Modelos de Regressão Linear Simples - parte II

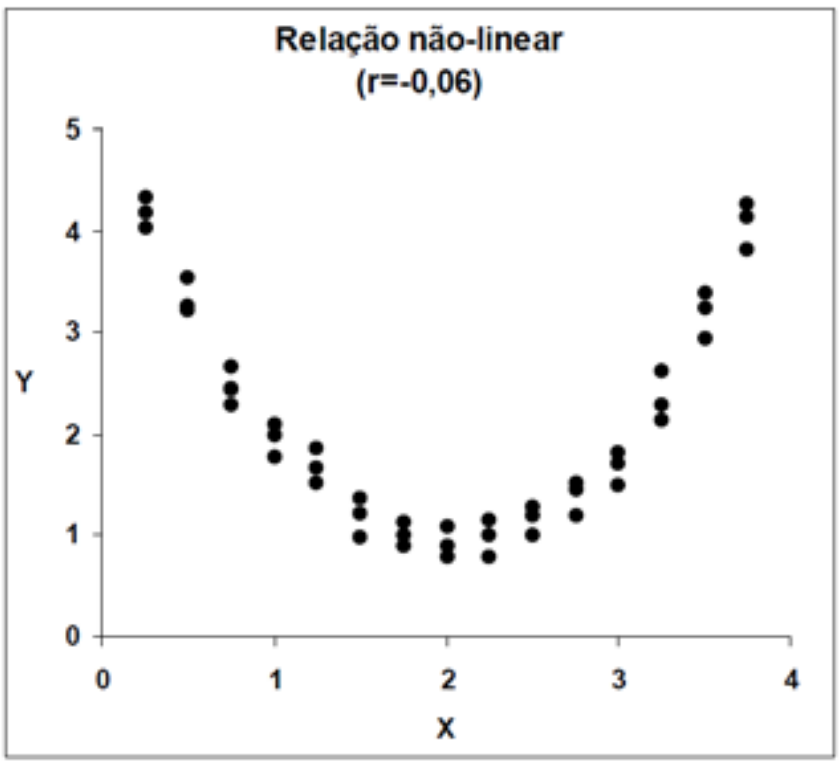
Erica Castilho Rodrigues

14 de Outubro de 2013

Erros Comuns que Envolvem a Análise de Correlação

Propriedade de linearidade

- Se $r \approx 0$ não significa que as variáveis não estão correlacionadas de forma alguma.
- O coeficiente de correlação mede apenas associação **linear** entre as variáveis.
- É preciso sempre fazer o diagrama de dispersão!

