Unidade 3. Regressão linear simples

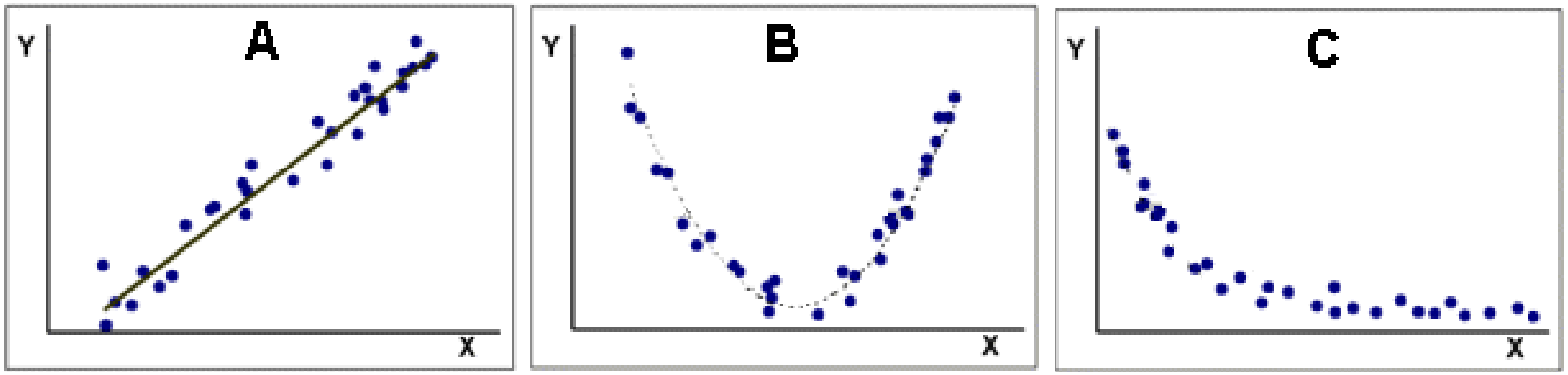
3.1. Introdução e modelo estatístico

3.2. Estimação dos parâmetros do modelo

3.3. Inferências sobre o coeficiente de regressão

Análise de regressão

→ A visualização do diagrama de dispersão sugere a existência de uma relação funcional entre as duas variáveis.



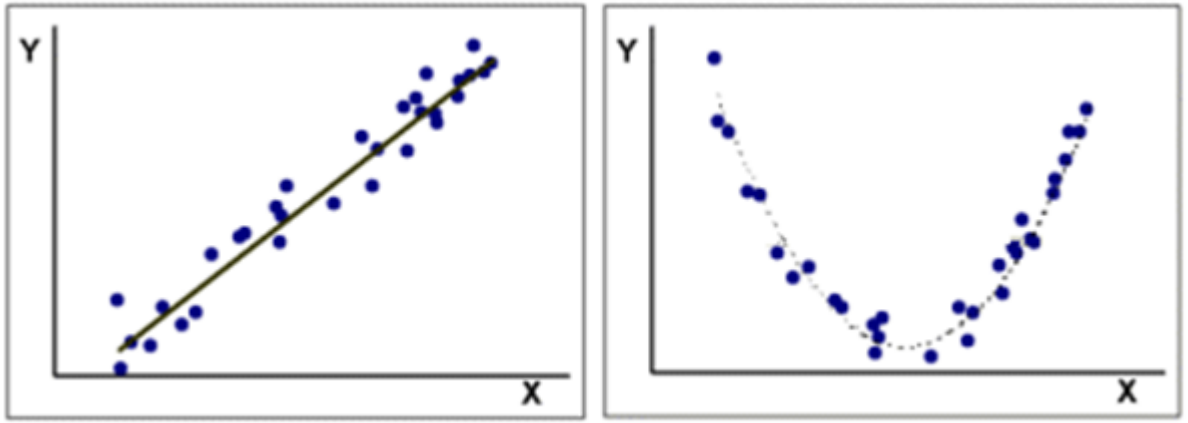
→ Problema: determinar uma função que exprima esse relacionamento.

→ A **análise de regressão** é uma técnica estatística cujo escopo é investigar e modelar a relação entre variáveis.

Análise de regressão

→ Existindo um relacionamento funcional entre os valores Y e X, essa função deverá explicar parcela significativa da variação de Y com X. Contudo, uma parcela da variação permanece inexplicada e deve ser atribuída ao acaso.

→ Admite-se a existência de uma função que explica, em termos médios, a variação de uma das variáveis com a variação da outra.



→ Os pontos observados apresentarão uma variação em torno da linha da função de regressão, devido à existência de uma **variação aleatória** adicional denominada de variação residual.

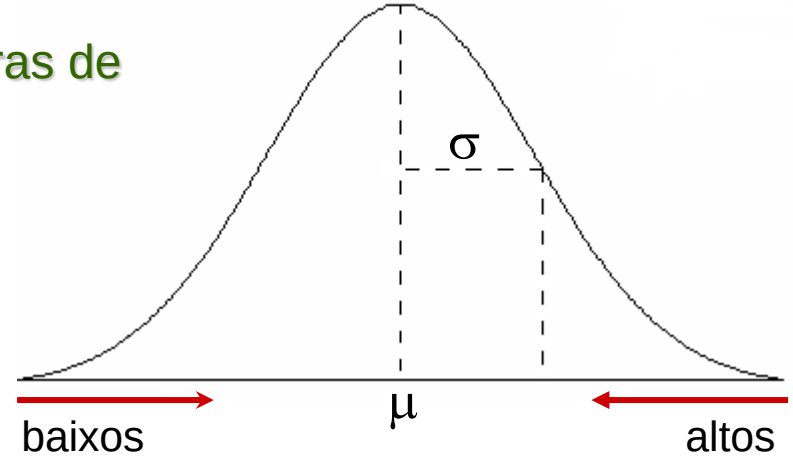
Origem do termo regressão → **Francis Galton**

As técnicas modernas de determinação da validade e da confiabilidade de testes são resultados diretos da descoberta da correlação, produzida quando Galton observou que as características herdadas tendem a regredir na direção da média.

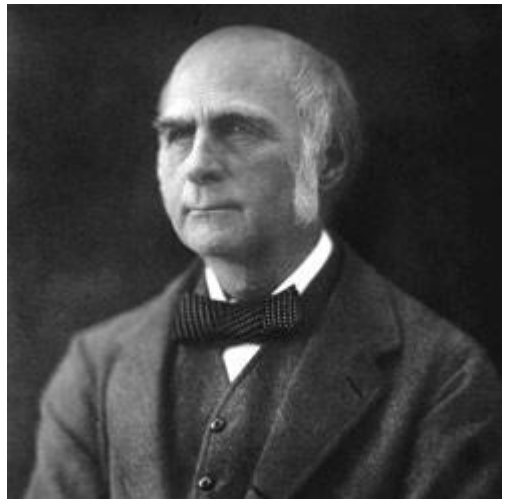
Estudo da relação entre as estaturas de pais e filhos

Y = estatura

$Y \sim N(\mu, \sigma^2)$



Regressão para a média



Francis Galton
(1822 – 1911)

Conclusão: filhos de pais altos não eram tão altos quanto seus pais e filhos de pais baixos não eram tão baixos quanto seus pais. A cada geração a estatura regride para a média da população. O fenômeno de retorno à média foi denominado regressão.

Por questões históricas o termo é usado até hoje, mas abriga uma série de técnicas.

A expressão **regressão linear simples** é utilizada por duas razões:

linear → relação entre X e Y é expressa por uma equação de primeiro grau, representada graficamente por uma reta

simples → envolve apenas duas variáveis: **X** que influencia e **Y** que é influenciada

Galton e os Métodos Estatísticos

Merece destaque o interesse de Galton pelas medidas e pela estatística. Ao longo de sua carreira, ele nunca parecia plenamente satisfeito com um problema até descobrir alguma maneira de quantificar os dados e analisá-los estatisticamente. Ele não se limitou a aplicar métodos estatísticos; também os desenvolveu.

Um estatístico belga, **Adolph Quetelet**, tinha sido o primeiro a aplicar a dados biológicos e sociais métodos estatísticos e a curva normal de probabilidade. A curva normal fora usada em trabalhos sobre a distribuição de medidas e erros na observação científica, mas o princípio da distribuição normal só veio a ser aplicado à variabilidade humana quando Quetelet demonstrou que medidas antropométricas de amostras aleatórias de pessoas geravam tipicamente uma curva normal. Ele mostrou que medidas da estatura de dez mil sujeitos se aproximavam da curva normal de distribuição, e usou a frase *l´homme moyen* (o homem médio) para exprimir a descoberta de que a maioria dos indivíduos se aglomera em torno da média ou centro de distribuição, e que um número cada vez menor vai sendo encontrado à medida que nos aproximamos dos extremos.

Galton ficou impressionado com os dados de Quetelet e supôs que resultados semelhantes poderiam ser encontrados para características mentais. Ele descobriu, por exemplo, que as notas dadas em exames universitários seguiam a mesma distribuição da curva normal dos dados de medida física de Quetelet. Devido à simplicidade da curva normal e à sua coerência em inúmeros traços, Galton propôs que um grande conjunto de medidas ou valores de características humanas poderia ser significativamente definido e resumido por dois números: o valor médio da distribuição (a média) e a dispersão ou gama de variação em torno desse valor médio (o desvio padrão).

A obra de Galton na estatística produziu uma das mais importantes medidas da ciência, a **correlação**. O primeiro relato sobre o que ele denominou “co-relações” apareceu em 1888. As técnicas modernas de determinação da validade e da confiabilidade de testes, bem como os métodos de análise fatorial, são resultados diretos da descoberta, por Galton, da correlação, produzida quando ele observou que as características herdadas tendem a regredir na direção da média. Por exemplo, **ele observou que os homens altos não são, em média, tão altos quanto os pais, enquanto os filhos de homens muito baixos são, em média, mais altos do que os pais**. Ele concebeu o meio gráfico de representar as propriedades básicas do coeficiente de correlação e desenvolveu uma fórmula de cálculo, hoje em desuso.

Galton aplicou o método da correlação a variações de medidas físicas, demonstrando, por exemplo, uma correlação entre a altura do corpo e o comprimento da cabeça. Com o estímulo de Galton, seu aluno **Karl Pearson** desenvolveu a fórmula matemática usada ainda hoje para o cálculo do coeficiente de correlação, chamada de coeficiente de correlação do produto-momento de Pearson. O símbolo do coeficiente de correlação, r, vem da primeira letra da palavra *regressão*, em reconhecimento à descoberta de Galton da tendência de as características humanas herdadas regredirem na direção da média ou mediana. A correlação é uma ferramenta fundamental das ciências sociais e do comportamento, bem como da engenharia e das ciências naturais. A partir da obra pioneira de Galton, foram desenvolvidas muitas outras técnicas estatísticas.

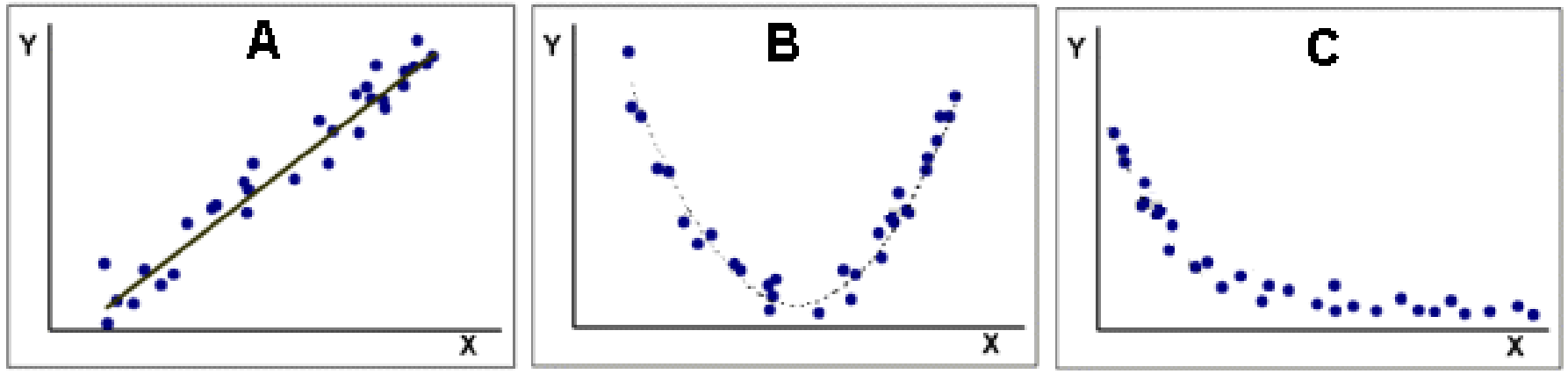
(Extraído do texto “As diferenças individuais: Francis Galton”, de **Suely Vieira Lopes**, Pontifícia Universidade Católica de Goiás.)

