Modelos de Regressão Linear Simples - parte I

Erica Castilho Rodrigues

27 de Setembro de 2017

Introdução

Coeficiente de Correlação Linear

Objetivos

Ao final deste capítulo você deve ser capaz de:

- Usar modelos de regressão para construir modelos para dados coletados.
- Entender como método de mínimos é usado para estimar parâmetros desconhecidos.

Introdução

- Podemos estar interessados em explorar a relação entre duas ou mais variáveis.
- Essa técnica é chamada **Análise de Regressão.**
- Exemplo: qual relação entre nível de escolaridade e renda?

- A ferrementa inicial para sabermos se existe relação entre as variáveis é o **gráfico de dispersão.**
- A correlação poderá:
  - não existir;
  - ser uma correlação linear (ao longo de uma reta);
  - ser uma correlação não linear (ao longo de uma curva).

Correlação

Há um relacionamento entre as variáveis?

- Elas aumentam juntas?
- Aumentando uma variável a outra aumenta ou diminui?
- Exemplo: nota na prova e horas de estudo.
- Variam juntas.
- Se uma aumenta, a outra também aumenta.

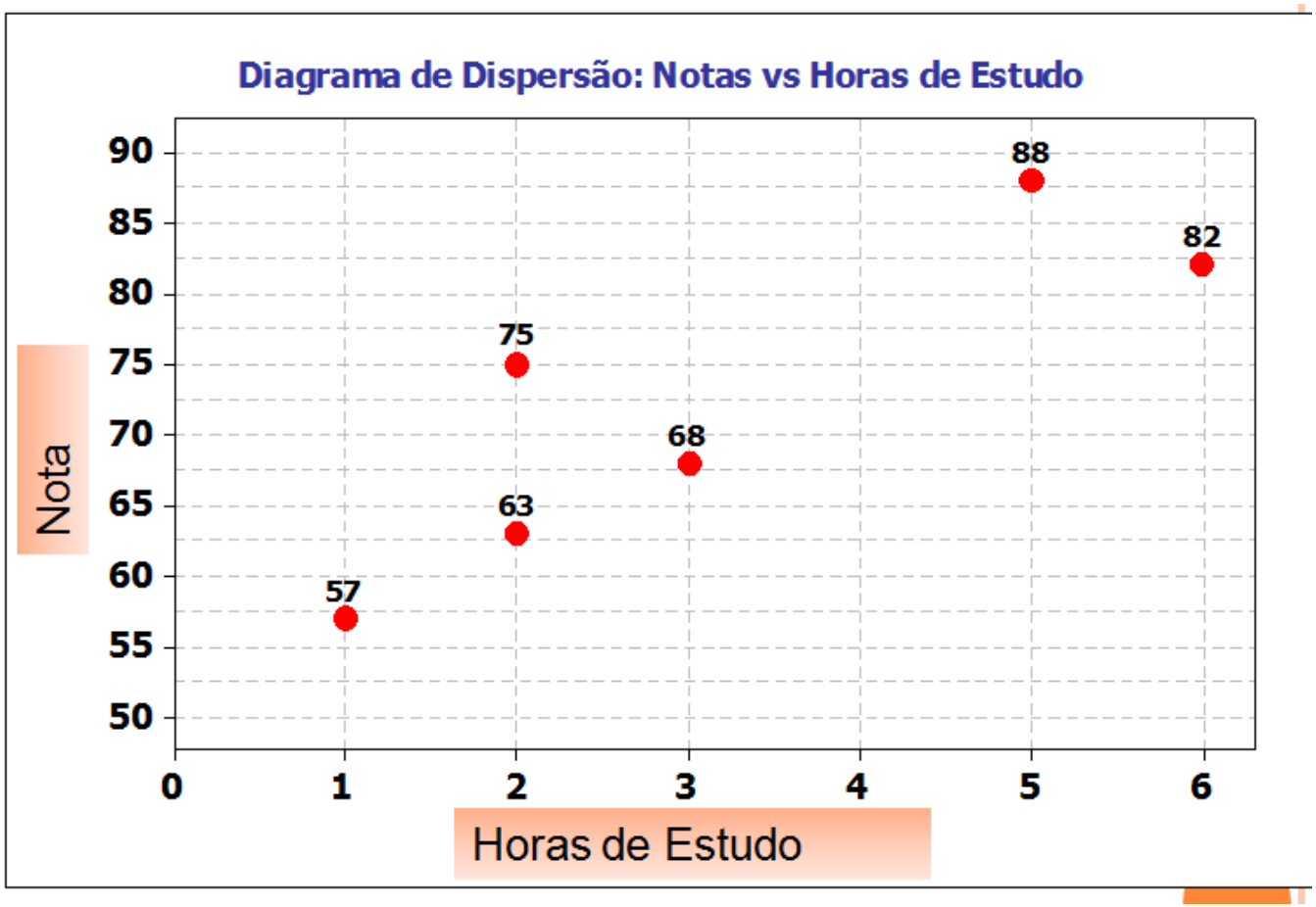
Exemplo:

- Vamos considerar as variáveis nota na prova e horas de estudo.
- Y - nota na prova (variável resposta).
- x - horas de estudo (variável explicativa ou preditora).
- Os dados são os seguintes

Aluno	Horas de Estudo	Nota na Prova
A	6	82
B	2	63
C	1	57
D	5	88
E	3	68
F	2	75

Exemplo: (continuação)

- O gráfico de dispersão é mostrado a seguir



- Parece existir uma correlação positiva (uma aumenta a outra aumenta).

Exemplo:

- 50 municípios de um estado são analisados.
- Deseja-se verificar se existe relação entre duas variáveis:
  - nível de pobreza da população;
  - taxa de roubos e furtos do município.
- Os dados coletados são apresentados a seguir.

Município	Índice de pobreza	Taxa de roubos e furtos
1	0,520243	4,6796
2	0,573892	5,8444
3	0,588446	6,5371
4	0,610826	5,8645
5	0,626467	6,5505
6	0,634336	7,3872
7	0,640390	6,0816
8	0,644050	6,569
9	0,663735	5,6804
10	0,676159	7,2103
11	0,692946	5,7382
12	0,719850	6,3769
13	0,721679	7,7359
14	0,764735	8,6539
15	0,781271	8,0238
16	0,786335	7,3774
17	0,797273	8,5611
18	0,806268	7,9516
19	0,811569	7,0226
20	0,821166	8,7187
21	0,821884	7,8224
22	0,828157	8,2076
23	0,845534	8,3666
24	0,857904	9,4827
25	0,858685	8,8655

Município	Índice de pobreza	Taxa de roubos e furtos
26	0,861767	8,0104
27	0,863136	7,644
28	0,881571	8,3115
29	0,886649	8,9649
30	0,895995	9,6845
31	0,909950	8,3012
32	0,921256	9,7942
33	0,924994	10,3435
34	0,927112	9,202
35	0,934594	8,5556
36	0,944849	10,3021
37	0,952311	9,909
38	0,954732	10,7073
39	0,956666	9,2998
40	0,958970	8,8892
41	0,958975	9,7285
42	0,965260	9,0648
43	0,968608	10,4605
44	0,969550	9,9478
45	0,980406	8,8263
46	0,982279	9,5566
47	0,982572	10,8394
48	0,983857	9,2011
49	0,994508	9,2753
50	0,995053	10,0008

Exemplo: (continuação)

Pergunta

Existe associação entre níveis de pobreza e taxas de roubos e furtos do município?

Ou seja...

O nível de pobreza de um município determina (explica, prediz, interfere no) nível de criminalidade (roubo/furto) do município?

Hipótese

Quanto maior os níveis de pobreza de um município, maior a taxa de criminalidade.

Exemplo: (continuação)

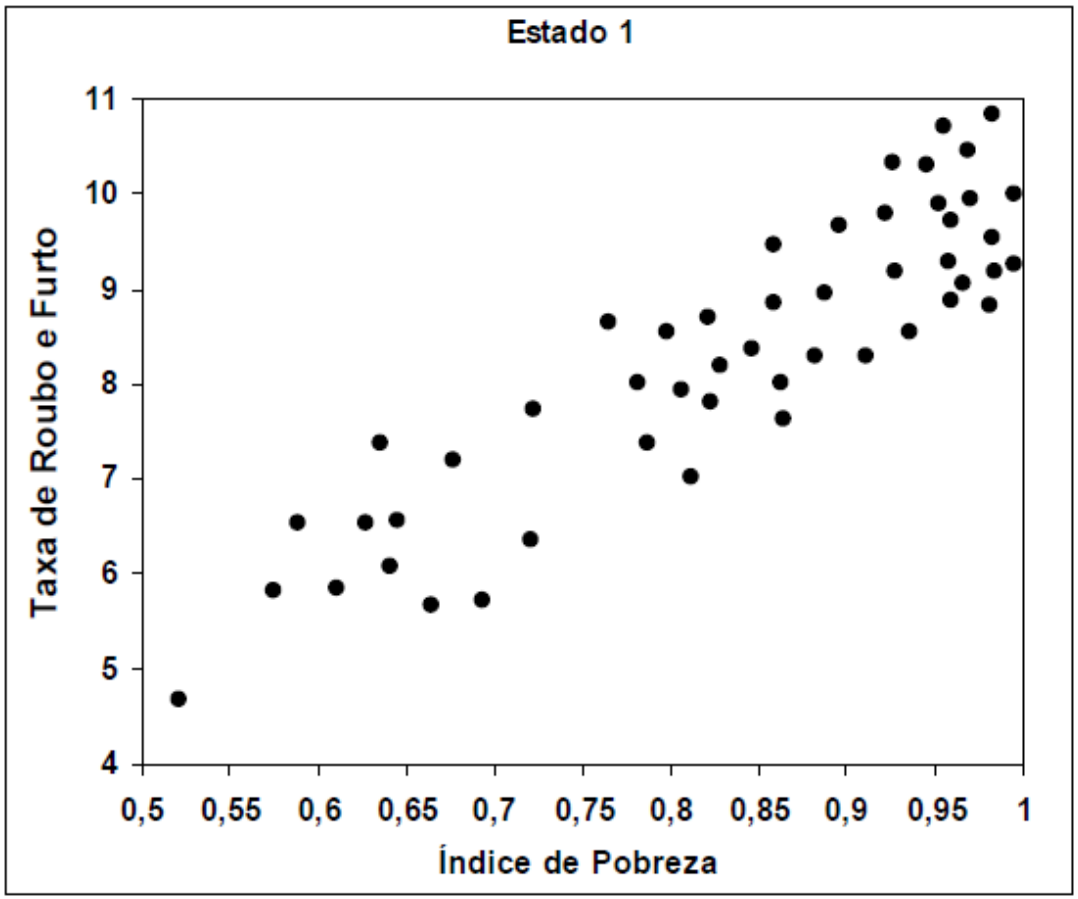
- **Variável Resposta ou Dependente:**

Exemplo: (continuação)

- **Variável Resposta ou Dependente:** criminalidade.
- **Variável Explicativa ou Independente:**

Exemplo: (continuação)

- **Variável Resposta ou Dependente:** criminalidade.
- **Variável Explicativa ou Independente:** pobreza.
- Gráfico de dispersão é mostrado a seguir:

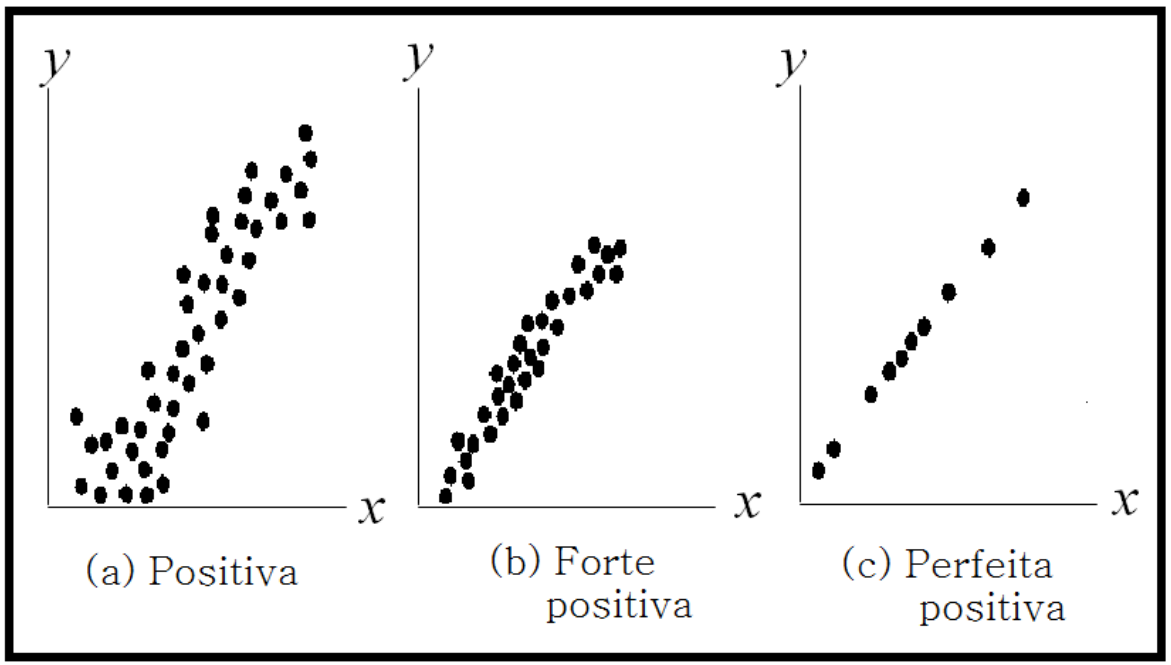


- Qual a conclusão que você tira?