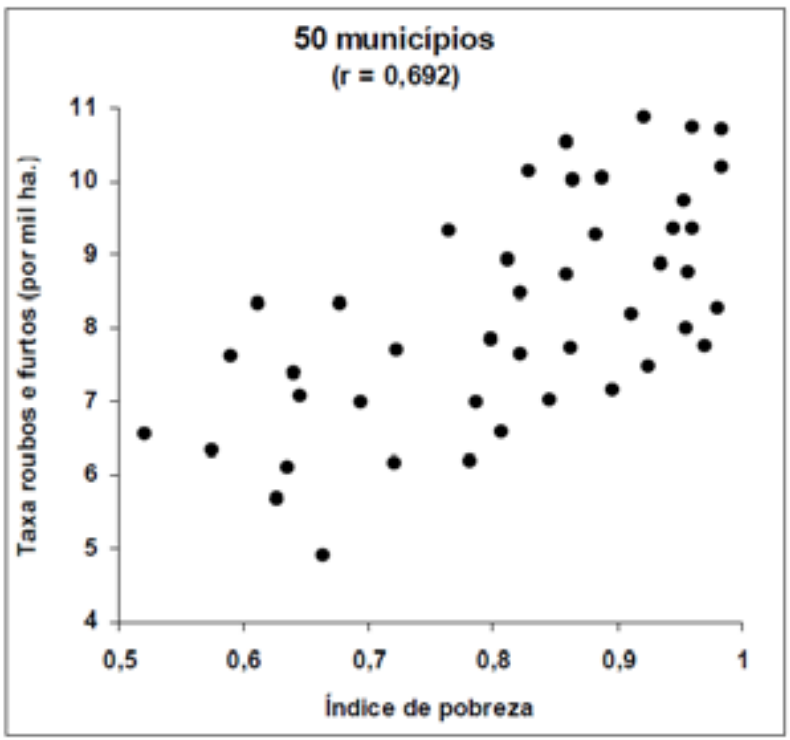


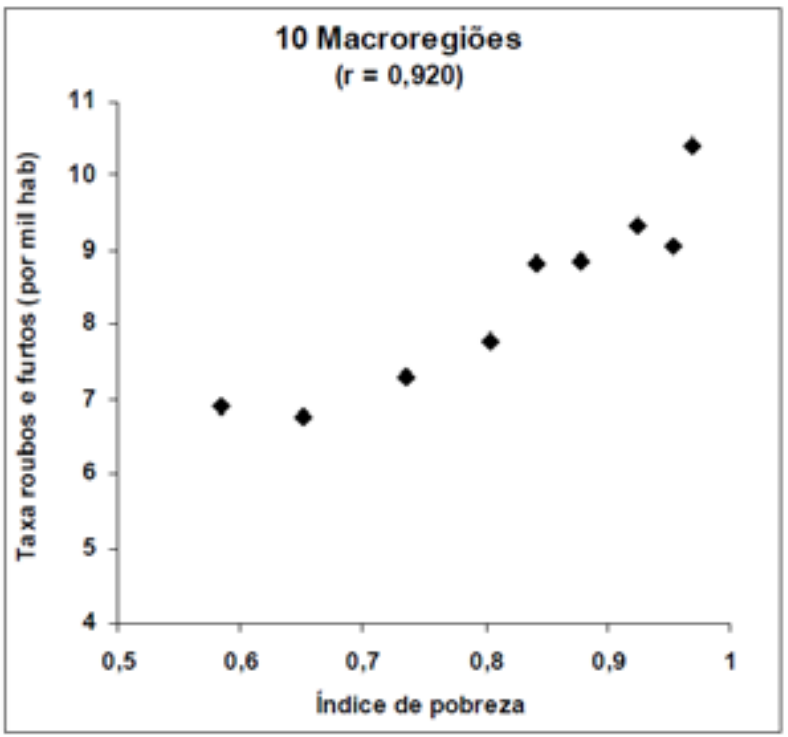
Dados agrupados em taxas ou em médias

- Quando agrupamos os dados, suprimimos variações entre os indivíduos.
- A variabilidade parece menor do que é na verdade.
- Isso pode inflacionar o coeficiente de correlação linear.
- Considere o exemplo da pobreza e criminalidade.
- Suponha que agrupamos os municípios em macroregiões.

- Tomamos a médias das variáveis em cada região.

- Quando agrupamos os dados a varibiliade diminui e a associação parece ser mais forte.
- Estamos escondendo uma fonte de variabilidade.





Concluir imediatamente que a correlação implica causalidade

- Uma pesquisa recente concluiu que:
  - países onde as pessoas consomem mais chocolate ganham um número maior de Prêmios Nobel.

FOOD & DRINK

Secret to Winning a Nobel Prize? Eat More Chocolate

By Olivia B. Waxman | Oct. 12, 2012 | 22 Comments

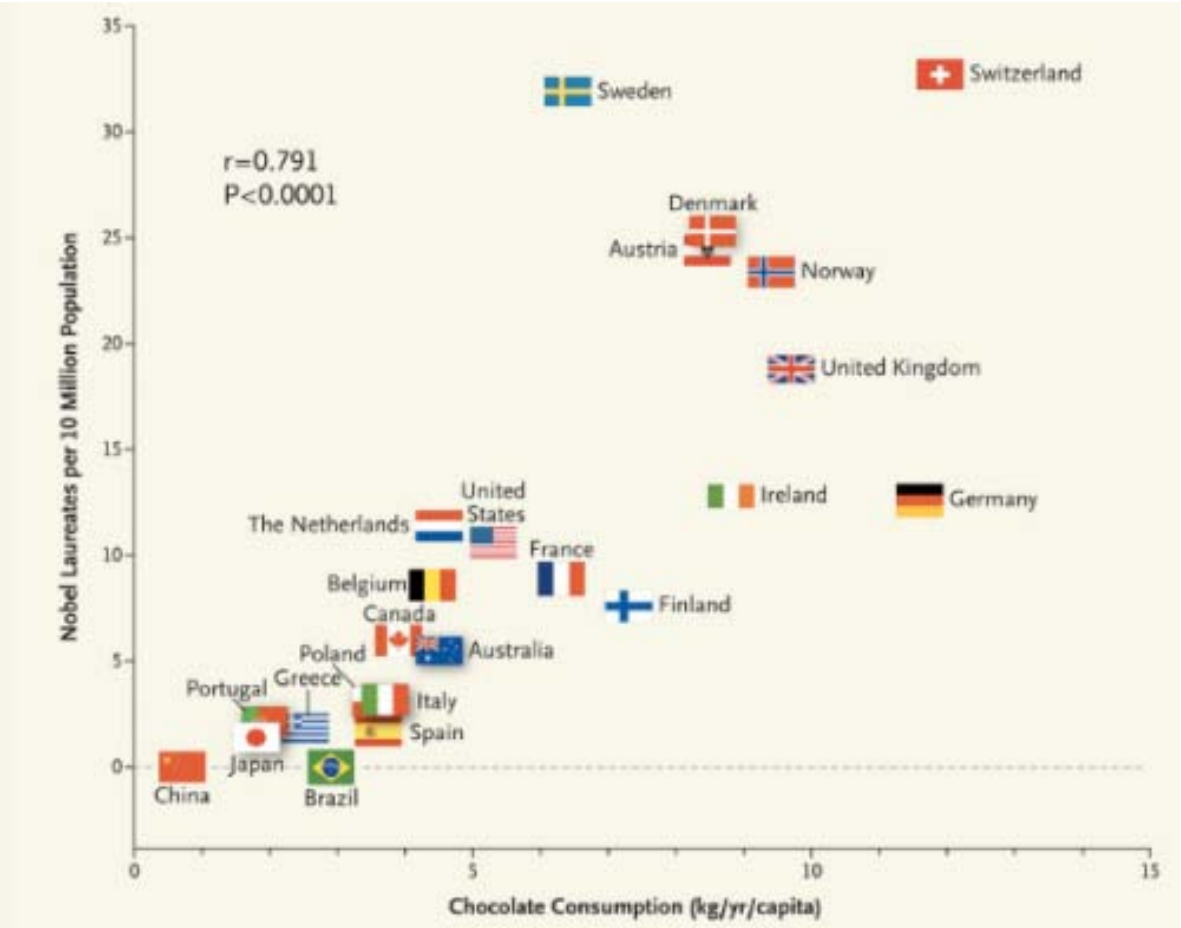
Tweet 625 +1 41 Share 26

As the Nobel Prizes are being awarded this week, one U.S. scientist asks: could eating chocolate have anything to do with becoming a laureate?