Ajustamento de curvas

Expressar através de uma equação matemática as relações entre variáveis conhecidas e variáveis que devem ser determinadas.

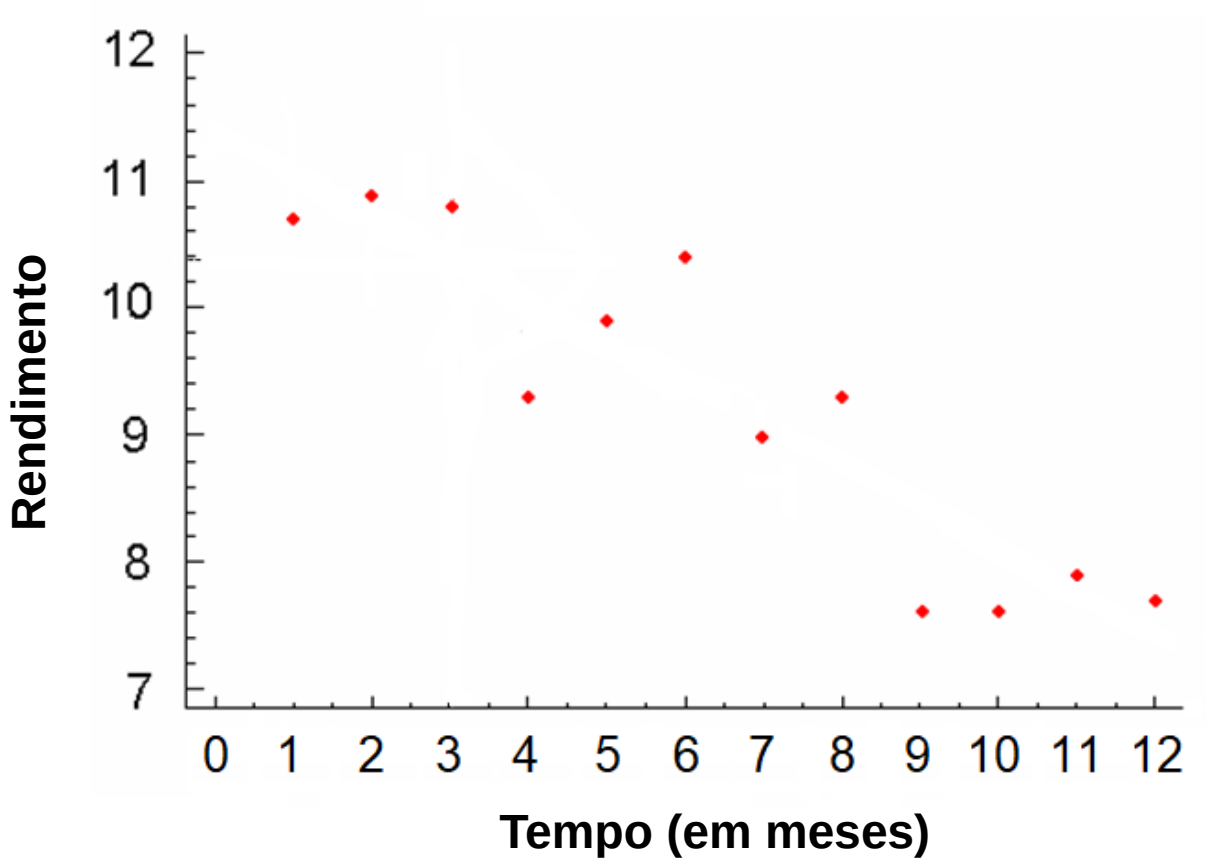
Como ajustar uma curva?

O ajustamento de curvas é feito com base em dados observados, de tal forma que a partir dessa curva (ou reta) ajustada se possa representar, graficamente ou analiticamente, a relação entre as variáveis.



Exemplo: Após uma regulagem eletrônica, um veículo apresenta um rendimento ideal quanto ao consumo de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados a seguir representam o rendimento medido mês a mês após a regulagem. Ajuste um modelo linear a esses dados.

Meses (X)	1	2	3	4	5	6	7	8	9	10	11	12
Rendimento (Y)	10,7	10,9	10,8	9,3	9,5	10,4	9,0	9,3	7,6	7,6	7,9	7,7



Cálculos iniciais

i	Meses (x)	Rendimento (y)	x^2	y^2	xy
1	1	10,7	1	114,49	10,7
2	2	10,9	4	118,81	21,8
3	3	10,8	9	116,64	32,4
4	4	9,3	16	86,49	37,2
5	5	9,5	25	90,25	47,5
6	6	10,4	36	108,16	62,4
7	7	9,0	49	81,00	63,0
8	8	9,3	64	86,49	74,4
9	9	7,6	81	57,76	68,4
10	10	7,6	100	57,76	76,0
11	11	7,9	121	62,41	86,9
12	12	7,7	144	59,29	92,4
Soma	78	110,7	650	1039,55	673,1
Média	6,5	9,225			

Cálculos – Coeficiente de Correlação

$\Sigma x = 78$ $\quad \bar{X} = 6{,}50$ $\quad \Sigma x^2 = 650$

$\Sigma y = 110{,}7$ $\quad \bar{Y} = 9{,}225$ $\quad \Sigma y^2 = 1039{,}55$

$\Sigma xy = 673{,}1$

$$SPXY = \sum x_i y_i - n\,\bar{x}\,\bar{y} = 673{,}1 - 12 \times 6{,}5 \times 9{,}225 = -46{,}45$$

$$SQX = \sum x_i^2 - n\,\bar{x}^2 = 650 - 12 \times 6{,}5^2 = 143$$

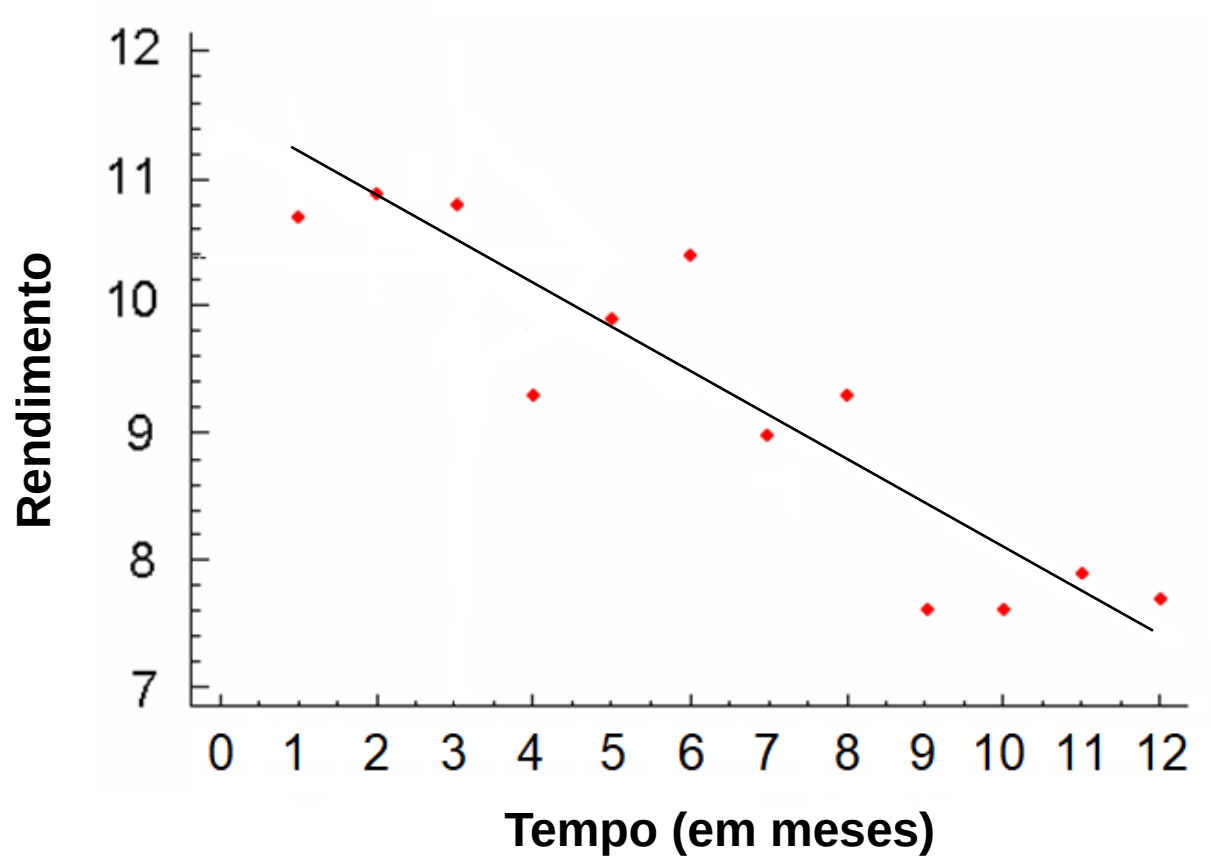
$$SQY = \sum y_i^2 - n\,\bar{y}^2 = 1039{,}55 - 12 \times 9{,}225^2 = 18{,}34$$

$$r = \frac{SPXY}{\sqrt{SQY \,.\, SQX}} = \frac{-46{,}45}{\sqrt{143{,}00 \times 18{,}34}} = -0{,}907$$

Interpretação: Existe uma correlação linear negativa na amostra entre tempo após a regulagem e rendimento. A intensidade desta correlação é forte.

Exemplo: Após uma regulagem eletrônica, um veículo apresenta um rendimento ideal quanto ao consumo de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados a seguir representam o rendimento medido mês a mês após a regulagem. Ajuste um modelo linear a esses dados.

Meses (X)	1	2	3	4	5	6	7	8	9	10	11	12
Rendimento (Y)	10,7	10,9	10,8	9,3	9,5	10,4	9,0	9,3	7,6	7,6	7,9	7,7

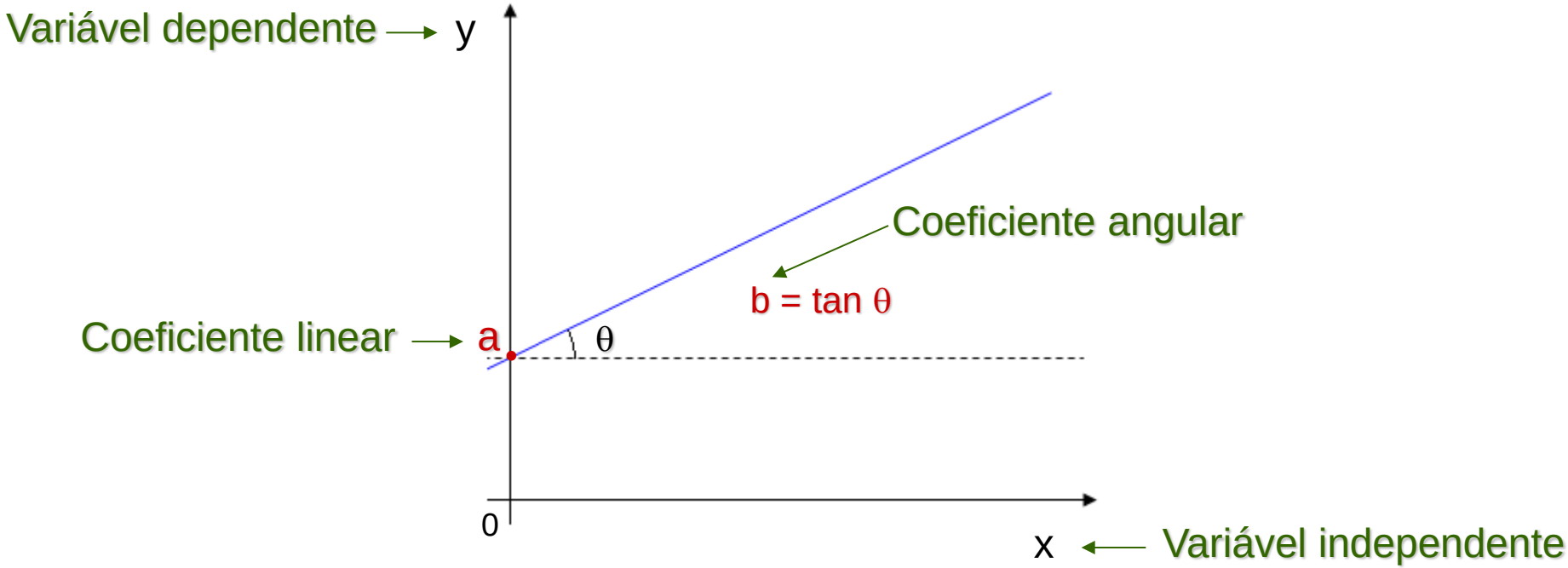


Ajustar uma curva (ou reta) é determinar a função matemática que melhor representa um conjunto de observações.