- Qual a conclusão que você tira?
- Parece existir uma associação positiva entre as variáveis.
- Vamos ver como podemos medir se essa associação é fraca ou forte.

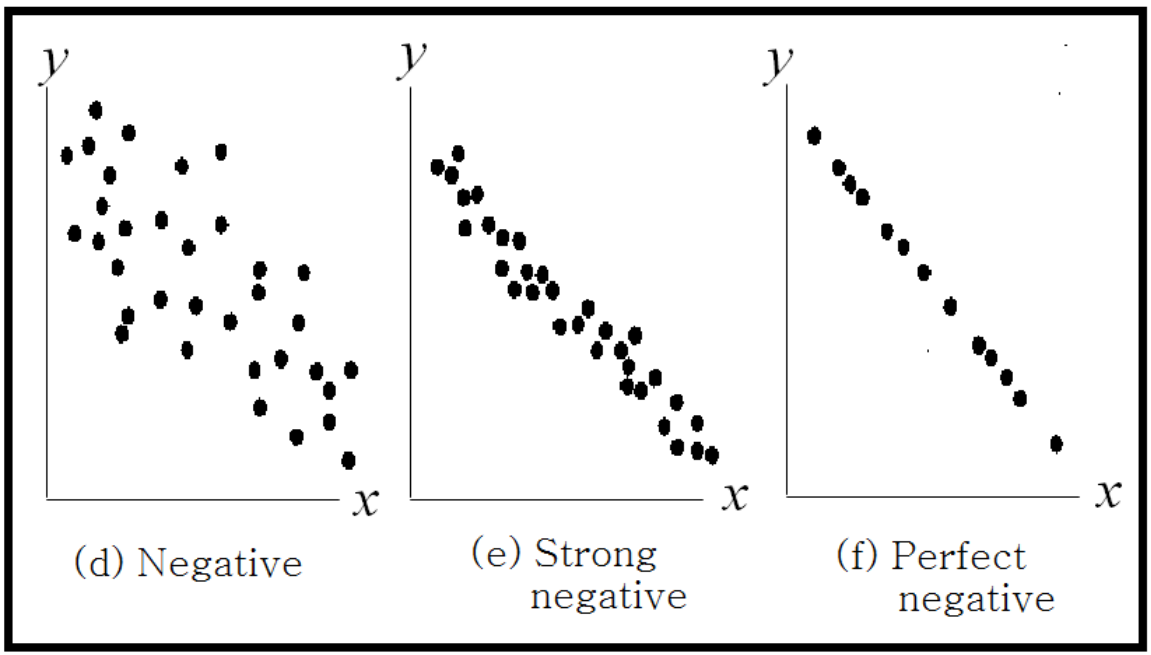
Correlação linear positiva

Uma variável aumenta a outra também aumenta.



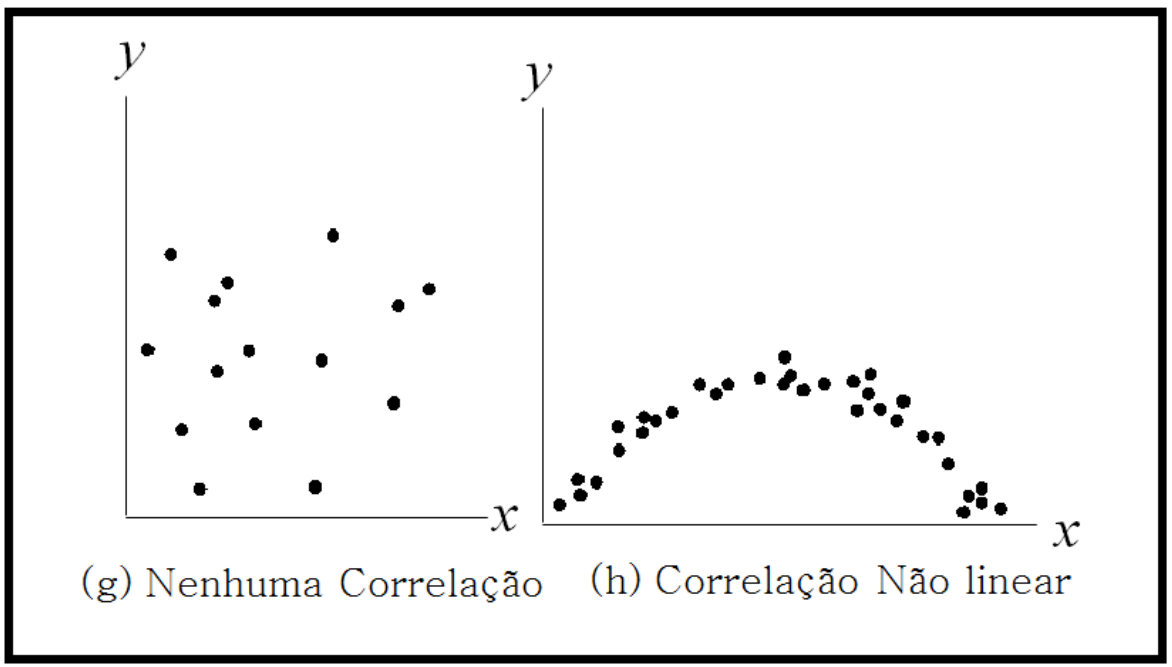
Correlação linear negativa

Uma variável aumenta a outra diminui.



Correlação não linear

Não existe relação linear entre as variáveis.



(g) Nenhuma Correlação (h) Correlação Não linear

Coeficiente de correlação linear

Mede a força da **relação linear** entre duas variáveis.

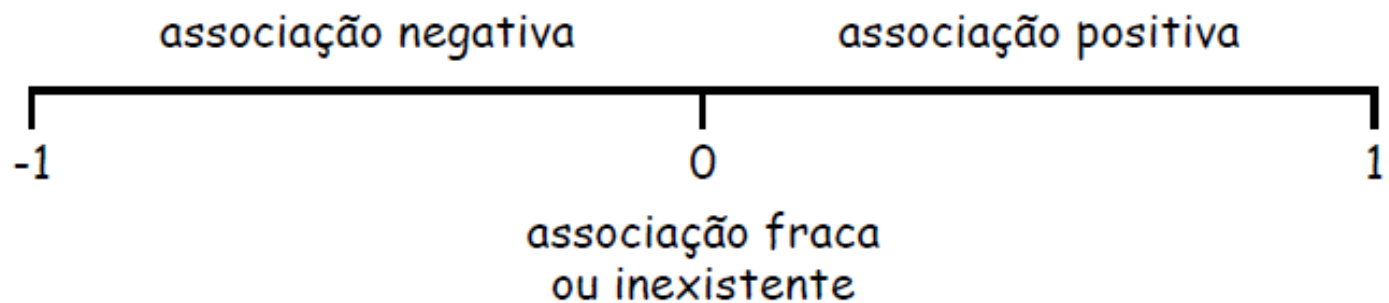
- Denotado por r se for amostral e ρ se for populacional.
- Mede o grau de associação linear entre duas variáveis.
- Indica também se essa associação é positiva ou negativa.

- É dado por

$$r = \frac{n\sum_i x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum_i x_i^2) - (\sum_i x_i)^2}\sqrt{n(\sum_i y_i^2) - (\sum_i y_i)^2}}$$

Propriedades do coeficiente de correlação linear

- $-1 \leq r \leq 1$
- Se $r \approx 1 \Rightarrow$ correlação forte positiva.
- Se $r \approx -1 \Rightarrow$ correlação forte negativa.
- Se $r \approx 0 \Rightarrow$ não existe correlação **linear**.
- O valor de r não é influenciado pelas escalas de x e y.



- A significância dos valores de *r* depende muito da área em que estamos trabalhando.
- Em algumas áreas, não se espera que uma variável explique bem a outra.
- Assim não se espera valores muito altos para *r*.
- A tabela a seguir apresenta um guia geral:

Valor de r	**Intensidade da associação**	Valor de r
$-0,2 < r \leq 0$	inexistente	$0 \leq r < 0,2$
$-0,4 < r \leq -0,2$	muito fraca	$0,2 \leq r < 0,4$
$-0,7 < r \leq -0,4$	fraca/moderada	$0,4 \leq r < 0,7$
$-0,9 < r \leq -0,7$	forte	$0,7 \leq r < 0,9$
$-1 < r \leq -0,9$	muito forte	$0,9 \leq r < 1$
$r = -1$	perfeita	$r = 1$

Exemplo

- Vamos retomar o exemplo dos níveis de pobreza e criminalidade.

Pergunta

Existe associação entre níveis de pobreza e taxas de roubos e furtos do município?

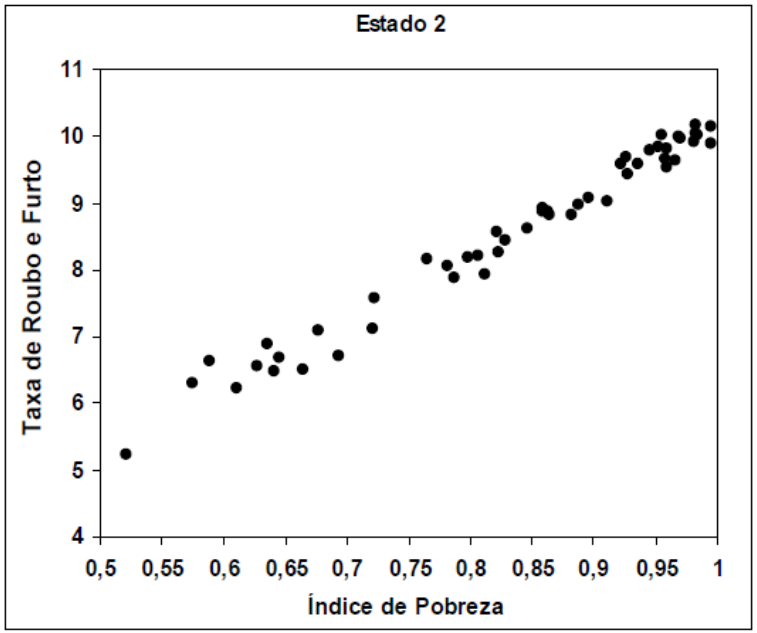
Ou seja...

O nível de pobreza de um município determina (explica, prediz, interfere no) nível de criminalidade (roubo/furto) do município?