GETTY IMAGES

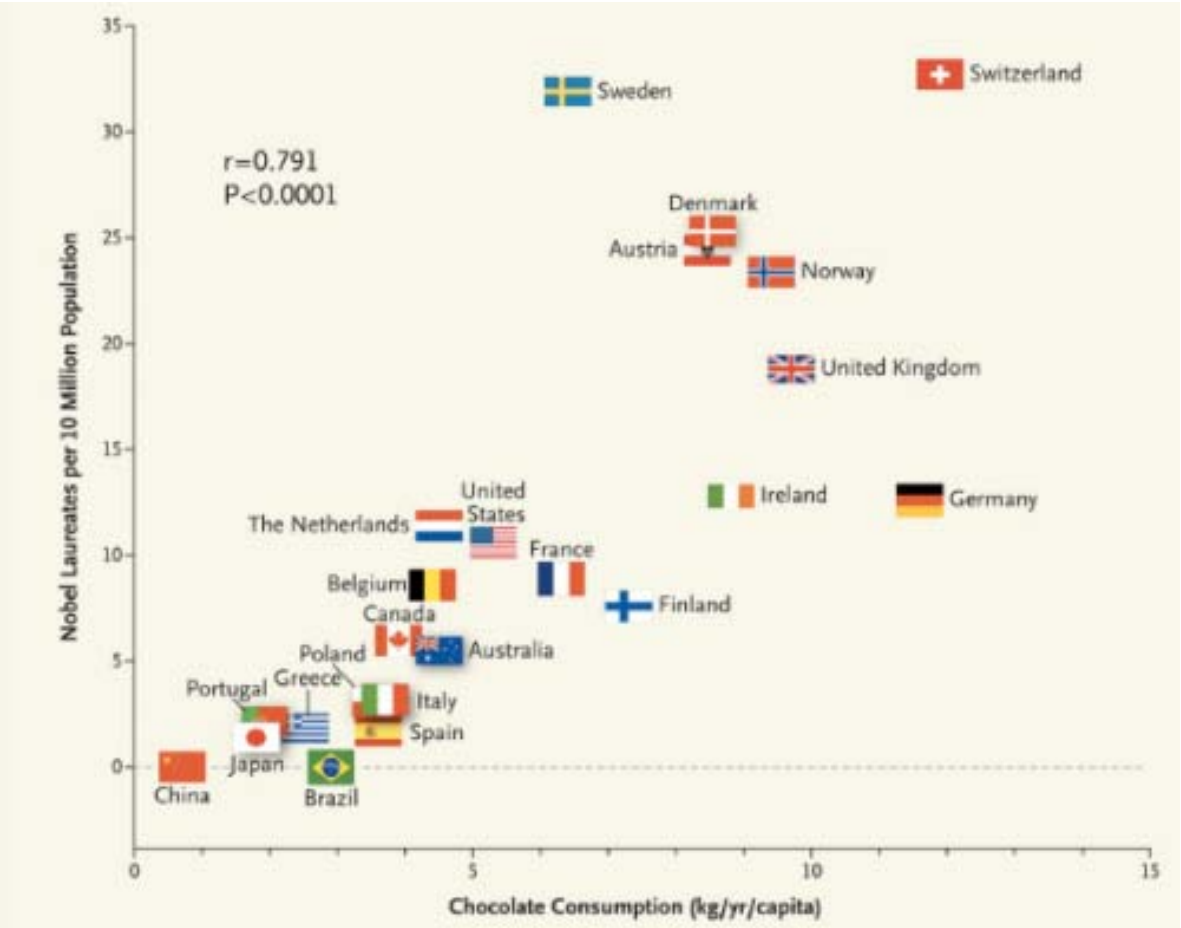
Why would the sweet treat be linked to winning the most prestigious intellectual award, you ask? In a "note" published in the *New England Journal of Medicine*, Dr. Franz H. Messerli, a cardiologist at St. Luke's–Roosevelt Hospital in New York City, writes that cocoa contains flavanols, plant-based compounds that previous studies have linked to the slowing or reversing of age-related cognitive decline. (You can also get flavonols in green tea, red wine and some fruits.)

- Temos a seguir o gráfico de dispersão das duas variáveis.



- Qual conclusão você tira?

- Temos a seguir o gráfico de dispersão das duas variáveis.



- Qual conclusão você tira? Parece existir uma associação positiva.

- Existe uma terceira variável oculta que leva ao aumento das duas quantidades.
- Qual variável é essa?

- Existe uma terceira variável oculta que leva ao aumento das duas quantidades.
- Qual variável é essa? Renda per capita.
- Países com maior renda per capita:
  - investem mais em educação, logo ganham mais Prêmios Nobel;
  - têm mais dinheiro para consumir chocolate.
- Portanto o consumo de chocolate não torna as pessoas mais inteligentes!

Método de Mínimos Quadrados

- Na regressão linear simples temos:
  - uma variável explicativa (x) e uma variável resposta(y).
- O valor esperado de Y para cada valor de x é

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

- A variável Y pode ser descrita pelo modelo

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

onde ϵ é um erro aleatório e

$$\epsilon \sim N(0, \sigma^2) \ .$$

- Os erros aleatório de diferentes observações são independentes.

- Suponha que temos n pares $(x_1, y_1), \ldots, (x_n, y_n)$.
- A figura mostra o gráfico de dispersão e uma candidata a reta de regressão.

