10.7 Sumário sobre a Inferência estatística no Modelo Regressão Linear Simples

Se você possui um conjunto de pares de valores

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) \tag{10.2}$$

e se você quer traçar uma reta que ajusta-se ao este conjunto pelo método chamado "mínimos quadrados", então a equação da reta ajustada será $y = \alpha + \beta x$, onde α e β são determinados pelos valores (10.2) via as seguintes fórmulas

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2}, \qquad \alpha = \overline{y} - \beta \overline{x} \qquad \left(\text{com } \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ e } \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} \right) \tag{10.3}$$

Recorde, que α chama-se *intercepto* e é igual à ordenada do ponto onde a reta $y = \alpha + \beta x$ corta o eixo y (quer dizer, o eixo das ordenadas); já β chama-se *coeficiente angular* e é igual ao tangente do ângulo entre a reta e o eixo x (quer dizer, o eixo de abcissas). Esse lembrete é importante se você for desenhar a reta ou interpretar o modelo – probabilístico ou determinístico – do qual adveiam os valores (10.2).

Recordo-lhe que *o resíduo* do ponto (x_i, y_i) em relação da reta $y = \gamma + \theta x$ (aqui, γ e θ são constantes quaisquer) é o valor de $y_i - (\gamma + \theta x_i)$. O conceito de resíduo dá-nós a seguinte maneira de caraterização da retá construída pelas fórmulas (10.3): α e β expressos em (10.3) correspondem àquela reta que apresenta o menor valor da soma dos quadrados de seus resíduos em relação aos pontos (10.2); é por isso, alias, que existe um nome alternativo para a reta ajustada que é *a reta de mínimos quadrados*. A propósito, eu não consegui entender por que o nome disse "mínimos quadrados" enquanto que o método procura pelo mínimo da soma dos quadrados; mas como meu disentendimento concerna a apelido só, então vamos deixar quieto.

Suponha agora que o conjunto (10.2) é uma amostra simples aleatória de um par de variáveis aleatórias (X, Y) ("simples aleatória" significa que cada (x_i, y_i) é uma realização deste par e que as realizações foram feitas de forma independente). E suponha também que X e Y satisfazem o modelo chamado *Regressão Linear Simples*, cujo significado no ámbito de nosso curso é assim:

$$Y = a + bX + \mathcal{E}, \tag{10.4}$$

onde a variável aleatória $\mathcal{E}$ é independente de X, tem esperança zero, quer dizer, $\mathbb{E}[\mathcal{E}] = 0$, e tem sua distribuição simêtrica em torno de 0. Neste caso – quer dizer, no caso quando a amostra foi gerada por X e Y relacionadas entre si pelo Modelo de Regressão Linerar Simples –, α e β dados por espressões (10.3) *estimam*, respectivamente, a e b da relação (10.4), ou, sendo dito da maneira mais formal, α e β são *as estimativas dos parâmetros a e b do Modelo de Regressão Linear Simples obtidas a partir da amostra (10.2) pelo método de mínimos quadrados.*

O uso da amostra que adveio de X e de Y para o fim de estimação de a e de b por α e β faz sentido quando as variáveis aleatórias X e Y de fato estão relacionadas segundo ao modelo (10.4). A confirmação/rejeição da presença desse relacionamento faz-se com base na amostra (10.2) e com o auxílio do coeficiente denotado por $r_{x,y}$ e calculado por uma das fórmulas apresenatadas abaixo:

$$\begin{aligned} r_{x,y} &= \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2) \cdot (\sum_{i=1}^{n} y_i^2 - n(\overline{y})^2)}} \end{aligned} \tag{10.5}$$

(você pode usar qualquer uma das duas expressões; geralmente, a segunda delas é a mais usada pois ela despensa o cálculo intermediário de $x_i - \overline{x}$ e de $y_i - \overline{y}$).

O procedimento que emprega $r_{x,y}$ no teste da relação (10.4) baséia-se em dois pilares. O primeiro deles é o fato que afirma que $r_{x,y}$ é uma estimativa para $\varrho(X, Y)$, coeficiente de corelação linear entre X e Y, enquanto que o segundo pilar está construído de algumas propriedades de $\varrho(X, Y)$. Antes de falar sobre tais propriedades, vale eu notar que o fato supramencionado sobre a estimação de $\varrho(X, Y)$ por $r_{x,y}$ motivou o nome para $r_{x,y}$; o nome é *o coeficiente de correlação linear amostral*. Observo que o termo "amostral" assinala a relação de $r_{x,y}$ com a amostra que adveio da variáveis aleatórias X e Y. E já que amostras denotam-se por letras minúsculas, então o índice de $r_{x,y}$ escreve-se em letras minúsculas. No contrário dessa notação, a coeficiente $\varrho(X, Y)$ usa letras maiúsculas o que é natural pois esse coeficiente concerna as variáveis aleatórias . Vale ainda eu mencionar que as vezes acrescenta-se a palavra "*de Pearson*" tanto ao nome de $r_{x,y}$ quanto ao de $\varrho(X, Y)$. Tal acrescimo deve-se ao fato que foi Karl Pearson (1857–1936) quem destacou esses coeficientes do grande grupo de diversas medidas de dependência, e estudou-os.