Hipótese

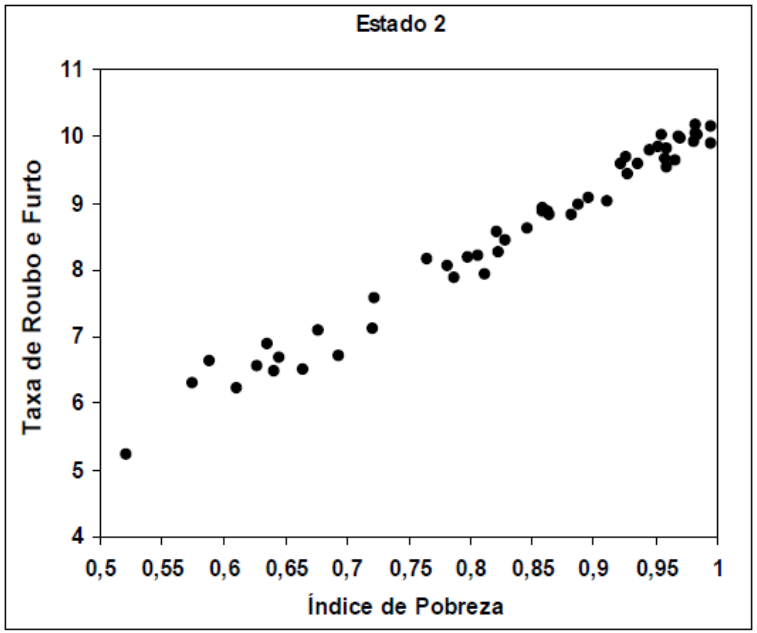
Quanto maior os níveis de pobreza de um município, maior a taxa de criminalidade.

Exemplo (continuação)



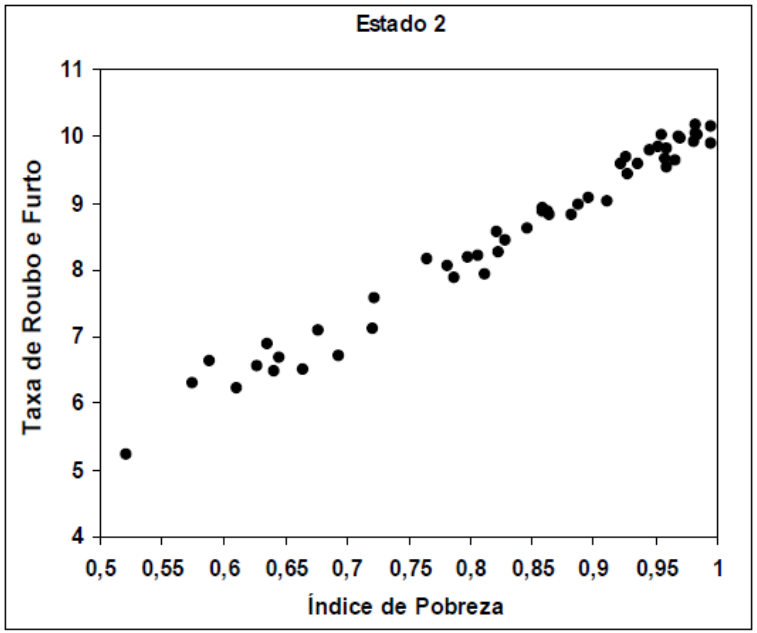
- Existe associação linear entre as variáveis?

Exemplo (continuação)



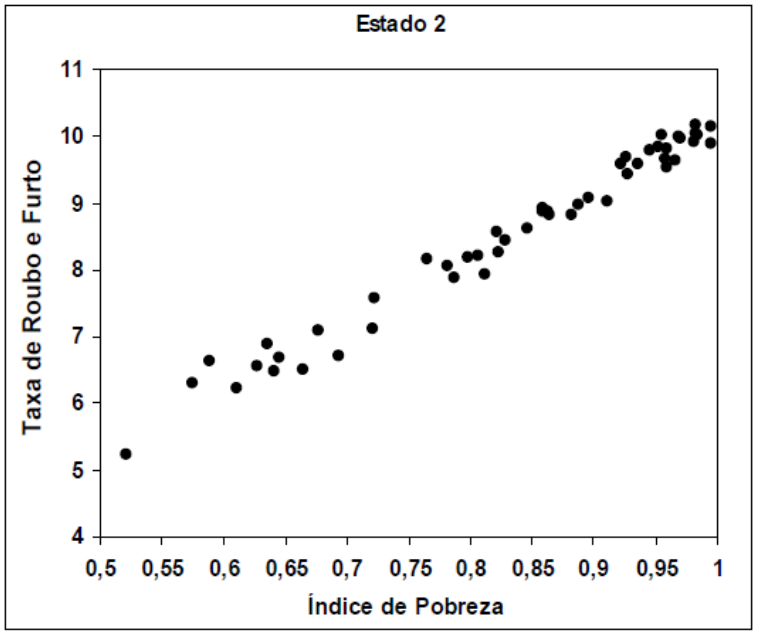
- Existe associação linear entre as variáveis? Sim.
- Ela é forte ou fraca?

Exemplo (continuação)



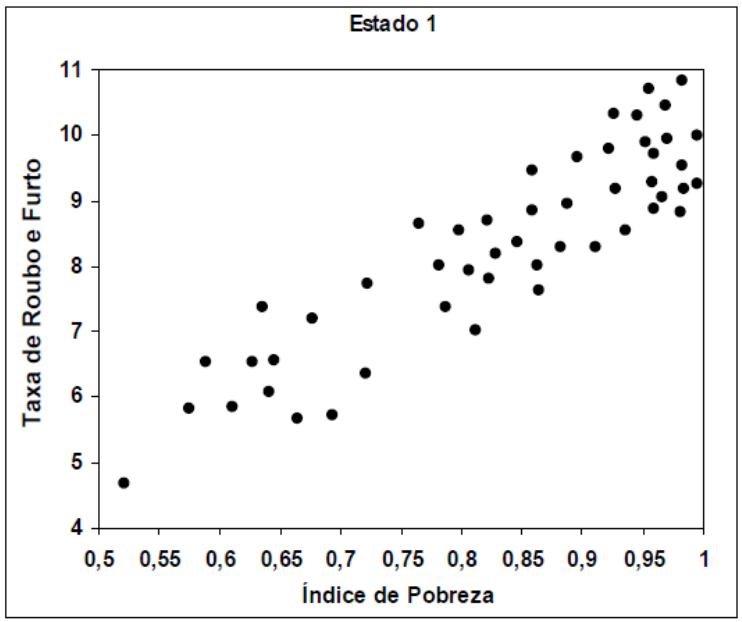
- Existe associação linear entre as variáveis? Sim.
- Ela é forte ou fraca? Forte.
- O valor de r deve estar em torno de quanto?

Exemplo (continuação)



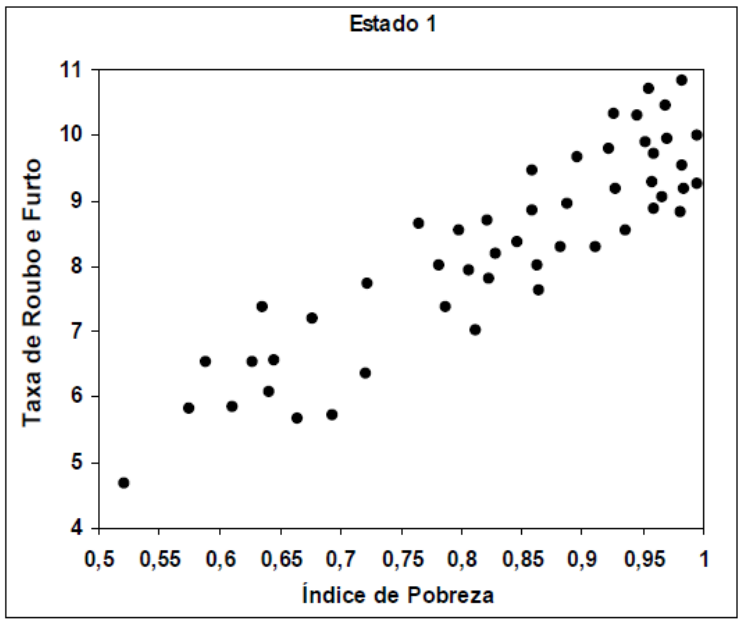
- Existe associação linear entre as variáveis? Sim.
- Ela é forte ou fraca? Forte.
- O valor de r deve estar em torno de quanto?
  $r = 0,989$.

Exemplo (continuação)



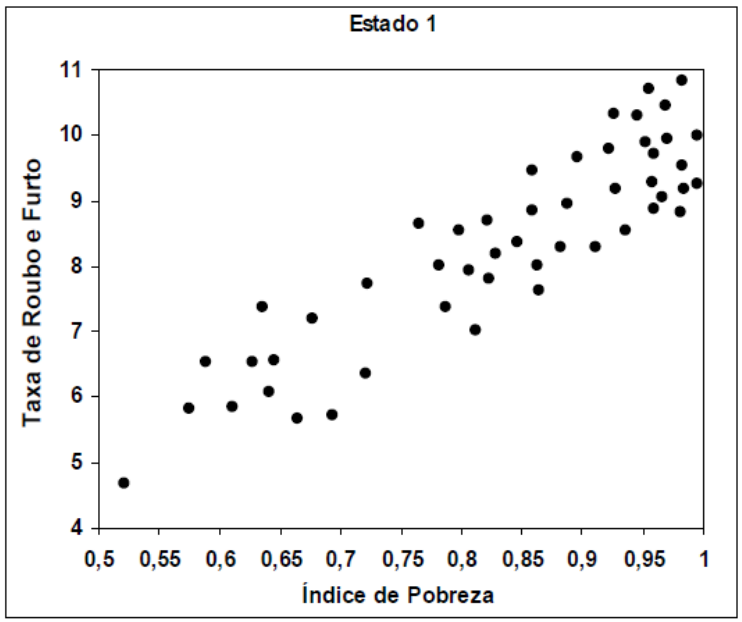
- Existe associação linear entre as variáveis?

Exemplo (continuação)



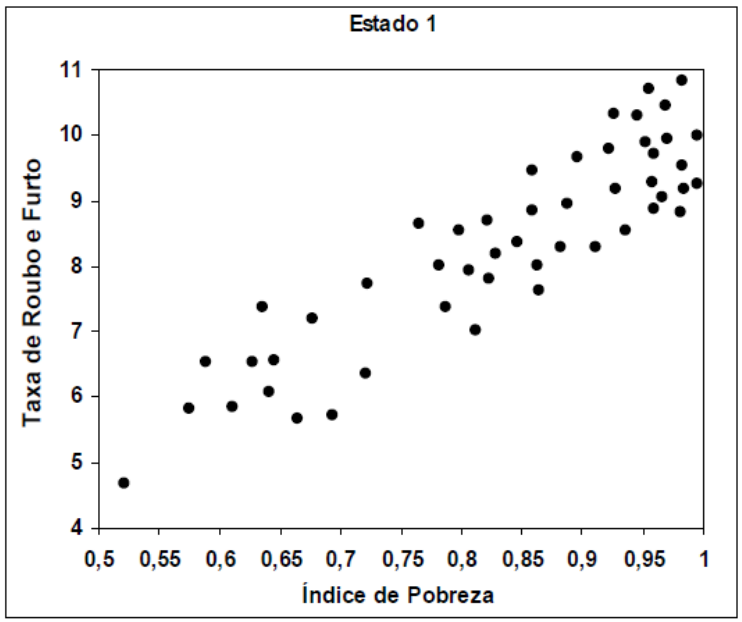
- Existe associação linear entre as variáveis? Sim.
- Ela é forte ou fraca?

