2.91	8.47
3	5	18	25	90	14.10	3.90	15.18
4	4	10	16	40	11.12	-1.12	1.25
5	3	7	9	21	8.13	-1.13	1.29
6	4	14	16	56	11.12	2.88	8.30
7	7	21	49	147	20.07	0.93	0.86
8	6	15	36	90	17.09	-2.09	4.37
9	5	13	25	65	14.10	-1.10	1.22
10	8	22	64	176	23.06	-1.06	1.12
Suma	53	150	301	855		0	58.90
Media	5.3	15					

Tabla 6.2: Cálculos de regresión simple para los datos de ventas

La ecuación de la recta de regresión es

$$y = -0.821 + 2.985x$$

o también

$$y - 15 = 2.985\,(x - 5.3)$$

Para calcular la precisión de estas estimaciones, primero debemos estimar la varianza del modelo.

Nota: Una aplicación de hojas de cálculo como *Microsoft Excel* tiene la función ESTIMACION.LINEAL que calcula de forma directa los coeficientes de regresión y algunos estadísticos más. Otra función matricial es TENDENCIA que permite calcular directamente las predicciones. Además, *Excel* lleva un conjunto de macros opcionales llamadas "Herramientas para análisis" que, entre otras cosas, calculan una regresión lineal completa.

En el ejemplo anterior, se comprueba que las suma de los residuos es cero, salvo problemas de redondeo. Esto no es una casualidad. Vamos a ver algunas propiedades adicionales para las predicciones $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y para los residuos $e_i = y_i - \hat{y}_i$, cuya demostración se deja para el lector (ver ejercicio 6.4).

(i) La suma de los residuos es cero: $\sum e_i = 0$.

(ii) La suma de los residuos ponderada por los valores de la variable regresora es cero: $\sum x_i e_i = 0$.

(iii) $\sum y_i = \sum \hat{y}_i$

(iv) La suma de los residuos ponderada por las predicciones de los valores observados es cero: $\sum \hat{y}_i e_i = 0$.

6.2. Medidas de ajuste

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

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

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. Otra medida de ajuste requiere una adecuada descomposición de la variabilidad de la variable respuesta.

Teorema 6.2.1

Consideremos el coeficiente de correlación muestral, cuyo significado es convencional,

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

Entonces se verifican las siguientes relaciones

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

(ii) $\text{SCR} = \sum (y_i - \hat{y}_i)^2 = (1 - r^2)\sum (y_i - \overline{y})^2 = (1 - r^2)S_y$

(iii) $\hat{\sigma}^2 = (\sum e_i^2)/(n-2) = (1 - r^2)S_y/(n-2)$

Demostración:

$$\begin{aligned}\sum (y_i - \overline{y})^2 &= \sum (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 \\ &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})\end{aligned}$$

pero $\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum (y_i - \hat{y}_i)\hat{y}_i - \overline{y}\sum (y_i - \hat{y}_i) = 0$ por las propiedades del apartado anterior. También podemos recordar la ortogonalidad de los subespacios de los errores y de las estimaciones. Queda así demostrada la relación (i).

Por otra parte, es fácil ver que

$$\sum (\hat{y}_i - \overline{y})^2 = \hat{\beta}_1^2 \sum (x_i - \overline{x})^2 = r^2 \sum (y_i - \overline{y})^2$$

de forma que finalmente

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

Luego

$$\sum (y_i - \hat{y}_i)^2 = (1 - r^2)\sum (y_i - \overline{y})^2$$

Como consecuencia tenemos que el estimador centrado de la varianza σ^2 del modelo 6.1 es

$$\hat{\sigma}^2 = \text{SCR}/(n-2) = (1-r^2)S_y/(n-2) \tag{6.6}$$

■

La 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, y el segundo término que 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

$$\text{VT} = \text{VNE} + \text{VE} \tag{6.7}$$

Definición 6.2.1

Una medida del ajuste de la recta de regresión a los datos es la proporción de variabilidad explicada que definimos con el nombre de coeficiente de determinación *así:*

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

Esta medida se puede utilizar en cualquier tipo de regresión, pero en el caso particular de la regresión lineal simple con una recta tenemos

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

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

El coeficiente de determinación R^2 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 las dos variables, cuando son aleatorias y sólo hay una variable regresora.

Ejemplo 6.2.1

Continuando con el ejemplo de los datos de ventas tenemos:

$$\text{SCR} = 58.896$$
$$\hat{\sigma}^2 = 58.896/8 = 7.362 \qquad \hat{\sigma} = 2.713$$
$$\text{VT} = S_y = 238$$
$$R^2 = 1 - \frac{58.896}{238} = 0.7525$$

6.3. Inferencia sobre los parámetros de regresión

Supongamos que el modelo 6.1 es un modelo lineal normal. Entonces (ver teorema 2.6.1) se verifica que

$$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)' \sim N_2(\boldsymbol{\beta}, \text{var}(\widehat{\boldsymbol{\beta}}))$$

donde

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} 1/n + \bar{x}/S_x & -\bar{x}/S_x \\ -\bar{x}/S_x & 1/S_x \end{pmatrix}$$

como hemos visto en 6.2–6.5. Además sabemos que $\widehat{\boldsymbol{\beta}}$ es independiente de SCR.

Como consecuencia de estas distribuciones hemos demostrado (ver 3.4 o 5.10) que para contrastar una hipótesis del tipo $H_0 : \mathbf{a}'\boldsymbol{\beta} = c$ se utiliza el estadístico

$$t = \frac{\mathbf{a}'\widehat{\boldsymbol{\beta}} - c}{(\hat{\sigma}^2(\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}))^{1/2}} \tag{6.8}$$

que seguirá una distribución t_{n-2}, si H_0 es cierta.

6.3.1. Hipótesis sobre la pendiente

El contraste de la hipótesis $H_0 : \beta_1 = b_1$ frente a $H_1 : \beta_1 \neq b_1$ se resuelve rechazando H_0 si

$$\left| \frac{\hat{\beta}_1 - b_1}{(\hat{\sigma}^2/S_x)^{1/2}} \right| > t_{n-2}(\alpha)$$

donde $P[|t_{n-2}| > t_{n-2}(\alpha)] = \alpha$.

En particular, estamos interesados en contrastar si la pendiente es cero, es decir, la hipótesis $H_0 : \beta_1 = 0$. Vamos a deducir este contraste directamente.

Si $H_0 : \beta_1 = 0$ es cierta, el modelo 6.1 se simplifica y se convierte en

$$y_i = \beta_0 + \epsilon_i$$

de donde

$$\text{SCR}_H = \sum (y_i - \hat{\beta}_{0|H})^2 = \sum (y_i - \bar{y})^2 = S_y \tag{6.9}$$

dado que $\hat{\beta}_{0|H} = \bar{y}$.

Por el teorema 6.2.1 sabemos que $\text{SCR} = (1 - r^2)S_y$, de manera que

$$F = \frac{\text{SCR}_H - \text{SCR}}{\text{SCR}/(n-2)} = \frac{S_y - (1-r^2)S_y}{(1-r^2)S_y/(n-2)} = (n-2)\frac{r^2}{1-r^2} \sim F_{1,n-2}$$

Finalmente,

$$t = \sqrt{F} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \tag{6.10}$$

sigue la distribución t de Student con $n-2$ grados de libertad.

Este contraste $H_0 : \beta_1 = 0$ se llama *contraste para la significación de la regresión* y se formaliza en una tabla de análisis de la varianza donde se explicita la descomposición de la suma de cuadrados 6.7.

El hecho de aceptar $H_0 : \beta_1 = 0$ puede implicar que la mejor predicción para todas las observaciones es $\bar{y}$, ya que la variable x no influye, y la regresión es inútil. Pero también podría pasar que la relación no fuera de tipo recta.

Rechazar la hipótesis $H_0 : \beta_1 = 0$ puede implicar que el modelo lineal 6.1 es adecuado. Pero también podría ocurrir que no lo sea. En todo caso, es muy importante no confundir la significación de

Fuente de variación	grados de libertad	suma de cuadrados	cuadrados medios	F
Regresión	1	$\hat{\beta}_1 S_{xy}$	CM_R	CM_R/ECM
Error	$n-2$	SCR	ECM	
Total	$n-1$	S_y		

Tabla 6.3: Análisis de la varianza para contrastar la significación de la regresión

la regresión con una prueba de causalidad. Los modelos de regresión únicamente cuantifican la relación lineal entre la variable respuesta y las variables explicativas, una en el caso simple, pero no justifican que éstas sean la causa de aquella.

Tanto la adecuación del modelo 6.1, como la hipótesis de normalidad han de estudiarse a través del análisis de los residuos.

6.3.2. Hipótesis sobre el punto de intercepción

Para el contraste de hipótesis $H_0 : \beta_0 = b_0$, se utiliza el estadístico

$$t = \frac{\hat{\beta}_0 - b_0}{(\hat{\sigma}^2(1/n + \overline{x}^2/S_x))^{1/2}}$$

que, si la hipótesis es cierta, sigue una distribución t de Student con $n-2$ grados de libertad.

