Estatística aplicada

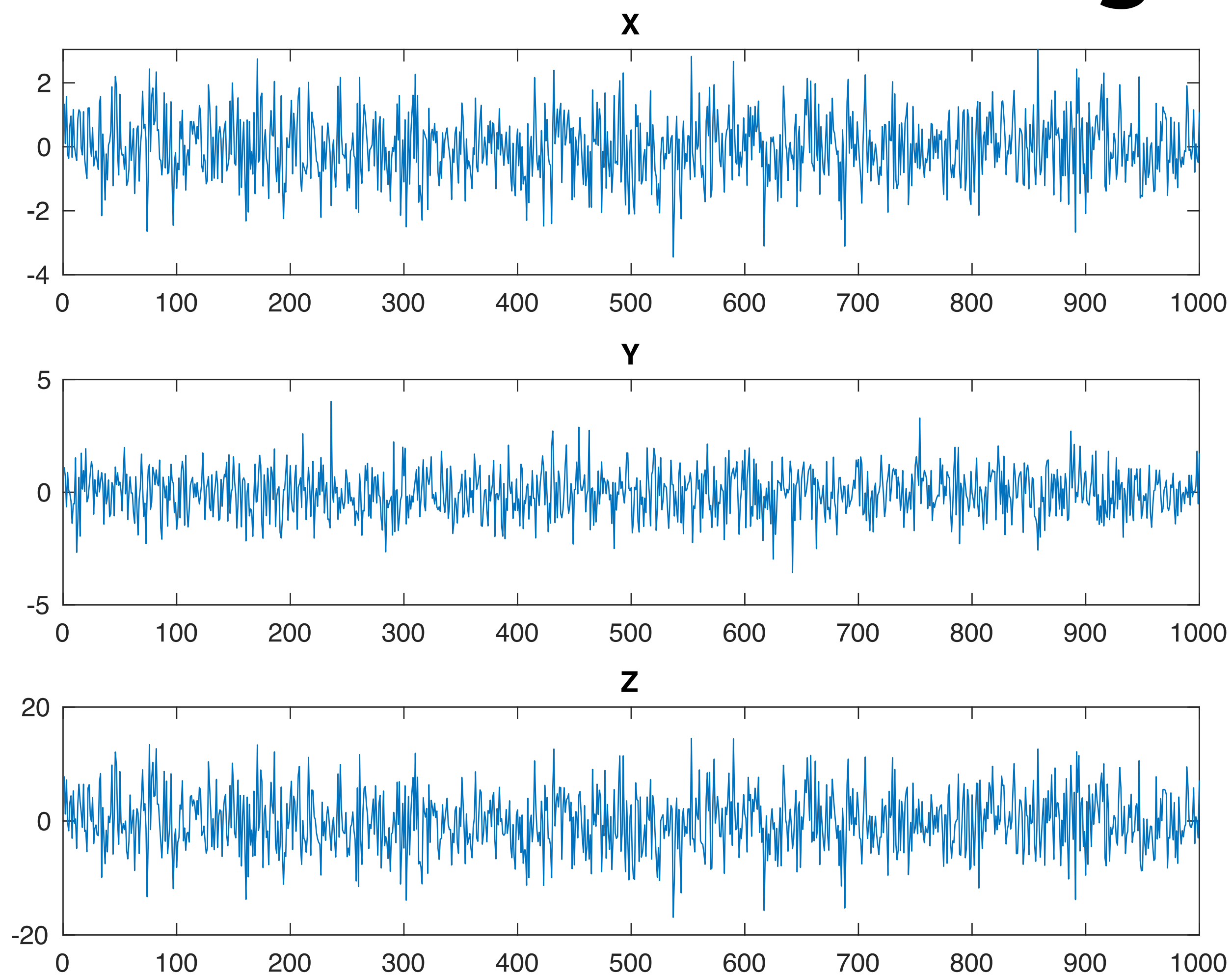
Tópicos especiais em Estatística Aplicada

Prof. Celso J. Munaro (cjmunaro@gmail.com)

IX - Regressão linear e correlação

Cap 11 de [1]

Introdução



Sejam as variáveis X,Y,Z

Busca-se modelos:

$$\hat{y} = b_0 + b_1 x$$

e

$$\hat{z} = b_0 + b_1 x$$

Introdução

No Matlab:
p1 = regress(y,[1 x])
p2 = regress(z,[1 x])

p1 $\hat{y} = -0.0205 - 0.0191x$

p2 $\hat{z} = -0.0205 + 5.0191x$

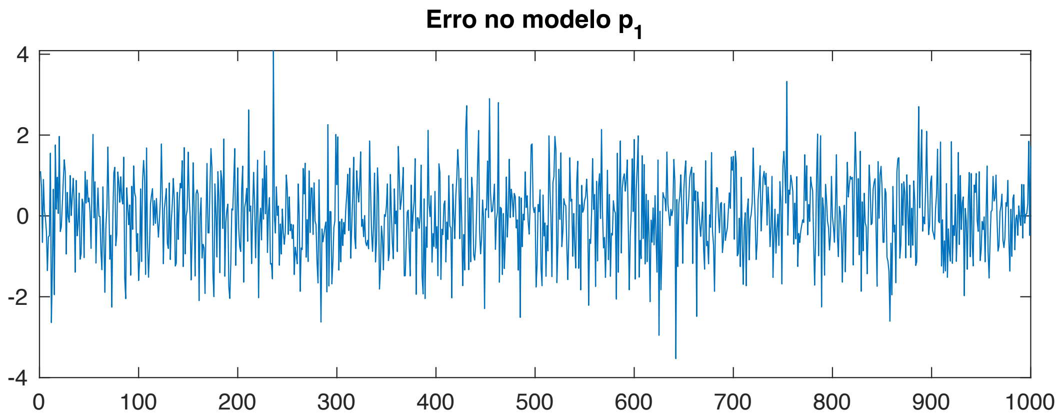
Que conclusões podemos obter destes modelos?

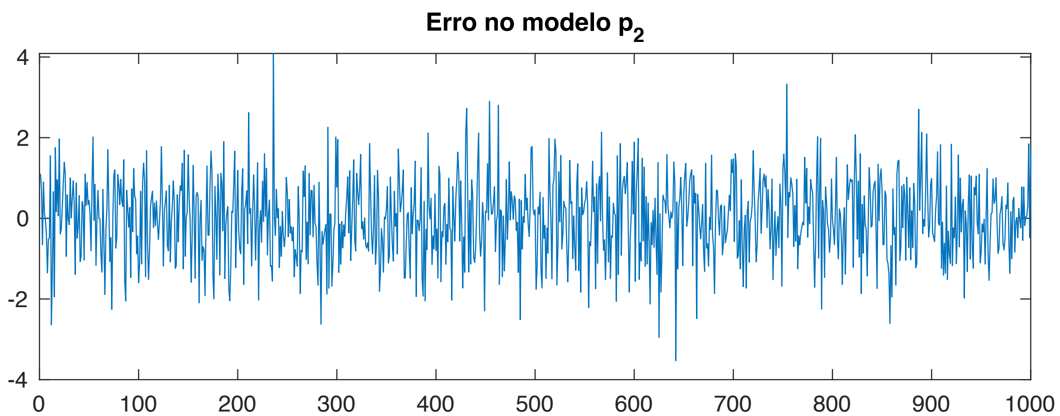
Introdução

$$\hat{y} = -0.0205 - 0.0191$$

$$\hat{z} = -0.0205 + 5.0191x$$

Que conclusões podemos obter destes modelos?





Introdução

Análise das estatísticas da regressão

[b,CI,stats] = regress(y,[o x])

$b = [b_0, b_1]$ **Assume-se que** $b_0 \neq 0$

CI = Confidence interval para b

stats=
R2 statistic,
the *F*-statistic and its
p-value,
estimate of the error variance.