Exemplo (continuação)



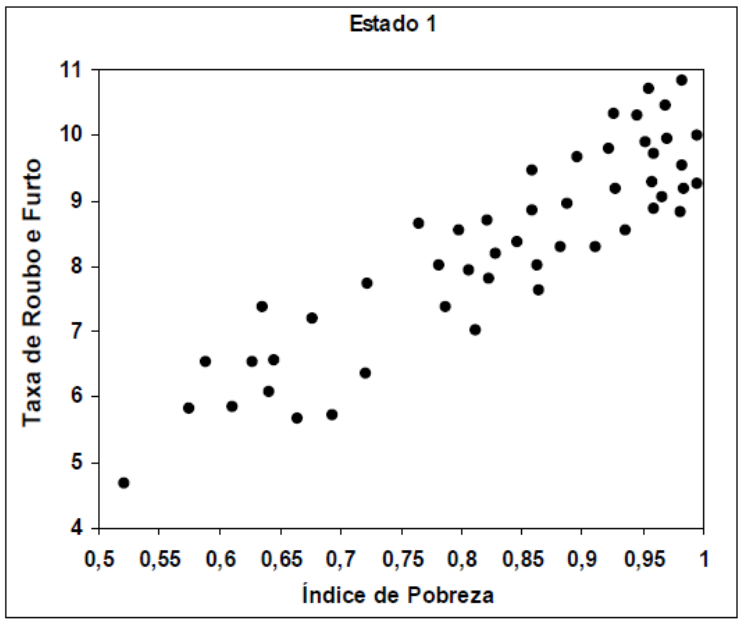
- Existe associação linear entre as variáveis? Sim.
- Ela é forte ou fraca? Forte.
- O valor de r é mais baixo ou mais alto que o anterior?

Exemplo (continuação)



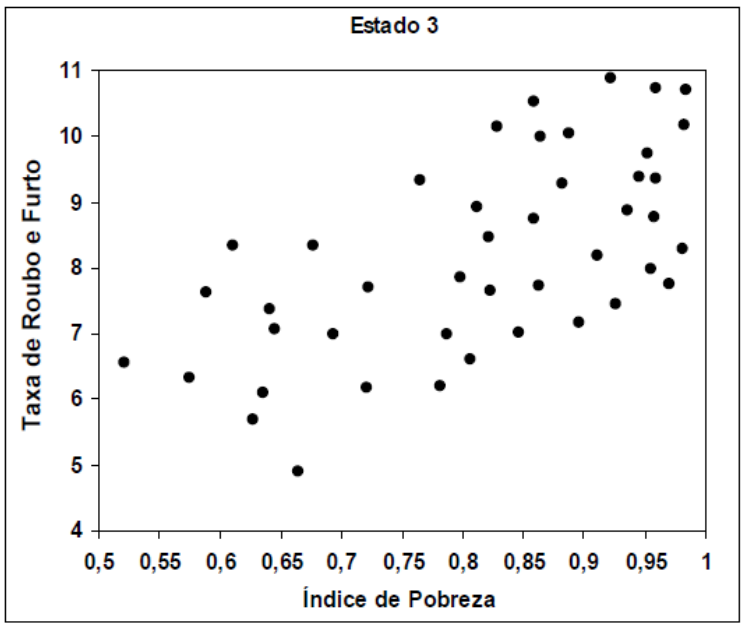
- Existe associação linear entre as variáveis? Sim.
- Ela é forte ou fraca? Forte.
- O valor de *r* é mais baixo ou mais alto que o anterior? Mais baixo.
- O valor de *r* deve estar em torno de quanto?

Exemplo (continuação)



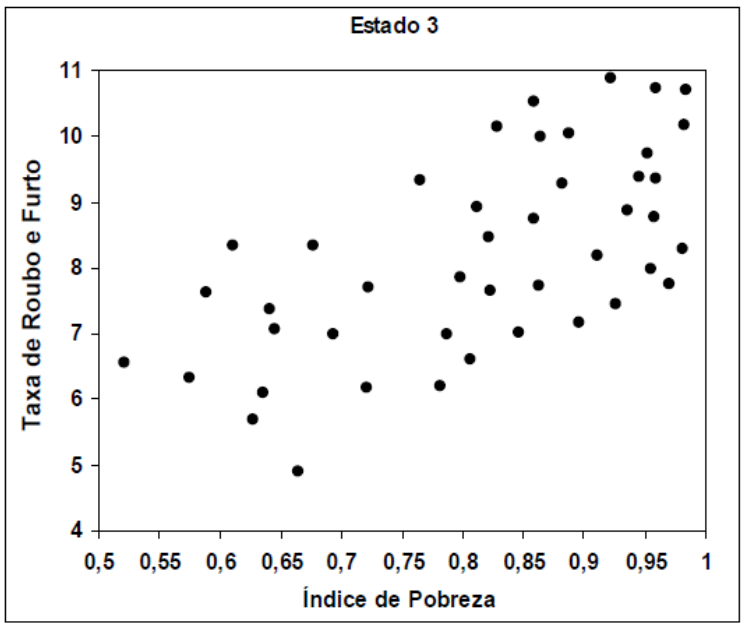
- Existe associação linear entre as variáveis? Sim.
- Ela é forte ou fraca? Forte.
- O valor de r é mais baixo ou mais alto que o anterior? Mais baixo.
- O valor de r deve estar em torno de quanto? $r = 0,898$.

Exemplo (continuação)



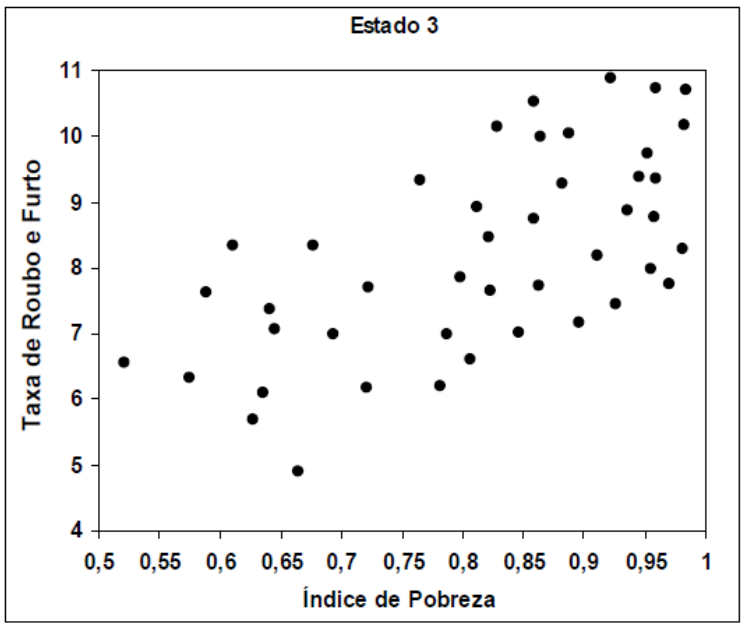
- Existe associação linear entre as variáveis?

Exemplo (continuação)



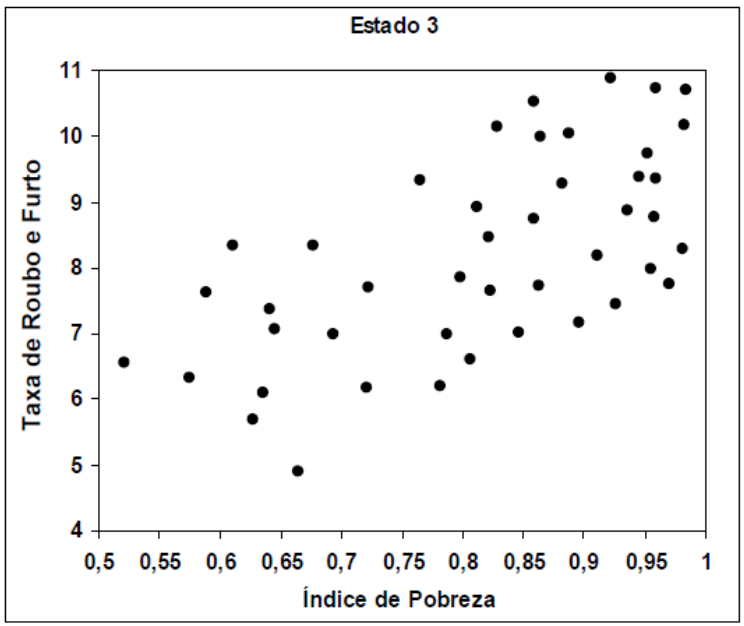
- Existe associação linear entre as variáveis? Sim.
- Ela é forte, moderada ou fraca?

Exemplo (continuação)



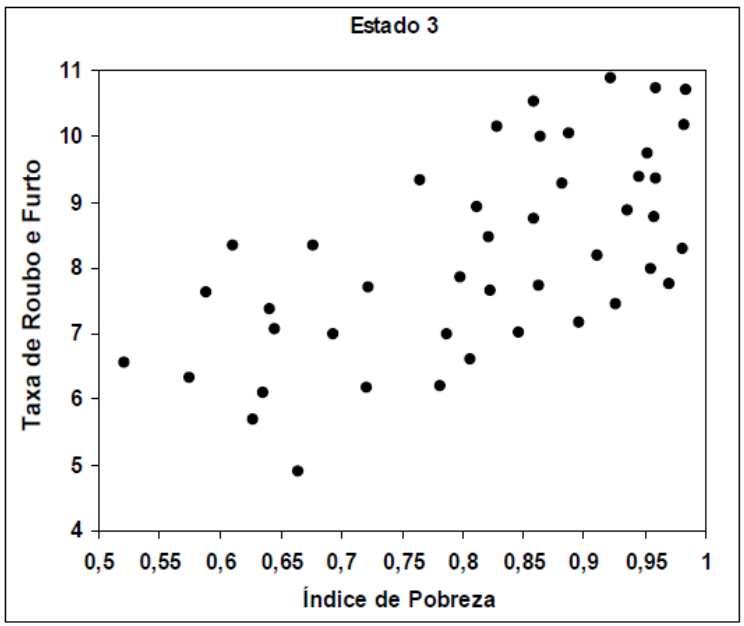
- Existe associação linear entre as variáveis? Sim.
- Ela é forte, moderada ou fraca? Moderada.
- O valor de r é mais baixo ou mais alto que o anterior?

Exemplo (continuação)



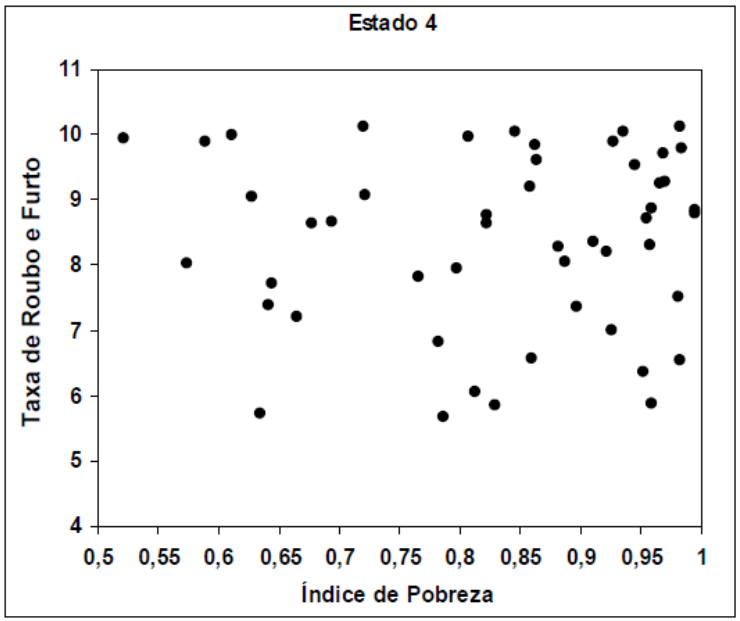
- Existe associação linear entre as variáveis? Sim.
- Ela é forte, moderada ou fraca? Moderada.
- O valor de *r* é mais baixo ou mais alto que o anterior? Mais baixo.
- O valor de *r* deve estar em torno de quanto?