Modelo matemático (determinístico)

Sendo X e Y duas variáveis que se relacionam de forma linear, a relação é expressa por:

$$y = a + bx$$



Exemplo: Salário de um vendedor

Modelo determinístico

Salário = fixo + 1% das vendas $\rightarrow$ $y = 1500 + 0{,}01\,x$

Modelo estatístico (probabilístico)

Se Y é uma variável aleatória, então ela está sujeita a variação (erro de observação). Este erro (e_i) deverá ser adicionado ao modelo. Assim, o modelo de regressão linear simples será:

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

onde:

y_i é a variável resposta (dependente)

x_i é a variável preditora (independente)

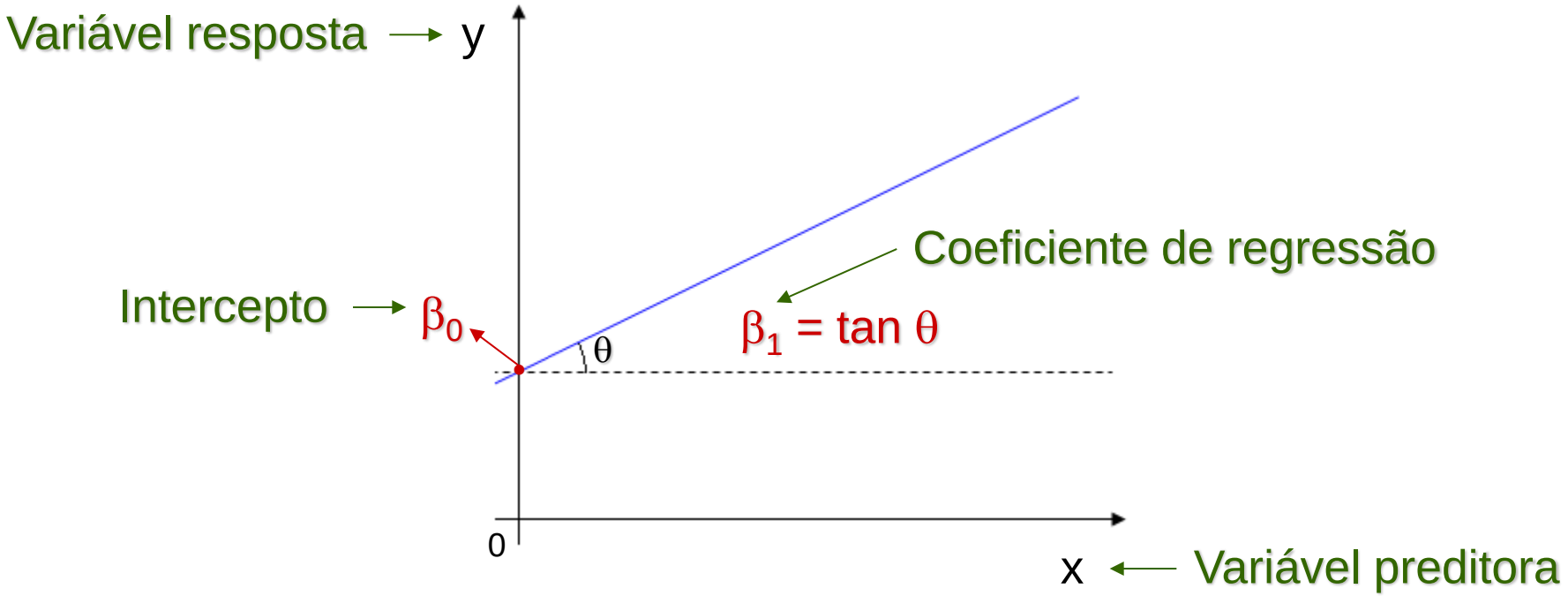
β_0 é o intercepto ou coeficiente linear (valor de Y para X=0)

β_1 é o coeficiente de regressão ou coeficiente angular

e_i é o erro (variação aleatória)

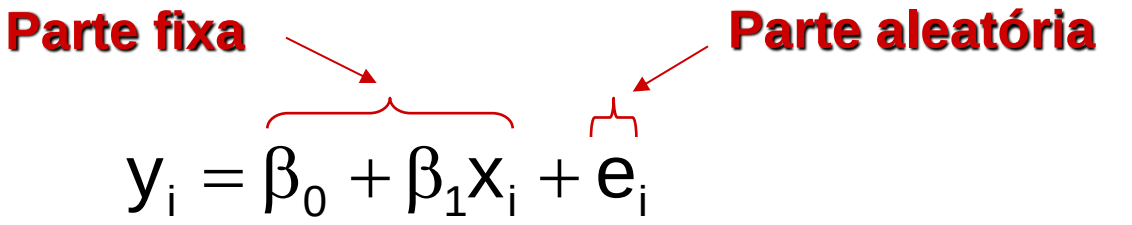
Modelo estatístico

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



Modelo estatístico

Este modelo é composto por uma parte fixa e uma parte aleatória:

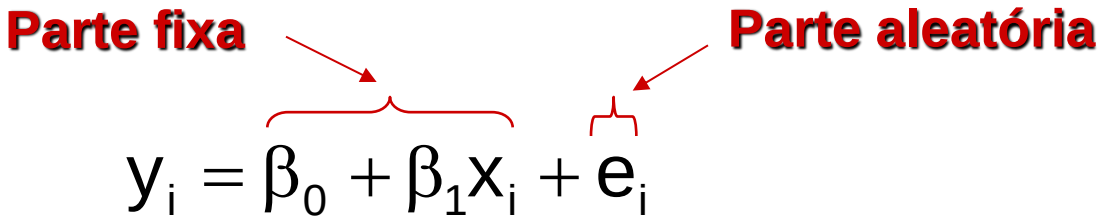


Parte fixa → informa como X influencia Y

Parte aleatória → mostra que Y possui uma variabilidade inerente, significando que X não é a única variável que influencia Y.

Modelo estatístico

Este modelo é composto por uma parte fixa e uma parte aleatória:



Parte fixa → informa como X influencia Y

Parte aleatória → mostra que Y possui uma variabilidade inerente, significando que X não é a única variável que influencia Y.

O modelo só será adequado quando a parte fixa for preponderante sobre a aleatória.

Exemplo: Após uma regulagem eletrônica, um veículo apresenta um rendimento ideal quanto ao consumo de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados a seguir representam o rendimento medido mês a mês após a regulagem.

Meses (X)	1	2	3	4	5	6	7	8	9	10	11	12
Rendimento (Y)	10,7	10,9	10,8	9,3	9,5	10,4	9,0	9,3	7,6	7,6	7,9	7,7

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

onde:

y_i é o rendimento do veículo (variável resposta)

x_i é o tempo após a regulagem (variável preditora) em meses

β_0 é o rendimento no tempo zero após a regulagem (intercepto)

β_1 é a taxa de variação no rendimento para cada unidade de tempo que passa (coeficiente de regressão)

e_i é o erro (variação aleatória)

Exemplo: Após uma regulagem eletrônica, um veículo apresenta um rendimento ideal quanto ao consumo de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados a seguir representam o rendimento medido mês a mês após a regulagem.

Meses (X)	1	2	3	4	5	6	7	8	9	10	11	12
Rendimento (Y)	10,7	10,9	10,8	9,3	9,5	10,4	9,0	9,3	7,6	7,6	7,9	7,7

Parâmetros (constantes)

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, 2, \ldots, 12$$

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_1 + e_1 \\ y_2 = \beta_0 + \beta_1 x_2 + e_2 \\ \quad \ldots \\ y_{12} = \beta_0 + \beta_1 x_{12} + e_{12} \end{cases}$$

Os coeficientes β_0 e β_1 são os parâmetros do modelo que serão estimados a partir dos valores da amostra.

Pressuposições

O termo do erro (e_i) é adicionado ao modelo, desde que se admitam como verdadeiras as seguintes pressuposições:

1. Os erros são aleatórios, têm média zero e variância constante, ou seja, $E(e_i) = 0$ e $V(e_i) = \sigma^2$.

2. Os erros têm distribuição normal e são independentes entre si.

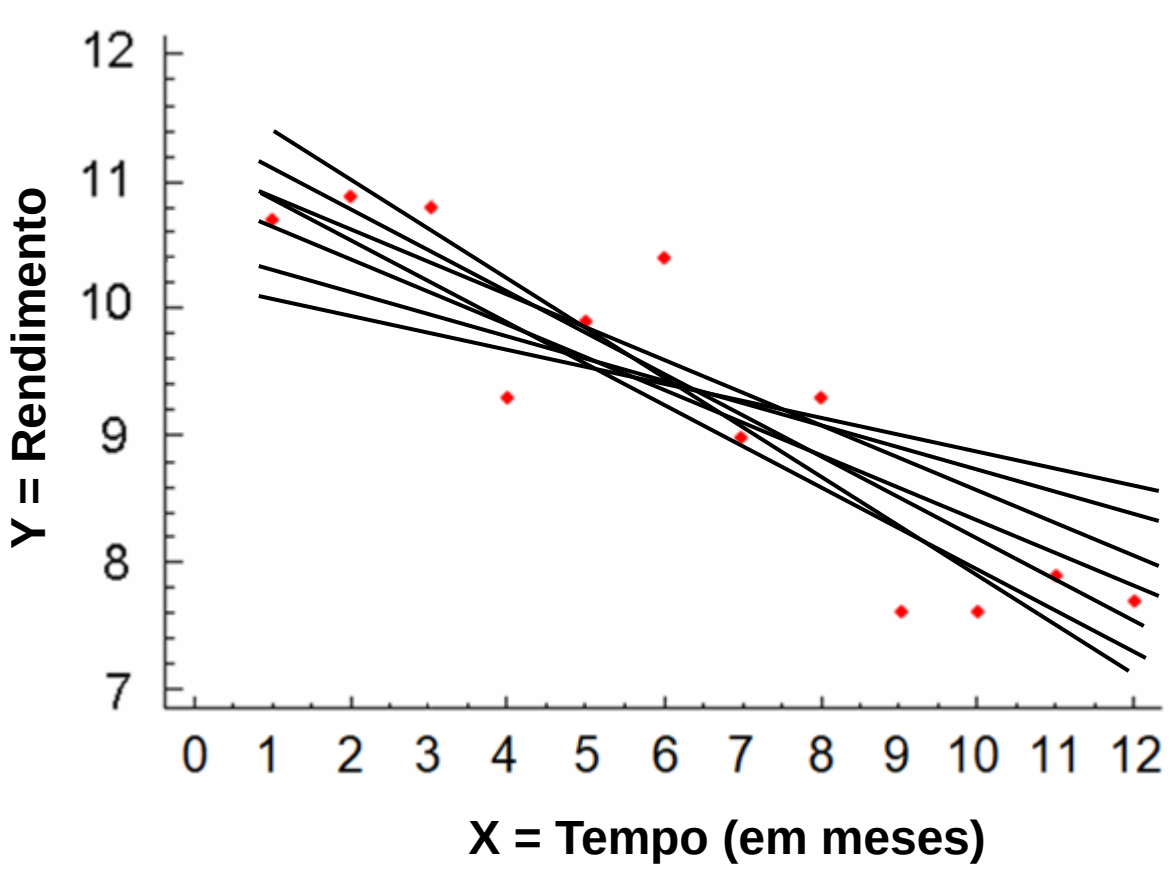
3. O modelo é adequado para todas as observações, não podendo haver nenhum valor de X que produza um valor de Y discrepante dos demais.

