Regressão Linear

Algumas vezes estamos interessados não apenas se existe associação entre duas variáveis quantitativas x e y, mas nós temos também uma hipótese a respeito de uma provável relação de causa e efeito entre variáveis. Desejamos saber se y "depende" de x. Neste caso, y é chamado de variável dependente ou variável resposta e x é chamado de variável independente ou explanatória que, na linguagem epidemiológica, é denominada "fator de risco". Na forma de regressão mais comumente utilizada, a regressão linear, temos a hipótese de que o valor de y depende do valor de x e expressamos matematicamente esta relação por meio de uma equação, assumindo que a associação entre x e y é linear, ou seja, descrita adequadamente por uma reta. Quando temos uma variável resposta y e uma variável explanatória x a regressão é dita simples. Quando temos uma variável resposta y e mais de uma variável explanatória, x1, x2, x3... a regressão é chamada múltipla.

A regressão é usada basicamente com duas finalidades: de previsão (prever o valor de y a partir do valor de x) e estimar o quanto x influencia ou modifica y.

Vejamos o exemplo abaixo. No diagrama de dispersão vemos que, à medida em que aumenta a porcentagem de crianças imunizadas contra DPT (difteria, coqueluche e tétano) em amostra de 20 países do mundo em 1992 diminui a taxa de mortalidade infantil de crianças menores de 5 anos. Esta relação pode ser descrita razoavelmente por uma reta. Temos a hipótese que a percentagem de imunização contra DPT pode influenciar a mortalidade infantil, mas desejamos medir esta associação, que pode ser descrita com a fórmula:

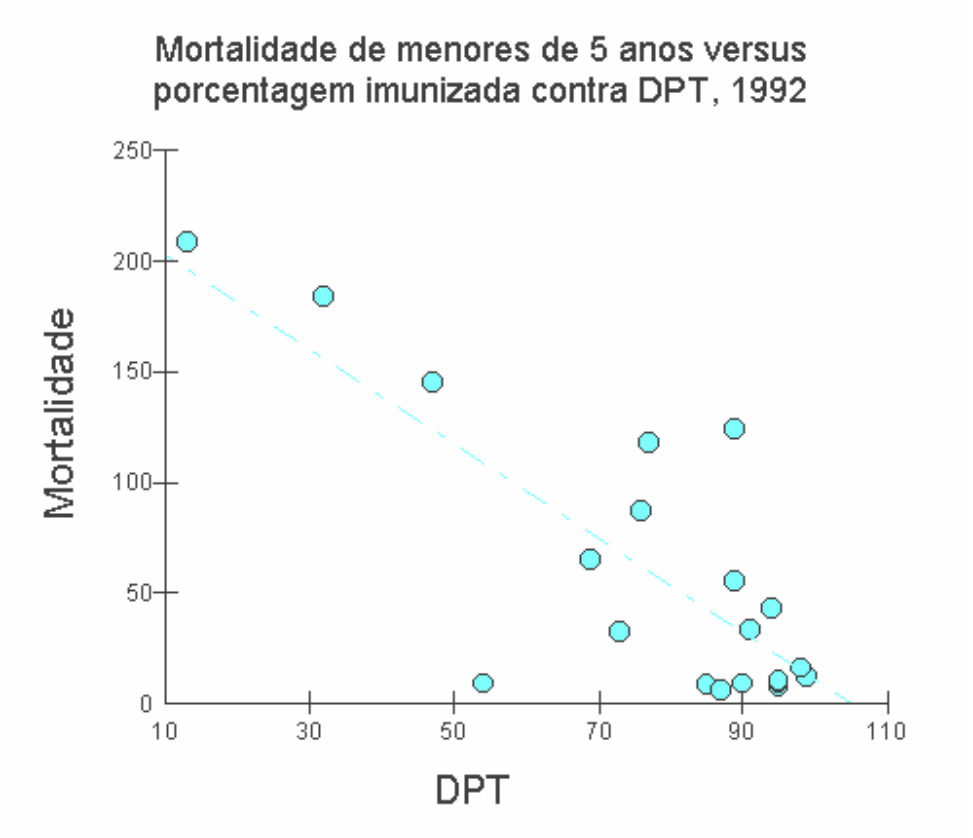
$Y = a + b\,x$

a= coeficiente linear (também chamado intercepto, é o valor que y assume quando x for zero)
b= coeficiente angular (é a inclinação da reta, mede o aumento ou redução em y para cada aumento de uma unidade em x).