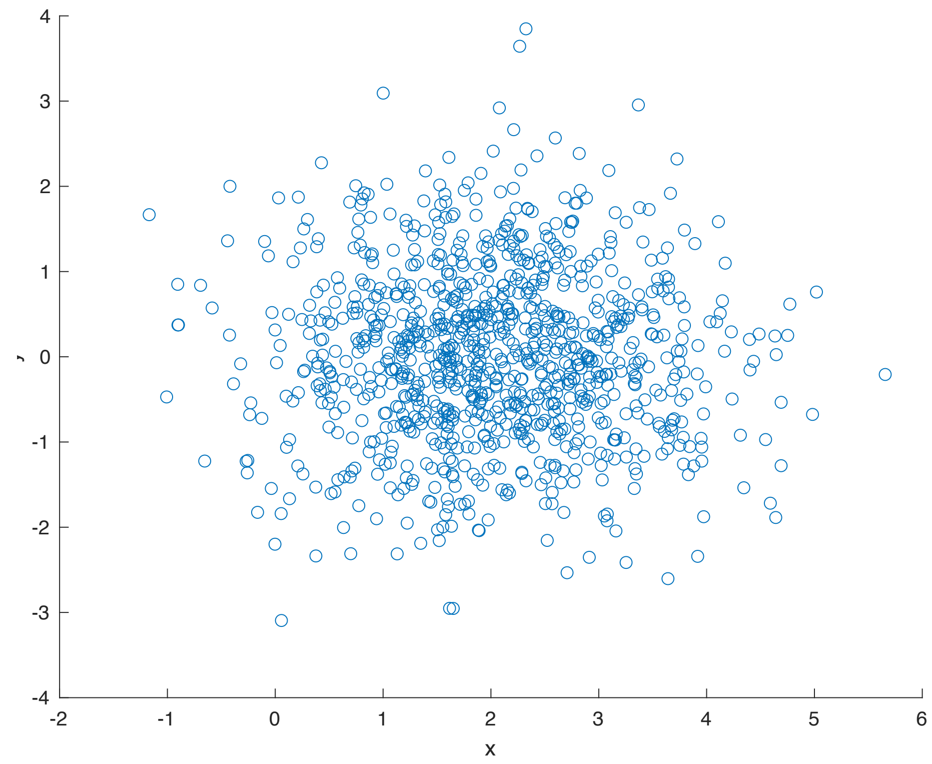
The *F*-test looks for a significant linear regression relationship between the response variable and the predictor variables.

Se $R^2 \approx 1$ e p-value << $(1 - \alpha)$, há evidências de que haja uma regressão linear

Introdução

$\hat{y} = -0.0205 - 0.0191x$

```
b =

   -0.0205
    0.0191


bint =

   -0.0821    0.0412
   -0.0425    0.0808


stats =

    0.0004    0.3699    0.5432    0.9865
```

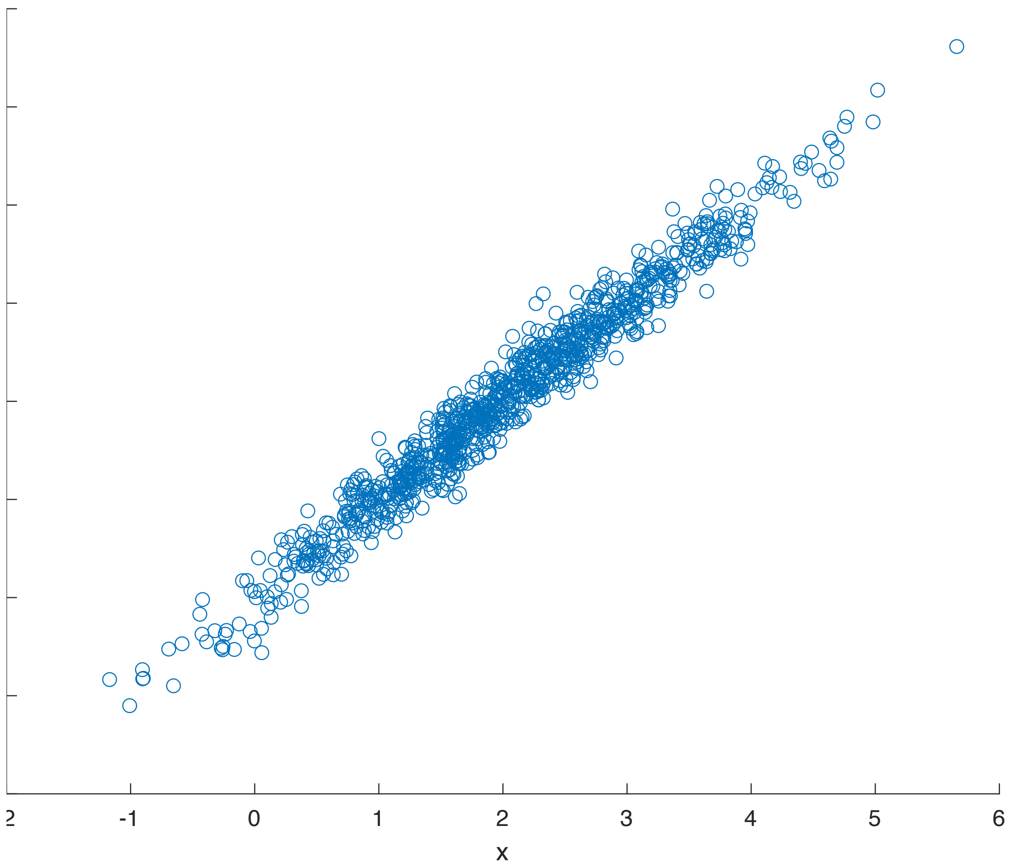


$R^2, F, p - value, \sigma$

Introdução

$\hat{z} = -0.0205 + 5.0191x$

```
b =

   -0.0205
    5.0191


bint =

   -0.0821    0.0412
    4.9575    5.0808


stats =

   1.0e+04 *

    0.0001    2.5514         0    0.0001
```



$R^2, F, p-value, \sigma$

Introdução

Outra alternativa é testar a correlação entre (x,y) e (x,z)

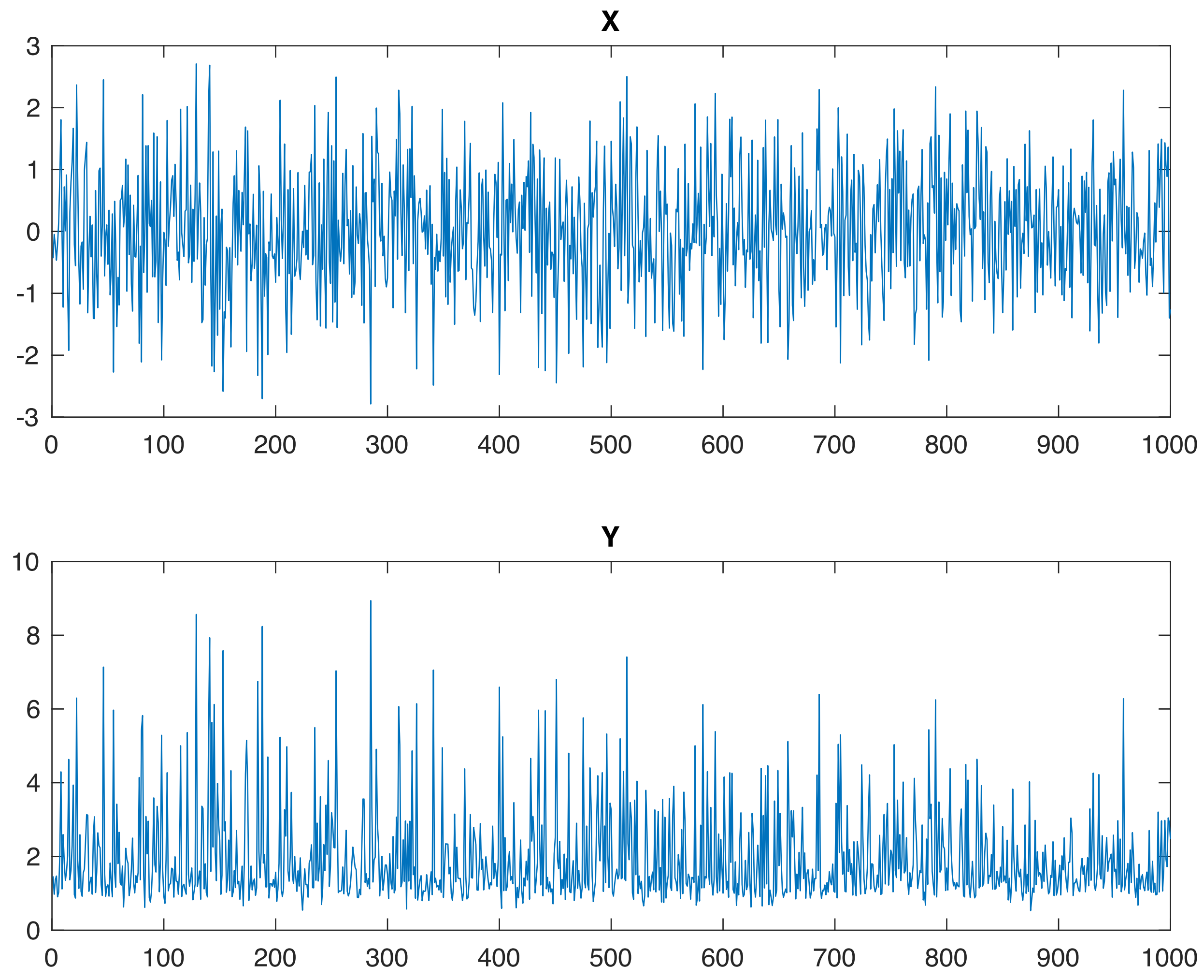
[r, p]=corr(y,x)
r=-0.0047
p=0.8827

[r, p]=corr(z,x)
r=1
p=0

Portanto, os testes estatísticos tem grande importância para decisões sobre a existência ou não de modelos de correlação