4. A variável X é fixa (não aleatória).

Análise de regressão

Quantas retas é possível traçar entre os pontos?

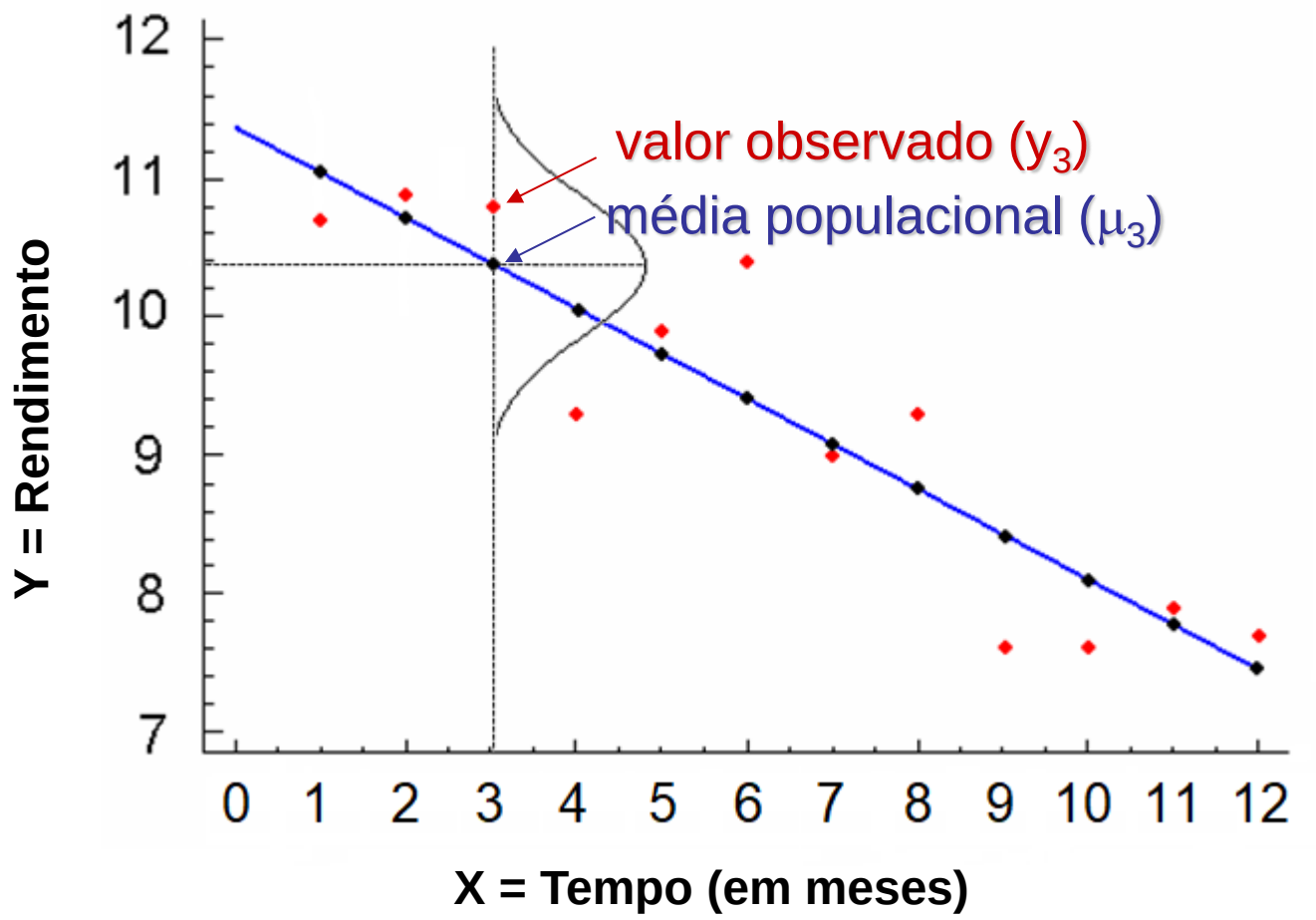
Infinitas!!!



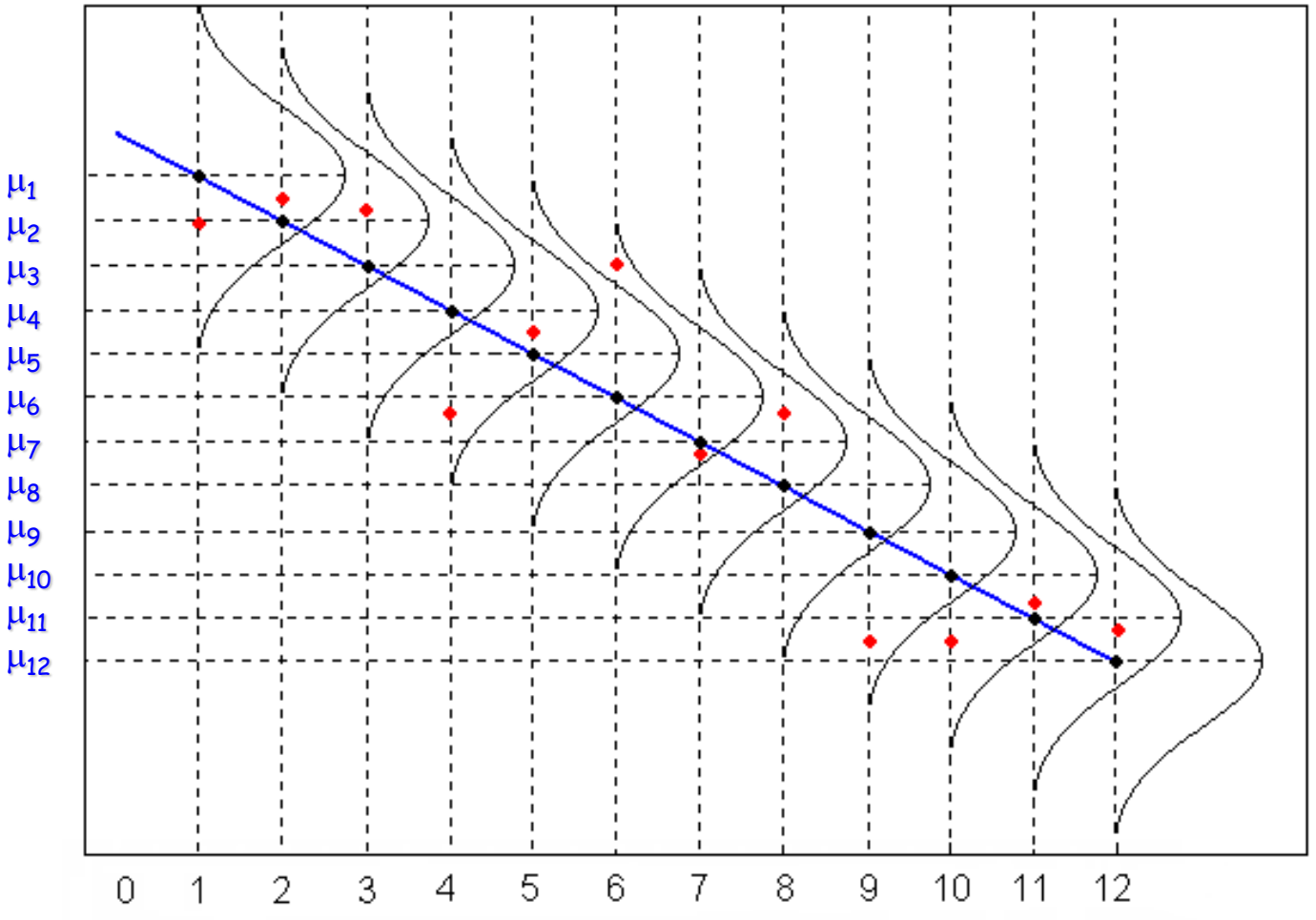
Determinar a reta que melhor representa o conjunto de observações!

Pressupomos que, para um valor qualquer de X, Y é uma variável aleatória com distribuição normal, média μ e variância σ^2.

Por exemplo, para X=3, Y tem distribuição normal, com média $\mu_3=E(y/x_3)$ e variância σ^2.



Médias populacionais (μ_i)



Assim, para cada valor de X, Y é uma variável aleatória com distribuição normal e média $\mu_i=E(y/x_i)$.

A variância de Y é constante (σ^2) para todos os valores de X.

Análise de regressão

O objetivo da análise de regressão é **determinar a equação (reta) que melhor representa a relação existente entre duas variáveis** e, a partir desta equação, fazer predições para a variável resposta.

Para isso, uma sequência de passos deve ser seguida:

1. Obtenção das estimativas (pontuais) dos coeficientes β_0 e β_1 para ajustar a equação da regressão.

2. Aplicação de **testes de hipóteses** para as estimativas obtidas, a fim de verificar se a equação de regressão é adequada.

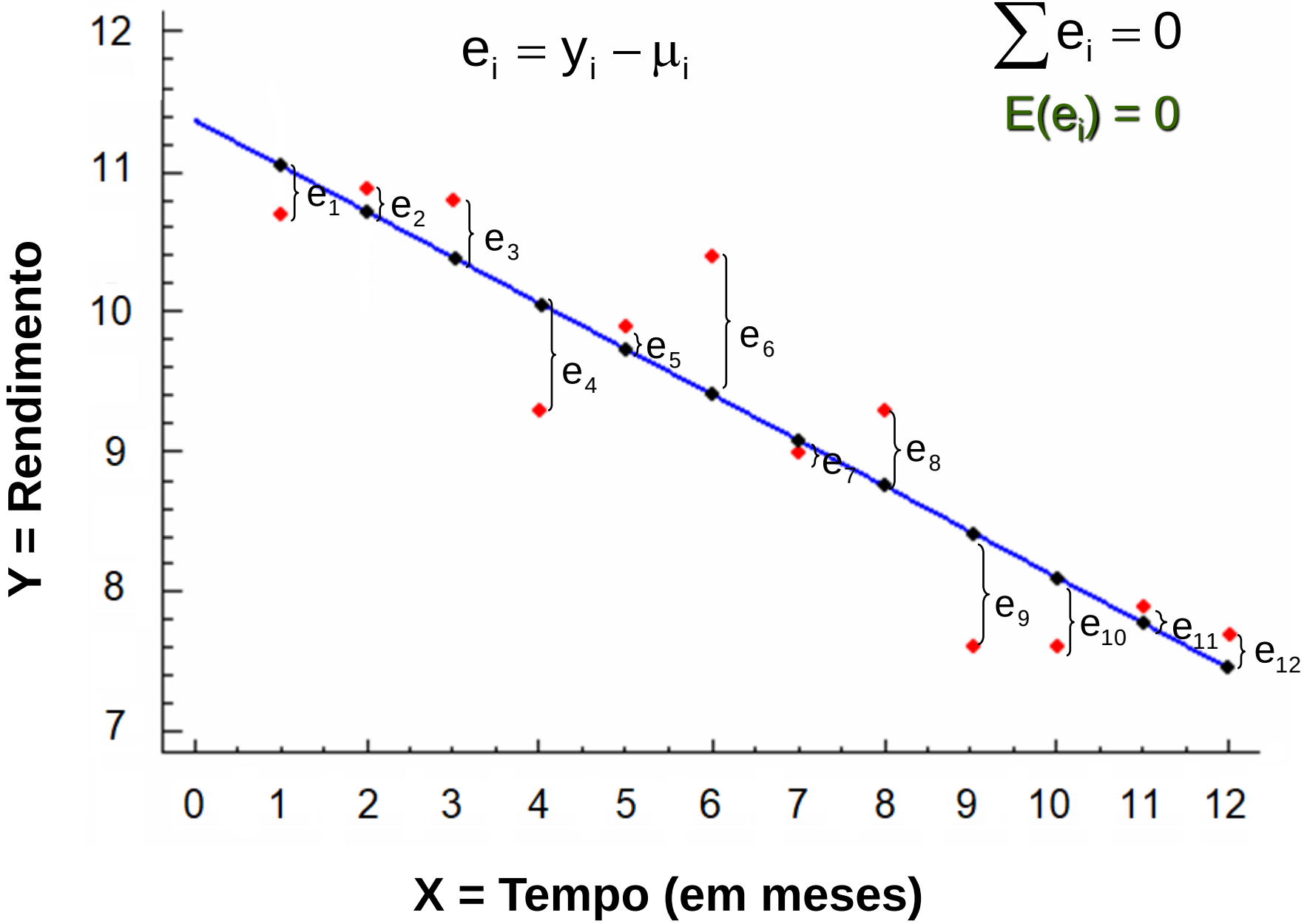
3. Construção de **intervalos de confiança** para os valores estimados pela equação de regressão.