Tabela 1. Porcentagem de crianças imunizadas contra DPT e taxa de mortalidade de menores de 5 anos para 20 países, 1992.

País	Porcentagem imunizada	Taxa de mortalidade por 1000 nascidos vivos
pais	**dpt**	**mort**
Bolivia	77	118
Brasil	69	65
Camboja	32	184
Canada	85	8
China	94	43
Republica Tcheca	99	12
Egito	89	55
Etiopia	13	208
Finlandia	95	7
Franca	95	9

Grecia	54	9
India	89	124
Italia	95	10
Japao	87	6
Mexico	91	33
Polonia	98	16
Federacao Russa	73	32
Senegal	47	145
Turquia	76	87
Reino Unido	90	9



Vamos analisar os cálculos abaixo realizados no Stata, com o comando abaixo:

regress mort dpt

```
    Source |       SS       df       MS          Number of obs =      20
-------------+------------------------------       F(  1,    18) =   30.10
     Model |  48497.0497     1  48497.0497        Prob > F      =  0.0000
  Residual |  29000.9503    18  1611.16391        R-squared     =  0.6258
-------------+------------------------------       Adj R-squared =  0.6050
     Total |      77498    19  4078.84211        Root MSE      =  40.139

------------------------------------------------------------------------------
      mort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dpt |  -2.135869   .3893022    -5.49   0.000    -2.953763   -1.317976
     _cons |   224.3163   31.44034     7.13   0.000     158.2626     290.37
------------------------------------------------------------------------------
```

O intercepto (a) deu o valor 224 e o coeficiente de regressão (b) produziu –2,14. A equação então ficou:

Y= a + bx
Y= 224 + (-2,14) x

A regressão é usada para previsão. Supondo que um determinado país tenha porcentagem de imunização contra DPT de 80% qual seria a sua mortalidade infantil esperada? Seria 52,8, conforme cálculo realizado abaixo.

Y= 224 –2,14 . 80
Y= 52,8

Outras perguntas que são respondidas pela regressão:

1) O quanto a variação de x influencia na variação de y?

Respondemos a esta pergunta usando o coeficiente b. Para cada variação de uma unidade em x (porcentagem de imunização por DPT) a taxa de mortalidade infantil em menores de cinco anos cai 2,14.

2) Qual a probabilidade desta redução da taxa de mortalidade em menores de cinco anos associada à imunização ser explicada pelo acaso?

Esta pergunta é respondida realizando-se um teste t para testar se o coeficiente angular (b) é diferente de zero. Se ele for zero a reta não tem inclinação alguma, então x não interfere em y. Neste exemplo o teste t resultou –5,49 e o valor de P é extremamente baixo (o programa fornece p=0.0000, ou seja, bem próximo de zero). Neste caso dizemos que o acaso é uma explicação pouco provável para este fenômeno.

3) Qual o percentual de variação de y explicado pela variação de x?

Esta resposta é dada pelo coeficiente de determinação. Neste exemplo, 63% da variação de y é explicado pela variação de x.

Agora que nós já vimos resumidamente por que se usa uma regressão e demos uma olhada em um exemplo, vamos ver como se faz os cálculos.

O método mais usado para estimar os parâmetros A e B é o método dos mínimos quadrados. Este método garante que a reta obtida é aquela para a qual se tem as menores distâncias (ao quadrado) entre os valores observados de y e a própria reta.