Conforme dito acima, $\varrho(X, Y)$ desempenha um papel importante no teste estatístico sobre a validade da relação linear entre duas variáveis aleatórias . Por isso, é importante conhecer as fórmulas de cálculo de $\varrho(X, Y)$: O passo intermediário é o cálculo da *covariância* entre X e Y:

$$\mathrm{Cov}(X, Y) = I\!E\left\{(X - I\!E[X]) \times (Y - I\!E[Y])\right\} = I\!E\left[X \times Y\right] - I\!E[X] \times I\!E[Y] \tag{10.6}$$

Você pode usar qualquer uma das expressões, mas parece-me que a segunda delas é menos trabalhosa computacionalmente. Compensa detalhar ela: os valores de $I\!E[X]$ e de $I\!E[Y]$ calculam-se a partir das distribuições marginais de X e de Y, enquanto que

$$I\!E\left[X \times Y\right] = \sum_{\substack{i=1,\ldots,k \\ j=1,\ldots,\ell}} p(x_i, y_j) x_i y_j \tag{10.7}$$

onde $x_1, \ldots, x_k$ são todos os valores que X pode assumir, k em número, onde $y_1, \ldots, y_\ell$ são todos os valores que Y pode assumir, ℓ, em número, e onde $p(x_i, y_j) = I\!P[X = x_i,\, Y = y_j]$ (não é para confundir os valores de X e Y com amostra de (X, Y)). E, finalmente,

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[x] \times \mathrm{Var}[Y]}} \tag{10.8}$$

Prosseguimos com a lista das propriedades de $\varrho(X, Y)$ que nós interessam na perspectiva da construção e execução do teste da presença do Modelo de Regressão Linear Simples:

(i) $\varrho(X, Y) = 1$ se e somente se $Y = a + bX$ e $b > 0$; e $\varrho(X, Y) = -1$ se e somente se $Y = a + bX$ e $b < 0$; observe que em ambos os casos o ruído está ausente;

(ii) se X e Y satisfazem ao Modelo de Regressão Linear Simples (10.4) então $\varrho(X, Y)$ esta estritamente entre -1 e 1; quanto maior o ruido $\mathcal{E}$, mais longe de 1 está $|\varrho(X, Y)|$ (a grandeza do ruido mede-se por $\mathrm{Var}[\mathcal{E}]$);

(iii) se X e Y são variáveis aleatórias independentes, então $\varrho(X, Y) = 0$; a recíproca, porém, não é válida: se $\varrho(X, Y) = 0$ então X e Y podem ser dependentes;

(iv) seja θ um valor fixado antemão que está estritamente entre -1 e 1; então existe variáveis aleatórias V e W e uma relação determinística (tipo $W = V^2$) tais que $\varrho(V, W) = \theta$;

(v) seja θ um valor fixado antemão que está estritamente entre -1 e 1; então existe variáveis aleatórias V e W e uma relação determinística com ruído (tipo $W = V^2 + \mathcal{E}$) tais que $\varrho(V, W) = \theta$.

Tudo que formulamos até o momento, pondo junto, justifica as regras (a)–(d) abaixo que regulamentam o uso de $r_{x,y}$ na identificação da dependência via Regressão Linear Simples, assim como o emprego dos coeficientes α e β. Antes de formular as regras, é precisa avisar que $|r_{x,y}|$ não pode ser maior que 1 para qualquer que seja conjunto (10.2).

(a) Se $r_{x,y} = 1$ então todos os pontos de (10.2) "deitam" numa reta cujo coeficiente angular é positivo; se $r_{x,y} = 1$ então todos os pontos de (10.2) "deitam" numa reta cujo coeficiente angular é negativo; tais casos ocorrem raramente e por isso aquí não nós interessam.

