



Probabilidade e inferência estatística com R - Módulo 3

Prof. Suellen Teixeira Zavadzki de Pauli

Objetivos





- Compreender os conceitos de correlação e regressão;
- Avaliar a correlação entre variáveis por meio de gráficos e teste;
- Estimar e visualizar um modelo de regressão;
- Interpretar coeficientes de regressão e estatísticas no contexto de problemas reais;
- Compreender os conceitos da Análise de Variância.

Correlação e Regressão Linear Simples

- Em determinadas situações, estamos interessados em descrever a relação entre duas variáveis ou até prever o valor de uma a partir da outra.
- **Exemplos:**
 - Qual o peso de determinado indivíduo se sabemos que a altura dele é X ?
 - Qual o consumo de combustível, em litros, dado que o carro percorreu uma distância de X km?
 - Qual a relação entre a renda semanal de uma família e as despesas de consumo?

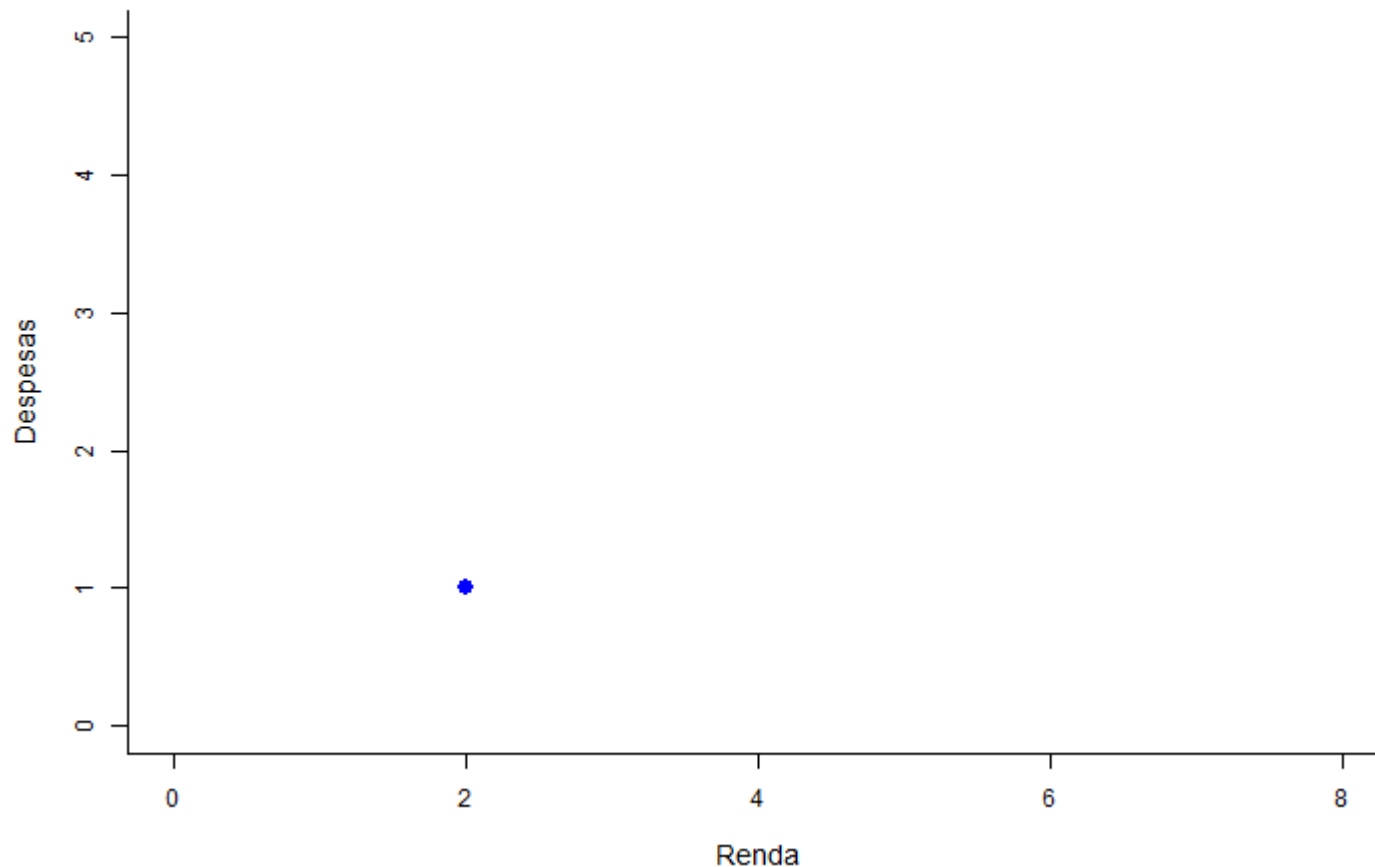
Introdução












	RENDA	DESPESAS
FAMÍLIA 1		

Introdução

Renda x Despesas

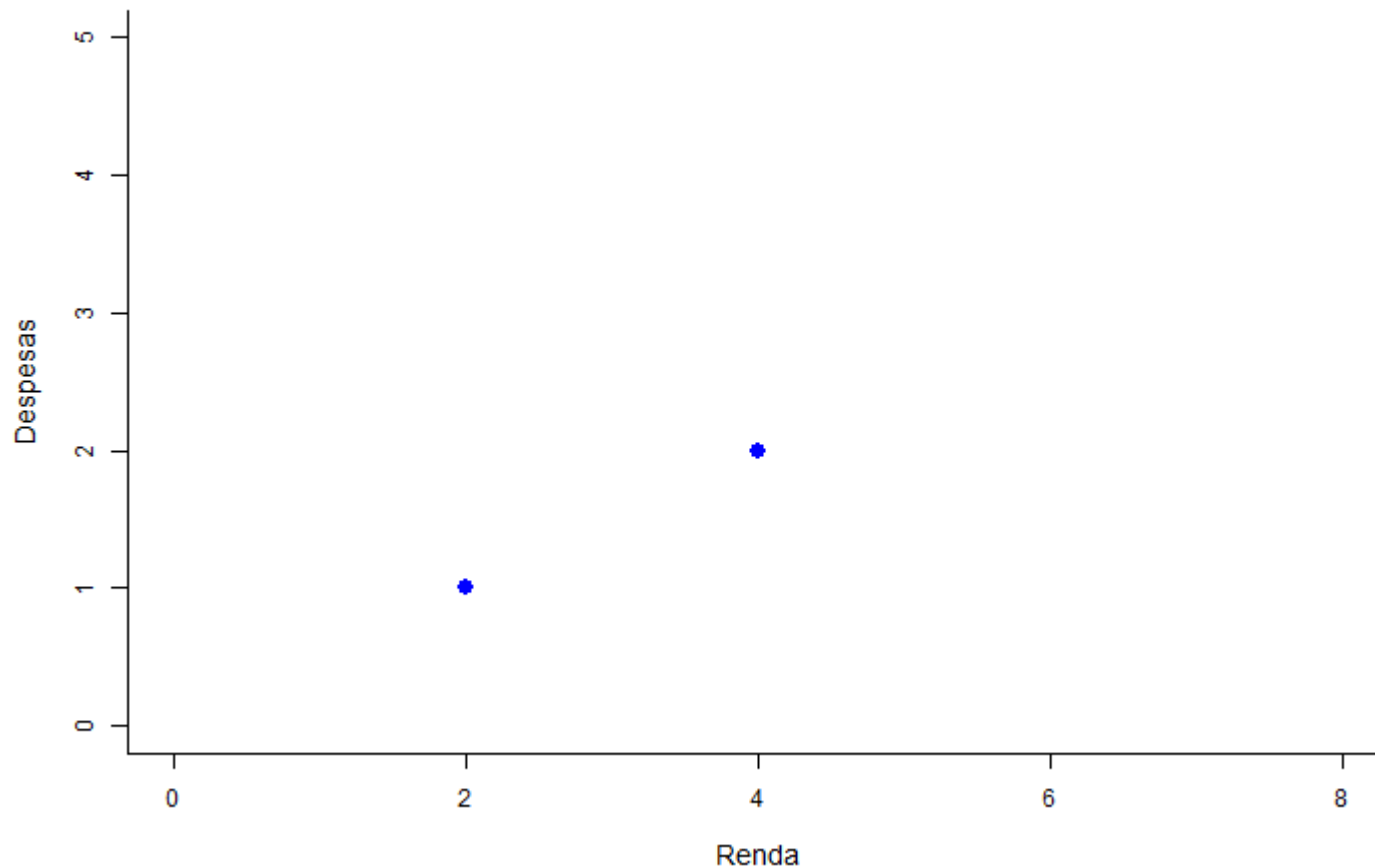


Introdução







	RENDA	DESPESAS
FAMÍLIA 1	 	
FAMÍLIA 2	   	 

Introdução

Renda x Despesas

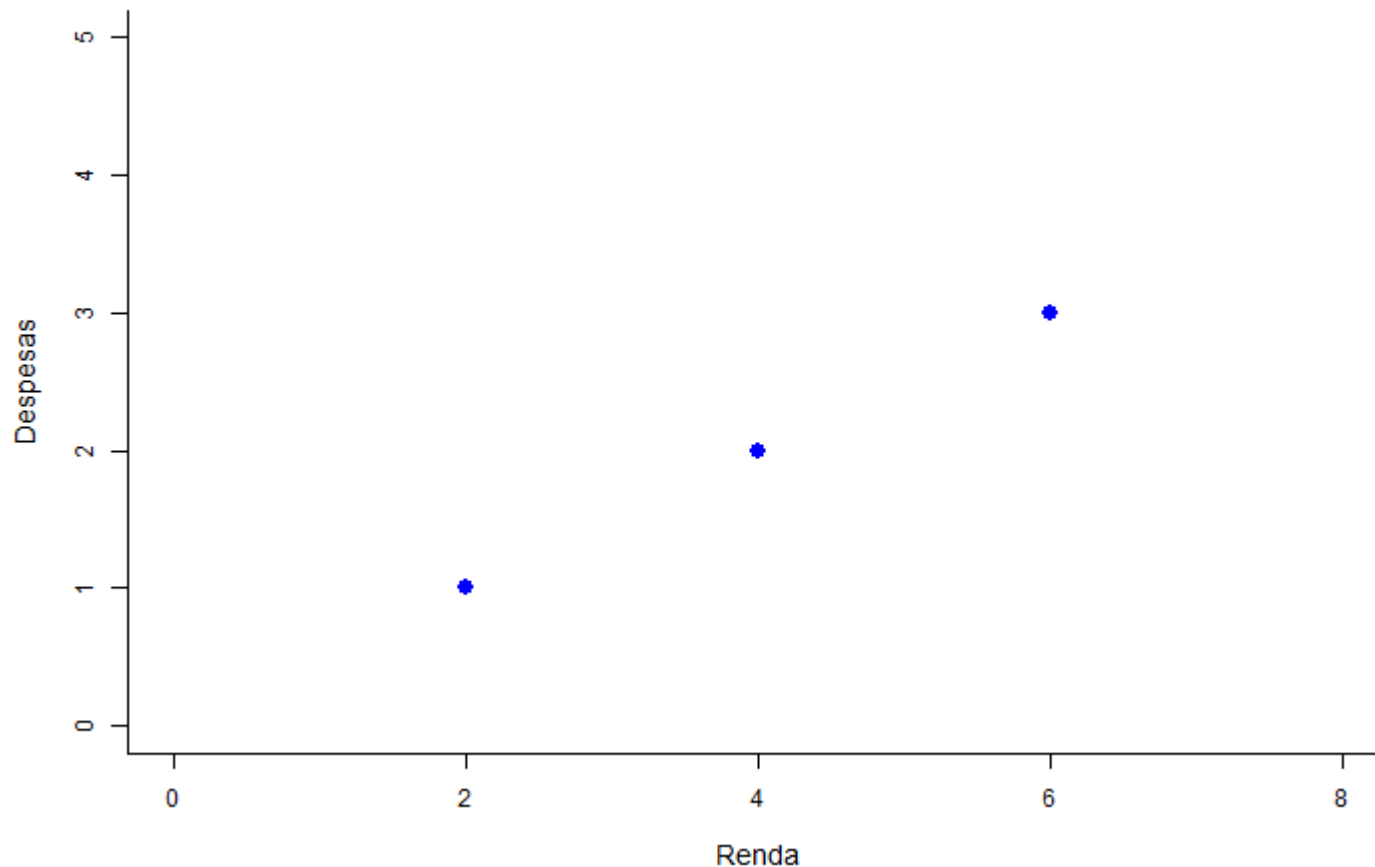


Introdução

	RENDA	DESPESAS
FAMÍLIA 1		
FAMÍLIA 2		
FAMÍLIA 3		

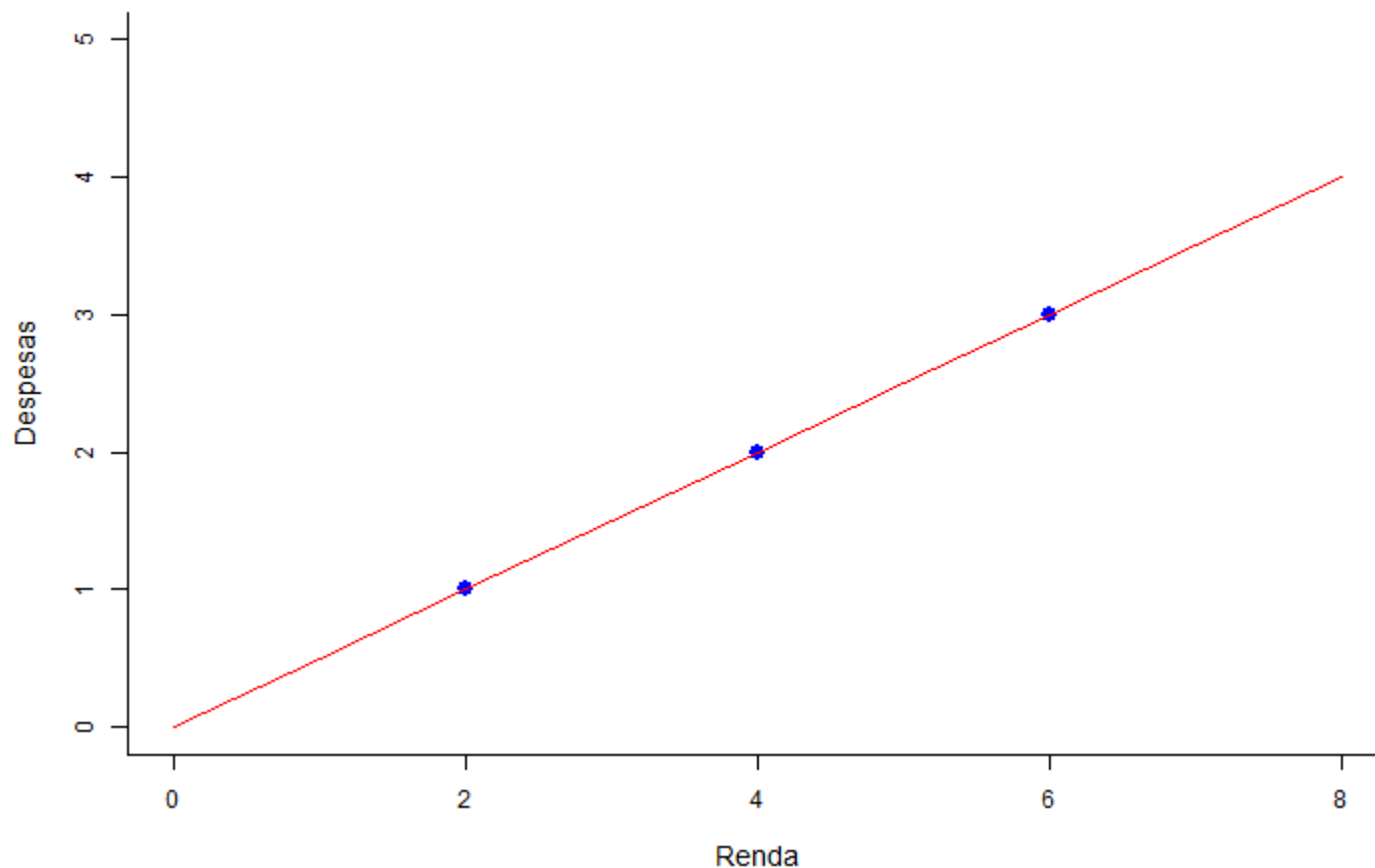
Introdução

Renda x Despesas



Introdução

Renda x Despesas

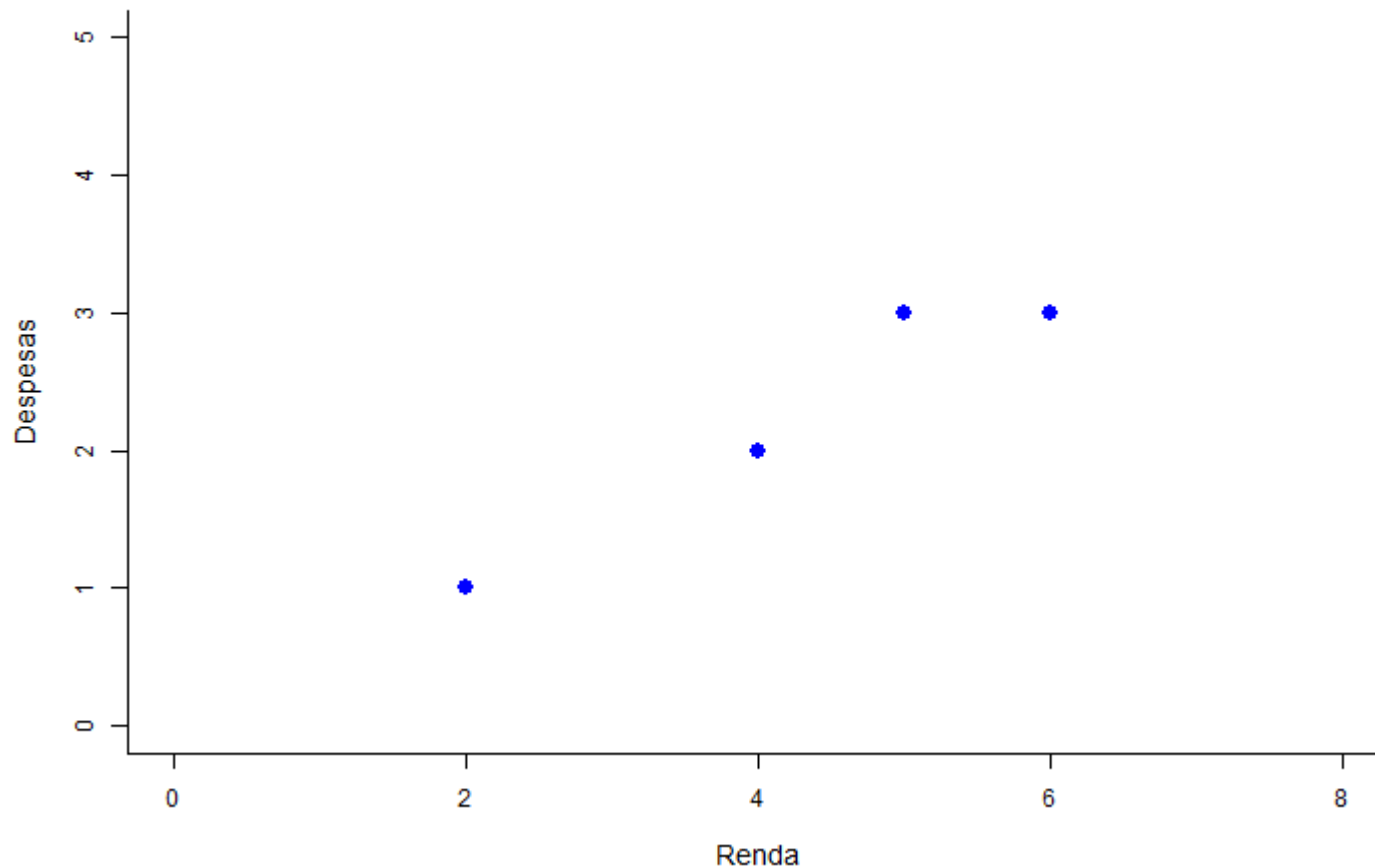


Introdução

	RENDA	DESPESAS
FAMÍLIA 1	2	1
FAMÍLIA 2	5	2
FAMÍLIA 3	8	3
FAMÍLIA 4	6	3

Introdução

Renda x Despesas

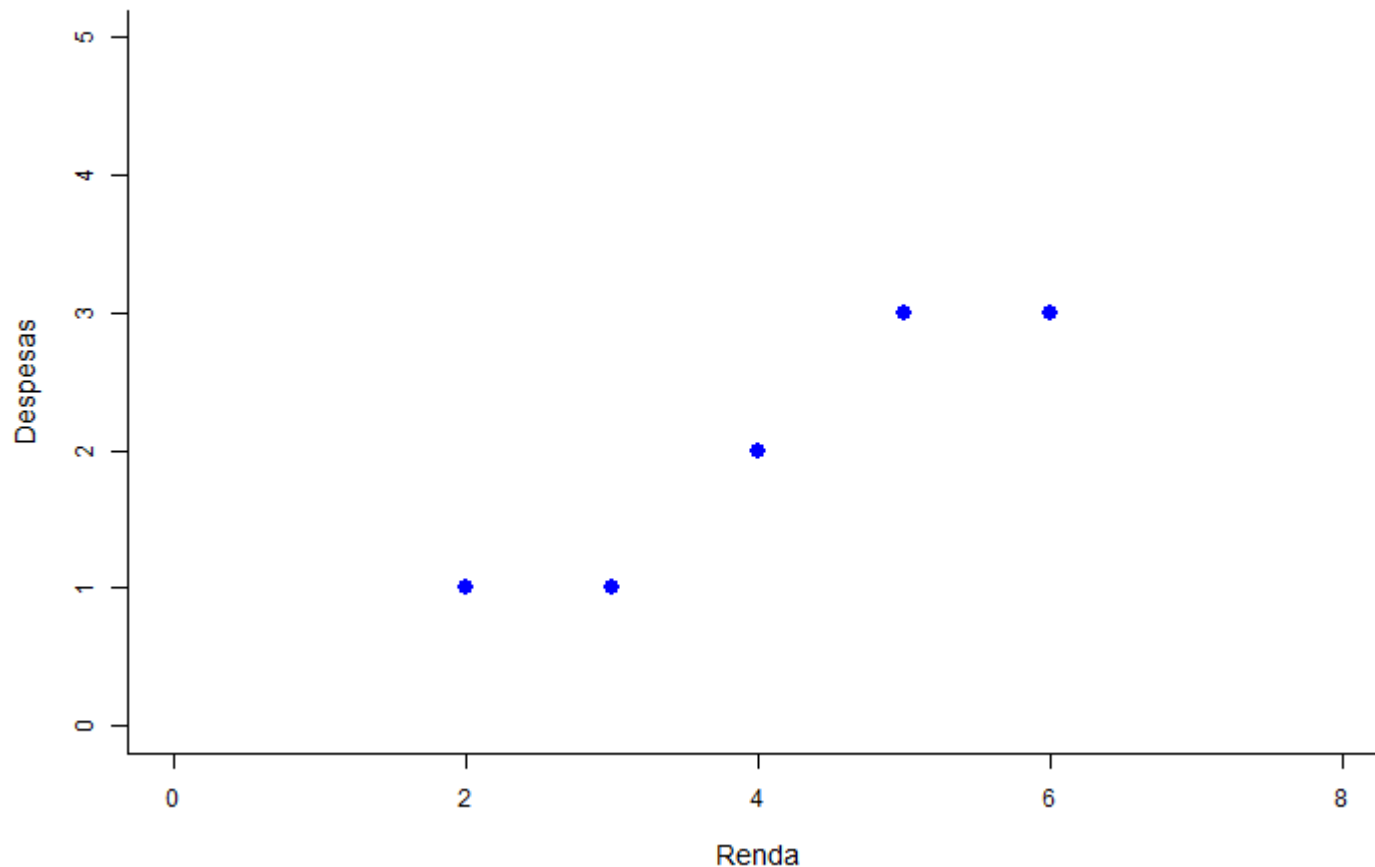


Introdução

	RENDA	DESPESAS
FAMÍLIA 1	2	1
FAMÍLIA 2	5	2
FAMÍLIA 3	7	3
FAMÍLIA 4	6	3
FAMÍLIA 5	3	1

Introdução

Renda x Despesas

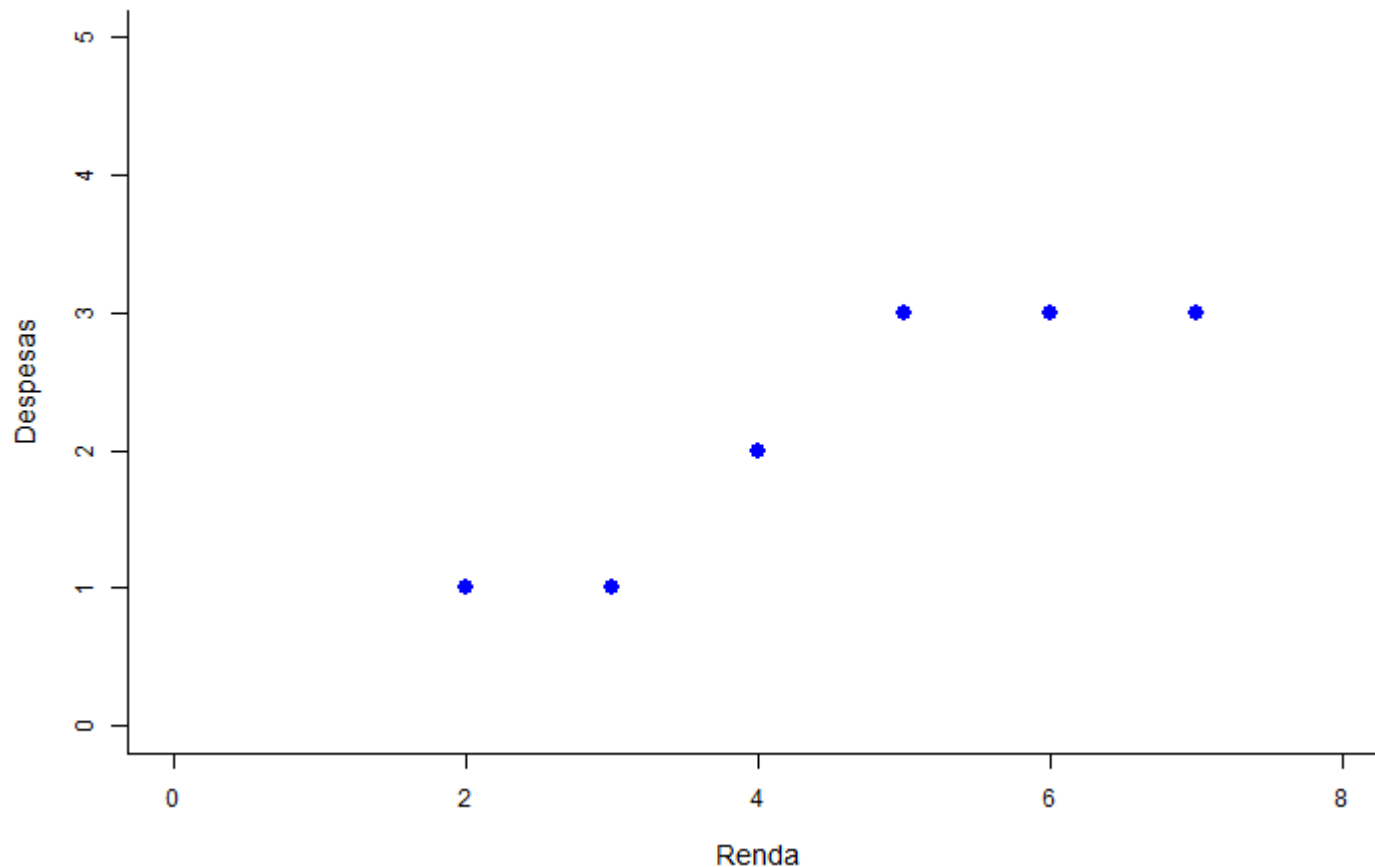


Introdução

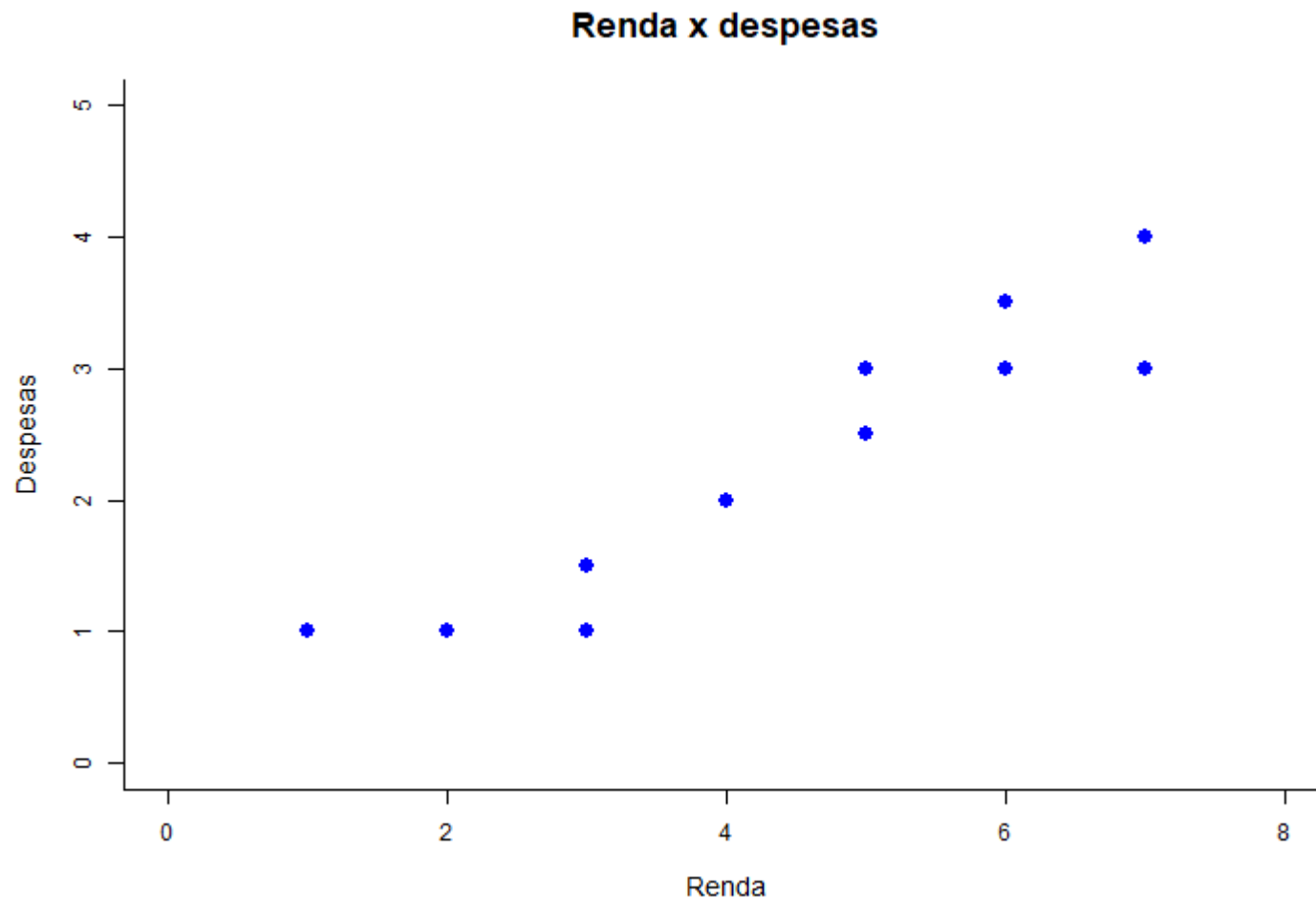