O coeficiente angular é estimado pela fórmula:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

O intercepto é estimado pela fórmula:

$$a = \bar{y} - b\bar{x}$$

Pressupostos para uso da regressão linear:

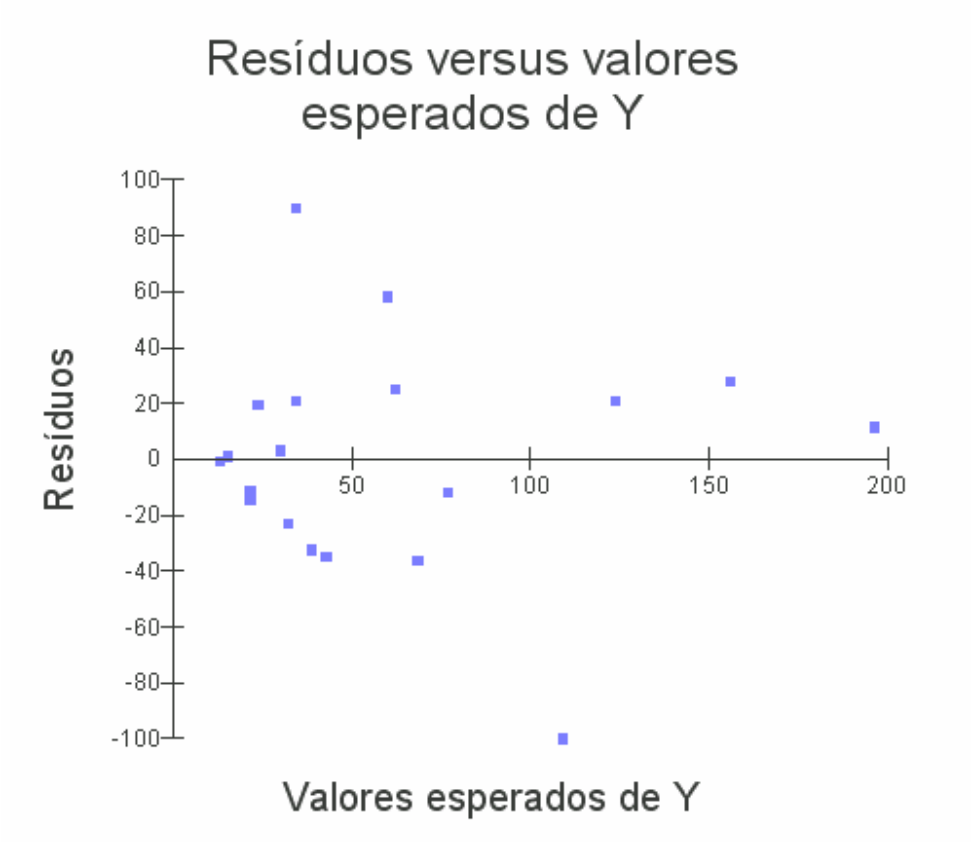
1) A variável y deve ter distribuição normal ou aproximadamente normal. Se a distribuição não for normal pode-se realizar uma transformação.
2) A variação de x deve ser a mesma para cada valor de y (homocedasticidade). Se não houver homocedasticidade é necessário transformar os dados.
3) Os pontos no diagrama de dispersão devem apresentar tendência linear. Se a relação for expressa por uma curva pode-se transformar os dados para tentar linearizar a associação ou então usa-se outra forma de regressão não linear.
4) Os valores de y foram obtidos ao acaso da população e são independentes uns dos outros
5) A variável x foi medida sem erro.

Análise de resíduos:

É importante, após se realizar a regressão, testar se os pressupostos acima se aplicam ao nosso caso. Isto se faz com a análise dos resíduos. Resíduos representam a diferença entre o valor observado de y e o que foi predito pelo modelo de regressão.

$$e_i = y_i - \hat{y}_i$$

A primeira forma de se avaliar resíduos é plotar um gráfico no qual os resíduos ($y - \hat{y}$) são colocados no eixo vertical (y) e os valores esperados de y ($\hat{y}$) no eixo horizontal (x).



