
Tera

Estatística e Modelagem de Dados

Aula 16:

Regressão Linear

Instrutora

Cristiane Rodrigues

- **Bacharel em Matemática – UNESP Rio Claro.**
- **Mestre em Estatística – USP Piracicaba**
- Experiências Profissionais:
  - Modelagem de Credito para PF e PJ – Banco Bradesco
  - Experiência com Segmentação e Análise de Series temporais – Atento
  - Consultora Analítica - SAS Institute Brasil
  - Consultora de Pré Vendas - SAS Institute Brasil
  - Coach/Professora do curso SAS Academy Data Science

crisrodrigues_27@hotmail.com

Índice

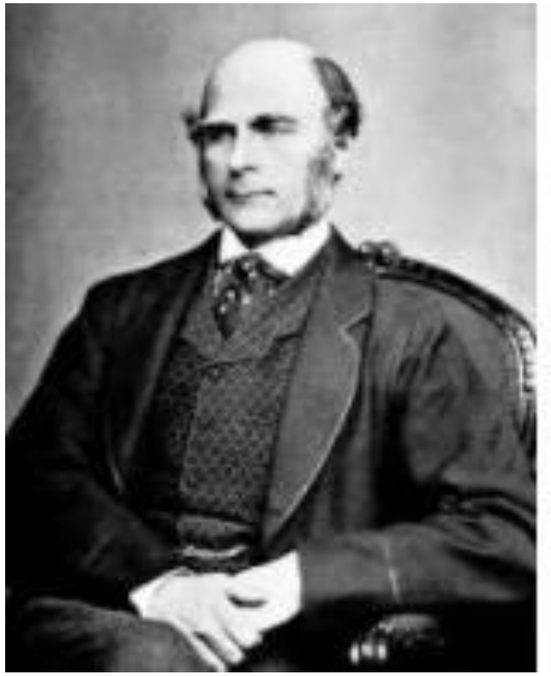
- Motivação
- Coeficiente de Correlação

- Regressão Linear Simples
  - Modelo
  - Equação Estimada
  - Interpretação dos Parâmetros
  - Qualidade do Ajuste
  - Teste de Significância: Teste F e Teste t

- Regressão Linear Multipla
- Suposições do Modelo
- Multicolinearidade

Motivação – Galton e o estudo sobre altura

Galton (1822-1911) foi um antropólogo, meterologista, matemático e estatístico inglês. Era primo de Charles Darwin e, baseado em sua obra, criou o conceito de "eugenia“ que seria a melhora de uma determinada espécie através da seleção artificial.



Questionamentos:

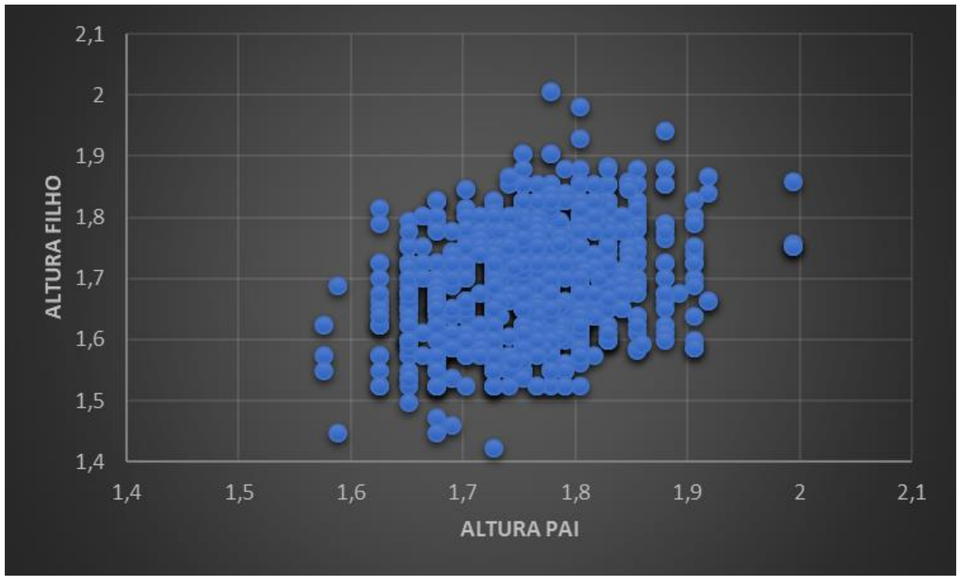
- Pais altos tem sempre filhos altos?
- Existe uma tendência de crescimento da altura média das pessoas?

