Capítulo 9 - Regressão Linear Simples (RLS): Notas breves

Regressão Linear Simples

Estrutura formal do modelo de Regressão Linear Simples (RLS):

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

onde

Y_i: variável resposta (ou variável dependente) associada à i-ésima observação; (aleatória).

x_i: constante conhecida, designada por variável explicativa (independente ou regressora) associada à i-ésima observação; (não aleatória).

β_0: parâmetro regressor (desconhecido);

β_1: parâmetro regressor (desconhecido);

ε_i: erro aleatório associado à i-ésima observação, verificando:

(i) $E[\varepsilon_i] = 0$ (valor esperado nulo);

(ii) $Var[\varepsilon_i] = \sigma^2$ (variância constante);

(iii) $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall_{i,j} \ \ (i \neq j)$ (não correlacionados);

$i = 1, 2, \ldots, n$.

Como $E[\varepsilon_i] = 0$ então $E[Y|x_i] = \beta_0 + \beta_1 x_i$ (recta de regressão). Ou seja, para cada valor x_i o ponto sobre a recta tem ordenada $E[Y|x_i] = \beta_0 + \beta_1 x_i$. Como consequência, β_0 é a ordenada na origem e β_1 é o declive ou coeficiente angular da recta.

Estimadores pontuais de mínimos quadrados de β_0, β_1, σ^2 e $E[Y|x_i] = \beta_0 + \beta_1 x_i$:

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$;
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$;

  onde

  $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.
- $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 =$

  $= \frac{1}{n-2} \left[\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \right) - \left(\hat{\beta}_1\right)^2 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \right]$,

  onde
- $\hat{Y}_i = \hat{E}[Y|x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i$, (ponto na recta estimada) e a recta de regressão estimada é $\hat{Y} = \hat{E}[Y|x] = \hat{\beta}_0 + \hat{\beta}_1 x$, $\forall x \in (\min(x_i), \max(x_i))$.

Observação: O resíduo da i-ésima observação é denotado por $e_i = r_i = (y_i - \hat{y}_i)$.

Inferências: intervalos e testes de hipóteses em RLS

Para se fazerem inferências para o modelo de RLS é necessário admitir que os erros têm distribuição normal.

Atrás admitiu-se que $E[\varepsilon_i] = 0$, $Var[\varepsilon_i] = \sigma^2$ e $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall_{i,j}$ $(i \neq j)$ e com a nova suposição da normalidade tem-se agora que $\varepsilon_i \underset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$.

Uma vez que $\varepsilon_i \underset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$ então $Y_i \underset{ind}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$. Como os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são combinações lineares das variáveis aleatórias Y_i então também se tem:

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\sigma^2\right)$$

e

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}\right).$$

Como σ^2 é desconhecido utiliza-se $\hat{\sigma}^2$. Pode mostrar-se que

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

e ainda que $\hat{\beta}_0$ e $\hat{\beta}_1$ são independentes de σ^2, o que conduz a

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)},$$

e

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}.$$

A) Inferências relativas à ordenada na origem, β_0.

Teste de hipóteses:

Hipótese nula: $H_0 : \beta_0 = \beta_{0,0}$

Quando H_0 é verdadeira a estatística de teste é:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

Pretende-se testar $H_0 : \beta_0 = \beta_{0,0}$ contra uma das alternativas:

(a) $H_1 : \beta_0 \neq \beta_{0,0}$ (teste bilateral)

(b) $H_1 : \beta_0 > \beta_{0,0}$ (teste unilateral superior/direita)

(c) $H_1 : \beta_0 < \beta_{0,0}$ (teste unilateral inferior/esquerda)

Região Rejeição (ao nível de significância α):

(a) $RR_\alpha : |T_0| > c$, com $c = F^{-1}_{t_{(n-2)}}(1-\alpha/2)$.

(b) $RR_\alpha : T_0 > c$, com $c = F^{-1}_{t_{(n-2)}}(1-\alpha)$.

(c) $RR_\alpha : T_0 < c$, com $c = F^{-1}_{t_{(n-2)}}(\alpha)$.

Decisão usual.

Intervalo de confiança:

Usando a seguinte variável aleatória fulcral:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}$$

obtém-se, após dedução, o seguinte intervalo para β_0 a $(1-\alpha)\times 100\%$ de confiança

$$I.C._{(1-\alpha)\times 100\%}(\beta_0) = \left(\hat{\beta}_0 \pm F^{-1}_{t_{n-2}}(1-\alpha/2)\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}\right)$$

B) Inferências relativas ao declive, β_1.

Teste de hipóteses:

Hipótese nula: $H_0 : \beta_1 = \beta_{1,0}$

Quando H_0 é verdadeira a estatística de teste é:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}.$$

Pretende-se testar $H_0 : \beta_1 = \beta_{1,0}$ contra uma das alternativas:

(a) $H_1 : \beta_1 \neq \beta_{1,0}$ (teste bilateral)

(b) $H_1 : \beta_1 > \beta_{1,0}$ (teste unilateral superior/direita)

(c) $H_1 : \beta_1 < \beta_{1,0}$ (teste unilateral inferior/esquerda)

Região Rejeição (ao nível de significância α):

(a) $RR_\alpha : |T_0| > c$, com $c = F^{-1}_{t_{(n-2)}}(1-\alpha/2)$.