6.3.3. Intervalos de confianza para los parámetros

Además de los estimadores puntuales de β_0, β_1 y σ^2, con las distribuciones estudiadas podemos proporcionar intervalos de confianza para estos parámetros. El ancho de estos intervalos estará en función de la calidad de la recta de regresión.

Con la hipótesis de normalidad y teniendo en cuenta las distribuciones de $\hat{\beta}_0$ y $\hat{\beta}_1$ estudiadas, un intervalo de confianza para la pendiente β_1 con nivel de confianza $100(1-\alpha)\,\%$ es

$$\hat{\beta}_1 \pm t_{n-2}(\alpha) \cdot (\hat{\sigma}^2/S_x)^{1/2}$$

donde $t_{n-2}(\alpha)$ es tal que $P[|t_{n-2}| < t_{n-2}(\alpha)] = 1-\alpha$.

Análogamente, para β_0 es

$$\hat{\beta}_0 \pm t_{n-2}(\alpha) \cdot (\hat{\sigma}^2(1/n + \overline{x}^2/S_x))^{1/2}$$

Las cantidades

$$ee(\hat{\beta}_1) = (\hat{\sigma}^2/S_x)^{1/2} \qquad ee(\hat{\beta}_0) = (\hat{\sigma}^2(1/n + \overline{x}^2/S_x))^{1/2}$$

son los *errores estándar* de la pendiente $\hat{\beta}_1$ y la intercepción $\hat{\beta}_0$, respectivamente. Se trata de estimaciones de la desviación típica de los estimadores. Son medidas de la precisión de la estimación de los parámetros.

Como sabemos

$$\hat{\sigma}^2 = \frac{\text{SCR}}{n-2} = \frac{1}{n-2} S_y(1-r^2)$$

es el estimador insesgado de σ^2 y la distribución de SCR/σ^2 es $\sim \chi^2_{n-2}$. Así, el intervalo de confianza al $100(1-\alpha)\,\%$ de σ^2 es

$$\frac{\text{SCR}}{\chi^2_{n-2}(\alpha/2)} \leq \sigma^2 \leq \frac{\text{SCR}}{\chi^2_{n-2}(1-\alpha/2)}$$

donde $\chi^2_{n-2}(\alpha/2)$ y $\chi^2_{n-2}(1-\alpha/2)$ son los valores de una χ^2_{n-2} para que la suma de las probabilidades de las colas sea α.

6.3.4. Intervalo para la respuesta media

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

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

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.

6.3.5. 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, 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

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

Este resultado se puede generalizar al caso de un intervalo de predicción al $100(1-\alpha)\,\%$ para la media de k futuras observaciones de la variable respuesta cuando $x = x_0$. Si $\bar{y}_0$ es la media de k futuras observaciones para $x = x_0$, un estimador de $\bar{y}_0$ es $\hat{y}_0$ de forma que el intervalo es

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

6.3.6. Región de confianza y intervalos de confianza simultáneos

Habitualmente, los intervalos de confianza se dan de forma conjunta para los dos parámetros β_0, β_1 de la regresión simple. Sin embargo, la confianza conjunta de ambos intervalos no es $100(1-\alpha)\,\%$, aunque los dos se hayan construido para verificar ese nivel de confianza. Si deseamos que el nivel de confianza conjunta sea el $100(1-\alpha)\,\%$ debemos construir una región de confianza o, alternativamente, los llamados intervalos de confianza simultáneos.

A partir de la distribución de la ecuación 5.9 sabemos que, en general,

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}}-\mathbf{A}\boldsymbol{\beta})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}}-\mathbf{A}\boldsymbol{\beta})/q}{\mathrm{SCR}/(n-r)} \sim F_{q,n-r}$$

donde, en este caso, $\mathbf{A}\widehat{\boldsymbol{\beta}} = \mathbf{I}\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ y $q = 2$. Así pues

$$\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{2\mathrm{ECM}} \sim F_{2,n-2}$$

y

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}$$

Con esta distribución se puede construir una región de confianza al $100(1-\alpha)\,\%$ para β_0, β_1 conjuntamente que viene dada por la elipse

$$\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{2\mathrm{ECM}} \leq F_{2,n-2}(\alpha)$$

Con el mismo objetivo, se pueden utilizar diversos métodos de obtención de intervalos simultáneos del tipo

$$\hat{\beta}_j \pm \Delta\cdot \mathrm{ee}(\hat{\beta}_j) \qquad j = 0, 1$$

Por ejemplo, el método de Scheffé proporciona los intervalos simultáneos

$$\hat{\beta}_j \pm (2F_{2,n-2}(\alpha))^{1/2}\cdot \mathrm{ee}(\hat{\beta}_j) \qquad j = 0, 1$$

6.4. Regresión pasando por el origen

Supongamos que, por alguna razón justificada, el experimentador decide proponer el modelo de regresión simple

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

que carece del término β_0.

El estimador MC del parámetro β_1 es

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

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

$$\hat{\sigma}^2 = \text{SCR}/(n-1) = \frac{1}{n-1}\left(\sum y_i^2 - \hat{\beta}_1 \sum x_i y_i\right) \tag{6.11}$$

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 estar muy seguros 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 estamos seguros, 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 6.11 que se puede comparar con el del modelo completo 6.6. El coeficiente 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

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

que no es comparable con el R^2 de la definición 6.2.1. De hecho puede ocurrir que $R_0^2 > R^2$, aunque $\text{ECM}_0 < \text{ECM}$.

6.5. Correlación

Consideremos la situación en la que las dos variables son aleatorias, tanto la la variable respuesta como la variable explicativa o regresora. De modo que tomamos una muestra aleatoria simple de tamaño n formada por las parejas $(x_1, y_1), \ldots, (x_n, y_n)$ de dos variables aleatorias (X, Y) con distribución conjunta normal bivariante

$$(X, Y)' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\mu} = (\mu_1, \mu_2)' \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

donde $\text{cov}(X, Y) = \sigma_1\sigma_2\rho$ y ρ es el coeficiente de correlación entre Y y X.

La distribución condicionada de Y dado un valor de $X = x$ es

$$Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2_{2\cdot1})$$

donde

$$\beta_0 = \mu_1 - \mu_2 \frac{\sigma_2}{\sigma_1}\rho$$
$$\beta_1 = \frac{\sigma_2}{\sigma_1}\rho$$
$$\sigma^2_{2\cdot1} = \sigma^2_2(1-\rho^2)$$

De modo que la esperanza de $Y|X = x$ es el modelo de regresión lineal simple

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

Además, hay una clara relación entre β_1 y ρ, $\rho = 0 \Leftrightarrow \beta_1 = 0$, en cuyo caso no hay regresión lineal, es decir, el conocimiento de $X = x$ no nos ayuda a predecir Y.

El método de la máxima verosimilitud proporciona estimadores de β_0 y β_1 que coinciden con los estimadores MC.

Ahora también es posible plantearse inferencias sobre el parámetro ρ. En primer lugar, el estimador natural de ρ es

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

y

$$\hat{\beta}_1 = \left(\frac{S_y}{S_x}\right)^{1/2} r$$

Así, $\hat{\beta}_1$ y r están relacionados, pero mientras r representa una medida de la asociación entre X e Y, $\hat{\beta}_1$ mide el grado de predicción en Y por unidad de X.

Nota: Ya hemos advertido de que cuando X es una variable controlada, r tiene un significado convencional, porque su magnitud depende de la elección del espaciado de los valores x_i. En este caso, ρ no existe y r no es un estimador.

También sabemos que $r^2 = R^2$, de modo que el coeficiente de determinación es el cuadrado de la correlación.

Finalmente, el principal contraste sobre ρ es el de incorrelación $H_0 : \rho = 0$ que es equivalente a $H_0 : \beta_1 = 0$ y se resuelve con el estadístico

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

que, si H_0 es cierta, sigue una distribución t_{n-2}.

6.6. Carácter lineal de la regresión simple

Supongamos ahora que estamos interesados en decidir si la regresión de Y sobre x es realmente lineal. 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, vamos a ver que podemos reconducir el contraste a la situación prevista en la sección 5.6.2 para la elección entre dos modelos lineales.

Necesitamos n_i valores de Y para cada x_i. Con un cambio de notación, para cada $i = 1, \ldots, k$, sean

$$x_i : y_{i1}, \ldots, y_{in_i} \qquad \overline{y}_i = (1/n_i) \sum_j y_{ij} \qquad s^2_{y_i} = (1/n_i) \sum_j (y_{ij} - \overline{y}_i)^2$$
$$\overline{y} = (1/n) \sum_{i,j} y_{ij} \qquad s^2_y = (1/n) \sum_{i,j} (y_{ij} - \overline{y})^2 \qquad n = n_1 + \cdots + n_k$$

Introducimos a continuación el coeficiente

$$\hat{\eta}^2 = 1 - \frac{1}{n} \sum_{i=1}^{k} n_i \frac{s^2_{y_i}}{s^2_y} \tag{6.12}$$

que verifica $0 \leq \hat{\eta}^2 \leq 1$, y mide el grado de concentración de los puntos (x_i, y_{ij}) a lo largo de la curva $y = g(x)$ (ver figura 6.1).