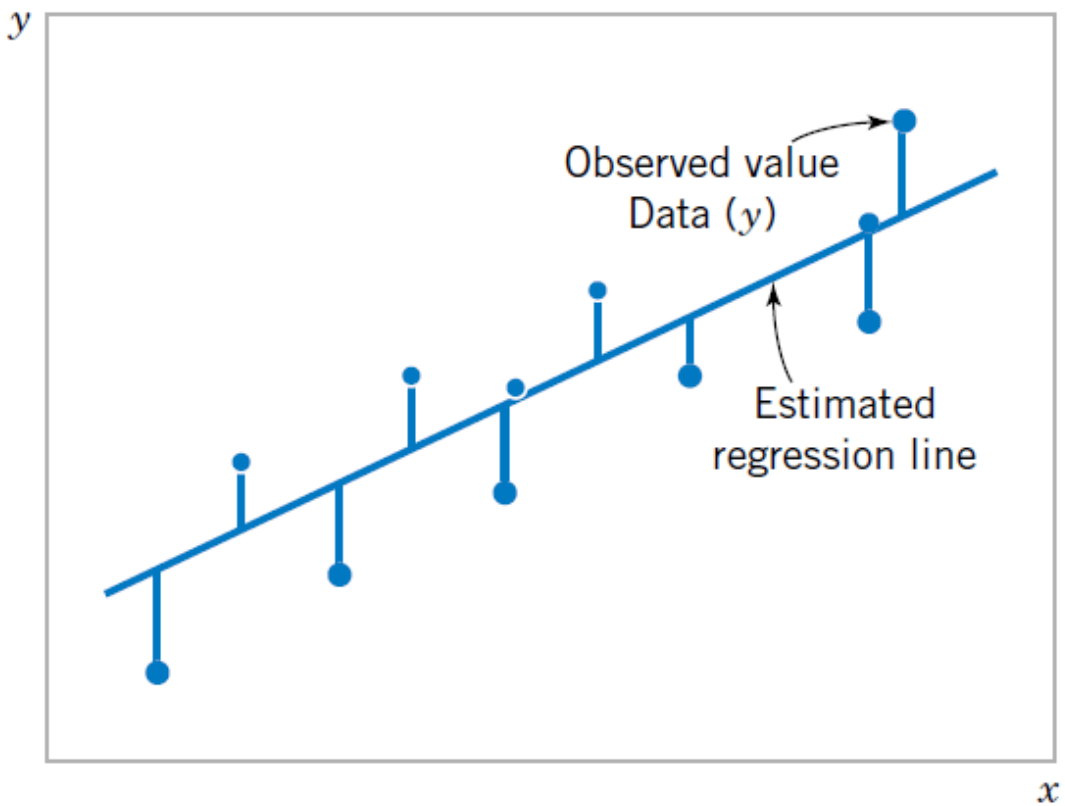


Figure 11-3 Deviations of the data from the estimated regression model.

Método de Mínimos Quadrados

Encontrar valores de β_0 e β_1 que minimizem a soma dos desvios ao quadrado.

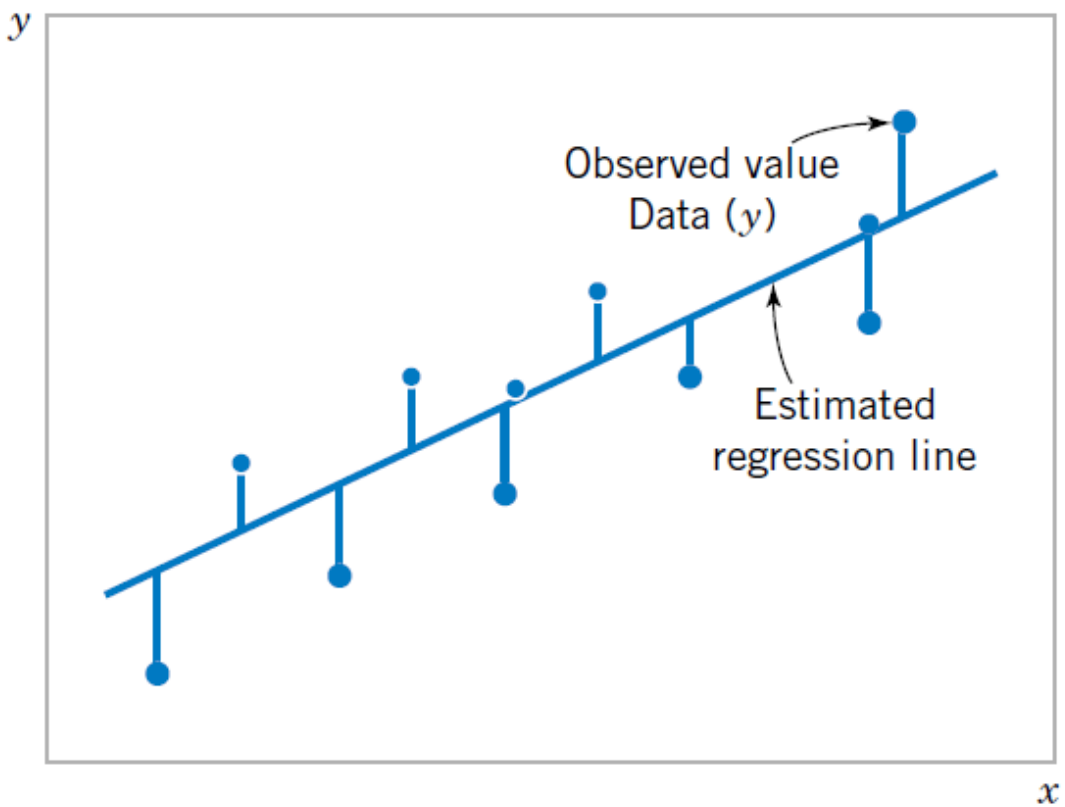


Figure 11-3 Deviations of the data from the estimated regression model.