(b) Se $r_{x,y}$ está próximo ao 1 ou ao -1 então pode-se acreditar que a amostra (10.2) adveio de X e Y que satisfazem ao Modelo de Regressão Linear Simples (10.4) e que o ruído $\mathcal{E}$ desse modelo é pequeno; ainda mais, o sinal de $r_{x,y}$ coincide com o sinal de β da reta dos Mínimos Quadrados. Infelizmente, os itens (iv) e (v) da lista de propriedades de $\varrho(X, Y)$ mostram que além da crença supraformulada podem existir outras crenças que sugerem que a dependência entre X e Y pode ser não linear. Existem métodos estatísticos que permitem identificar a verdadeira dependência com bom grau de precisão, mas tudo isso fica fora do âmbito do presente texto. Portanto, no texto será aceita a crença que aponta ao Modelo de Regessão Linear Simples. Essa aceitação faz sentido também do ponto de vista prática pois a dependência linear é a aproximação de primeira ordem para qualquer dependência mais complexa, e essa aproximação já é capaz de mostrar a tendência da dependência, isto é, mostrar se os valores de Y aumentam-se com o auamento dos valores de X, ou se, pelo contrário, diminuem-se Em muitos estudos práticos, o descobrimento da tendência já é um resultado significativo e útil.

(c) Se $r_{x,y}$ está próximo ao zero, há indicação de que X e Y são independentes. Mas, de acordo com (iii), é só uma indicação, e para confirmar a independência é precisa usar métodos específicos que não estão ensinados nesse curso.

(d) Faltou falar sobre o caso que complementa aos casos (a), (b), (c), isto é, o caso quando $r_{x,y}$ não está nem perto de zero, nem perto de 1 nem de -1. Qualquer de tais valores de $r_{x,y}$ pode ser gerado por qualquer uma das três seguintes relações entre X e Y: *(i)* X e Y satisfazem ao Modelo de Regressão Linear Simples (10.4) mas o ruído $\mathcal{E}$ é grande, *(ii)* X e Y estão em relação não linear e com ruído grande, *(iii)* X e Y estão em relação não linear bem complexa e com ruído possivelmente pequeno. O estudo da terceira das relações exige um conhecimento da Estatística que vai muito além dos limites do presente curso. A segunda relação também não será estudada em detalhes em nosso curso. Então, já que a única ferramenta da área ensinada no presente texto é o ajuste da reta de mínimos quadrados, então a dúvida é se ela é aplicável no presente caso do valor de $r_{x,y}$. A resposta é clara: A reta de mínimos quadrados existe para qualquer conjunto de pontos (10.2), já que para qualquer conjunto de tais pontos é possível calcular os valores de α e de β. A questão principal é o que essa reta nós diz sobre X e Y. A resposta também é clara: Se X e Y satisfazem ao Modelo de Regressão Linear Simples então a e b do modelo são estimados por, respectivamente, α e β. Entretanto, o valor de $r_{x,y}$ mostra que o ruído da relação linear entre X e Y é grande e isso faz com que a precisão da estimação pode ser muito, muito ruim. Já se X e Y não satisfazem ao Modelo de Regressão Linear Simples então a reta ajustada $y = \alpha + \beta x$ pode ser interpretada como a estimativa – possivlemente não muito precisa – para a reta $y = a + bx$ que expressa a tendência principal da relação entre X e Y.

Falta só eu comentar sobre a quantificação do conceito "$|r_{x,y}|$ está perto/longe de 1". Aqui não há uma regra aceita por todos. Consequentemente, os exercícios e as provas do curso não tocam nesse assunto; suas questões pedirão simplesmente calcular $r_{x,y}$ e, em seguida, ajustar a reta de mínimos quadrados; as vezes, pedira-se-á também interpretar os coeficientes α e β dessa reta. Quanto à quantificação, eu me lembro da frase "as experiências dos estatísticos mostram que o valor de $|r_{x,y}|$ maior que $0,8$ já pode ser considerado próximo ao 1". Entretanto, na minha experiência, eu vi que valores de $|r_{x,y}|$ na faixa de $0,6$ já são acompanhados por uma boa estimativa dos coeficientes a e b por, respectivamente, α eβ. Em relação dessa discussão, é curioso mencionar que Dr. Jordan Peterson[1] ao falar de conclusões baseadas em dados da área de sociologia, disse que o valor $0,2$ para $|r_{x,y}|$ já é suficiente para afirmar que as variáveis em estudo satisfazem ao Modelo de Regressão Linear Simples é consequentemente aproveitar do modelo para deduzir conclusões. Ainda mais, segundo as palavras de Peterson, quando $|r_{x,y}| = 0,5$, já tem-se uma "tremenda prova" de associação entre as variáveis. A citação da fala de Peterson indica que nossa expectativa acerca da capacidade indicativa de $|r_{x,y}|$ depende muito da magnitude do erro envolvido nos dados estudados. Na área da Sociologia, as "medições" são propensos ao ruído muito grande, e é por isso que os requisitos sobre $|r_{x,y}|$ foram relaxados.

[1]Dr. Jordan B. Peterson is a professor of psychology at the University of Toronto, a clinical psychologist and the author of the multi-million copy bestseller "12 Rules for Life: An Antidote to Chaos".