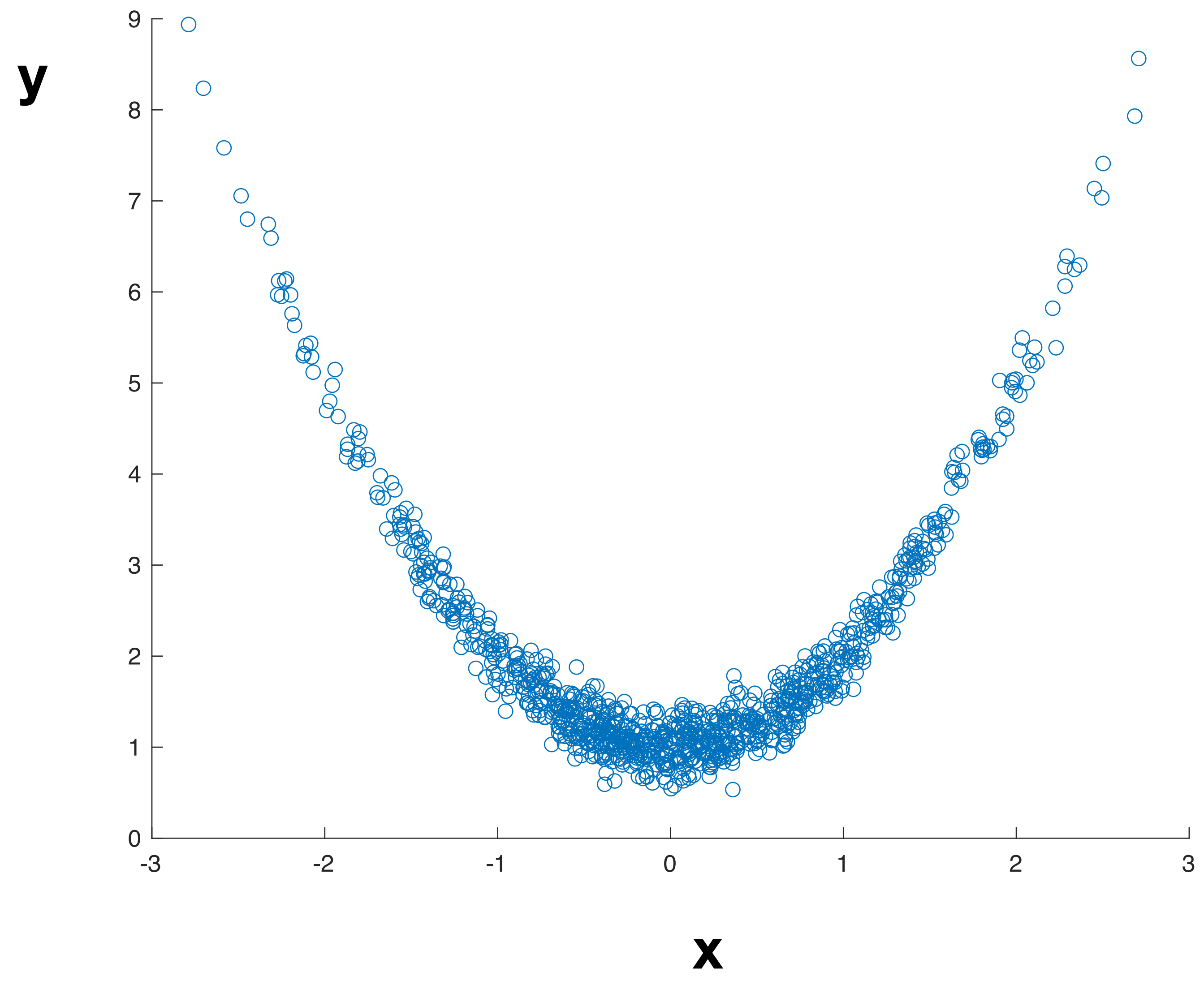
Introdução

Atenção: não linearidade



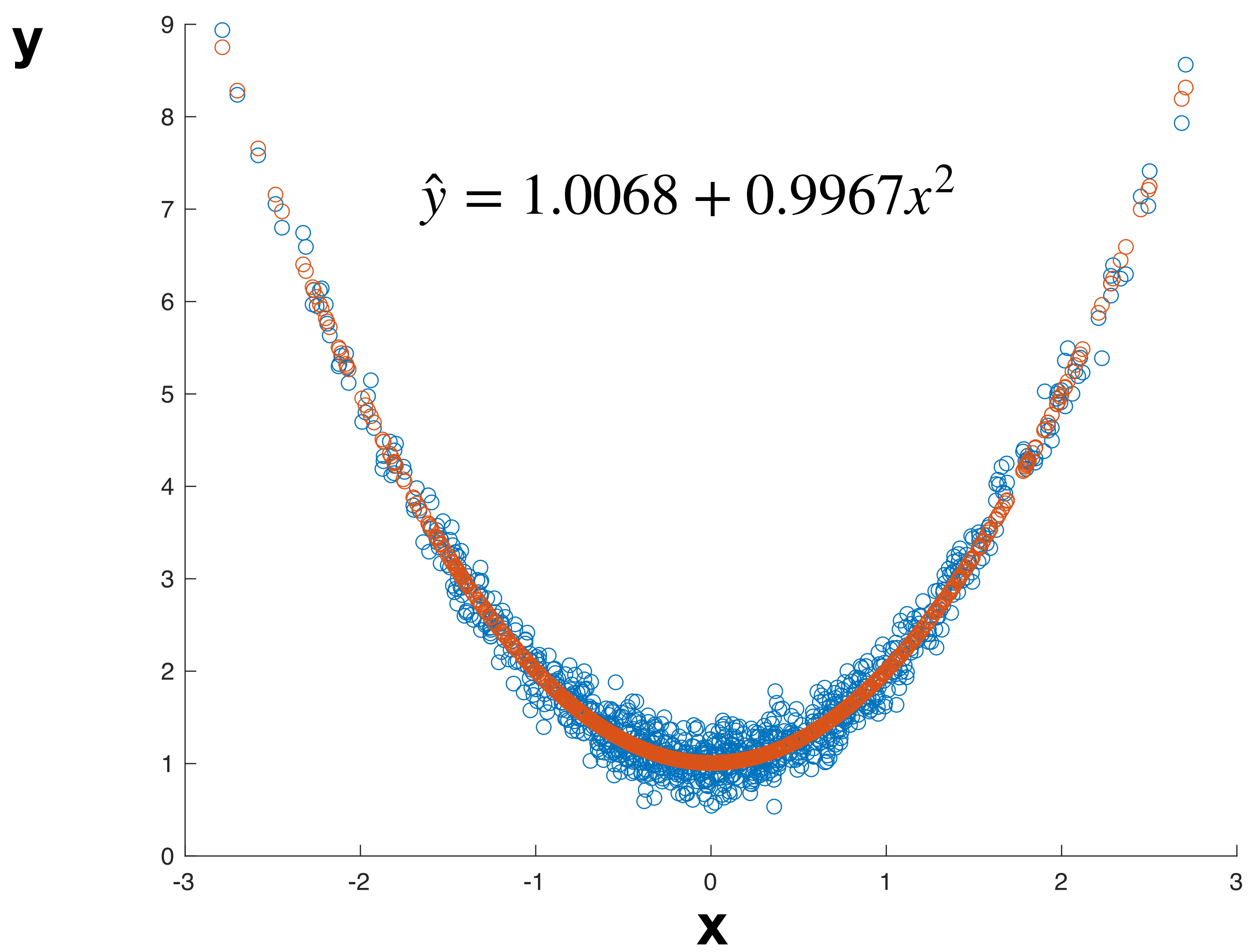
[r, p]=corr(y,x)
r=0.0322
p=0.3091

Introdução



[r, p]=corr(y,x)
r=0.0322
p=0.3091

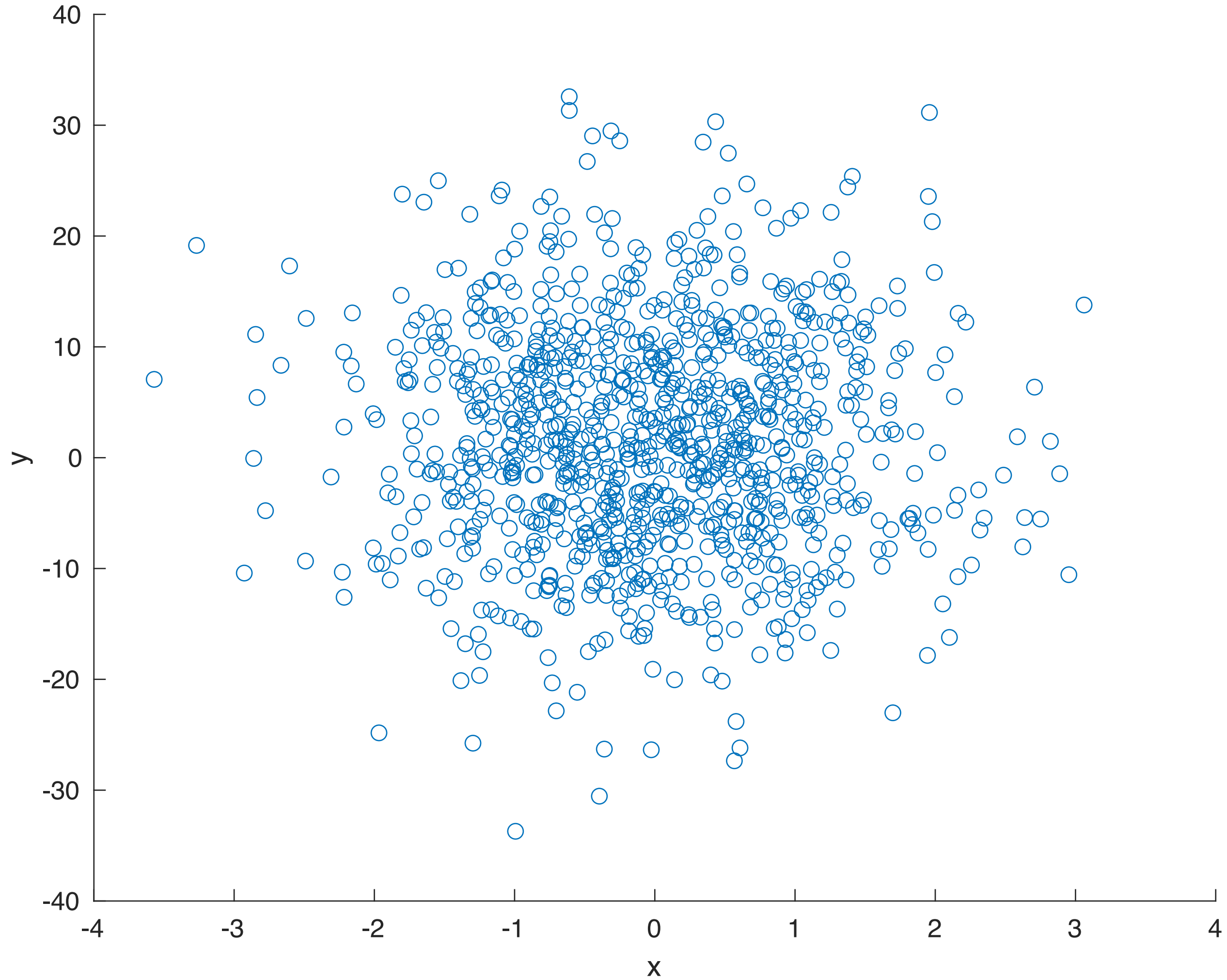
Introdução



[r, p]=corr(y,x)
r=0.0322
p=0.3091

Introdução

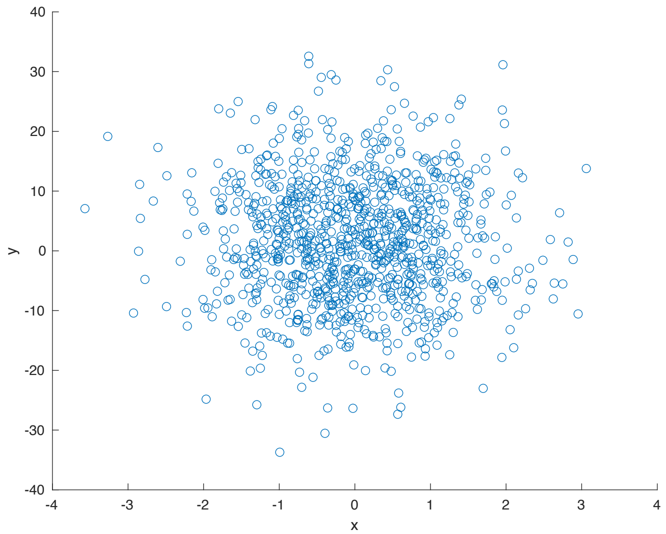
Outro caso: atraso entre x e y



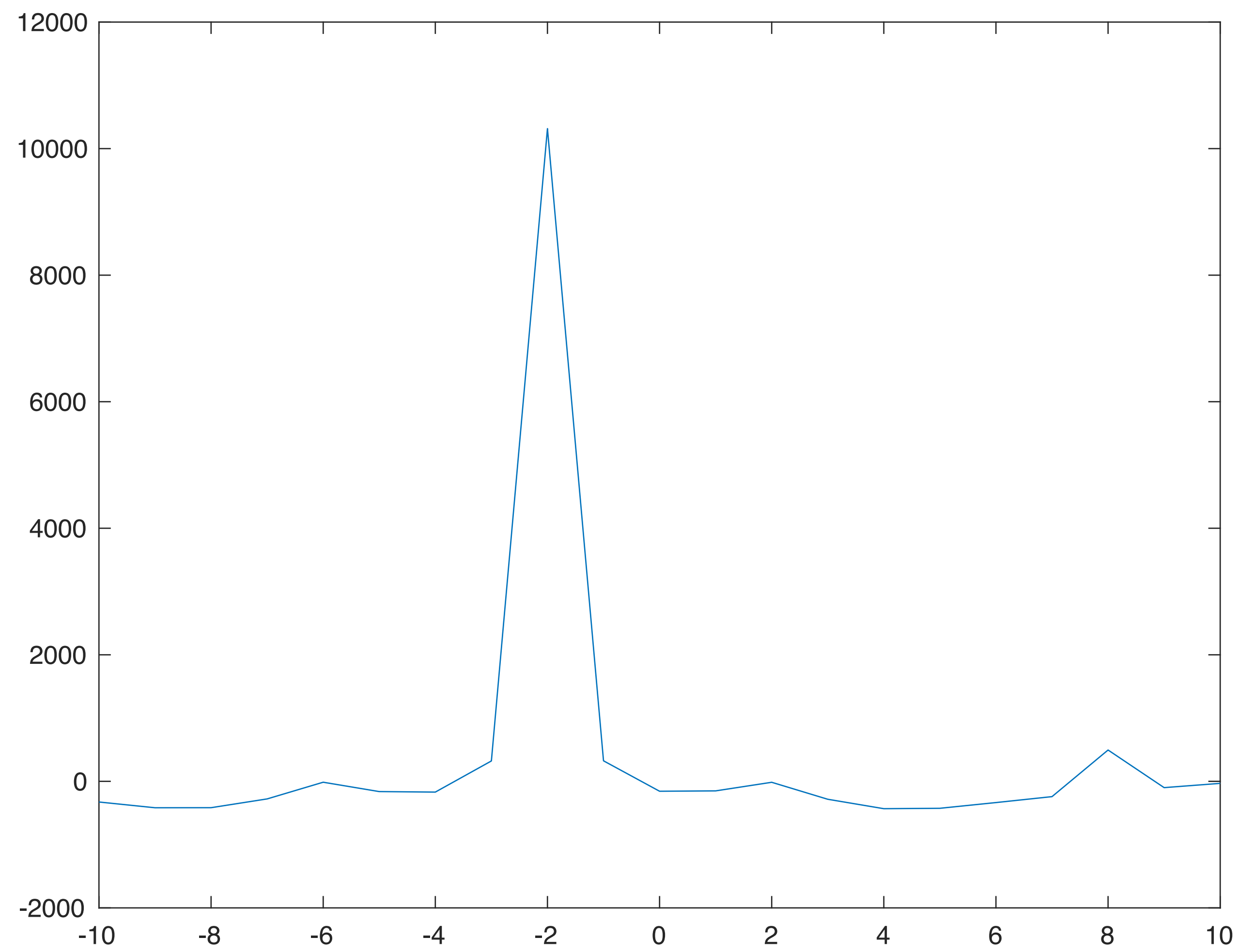
[r, p]=corr(y,x)
r=-0.0093
p=0.7699

Introdução

Outro caso:atraso entre y e x



[r, p]=corr(y,x-2)
r=1
p=0



Introdução

Usar testes de correlação não-linear também:

Kendall, Spearman, são exemplos.

No Matlab,

```
[r, p]=corr(y,x, 'Type','Spearman')
```

Conclusão:
Usar:

- **Testes estatísticos**
- **Investigar atrasos**
- **Análise visual**
- **Testes nos modelos ajustados**

Introdução

Suponha que haja evidências de que os pontos estão distribuídos em torno de uma reta, para um intervalo de valores de x.

Neste caso, é razoável acreditar que os valores médios de Y podem ser previstos a partir dos valores de X através da equação

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

onde a declividade e interseção são os parâmetros da regressão.

Introdução

Enquanto o valor médio de Y seja uma função x, o valor observado de y não cai em geral sobre essa reta.