- Utilizando a equação

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

as n observações da amostra podem ser expressas como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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

- A soma dos quadrados dos desvios das observações em relação a reta é dada por

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Defina $\hat{\beta}_0$ e $\hat{\beta}_1$ os estimadores de mínimos quadrados de β_1 e β_2.
- Queremos encontrar os valores de β_0 e β_1 que minimizam o erro.
- Como fazemos isso?

- Defina $\hat{\beta}_0$ e $\hat{\beta}_1$ os estimadores de mínimos quadrados de β_1 e β_2.
- Queremos encontrar os valores de β_0 e β_1 que minimizam o erro.
- Como fazemos isso? Deriva e iguala a zero.
- $\hat{\beta}_0$ e $\hat{\beta}_1$ devem satisfazer

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} =$$

- Defina $\hat{\beta}_0$ e $\hat{\beta}_1$ os estimadores de mínimos quadrados de β_1 e β_2.
- Queremos encontrar os valores de β_0 e β_1 que minimizam o erro.
- Como fazemos isso? Deriva e iguala a zero.
- $\hat{\beta}_0$ e $\hat{\beta}_1$ devem satisfazer

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} =$$

- Defina $\hat{\beta}_0$ e $\hat{\beta}_1$ os estimadores de mínimos quadrados de β_1 e β_2.
- Queremos encontrar os valores de β_0 e β_1 que minimizam o erro.
- Como fazemos isso? Deriva e iguala a zero.
- $\hat{\beta}_0$ e $\hat{\beta}_1$ devem satisfazer

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

- Simplificando essas equações temos que

- Simplificando essas equações temos que

$$n\hat{\beta}_0 + \hat{\beta}_i \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \ .$$

- Essas equações são chamadas equações normais.
- A solução dessas equações resulta nos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$.

Estimativas de Mínimos Quadrados

- A estimativa de mínimos quadrados do intersepto é dada por

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \, .$$

- A estimativa da inclinação é dada por

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \, .$$

- Em que

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \, .$$

- A linha de regressão ajustada é dada por

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

- Cada par de observações satisfaz

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, \ldots, n$$

onde e_i é chamado **resíduo** e é dado por

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

- O **resíduo** descreve o erro no ajuste do modelo para a i-ésima observação y_i.
- Veremos adiante:
  - como usar os resíduos para verificar adequação do modelo.

- Podemos usar símbolos especiais para o numerador e denominador das expressões de $\hat{\beta}_0$ e $\hat{\beta}_1$.
- Considere os dados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Defina as sequintes quantidades

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})$$

$$= \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

- Vejamos as demonstrações dos resultados:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2)$$

$$\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i} x_i + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\left(\frac{\sum_i x_i}{n}\right)^2$$

$$\sum_{i=1}^{n} x_i^2 - \frac{(\sum_i x_i)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) = \sum_{i=1}^{n}(y_i x_i - y_i\overline{x} - \overline{y}x_i + \overline{yx})$$

$$= \sum_i y_i x_i - \overline{x}\sum_i y_i - \overline{y}\sum_i x_i + n\overline{xy}$$

$$= \sum_i y_i x_i - n\overline{xy} - n\overline{xy} + n\overline{xy}$$

$$= \sum_i y_i x_i - n\overline{xy} = \sum_i y_i x_i - n\frac{\sum_i x_i \sum_i y_i}{n^2} = \sum_i y_i x_i - \frac{\sum_i x_i \sum_i y_i}{n}$$

Exemplo:

- Considere os dados de pureza de oxigênio:

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

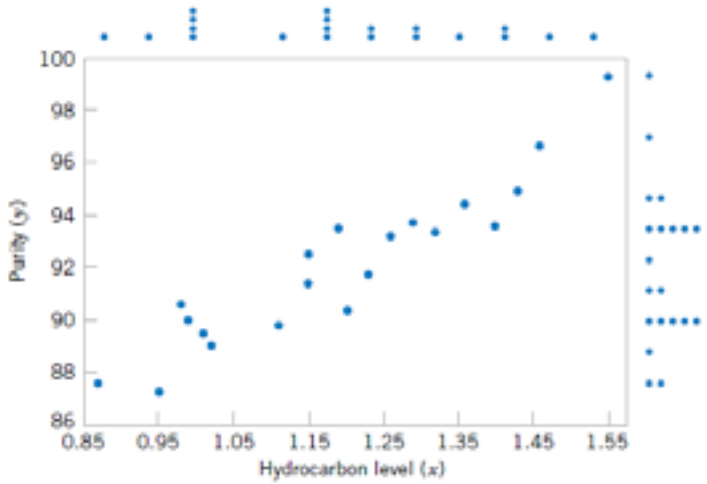


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Exemplo:

- Vamos ajustar o modelo de regressão linear simples para esses dados.
- Temos que

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23,92 \quad \sum_{i=1}^{20} y_i = 1842,21$$

$$\overline{x} = 1,1960 \quad \overline{y} = 92,1605\,.$$

$$\sum_{i=1}^{20} y_i^2 = 170044,5321 \quad \sum_{i=1}^{20} x_i^2 = 29,2892$$

$$\sum_{i=1}^{20} x_i y_i = 2214,6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29,2892 - \frac{(23,92)^2}{20} = 0,68088$$

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{(\sum_{i=1}^{20} x_i)(\sum_{i=1}^{20} y_i)}{20} = 2214,6566 - \frac{(23,92)(1843,21)}{20} = 10,1$$

- Então as estimativas de mínimos quadrados são

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10,17744}{0,68088} = 19,94748$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 92,1605 - (14,94748)1,196 = 74,28331\,.$$

- O modelo de regressão linear simples é:

$$\hat{y} = 74,283 - 14,947x\ .$$

- A figura a seguir mostra reta de regressão ajustada.