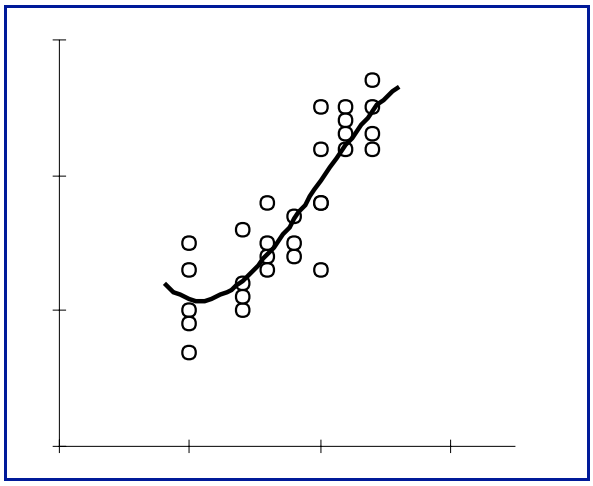

Figura 6.1: Curva que mejor se ajusta a los datos

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 $\hat{\delta}_i = \overline{y}_i$. La identidad

$$\mathrm{SCR}_H = \mathrm{SCR} + (\mathrm{SCR}_H - \mathrm{SCR})$$

es entonces

$$\sum_{i,j} (y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i,j} (y_{ij} - \overline{y}_i)^2 + \sum_i n_i (\overline{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Dividiendo por n tenemos

$$s^2_y (1 - r^2) = s^2_y (1 - \hat{\eta}^2) + s^2_y (\hat{\eta}^2 - r^2)$$

y el contraste para decidir si la regresión es lineal se resuelve a través del estadístico

$$F = \frac{(\hat{\eta}^2 - r^2)/(k-2)}{(1 - \hat{\eta}^2)/(n-k)} \tag{6.13}$$

que tiene $(k-2)$ y $(n-k)$ grados de libertad. Si F resulta significativa, rechazaremos el carácter lineal de la regresión.

Observaciones:

1) Solamente se puede aplicar este test si se tienen $n_i > 1$ observaciones de Y para cada x_i $(i = 1, \ldots, k)$.

2) $\hat{\eta}^2$ es una versión muestral de la llamada *razón de correlación* entre dos variables aleatorias X, Y

$$\eta^2 = \frac{E[(g(X) - E(Y))^2]}{\mathrm{var}(Y)}$$

siendo

$$y = g(x) = E(Y|X = x)$$

la curva de regresión de la media de Y sobre X. Este coeficiente η^2 verifica:

a) $0 \leq \eta^2 \leq 1$

b) $\eta^2 = 0 \Rightarrow y = E(Y)$ (la curva es la recta y = constante).

c) $\eta^2 = 1 \Rightarrow y = g(X)$ (Y es función de X)

3) Análogamente, podemos también plantear la hipótesis de que Y es alguna función (no lineal) de x frente a la hipótesis nula de que no hay ningún tipo de relación. Las hipótesis son:

$$H_0 : y_i = \mu + \epsilon_i$$
$$H_1 : y_i = g(x_i) + \epsilon_i$$

siendo μ constante. Entonces, con las mismas notaciones de antes,

$$\text{SCR}_H = \sum_{i,j} (y_{ij} - \bar{y})^2 \qquad \text{con } n-1 \text{ g.l.}$$
$$\text{SCR} = \sum_{i,j} (y_{ij} - \bar{y}_i)^2 \qquad \text{con } n-k \text{ g.l.}$$

Operando, se llega al estadístico

$$F = \frac{\hat{\eta}^2/(k-1)}{(1-\hat{\eta}^2)/(n-k)} \tag{6.14}$$

Comparando 6.14 con 6.10, podemos interpretar 6.14 como una prueba de significación de la razón de correlación.

Ejemplo 6.6.1

Se mide la luminosidad (en lúmenes) de un cierto tipo de lámparas después de un tiempo determinado de funcionamiento (en horas). Los resultados para una serie de 3, 2, 3, 2 y 2 lámparas fueron:

Tiempo (x)	Luminosidad (Y)			
250	5460	5475	5400	($n_1 = 3$)
500	4800	4700		($n_2 = 2$)
750	4580	4600	4520	($n_3 = 3$)
1000	4320	4300		($n_4 = 2$)
1250	4000	4010		($n_5 = 2$)

Con estos datos podemos ilustrar algunos aspectos de la regresión lineal de la luminosidad sobre el tiempo de funcionamiento.

- *Recta de regresión y coeficiente de correlación:*

$$\bar{x} = 708.33 \quad \bar{y} = 4680.42 \quad n=12$$
$$s_x = 351.09 \quad s_y = 500.08 \quad s_{xy} = -170190.97$$
$$r = -0.969 \qquad \hat{\beta}_1 = -1.381$$
$$y - 4680.42 = -1.381(x - 708.33)$$

La hipótesis $H_0 : \beta_1 = 0$ debe ser rechazada pues (ver 6.10) obtenemos $t = 12.403$ (10 g.l.) que es muy significativo.

- *Razón de correlación y carácter lineal de la regresión:*

$$\begin{array}{lllll} \bar{y}_1 = 5445 & \bar{y}_2 = 4750 & \bar{y}_3 = 4566.7 & \bar{y}_4 = 4310 & \bar{y}_5 = 4005 \\ s^2_{y_1} = 1050 & s^2_{y_2} = 2500 & s^2_{y_3} = 1155.5 & s^2_{y_4} = 100 & s^2_{y_5} = 25 \end{array}$$

$$\bar{y} = 4680.42 \qquad s^2_y = 250077 \qquad n = 12 \quad k = 5$$

$$\hat{\eta}^2 = 1 - \frac{1}{n}\sum_{i=1}^{k} n_i \frac{s^2_{y_i}}{s^2_y} = 0.996$$

Aplicando 6.13

$$F = \frac{(0.996 - 0.939)/3}{(1 - 0.996)/7} = 33.3$$

con 3 y 7 g.l. Se puede rechazar que la regresión es lineal.

Aplicando ahora 6.14

$$F = \frac{0.996/4}{(1 - 0.996)/7} = 435.7$$

vemos que la razón de correlación es muy significativa.

6.7. Comparación de rectas

En primer lugar, vamos a estudiar detalladamente la comparación de dos rectas, ya que en este caso las fórmulas son un poco más sencillas. A continuación presentaremos el caso general cuyos detalles pueden verse en Seber[66] pág. 197-205.

6.7.1. Dos rectas

Consideremos 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

$$y_{1i} = \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \qquad i = 1, \ldots, n_1$$
$$y_{2i} = \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \qquad i = 1, \ldots, n_2$$

y sus estimadores MC son

$$\hat{\alpha}_h = \bar{y}_h - \hat{\beta}_h \bar{x}_h \qquad \hat{\beta}_h = r_h \left(\frac{S_{yh}}{S_{xh}}\right)^{1/2} \qquad h = 1, 2$$

donde $\bar{x}_h, S_{xh}, \bar{y}_h, S_{yh}, r_h$ son las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación para cada una de las muestras $h = 1, 2$ respectivamente.

También deberemos considerar $\bar{x}, S_x, \bar{y}, S_y, r$ las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación de las dos muestras conjuntamente.

Vamos a considerar las dos regresiones simples como un único modelo lineal. Para ello hacemos

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

y

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

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

Así pues, el modelo que presenta a las dos regresiones simples conjuntamente $\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ es un modelo lineal siempre que los errores verifiquen las condiciones de Gauss-Markov. Entonces 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 $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2$ y la suma de cuadrados residual es

$$\begin{aligned} \text{SCR} &= \sum_{i=1}^{n_1} (y_{1i} - \hat{\alpha}_1 - \hat{\beta}_1 x_{1i})^2 + \sum_{i=1}^{n_2} (y_{2i} - \hat{\alpha}_2 - \hat{\beta}_2 x_{2i})^2 \\ &= \text{SCR}_1 + \text{SCR}_2 = S_{y1}(1 - r_1^2) + S_{y2}(1 - r_2^2) \\ &= S_{y1} - \hat{\beta}_1^2 S_{x1} + S_{y2} - \hat{\beta}_2^2 S_{x2} \end{aligned} \tag{6.15}$$

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

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

También se pueden utilizar los contrastes que se explican en la sección 6.7.3.