Como a soma dos erros é nula, a média dos erros também é nula.

$$e_i = y_i - \mu_i$$

$$\sum e_i = 0$$

$$E(e_i) = 0$$



$E(e_i) = 0$

$V(e_i) = \sigma^2$

Se $y_i = \beta_0 + \beta_1 x_i + e_i$, então

$$\mu_{y/x_i} = E(y/x_i) = E(\beta_0 + \beta_1 x_i + e_i)$$

$$\mu_{y/x_i} = E(y/x_i) = E(\beta_0) + E(\beta_1 x_i) + E(e_i)$$

$$\mu_{y/x_i} = E(y/x_i) = \beta_0 + \beta_1 x_i$$

Sendo assim, se $y_i = \beta_0 + \beta_1 x_i + e_i$, então

$$y_i = \mu_i + e_i$$

logo, $e_i = y_i - \mu_i$

valor observado

valor esperado

A estimação dos parâmetros β_0 e β_1 é efetuada através do método dos mínimos quadrados.

Método dos mínimos quadrados

Este método tem como objetivo obter estimativas de tal forma que a soma dos quadrados dos erros ($\sum e_i^2$) seja o menor valor possível.

Sendo $e_i = y_i - \mu_i$ e $\mu_i = \beta_0 + \beta_1 x_i$,

podemos escrever:

$$\sum e_i^2 = \sum (y_i - \mu_i)^2$$
$$= \sum \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2$$

Método dos mínimos quadrados

Para encontrar os valores de β_0 e β_1 que tornam mínima a soma de quadrados dos erros, o método consiste em três passos:

1. Encontrar as derivadas parciais em relação a β_0 e β_1 para a equação:

$$\sum e_i^2 = \sum \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2$$

2. Igualar as derivadas a zero para obter os pontos críticos.

3. Determinar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ por um sistema de equações normais.

A conclusão do terceiro passo resulta nas seguintes expressões:

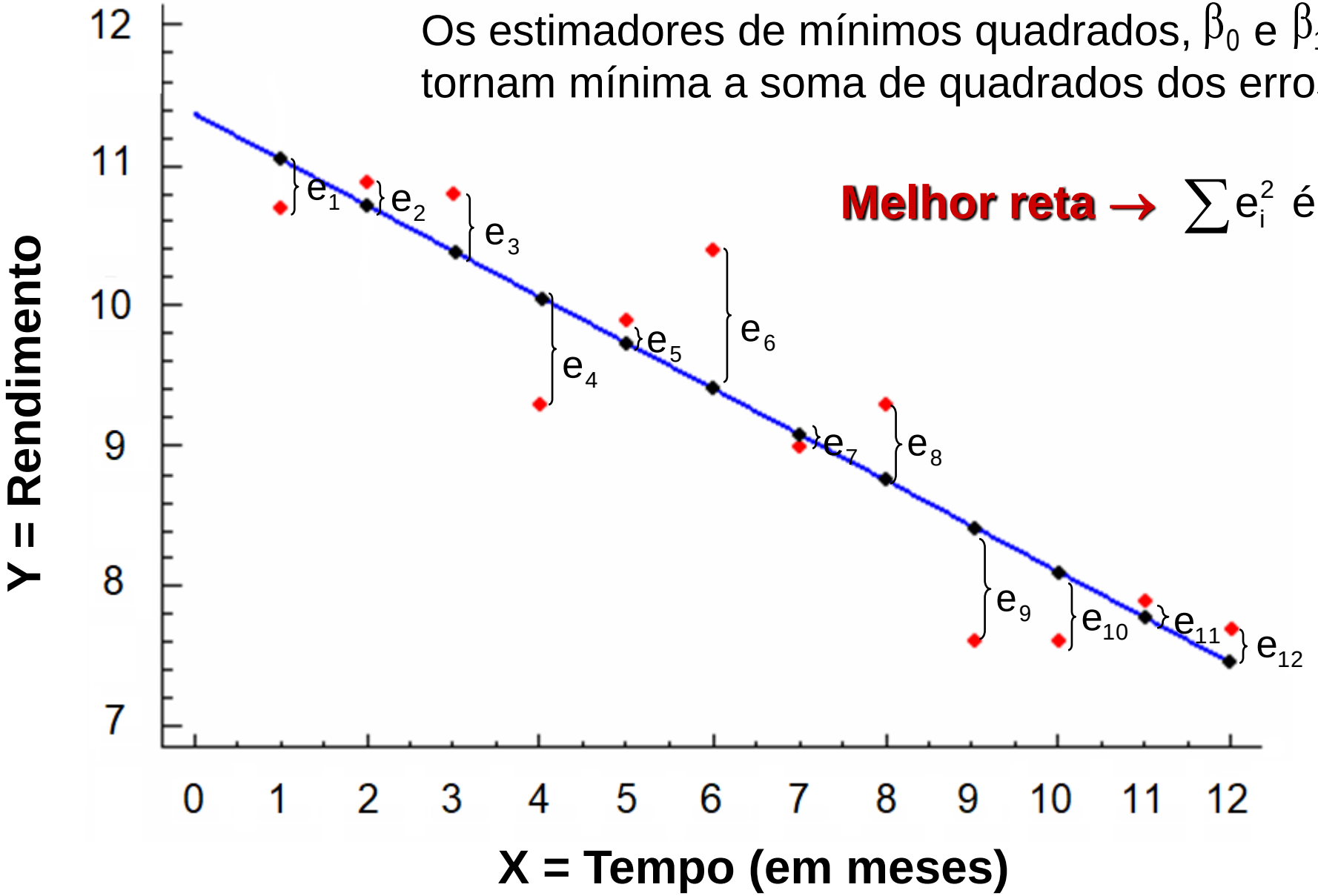
$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

Método dos mínimos quadrados

Os estimadores de mínimos quadrados, $\hat{\beta}_0$ e $\hat{\beta}_1$, tornam mínima a soma de quadrados dos erros $\sum e_i^2$.

Melhor reta → $\sum e_i^2$ é mínima



Estimadores de mínimos quadrados

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SPXY}{SQX}$$ ← Estimador do coeficiente de regressão

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$ ← Estimador do intercepto

$$y_i = \beta_0 + \beta_1 x + e_i$$ ← Valor observado de y para um dado valor x_i

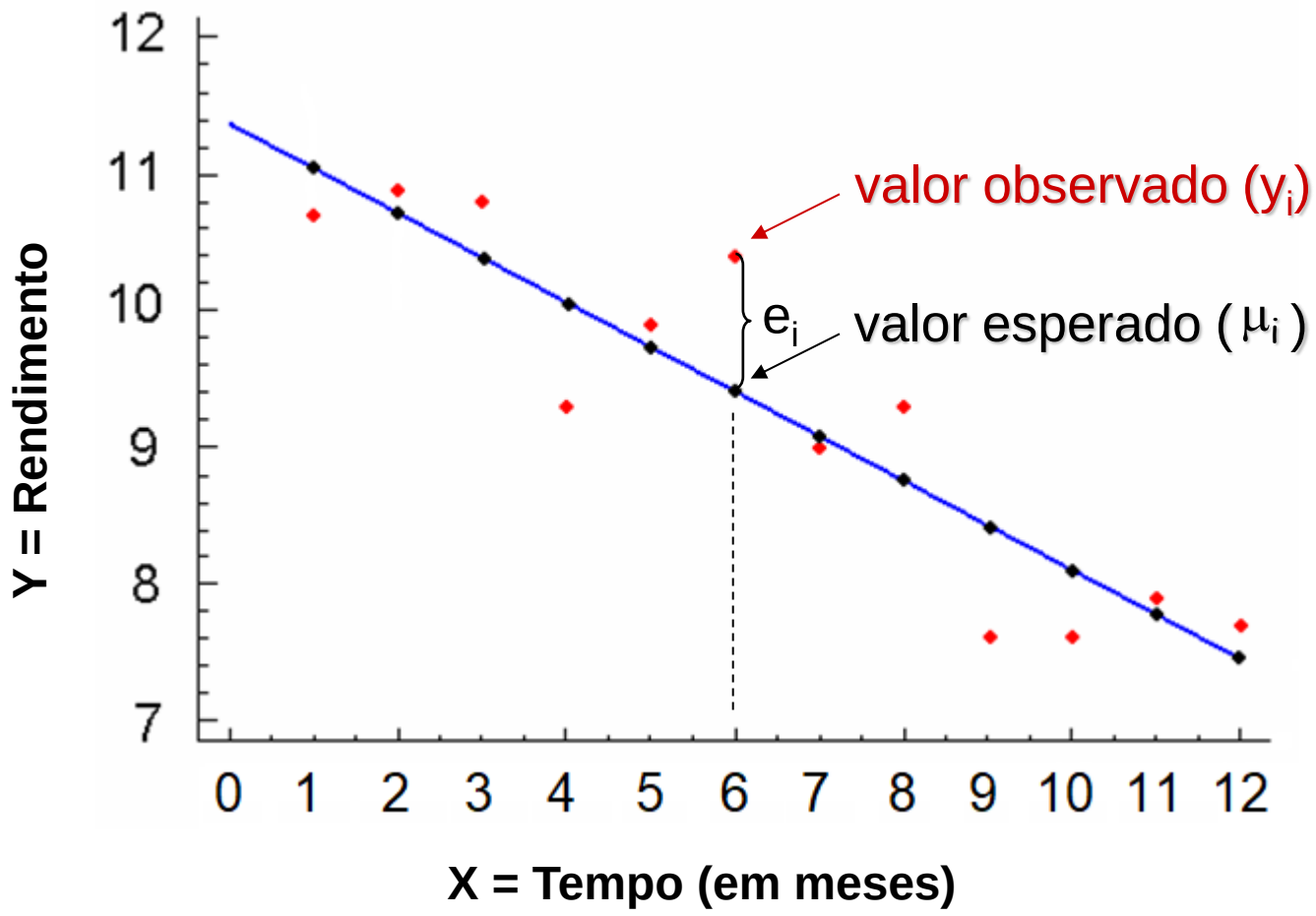
$$\mu_i = \beta_0 + \beta_1 x$$ ← Valor esperado de y para um dado valor x_i

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x$$ ← Estimativa do valor esperado de y para um dado valor x_i

$$e_i = y_i - \mu_i$$ ← Erro aleatório da observação i

$$\hat{e}_i = y_i - \hat{\mu}_i$$ ← Estimativa do erro aleatório da observação i (resíduo)

O erro de cada observação é definido pela diferença entre o valor observado e o valor esperado: $e_i = y_i - \mu_i$



Estimadores de mínimos quadrados

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SPXY}{SQX}$$ ← Estimador do coeficiente de regressão

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$ ← Estimador do intercepto

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x$$ ← Estimativa do valor esperado de y para um dado valor x_i

$$\hat{e}_i = y_i - \hat{\mu}_i$$ ← Estimativa do erro aleatório da observação i (resíduo)

Regressão e correlação

A análise de regressão simples e a análise de correlação simples são estreitamente relacionadas. Assim, no modelo de regressão simples, o coeficiente de regressão é **proporcional** e tem o **mesmo sinal** do coeficiente de correlação simples.

$$r_{xy} = \frac{SPXY}{\sqrt{SQX.\,SQY}}$$

$$r_{xy}\sqrt{SQX.\,SQY} = SPXY$$

$$\hat{\beta}_1 = \frac{SPXY}{SQX} = \frac{r_{xy}\sqrt{SQX.\,SQY}}{SQX} = r_{xy}\frac{\sqrt{SQX.\,SQY}}{SQX}$$

$$\hat{\beta}_1 = r_{xy}\frac{\sqrt{SQX.\,SQY}}{SQX}$$

Exemplo: Após uma regulagem eletrônica, um veículo apresenta um rendimento ideal quanto ao consumo de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados a seguir representam o rendimento medido mês a mês após a regulagem. Ajuste um modelo linear a esses dados.