(b) $RR_\alpha : T_0 > c$, com $c = F^{-1}_{t_{(n-2)}}(1-\alpha)$.

(c) $RR_\alpha : T_0 < c$, com $c = F^{-1}_{t_{(n-2)}}(\alpha)$.

Decisão usual.

Um teste importante em RLS é

$$H_0 : \beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$$

pois não rejeitar H_0 significa que há evidência para a não existência de uma associação linear entre x e y, ou seja não há associação ou a associação não é linear. A este teste costuma designar-se por teste à significância da regressão.

Intervalo de confiança:

Usando a seguinte variável aleatória fulcral:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(n-2)}$$

obtém-se, após dedução, o seguinte intervalo para β_1 a $(1-\alpha)\times 100\%$ de confiança é

$$I.C._{(1-\alpha)\times 100\%}(\beta_1) = \left(\hat{\beta}_1 \pm F^{-1}_{t_{n-2}}(1-\alpha/2) \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} \right)$$

C) Inferências para um ponto na recta de regressão: $E[Y|x_0] = \beta_0 + \beta_1 x_0$

Um estimador pontual de $E[Y|x_0] = \beta_0 + \beta_1 x_0$ é $\hat{E}[Y|x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Pode mostrar-se que

$$T = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(n-2)}.$$

Testes de hipóteses e intervalos de confiança são então baseados nesta variável, e o procedimento é o habitual.

Coeficiente de determinação

O coeficiente de determinação é uma medida descritiva indicadora da qualidade do ajustamento da recta estimada. Mostra-se que:

$$\underset{SST}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = \underset{SSE}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} + \underset{SSR,}{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}$$

onde SST corresponde à soma de quadrados total, SSE corresponde à soma dos quadrados dos resíduos e SSR é a soma dos quadrados da regressão.

O coeficiente de determinação é:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} =$$

$$= \frac{\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)} = \hat{\beta}_1 \frac{\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}\right)}{\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}.$$

$R^2 \times 100\%$ representa a percentagem de variabilidade total de Y que é explicada pelo modelo de regressão. $0 \leq R^2 \leq 1$, quando todos os pontos $\hat{y}_i = y_i$ tem-se $r^2 = 1$ o que significa que existe um ajustamento perfeito. Quando $\hat{\beta}_1 = 0$ tem-se $r^2 = 0$ então o ajustamento não é adequado, a variável x não explica nada da variabilidade de y.

Exemplo: Para explorar a relação entre a massa muscular e a idade (no sexo feminino) um nutricionista seleccionou aleatoriamente 16 mulheres com idades compreendidas entre os 40 e os 79 anos. Os resultados observados encontram-se na tabela seguinte (x representa a idade e y é um índice de massa muscular):

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_i	71	64	43	67	56	73	68	56	76	65	45	58	45	53	49	78
y_i	82	91	100	68	87	73	78	80	65	84	116	76	97	100	105	77

$\sum_{i=1}^{16} x_i = 967$, $\sum_{i=1}^{16} y_i = 1379$, $\sum_{i=1}^{16} x_i^2 = 60409$, $\sum_{i=1}^{16} y_i^2 = 121887$, $\sum_{i=1}^{16} x_i y_i = 81331$.
Admita que o modelo de regressão linear simples ($Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) é adequado.

a) Calcule:

   i) uma estimativa pontual da diferença entre as massas musculares médias de mulheres cujas idades diferem de um ano;

   ii) uma estimativa pontual da massa muscular média para as mulheres de 60 anos;

   iii) o valor do resíduo para a 8ª observação;

   iv) uma estimativa pontual de $Var(\varepsilon_i) = \sigma^2$;

   v) o coeficiente de determinação e interprete o valor obtido.

b) O nutricionista pensa que (na gama de idades considerada) a massa muscular é significativamente influenciada pela idade. Acha que as observações feitas confirmam esta hipótese ? Use um nível de significância de 5% e indique as hipóteses de trabalho de que necessita para efectuar o teste.

c) Calcule o intervalo de confiança a 99% para o valor esperado da massa muscular para uma mulher de 45 anos. Acha legítimo usar o mesmo procedimento tratando-se de uma mulher com 20 anos em vez de 45?

Resolução:

Modelo de RLS:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

onde

Y_i: variável aleatória que representa o índice da massa muscular associada à i-ésima mulher;

x_i: variável explicativa que representa a idade da i-ésima mulher;

β_0: ordenada na origem;

β_1: parâmetro regressor;

ε_i: erro aleatório associado à i-ésima mulher, verificando:

(i) $E[\varepsilon_i] = 0$;

(ii) $Var[\varepsilon_i] = \sigma^2$;

(iii) $Cov[\varepsilon_i, \varepsilon_j] = 0, \forall_{i,j} \ \ (i \neq j)$;

$i = 1, 2, \ldots, 16$.