Os pontos devem ficar distribuídos de forma equilibrada acima e abaixo da linha que passe no ponto de resíduo 0, formando uma nuvem retangular de pontos. Quando não há homocedasticidade (que é o caso acima), observa-se uma nuvem em forma de cone. A dispersão dos valores é maior na primeira parte da distribuição. Quando a relação não for linear, observa-se uma nuvem curva.

Neste caso, porém existem valores atípicos, Grécia, que tem uma baixa cobertura de DPT e uma mortalidade infantil baixa e Índia que tem alta cobertura de DPT e uma mortalidade alta. Pode ser que a retirada de pontos extremos, com resíduos altos melhore a homocedasticidade. Entretanto só se deve retirar pontos extremos com uma boa justificativa (erro de leitura ou anotação dos dados, problemas durante a realização do experimento). Se eles realmente fazem parte da realidade é melhor tentar uma transformação, pois a eliminação do ponto vai distorcer a análise do fenômeno.

Para realizar estes cálculos no Stata, digite:

```
predict morte

gen res=mort-morte
```

gen str5 letra = substr(pais,1,5)

twoway scatter res morte, mlabel(letra)

O gráfico dos resíduos versus cada variável explanatória também é muito elucidativo para testar os pressupostos do modelo. A presença de uma relação curvilinear, por exemplo, sugere que a adição de um termo quadrático à variável explanatória deve ser adicionado ao modelo.

rvpplot dpt, mlabel(letra)

O gráfico de probabilidade normal dos resíduos também é muito útil. Depois que toda a variável sistemática for removida do modelo, os resíduos devem ter distribuição normal.

pnorm res

EXERCÍCIOS

Medidas de comprimento (em cm) e de peso (em gramas) de uma amostra de 20 bebês nascidos com baixo peso estão na tabela abaixo:

Comprimento	Peso
comp	**peso**
41	1360
40	1490
38	1490
38	1180
38	1200
32	680
33	620
38	1060
34	830
32	880
39	1130
38	1140
39	1350
37	950
39	1220
38	980
42	1480
39	1250
38	1250
30	1320

1) Primeiro digite os dados acima no Stata.

2) Verifique se as variáveis têm distribuição normal ou se há valores extremos. Plote o histograma, o Box-plot e o gráfico da probabilidade normal para cada variável. Explore o menu Graphics do Stata.

Graphics / Histogram / Variable: peso / OK
Graphics / Box Plot / Variable: comp / OK
Graphics / Distributional Graphs / Normal quantile plot / Variable: peso / OK

Alternativamente você poderia digitar:

histogram peso

graph box peso, medtype(line)

qnorm peso

3) Construa um gráfico de dispersão bidimensional do peso (x) versus o comprimento (y) e avalie se esta relação pode ser descrita por uma reta.

   Graphics / Twoway graphs / Type: scatter X: peso Y: comp / OK

   Ou

   twoway (scatter comp peso)

4) Há alguma evidência de uma relação linear entre as variáveis? Há algum ponto extremo?

5) É possível, a partir do conhecimento do peso do recém-nascido prever o seu comprimento? Usando o comprimento como variável resposta e o peso como variável explicativa, faça os cálculos da regressão linear.

   regress comp peso

6) Quais os valores obtidos para o intercepto e para o coeficiente angular? Ao nível de significância de 0.05, teste a hipótese nula de que a verdadeira inclinação da reta (b) é igual a 0. O que você conclui?

7) Qual o comprimento estimado pelo modelo para um bebê que pesou 1320 gramas? Qual o resíduo neste caso (a diferença entre o comprimento observado, no caso 30 e o comprimento estimado pelo modelo)?

8) O modelo de regressão de mínimos quadrados parece se ajustar aos dados observados? Comente os coeficientes de determinação e o gráfico dos resíduos versus os valores ajustados do comprimento, o gráfico dos resíduos versus a variável explanatória e o gráfico da probabilidade normal dos resíduos.

rvfplot

rvpplot peso

pnorm res

9) Apague o ponto extremo e refaça todos os cálculos. O que se alterou quando você removeu o ponto atípico do conjunto de dados?