Meses (X)	1	2	3	4	5	6	7	8	9	10	11	12
Rendimento (Y)	10,7	10,9	10,8	9,3	9,5	10,4	9,0	9,3	7,6	7,6	7,9	7,7

$r = -0{,}907$

Tabelas auxiliar - Cálculos iniciais

i	Meses (x)	Rendimento (y)	x^2	y^2	xy
1	1	10,7	1	114,49	10,7
2	2	10,9	4	118,81	21,8
3	3	10,8	9	116,64	32,4
4	4	9,3	16	86,49	37,2
5	5	9,5	25	90,25	47,5
6	6	10,4	36	108,16	62,4
7	7	9,0	49	81,00	63,0
8	8	9,3	64	86,49	74,4
9	9	7,6	81	57,76	68,4
10	10	7,6	100	57,76	76,0
11	11	7,9	121	62,41	86,9
12	12	7,7	144	59,29	92,4
Soma	78	110,7	650	1039,55	673,1
Média	6,5	9,225			

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{673,1 - 12 \text{ x } 6,5 \text{ x } 9,225}{650 - 12 \text{ x } 6,5^2} = -0,325$$

Estimativas dos parâmetros:

$\hat{\beta}_1 = -0{,}325 \qquad \bar{y} = 9{,}225 \qquad \bar{x} = 6{,}5$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$= 9{,}225 - (-0{,}325) \times 6{,}5$$
$$= 11{,}34$$

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Equação da reta ajustada ou equação de regressão

$$\hat{\mu}_i = 11{,}34 - 0{,}325\, x$$

Estimativas pontuais

Estimativas dos valores esperados de y/x_i:

$$\hat{\mu}_i = 11,34 - 0,325 x_i$$

$$\hat{\mu}_1 = 11,34 - 0,325 \times 1 = 11,02$$

$$\hat{\mu}_2 = 11,34 - 0,325 \times 2 = 10,69$$

$$\hat{\mu}_3 = 11,34 - 0,325 \times 3 = 10,37$$

$$\hat{\mu}_4 = 11,34 - 0,325 \times 4 = 10,04$$

$$\hat{\mu}_5 = 11,34 - 0,325 \times 5 = 9,72$$

$$\hat{\mu}_6 = 11,34 - 0,325 \times 6 = 9,39$$

$$\hat{\mu}_7 = 11,34 - 0,325 \times 7 = 9,07$$

$$\hat{\mu}_8 = 11,34 - 0,325 \times 8 = 8,74$$

$$\hat{\mu}_9 = 11,34 - 0,325 \times 9 = 8,42$$

$$\hat{\mu}_{10} = 11,34 - 0,325 \times 10 = 8,09$$

$$\hat{\mu}_{11} = 11,34 - 0,325 \times 11 = 7,77$$

$$\hat{\mu}_{12} = 11,34 - 0,325 \times 12 = 7,44$$

Estimativas do erros (resíduos):

$$\hat{e}_i = y_i - \hat{\mu}_i$$

$$\hat{e}_1 = 10,7 - 11,02 = -0,315$$

$$\hat{e}_2 = 10,9 - 10,69 = 0,210$$

$$\hat{e}_3 = 10,8 - 10,37 = 0,435$$

$$\hat{e}_4 = 9,3 - 10,04 = -0,740$$

$$\hat{e}_5 = 9,5 - 9,72 = -0,210$$

$$\hat{e}_6 = 10,4 - 9,39 = 1,010$$

$$\hat{e}_7 = 9,0 - 9,07 = -0,065$$

$$\hat{e}_8 = 9,3 - 8,74 = 0,560$$

$$\hat{e}_9 = 7,6 - 8,42 = -0,815$$

$$\hat{e}_{10} = 7,6 - 8,09 = -0,490$$

$$\hat{e}_{11} = 7,9 - 7,77 = 0,135$$

$$\hat{e}_{12} = 7,7 - 7,44 = 0,260$$

Análise de regressão

O objetivo da análise de regressão é **determinar a equação (reta) que melhor representa a relação existente entre duas variáveis** e, a partir desta equação, fazer predições para a variável resposta.

Esta análise compreende uma sequência de passos:

1. Obtenção das estimativas (pontuais) dos coeficientes β_0 e β_1 para ajustar a equação da regressão.

2. Aplicação de **testes de hipóteses** para as estimativas obtidas, a fim de verificar se a equação de regressão é adequada.

3. Construção de **intervalos de confiança** para os valores estimados pela equação de regressão.

Testes de hipóteses e intervalos de confiança para β_1

O parâmetro mais importante na equação de regressão é o coeficiente de regressão β_1 porque determina a declividade da reta.

Assim, quando estimamos β_1, devemos verificar se esta estimativa difere significativamente de zero.

Esta verificação pode ser feita por um teste de hipóteses, em que as hipóteses de interesse são:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Testamos se a inclinação é igual a zero, o que equivale a testar se existe uma relação linear entre Y e X.

Metodologias para testar H_0
- Teste t
- Análise da variância

Teste t

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Podemos testar H_0 utilizamos a estatística T com distribuição de t de Student.

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} \sim t\,(\nu)$$

onde:

$$\theta = \beta_1$$

$$\hat{\theta} = \hat{\beta}_1$$

$$S(\hat{\theta}) = S(\hat{\beta}_1) = \sqrt{S^2(\hat{\beta}_1)} = \sqrt{\frac{\frac{\sum \hat{e}_i^2}{n-2}}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{SQX}}}$$

$$\nu = n - 2$$

A variância do estimador $\hat{\beta}_1$ é obtida da seguinte forma:

$$V(\hat{\beta}_1) = V\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right) = V\left(\frac{\sum y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right)$$

$$V(\hat{\beta}_1) = V\left(\frac{1}{\sum(x_i - \bar{x})^2}\sum y_i(x_i - \bar{x})\right) = \left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 V\left[\sum y_i(x_i - \bar{x})\right]$$

$$V(\hat{\beta}_1) = \frac{1}{\left[\sum(x_i - \bar{x})^2\right]^2}\sum(x_i - \bar{x})^2 V(y_i) = \frac{\sum(x_i - \bar{x})^2}{\left[\sum(x_i - \bar{x})^2\right]^2} V(y_i)$$

$$V(\hat{\beta}_1) = \frac{V(y_i)}{\sum(x_i - \bar{x})^2}$$

$$V(\hat{\beta}_1) = \frac{\sigma_y^2}{\sum(x_i - \bar{x})^2}$$

Sendo σ^2 um parâmetro desconhecido, utilizamos o seu estimador $S^2 = \frac{\sum \hat{e}_i^2}{n-2}$ para obter a estimativa da variância de $\hat{\beta}_1$:

$$S^2(\hat{\beta}_1) = \frac{\frac{\sum \hat{e}_i^2}{n-2}}{\sum(x_i - \bar{x})^2}$$

Assim, sob $H_0 : \beta_1 = 0$ verdadeira, temos

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} = \frac{\hat{\beta}_1 - 0}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} \sim t(\nu)$$

onde: $\theta = \beta_1 = 0$

$$\hat{\theta} = \hat{\beta}_1$$

$$S(\hat{\theta}_1) = S(\hat{\beta}_1) = \sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{SQX}}}$$

$$\nu = n - 2$$

Estatística do teste

$$T = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{SQX}}}} \sim t(\nu)$$

Critério de decisão

Comparação com o t crítico (tabelado): $t_{\alpha/2(n-2)}$

Rejeitamos H_0 se: $|t| > t_{\alpha/2(n-2)}$

Distribuição t de Student

$$T = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{SQX}}}} \sim t(\nu)$$

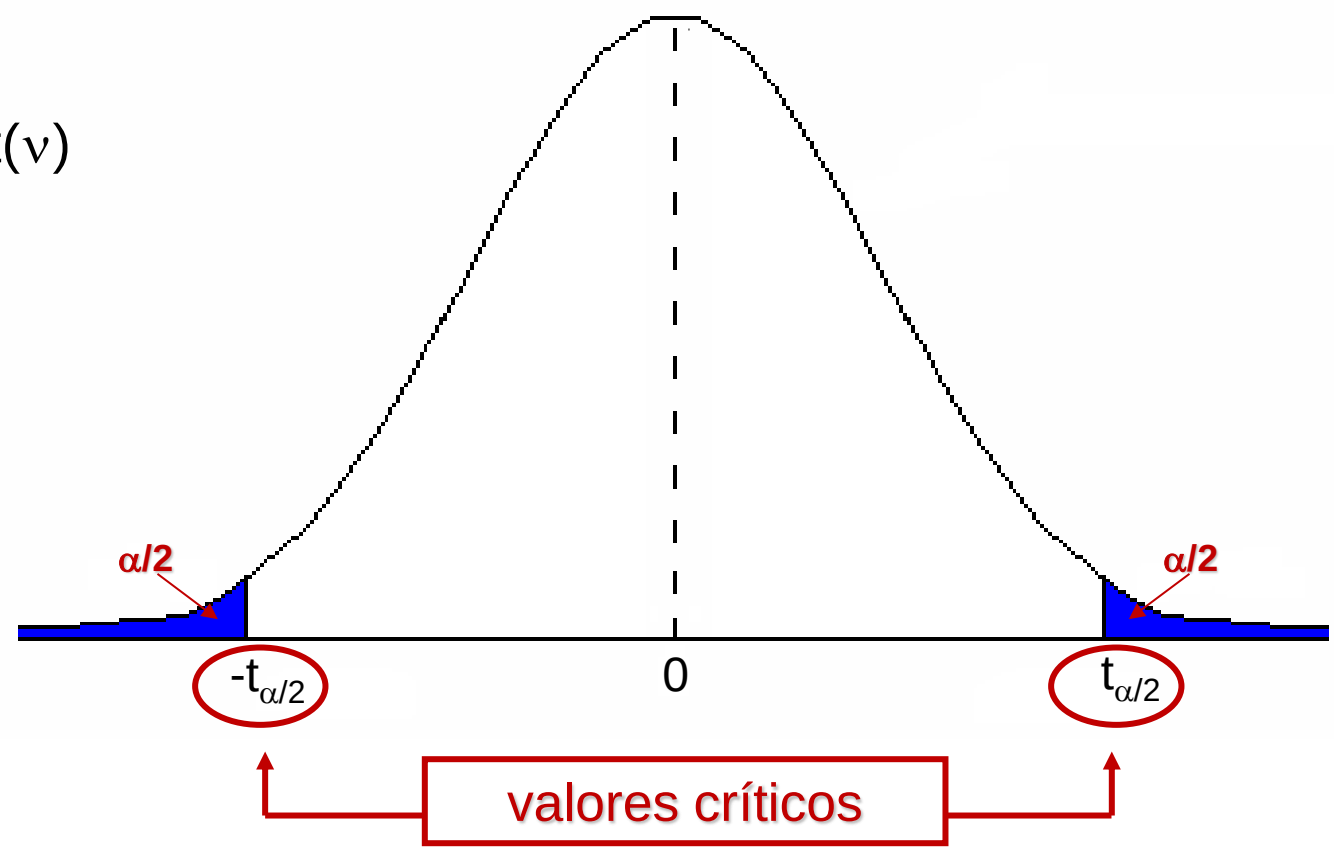


Tabela auxiliar

i	Meses (x)	Rendimento (y)	x^2	$\hat{\mu}_i$	$\hat{e}_i$	$\hat{e}_i^2$
1	1	10,7	1	11,02	-0,32	0,0992
2	2	10,9	4	10,69	0,21	0,0441
3	3	10,8	9	10,37	0,44	0,1892
4	4	9,3	16	10,04	-0,74	0,5476
5	5	9,5	25	9,72	-0,22	0,0462
6	6	10,4	36	9,39	1,01	1,0201
7	7	9,0	49	9,07	-0,06	0,0042
8	8	9,3	64	8,74	0,56	0,3136
9	9	7,6	81	8,42	-0,82	0,6642
10	10	7,6	100	8,09	-0,49	0,2401
11	11	7,9	121	7,77	0,14	0,0182
12	12	7,7	144	7,44	0,26	0,0676
Soma	78	110,7	650	11,02	0,0	3,2545
Média	6,5	9,225				

Estatística do teste

$$T = \frac{\hat{\beta}_1}{\sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{SQX}}}}$$

$$t = \frac{-0{,}325}{\sqrt{\frac{\frac{3{,}2545}{10}}{143}}} = \frac{-0{,}325}{\sqrt{\frac{0{,}32545}{143}}} = -6{,}81$$

Teste t

Hipóteses estatísticas

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Estatística do teste

$$t = \frac{-0{,}325}{\sqrt{\frac{3{,}2545}{\frac{10}{143}}}} = \frac{-0{,}325}{\sqrt{\frac{0{,}32545}{143}}} = -6{,}81$$

Decisão e conclusão

$\alpha = 0{,}05$ $\qquad$ $t_{\alpha/2(10)} = 2{,}228$

$\nu = (12 - 2) = 10$ $\qquad$ $|t = -6{,}81| > t_{\alpha/2(n-2)} = 2{,}228$ $\leftarrow$ Rejeita-se H_0

Concluímos, ao nível de 5% de significância, que o coeficiente de regressão populacional difere de zero. Portanto, existe relação linear significativa entre o tempo após a regulagem do motor e o rendimento quanto ao consumo de combustível.