	RENDA	DESPESAS
FAMÍLIA 1	2	1
FAMÍLIA 2	4	2
FAMÍLIA 3	6	3
FAMÍLIA 4	5	3
FAMÍLIA 5	3	1
FAMÍLIA 6	6	3

Introdução

Renda x Despesas

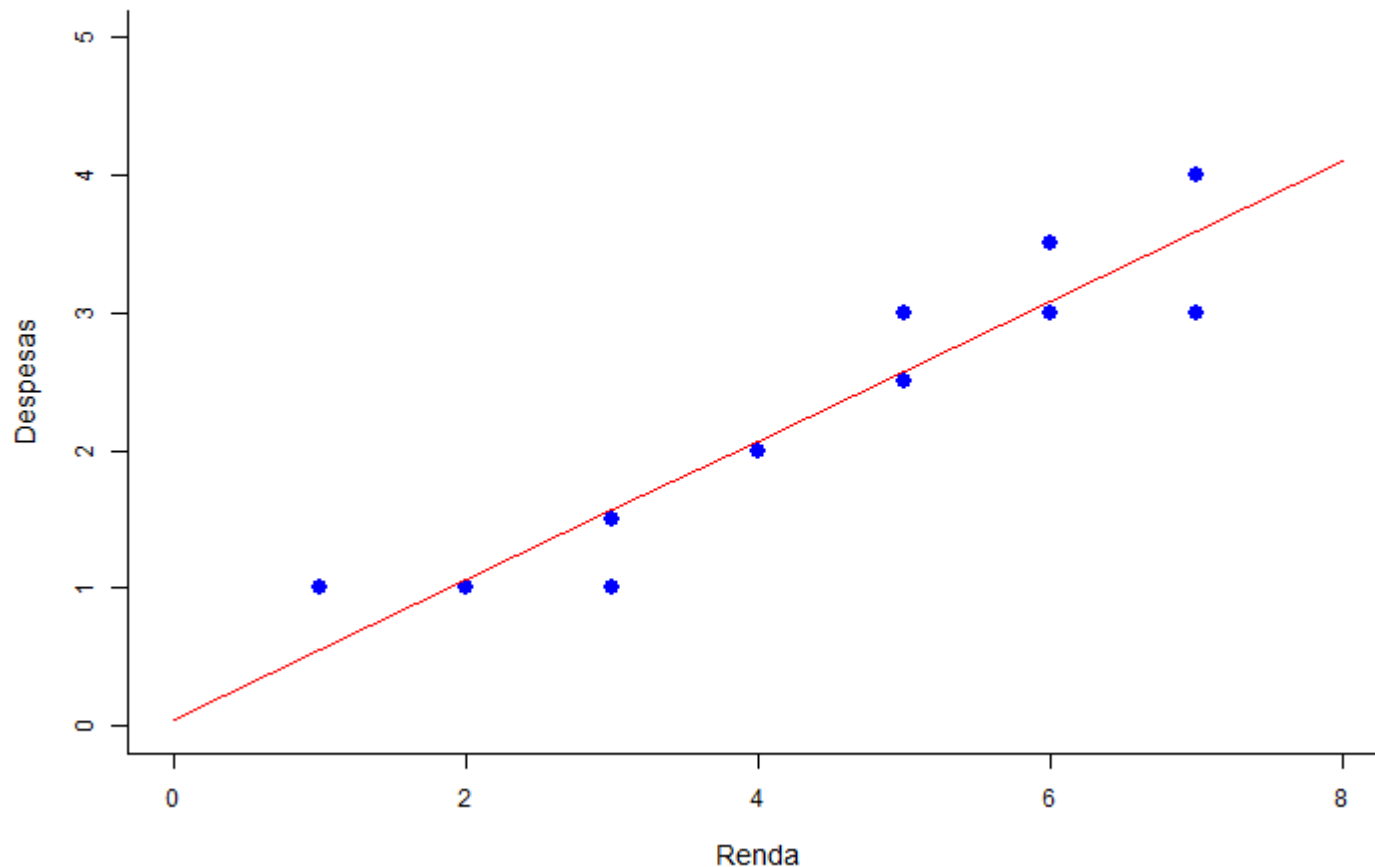


Introdução



Introdução

Renda x despesas



"All models are wrong but some are useful"

George Box

- Estudar a relação linear entre duas variáveis quantitativas:
 - Explicitando a forma dessa relação: **regressão**
 - É indispensável identificar qual variável é a variável dependente.
 - Quantificando a força ou o grau dessa relação: **correlação**
 - Não é necessário identificar qual variável é a variável dependente, pois queremos estudar o grau de relacionamento entre as variáveis X e Y , ou seja, uma medida de covariabilidade entre elas.
 - A correlação é considerada como uma medida de influência mútua entre variáveis, por isso não é necessário especificar quem influencia e quem é influenciado.

