Regressão Linear

Estatística II - 2024/2025
ISCTE-IUL

Afonso Moniz Moreira[1,2]

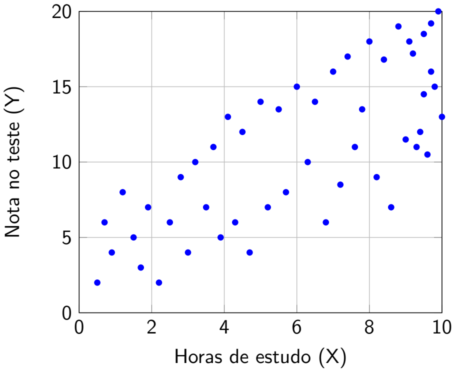
[1]ISCTE-IUL, Departamento de Métodos Quantitativos para a Economia e Gestão
[2]CMVM - Comissão do Mercado de Valores Mobiliários, Departamento de Supervisão de Mercados

Aviso/Disclaimer

- Este conjunto de slides não são, nem pretendem ser uma substituição à bibliografia principal da cadeira de Estatística II.
- Este conjunto de slides não são, nem pretendem ser uma fonte rigorosa de estudo dos tópicos da cadeira.
- O único propósito destes slides é ajudar o autor a guiar as aulas da forma mais coloquial possível sem ter de carregar formalismos desnecessários.
- Assim sendo, o formalismo estatístico é eliminado sempre que possível para agilizar uma primeira aprendizagem por parte dos estudantes.

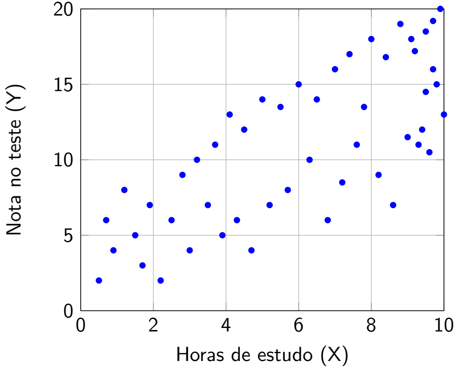
Para que serve a Regressão Linear ? I

- Vamos considerar as seguintes observações para 50 alunos do ISCTE, em que o par (X, Y) são o número de horas de estudo e a nota obtida no teste, respectivamente.



Para que serve a Regressão Linear ? I

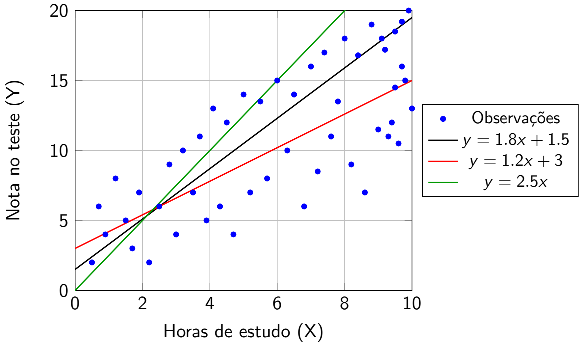
- Vamos considerar as seguintes observações para 50 alunos do ISCTE, em que o par (X, Y) são o número de horas de estudo e a nota obtida no teste, respectivamente.



- Que conclusão se pode retirar deste diagrama ? Talvez com $\rho_{X,Y}$

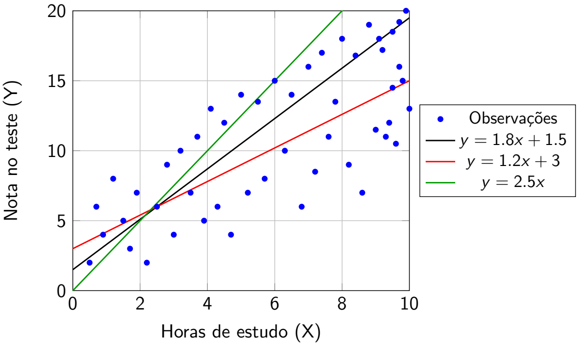
Para que serve a Regressão Linear ? II

- No intuito de dar uma ordem à nuvem de observações, podemos considerar várias relações lineares:



Para que serve a Regressão Linear ? II

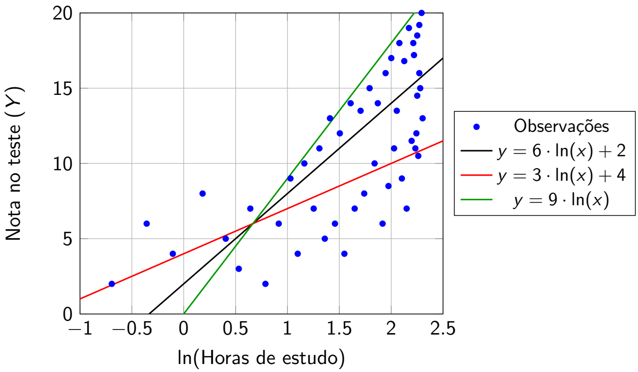
- No intuito de dar uma ordem à nuvem de observações, podemos considerar várias relações lineares:



- Certas relações vão atingir melhor ajustamento que outras...