Test de coincidencia

Se trata de investigar si las dos rectas se pueden considerar iguales, es decir, vamos a contrastar la hipótesis

$$H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$$

Ésta es una hipótesis lineal contrastable (el modelo es de rango máximo) del tipo $H_0 : \mathbf{A}\boldsymbol{\gamma} = \mathbf{0}$ con

$$\mathbf{A}\boldsymbol{\gamma} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

donde $\mathbf{A}$ es 2×4 y $q = \text{rg}\,\mathbf{A} = 2$. Luego podríamos utilizar las fórmulas obtenidas para el contraste. Sin embargo, en este caso es mucho más fácil calcular directamente la suma de cuadrados bajo la hipótesis.

Bajo H_0 la estimación MC de los parámetros comunes $\alpha = \alpha_1 = \alpha_2$ y $\beta = \beta_1 = \beta_2$ es sencillamente la que se obtiene del modelo lineal conjunto, es decir, una única recta de regresión con todos los datos juntos:

$$\alpha^* = \bar{y} - \beta^* \bar{x}$$

$$\beta^* = r \left(\frac{S_y}{S_x} \right)^{1/2}$$

Luego

$$\begin{aligned}\text{SCR}_H &= \sum_{i=1}^{n_1}(y_{1i}-\alpha^*-\beta^* x_{1i})^2+\sum_{i=1}^{n_2}(y_{2i}-\alpha^*-\beta^* x_{2i})^2\\ &= S_y(1-r^2)\end{aligned}$$

De modo que el estadístico F es

$$F=\frac{(\text{SCR}_H-\text{SCR})/2}{\text{SCR}/(n_1+n_2-4)}=\frac{(S_y(1-r^2)-\text{SCR})/2}{\text{ECM}} \tag{6.16}$$

con distribución F_{2,n_1+n_2-4}, si H_0 es cierta.

Test de paralelismo

Ahora queremos comprobar la hipótesis

$$H_0' : \beta_1 = \beta_2$$

para la que $\mathbf{A}$ es 1×4 y $q = \text{rg}\,\mathbf{A} = 1$.
Bajo H_0', la estimación MC de los parámetros α_1, α_2 y $\beta=\beta_1=\beta_2$ se obtiene de la minimización de

$$\xi=\sum_{i=1}^{n_1}(y_{1i}-\alpha_1-\beta x_{1i})^2+\sum_{i=1}^{n_2}(y_{2i}-\alpha_2-\beta x_{2i})^2$$

Las derivadas parciales son

$$\begin{aligned}\frac{\partial\xi}{\partial\alpha_1}&=\sum_{i=1}^{n_1}2(y_{1i}-\alpha_1-\beta x_{1i})(-1)\\ \frac{\partial\xi}{\partial\alpha_2}&=\sum_{i=1}^{n_2}2(y_{2i}-\alpha_2-\beta x_{2i})(-1)\\ \frac{\partial\xi}{\partial\beta}&=\sum_{i=1}^{n_1}2(y_{1i}-\alpha_1-\beta x_{1i})(-x_{1i})+\sum_{i=1}^{n_2}2(y_{2i}-\alpha_2-\beta x_{2i})(-x_{2i})\end{aligned}$$

Al igualar a cero, de las dos primeras ecuaciones tenemos

$$\tilde\alpha_1=\bar y_1-\tilde\beta\bar x_1 \qquad \tilde\alpha_2=\bar y_2-\tilde\beta\bar x_2$$

y si sustituimos en la tercera ecuación

$$\begin{aligned}\tilde\beta&=\frac{\sum_{i=1}^{n_1}x_{1i}(y_{1i}-\bar y_1)+\sum_{i=1}^{n_2}x_{2i}(y_{2i}-\bar y_2)}{\sum_{i=1}^{n_1}x_{1i}(x_{1i}-\bar x_1)+\sum_{i=1}^{n_2}x_{2i}(x_{2i}-\bar x_2)}\\ &=\frac{\sum_{h=1}^{2}\sum_{i=1}^{n_h}(x_{hi}-\bar x_h)(y_{hi}-\bar y_h)}{\sum_{h=1}^{2}\sum_{i=1}^{n_h}(x_{hi}-\bar x_h)^2}\\ &=\frac{r_1(S_{x1}S_{y1})^{1/2}+r_2(S_{x2}S_{y2})^{1/2}}{S_{x1}+S_{x2}}\end{aligned}$$

De modo que la suma de cuadrados es

$$\begin{aligned}\text{SCR}_{H'}&=\sum_{i=1}^{n_1}(y_{1i}-\tilde\alpha_1-\tilde\beta x_{1i})^2+\sum_{i=1}^{n_2}(y_{2i}-\tilde\alpha_2-\tilde\beta x_{2i})^2\\ &=\sum_{h=1}^{2}\sum_{i=1}^{n_h}(y_{hi}-\bar y_h-\tilde\beta(x_{hi}-\bar x_h))^2\\ &=\sum_{h=1}^{2}\sum_{i=1}^{n_h}(y_{hi}-\bar y_h)^2-\tilde\beta^2\sum_{h=1}^{2}\sum_{i=1}^{n_h}(x_{hi}-\bar x_h)^2\end{aligned}$$

y el numerador del test F es

$$\mathrm{SCR}_{H'} - \mathrm{SCR} = \sum_{h=1}^{2} \hat{\beta}_h^2 \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2 - \tilde{\beta}^2 \sum_{h=1}^{2} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$$

Finalmente el estadístico F se puede escribir

$$F = \frac{\hat{\beta}_1^2 S_{x1} + \hat{\beta}_2^2 S_{x2} - \tilde{\beta}^2 (S_{x1} + S_{x2})}{\mathrm{ECM}}$$

que bajo la hipótesis sigue una distribución F_{1,n_1+n_2-4}.

En la práctica, primero se realiza un test de paralelismo y, si se acepta, se realiza el test cuyo estadístico es

$$F = \frac{\mathrm{SCR}_{H'} - \mathrm{SCR}_H}{\mathrm{SCR}_H/(n_1 + n_2 - 3)}$$

Finalmente, y si este último ha sido no significativo, procederemos con el contraste de coincidencia.

Test de concurrencia

Se trata de comprobar la igualdad de los términos independientes de las dos rectas, es decir

$$H_0'' : \alpha_1 = \alpha_2$$

Como en el apartado anterior, se puede ver que el mínimo de la función

$$\xi^* = \sum_{i=1}^{n_1} (y_{1i} - \alpha - \beta_1 x_{1i})^2 + \sum_{i=1}^{n_2} (y_{2i} - \alpha - \beta_2 x_{2i})^2$$

se alcanza cuando

$$\breve{\alpha} = \left(n_1 + n_2 - \frac{x_{1\cdot}^2}{\sum_{i=1}^{n_1} x_{1i}^2} - \frac{x_{2\cdot}^2}{\sum_{i=1}^{n_2} x_{2i}^2} \right)^{-1} \left(y_{\cdot\cdot} - \frac{x_{1\cdot} \sum_{i=1}^{n_1} x_{1i} y_{1i}}{\sum_{i=1}^{n_1} x_{1i}^2} - \frac{x_{2\cdot} \sum_{i=1}^{n_2} x_{2i} y_{2i}}{\sum_{i=1}^{n_2} x_{2i}^2} \right)$$

$$\breve{\beta}_1 = \frac{\sum_{i=1}^{n_1} (y_{1i} - \breve{\alpha}) x_{1i}}{\sum_{i=1}^{n_1} x_{1i}^2} \qquad \breve{\beta}_2 = \frac{\sum_{i=1}^{n_2} (y_{2i} - \breve{\alpha}) x_{2i}}{\sum_{i=1}^{n_2} x_{2i}^2}$$

donde $y_{\cdot\cdot} = \sum_{h=1}^{2} \sum_{i=1}^{n_h} y_{hi}$, $x_{1\cdot} = \sum_{i=1}^{n_1} x_{1i}$ y $x_{2\cdot} = \sum_{i=1}^{n_2} x_{2i}$.

Con estos resultados se puede calcular la suma de cuadrados

$$\mathrm{SCR}_{H''} = \sum_{h=1}^{2} \sum_{i=1}^{n_h} (y_{hi} - \breve{\alpha} - \breve{\beta}_h x_{hi})^2$$

y el estadístico

$$F = \frac{\mathrm{SCR}_{H''} - \mathrm{SCR}}{\mathrm{ECM}}$$

que, bajo H_0'', sigue una distribución F_{1,n_1+n_2-4}.

El test que acabamos de ver contrasta la concurrencia de las dos rectas en $x = 0$. Si deseamos comprobar la concurrencia en un punto $x = c$, bastará aplicar este mismo test sustituyendo los datos x_{hi} por $x_{hi} - c$. Si lo que queremos es saber simplemente si las rectas se cortan (en algún punto), es suficiente con rechazar la hipótesis de paralelismo.