Uma forma adequada para generalizar é supor um modelo probabilístico, em qual o valor médio de Y seja uma função de x mais um termo aleatório,

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

Podemos considerar este modelo como empírico

Introdução

- Lembrando que:

Valor Esperado

-
  - $E[c] = c$
  - $E[cX] = cE[X]$
  - $E[X + Y] = E[X] + E[Y]$
  - $E[XY] = E[X]E[Y]$
  - $E[aX + b] = aE[X] + b$

Variância

- Definição: $V(X) = E[X^2] - E[X]^2 = E[X - E[X]]^2$
  - $V(c) = 0$
  - $V(cX) = c^2 V(X)$
  - $V(X + c) = V(X)$
  - $V(X + Y) = V(X) + V(Y)$ se X e Y forem independentes
  - $V(X - Y) = V(X) + V(Y)$ se X e Y forem independentes
  - $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$
  - $V(X - Y) = V(X) + V(Y) - 2Cov(X, Y)$
  - $V(\sum_i X_i) = \sum_i V(X_i) + 2 \sum \sum_{i<j} Cov(X_i, X_j)$
  - $V(X) = Cov(X, X)$

-

Introdução

A componente aleatória ϵ determina as propriedades de Y.

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

$$V(Y/x) = V(\beta_0 + \beta_1 x + \epsilon) = V(\beta_0 + \beta_1 x) + V(\epsilon) = 0 + V(\epsilon) = \sigma^2$$

A variabilidade de Y para um valor de x é determinada pela variância σ^2

Introdução

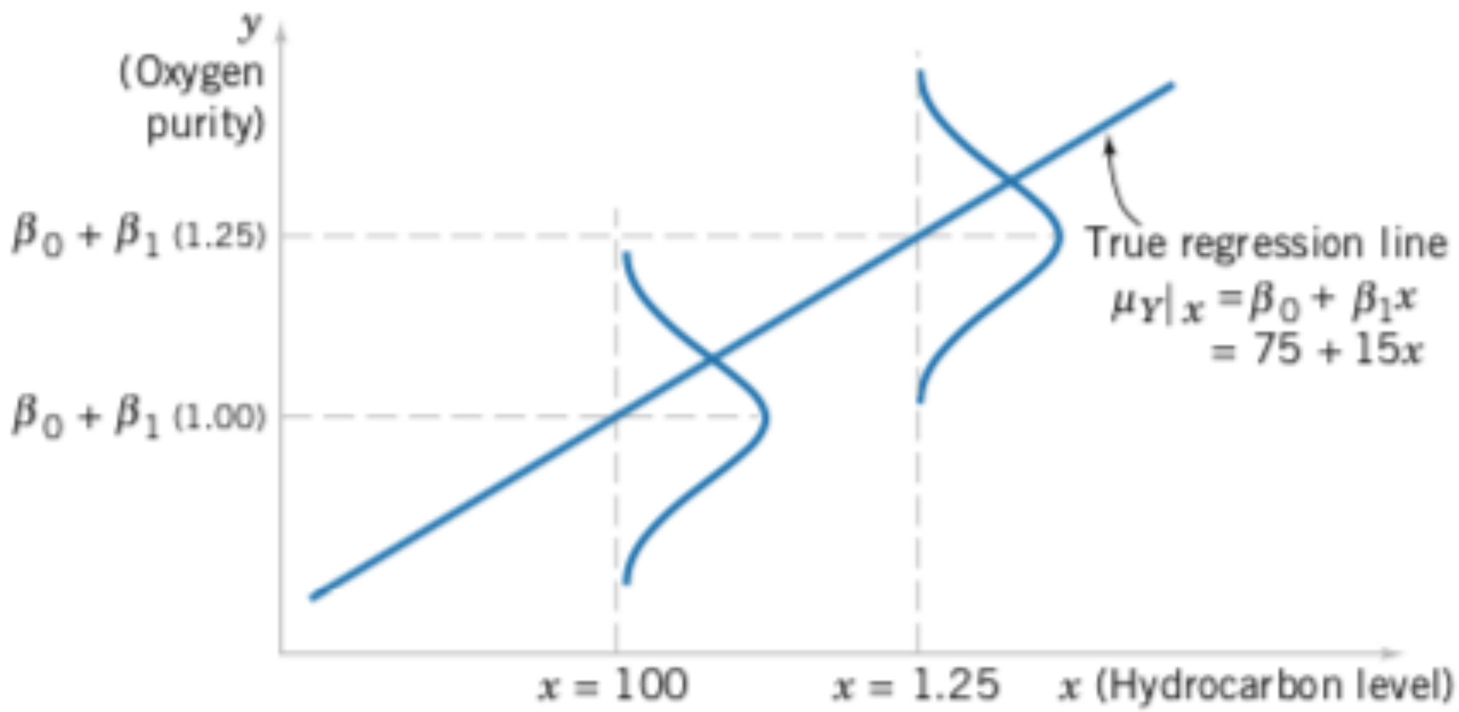


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity-hydrocarbon data.

Regressão linear simples

$E(Y/x) = \beta_0 + \beta_1 x$

Assumimos que cada nova observação Y pode ser dada por

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

onde ϵ é um erro aleatório com média zero e variância σ^2. Assume-se que os erros de diferentes observações não são correlacionados.

Regressão linear simples

Como obter os parâmetros β_0 e β_1 do modelo?

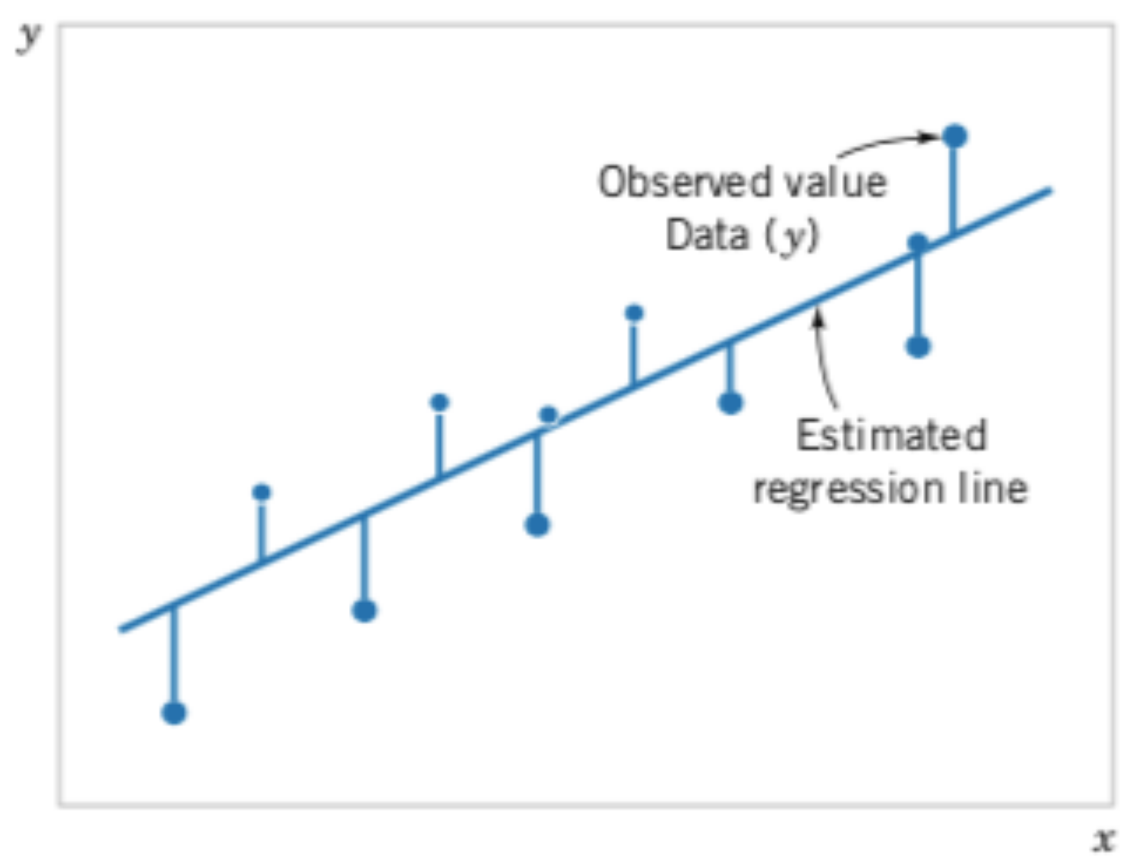


Figure 11-3 Deviations of the data from the estimated regression model.

Regressão linear simples

Os parâmetros β_0 e β_1 do modelo são obtidos pelo método dos mínimos quadrados.

As n observações podem ser escritas por

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1,2,...,n$$

Os desvios das observações em relação ao modelo são dadas por

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

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

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

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

Regressão linear simples

Simplificando,

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

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

Essas duas equações fornecem os dois estimadores de mínimos quadrados, $\hat{\beta}_0, \hat{\beta}_1$.

Regressão linear simples

Definição

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11-8}$$

where $\bar{y} = (1/n) \sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$.

Regressão linear simples

.