Diagrama de dispersão



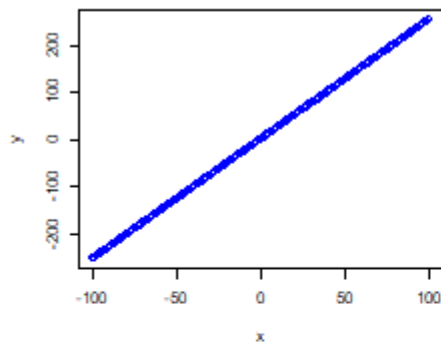
- Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

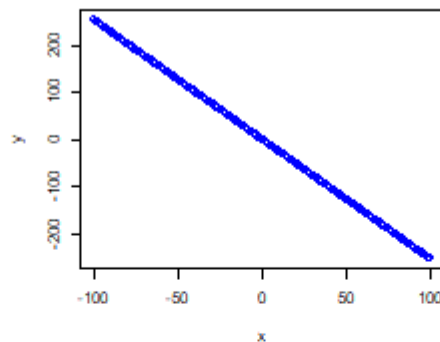
- Com base no conjunto de dados é possível construir um diagrama de dispersão, o qual deve exibir uma tendência linear para que se possa usar a regressão linear;
- Com isso podemos decidir empiricamente se um relacionamento linear entre X e Y pode ser assumido;
- É possível verificar se o grau de relacionamento linear entre as variáveis é forte ou fraco.

Diagrama de dispersão

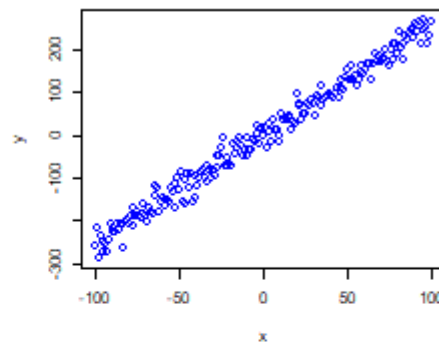
Correlação perfeita positiva



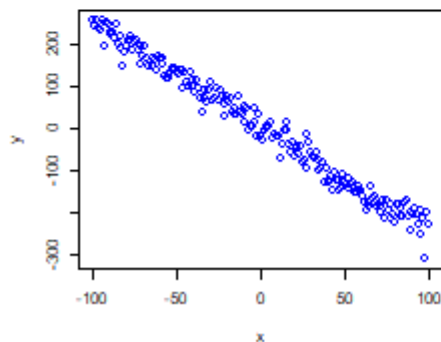
Correlação perfeita negativa



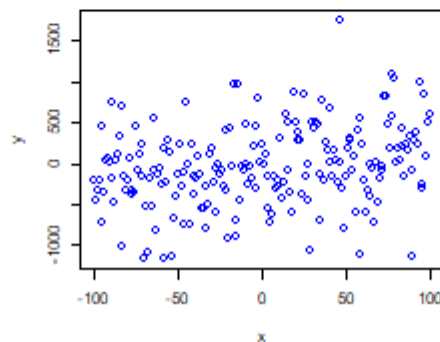
Alta correlação positiva



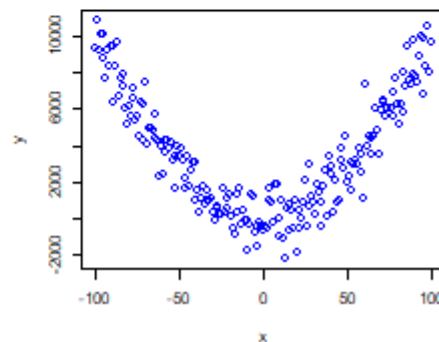
Alta correlação negativa



Baixa correlação



Correlação não linear



Coeficiente de correlação linear

O grau de relação entre duas variáveis pode ser medido através do coeficiente de correlação linear (r), dado por

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

onde $-1 \leq r \leq 1$;

- $r = 1$: relação linear perfeita positiva entre X e Y ;
- $r = 0$: : inexistência de relação linear entre X e Y ;
- $r = -1$: relação linear perfeita negativa entre X e Y ;
- $r > 0$: relação linear positiva entre X e Y ;
- $r < 0$: relação linear negativa entre X e Y ;

Coeficiente de determinação



- Existem muitos tipos de associações possíveis, e o coeficiente de correlação avalia o quanto uma nuvem de pontos no gráfico de dispersão se aproxima de uma reta;
- O coeficiente de determinação (r^2) é o quadrado do coeficiente de correlação, por consequência;

$$0 \leq r^2 \leq 1$$

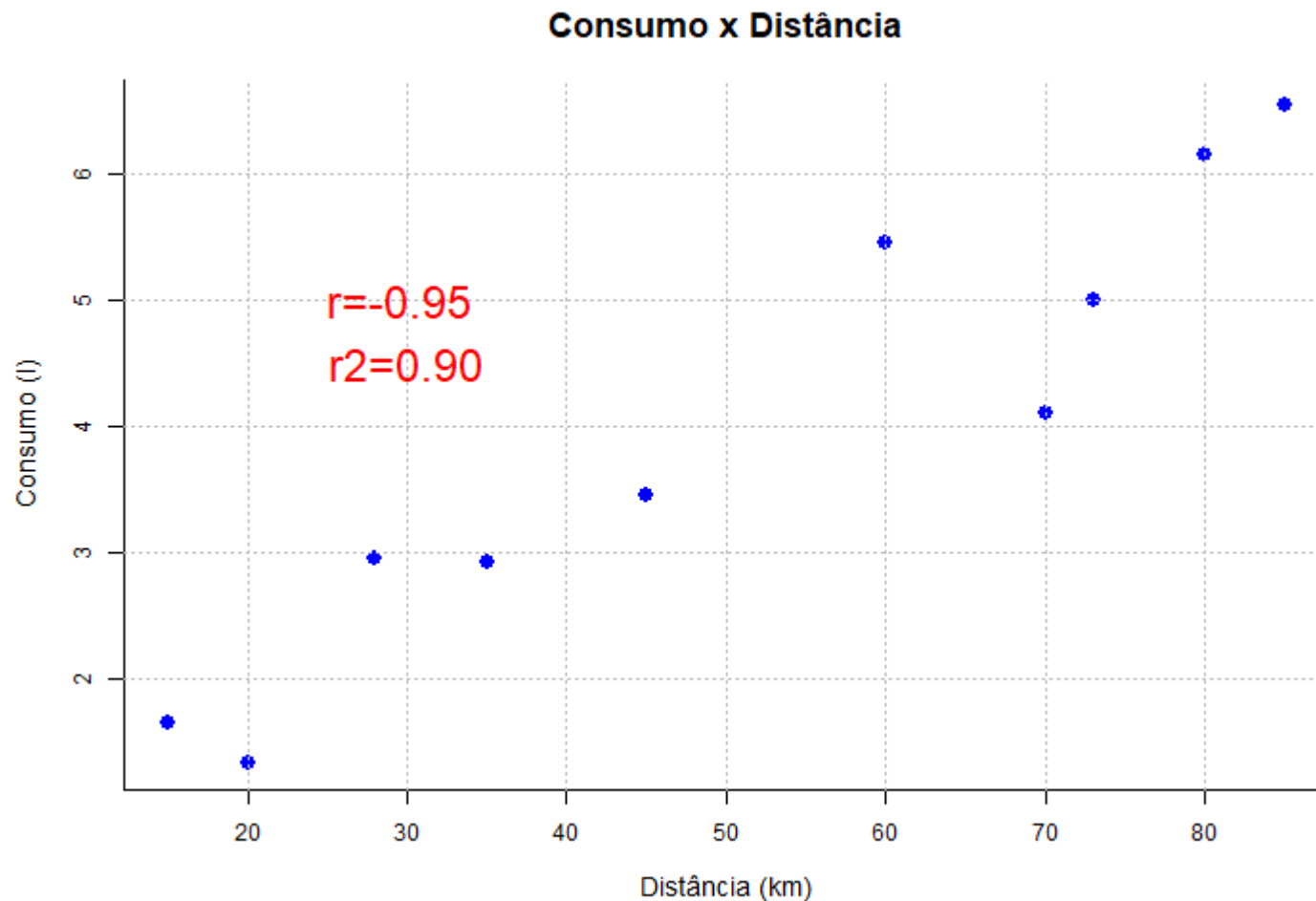
- O r^2 nos dá a porcentagem de variação em Y que pode ser explicada pela variável independente X .
- Quanto mais próximo de 1, maior é a explicação da variável Y pela variável X .

Exercício 1

- A tabela a seguir relaciona as distâncias percorridas por carros (km) e seus consumos de combustível (litros), em uma amostra de 10 carros novos.
 - Calcule o coeficiente de correlação linear e o coeficiente de determinação.
 - Faça um diagrama de dispersão.

<i>Distância</i>	20.00	60.00	15.00	45.00	35.00	80.00	70.00	73.00	28.00	85.00
<i>Consumo</i>	1.33	5.45	1.66	3.46	2.92	6.15	4.11	5.00	2.95	6.54

Exercício 1



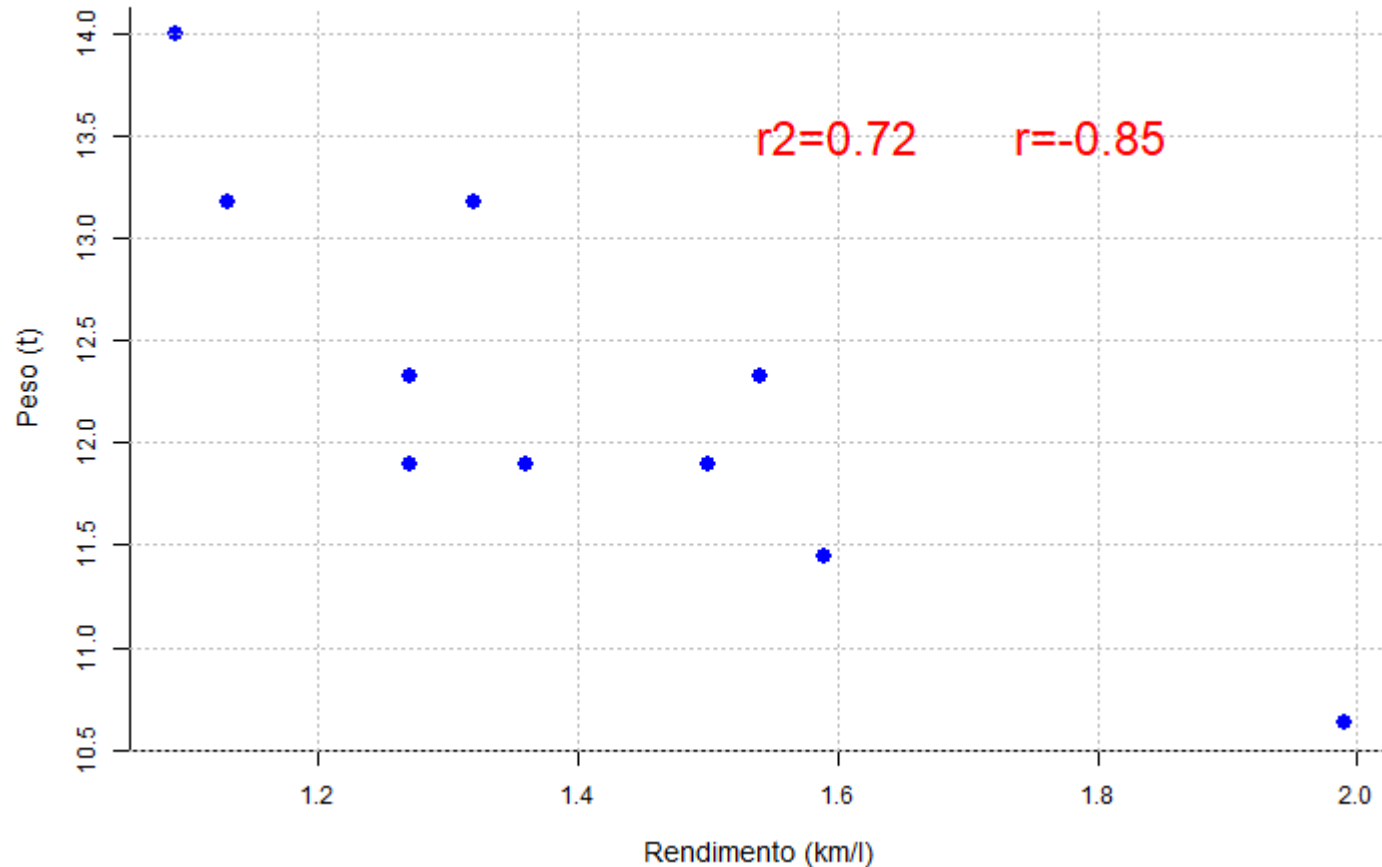
Exercício 2

- A tabela a seguir relaciona os pesos de carros (t) e o rendimento de combustível (em km/l), para uma amostra de 10 carros.
 - Calcule o coeficiente de correlação linear e o coeficiente de determinação.
 - Faça um diagrama de dispersão.