Exemplo (continuação)



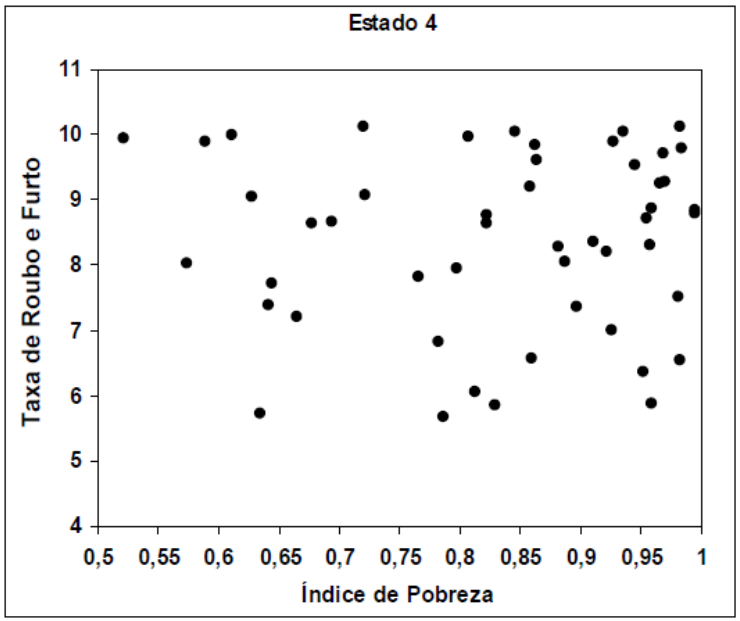
- Existe associação linear entre as variáveis? Sim.
- Ela é forte, moderada ou fraca? Moderada.
- O valor de r é mais baixo ou mais alto que o anterior? Mais baixo.
- O valor de r deve estar em torno de quanto? $r = 0,692$.

Exemplo (continuação)



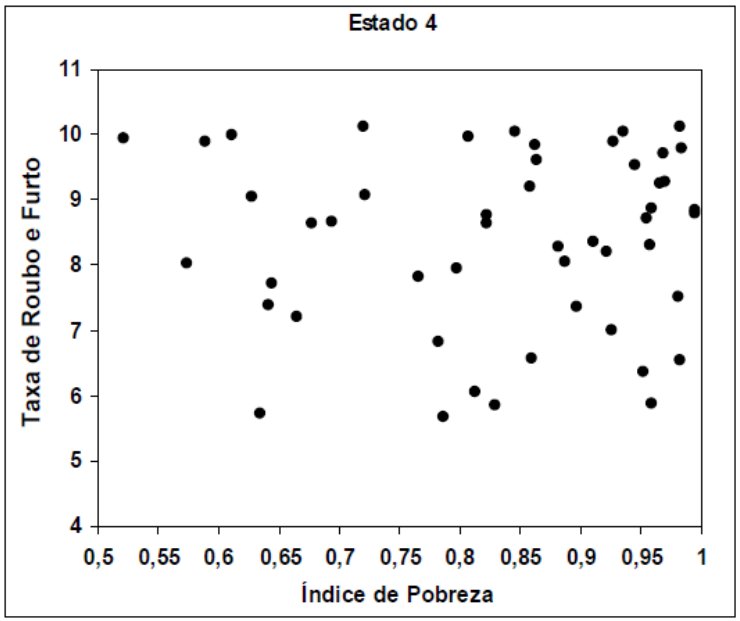
- Existe associação linear entre as variáveis?

Exemplo (continuação)



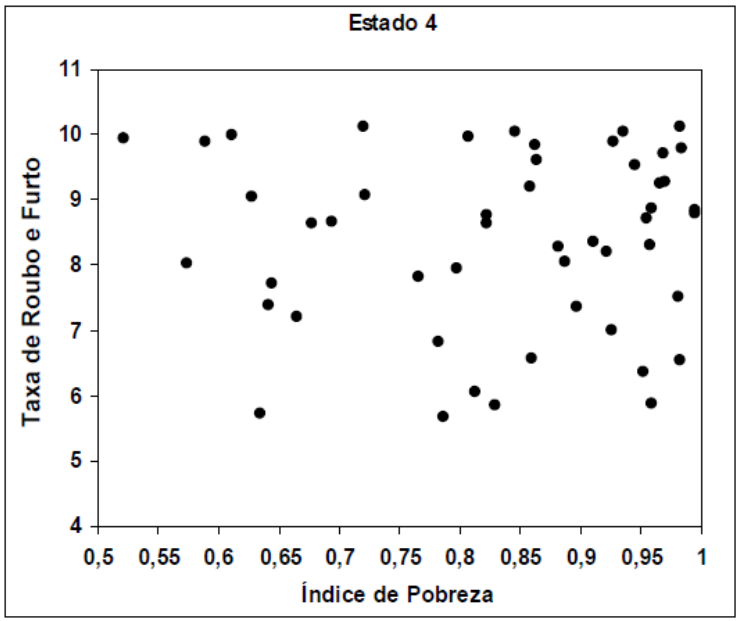
- Existe associação linear entre as variáveis? Não.
- O valor de r é mais baixo ou mais alto que o anterior?

Exemplo (continuação)



- Existe associação linear entre as variáveis? Não.
- O valor de r é mais baixo ou mais alto que o anterior? Mais baixo.
- O valor de r deve estar em torno de quanto?

Exemplo (continuação)



- Existe associação linear entre as variáveis? Não.
- O valor de r é mais baixo ou mais alto que o anterior? Mais baixo.
- O valor de r deve estar em torno de quanto? $r = 0,019$.

Exemplo

- Queremos verificar se existe associação entre idade e pressão sanguínea.
- Os dados são mostrados a seguir:

Aluno	Age	Blood Pressure	Age* BP	age^2	BP^2
A	43	128	5504	1849	16384
B	48	120	5760	2304	14400
C	56	135	7560	3136	18225
D	61	143	8723	3721	20449
E	67	141	9447	4489	19881
F	70	152	10640	4900	23104
Soma	345	819	4764	20399	112443

Exemplo (continuação)

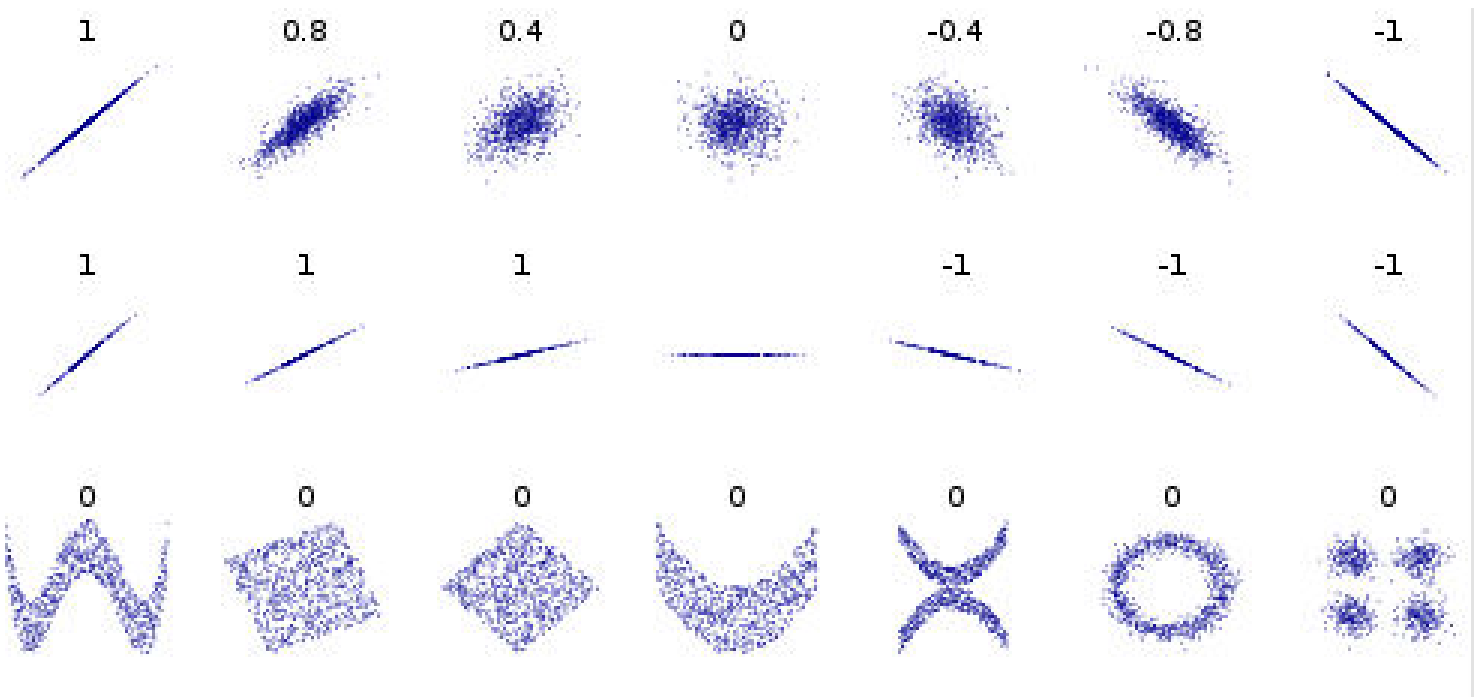
- Temos então que

$$r = \frac{(6)(47634) - (345)(819)}{\sqrt{(6)(20399) - 345^2}\sqrt{(6)(112443) - 819^2}}$$

$$= 0,897$$

- As variáveis estão fortemente associadas.
- A associação é positiva:
  - quanto maior a idade $\Rightarrow$ maior a pressão sanguínea.

- A figura abaixo mostra valores de *r* para vários conjuntos de dados distintos.



Coeficiente de correlação populacional ρ

- É uma característica da população e não da amostra.
- Não sabemos seu valor verdadeiro.
- Mas podemos estimá-lo pelo coeficiente amostral.
- Pode-se fazer testes e intervalos de confiança para esse parâmetro.
- Pode ser que na população $\rho = 0$,
  - mas na amostra $r \neq 0$ por mero acaso.
- Vamos ver a seguir como podemos testar se o coeficiente é significativo.

Teste do Coeficiente de Correlação

- O coeficiente de correlação r é apenas uma estimativa amostral.
- Ele é calculado com base em uma amostra de tamanho n.
- Os valores amostrais podem apresentar uma correlação, mas a população não.
- Se $r \neq 0$ não garante que $\rho \neq 0$.
- Podemos fazer um teste de hipótese para verificar se de fato $\rho \neq 0$.

- As hipóteses a serem testadas são as seguintes:

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0 \, .$$

- Sob H_0 temos que $\rho = 0$.
- Além disso, pode-se mostrar que

$$Var(r) = \frac{1 - \rho^2}{n - 2}$$

que é estimada por

$$Var(r) = \frac{1 - r^2}{n - 2} \, .$$

- A estatística de teste é dada por:

$$t = \frac{r - \text{valor sob } H_0}{\sqrt{Var(r)}}$$

- A estatística de teste fica

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

- Como a variância está sendo estimada, essa estatística tem uma distribuição t-student com $n-2$ graus de liberdade.

Exemplo

- Considere o exemplo de índices de pobreza e violência.
- Temos que

$$r = \frac{n\sum_i x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum_i x_i^2) - (\sum_i x_i)^2}\sqrt{n(\sum_i y_i^2) - (\sum_i y_i)^2}} = 0,898\,.$$

- Queremos testar

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0\,.$$

- A estatística de teste é dada por

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,898\sqrt{50-2}}{\sqrt{1-0,898^2}} = 14,25\,.$$

Exemplo (continuação)

- Vamos considerar $\alpha = 0,05$.
- Temos $t_{48;0,025} \approx z_{0,025}$ pois n é grande.
- $z_{0,025} = 1,96$
- Qual a região crítica do teste?

Exemplo (continuação)

- Vamos considerar $\alpha = 0,05$.
- Temos $t_{48;0,025} \approx z_{0,025}$ pois n é grande.
- $z_{0,025} = 1,96$
- Qual a região crítica do teste?

$$t > 1,96 \text{ ou } t < -1,96 .$$

- Qual a conclusão?

Exemplo (continuação)

- Vamos considerar $\alpha = 0,05$.
- Temos $t_{48;0,025} \approx z_{0,025}$ pois n é grande.
- $z_{0,025} = 1,96$
- Qual a região crítica do teste?

$$t > 1,96 \text{ ou } t < -1,96 \, .$$

- Qual a conclusão?
- Como $t = 14,25 > 1,96$, rejeitamos H_0.
- Com 5% de significância há evidências de que $\rho \neq 0$, ou seja, de que existe associação entre as variáveis.