Graus de Liberdade (ν)	**Limites bilaterais**: $P(\lvert t \rvert > t_{\alpha/2})$							
	Nível de Significância (α)							
	0,50	0,20	0,10	0,05	0,025	0,02	0,01	0,005
1	1,000	3,078	6,314	12,706	25,542	31,821	63,657	127,320
2	0,816	1,886	2,920	4,303	6,205	6,965	9,925	14,089
3	0,715	1,638	2,353	3,183	4,177	4,541	5,841	7,453
4	0,741	1,533	2,132	2,776	3,495	3,747	4,604	5,598
5	0,727	1,476	2,015	2,571	3,163	3,365	4,032	4,773
6	0,718	1,440	1,943	2,447	2,969	3,143	3,707	4,317
7	0,711	1,415	1,895	2,365	2,841	2,998	3,500	4,029
8	0,706	1,397	1,860	2,306	2,752	2,896	3,355	3,833
9	0,703	1,383	1,833	2,262	2,685	2,821	3,250	3,690
10	0,700	1,372	1,813	2,228	2,634	2,764	3,169	3,581
11	0,697	1,363	1,796	2,201	2,503	2,718	3,106	3,497
12	0,695	1,356	1,782	2,179	2,560	2,681	3,055	3,428
13	0,694	1,350	1,771	2,160	2,533	2,650	3,012	3,373
14	0,692	1,345	1,761	2,145	2,510	2,624	2,977	3,326
15	0,691	1,341	1,753	2,132	2,490	2,602	2,947	3,286
16	0,690	1,337	1,746	2,120	2,473	2,583	2,921	3,252
17	0,689	1,333	1,740	2,110	2,458	2,567	2,898	3,223
18	0,688	1,330	1,734	2,101	2,445	2,552	2,878	3,197
19	0,688	1,328	1,729	2,093	2,433	2,539	2,861	3,174
20	0,687	1,325	1,725	2,086	2,423	2,528	2,845	3,153
21	0,686	1,323	1,721	2,080	2,414	2,518	2,831	3,135
22	0,686	1,321	1,717	2,074	2,406	2,508	2,819	3,119
23	0,685	1,319	1,714	2,069	2,398	2,500	2,807	3,104
24	0,685	1,318	1,711	2,064	2,391	2,492	2,797	3,091
25	0,684	1,316	1,708	2,060	2,385	2,485	2,787	3,078
26	0,684	1,315	1,706	2,056	2,379	2,479	2,779	3,067
27	0,684	1,314	1,703	2,052	2,373	2,473	2,771	3,057

Coeficiente de determinação (r^2)

- Após ajustar a equação da reta é necessário verificar a qualidade deste ajustamento.

- O coeficiente de determinação, denotado por r^2, expressa a proporção da variação total de Y que é explicada pela regressão, ou seja, pelo efeito linear de X sobre Y.

- O coeficiente de determinação é sempre positivo, variando entre 0 e 1 ($0 \leq r^2 \leq 1$), e deve ser interpretado como a proporção da variação total da variável dependente Y que é explicada pelo modelo de regressão. Quanto mais próximo de 1, melhor o ajustamento.

- O coeficiente de determinação (r^2) pode ser obtido elevando-se ao quadrado o coeficiente de correlação linear de Pearson (r).

Análise de regressão

O objetivo da análise de regressão é **determinar a equação (reta) que melhor representa a relação existente entre duas variáveis** e, a partir desta equação, fazer predições para a variável resposta.

Para isso, uma sequência de passos deve ser seguida:

1. Obtenção das estimativas (pontuais) dos coeficientes β_0 e β_1 para ajustar a equação da regressão.

2. Aplicação de **testes de hipóteses** para as estimativas obtidas, a fim de verificar se a equação de regressão é adequada.

3. Construção de **intervalos de confiança** para os valores estimados pela equação de regressão.

Intervalo de confiança para β_1

Podemos verificar se a estimativa de β_1 difere significativamente de zero construindo o intervalo de confiança para β_1.

Para isso, utilizamos a expressão geral do intervalo de confiança:

$$IC(\theta;1-\alpha): \hat{\theta} \pm t_{\alpha/2} S(\hat{\theta})$$

Fazendo as substituições referentes ao parâmetro β_1, temos

$$IC(\beta_1;1-\alpha): \hat{\beta}_1 \pm t_{\alpha/2} S(\hat{\beta}_1)$$

$$IC(\beta_1;1-\alpha): \hat{\beta}_1 \pm t_{\alpha/2} \sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{SQX}}}$$

No exemplo:

$$IC(\beta_1;1-\alpha): \hat{\beta}_1 \pm t_{\alpha/2}\sqrt{\frac{\sum \hat{e}_i^2}{n-2}}{SQX}$$

$$IC(\beta_1;0,95): -0,325 \pm 2,228 \times \sqrt{\frac{\frac{3,2545}{10}}{143}}$$

$$IC(\beta_1;0,95): -0,325 \pm 0,0996$$

$$LI = -0,325 - 0,0996 = -0,4246$$

$$LS = -0,325 + 0,0996 = -0,2254$$

$$IC(\beta_1;0,95): [-0,4246; -0,2254]$$

Concluímos, com 95% de confiança, que o coeficiente de regressão populacional é coberto pelos limites -0,4246 a -0,2254.
Esse intervalo não inclui o zero, portanto, esta conclusão está de acordo com a do teste de hipótese que rejeitou a hipótese de que $\beta_1 = 0$.

Intervalo de confiança para β_0

$$IC(\theta;1-\alpha): \hat{\theta} \pm t_{\alpha/2} S(\hat{\theta})$$

$$IC(\beta_0;1-\alpha): \hat{\beta}_0 \pm t_{\alpha/2} S(\hat{\beta}_0)$$

$$IC(\beta_0;1-\alpha): \hat{\beta}_0 \pm t_{\alpha/2} \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right] \frac{\sum \hat{e}_i^2}{n-2}}$$

$$IC(\beta_0;1-\alpha): \hat{\beta}_0 \pm t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SQX}\right) \frac{\sum \hat{e}_i^2}{n-2}}$$

No exemplo:

$$IC(\beta_0;1-\alpha): \hat{\beta}_0 \pm t_{\alpha/2} S(\hat{\beta}_0)$$

$$IC(\beta_0;1-\alpha): \hat{\beta}_0 \pm t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right) \frac{\sum \hat{e}_i^2}{n-2}}$$

$$IC(\beta_0;0{,}95): 11{,}34 \pm 2{,}228 \sqrt{\left(\frac{1}{12} + \frac{6{,}5^2}{143}\right) \frac{3{,}2545}{10}}$$

$$IC(\beta_0;0{,}95): 11{,}34 \pm 2{,}228 \sqrt{0{,}3788 \times 0{,}32545}$$

$$IC(\beta_0;0{,}95): 11{,}34 \pm 0{,}7823$$

$$LI = 11{,}34 - 0{,}7823 = 10{,}56$$

$$LS = 11{,}34 + 0{,}7823 = 12{,}12$$

$$IC(\beta_0;0{,}95): [10{,}56;\ 12{,}12]$$

Concluímos, com 95% de confiança, que o intervalo de 10,56 a 12,12 contém o intercepto populacional, ou seja, no tempo zero após a regulagem o rendimento populacional está entre 10,56 e 12,12.

Predição de médias de Y

Intervalo de confiança para médias da população μ_i

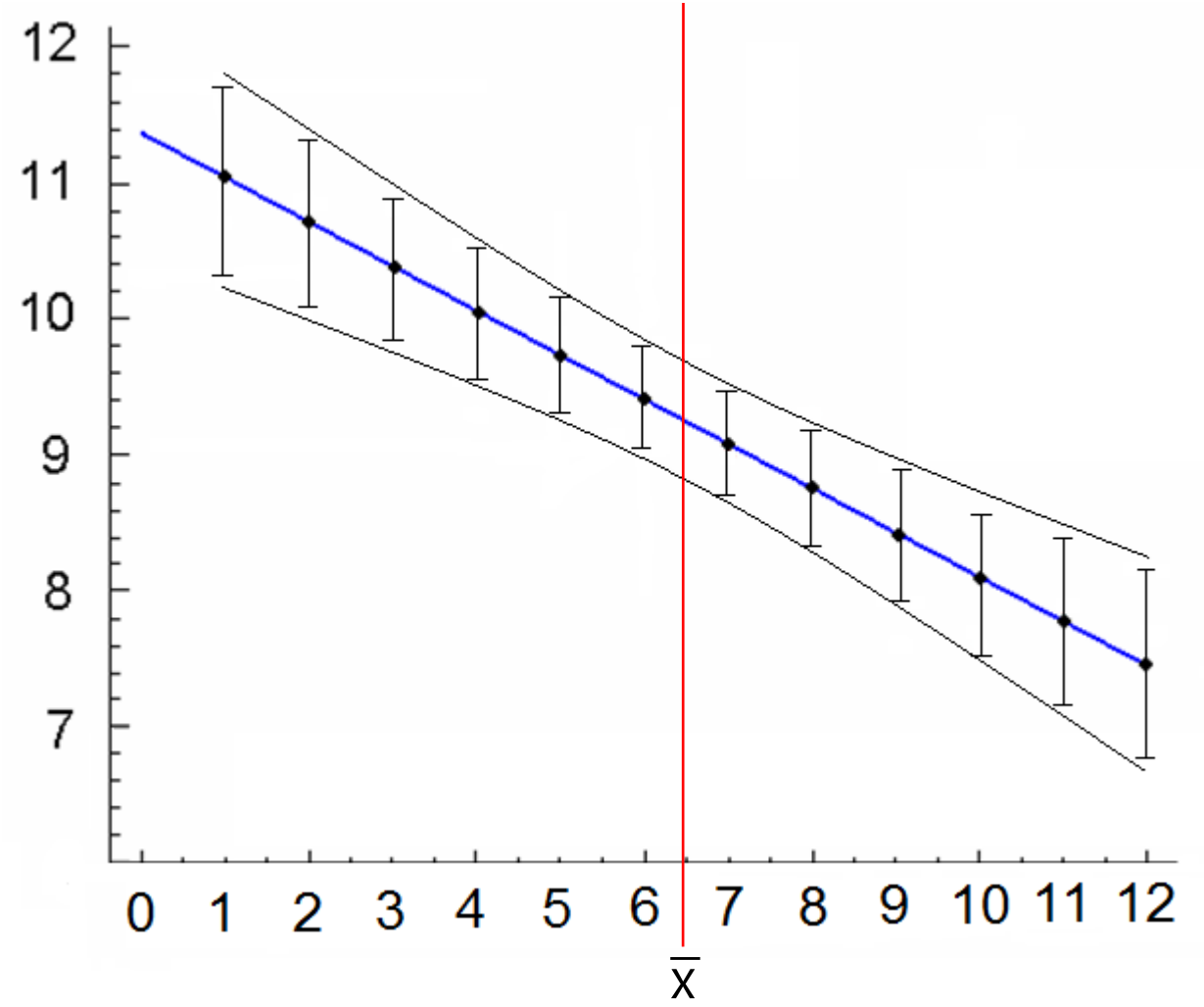
$$IC(\theta;1-\alpha): \hat{\theta} \pm t_{\alpha/2} S(\hat{\theta})$$

$$IC(\mu_i;1-\alpha): \hat{\mu}_i \pm t_{\alpha/2} S(\hat{\mu}_i)$$

$$IC(\mu_i;1-\alpha): \hat{\mu}_i \pm t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)}$$

$$IC(\mu_i;1-\alpha): \hat{\mu}_i \pm t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SQX}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)}$$

Como pode ser visto, a variância da predição é mínima quando $x_i = \bar{x}$ e aumenta quando x_i se afasta de $\bar{x}$.