6.7.2. Varias rectas

Supongamos que tenemos la intención de comparar H rectas de regresión

$$Y = \alpha_h + \beta_h x_h + \epsilon \qquad h = 1, \ldots, H$$

donde $E(\epsilon) = 0$ y $\text{var}(\epsilon) = \sigma^2$ es la misma para cada recta. Esta última condición es absolutamente imprescindible para poder aplicar los contrastes estudiados al modelo lineal conjunto que ahora describiremos.

Para cada h, consideremos los n_h pares (x_{hi}, y_{hi}) $i = 1, \ldots, n_h$ de modo que

$$y_{hi} = \alpha_h + \beta_h x_{hi} + \epsilon_{hi} \qquad i = 1, \ldots, n_h$$

con ϵ_{hi} independientes e idénticamente distribuidos como $N(0, \sigma^2)$.

Sea $\mathbf{Y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{H1}, \ldots, y_{Hn_2})'$ y

$$\mathbf{X}\boldsymbol{\gamma} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_H \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_H \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_H \end{pmatrix}$$

donde $\mathbf{x}_h = (x_{h1}, \ldots, x_{hn_h})'$, para cada $h = 1, \ldots, H$.

Con todo ello disponemos del modelo lineal

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

donde $\mathbf{X}$ es $N \times 2H$, con $\text{rg}(\mathbf{X}) = 2H$ y $N = \sum_{h=1}^{H} n_h$.

De esta forma podemos contrastar cualquier hipótesis lineal de la forma $H_0 : \mathbf{A}\boldsymbol{\gamma} = \mathbf{c}$.

La estimación MC de los parámetros α_h, β_h de este modelo se obtiene de cada recta particular

$$\hat{\beta}_h = \frac{\sum_i (y_{hi} - \bar{y}_{h\cdot})(x_{hi} - \bar{x}_{h\cdot})}{\sum_i (x_{hi} - \bar{x}_{h\cdot})^2} = r_h \left(\frac{S_{yh}}{S_{xh}} \right)^{1/2}$$

$$\hat{\alpha}_h = \bar{y}_{h\cdot} - \hat{\beta}_h \bar{x}_{h\cdot}$$

donde $\bar{x}_{h\cdot}, S_{xh}, \bar{y}_{h\cdot}, S_{yh}, r_h$ son las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación para cada una de las muestras $h = 1, \ldots, H$ respectivamente.

También la suma de cuadrados general SCR es simplemente la suma de las sumas de cuadrados de los residuos de cada recta de regresión por separado

$$\begin{aligned} \text{SCR} &= \sum_{h=1}^{H} \left(\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_{h\cdot})^2 - \hat{\beta}_h^2 \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_{h\cdot})^2 \right) \\ &= \sum_{h=1}^{H} \text{SCR}_h = \sum_{h=1}^{H} S_{yh}(1 - r_h^2) \\ &= \sum_{h=1}^{H} S_{yh} - \hat{\beta}_h^2 S_{xh} \end{aligned}$$

Test de coincidencia

Se trata de investigar si las rectas son iguales, es decir, si

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_H (= \alpha) \; ; \; \beta_1 = \beta_2 = \cdots = \beta_H (= \beta)$$

que podemos escribir matricialmente con una matriz $\mathbf{A}$ de tamaño $(2H-2) \times 2H$ de rango $2H-2$.
A partir de las estimaciones MC de los parámetros α, β que se obtienen de la recta ajustada con todos los puntos reunidos en una única muestra, la suma de cuadrados residual es

$$\begin{aligned} \text{SCR}_H &= \sum_{h=1}^{H} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_{..} - \beta^* (x_{hi} - \bar{x}_{..}))^2 \\ &= \sum_{h=1}^{H} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_{..})^2 - (\beta^*)^2 \sum_{h=1}^{H} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_{..})^2 \\ &= S_y (1 - r^2) \end{aligned}$$

donde

$$\beta^* = \frac{\sum_h \sum_i (y_{hi} - \bar{y}_{..})(x_{hi} - \bar{x}_{..})}{\sum_h \sum_i (x_{hi} - \bar{x}_{..})^2} = r \left(\frac{S_y}{S_x} \right)^{1/2}$$

y los estadísticos $\bar{x}_{..}, S_x, \bar{y}_{..}, S_y, r$ son las medias, sumas de cuadrados de las desviaciones y coeficiente de correlación de la muestra conjunta.
Entonces el estadístico F para el contraste de esta hipótesis es

$$F = \frac{(\text{SCR}_H - \text{SCR})/(2H-2)}{\text{SCR}/(N-2H)} \tag{6.17}$$

Contraste de paralelismo

Ahora se trata de investigar si las pendientes de las rectas son iguales, es decir, si

$$H_0' : \beta_1 = \beta_2 = \cdots = \beta_H$$

que matricialmente es equivalente a

$$H_0' : \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \mathbf{0}$$

En este caso, la matriz $\mathbf{A}$ que representa las restricciones de los parámetros es $(H-1) \times 2H$ y su rango es $H-1$. De modo que tomando, en el contraste F, los valores $q = H-1$, $n = N$ y $k = 2H$, el estadístico especificado para este contraste es

$$F = \frac{(\text{SCR}_{H'} - \text{SCR})/(H-1)}{\text{SCR}/(N-2H)}$$

Para calcular el numerador de este estadístico podemos proceder con las fórmulas generales estudiadas u observar las peculiaridades de este modelo que permiten obtener $\text{SCR}_{H'}$.
Primero hay que minimizar $\sum_h \sum_i (y_{hi} - \alpha_h - \beta x_{hi})^2$, de donde se obtienen los estimadores

$$\tilde{\alpha}_h = \bar{y}_{h\cdot} - \tilde{\beta} \bar{x}_{h\cdot} \qquad h = 1, \ldots, H$$

$$\tilde{\beta} = \frac{\sum_h \sum_i x_{hi}(y_{hi} - \bar{y}_{h\cdot})}{\sum_h \sum_i x_{hi}(x_{hi} - \bar{x}_{h\cdot})} = \frac{\sum_h \sum_i (y_{hi} - \bar{y}_{h\cdot})(x_{hi} - \bar{x}_{h\cdot})}{\sum_h \sum_i (x_{hi} - \bar{x}_{h\cdot})^2} = \frac{\sum_h r_h (S_{xh} S_{yh})^{1/2}}{\sum_h S_{xh}}$$

Este último estimador es un estimador conjunto (*pooled*) de la pendiente común.
Con estas estimaciones se procede a calcular la suma de cuadrados

$$\mathrm{SCR}_{H'} = \sum_{h=1}^{H} S_{yh} - \tilde{\beta}^2 \sum_{h=1}^{H} S_{xh}$$

y el estadístico F es

$$F = \frac{(\sum_h \hat{\beta}_h^2 S_{xh} - \tilde{\beta}^2 \sum_h S_{xh})/(H-1)}{\mathrm{SCR}/(N-2H)}$$

que bajo H_0' sigue una distribución $F_{H-1,N-2H}$.
En la práctica, es aconsejable comenzar por un contraste de paralelismo y, si se acepta, continuar con el contraste cuyo estadístico es

$$F = \frac{(\mathrm{SCR}_{H'} - \mathrm{SCR}_H)/(H-1)}{\mathrm{SCR}_H/(N-H-1)}$$

Finalmente, y si este último ha sido no significativo, procederemos con el contraste 6.17.

Test de concurrencia

Deseamos contrastar la hipótesis de que todas las rectas se cortan en un punto del eje de las Y, es decir, para $x = 0$:

$$H_0'' : \alpha_1 = \alpha_2 = \cdots = \alpha_H (= \alpha)$$

En este caso, las estimaciones de los parámetros bajo la hipótesis son

$$\breve{\alpha} = \left(N - \frac{x_{1\cdot}^2}{\sum_i x_{1i}^2} - \cdots - \frac{x_{H\cdot}^2}{\sum_i x_{Hi}^2}\right)^{-1} \left(y_{\cdot\cdot} - \frac{x_{1\cdot} \sum_i x_{1i} y_{1i}}{\sum_i x_{1i}^2} - \cdots - \frac{x_{H\cdot} \sum_i x_{Hi} y_{Hi}}{\sum_i x_{Hi}^2} -\right)$$

$$\breve{\beta}_h = \frac{\sum_i (y_{hi} - \breve{\alpha}) x_{hi}}{\sum_i x_{hi}^2} \qquad h = 1, 2, \ldots, H$$

donde $x_{h\cdot} = \sum_i x_{hi}$ y $y_{\cdot\cdot} = \sum_h \sum_i y_{hi}$.
La suma de cuadrados residual es

$$\mathrm{SCR}_{H''} = \sum_h \sum_i (y_{hi} - \breve{\alpha} - \breve{\beta}_h x_{hi})^2$$

y con ella se puede calcular el estadístico F para el contraste

$$F = \frac{(\mathrm{SCR}_{H''} - \mathrm{SCR})/(H-1)}{\mathrm{SCR}/(N-2H)}$$

Cuando los valores de las x son los mismos para todas las rectas, tenemos que $n_h = n$ y $x_{hi} = x_i$ para toda $h = 1, \ldots, H$ y así las fórmulas son más simples

$$\breve{\alpha} = \left(Hn - \frac{Hx_{\cdot}^2}{\sum_i x_i^2}\right)^{-1} \left(y_{\cdot\cdot} - \frac{x_{\cdot} \sum_i x_i y_{\cdot i}}{\sum_i x_i^2}\right) = \bar{y}_{\cdot\cdot} - \frac{\bar{x} \sum_h \sum_i y_{hi}(x_i - \bar{x})}{H \sum_i (x_i - \bar{x})^2} = \bar{y}_{\cdot\cdot} - \bar{x} \frac{\sum_h \hat{\beta}_h}{H}$$

donde cada $\hat{\beta}_h$ es la estimación de la pendiente de la h-ésima recta, mientras que $\breve{\alpha}$ es el corte de la recta de regresión *media*.