Exemplo (continuação)

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$

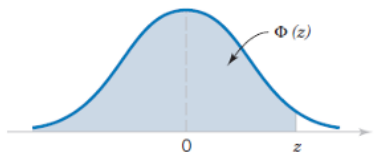
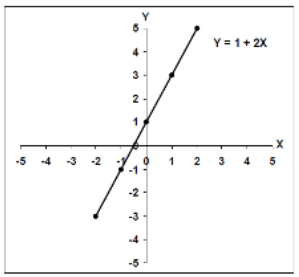


Table II Cumulative Standard Normal Distribution (*continued*)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705

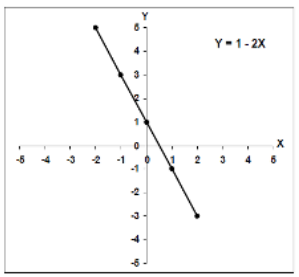
Revisão de matemática: equação da reta

$y = a + bx$



$y = 1 + 2x$

X	Y
-2	-3
-1	-1
0	1
1	3
2	5



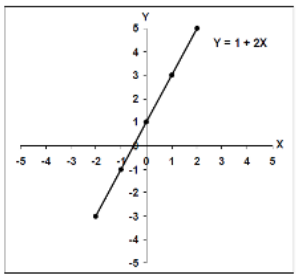
$y = 1 - 2x$

X	Y
-2	5
-1	3
0	1
1	-1
2	-3

- Coeficiente Linear ou Intercepto (a):

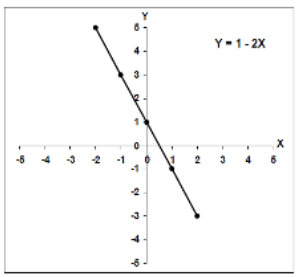
Revisão de matemática: equação da reta

$y = a + bx$



y = 1 + 2x

X	Y
-2	-3
-1	-1
0	1
1	3
2	5



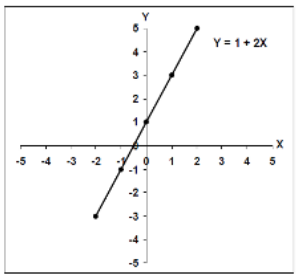
y = 1 - 2x

X	Y
-2	5
-1	3
0	1
1	-1
2	-3

- Coeficiente Linear ou Intercepto (a): valor de y quando $x = 0$.
- Coeficiente Angular ou Inclinação da reta (b):
  - $b > 0$ a reta é

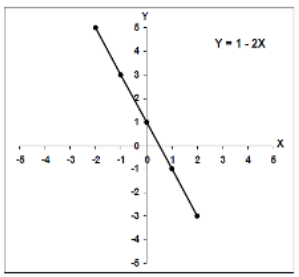
Revisão de matemática: equação da reta

$y = a + bx$



y = 1 + 2x

X	Y
-2	-3
-1	-1
0	1
1	3
2	5



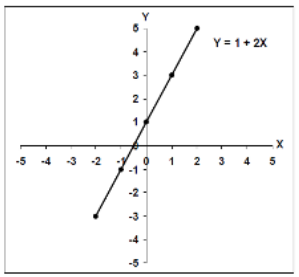
y = 1 - 2x

X	Y
-2	5
-1	3
0	1
1	-1
2	-3

- Coeficiente Linear ou Intercepto (a): valor de y quando $x = 0$.
- Coeficiente Angular ou Inclinação da reta (b):
  - $b > 0$ a reta é crescente (x cresce, y cresce)
  - $b < 0$ a reta é

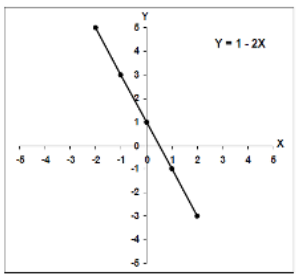
Revisão de matemática: equação da reta

$y = a + bx$



y = 1 + 2x

X	Y
-2	-3
-1	-1
0	1
1	3
2	5



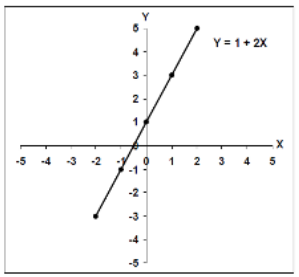
y = 1 - 2x

X	Y
-2	5
-1	3
0	1
1	-1
2	-3

- Coeficiente Linear ou Intercepto (a): valor de y quando $x = 0$.
- Coeficiente Angular ou Inclinação da reta (b):
  - $b > 0$ a reta é crescente (x cresce, y cresce)
  - $b < 0$ a reta é decrescente (x cresce, y decresce)
  - $b = 0$ reta é

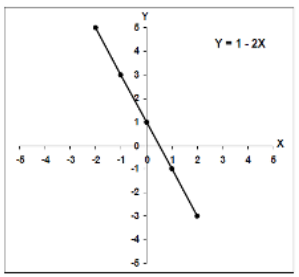
Revisão de matemática: equação da reta

$y = a + bx$



y = 1 + 2x

X	Y
-2	-3
-1	-1
0	1
1	3
2	5



y = 1 - 2x

X	Y
-2	5
-1	3
0	1
1	-1
2	-3

- Coeficiente Linear ou Intercepto (a): valor de y quando $x = 0$.
- Coeficiente Angular ou Inclinação da reta (b):
  - $b > 0$ a reta é crescente (x cresce, y cresce)
  - $b < 0$ a reta é decrescente (x cresce, y decresce)
  - $b = 0$ reta é paralela ao eixo x (x cresce, y não muda).

- Qual a derivada de $a + bx$ em relação a x?

- Qual a derivada de $a + bx$ em relação a x? b.
- Qual interpretação da derivada?

- Qual a derivada de $a + bx$ em relação a x? b.
- Qual interpretação da derivada?
- Quanto y varia quando x varia.
- O b representa o número de unidades que y aumenta ou diminui quando x aumenta em uma unidade.

Regressão Linear

- Verificamos até agora se existe correlação ou não entre as variáveis.
- Se existe, podemos querer descobrir a forma dessa associação.
- Queremos estimar a função que determina a relação entre as variáveis.
- Podemos usar a equação ajustada para prever valores da variável resposta.

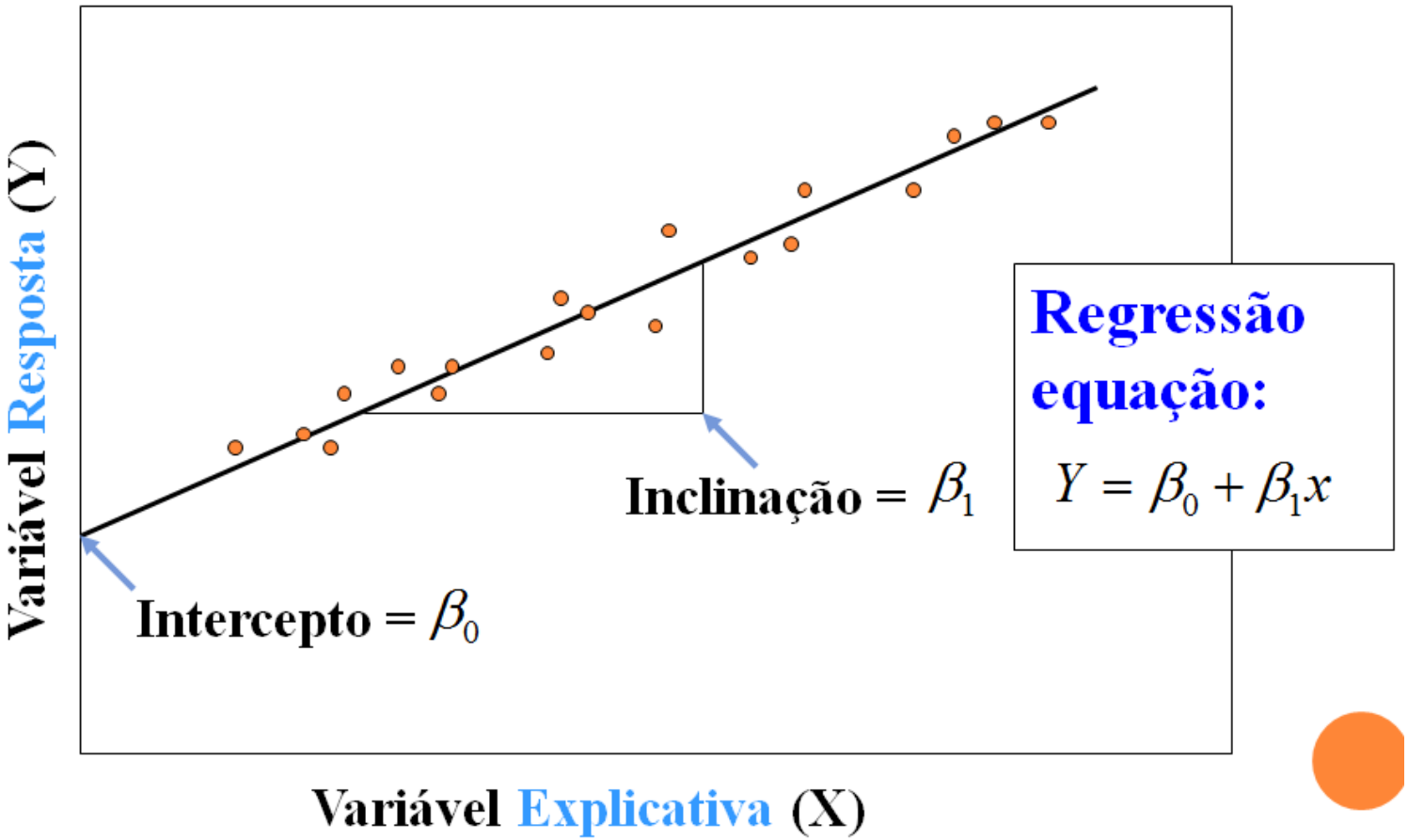
Exemplo:

- Estamos analisando um processo químico.
- O rendimento do produto está relacionado com a temperatura do processo.
- Podemos construir um modelo que seja capaz de:
  - prever o rendimento para uma dada temperatura.
- Esse modelo pode ser usado na otimização do processo:
  - encontrar a temperatura que maximiza o rendimento.

- É estabelecida uma equação

$$\underbrace{Y}_{\textit{resposta}} = \beta_0 + \beta_1 \underbrace{x}_{\textit{explicativa}}$$

  onde
  - β_0 é o intercepto em Y (x=0);
  - β_1 é inclinação (taxa de mudança).
- Veremos que essa equação não é exata.
- Precisamos incluir um erro aleatório.
- Vamos considerar assim por enquanto.