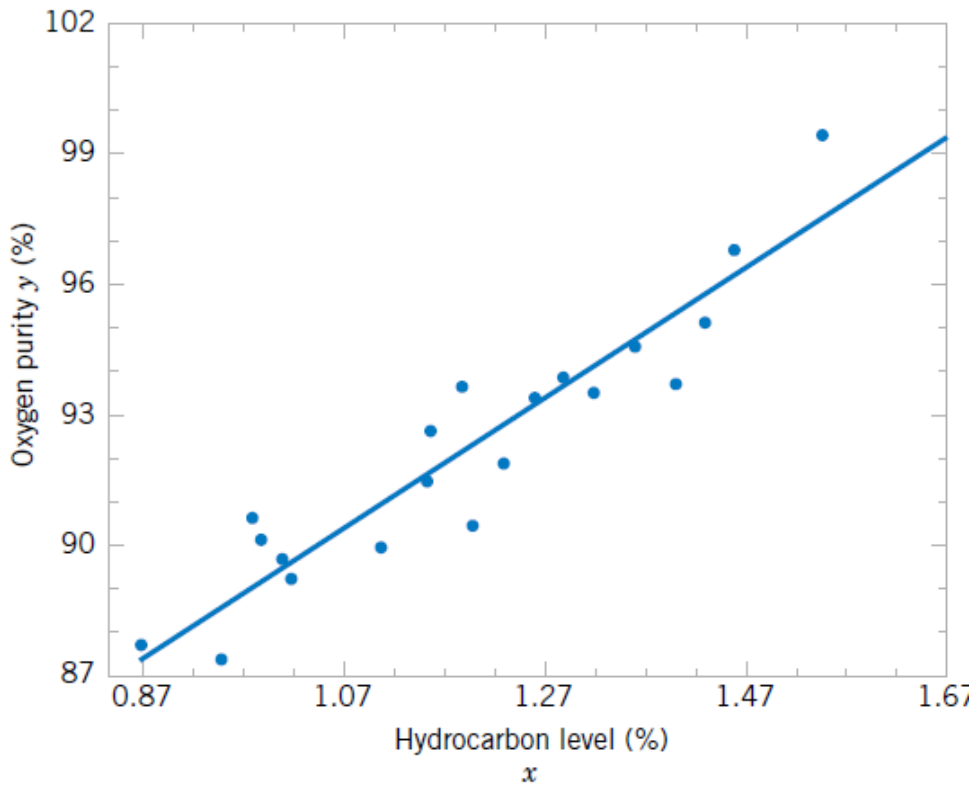


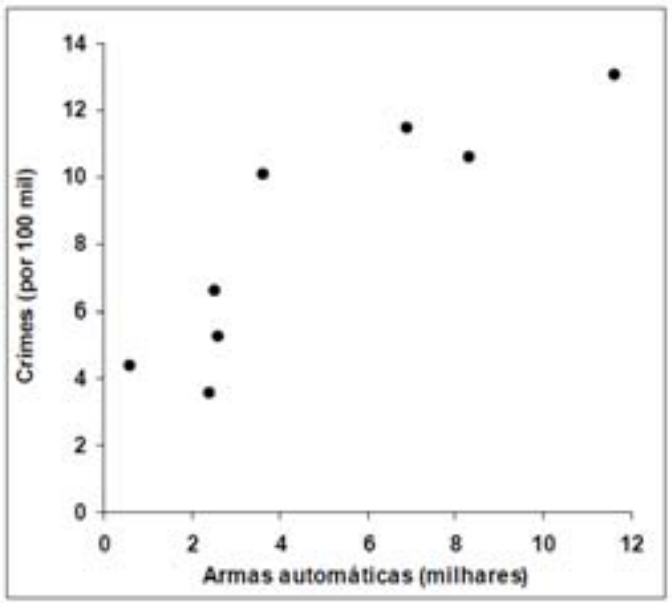
Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.20 + 14.97x$.

Exemplo:

- Vamos considerar as seguintes variáveis:
  - número de armas automáticas registradas;
  - taxa de criminalidade.
- Essas variáveis são observadas em 8 municípios americanos.

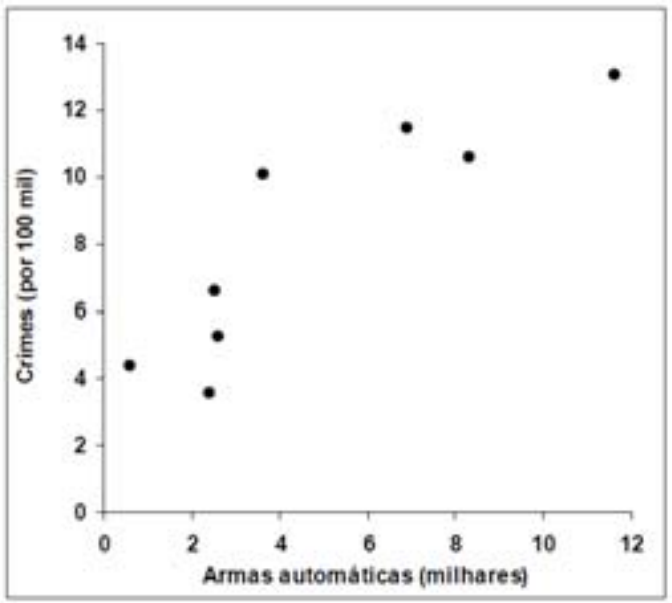
Armas (milhares):	11,6	8,3	3,6	0,6	6,9	2,5	2,4	2,6	$\overline{x} = 4{,}81$
Crimes (100 mil hab):	13,1	10,6	10,1	4,4	11,5	6,6	3,6	5,3	$\overline{y} = 8{,}15$