<i>Peso</i>	1.32	1.59	1.27	1.99	1.13	1.54	1.36	1.5	1.27	1.09
<i>Rendimento</i>	13.18	11.45	12.33	10.63	13.18	12.33	11.90	11.9	11.90	14.00

Exercício 2

Peso x Rendimento



Teste para o coeficiente de correlação



- Usualmente definimos o coeficiente de correlação para uma amostra, pois desconhecemos esse valor para a população.
- Uma população que tenha duas variáveis não correlacionadas pode produzir uma amostra com coeficiente de correlação diferente de zero.
- Para testar se uma amostra foi colhida de uma população para a qual o coeficiente de correlação entre duas variáveis é nulo, precisamos obter a distribuição amostral da estatística r .

Teste para o coeficiente de correlação

- Seja ρ o verdadeiro coeficiente de correlação populacional desconhecido. Seja ρ o verdadeiro coeficiente de correlação populacional desconhecido.
- Para testar se o coeficiente de correlação populacional é igual a zero, realizamos um teste de hipótese com

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

A estatística de teste utilizada é

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

que tem distribuição t de Student com $n - 2$ graus de liberdade.

Teste para o coeficiente de correlação

Procedimentos gerais

- Hipóteses $H_0 : \rho = 0, H_1 : \rho \neq 0$
- Nível de significância α
- Verificar a região de rejeição com base no nível de significância t_{crit} , com $n - 2$ graus de liberdade
- Cálculo da estatística do teste sob a hipótese nula

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

- Rejeitar a hipótese nula se a estatística de teste calculada estiver dentro da região de rejeição ou $|t_{calc}| > |t_{crit}|$

Exercício 1 - continuação



Com base nas informações do exercício 1, realize o teste de hipótese para o coeficiente de correlação ρ , usando um nível de 5% de significância.

$$r = 0.95$$

Exercício 1 - continuação



Com base nas informações do exercício 1, realize o teste de hipótese para o coeficiente de correlação ρ , usando um nível de 5% de significância.

$$r = 0.95$$

$$t_{calc} > t_{crit}$$

$$8.605 > 2.306$$

Rejeitamos a hipótese nula de que não há correlação entre as variáveis, com 95% de confiança.

Exercício 2 - continuação



Com base nas informações do exercício 2, realize o teste de hipótese para o coeficiente de correlação ρ , usando um nível de 5% de significância.

$$r = 0.85$$

Exercício 2 - continuação



Com base nas informações do exercício 2, realize o teste de hipótese para o coeficiente de correlação ρ , usando um nível de 5% de significância.

$$r = -0.85$$

$$t_{calc} = -4.563861$$

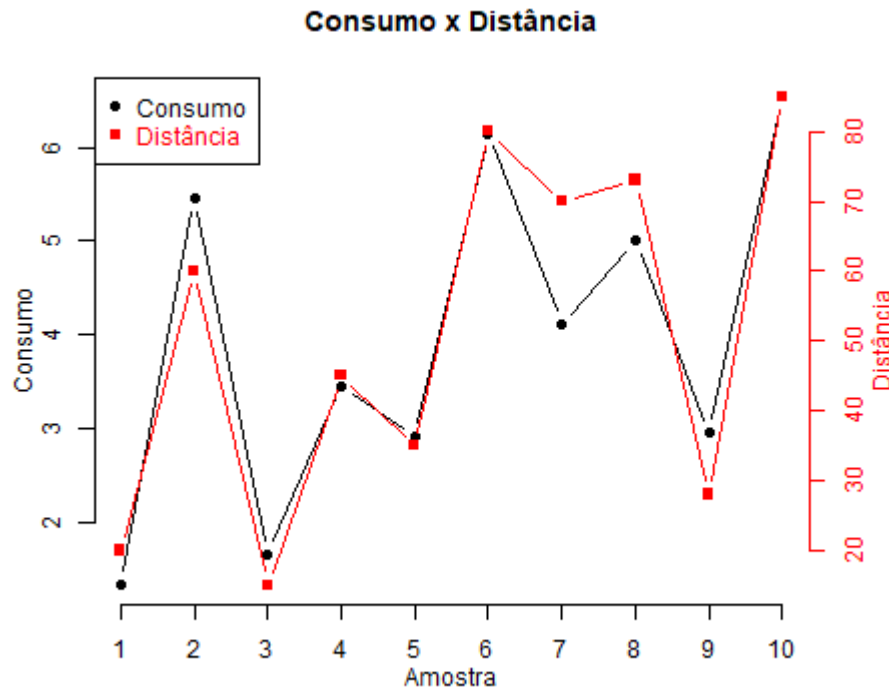
$$|t_{calc}| > |t_{crit}|$$

$$4.56 > 2.306$$

Rejeitamos a hipótese nula de que não há correlação entre as variáveis, com 95% de confiança.

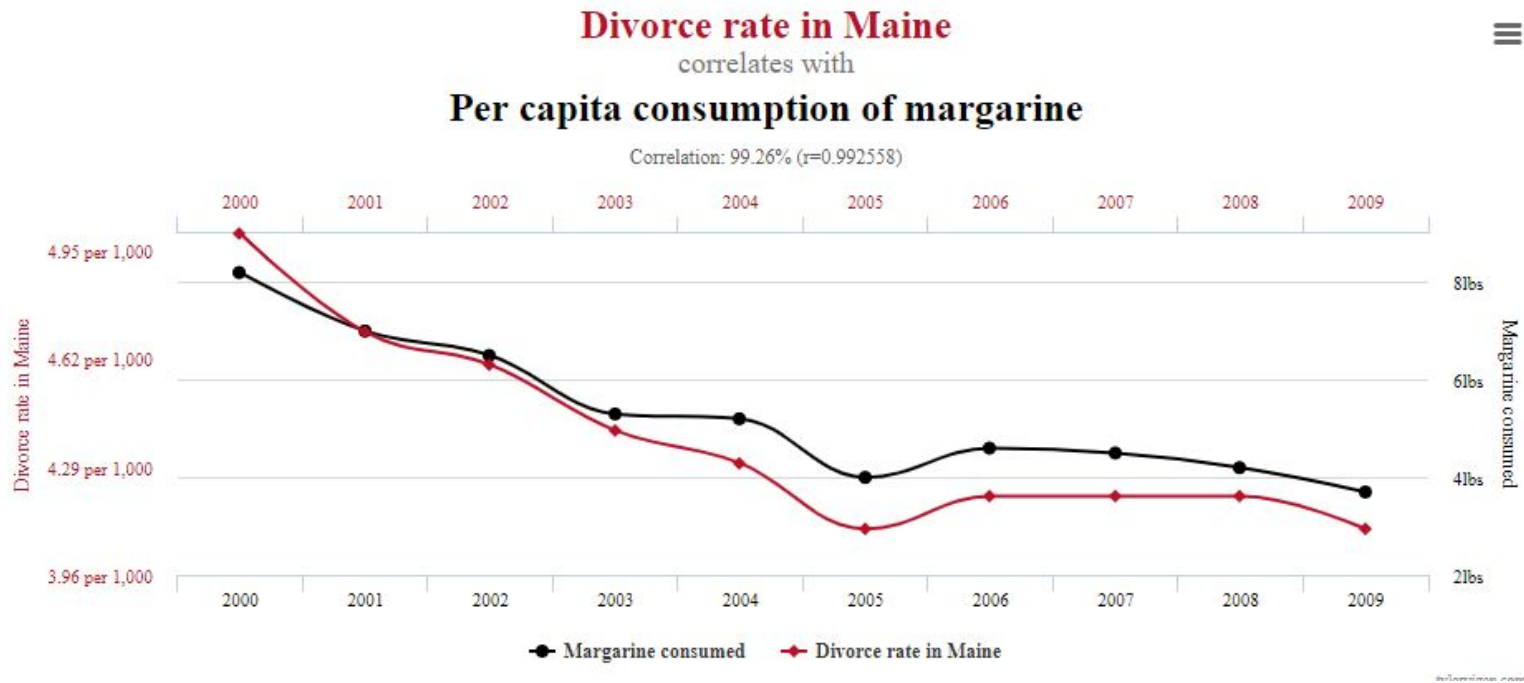
Exercício 1 - continuação

- Construa um gráfico no qual seja possível visualizar os valores das duas variáveis no eixo y.



Correlação x Causalidade ?

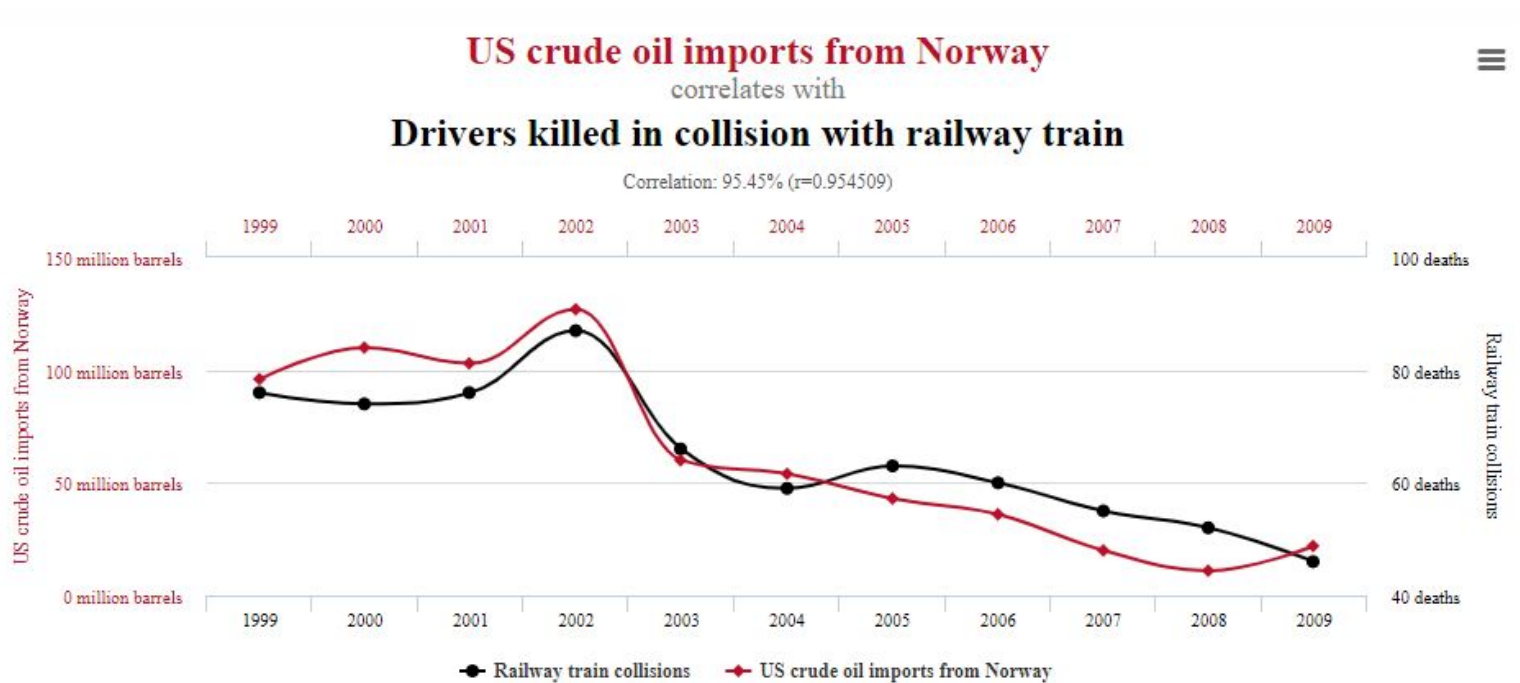
Taxa de divórcio no Maine se correlaciona com o consumo per capita de margarina



Fonte: <https://www.tylervigen.com/spurious-correlations>

Correlação x Causalidade ?