Para que serve a Regressão Linear? III

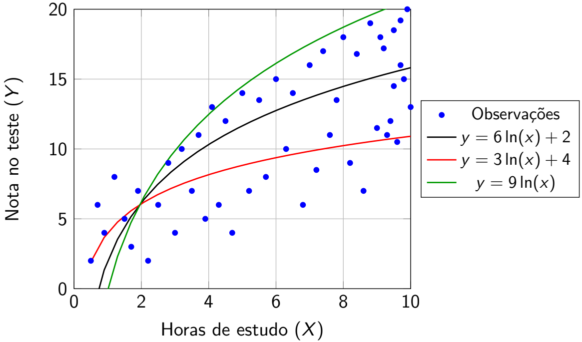
- Também podemos testar relações lineares entre Y e $\ln(X)$



- As relações lineares tornam-se visíveis após transformação logarítmica de X.

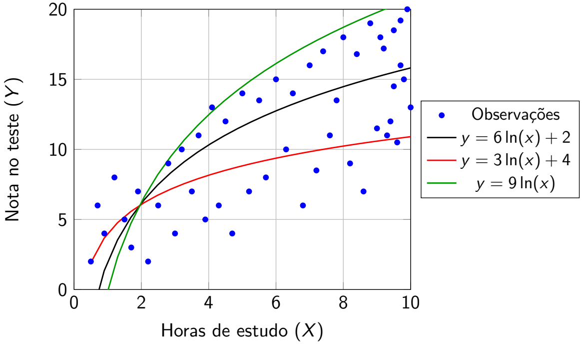
Para que serve a Regressão Linear ? III

- Ou não lineares entre Y e X, se não considerarmos a transformação logarítmica de X.



Para que serve a Regressão Linear ? III

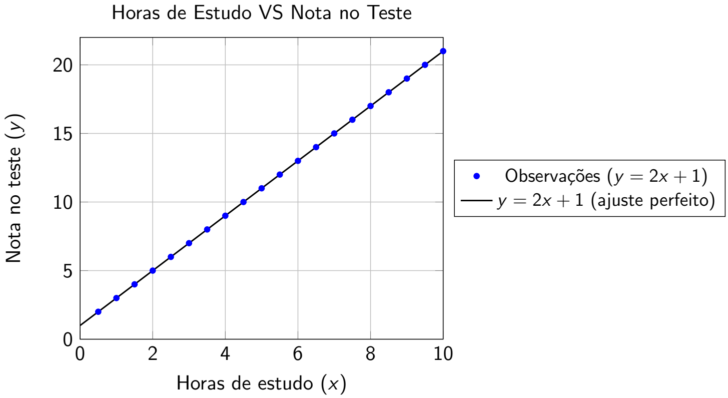
- Ou não lineares entre Y e X, se não considerarmos a transformação logarítmica de X.



- É cada vez mais difícil conseguir melhores resultados a partir de certo ponto... $\frac{dY^2}{dX^2} \leq 0, \forall X \in \mathbb{R}_0^+$

Outros Exemplos I

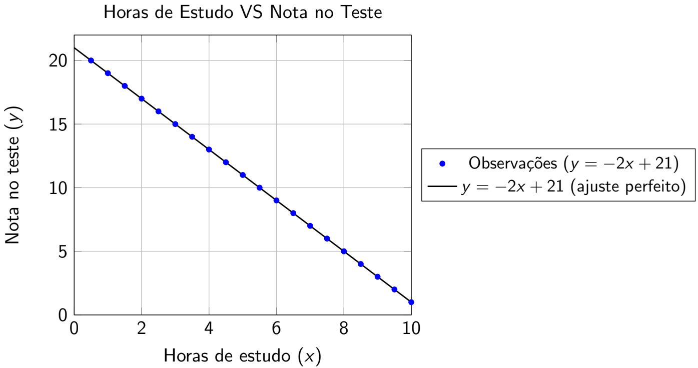
- Uma explicação perfeita de uma relação linear positiva



- Todas as observações alinham-se perfeitamente com a reta: correlação linear perfeita positiva $\rho_{X,Y} = 1$.

Outros Exemplos II

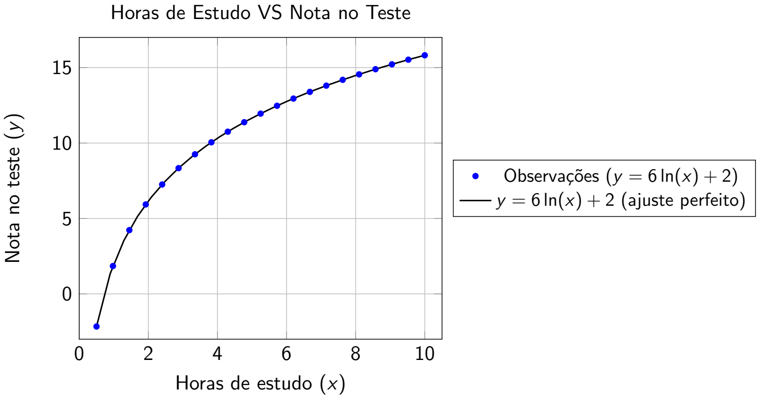
- Uma explicação perfeita de uma relação linear negativa



- Todas as observações alinham-se perfeitamente com uma reta decrescente: correlação linear perfeita negativa $\rho_{X,Y} = -1$.