Exemplo: (continuação)



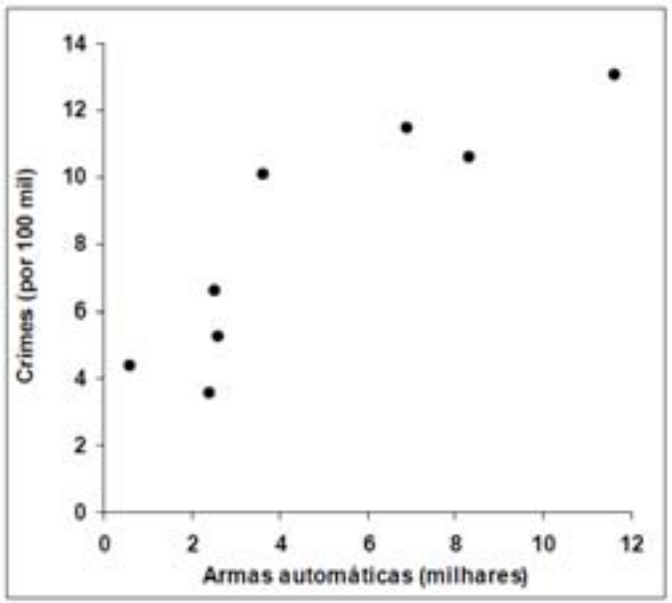
- O gráfico de dispersão é apresentado ao lado.
- Os crimes com armas de fogo parecem estar relacionados com a quantidade de armas automáticas?

Exemplo: (continuação)



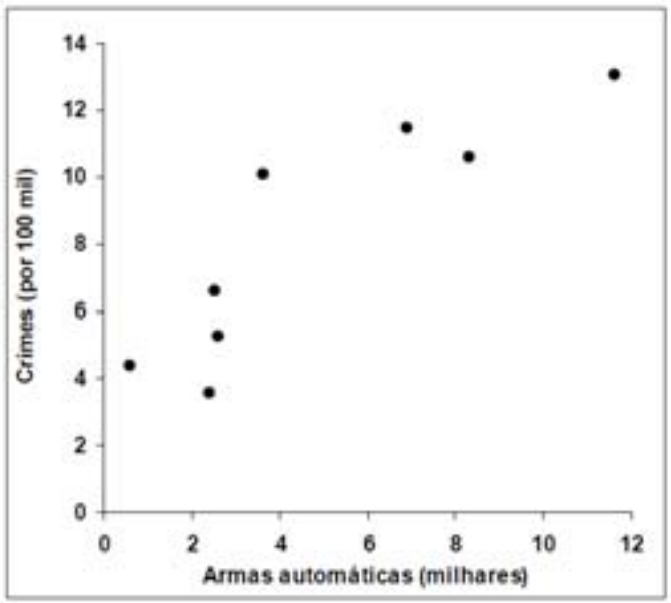
- O gráfico de dispersão é apresentado ao lado.
- Os crimes com armas de fogo parecem estar relacionados com a quantidade de armas automáticas? Sim.
- Esta relação é linear ?

Exemplo: (continuação)



- O gráfico de dispersão é apresentado ao lado.
- Os crimes com armas de fogo parecem estar relacionados com a quantidade de armas automáticas? Sim.
- Esta relação é linear ? Sim.
- É crescente ou decrescente ?

Exemplo: (continuação)



- O gráfico de dispersão é apresentado ao lado.
- Os crimes com armas de fogo parecem estar relacionados com a quantidade de armas automáticas? Sim.
- Esta relação é linear ? Sim.
- É crescente ou decrescente ? Crescente (ou positiva).

Exemplo: (continuação)

- Qual a intensidade ?

Exemplo: (continuação)

- Qual a intensidade ? Forte.

$$r = 0,885$$

- Esse valor é significativo?
- Vamos testar as hipóteses

Exemplo: (continuação)

- Qual a intensidade ? Forte.

$$r = 0,885$$

- Esse valor é significativo?
- Vamos testar as hipóteses

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

- Vimos que a estatística de teste é dada por

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

Exemplo: (continuação)

- Qual a intensidade ? Forte.

$$r = 0,885$$

- Esse valor é significativo?
- Vamos testar as hipóteses

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

- Vimos que a estatística de teste é dada por

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{(0,885)\sqrt{8-2}}{\sqrt{1-(0,885)^2}} = 4.656$$

- Sob H_0 temos que

$$t \sim t_6 \, .$$

- Fixando $\alpha = 0,05$, $t_{0,025;6} =$

Exemplo: (continuação)

- Fixamos o nível de significância em $\alpha = 0,05$.

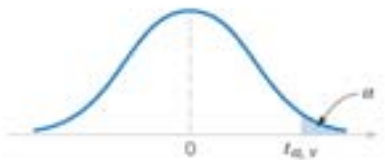


Table IV Percentage Points $t_{\alpha,\nu}$ of the t-Distribution

ν \ α	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959

- Temos então que $t_{0,025;6} =$

Exemplo: (continuação)

- Fixamos o nível de significância em $\alpha = 0,05$.

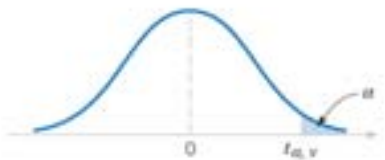


Table IV Percentage Points $t_{\alpha,\nu}$ of the t-Distribution

ν \ α	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959

- Temos então que $t_{0,025;6} = 2,447$.

Exemplo: (continuação)

- A região crítica é dada por

Exemplo: (continuação)

- A região crítica é dada por

$$t < -2,447 \quad \text{ou} \quad t > 2,447 \, .$$

- Qual a conclusã do teste?