The **fitted** or **estimated regression line** is therefore

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

Note that each pair of observations satisfies the relationship

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \qquad i = 1, 2, \ldots, n$$

Resíduo: $e_i = y_i - \hat{y}_i$

Regressão linear simples

Símbolos para a equação (11.8):

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{11-10}$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \tag{11-11}$$

Regressão linear simples

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

Exemplo 11.1

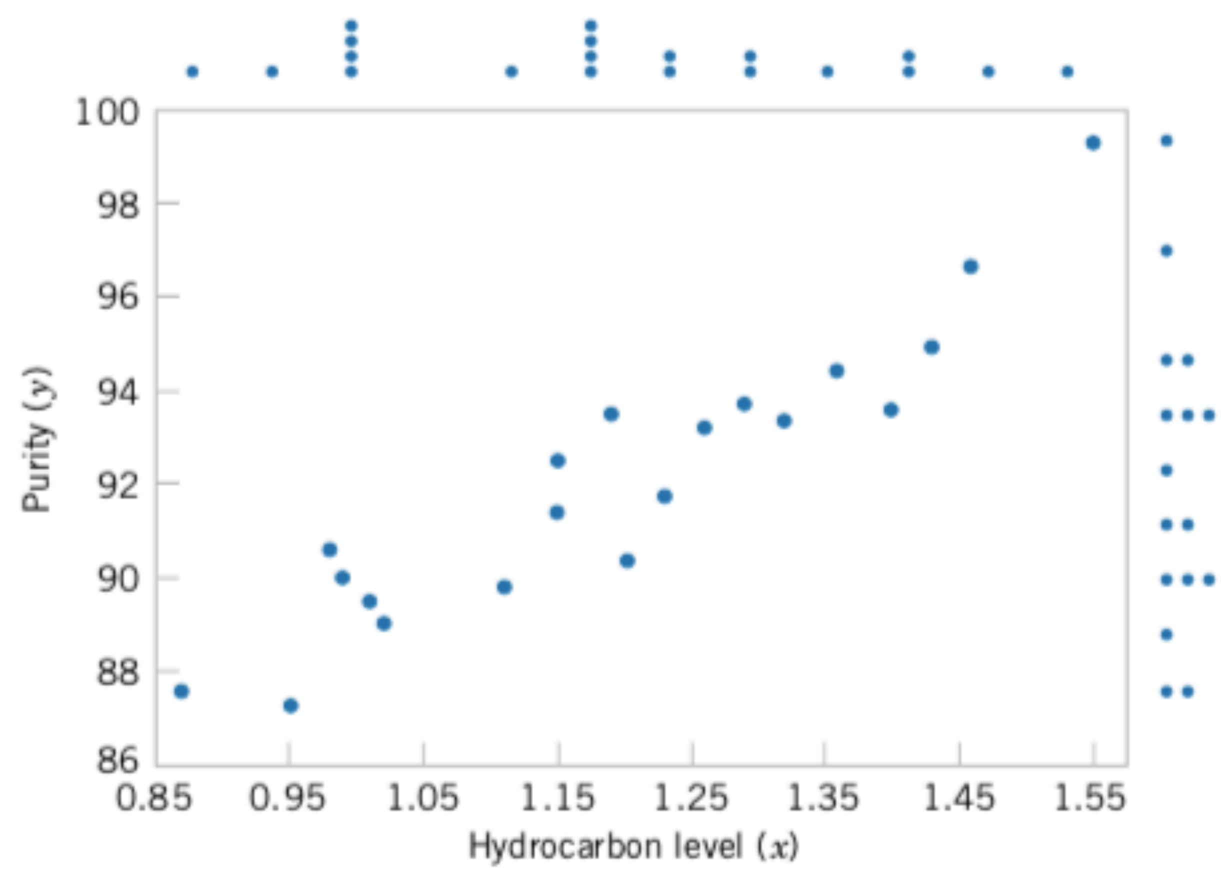


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Regressão linear simples

Exemplo 11.1

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21 \quad \bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892 \quad \sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} = 0.68088$$

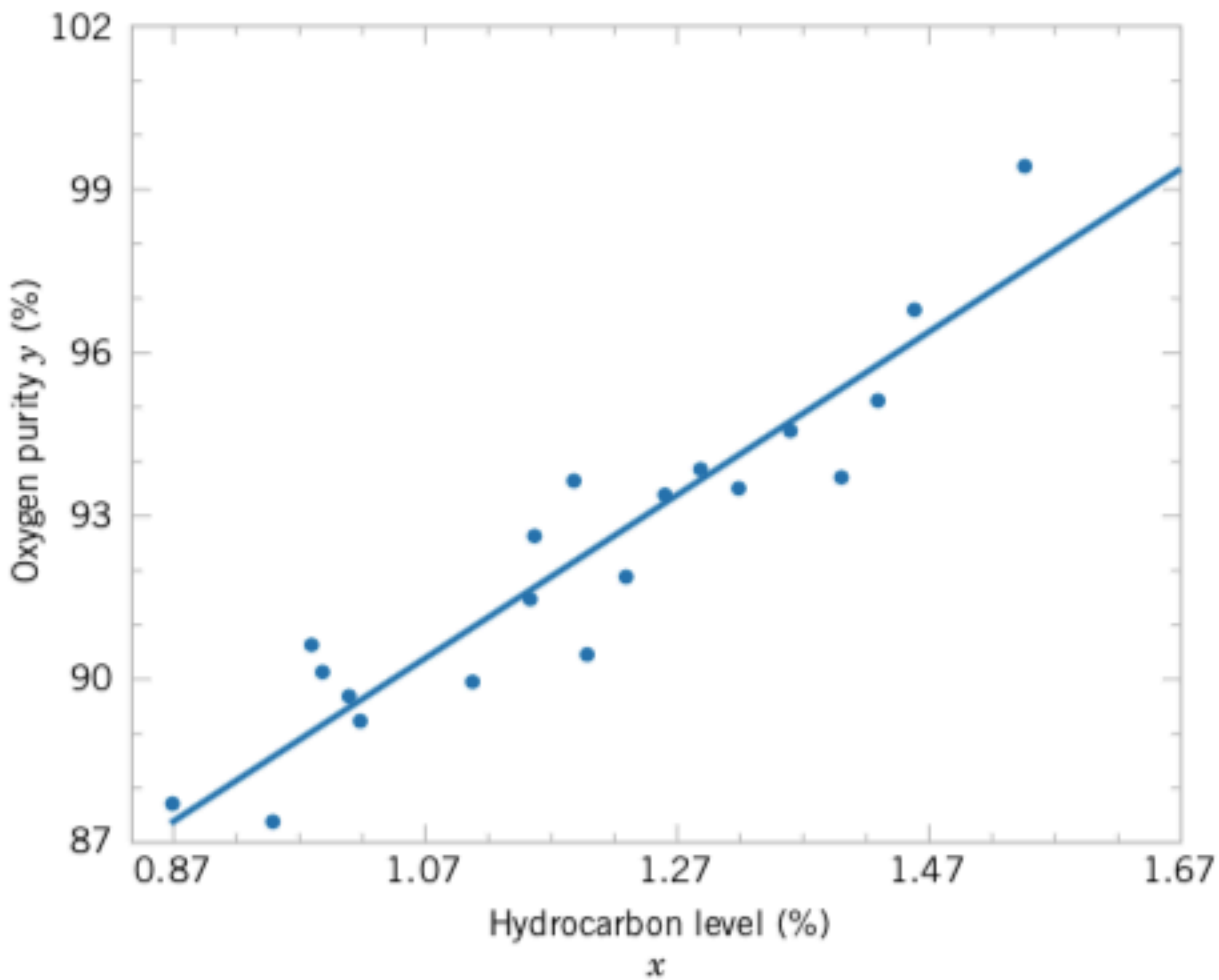
and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 2{,}214.6566 - \frac{(23.92)(1{,}843.21)}{20} = 10.17744$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331 \qquad \hat{y} = 74.283 + 14.947x$$

Regressão linear simples



Exemplo 11.1

Regressão linear simples

Estimativa da variância

Exemplo 11.1

There is actually another unknown parameter in our regression model, σ^2 (the variance of the error term ϵ). The residuals $e_i = y_i - \hat{y}_i$ are used to obtain an estimate of σ^2. The sum of squares of the residuals, often called the **error sum of squares,** is

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (11\text{-}12)$$

We can show that the expected value of the error sum of squares is $E(SS_E) = (n-2)\sigma^2$. Therefore an **unbiased estimator** of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad (11\text{-}13)$$

Regressão linear simples

Estimativa da variância

Exemplo 11.1

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (11\text{-}12)$$

Computing SS_E using Equation 11-12 would be fairly tedious. A more convenient computing formula can be obtained by substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into Equation 11-12 and simplifying.

The resulting computing formula is

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \qquad (11\text{-}14)$$

where $SS_T = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ is the **total sum of squares of the response variable** y. The error sum of squares and the estimate of σ^2 for the oxygen purity data, $\hat{\sigma}^2 = 1.18$, are highlighted in the Minitab output in Table 11-2.