Outros Exemplos III

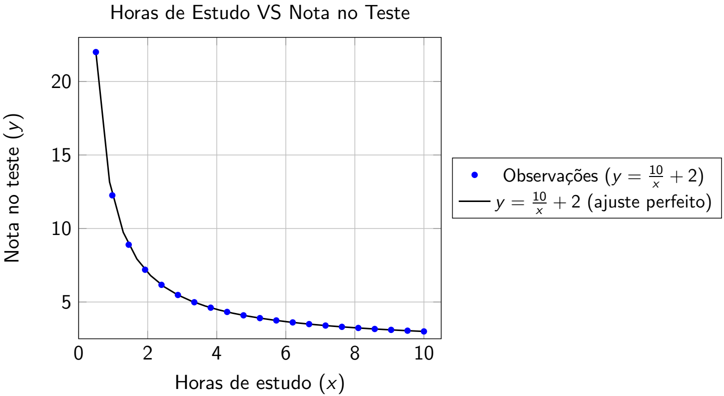
- Uma explicação perfeita de uma relação logarítmica positiva



- Todas as observações seguem exatamente a curva logarítmica: relação funcional perfeita.

Outros Exemplos IV

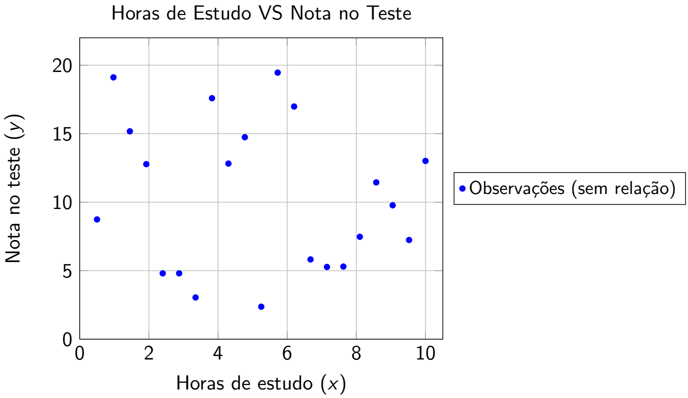
- Uma explicação perfeita de uma relação hiperbólica negativa



- Todas as observações alinham-se com uma curva hiperbólica — relação funcional perfeita.

Outros Exemplos V

- Um exemplo de ausência total de relação entre duas variáveis



- Não existe qualquer padrão ou estrutura

Correlação e Causalidade

- Qual é exatamente a diferença ? Considere-se estes exemplos:

Correlação e Causalidade

- Qual é exatamente a diferença ? Considere-se estes exemplos:
- Quando as temperaturas estão elevadas o consumo de gelados, em média, aumenta.

Correlação e Causalidade

- Qual é exatamente a diferença ? Considere-se estes exemplos:
- Quando as temperaturas estão elevadas o consumo de gelados, em média, aumenta.
- Quando as temperaturas estão elevadas, as queimaduras solares, em média, aumentam.

Correlação e Causalidade

- Qual é exatamente a diferença ? Considere-se estes exemplos:
- Quando as temperaturas estão elevadas o consumo de gelados, em média, aumenta.
- Quando as temperaturas estão elevadas, as queimaduras solares, em média, aumentam.
- Ao calcularmos um coeficiente de correlação linear, ρ de person, entre o consumo de gelados e as queimaduras solares vamos verificar, com um elevado grau de confiança, uma correlação positiva.

Correlação e Causalidade

- Qual é exatamente a diferença ? Considere-se estes exemplos:
- Quando as temperaturas estão elevadas o consumo de gelados, em média, aumenta.
- Quando as temperaturas estão elevadas, as queimaduras solares, em média, aumentam.
- Ao calcularmos um coeficiente de correlação linear, ρ de person, entre o consumo de gelados e as queimaduras solares vamos verificar, com um elevado grau de confiança, uma correlação positiva.
- Então podemos de facto assumir que comer gelados aumenta as queimaduras solares...

Correlação e Causalidade

- Qual é exatamente a diferença ? Considere-se estes exemplos:
- Quando as temperaturas estão elevadas o consumo de gelados, em média, aumenta.
- Quando as temperaturas estão elevadas, as queimaduras solares, em média, aumentam.
- Ao calcularmos um coeficiente de correlação linear, ρ de person, entre o consumo de gelados e as queimaduras solares vamos verificar, com um elevado grau de confiança, uma correlação positiva.
- Então podemos de facto assumir que comer gelados aumenta as queimaduras solares...
- **NÃO!** Existe uma terceira variável que causa estas duas... o nível exposição solar!

Correlação e Causalidade

- Qual é exatamente a diferença ? Considere-se estes exemplos:
- Quando as temperaturas estão elevadas o consumo de gelados, em média, aumenta.
- Quando as temperaturas estão elevadas, as queimaduras solares, em média, aumentam.
- Ao calcularmos um coeficiente de correlação linear, ρ de person, entre o consumo de gelados e as queimaduras solares vamos verificar, com um elevado grau de confiança, uma correlação positiva.
- Então podemos de facto assumir que comer gelados aumenta as queimaduras solares...
- **NÃO!** Existe uma terceira variável que causa estas duas... o nível exposição solar!
- **Correlação não implica causalidade!!!** A relação pode ser expúria.

Modelo de Regressão Linear Simples

- Vamos considerar que temos duas amostras aleatórias de duas variáveis Y e X de uma determinada população, em que cada uma delas tem dimensão n, portanto $(Y_1, ..., Y_n)$ e $(X_1, ..., X_n)$.