En este caso

$$\mathrm{SCR}_{H''} = \sum_h \sum_i y_{hi}^2 - \frac{\left(\sum_h \sum_i x_i y_{hi}\right)^2}{\sum_i x_i^2} - \breve{\alpha}^2 Hn \frac{\sum_i (x_i - \overline{x})^2}{\sum_i x_i^2}$$

Además, como $\overline{y}_{\cdot\cdot}$ y $\hat{\beta}_h$ están incorrelacionados

$$\begin{aligned} \mathrm{var}(\breve{\alpha}) &= \mathrm{var}(\overline{y}_{\cdot\cdot}) + H\overline{x}^2 \frac{\mathrm{var}(\hat{\beta}_h)}{H^2} \\ &= \frac{\sigma^2}{H}\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_i (x_i - \overline{x})^2}\right) = \frac{\sigma^2 \sum_i x_i^2}{nH \sum_i (x_i - \overline{x})^2} \end{aligned}$$

de modo que tenemos la posibilidad de construir un intervalo de confianza para α ya que

$$(\breve{\alpha} - \alpha)\left(\frac{nH \sum_i (x_i - \overline{x})^2}{\mathrm{ECM} \sum_i x_i^2}\right)^{1/2} \sim t_{H(n-2)}$$

donde $\mathrm{ECM} = \mathrm{SCR}/(nH - 2H)$.

Por otra parte, también podemos estudiar si las rectas se cortan en un punto $x = c$ distinto del cero. Simplemente reemplazaremos x_{hi} por $x_{hi} - c$ en todas las fórmulas anteriores. La coordenada y del punto de corte sigue siendo estimada por $\breve{\alpha}$.

Sin embargo, si el punto de corte es desconocido $x = \phi$, la hipótesis a contrastar es mucho más complicada

$$H_0''' : \alpha_h + \beta_h \phi = \text{cte.} = \overline{\alpha} + \overline{\beta}\phi \qquad h = 1, 2, \ldots, h$$

o también

$$H_0''' : \frac{\alpha_1 - \overline{\alpha}}{\beta_1 - \overline{\beta}} = \cdots = \frac{\alpha_H - \overline{\alpha}}{\beta_H - \overline{\beta}}$$

y desgraciadamente no es lineal.

6.7.3. Contraste para la igualdad de varianzas

En los contrastes de comparación de rectas se hace la suposición de la igualdad de las varianzas σ_h^2 de los modelos lineales simples $h = 1, \ldots, H$.

Los estimadores de dichas varianzas son los errores cuadráticos medios particulares

$$S_h^2 = \frac{\sum_i (y_{hi} - \overline{y}_{h\cdot} - \hat{\beta}_h (x_{hi} - \overline{x}_{h\cdot}))^2}{n_h - 2}$$

y sabemos que

$$(n_h - 2) S_h^2 / \sigma_h^2 \sim \chi^2_{n_h - 2} \qquad h = 1, \ldots, H \qquad \text{indep.}$$

Para contrastar la hipótesis

$$H_0 : \sigma_1^2 = \cdots = \sigma_H^2$$

hay varios métodos, desde los más clásicos de Bartlett(1937) o Hartley(1950), muy sensibles a la no normalidad de los datos, hasta los más robustos entre los que destaca el de Levene con sus variantes.

Si hacemos $f_h = n_h - 2$, el test de Bartlett es

$$T = \frac{(\sum f_h) \log S^2 - \sum (f_h \log S_h^2)}{C}$$

donde

$$S^2 = \frac{\sum f_h S_h^2}{\sum f_h} \qquad C = 1 + \frac{\sum f_h^{-1} - (\sum f_h)^{-1}}{3(H-1)}$$

Si H_0 es cierta, aproximadamente $T \sim \chi^2_{H-1}$.

Cuando los f_h son todos iguales, Hartley propone el estadístico

$$F = \frac{\text{máx}\{S_1^2, \ldots, S_H^2\}}{\text{mín}\{S_1^2, \ldots, S_H^2\}}$$

Sin embargo, como se trata de comparar las varianzas a partir de las observaciones o *réplicas* de H poblaciones, es mejor considerar el problema como un análisis de la varianza de un factor. La prueba robusta de Levene sobre la homogeneidad de varianzas se basa en el análisis de la varianza de un factor con los datos $z_{hi} = |y_{hi} - \bar{y}_{h\cdot}|$. Para reforzar la resistencia del método se puede utilizar como medida de localización la mediana.

Finalmente podemos añadir que, cuando la heterogeneidad de las varianzas es evidente, siempre es posible estudiar alguna transformación potencia de los datos originales y_{hi} que mejore la situación.

6.8. Un ejemplo para la reflexión

La siguiente tabla presenta cinco conjuntos de datos para cinco modelos de regresión simple diferentes: los datos bajo el encabezamiento x_1(a-d) son los valores de una variable regresora que es común en las cuatro regresiones con las variables respuesta y(a), y(b), y(c) y y(d). Las series de datos x(e) y y(e) definen otra regresión.

obs.	x_1(a-d)	y(a)	y(b)	y(c)	y(d)	x(e)	y(e)
1	7	5,535	0,103	7,399	3,864	13,715	5,654
2	8	9,942	3,770	8,546	4,942	13,715	7,072
3	9	4,249	7,426	8,468	7,504	13,715	8,496
4	10	8,656	8,792	9,616	8,581	13,715	9,909
5	12	10,737	12,688	10,685	12,221	13,715	9,909
6	13	15,144	12,889	10,607	8,842	13,715	9,909
7	14	13,939	14,253	10,529	9,919	13,715	11,327
8	14	9,450	16,545	11,754	15,860	13,715	11,327
9	15	7,124	15,620	11,676	13,967	13,715	12,746
10	17	13,693	17,206	12,745	19,092	13,715	12,746
11	18	18,100	16,281	13,893	17,198	13,715	12,746
12	19	11,285	17,647	12,590	12,334	13,715	14,164
13	19	21,385	14,211	15,040	19,761	13,715	15,582
14	20	15,692	15,577	13,737	16,382	13,715	15,582
15	21	18,977	14,652	14,884	18,945	13,715	17,001
16	23	17,690	13,947	29,431	12,187	33,281	27,435

Tabla 6.4: Datos de cinco regresiones simples

Se puede comprobar que, en los cinco casos, la regresión de y sobre x conduce exactamente a la misma recta

$$y = 0.520 + 0.809x$$

La varianza explicada, la no explicada i la varianza residual son idénticas en todas las regresiones, así como también el coeficiente de determinación.

Por lo tanto, las cinco regresiones parecen ser formalmente idénticas. A pesar de ello, si dibujamos en cada caso los diagramas de dispersión y la recta de regresión, observaremos que nuestra impresión se modifica radicalmente: en la página 114 tenemos los gráficos para los cinco conjuntos de datos.

número de obs.	$n = 16$	$\hat{\beta}_1 = 0.809$	$ee(\hat{\beta}_1)=0{,}170$
media de las x_1	$\bar{x}_1 = 14.938$	$\hat{\beta}_0 = 0.520$	$ee(\hat{\beta}_0)=2{,}668$
media de las y	$\bar{y} = 12.600$	$R^2 = 0.617$	

$\sum (y_i - \bar{y})^2 = 380.403$ con 15 g.l. $\quad \sum (y_i - \hat{y}_i)^2 = 145.66$ con 14 g.l.