Cálculos auxiliares:

$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 81331 - 16 \times \frac{967}{16} \times \frac{1379}{16} = -2012.3125$

$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 60409 - (967^2)/16 = 1965.9375$

$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 121887 - (1379^2)/16 = 3034.4375$

a) i) Como $E[Y|(x_j + 1)] - E[Y|(x_j)] = \beta_0 + \beta_1(x_j + 1) - \beta_0 - \beta_1 x_j = \beta_1$, então uma estimativa dessa diferença será

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{-2012.3125}{1965.9375} \approx -1.0236$$

ii) A estimativa pretendida é
$\hat{E}[Y|x=60] = \hat{\beta}_0 + \hat{\beta}_1 \times 60 = 148.05 - 1.0236 \times 60 \approx 86.63$, pois
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1379/16 + 1.0236 \times (967/16) \approx 148.05$.

iii) Representando $\hat{y}_i = \hat{E}[Y|x_i]$ sabe-se que o oitavo resíduo é dado por $e_8 = y_8 - \hat{y}_8 = 80 - 90.728 \approx -10.7284$, já que $\hat{y}_8 = 148.05 - 1.0236 \times 56 = 90.728$.

iv) Uma estimativa pontual de $Var(\varepsilon_i) = \sigma^2$, $\forall i$, $i = 1, 2, \ldots, 16$ é
$\hat{\sigma}^2 = \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) - \left(\hat{\beta}_1\right)^2 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\right] =$

$= \frac{1}{14}(3034.4375) - (-1.0236)^2 (1965.9375) \approx 69.62$

v) O valor para o coeficiente de determinação é
$r^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)} = \frac{(-2012.3125)^2}{(1965.9375)(3034.4375)} \approx 0.679.$
Este valor indica que cerca de 67% da variabilidade deste índice de massa muscular é explicada pela idade (da mulher).

b) O que o nutricionista pretende testar pode ser traduzido no seguinte teste de hipóteses:

$$H_0 : \beta_1 = 0 \;\; vs \;\; H_1 : \beta_1 \neq 0$$

Para se efectuar este teste de hipóteses é necessário considerar a seguinte hipótese de trabalho:

$$\varepsilon_i \underset{i.i.d.}{\sim} N\left(0, \sigma^2\right), \quad \forall_i, \quad i = 1, 2, \ldots, 16.$$

Nível de significância: $\alpha = 0.05$;

Quando H_0 é verdadeira a estatística de teste é:

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \sim t_{(14)}.$$

Região Rejeição para T_0:

$RR_{0.05} = (-\infty, -c) \cup (c, +\infty)$, onde $c = P(T_0 \in RC|H_0) = 0.05$. Então o valor de c será dado por

$c = F^{-1}_{t(14)}\left(1 - \frac{\alpha}{2}\right) = F^{-1}_{t(14)}(1 - 0.025) = F^{-1}_{t(14)}(0.975) = 2.145,$

logo

$RR_{0.05} = (-\infty, -2.145) \cup (2.145, +\infty).$

Valor observado da estatística de teste:

$$t_0 = \frac{-1.0236}{\sqrt{\frac{69.62}{1965.9375}}} \approx -5.44$$

Decisão:

Como o valor observado da estatística de teste, $t_0 = -5.44 \in RR_{0.05}$, devemos rejeitar H_0 ao nível de significância de 5%, ou seja, parece haver evidência de que a idade da mulher influencia a sua massa muscular.

c) Pretende-se um intervalo de confiança a 99% para $E[Y|x = 45]$. Observar que $x = 45 \in (\min(x_i), \max(x_i))$.

Variável aleatória fulcral:

$$T = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \sim t_{(14)},$$

com $x_0 = 45$.

Como $1 - \alpha = 0.99$ então $\alpha = 0.01$ e $a = F^{-1}_{t(14)}\left(1 - \frac{\alpha}{2}\right) = F^{-1}_{t(14)}(1 - 0.005) = F^{-1}_{t(14)}(0.995) = 2.977$.

Como a distribuição da t-student é simétrica vem:

$$P\left(-2.977 \leq \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) - (\beta_0 + \beta_1 x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right)\hat{\sigma}^2}} \leq 2.977\right) = 0.99$$

Obtendo-se então o intervalo aleatório com 99% de confiança:

$$ICAleat._{99\%}\left(\beta_0+\beta_1 x_0\right)=\left(\left(\hat{\beta}_0+\hat{\beta}_1 x_0\right)\pm 2.977\sqrt{\left(\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\sum\limits_{i=1}^{n}x_i^2-n\bar{x}^2}\right)\hat{\sigma}^2}\right)$$

Concretização: O intervalo de confiança pretendido é dado por

$$IC_{99\%}\left(\beta_0+45\beta_1\right)=\left((148.05-1.0236\times 45)\pm 2.977\sqrt{\left(\frac{1}{16}+\frac{\left(\frac{967}{16}-45\right)^2}{1965.9375}\right)69.62}\right)$$

$$=(91.34,112.63)\,.$$

Para $x=20$ anos não é correcto utilizar o mesmo procedimento pois este valor não pertence ao $(\min(x_i),\max(x_i))$. Qualquer extrapolação para valores que não fizeram parte do ajuste do modelo será um abuso.