Modelo de Regressão Linear Simples

- Vamos considerar que temos duas amostras aleatórias de duas variáveis Y e X de uma determinada população, em que cada uma delas tem dimensão n, portanto $(Y_1, ..., Y_n)$ e $(X_1, ..., X_n)$.
- Há muitas luas atrás... numa cadeira chamada Estatística I... aprendemos um conceito denominado coeficiente de correlação linear, ou ρ de pearson:

Modelo de Regressão Linear Simples

- Vamos considerar que temos duas amostras aleatórias de duas variáveis Y e X de uma determinada população, em que cada uma delas tem dimensão n, portanto $(Y_1, ..., Y_n)$ e $(X_1, ..., X_n)$.
- Há muitas luas atrás... numa cadeira chamada Estatística I... aprendemos um conceito denominado coeficiente de correlação linear, ou ρ de pearson:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1] \tag{1}$$

Modelo de Regressão Linear Simples

- Vamos considerar que temos duas amostras aleatórias de duas variáveis Y e X de uma determinada população, em que cada uma delas tem dimensão n, portanto $(Y_1, ..., Y_n)$ e $(X_1, ..., X_n)$.
- Há muitas luas atrás... numa cadeira chamada Estatística I... aprendemos um conceito denominado coeficiente de correlação linear, ou ρ de pearson:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1] \tag{1}$$

- No entanto, isto traduz-se apenas num número e estamos interessados em perceber, se Y (i.e., a variável dependente) pode ser explicado por X (i.e., a variável independente), ou seja será que:

Modelo de Regressão Linear Simples

- Vamos considerar que temos duas amostras aleatórias de duas variáveis Y e X de uma determinada população, em que cada uma delas tem dimensão n, portanto $(Y_1, ..., Y_n)$ e $(X_1, ..., X_n)$.
- Há muitas luas atrás... numa cadeira chamada Estatística I... aprendemos um conceito denominado coeficiente de correlação linear, ou ρ de pearson:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1] \tag{1}$$

- No entanto, isto traduz-se apenas num número e estamos interessados em perceber, se Y (i.e., a variável dependente) pode ser explicado por X (i.e., a variável independente), ou seja será que:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{2}$$

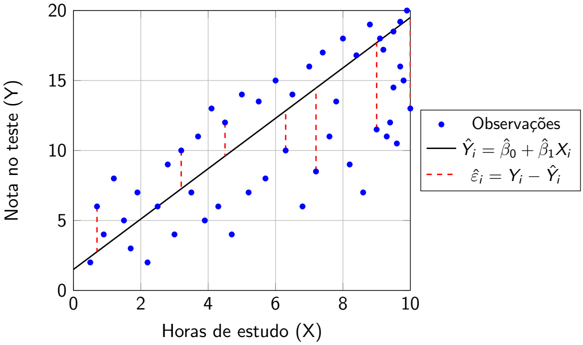
onde ε_i é o erro não observado da observação i e $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \mathbb{N}$.

Modelo de Regressão Linear Simples

- Por outras palavras será que existe uma relação causa-efeito ? Será que X causa Y de forma linear ?

Modelo de Regressão Linear Simples

- Por outras palavras será que existe uma relação causa-efeito ? Será que X causa Y de forma linear ?
- Vamos recuperar a nossa relação entre notas e tempo de estudo, agora com uma possível recta de regressão dada por: $\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0$



Estimador dos Mínimos Quadrados - OLS I

- Já verificamos anteriormente que para a mesma núvem de pontos podem existir múltiplas rectas... Então como vamos decidir ?

Estimador dos Mínimos Quadrados - OLS I

- Já verificamos anteriormente que para a mesma núvem de pontos podem existir múltiplas rectas... Então como vamos decidir ?
- Vamos escolher a recta que minimiza o somatório de todos os resíduos quadrados, ou seja:

Estimador dos Mínimos Quadrados - OLS I

- Já verificamos anteriormente que para a mesma núvem de pontos podem existir múltiplas rectas... Então como vamos decidir ?
- Vamos escolher a recta que minimiza o somatório de todos os resíduos quadrados, ou seja:

$$\min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3}$$

Estimador dos Mínimos Quadrados - OLS I

- Já verificamos anteriormente que para a mesma núvem de pontos podem existir múltiplas rectas... Então como vamos decidir ?
- Vamos escolher a recta que minimiza o somatório de todos os resíduos quadrados, ou seja:

$$\min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3}$$

- No fundo, queremos a melhor estimativa $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ que surge de minimizar $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, obtendo-se assim a melhor recta possível. A que tem melhor ajustamento aos pares (y_i, x_i) observados.

Estimador dos Mínimos Quadrados - OLS I

- Já verificamos anteriormente que para a mesma núvem de pontos podem existir múltiplas rectas... Então como vamos decidir ?
- Vamos escolher a recta que minimiza o somatório de todos os resíduos quadrados, ou seja:

$$\min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{(\hat{\beta}_0,\hat{\beta}_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3}$$

- No fundo, queremos a melhor estimativa $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ que surge de minimizar $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, obtendo-se assim a melhor recta possível. A que tem melhor ajustamento aos pares (y_i, x_i) observados.
- Como os erros ε_i não são diretamente observados o máximo que se pode fazer é minimizar a sua estimativa $\hat{\varepsilon}_i$, ou seja os resíduos.