$\hat{\sigma} = 3.226$

Tabla 6.5: Principales resultados de la regresión simple

- La figura **a** es la que representan todos los manuales que explican la regresión simple. El modelo de la regresión lineal simple parece correcto y adaptado a los datos que permite describir correctamente. El modelo parece válido.
- La figura **b** sugiere que el modelo lineal simple no está absolutamente adaptado a los datos que pretende describir. Más bien, la forma adecuada es la cuadrática con una débil variabilidad. El modelo lineal simple es incorrecto; en particular, las predicciones que él proporciona son sesgadas: subestimaciones para los valores próximos a la media de x y sobreestimaciones para el resto.
- La figura **c** sugiere todavía que el modelo lineal simple no se adapta a los datos, pero una única observación parece ser la causa. Por contra, las otras observaciones están bien alineadas pero respecto a otra recta de ecuación $y = 4.242 + 0.503x_1$. Hay pues, un dato verdaderamente sospechoso. La reacción natural del experimentador será la de investigar con detalle la razón de esta desviación. ¿No será un error de transcripción? ¿Hay alguna causa que justifique la desviación y que no tiene en cuenta el modelo lineal simple?
- La figura **d** tiene un análisis más sutil: los puntos rodean la recta, pero aumentan las desviaciones a medida que crecen los valores de la variable regresora. Se hace evidente que la suposición de una varianza común de los residuos no se verifica.
- Finalmente, la figura **e** es más contundente: el modelo parece correcto. Si la calidad de los datos no puede ponerse en duda, este conjunto es tan válido como el primero y los resultados numéricos de la regresión son correctos. Pero nosotros intuimos que este resultado no es lo suficientemente satisfactorio: todo depende de la presencia de un único punto, si lo suprimimos, incluso no será posible calcular la pendiente de la recta, ya que la suma de los cuadrados de las desviaciones de las x es cero. Éste no es el caso del primer conjunto de datos, donde la supresión de un punto no conduce más que a una ligera modificación de los resultados. Así pues, deberíamos ser extremadamente cautos con las posibles utilizaciones de este modelo. Además, debemos indicar que el experimento definido por los valores de x es muy malo.

Naturalmente, los conjuntos de datos **b**, **c**, **d** y **e** muestran casos extremos que, en la práctica, no hallaremos de forma tan clara. Ésta es una razón suplementaria para dotar al experimentador de medios para detectarlos. Cuando las desviaciones de las suposiciones del modelo son débiles, los resultados no serán erróneos, pero si las suposiciones son groseramente falsas, las conclusiones pueden incluso no tener sentido. La herramienta fundamental para la validación de las hipótesis del modelo es el análisis de los residuos del modelo estimado.

El **análisis de los residuos** (ver capítulo 9) tiene como objetivo contrastar a posteriori las hipótesis del modelo lineal. Es especialmente importante cuando, si tenemos un único valor de y para cada x, los contrastes de homocedasticidad, normalidad e independencia no se pueden hacer a priori. Analizaremos los residuos para comprobar:

a) si la distribución es aproximadamente normal;

b) si su variabilidad es constante y no depende de x o de otra causa;

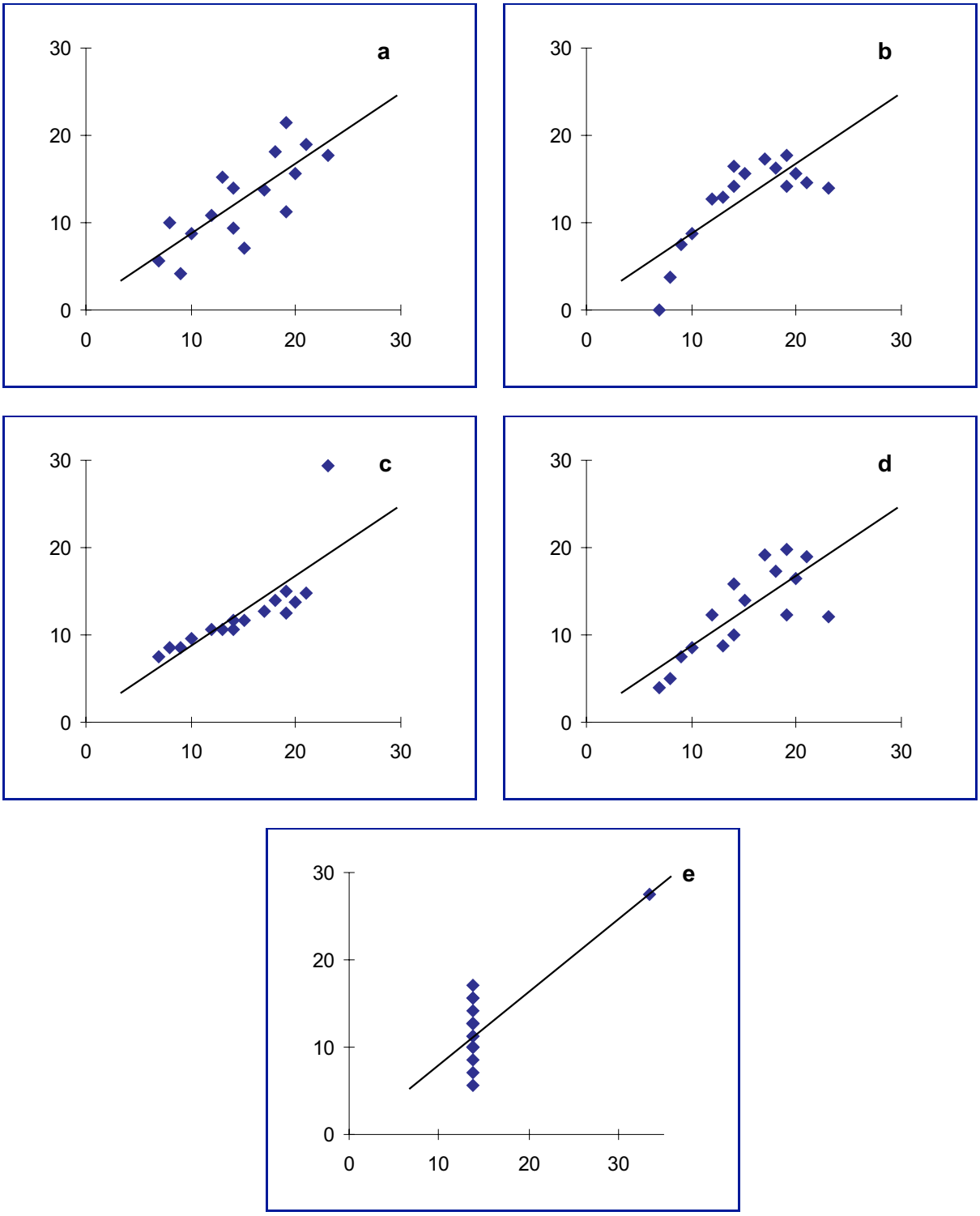

Figura 6.2: Gráficos de los cinco conjuntos de datos con la recta de regresión

c) si presentan evidencia de una relación no lineal entre las variables;

d) si existen observaciones atípicas o heterogéneas respecto a la variable x, la y o ambas.

6.9. Ejemplos con R

Vamos a recuperar el ejemplo de la sección 1.8 donde se calculan algunas regresiones a partir del ejemplo inicial con los datos de la tabla 1.1. En esa sección, el cálculo de la regresión simple se realiza con la función `lsfit(x,y)` que asignamos al objeto `recta.ls`

```
> recta.ls<-lsfit(dens,rvel)
```

Ahora utilizaremos la función `lm` que define el modelo de regresión simple.

```
> recta<-lm(rvel~dens)
> recta
Call:
lm(formula = rvel ~ dens)

Coefficients:
 (Intercept)         dens
    8.089813  -0.05662558

Degrees of freedom: 24 total; 22 residual
Residual standard error: 0.2689388
```

También se pueden obtener otros datos importantes con la función `summary`:

```
> recta.resumen<-summary(recta)
> recta.resumen

Call: lm(formula = rvel ~ dens)
Residuals:
     Min       1Q   Median       3Q      Max
 -0.3534  -0.2272  -0.03566  0.1894  0.5335

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  8.0898    0.1306   61.9295   0.0000
       dens -0.0566    0.0022  -26.0076   0.0000

Residual standard error: 0.2689 on 22 degrees of freedom
Multiple R-Squared: 0.9685
F-statistic: 676.4 on 1 and 22 degrees of freedom, the p-value is 0

Correlation of Coefficients:
     (Intercept)
dens -0.9074
```

Además se puede acceder a muchos valores de los objetos `recta` y `recta.resumen` de forma directa.

```
> recta$coef
 (Intercept)         dens
    8.089813  -0.05662558
> recta.resumen$sigma
[1] 0.2689388
```

En general, podemos saber los diferentes resultados que se obtienen con el comando `lm` si escribimos `names(recta)` o `names(summary(recta))`.

```
> names(recta)
 [1] "coefficients" "residuals" "fitted.values" "effects" "R" "rank"
```

```
 [7] "assign" "df.residual" "contrasts" "terms" "call"
> names(summary(recta))
 [1] "call" "terms" "residuals" "coefficients" "sigma" "df"
 [7] "r.squared" "fstatistic" "cov.unscaled" "correlation"
```