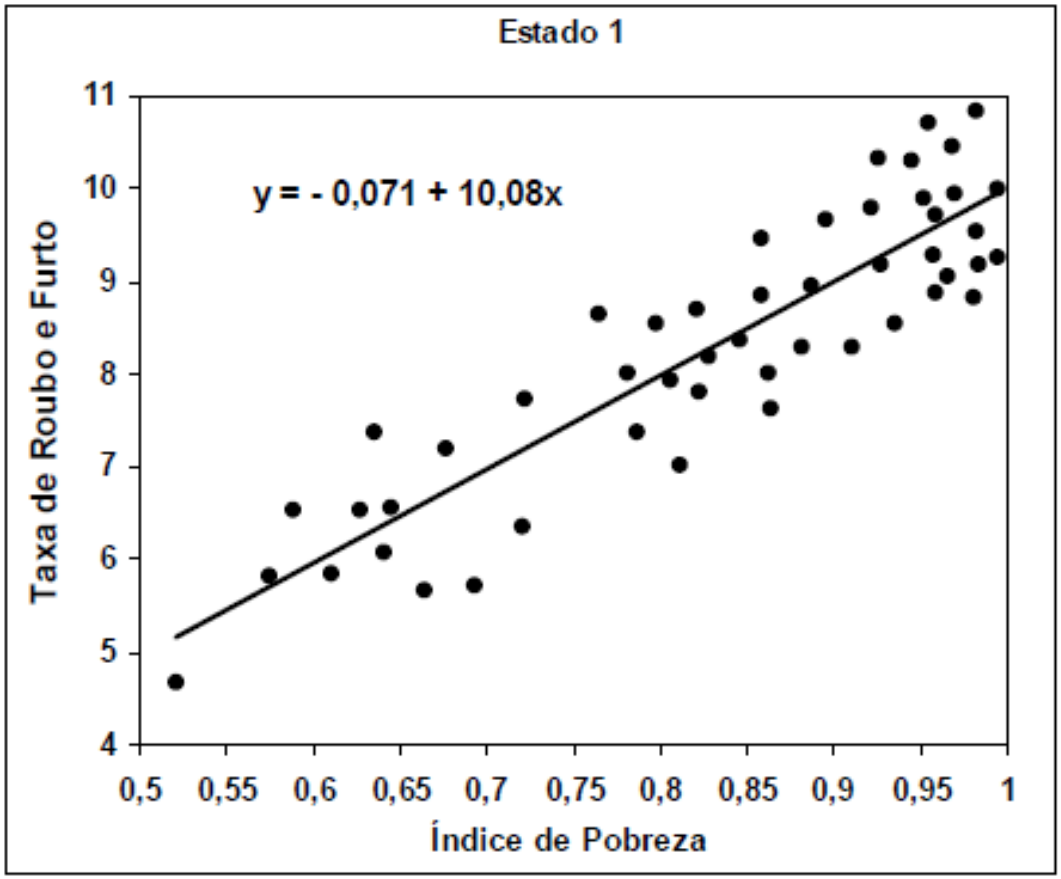

Variável Resposta (Y)
Inclinação = β_1
Intercepto = β_0
Regressão equação:
$Y = \beta_0 + \beta_1 x$
Variável Explicativa (X)

Exemplo:

- Considere novamente o exemplo de associação entre pobreza e criminalidade.
- Vimos que existe uma forte associação entre as variáveis.
- Podemos escrever a variável taxa de criminalidade em função da variável pobreza:

$$\text{Taxa Criminalidade} = -0,7 + 10,08\text{Pobreza}$$

Exemplo: (continuação)

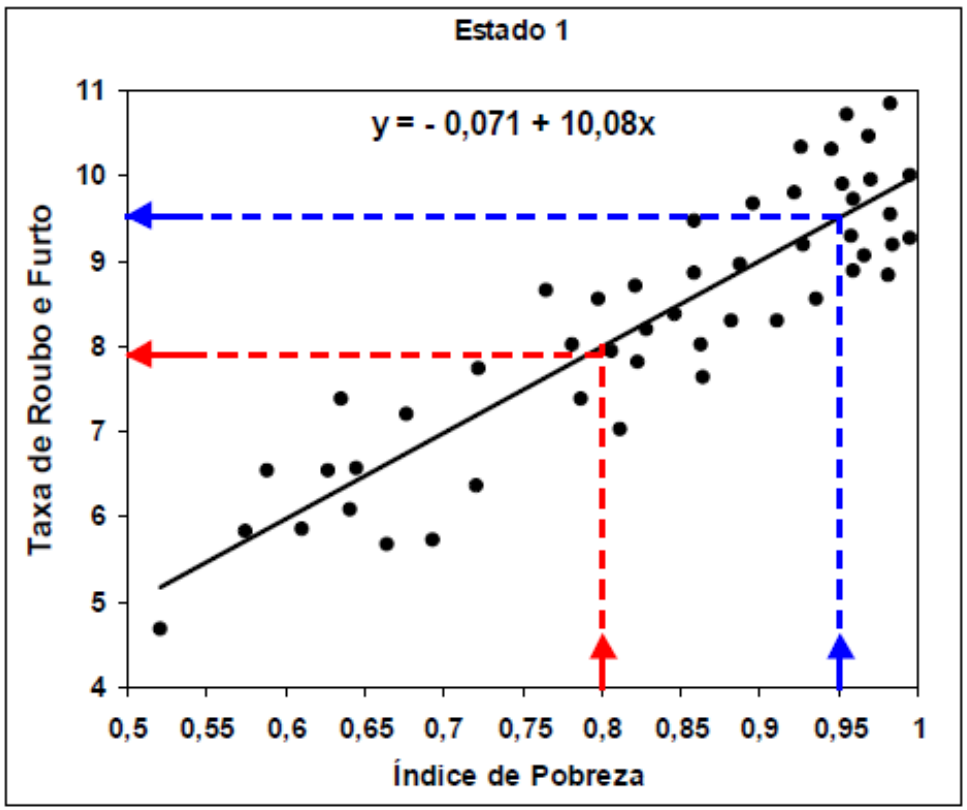


$r = 0,898$

$R^2 = 80,7\%$

- Considere um município com índice de pobreza $X = 0,8$.
- O valor esperado da taxa de furto é

$$-0,7 + 10,08 * (0,8) = 7,99 \text{ casos por mil habitantes}$$



Exemplo:

- Vamos olhar a relação entre:
  - y - pureza do oxigênio produzido em um processo de destilação;
  - x - porcentagem de hidrocarbonetos presentes no condensador.

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

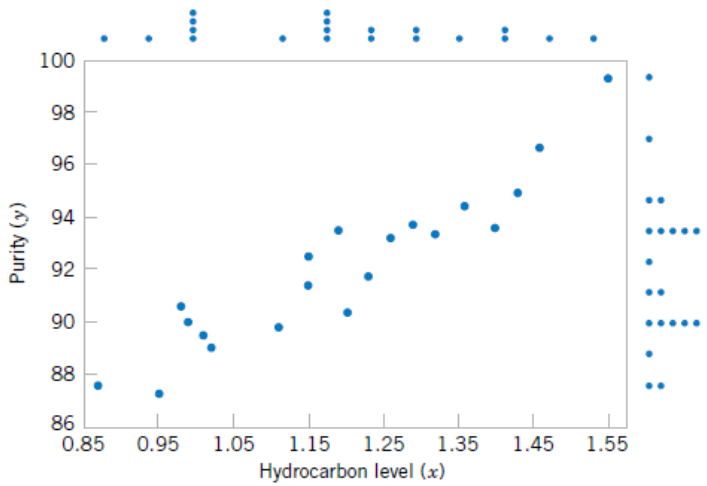


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Exemplo: (continuação)

- O **gráfico de dispersão** que representa cada par (x_i, y_i) como um ponto.
- Nenhuma curva simples passa exatamente por todos pontos.
- Os pontos parecem estar dispertsos aleatoriamente em torno de uma reta.
- É razoável considerar que a média de Y esteja relacionada linearmente com x

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x \, .$$

- β_0 e β_1 são chamados **coeficientes de regressão.**
- O valor de y não cai exatamente sobre a reta.

- O modelo

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x \, .$$

descreve a média de Y e não seu valor observado.

- Podemos generalizar para um modelo probabilístico.
- Consideramos que o valor esperado de x é função linear de Y.
- Para um valor fixo de x, o valor de Y é dado pela função do valor médio mais um erro aleatório

$$Y = \underbrace{\beta_0 + \beta_1 x}_{\text{Valor médio}} + \epsilon$$

onde ϵ é um erro aleatório.

- Esse modelo é chamado **modelo de regressão linear simples.**

- Se tivéssemos várias variáveis $(x_1, x_2, \ldots, x_p)$ no modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

é chamado **modelo de regressão linear múltipla.**

- Exemplo: renda explicada pelo sexo, faixa etária, escolaridade, etc.

- O modelo pode aparecer a parte de uma relação teórica.
- Exemplo:

$$p = x_0 + v \times t$$

onde

  - p é posição (y)
  - t tempo (x)
  - v velocidade (β_1)
  - x_0 posição inicial (β_0) .

- Em outros casos, não sabemos qual relação entre y e x.
- Então o modelo é escolhido a partir do diagrama de dispersão.
- Como foi feito para os dados do oxigênio.

- Quando não conhecemos uma relação teórica chamamos de **modelo empírico.**
- Escolhemos a forma mais adequada a partir de uma análise empírica dos dados.
- Essa forma não precisa ser necessariamente uma reta.
- Só iremos tratar aqui esse caso mais simples.

- Considere novamente o modelo

$$Y = \beta_0 + \beta_1 x + \epsilon \, .$$

- Consideramos que $\epsilon \sim N(0, \sigma^2)$.
- Se fixarmos x temos que

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) =$$

- Considere novamente o modelo

$$Y = \beta_0 + \beta_1 x + \epsilon \, .$$

- Consideramos que $\epsilon \sim N(0, \sigma^2)$.
- Se fixarmos x temos que

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = E(\beta_0) + E(\beta_1 x) + E(\epsilon)$$

$$=$$

- Considere novamente o modelo

$$Y = \beta_0 + \beta_1 x + \epsilon \, .$$

- Consideramos que $\epsilon \sim N(0, \sigma^2)$.
- Se fixarmos x temos que

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = E(\beta_0) + E(\beta_1 x) + E(\epsilon)$$

$$= \beta_0 + \beta_1 x + 0 = \beta_0 + \beta_1 x$$

(como na definição anterior).

- A variância é dada por

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + \epsilon) =$$

- Considere novamente o modelo

$$Y = \beta_0 + \beta_1 x + \epsilon \, .$$

- Consideramos que $\epsilon \sim N(0, \sigma^2)$.
- Se fixarmos x temos que

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = E(\beta_0) + E(\beta_1 x) + E(\epsilon)$$

$$= \beta_0 + \beta_1 x + 0 = \beta_0 + \beta_1 x$$

(como na definição anterior).

- A variância é dada por

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + \epsilon) = Var(\beta_0) + Var(\beta_1 x) + Var(\epsilon)$$

$$=$$

- Considere novamente o modelo

$$Y = \beta_0 + \beta_1 x + \epsilon \, .$$

- Consideramos que $\epsilon \sim N(0, \sigma^2)$.
- Se fixarmos x temos que

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = E(\beta_0) + E(\beta_1 x) + E(\epsilon)$$

$$= \beta_0 + \beta_1 x + 0 = \beta_0 + \beta_1 x$$

(como na definição anterior).

- A variância é dada por

$$Var(Y|x) = Var(\beta_0 + \beta_1 x + \epsilon) = Var(\beta_0) + Var(\beta_1 x) + Var(\epsilon)$$

$$= 0 + 0 + \sigma^2 = \sigma^2 \, .$$

Interpretação do β_1

- Temos que β_1 é a inclinação da reta.
- Então β_1 representa:
  - o aumento **esperado** em Y quando x aumenta uma unidade.
- Exemplo: Y (renda em mil reais) e x (escolaridade em anos)

$$Y = 0,5 + 1,5x + \epsilon \, .$$

- Espera-se um aumento de R$ 1500,00 no salário para cada ano a mais de estudo.

Interpretação do β_0

- β_0 é o intercepto da reta.
- Então β_0 representa:
  - o valor **esperado** de Y quando $x = 0$.
- Exemplo: Y (renda em mil reais) e x (escolaridade em anos)

$$Y = 0,5 + 1,5x + \epsilon \, .$$

- A renda esperada de uma pessoa sem estudo algum é de R$ 500,00.