Estimador dos Mínimos Quadrados - OLS II

- Então para se obter $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ trata-se de um problema de optimização simples:

Estimador dos Mínimos Quadrados - OLS II

- Então para se obter $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ trata-se de um problema de optimização simples:

$$\min_{(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{4}$$

Estimador dos Mínimos Quadrados - OLS II

- Então para se obter $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ trata-se de um problema de optimização simples:

$$\min_{(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{4}$$

- Iniciamos com $\hat{\beta}_0$:

$$\frac{\partial}{\partial \hat{\beta}_0} \left[\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right] = 0 \iff -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\iff \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \iff \hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\iff \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

Estimador dos Mínimos Quadrados - OLS III

- Agora para $\hat{\beta}_1$:

Estimador dos Mínimos Quadrados - OLS III

- Agora para $\hat{\beta}_1$:

$$\frac{\partial}{\partial \hat{\beta}_1}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right] = 0 \Longleftrightarrow -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Longleftrightarrow \quad \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Longleftrightarrow \quad \hat{\beta}_1 = \boxed{\frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}}$$

Estimador dos Mínimos Quadrados - OLS III

- Agora para $\hat{\beta}_1$:

$$\frac{\partial}{\partial \hat{\beta}_1}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right] = 0 \iff -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\iff \quad \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\iff \quad \hat{\beta}_1 = \boxed{\frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}}$$

- Finalmente juntamos $\hat{\beta}_0$ e $\hat{\beta}_1$ no mesmo sistema para que ambos os estimadores dependam apenas das amostras realizadas $(y_1, \ldots, y_n)$ e $(x_1, \ldots, x_n)$:

Estimador dos Mínimos Quadrados - OLS IV

Começamos com o sistema:

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \end{cases}$$

Substituímos $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ na equação de $\hat{\beta}_1$:

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

Estimador dos Mínimos Quadrados - OLS V

Desenvolvemos o numerador:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{Y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{X} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

Multiplicamos ambos os lados por $\sum_{i=1}^{n} x_i^2$:

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \bar{Y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{X} \sum_{i=1}^{n} x_i$$

Agrupamos os termos com $\hat{\beta}_1$:

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \bar{X} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i y_i - \bar{Y} \sum_{i=1}^{n} x_i$$

Estimador dos Mínimos Quadrados - OLS VI

Sabendo que $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \iff \sum_{i=1}^{n} x_i = n\bar{X}$ temos:

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{X}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\bar{X}\,\bar{Y}$$

Logo, os estimadores OLS são:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\,\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}} \quad \text{e} \quad \boxed{\hat{\beta}_0 = \bar{Y} - \left(\frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\,\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2} \right) \bar{X}}$$

Qualidade do Ajustamento I

- Como se mede se mede a qualidade do ajustamento ?

Qualidade do Ajustamento I

- Como se mede se mede a qualidade do ajustamento ?
- Através da seguinte relação:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{5}$$

Qualidade do Ajustamento I

- Como se mede se mede a qualidade do ajustamento ?
- Através da seguinte relação:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{5}$$

- Também podemos apresentar a relação anterior da seguinte maneira:

Qualidade do Ajustamento I

- Como se mede se mede a qualidade do ajustamento ?
- Através da seguinte relação:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{5}$$

- Também podemos apresentar a relação anterior da seguinte maneira:

$$\underbrace{SST}_{\text{Sum of Squares Total}} = \underbrace{SSR}_{\text{Sum of Squares Regression}} + \underbrace{SSR}_{\text{Sum of Squares Errors/Residuals}} \tag{6}$$

Qualidade do Ajustamento I

- Como se mede se mede a qualidade do ajustamento ?
- Através da seguinte relação:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{5}$$

- Também podemos apresentar a relação anterior da seguinte maneira:

$$\underbrace{SST}_{\text{Sum of Squares Total}} = \underbrace{SSR}_{\text{Sum of Squares Regression}} + \underbrace{SSR}_{\text{Sum of Squares Errors/Residuals}} \tag{6}$$

- No fundo a relação anterior indica que a variação total pode ser decomposta em variação relativa à regressão e variação relativa aos resíduos.

Qualidade do Ajustamento II

- Quanto maior o peso da componente SSR no SST, maior é a capacidade explicativa da regressão que estamos a efectuar e portanto maior é a qualidade do ajustamento da recta à nuvem de pontos.

Qualidade do Ajustamento II

- Quanto maior o peso da componente SSR no SST, maior é a capacidade explicativa da regressão que estamos a efectuar e portanto maior é a qualidade do ajustamento da recta à nuvem de pontos.
- Podemos condensar esta medida de ajustamento num número que se denomina de coeficiente de determinação que denomina por $R^2 \in [0, 1]$.

Qualidade do Ajustamento II

- Quanto maior o peso da componente SSR no SST, maior é a capacidade explicativa da regressão que estamos a efectuar e portanto maior é a qualidade do ajustamento da recta à nuvem de pontos.
- Podemos condensar esta medida de ajustamento num número que se denomina de coeficiente de determinação que denomina por $R^2 \in [0, 1]$.
- Seguindo a descrição efectuada anteriormente o coeficiente de determinação $R^2 \in [0, 1]$ é dado pelo seguinte rácio:

Qualidade do Ajustamento II

- Quanto maior o peso da componente SSR no SST, maior é a capacidade explicativa da regressão que estamos a efectuar e portanto maior é a qualidade do ajustamento da recta à nuvem de pontos.
- Podemos condensar esta medida de ajustamento num número que se denomina de coeficiente de determinação que denomina por $R^2 \in [0, 1]$.
- Seguindo a descrição efectuada anteriormente o coeficiente de determinação $R^2 \in [0, 1]$ é dado pelo seguinte rácio:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{7}$$

Qualidade do Ajustamento II

- Quanto maior o peso da componente SSR no SST, maior é a capacidade explicativa da regressão que estamos a efectuar e portanto maior é a qualidade do ajustamento da recta à nuvem de pontos.
- Podemos condensar esta medida de ajustamento num número que se denomina de coeficiente de determinação que denomina por $R^2 \in [0, 1]$.
- Seguindo a descrição efectuada anteriormente o coeficiente de determinação $R^2 \in [0, 1]$ é dado pelo seguinte rácio:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{7}$$

- Na regressão linear **simples**, quando existe apenas um regressor/variável independente (X), o coeficiente de determinação $R^2 \in [0, 1]$ é igual ao quadrado do coeficiente de correlação linear (i.e., $\rho_{X,Y}$ de Pearson).