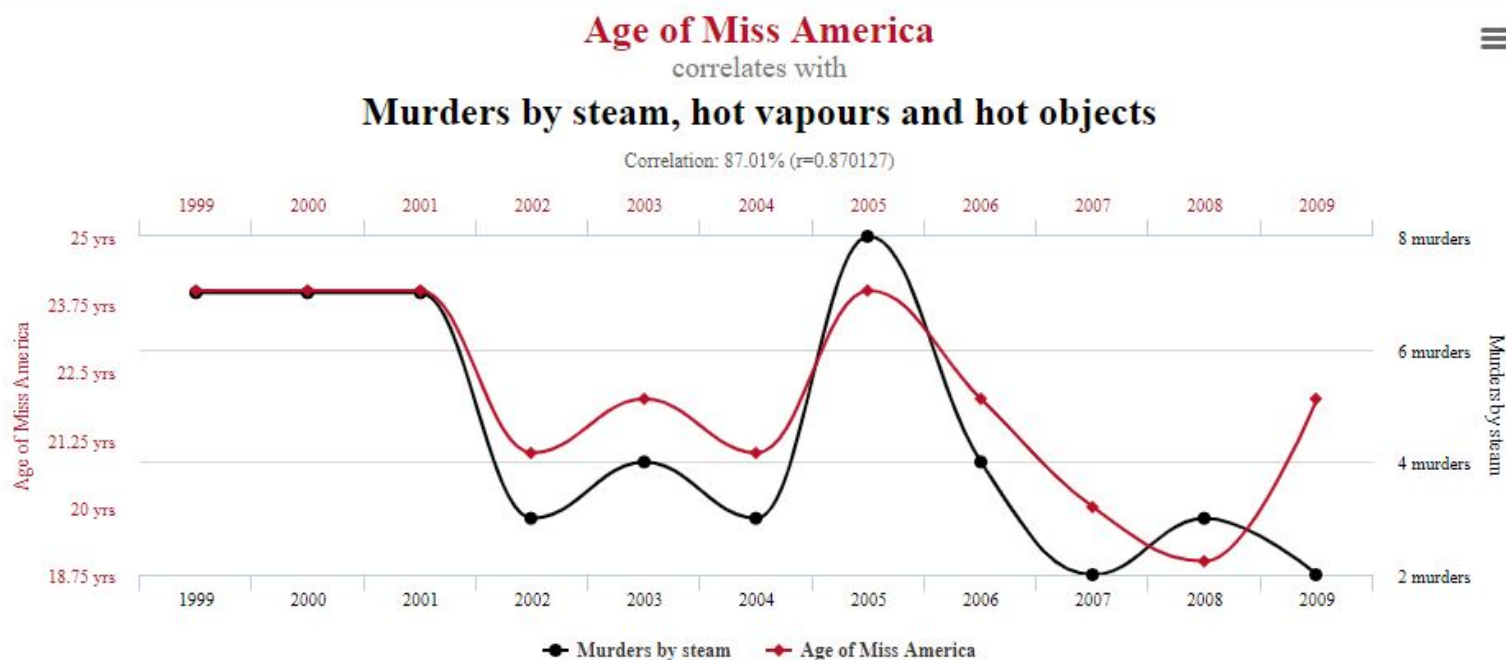
As importações de petróleo bruto dos EUA da Noruega se correlacionam com motoristas mortos em colisão com trem ferroviário



Fonte: <https://www.tylervigen.com/spurious-correlations>

Correlação x Causalidade ?

A idade da miss America correlaciona-se com assassinatos por vapor, vapores quentes e objetos quentes



Fonte: <https://www.tylervigen.com/spurious-correlations>

Matriz de correlação

- Base de dados mtcars

```
##           mpg           cyl           disp           hp           drat
## mpg      1.00000000 -0.8521620 -0.8475514 -0.7761684  0.6811719
## cyl     -0.8521620  1.00000000  0.9020329  0.8324475 -0.6999381
## disp    -0.8475514  0.9020329  1.00000000  0.7909486 -0.7102139
## hp      -0.7761684  0.8324475  0.7909486  1.00000000 -0.4487591
## drat     0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000
```

Análise de regressão



- A análise de regressão estuda a relação entre uma variável chamada de **variável dependente** Y e outras variáveis chamadas **variáveis independentes** X ;
- A relação entre Y e X é representada por um **modelo**, que associa a variável **variável dependente** com as variáveis **variáveis independentes**;
 - **Variável dependente**: Variável que desejamos prever ou explicar
 - **Variável independente**: Variável usada para explicar a variável dependente

Análise de regressão



- A análise de regressão é usada para:
 - Predizer valores de uma variável dependente baseado no valor de ao menos uma variável independente;
 - Explicar o impacto de mudanças em uma variável independente na variável dependente
- Regressão Linear Simples
- Regressão Linear Múltipla

Análise de regressão



- Como utilizar a regressão linear para prever valores de uma variável dependente Y com base em informações de uma variável independente X ?

- Na prática, procura-se uma função de X que explique Y , ou seja,

$$X; Y \rightarrow Y \simeq f(X)$$

- Essa relação, em geral, não é perfeita, ou seja, existem erros associados.

Análise de regressão



- Uma das preocupações estatísticas ao analisar dados é a de criar modelos do fenômeno em observação;
- As observações frequentemente estão misturadas com variações acidentais ou aleatórias;
- Assim, é conveniente supor que cada observação é formada por duas partes: uma previsível (ou controlada) e outra aleatória (ou não previsível), ou seja

$$(observação) = (previsível) + (aleatório)$$

Análise de regressão



- A parte previsível incorpora o conhecimento sobre o fenômeno, e é usualmente expressa por uma função matemática com parâmetros desconhecidos.
- As observações frequentemente estão misturadas com variações acidentais ou aleatórias.
- A parte aleatória deve obedecer algum modelo de probabilidade
- Com isso, o trabalho é produzir estimativas para os parâmetros desconhecidos, com base em amostras observadas.

Regressão Linear Simples

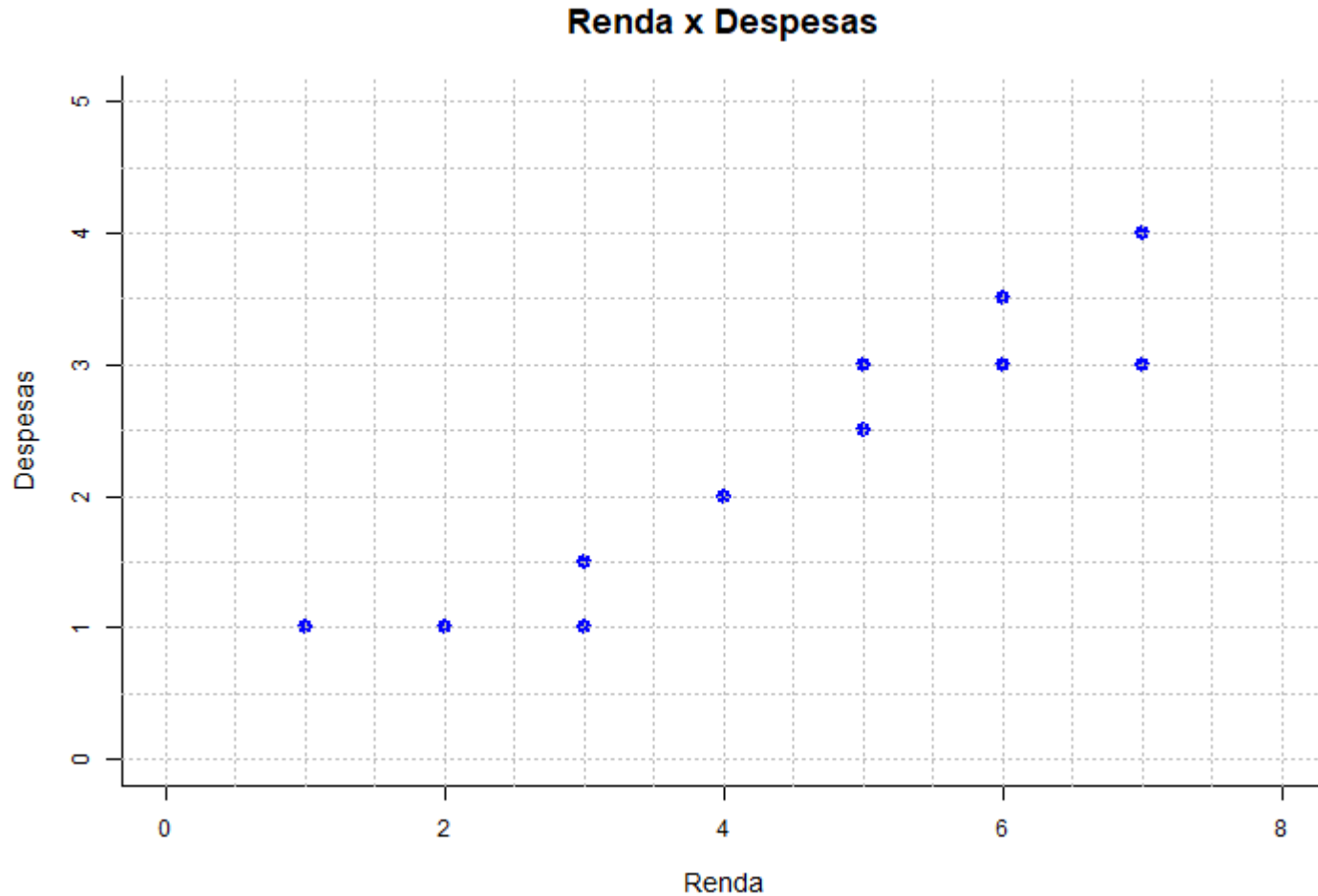
$$(observação) = (previsível) + (aleatório)$$

Matematicamente, podemos escrever

$$y_i = \theta + e_i$$

- y_i = observação i ;
- θ = efeito fixo, comum a todos os indivíduos
- e_i = efeito residual da observação i , pode ser considerado como o efeito resultante de várias características que não estão explícitas no modelo

Exemplo Renda x Despesas



Exemplo Renda x Despesas

Considerando que as despesas médias da população é de $\mu=2.2$, então a despesa de cada pessoa y_i pode ser descrita pelo seguinte modelo:

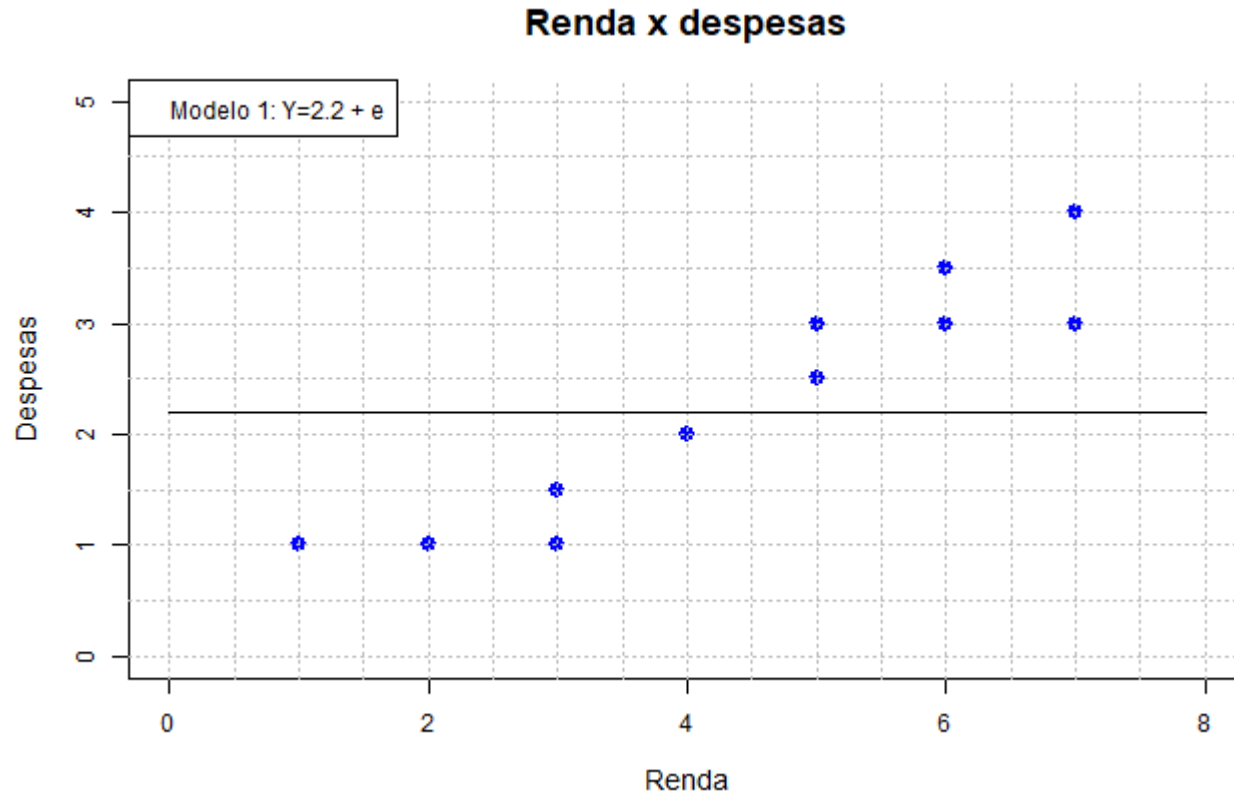
$$y_i = 2.2 + e_i$$

onde $\theta = \mu$, e cada e_i determinará as despesas de cada pessoa, em função de diversos fatores como: renda, idade, país, . . . , ou seja

$$e_i = f(Renda, idade, país \dots)$$

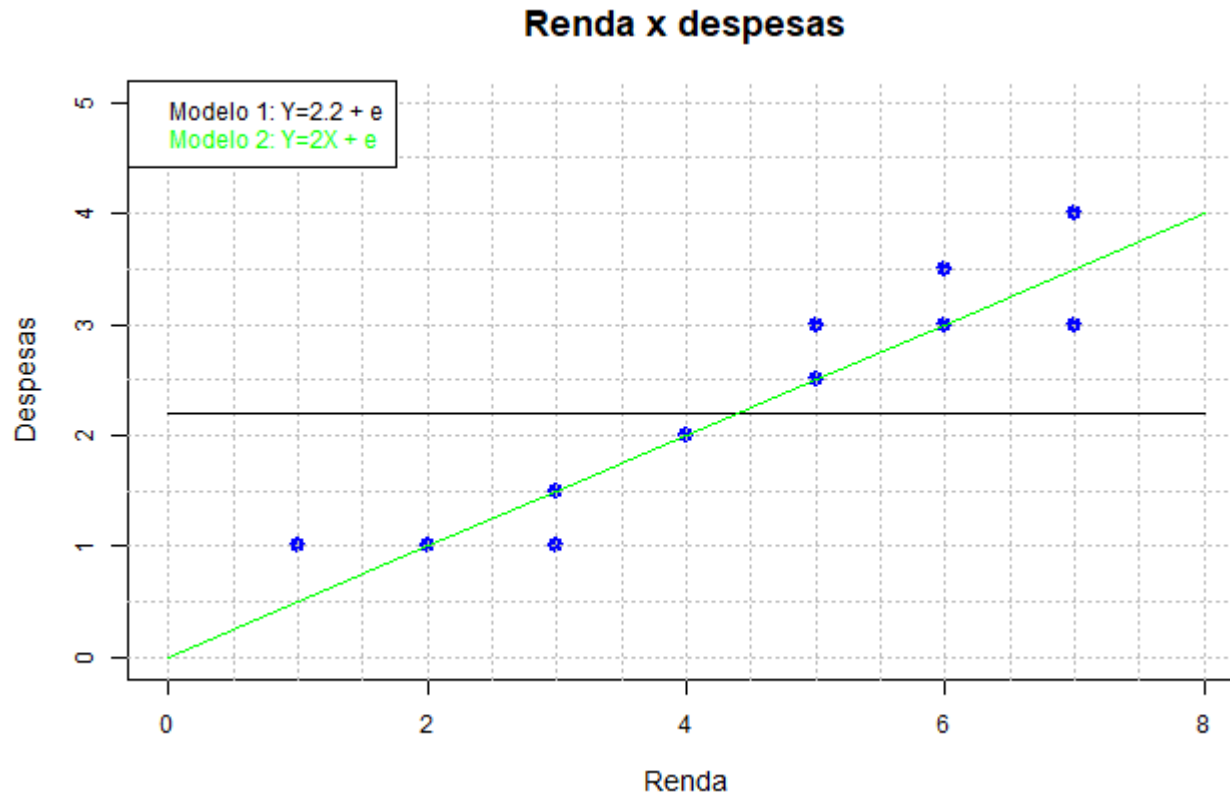
Ou seja, à medida que relacionamos as despesas com outras variáveis, ganhamos informação e diminuimos o erro.

Exemplo Renda x Despesas



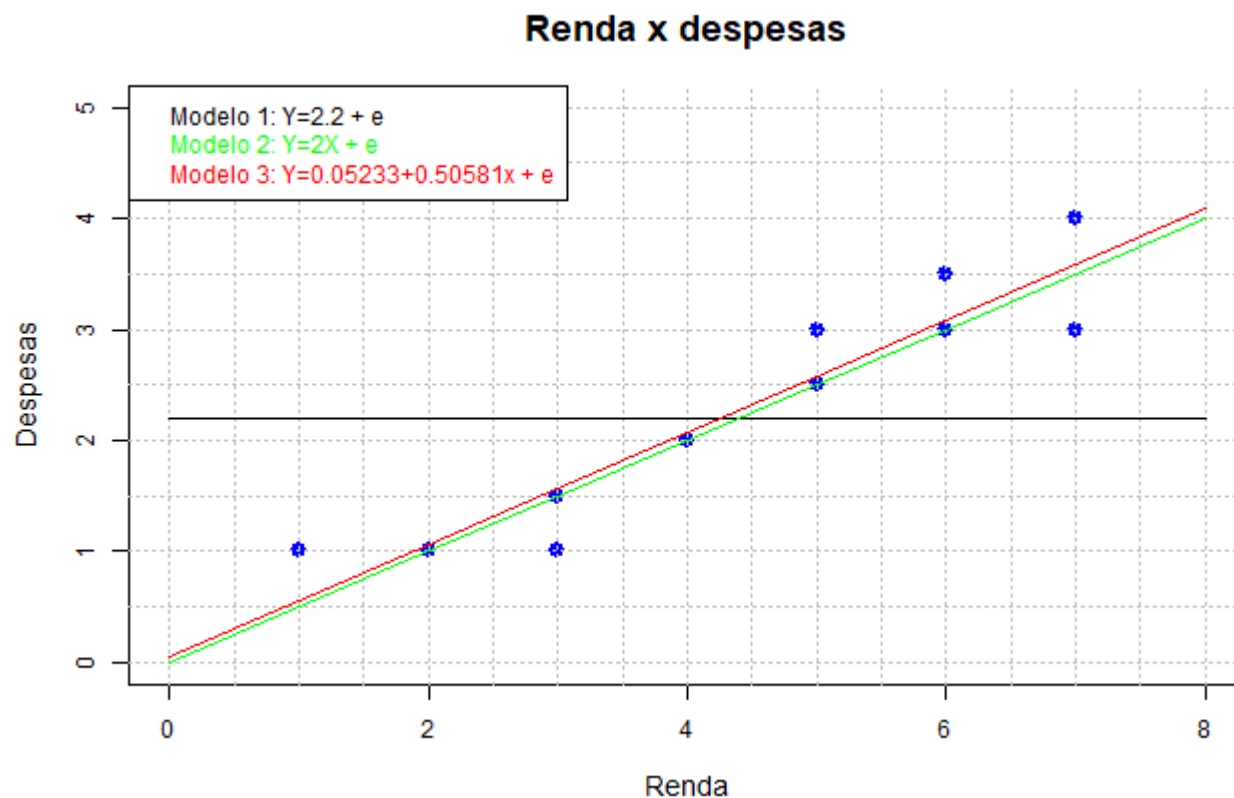
Exemplo Renda x Despesas

- Como as despesas dependem da renda de maneira linear, podemos então aprimorar o modelo anterior incorporando essa informação.

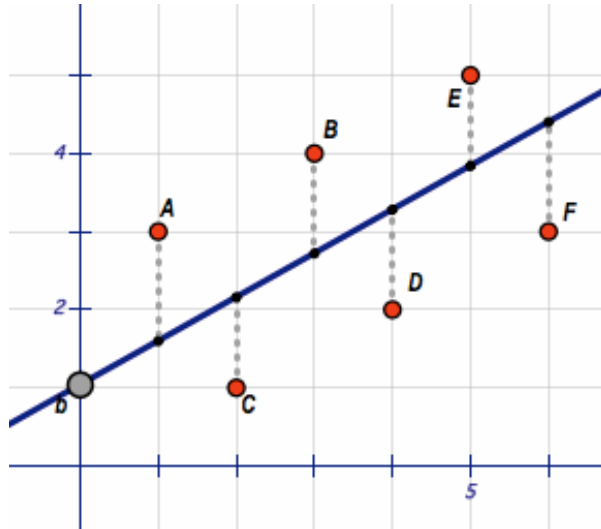


Exemplo Renda x Despesas

- Como as despesas dependem da renda de maneira linear, podemos então aprimorar o modelo anterior incorporando essa informação.



Regressão Linear



Regressão Linear Simples



Um modelo linear entre duas variáveis X e Y é definido matematicamente como uma equação com dois parâmetros desconhecidos,

$$Y = \beta_0 + \beta_1 X$$

Sendo assim, o modelo anterior onde conhecíamos só a média μ ,

$$y_i = \mu + e_i$$

pode ser reescrito como

$$y_i = \beta_0 + \beta_1 \text{Renda} + e_i$$

Note que o erro deve diminuir, pois agora

$$e_i = f(\text{idade}, \text{país}, \dots)$$

ou seja, incorporamos uma informação para explicar o peso, que antes estava inserida no erro.

Regressão Linear



- No exemplo anterior, notamos que Despesas é uma variável dependente (linearmente) da Renda.
- A análise de regressão é a técnica estatística que analisa as relações existentes entre uma única variável dependente, e uma ou mais variáveis independentes.
- O objetivo é estudar as relações entre as variáveis, a partir de um modelo matemático, permitindo estimar o valor de uma variável a partir da outra. Exemplo: sabendo a renda podemos determinar a despesa de uma família, se conhecemos os parâmetros do modelo anterior.