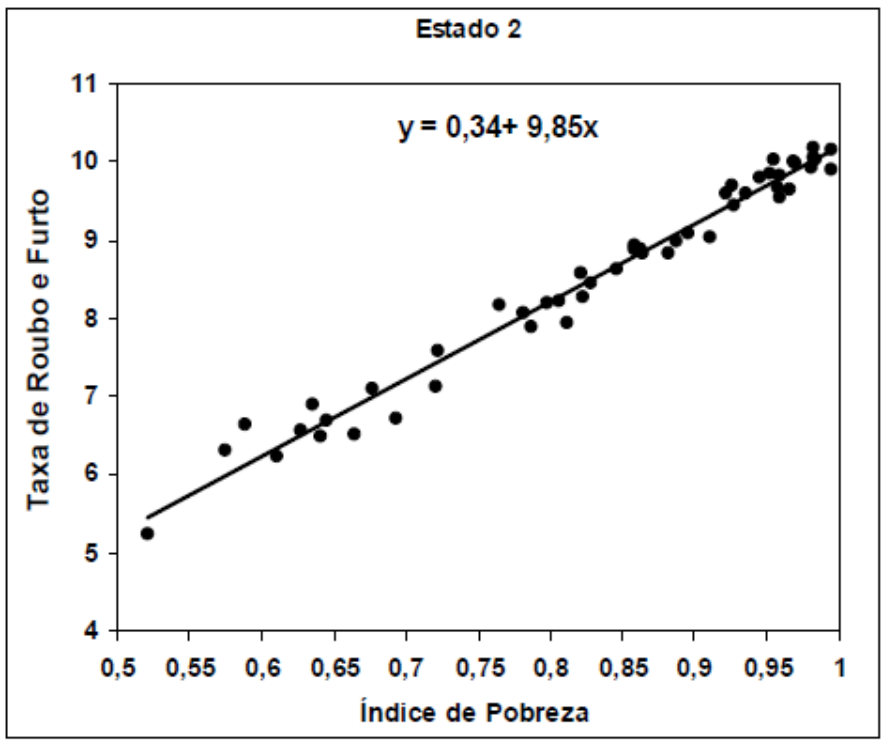
Exemplo:

- Considere novamente o exemplo de associação entre pobreza e criminalidade.
- Vimos que existe uma forte associação entre as variáveis.
- Temos que $\beta_0 = -0,7$.
- Esse coeficiente não tem sentido prático.
- Não existe taxa negativa.
- $\beta_1 = 10,08$.
- Interpretação: O aumento em uma unidade nos índices de pobreza aumenta em 10,08 o número de casos esperados de casos de roubos/furtos por mil habitantes.

Exemplo: (continuação)

- Para o Estado 2 temos que a reta de regressão é dada por:

$$\text{Taxa Criminalidade} = 0,34 + 9,85\text{Pobreza}$$



Exemplo: (continuação)

- Nesse caso a reta é bastante informativa.
- Há pouca dispersão dos pontos em torno dela.
- β_0 não tem interpretação.
- $\beta_1 = 9,85$.
- Qual interpretação?

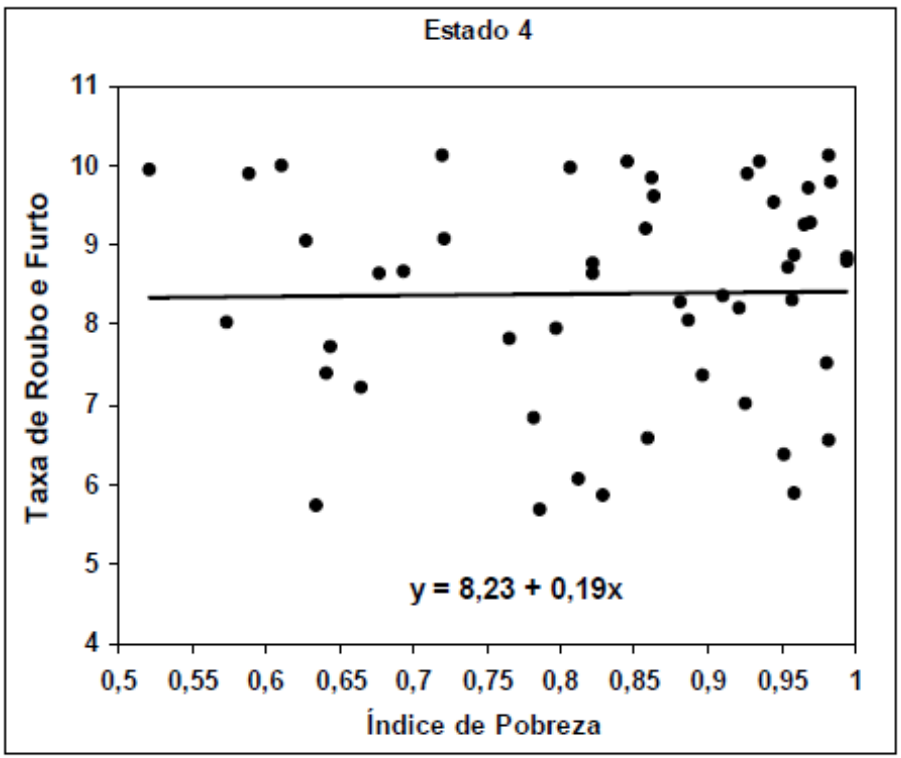
Exemplo: (continuação)

- Nesse caso a reta é bastante informativa.
- Há pouca dispersão dos pontos em torno dela.
- β_0 não tem interpretação.
- $\beta_1 = 9,85$.
- Qual interpretação?
- O aumento em uma unidade nos índices de pobreza aumenta em 9,85 o número de casos esperados de casos de roubos/furtos por mil habitantes.

Exemplo: (continuação)

- Para o Estado 4 temos que a reta de regressão é dada por:

$$\text{Taxa Criminalidade} = 8,23 + 0,19\text{Pobreza}$$



Exemplo: (continuação)

- A reta de regressão não é útil nesse caso.
- Não deve ser usada.
- Para esse estado, os níveis de pobreza não dizem nada sobre a criminalidade.
- Não podemos interpretar os coeficientes.

Exemplo:

- Considere o exemplo de destilação do oxigênio:
  - y - pureza do oxigênio produzido em um processo de destilação;
  - x - porcentagem de hidrocarbonetos presentes no condensador.
- Suponha que o verdadeiro modelo é dado por

$$Y = 75 + 15x + \epsilon$$

  onde $\epsilon \sim N(0, 2)$.
- Então

$$Y \sim$$

Exemplo:

- Considere o exemplo de destilação do oxigênio:
  - y - pureza do oxigênio produzido em um processo de destilação;
  - x - porcentagem de hidrocarbonetos presentes no condensador.
- Suponha que o verdadeiro modelo é dado por

$$Y = 75 + 15x + \epsilon$$

  onde $\epsilon \sim N(0, 2)$.
- Então

$$Y \sim N(75 + 15x, 2)\,.$$

- A variância do efeito aleatório σ^2
  - determina a variablidade do Y em torno da reta.
- Se σ^2 é grande $\Rightarrow$

Exemplo:

- Considere o exemplo de destilação do oxigênio:
  - y - pureza do oxigênio produzido em um processo de destilação;
  - x - porcentagem de hidrocarbonetos presentes no condensador.
- Suponha que o verdadeiro modelo é dado por

$$Y = 75 + 15x + \epsilon$$

onde $\epsilon \sim N(0, 2)$.

- Então

$$Y \sim N(75 + 15x, 2)\,.$$

- A variância do efeito aleatório σ^2
  - determina a variablidade do Y em torno da reta.
- Se σ^2 é grande $\Rightarrow$ as observações ficam longe da reta.
- Se σ^2 é pequeno $\Rightarrow$

Exemplo:

- Considere o exemplo de destilação do oxigênio:
  - y - pureza do oxigênio produzido em um processo de destilação;
  - x - porcentagem de hidrocarbonetos presentes no condensador.
- Suponha que o verdadeiro modelo é dado por

$$Y = 75 + 15x + \epsilon$$

onde $\epsilon \sim N(0, 2)$.

- Então

$$Y \sim N(75 + 15x, 2)\,.$$

- A variância do efeito aleatório σ^2
  - determina a variablidade do Y em torno da reta.
- Se σ^2 é grande $\Rightarrow$ as observações ficam longe da reta.
- Se σ^2 é pequeno $\Rightarrow$ as observações ficam perto da reta.

Exemplo: O modelo pode ser representado graficamente da seguinte forma:

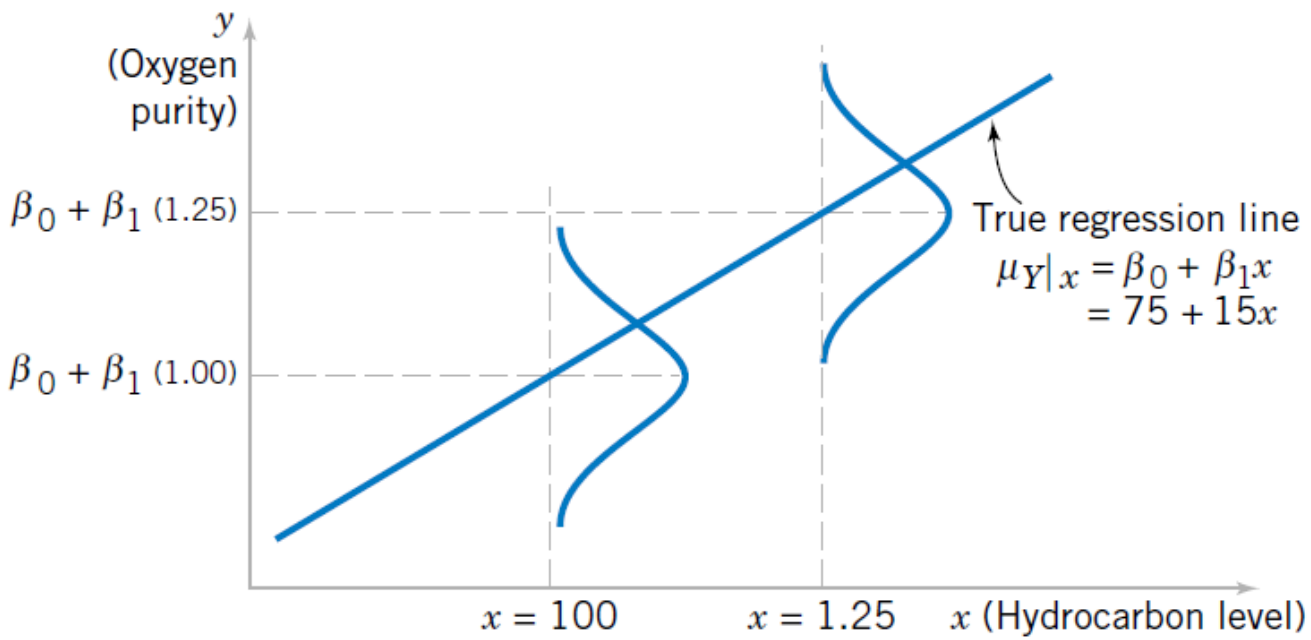


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity-hydrocarbon data.

Exemplo:

- Podemos usar o modelo para respondermos :
  - qual a pureza esperada do oxigênio para uma determinada porcentagem de hidrocarbonetos?
- Considere que a porcentagem de hidrocarbonetos é 1,25% ($x = 1,25$).
- Então a pureza esperada do oxigênio é de:

$$E(Y|x) = 75 + 15(1,25) = 93,75\,.$$

- Esse é um exemplo hipotético.
- Geralmente não saberemos o valor real de (β_0, β_1) e σ^2.
- São estimados a partir de dados da amostra.
- Veremos a seguir o método mais usado.
- O método de mínimos quadrados.

Abusos sobre a regressão

- Associação entre variáveis não implica relação causal.
- Planejamento de experimentos é a única forma de determinar relações causais.
- Relações de regrssão são válidas apenas dentro da faixa dos dados coletados.
- Modelos de regressão podem não ser válidos para extrapolação.

Uso da equação de regressão

- Podem ser úteis para predizer o valor de uma variável dado o valor de outra.
- Só podemos usá-la se o valor de r indica uma associação **linear** entre as variáveis.
- A reta de regressão precisa se ajustar bem aos dados.
- Se não existe uma correlação linear, a nossa melhor estimativa para Y é sua média.

- Se queremos predizer o valor de Y usando x:
  - se não existe relação linear entre x e y, o melhor valor é a média;
  - se existe relação linear, substitui o valor de x na reta de regressão.
- Uma reta ajustada no passado pode não ser útil hoje.
- Não devemos fazer predições para populações diferentes daquela de onde provem os dados amostrais.