Qualidade do Ajustamento III

- No seguimento da aferição da qualidade do modelo, são executados os seguintes ensaios de hipóteses bilaterais aos parâmetros estimados:

Qualidade do Ajustamento III

- No seguimento da aferição da qualidade do modelo, são executados os seguintes ensaios de hipóteses bilaterais aos parâmetros estimados:
- $H_0 : \beta_0 = 0$ VS $H_0 : \beta_0 \neq 0$ e $H_0 : \beta_1 = 0$ VS $H_0 : \beta_1 \neq 0$

Qualidade do Ajustamento III

- No seguimento da aferição da qualidade do modelo, são executados os seguintes ensaios de hipóteses bilaterais aos parâmetros estimados:
- $H_0 : \beta_0 = 0$ VS $H_0 : \beta_0 \neq 0$ e $H_0 : \beta_1 = 0$ VS $H_0 : \beta_1 \neq 0$
- As estatísticas dos testes e respectivas distribuições amostrais são dadas, respectivamente para β_0 e β_1, por:

Qualidade do Ajustamento III

- No seguimento da aferição da qualidade do modelo, são executados os seguintes ensaios de hipóteses bilaterais aos parâmetros estimados:
- $H_0 : \beta_0 = 0$ VS $H_0 : \beta_0 \neq 0$ e $H_0 : \beta_1 = 0$ VS $H_0 : \beta_1 \neq 0$
- As estatísticas dos testes e respectivas distribuições amostrais são dadas, respectivamente para β_0 e β_1, por:

$$\frac{\hat{\beta}_0 - 0}{\sqrt{VAR[\hat{\beta}_0]}} \sim t_{(n-2)} \text{ e } \frac{\hat{\beta}_1 - 0}{\sqrt{VAR[\hat{\beta}_1]}} \sim t_{(n-2)} \tag{8}$$

Qualidade do Ajustamento III

- No seguimento da aferição da qualidade do modelo, são executados os seguintes ensaios de hipóteses bilaterais aos parâmetros estimados:
- $H_0 : \beta_0 = 0$ VS $H_0 : \beta_0 \neq 0$ e $H_0 : \beta_1 = 0$ VS $H_0 : \beta_1 \neq 0$
- As estatísticas dos testes e respectivas distribuições amostrais são dadas, respectivamente para β_0 e β_1, por:

$$\frac{\hat{\beta}_0 - 0}{\sqrt{VAR[\hat{\beta}_0]}} \sim t_{(n-2)} \text{ e } \frac{\hat{\beta}_1 - 0}{\sqrt{VAR[\hat{\beta}_1]}} \sim t_{(n-2)} \quad (8)$$

- A rejeição de H_0 em ambos os testes anteriores significa que quer a constante β_0, quer a variável independente (X) em análise contribuem, de forma individual, para explicar a variável dependente (Y).

Estimador dos Mínimos Quadrados - Pressupostos I

- As estimativas $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ **não são** fidedignas sem os seguintes pressupostos verificados:

Estimador dos Mínimos Quadrados - Pressupostos I

- As estimativas $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ **não são** fidedignas sem os seguintes pressupostos verificados:
- **Linearidade entre a Variável Dependente e Variável Independente**

Estimador dos Mínimos Quadrados - Pressupostos I

- As estimativas $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ **não são** fidedignas sem os seguintes pressupostos verificados:
- **Linearidade entre a Variável Dependente e Variável Independente**
  - A variável dependente Y tem de ser explicada pela variável dependente X de forma linear. Verificado através da correlação linear (i.e., ρ de person) que deve ser estatisticamente significativa. Pode ser complementado pela forma do diagrama (Y_i, X_i).

Estimador dos Mínimos Quadrados - Pressupostos I

- As estimativas $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ **não são** fidedignas sem os seguintes pressupostos verificados:
- **Linearidade entre a Variável Dependente e Variável Independente**
    - A variável dependente Y tem de ser explicada pela variável dependente X de forma linear. Verificado através da correlação linear (i.e., ρ de person) que deve ser estatisticamente significativa. Pode ser complementado pela forma do diagrama (Y_i, X_i).
- **Normalidade dos erros** - $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \mathbb{N}$

Estimador dos Mínimos Quadrados - Pressupostos I

- As estimativas $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ **não são** fidedignas sem os seguintes pressupostos verificados:
- **Linearidade entre a Variável Dependente e Variável Independente**
  - A variável dependente Y tem de ser explicada pela variável dependente X de forma linear. Verificado através da correlação linear (i.e., ρ de person) que deve ser estatisticamente significativa. Pode ser complementado pela forma do diagrama (Y_i, X_i).
- **Normalidade dos erros** - $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \mathbb{N}$
  - Verificado através de um teste de ajustamento à normalidade dos resíduos que são o melhor estimador dos erros. Usa-se os testes de ajustamento à Normalidade: Kolmorogov-Smirnov e Shapiro-Wilk. Pode ser complementado com um diagrama Probability-Probability (P-P plot).