Regressão Linear



- Em uma análise de regressão linear consideraremos apenas as variáveis que possuem uma relação linear entre si.
- Uma análise de regressão linear múltipla pode associar k variáveis independentes (X) para “explicar” uma única variável dependente (Y),

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_k * X_k + e$$

- Uma análise de regressão linear simples associa uma única variável independente (X) com uma variável dependente (Y),

$$Y = \beta_0 + \beta_1 * X_1 + e$$

Regressão Linear



- Dados n pares de valores, $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, se for admitido que Y é função linear de X pode-se estabelecer uma regressão linear simples, cujo modelo estatístico é

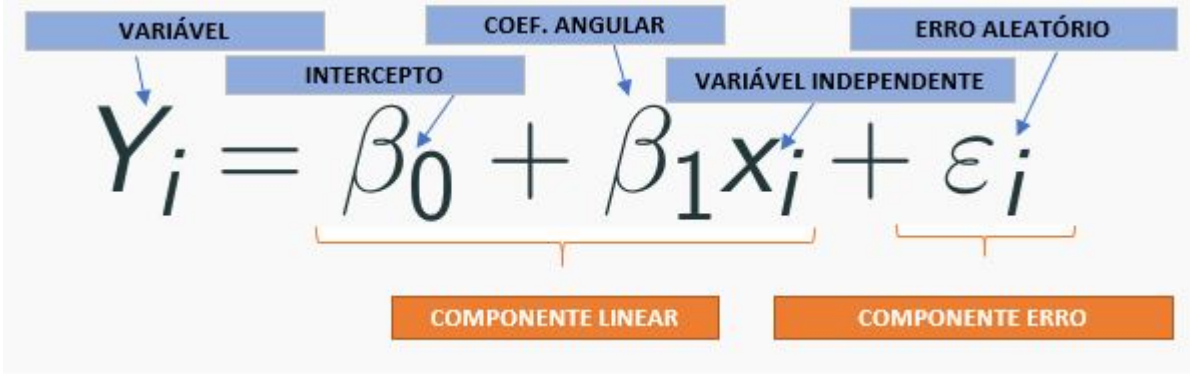
$$Y_i = \beta_0 + \beta_1 * X_i + e_i, i = 1, 2, \dots, n$$

onde,

- Y é variável resposta
- X é variável explicativa
- β_0 é o intercepto da reta (valor de Y quando $X=0$)
- β_1 é o coeficiente angular da reta (efeito de X sobre Y)
- $e \sim N(0, \sigma^2)$ é o resíduo

Devemos estimar os parâmetros de β_0 e β_1

Regressão Linear



The diagram illustrates the components of the linear regression equation $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Labels in blue boxes point to specific terms: 'VARIÁVEL' points to Y_i , 'INTERCEPTO' points to β_0 , 'COEF. ANGULAR' points to β_1 , 'VARIÁVEL INDEPENDENTE' points to x_i , and 'ERRO ALEATÓRIO' points to ε_i . Brackets below the equation group the terms into two categories: 'COMPONENTE LINEAR' (covering $\beta_0 + \beta_1 x_i$) and 'COMPONENTE ERRO' (covering ε_i).

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

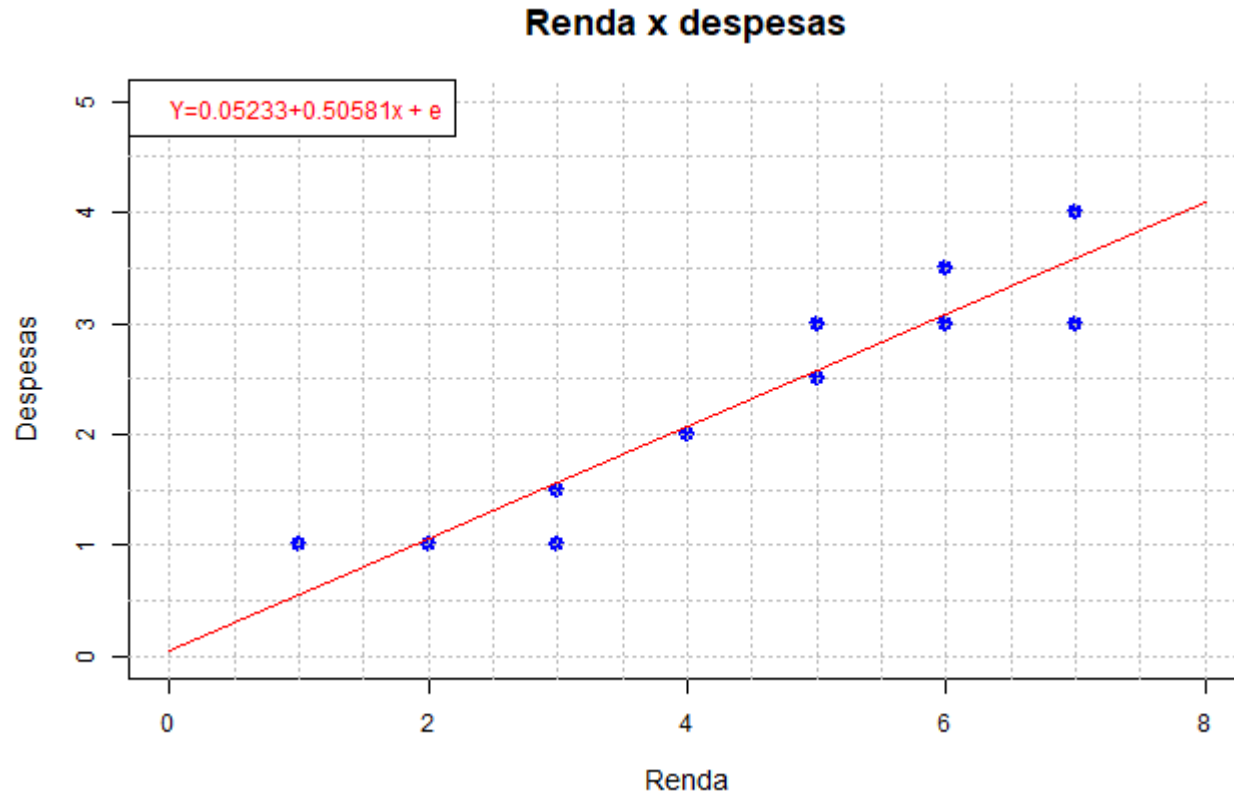
COMPONENTE LINEAR COMPONENTE ERRO

Regressão Linear



- β_0 representa o ponto onde a reta corta o eixo Y (na maioria das vezes não possui interpretação prática)
- β_1 representa a variabilidade em Y causada pelo aumento de uma unidade em X . Além disso,
- $\beta_1 > 0$ mostra que com o aumento de X , também há um aumento em Y ,
- $\beta_1 = 0$ mostra que não há efeito de X sobre Y
- $\beta_1 < 0$ mostra que com o aumento de X , há uma diminuição em Y

Regressão Linear



Estimação dos parâmetros



- Como através de uma amostra obtemos uma estimativa da verdadeira equação de regressão, denominamos

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * X_i + e_i, i = 1, 2, \dots, n$$

- ou seja, \hat{Y}_i é o valor estimado de Y_i , através das estimativas de β_0 e β_1 , que chamaremos de $\hat{\beta}_0$ e $\hat{\beta}_1$. Para cada valor de Y_i , temos um valor \hat{Y}_i estimado pela equação de regressão,

$$Y_i = \hat{Y}_i + e_i,$$

Estimação dos parâmetros



Portanto, o erro (ou desvio) de cada observação em relação ao modelo adotado será

$$e_i = Y_i - \hat{Y}_i$$

$$e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Devemos então adotar um modelo cujos parâmetros β_0 e β_1 tornem essa diferença a menor possível. Isso equivale a minimizar a soma de quadrados dos resíduos (SQR), ou do erro,

$$SQR = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

Estimação dos parâmetros

O método de minimizar a soma de quadrados dos resíduos é denominado de método dos mínimos quadrados. Para se encontrar o ponto mínimo de uma função, temos que obter as derivadas parciais em relação a cada parâmetro,

$$\frac{\partial SQR}{\partial \beta_0} = 2 \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)](-1)$$

$$\frac{\partial SQR}{\partial \beta_1} = 2 \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)](-X_i)$$

e igualar os resultados a zero

$$\hat{\beta}_0 = \frac{\partial SQR}{\partial \beta_0} = 0$$

$$\hat{\beta}_1 = \frac{\partial SQR}{\partial \beta_1} = 0$$

Estimação dos parâmetros

- Dessa forma, chegamos às estimativas de mínimos quadrados para os parâmetros β_0 e β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

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

onde

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

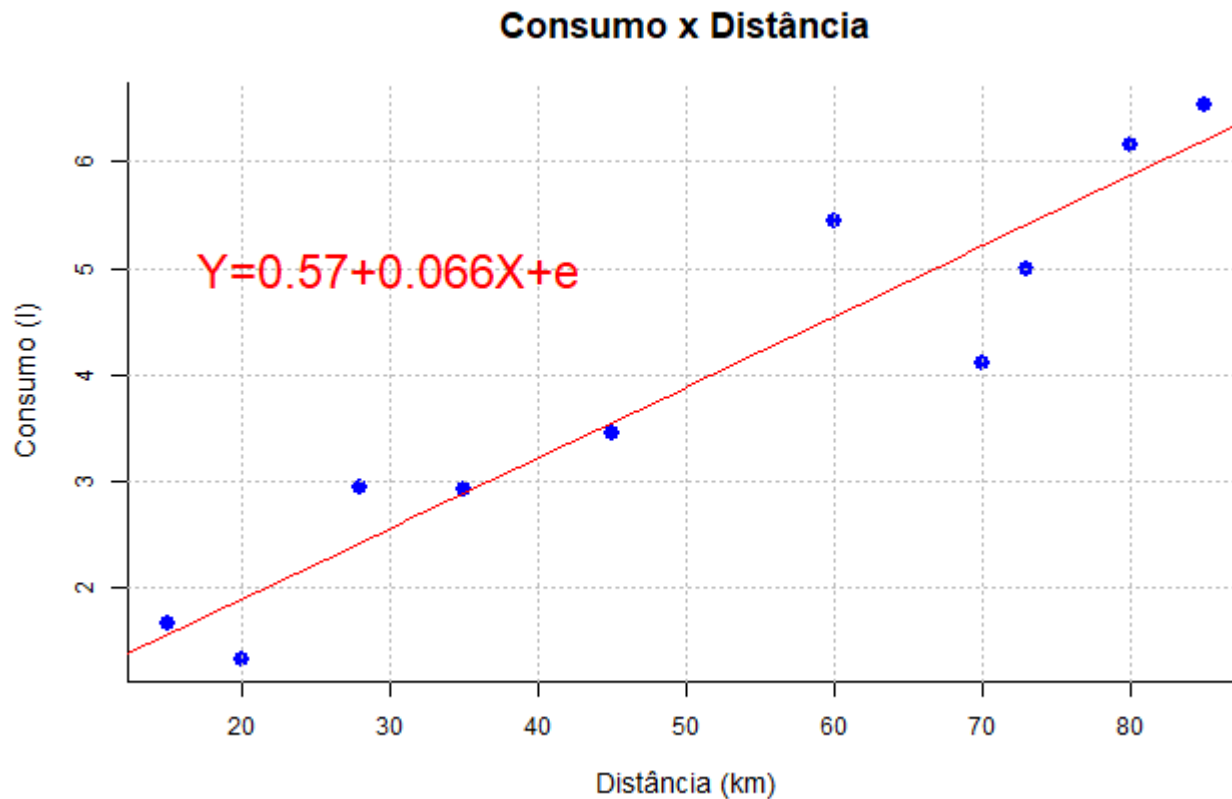
Exercício 1 continuação

- A tabela a seguir relaciona as distâncias percorridas por carros (km) e seus consumos de combustível (litros), em uma amostra de 10 carros novos.
 - Estime os parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$
 - Interprete o resultado
 - Trace o modelo linear aproximado,
 - Faça uma predição do consumo de combustível para uma distância de 37.00 km

<i>Distância</i>	20.00	60.00	15.00	45.00	35.00	80.00	70.00	73.00	28.00	85.00
<i>Consumo</i>	1.33	5.45	1.66	3.46	2.92	6.15	4.11	5.00	2.95	6.54

Exercício 1 continuação

- $\beta_0 = 0.57$ e $\beta_1 = 0.066$



Exercício 2 continuação

- A tabela a seguir relaciona os pesos de carros (t) e o rendimento de combustível (em km/l), para uma amostra de 10 carros.
 - Estime os parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$
 - Interprete o resultado
 - Trace o modelo linear aproximado,
 - Faça uma predição do rendimento de combustível para um carro de 1.75t

<i>Peso</i>	1.32	1.59	1.27	1.99	1.13	1.54	1.36	1.5	1.27	1.09
<i>Rendimento</i>	13.18	11.45	12.33	10.63	13.18	12.33	11.90	11.9	11.90	14.00

Exercício 2 continuação

- $\beta_0 = 16.68$ e $\beta_1 = -3.13$
- Para 1.75t o rendimento previsto é de 11.20