Propriedades do estimador de mínimos quadrados

Os valores dos estimadores β_1 e β_0 dependem dos valores observados y

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

Portanto, os estimadores por mínimos quadrados da regressão linear podem ser considerados variáveis aleatórias.

Podemos investigar a polarização (bias) e a variância de $\hat{\beta}_0$ e $\hat{\beta}_1$

Propriedades do estimador de mínimos quadrados

A propriedade da não polarização garante que os valores estimados do parâmetro estão em torno do valor verdadeiro.

Espera-se que um estimador seja não polarizado e tenha pequena variância.

Propriedades do estimador de mínimos quadrados

$E(\hat{\beta}_1) = \beta_1$

Portanto, $\hat{\beta}_1$ é um estimador não polarizado da inclinação β_1

Como $V(\epsilon_1) = \sigma^2$, segue que $V(Y_i) = \sigma^2$ e

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Propriedades do estimador de mínimos quadrados

Para $\hat{\beta}_0$

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

Portanto, $\hat{\beta}_0$ é um estimador não polarizado de β_0.

Por fim,

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \bar{x} / S_{xx}.$$

Propriedades do estimador de mínimos quadrados

Definição:

In simple linear regression the **estimated standard error of the slope** and the **estimated standard error of the intercept** are

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

respectively, where $\hat{\sigma}^2$ is computed from Equation 11-13.

Propriedades do estimador de mínimos quadrados

Conclusões:

- $\hat{\beta}_1$ is normally distributed with mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$;
- $\hat{\beta}_0$ is normally distributed with mean β_0 and variance $\sigma^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}$;
- $\frac{(n-2)S^2}{\sigma^2}$ has a χ^2 distribution with $n-2$ degrees of freedom;
- S^2 is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Teste de hipóteses em regressão linear

Os resíduos tem distribuição normal e são independentemente distribuídos (normally and independently distributed-NID)

ϵ_i são $NID(0,\sigma^2)$

As observações Y_i são $NID(\beta_0 + \beta_1 x_i, \sigma^2)$

$\hat{\beta}_1 = NID(\beta_1, \dfrac{\sigma^2}{S_{xx}})$ uma combinação linear de variáveis aleatórias

$(n-2)\hat{\sigma}^2/\sigma^2$ tem distribuição chi-quadrado com n-2 graus de liberdade

Teste de hipóteses em regressão linear

A estatística

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \qquad\qquad T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

segue uma distribuição t com n-2 graus de liberdade

A hipótese $H_0 : \beta_1 = \beta_{1,0}$ seria rejeitada se

$$|t_0| > t_{\alpha/2,n-2}$$

Teste de hipóteses em regressão linear

A similar procedure can be used to test hypotheses about the intercept. To test

$$H_0\colon \beta_0 = \beta_{0,0}$$
$$H_1\colon \beta_0 \neq \beta_{0,0} \tag{11-21}$$

we would use the statistic

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)} \tag{11-22}$$

and reject the null hypothesis if the computed value of this test statistic, t_0, is such that $|t_0| > t_{\alpha/2,n-2}$. Note that the denominator of the test statistic in Equation 11-22 is just the standard error of the intercept.

Teste de hipóteses em regressão linear

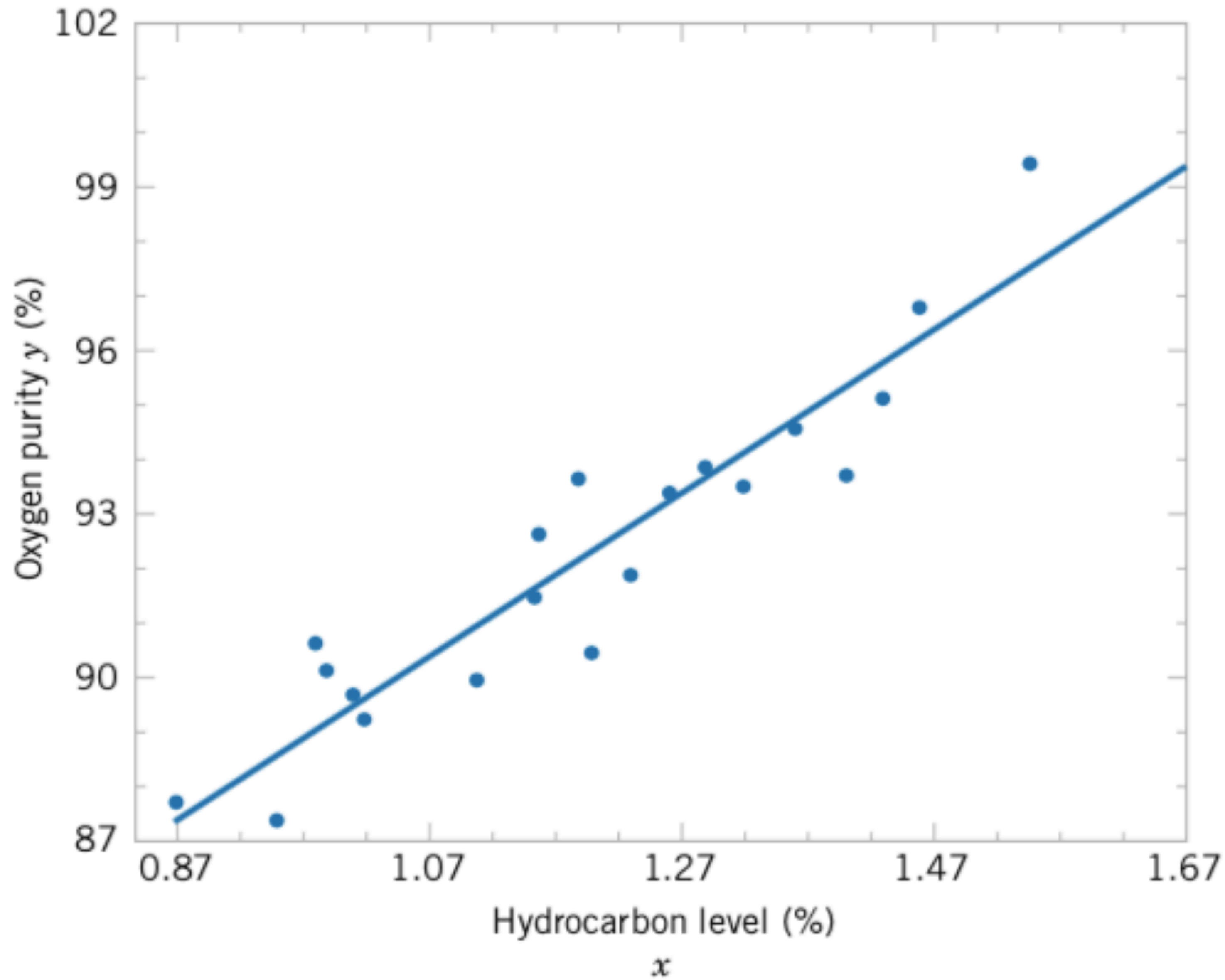
Significância da regressão

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

A falha em rejeitar $H_0 : \beta_1 = 0$ é equivalente a concluir que não há regressão entre x e y. Ou seja, x e y não estão relacionados, pelo menos pelo modelo de regressão.

Teste de hipóteses em regressão linear

Exemplo 11-2



Teste de hipóteses em regressão linear

Exemplo 11-2

We will test for significance of regression using the model for the oxygen purity data from Example 11-1. The hypotheses are

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0$$

and we will use $\alpha = 0.01$. From Example 11-1 and Table 11-2 we have

$$\hat{\beta}_1 = 14.97 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

so the t-statistic in Equation 10-20 becomes

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{14.947}{\sqrt{1.18/0.68088}} = 11.35$$

$t_{0.005,18} = 2.88$ **A hipotese H_0 deve ser rejeitada**

IC dos coeficientes de regressão linear

Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ confidence interval **on the slope** β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{11-29}$$

Similarly, a $100(1-\alpha)\%$ confidence interval **on the intercept** β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \tag{11-30}$$

IC dos coeficientes de regressão linear

Exemplo 11-1

We will find a 95% confidence interval on the slope of the regression line using the data in Example 11-1. Recall that $\hat{\beta}_1 = 14.947$, $S_{xx} = 0.68088$, and $\hat{\sigma}^2 = 1.18$ (see Table 11-2). Then, from Equation 10-31 we find

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

or

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \leq \beta_1 \leq 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

This simplifies to

$$12.197 \leq \beta_1 \leq 17.697$$

Análise de resíduos

É importante checar se a suposição sobre os resíduos é válida, de que são independentes com distribuição normal e variância constante.