Estimador dos Mínimos Quadrados - Pressupostos I

- As estimativas $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ **não são** fidedignas sem os seguintes pressupostos verificados:
- **Linearidade entre a Variável Dependente e Variável Independente**
  - A variável dependente Y tem de ser explicada pela variável dependente X de forma linear. Verificado através da correlação linear (i.e., ρ de person) que deve ser estatisticamente significativa. Pode ser complementado pela forma do diagrama (Y_i, X_i).
- **Normalidade dos erros** - $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \mathbb{N}$
  - Verificado através de um teste de ajustamento à normalidade dos resíduos que são o melhor estimador dos erros. Usa-se os testes de ajustamento à Normalidade: Kolmorogov-Smirnov e Shapiro-Wilk. Pode ser complementado com um diagrama Probability-Probability (P-P plot).
- **Valor esperado dos erros é nulo** - $\mathbb{E}[\varepsilon_i] = 0, \forall i \in \mathbb{N}$

Estimador dos Mínimos Quadrados - Pressupostos I

- As estimativas $(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2$ **não são** fidedignas sem os seguintes pressupostos verificados:
- **Linearidade entre a Variável Dependente e Variável Independente**
  - A variável dependente Y tem de ser explicada pela variável dependente X de forma linear. Verificado através da correlação linear (i.e., ρ de person) que deve ser estatisticamente significativa. Pode ser complementado pela forma do diagrama (Y_i, X_i).
- **Normalidade dos erros** - $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \mathbb{N}$
  - Verificado através de um teste de ajustamento à normalidade dos resíduos que são o melhor estimador dos erros. Usa-se os testes de ajustamento à Normalidade: Kolmorogov-Smirnov e Shapiro-Wilk. Pode ser complementado com um diagrama Probability-Probability (P-P plot).
- **Valor esperado dos erros é nulo** - $\mathbb{E}[\varepsilon_i] = 0, \forall i \in \mathbb{N}$
  - Este pressuposto não é passível de ser verificado. Os erros ε_i não são observados, a sua melhor estimativa são os resíduos estimados $\hat{\varepsilon}_i$ e estes já são ajustados de forma a terem soma e média nula.

Estimador dos Mínimos Quadrados - Pressupostos II

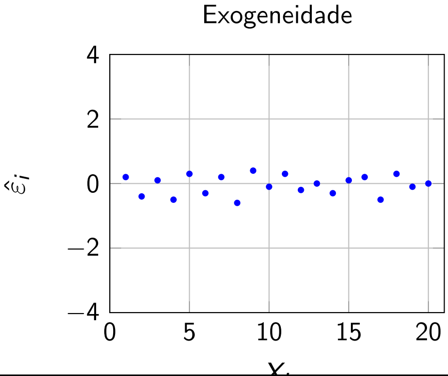
- **Exogeneidade dos Erros** - $\mathbb{E}[\varepsilon_i|X] = 0, \forall i \in \mathbb{N}$

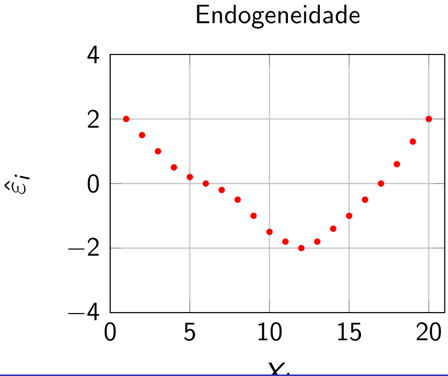
Estimador dos Mínimos Quadrados - Pressupostos II

- **Exogeneidade dos Erros** - $\mathbb{E}[\varepsilon_i|X] = 0, \forall i \in \mathbb{N}$
  - Este pressuposto não vai ser verificado nesta cadeira. No entanto a ideia é que não exista uma relação entre os resíduos (i.e., estimativas dos erros) e o regressor escolhido.

Estimador dos Mínimos Quadrados - Pressupostos II

- **Exogeneidade dos Erros** - $\mathbb{E}[\varepsilon_i|X] = 0, \forall i \in \mathbb{N}$
  - Este pressuposto não vai ser verificado nesta cadeira. No entanto a ideia é que não exista uma relação entre os resíduos (i.e., estimativas dos erros) e o regressor escolhido.





Estimador dos Mínimos Quadrados - Pressupostos III

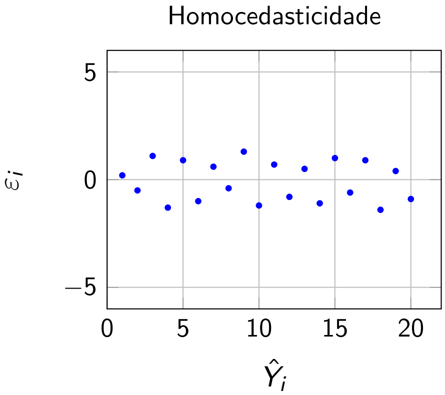
- **Homocedasticidade** - $VAR[\varepsilon_i] = \sigma^2, \forall i \in \mathbb{N}$

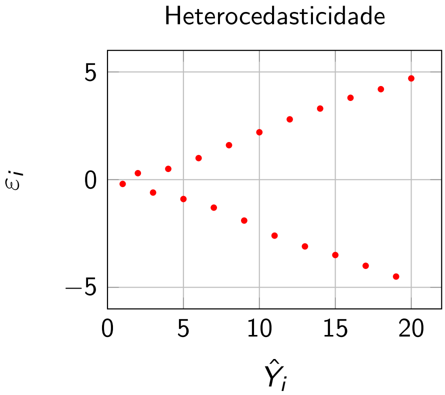
Estimador dos Mínimos Quadrados - Pressupostos III

- **Homocedasticidade** - $VAR[\varepsilon_i] = \sigma^2, \forall i \in \mathbb{N}$
  - Este pressuposto verifica-se, através do diagrama $(\hat{Y}_i, \varepsilon_i)$. Se a dispersão dos resíduos ε_i se mantiver constante, o pressuposto está verificado.