Exemplo: (continuação)

- A região crítica é dada por

$$t < -2,447 \quad \text{ou} \quad t > 2,447 \, .$$

- Qual a conclusã do teste?
- Como $t > 2,447$ rejeitamos H_0.
- Com 5% de significância pode-se dizer que existe uma associação significativa entre número de armas automáticas registradas e taxa de criminalidade.

Exemplo: (continuação)

- Vimos então que as variáveis estão **linearmente** associadas.
- Vamos agora ajustar um modelo de regressão linear simples.
- A variável resposta é

Exemplo: (continuação)

- Vimos então que as variáveis estão **linearmente** associadas.
- Vamos agora ajustar um modelo de regressão linear simples.
- A variável resposta é taxa de criminalidade.
- A variável explicativa é

Exemplo: (continuação)

- Vimos então que as variáveis estão **linearmente** associadas.
- Vamos agora ajustar um modelo de regressão linear simples.
- A variável resposta é taxa de criminalidade.
- A variável explicativa é número de armas automáticas registradas.
- O modelo de regressão é dada por

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

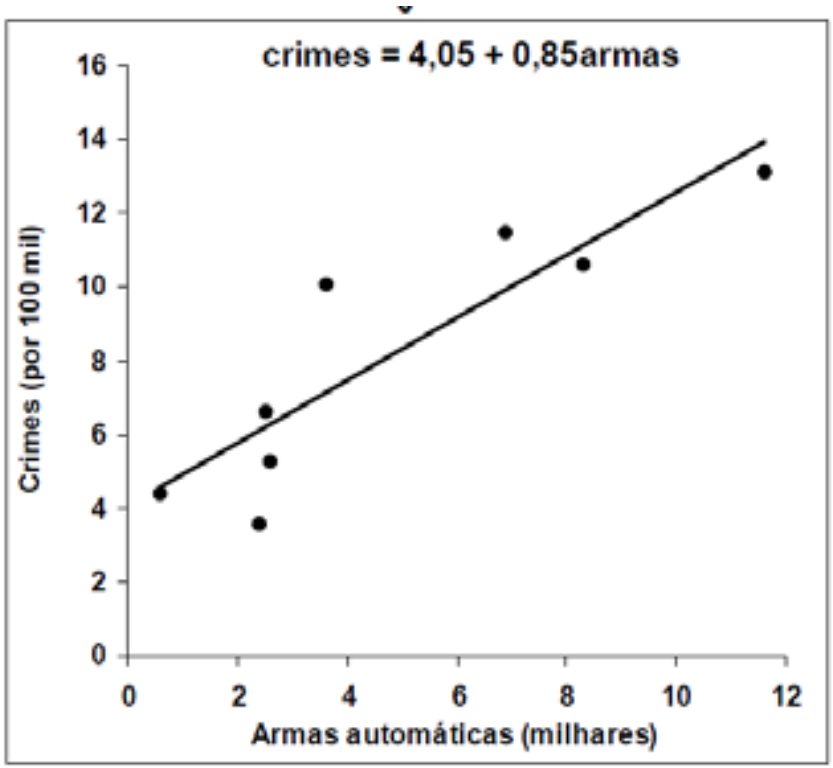
onde $\epsilon \sim^{iid} N(0, \sigma^2)$.

Exemplo: (continuação)

- As estimativas dos parâmetros β_0 e β_1 são dadas por

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0,85 \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 4,05.$$

- A reta ajustada é mostrada logo abaixo.



- O que significa $\beta_0 = 4,05$?

- O que significa $\beta_0 = 4,05$?
- Em um local com nenhuma arma registrada a taxa de criminalidade esperada é de 4,05 crimes por 100 mil habitantes.
- O que significa $\beta_1 = 0,85$?

- O que significa $\beta_0 = 4,05$?
- Em um local com nenhuma arma registrada a taxa de criminalidade esperada é de 4,05 crimes por 100 mil habitantes.
- O que significa $\beta_1 = 0,85$?
- Para cada aumento em uma unidade no número de armas registradas aumenta em 0,85 o número de crimes esperados por 100 mil habitantes.

Estimador de σ^2

- Um outro parâmetro desconhecido do modelo é a variância do erro

$$Var(\epsilon) = \sigma^2 .$$

- Podemos estimá-la usando os resíduos

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

- Soma dos quadrados dos resíduos SQ_E é dada por

$$SQ_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Pode-se mostrar que

$$E(SQ_E) = (n-2)\sigma^2 \quad \Rightarrow \quad \sigma^2 = \frac{E(SQ_E)}{n-2}\,.$$

- Um estimador de σ^2 é dado por

$$\hat{\sigma^2} = \frac{SQ_E}{n-2}$$

- Programas computacionais são utilizados para ajustar modelos de regressão.

Table 11-2 Minitab Output for the Oxygen Purity Data in Example 11-1

Regression Analysis

The regression equation is

Purity = 74.3 + 14.9 HC Level

Predictor	Coef	SE Coef	T	P
Constant	74.283 ← $\hat{\beta}_0$	1.593	46.62	0.000
HC Level	14.947 ← $\hat{\beta}_1$	1.317	11.35	0.000

S = 1.087 R-Sq = 87.7% R-Sq (adj) = 87.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	152.13	152.13	128.86	0.000
Residual Error	18	21.25 ← SS_E	1.18 ← $\hat{\sigma}^2$		
Total	19	173.38			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	89.231	0.354	(88.486, 89.975)	(86.830, 91.632)

Values of Predictors for New Observations

New Obs	HC Level
1	1.00

- Veremos como fazer esse ajuste usando o software R.