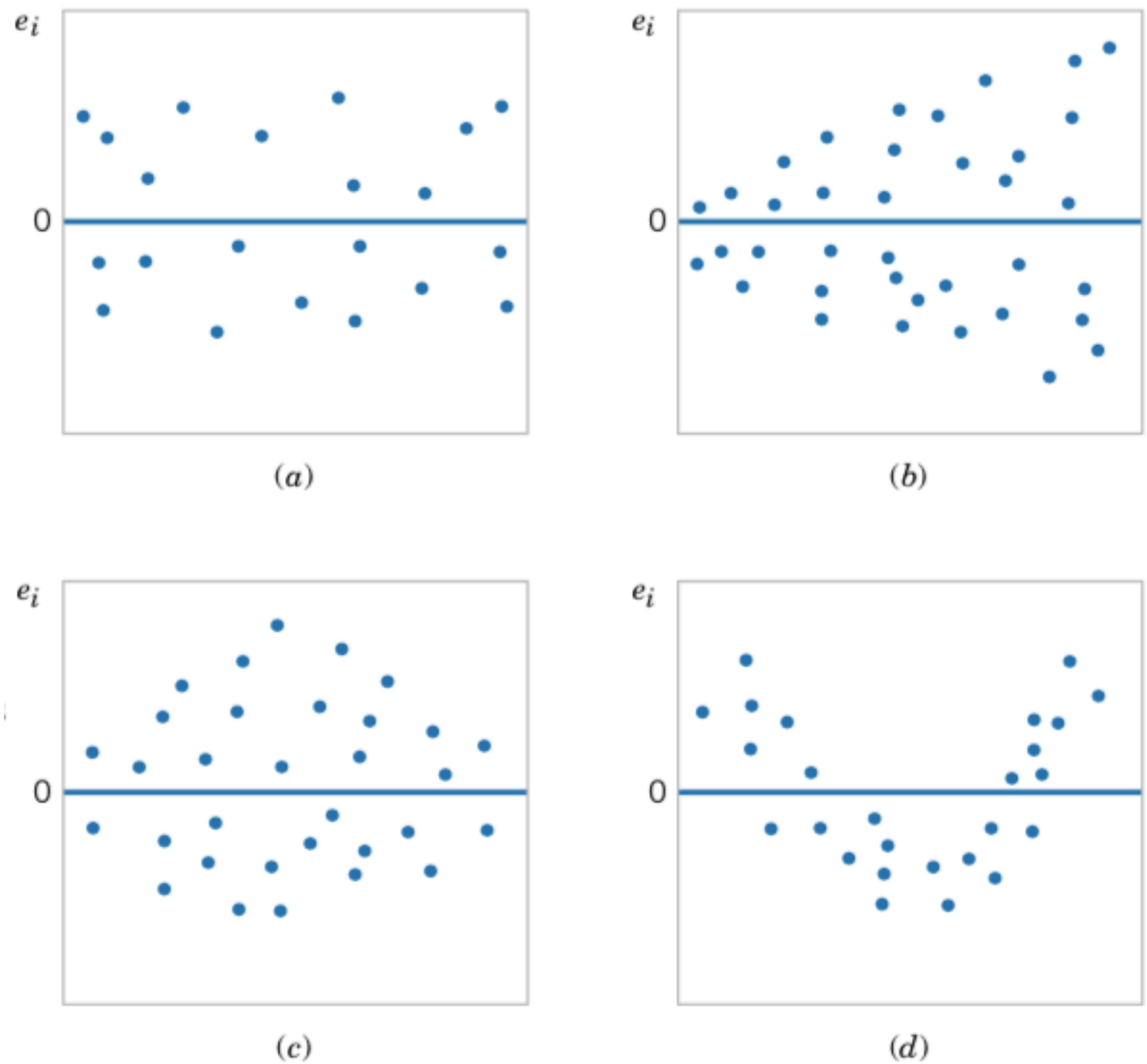
```
[c,lags]=xcorr(e)
```

Verificar a normalidade dos resíduos: normplot, histograma, valores dentro do intervalo de confiança

Também pode-se plotar os resíduos versus as observações y e a variável independe x

```
[c,lags]=xcorr(y,e)
```

Análise de resíduos



Coeficiente de regressão

A quantidade

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

é chamada coeficiente de regressão, sendo usada normalmente para verificar a qualidade do modelo

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

$$SS_T = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

R^2 : quanto mais próximo de 1, melhor

Coeficiente de regressão

Mau uso do coeficiente de regressão

Uso y Mal-Uso del Coeficiente de Regresión R^2 en Modelos de Correlación y Predicción

José O. Valderrama[1, 2,] y Richard A. Campusano [2,3]
(1) Univ. de La Serena, Dpto. de Ingeniería Mecánica, Casilla 554, La Serena-Chile
(2) Centro de Información Tecnológica, Casilla 724, La Serena – Chile
(3) Depto. de Física, Univ. de La Serena, Casilla 554, La Serena-Chile

Coeficiente de regressão

Mau uso do coeficiente de regressão

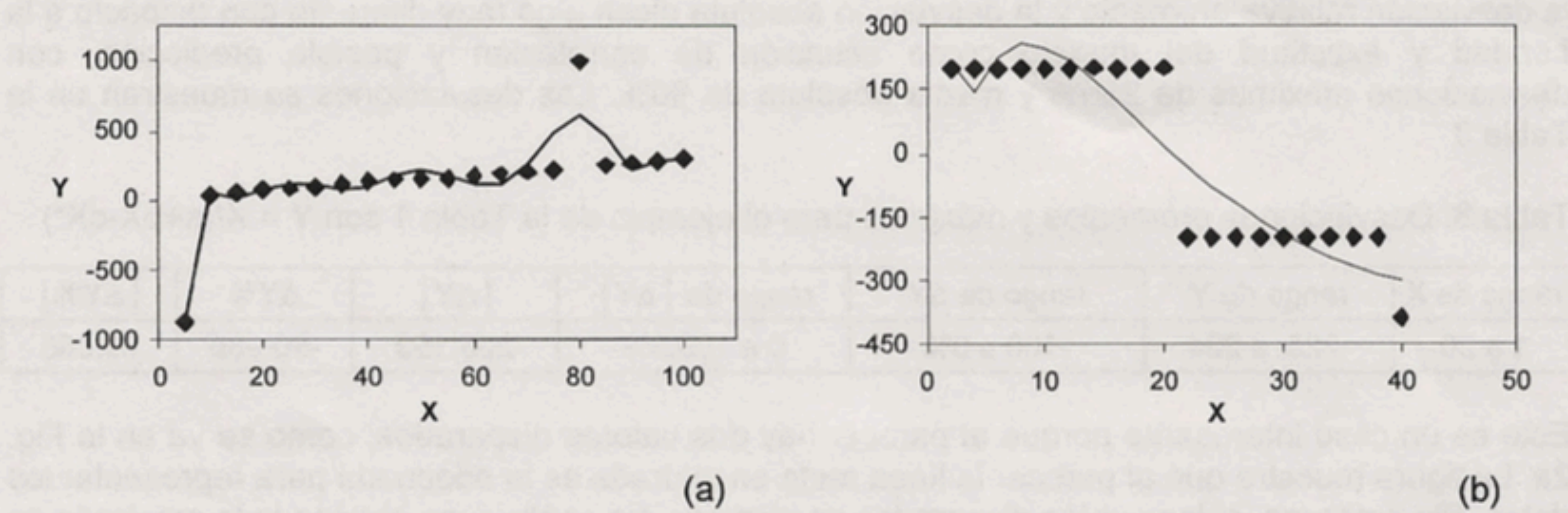


Fig. 1: Conjuntos de datos muy distintos con similar R^2: (a) R^2=0,8433 y (b) R^2=0,8433

Predição de novas observações

Uma importante aplicação da regressão linear é prever futuros valores Y para valores dados de x. Se x_0 é o valor da variável de regressão, então o valor futuro de Y_0 é dado pelo estimador pontual

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Predição de novas observações

Seja agora o problema de estimar um intervalo para futuros valores de Y. Essa nova observação é independente das observações usadas para construir o modelo de regressão.

O intervalo de confiança já calculado $\mu_{Y|x_0}$ não é apropriado pois é calculado usando os dados de regressão. Esse IC se refere ao valor real da resposta para $x = x_0$, que é um parâmetro da população, e não um valor futuro.

Predição de novas observações

Seja Y_0 o valor observado futuro para $x = x_0$ e seja Y_0 dado por $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

O erro de predição $e_p = Y_0 - \hat{Y}_0$ é uma variável aleatória com distribuição normal, média zero e variância

$$V(e_{\hat{p}}) = V(Y_0 - \hat{Y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

pois Y_0 é independente de $\hat{Y}_0$

Predição de novas observações

Usando $\hat{\sigma}^2$ para estimar σ, pode-se mostrar que

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}}$$

tem distribuição t com n-2 graus de liberdade.

Predição de novas observações

Definição:

A $100(1 - \alpha)$ % **prediction interval on a future observation** Y_0 at the value x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

$$\leq Y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \qquad (11\text{-}33)$$

The value $\hat{y}_0$ is computed from the regression model $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Observe que o intervalo é mínimo para $x = \bar{x}$

Predição de novas observações

O erro de predição depende do erro do modelo de regressão e do erro dos futuros valores observados.

Ver Exemplo 11.6