Estimador dos Mínimos Quadrados - Pressupostos III

- **Homocedasticidade** - $VAR[\varepsilon_i] = \sigma^2, \forall i \in \mathbb{N}$
  - Este pressuposto verifica-se, através do diagrama $(\hat{Y}_i, \varepsilon_i)$. Se a dispersão dos resíduos ε_i se mantiver constante, o pressuposto está verificado.





Estimador dos Mínimos Quadrados - Pressupostos IV

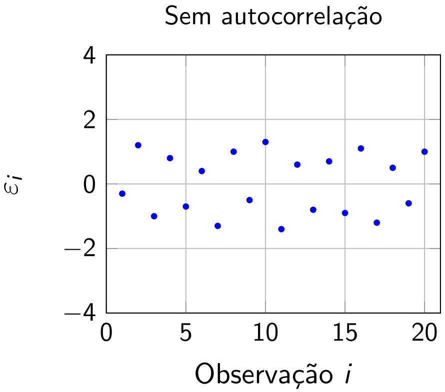
- **Inexistência de Autocorrelação nos Erros** - $COV(\varepsilon_i, \varepsilon_j), \forall i, j \in \mathbb{N}$

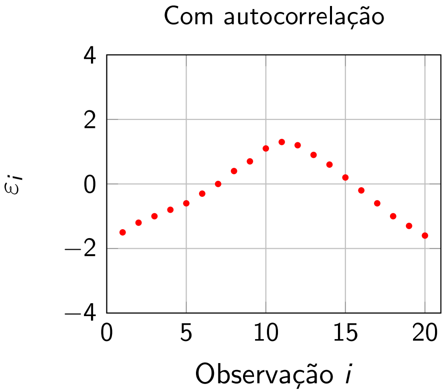
Estimador dos Mínimos Quadrados - Pressupostos IV

- **Inexistência de Autocorrelação nos Erros** - $COV(\varepsilon_i, \varepsilon_j), \forall i, j \in \mathbb{N}$
  - Este pressuposto verifica-se, através do diagrama $(\hat{Y}_i, \varepsilon_i)$. Se não existir um padrão/estrutura dos resíduos ε_i, o pressuposto está verificado.

Estimador dos Mínimos Quadrados - Pressupostos IV

- **Inexistência de Autocorrelação nos Erros** - $COV(\varepsilon_i, \varepsilon_j), \forall i, j \in \mathbb{N}$
  - Este pressuposto verifica-se, através do diagrama $(\hat{Y}_i, \varepsilon_i)$. Se não existir um padrão/estrutura dos resíduos ε_i, o pressuposto está verificado.





Modelo de Regressão Linear Múltipla

- Generalizando a regressão linear simples... Vamos considerar $K + 1$ amostras aleatórias de dimensão n de K+1 variáveis Y, X_1, X_2,..., X_K de uma determinada população, portanto $(Y_1, ..., Y_n)$, $(X_{11}, ..., X_{1n})$, $(X_{21}, ..., X_{2n})$, ..., $(X_{K1}, ..., X_{Kn})$.

Modelo de Regressão Linear Múltipla

- Generalizando a regressão linear simples... Vamos considerar $K+1$ amostras aleatórias de dimensão n de K+1 variáveis Y, X_1, X_2,..., X_K de uma determinada população, portanto $(Y_1, ..., Y_n)$, $(X_{11}, ..., X_{1n})$, $(X_{21}, ..., X_{2n})$, ..., $(X_{K1}, ..., X_{Kn})$.
- Agora existe interesse em perceber, se Y (i.e., a variável dependente) pode ser explicada pelas K variáveis independentes, ou seja será que:

Modelo de Regressão Linear Múltipla

- Generalizando a regressão linear simples... Vamos considerar $K+1$ amostras aleatórias de dimensão n de K+1 variáveis Y, X_1, X_2,..., X_K de uma determinada população, portanto $(Y_1, ..., Y_n)$, $(X_{11}, ..., X_{1n})$, $(X_{21}, ..., X_{2n})$, ..., $(X_{K1}, ..., X_{Kn})$.
- Agora existe interesse em perceber, se Y (i.e., a variável dependente) pode ser explicada pelas K variáveis independentes, ou seja será que:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + ... + \beta_1 X_{Ki} + \varepsilon_i, \tag{9}$$

onde ε_i é o erro não observado da observação i e $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \mathbb{N}$.

Modelo de Regressão Linear Múltipla

- Generalizando a regressão linear simples... Vamos considerar $K+1$ amostras aleatórias de dimensão n de K+1 variáveis Y, X_1, X_2,..., X_K de uma determinada população, portanto $(Y_1, ..., Y_n)$, $(X_{11}, ..., X_{1n})$, $(X_{21}, ..., X_{2n})$, ..., $(X_{K1}, ..., X_{Kn})$.
- Agora existe interesse em perceber, se Y (i.e., a variável dependente) pode ser explicada pelas K variáveis independentes, ou seja será que:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + ... + \beta_1 X_{Ki} + \varepsilon_i, \tag{9}$$

  onde ε_i é o erro não observado da observação i e $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \mathbb{N}$.
- Com $K = 1$ temos a regressão linear simples, pelo que a regressão linear múltipla exige que $K \geq 2$, ou seja pelo menos 2 regressores.