Como pode ser visto, a variância da predição é mínima quando $x_i = \overline{x}$ e aumenta quando x_i se afasta de $\overline{x}$.

No exemplo: Intervalo de confiança para μ_1

$$IC(\mu_1;0{,}95): \hat{\mu}_1 \pm t_{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{(x_1-\bar{x})^2}{SQX}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)}$$

$$IC(\mu_1;0{,}95): 11{,}02 \pm 2{,}228\sqrt{\left(\frac{1}{12}+\frac{(1-6{,}5)^2}{143}\right)\left(\frac{3{,}2545}{10}\right)}$$

$$IC(\mu_1;0{,}95): 11{,}02 \pm 2{,}228 \times 0{,}3097$$

$$IC(\mu_1;0{,}95): 11{,}02 \pm 0{,}6902$$

$$LI = 11{,}02 - 0{,}6902 = 10{,}32$$

$$LS = 11{,}02 + 0{,}6902 = 11{,}71$$

$$IC(\mu_1;0{,}95): [10{,}32;\ 11{,}71]$$

Concluímos, com 95% de confiança, que o intervalo de 10,32 a 11,71 contém a média populacional de Y para X=1. Isto significa que temos 95% de confiança de que um mês após a regulagem o rendimento médio deste modelo de carro estará no intervalo de 10,32 a 11,71.

No exemplo: Intervalo de confiança para μ_6

$$IC(\mu_6;0{,}95): \hat{\mu}_6 \pm t_{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{(x_6-\bar{x})^2}{SQX}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)}$$

$$IC(\mu_6;0{,}95): 9{,}39 \pm 2{,}228\sqrt{\left(\frac{1}{12}+\frac{(6-6{,}5)^2}{143}\right)\left(\frac{3{,}2545}{10}\right)}$$

$$IC(\mu_6;0{,}95): 9{,}39 \pm 2{,}228 \times 0{,}1664$$

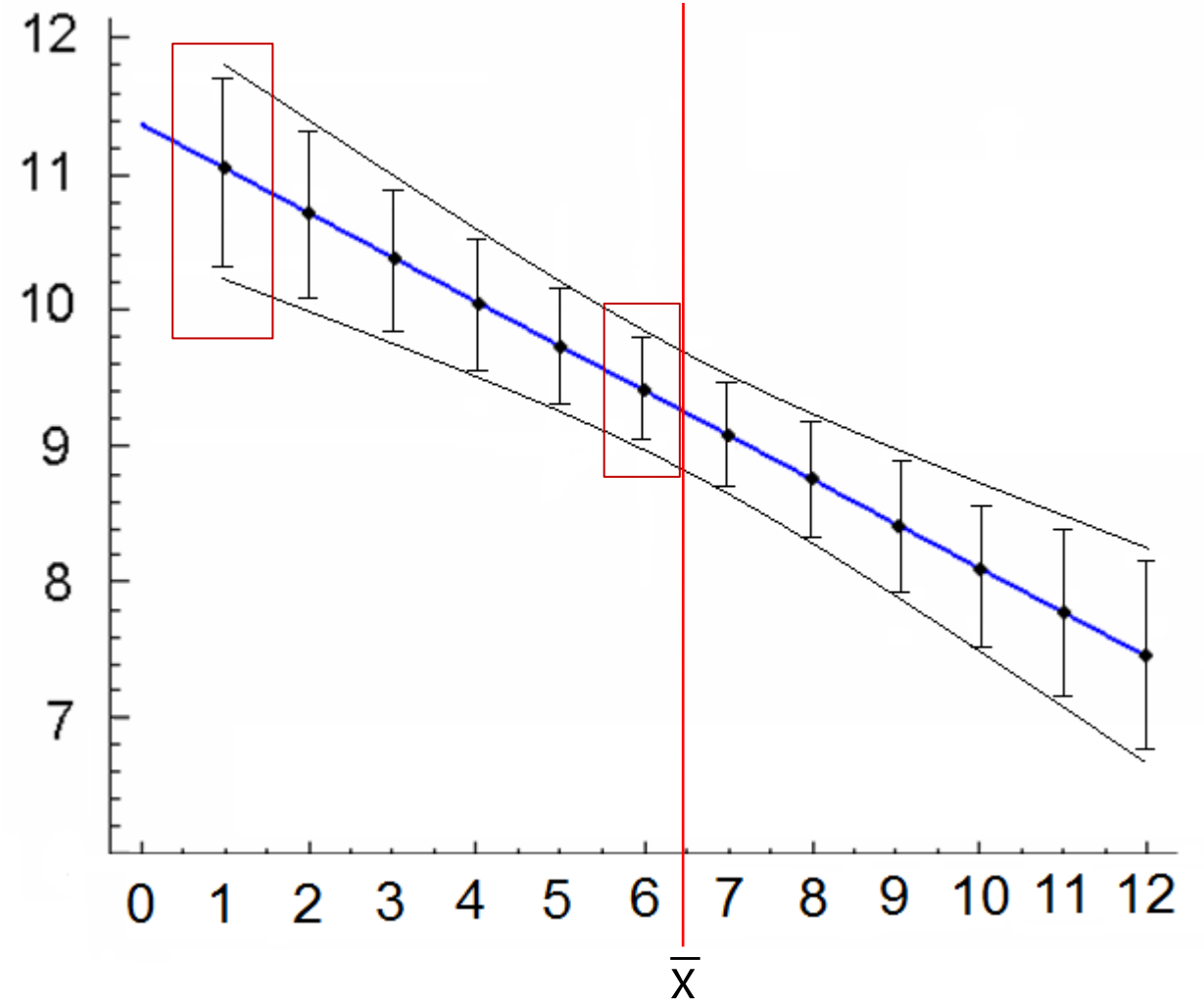
$$IC(\mu_6;0{,}95): 9{,}39 \pm 0{,}3707$$

$$LI = 9{,}39 - 0{,}3707 = 9{,}02$$

$$LS = 9{,}39 + 0{,}3707 = 9{,}76$$

$$IC(\mu_6;0{,}95): [9{,}02;\ 9{,}76]$$

Concluímos, com 95% de confiança, que o intervalo de 9,02 a 9,76 contém a média populacional de Y para X=6. Isto significa que temos 95% de confiança de que seis mês após a regulagem o rendimento médio deste modelo de carro estará no intervalo de 9,02 a 9,76.



$IC(\mu_1;0{,}95) : [10{,}32;\ 11{,}71]$

$IC(\mu_6;0{,}95) : [9{,}02;\ 9{,}76]$

Extrapolação da equação de regressão

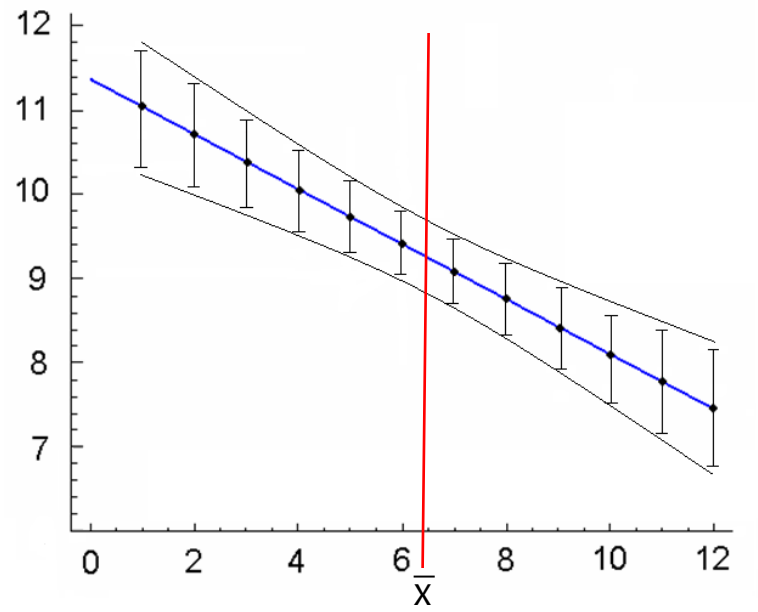
De uma forma geral, não é recomendada a extrapolação da equação de regressão para além dos limites dos dados amostrais utilizados na estimativa dos parâmetros do modelo de regressão linear.

O desestímulo à extrapolação apresenta basicamente dois motivos:

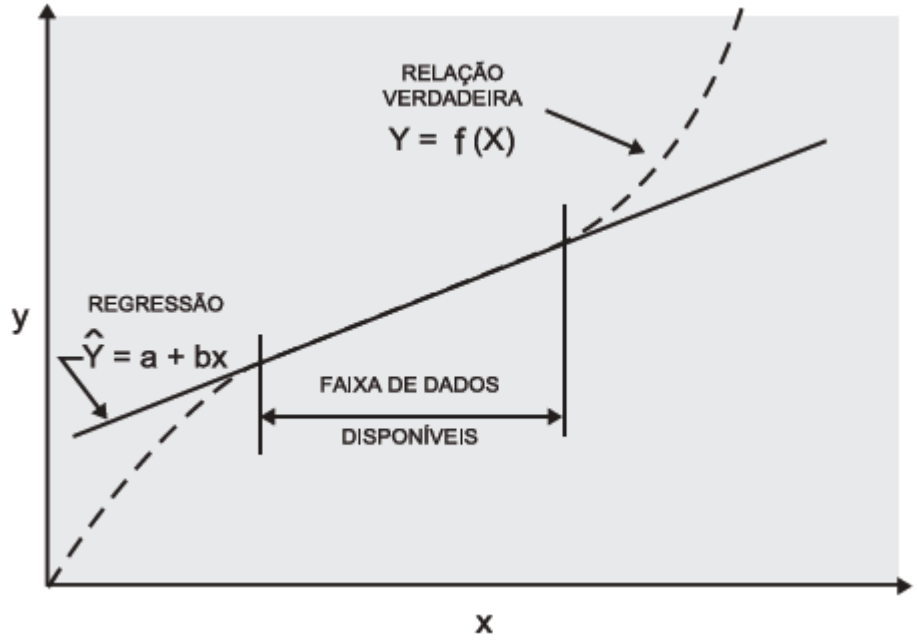
1. A amplitude do intervalo de confiança sobre a linha de regressão aumenta à medida que os valores da variável X se afastam da média;

2. A relação entre as variáveis X e Y pode não ser linear para valores que extrapolam os dados utilizados na regressão.

(1)



(2)



Considerações finais

- A variância da inclinação $\hat{\beta}_1$ aumenta quando se reduz o intervalo de variação de X.
- Se o intervalo é pequeno, $S(\hat{\beta}_1)$ será grande e nesse caso será difícil rejeitar a hipótese H_0: $\beta_1 = 0$. Em outras palavras, se a relação entre X e Y é medida em um intervalo reduzido de X, os parâmetros estimados não terão muito significado estatístico.
- **Se o objetivo é construir um modelo de regressão, deve-se coletar dados nos extremos do intervalo de X, ou seja, nos limites de interesse e viabilidade práticos ou nos limites em que se supõe válida a relação linear.**

Bibliografia consultada

NAGHETTINI, M.; PINTO, E. J. de A. **Hidrologia estatística**. Belo Horizonte: CPRM, 2007. 552 p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br