De modo que podemos utilizar estos datos para nuevos cálculos. Por ejemplo podemos calcular la matriz estimada de covarianzas entre los estimadores de los parámetros $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ así:

```
> cov.beta<-round(recta.resumen$sigma^2*recta.resumen$cov.unscaled,6)
> cov.beta
            (Intercept)      dens
(Intercept)    0.017064 -0.000258
       dens   -0.000258  0.000005
```

Por otra parte, y aunque el resumen proporcionado por la función `summary(recta)` incluye el test F de significación de la regresión, la tabla del Análisis de la Varianza se puede calcular con la función `aov`.

```
> summary(aov(recta))
            Df Sum of Sq  Mean Sq  F Value Pr(F)
     dens    1  48.92231 48.92231 676.3944     0
Residuals 22   1.59122  0.07233
```

También se pueden calcular intervalos de confianza al 95 % para los parámetros β_0, β_1.

```
> coef(recta)
 (Intercept)        dens
    8.089813 -0.05662558
> coef.recta<-coef(recta)
> names(coef.recta)
[1] "(Intercept)" "dens"
> names(coef.recta)<-NULL # Truco para utilizar mejor los coeficientes
> coef.recta
        1           2
 8.089813 -0.05662558
> ee0<-sqrt(cov.beta[1,1])
> ee1<-sqrt(cov.beta[2,2])
> c(coef.recta[1]+qt(0.025,22)*ee0,coef.recta[1]+qt(0.975,22)*ee0)
[1] 7.818905 8.360721
> c(coef.recta[2]+qt(0.025,22)*ee1,coef.recta[2]+qt(0.975,22)*ee1)
[1] -0.06126290 -0.05198826
```

Cabe señalar que si el modelo de regresión simple debe pasar por el origen, es decir, no tiene término de intercepción, podemos utilizar la función `lsfit(x,y,int=F)` o la función `lm(y ~ x - 1)`.

La predicción puntual o por intervalo de nuevos valores de la variable respuesta se puede hacer con la función `predict` del modelo lineal. Atención, porque los argumentos en R y S-PLUS difieren.

Por último, podemos añadir que en R existe un conjunto de datos similares a los explicados en la sección 6.8:

```
> data(anscombe)
> summary(anscombe)
```

6.10. Ejercicios

Ejercicio 6.1

Probar que bajo el modelo lineal normal $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ las estimaciones MC $\hat{\beta}_0, \hat{\beta}_1$ son estocásticamente independientes si y sólo si $\sum x_i = 0$.

Ejercicio 6.2

Comprobar que la pendiente de la recta de regresión es

$$\hat{\beta}_1 = r\frac{S_y^{1/2}}{S_x^{1/2}} = r\frac{s_y}{s_x}$$

donde r es el coeficiente de correlación

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

Ejercicio 6.3

Consideremos el modelo de regresión simple alternativo

$$y_i = \gamma_0 + \gamma_1(x_i - \bar{x}) + \epsilon_i \qquad i = 1, \ldots, n$$

La matriz de diseño asociada es $\mathbf{X}_* = (\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1})$ donde $\mathbf{1} = (1, \ldots, 1)'$ y $\mathbf{x} = (x_1, \ldots, x_n)'$. Este modelo es equivalente al modelo 6.1 ya que $\langle \mathbf{X}_* \rangle = \langle \mathbf{X} \rangle$.
Calcular las estimaciones $\widehat{\boldsymbol{\gamma}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{Y}$ para comprobar que

$$\hat{\gamma}_0 = \bar{y}$$
$$\hat{\gamma}_1 = \hat{\beta}_1 = \sum \frac{x_i - \bar{x}}{S_x} y_i$$

Calcular la matriz de varianzas-covarianzas $\text{var}(\widehat{\boldsymbol{\gamma}}) = \sigma^2(\mathbf{X}_*'\mathbf{X}_*)^{-1}$ y comprobar que $\hat{\gamma}_0 = \bar{y}$ está incorrelacionado con $\hat{\gamma}_1 = \hat{\beta}_1$. A partir de este resultado, calcular $\text{var}(\hat{\beta}_1) = \text{var}(\hat{\gamma}_1)$ y $\text{var}(\hat{\beta}_0) = \text{var}(\bar{y} - \hat{\beta}_1\bar{x})$.
Calcular también la matriz proyección $\mathbf{P} = \mathbf{X}_*(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Ejercicio 6.4

En un modelo de regresión simple, con β_0, demostrar que se verifican las siguientes propiedades para las predicciones $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y los residuos $e_i = y_i - \hat{y}_i$:

(i) La suma de los residuos es cero: $\sum e_i = 0$.

(ii) $\sum y_i = \sum \hat{y}_i$

(iii) La suma de los residuos ponderada por los valores de la variable regresora es cero: $\sum x_i e_i = 0$.

(iv) La suma de los residuos ponderada por las predicciones de los valores observados es cero: $\sum \hat{y}_i e_i = 0$.

Ejercicio 6.5 Modelo de regresión simple estandarizado

A partir de los datos observados de una variable respuesta y_i y de una variable regresora x_i se definen unas nuevas variables estandarizadas como

$$u_i = \frac{x_i - \bar{x}}{S_x^{1/2}} \qquad v_i = \frac{y_i - \bar{y}}{S_y^{1/2}} \qquad i = 1, \ldots, n$$

La estandarización significa que los datos transformados están centrados y los vectores $\mathbf{u} = (u_1, \ldots, u_n)'$, $\mathbf{v} = (v_1, \ldots, v_n)'$ son de longitud uno, es decir, $\|\mathbf{u}\| = 1$ y $\|\mathbf{v}\| = 1$.

Se define el modelo de regresión simple estandarizado como

$$v_i = b_1 u_i + \epsilon_i \qquad i = 1, \ldots, n$$

En este modelo desaparece de manera natural la ordenada en el origen al realizar el cambio de variables.

Comprobar que

$$\hat{\beta}_1 = \hat{b}_1 \sqrt{\frac{S_y}{S_x}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Además, la "matriz" $\mathbf{u}'\mathbf{u} = \|\mathbf{u}\|^2 = 1$ y la estimación de b_1 es muy sencilla $\hat{b}_1 = r$.

Ejercicio 6.6

En el caso de una regresión lineal simple pasando por el origen y con la hipótesis de normalidad, escribir el contraste de la hipótesis $H_0 : \beta_1 = b_1$, donde b_1 es una constante conocida.

Ejercicio 6.7

Para el modelo lineal simple consideremos la hipótesis

$$H_0 : y_0 = \beta_0 + \beta_1 x_0$$

donde (x_0, y_0) es un punto dado. Esta hipótesis significa que la recta de regresión pasa por el punto (x_0, y_0). Construir un test para esta hipótesis.

Ejercicio 6.8

Hallar la recta de regresión simple de la variable respuesta *raíz cuadrada de la velocidad* sobre la variable regresora *densidad* con los datos de la tabla 1.1 del capítulo 1.

Comprobar las propiedades del ejercicio 6.4 para estos datos.

Calcular la estimación de σ^2 y, a partir de ella, las estimaciones de las desviaciones estándar de los estimadores de los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$.

Escribir los intervalos de confianza para los parámetros con un nivel de confianza del 95 %.

Construir la tabla para la significación de la regresión y realizar dicho contraste.

Hallar el intervalo de la predicción de la respuesta media cuando la densidad es de 50 vehículos por km. Nivel de confianza: 90 %.

Ejercicio 6.9

Comparar las rectas de regresión de hombres y mujeres con los logaritmos de los datos del ejercicio 1.4.

Ejercicio 6.10

Se admite que una persona es *proporcionada* si su altura en cm es igual a su peso en kg más 100. En términos estadísticos si la recta de regresión de Y (altura) sobre X (peso) es

$$Y = 100 + X$$

Contrastar, con un nivel de significación $\alpha = 0.05$, si se puede considerar válida esta hipótesis a partir de los siguientes datos que corresponden a una muestra de mujeres jóvenes:

X:	55	52	65	54	46	60	54	52	56	65	52	53	60
Y:	164	164	173	163	157	168	171	158	169	172	168	160	172

Razonar la bondad de la regresión y todos los detalles del contraste.

Ejercicio 6.11

El período de oscilación de un péndulo es $2\pi\sqrt{\frac{l}{g}}$, donde l es la longitud y g es la constante de gravitación. En un experimento observamos t_{ij} $(j = 1, \ldots, n_i)$ períodos correspondientes a l_i $(i = 1, \ldots, k)$ longitudes.

(a) Proponer un modelo, con las hipótesis que se necesiten, para estimar la constante $\frac{2\pi}{\sqrt{g}}$ por el método de los mínimos cuadrados.

(b) En un experimento se observan los siguientes datos:

longitud	período
18.3	8.58 7.9 8.2 7.8
20	8.4 9.2
21.5	9.7 8.95 9.2
15	7.5 8

Contrastar la hipótesis $H_0 : \frac{2\pi}{\sqrt{g}} = 2$.