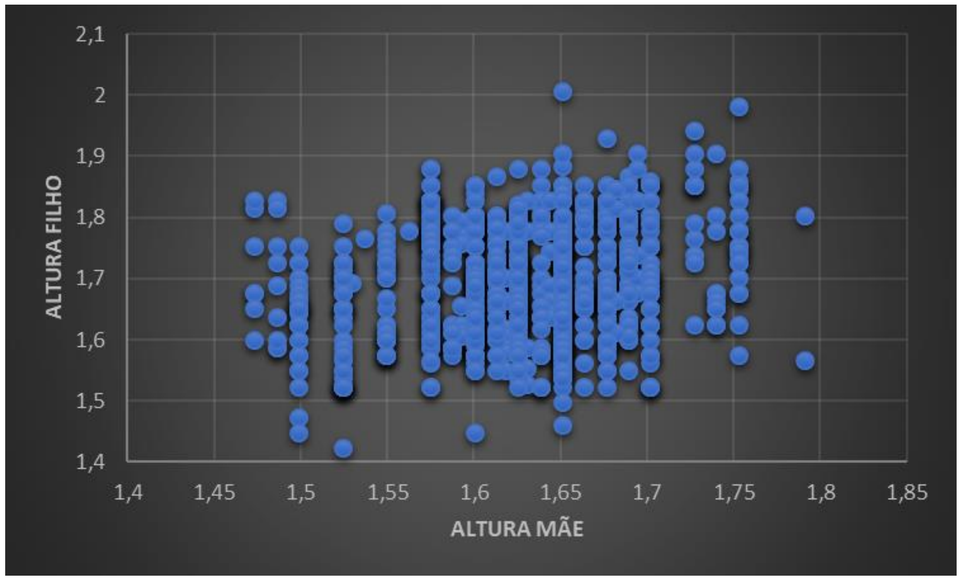
Galton usou a regressão e correlação para estudar a variação genética em humanos.

Motivação – Galton e o estudo sobre altura

ID Familia	Altura Pai	Altura Mãe	Sexo Filho	Altura Filho	Quantidade Filhos
1	1.9939	1.7018	M	1.85928	4
1	1.9939	1.7018	F	1.75768	4
1	1.9939	1.7018	F	1.75260	4
1	1.9939	1.7018	F	1.75260	4
2	1.9177	1.6891	M	1.86690	4
2	1.9177	1.6891	M	1.84150	4
2	1.9177	1.6891	F	1.66370	4
2	1.9177	1.6891	F	1.66370	4
3	1.9050	1.6256	M	1.80340	2
3	1.9050	1.6256	F	1.72720	2

Motivação - Galton e o estudo sobre altura





Existe uma corelação linear entre essas variáveis?

O que é o coeficiente de corelação linear?

Coeficiente de correlação

O coeficiente de correlação é uma medida descritiva da força da associação linear entre duas variáveis.

$$-1 \leq r \leq +1$$

$r = 0$	Não existe correlação entre as variáveis
$r = 1$	Existe correlação linear positiva perfeita entre as variáveis
$r = -1$	Existe correlação linear negativa perfeita entre as variáveis
$r \geq 0{,}70$	Existe uma “forte” correlação linear entre as variáveis
$0 \leq r \leq 0{,}70$	Existe uma “fraca” correlação linear entre as variáveis

Coeficiente de correlação

- Valores que se aproximam de 1 ou +1 indicam uma forte relação linear.
- Quanto mais próxima a correlação estiver de zero, mais fraca será esta relação.

Coeficiente de correlação Amostral

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Em que

$$S_{xy} = \ Covariância\ Amostral = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$S_x = Desvio\ padrão\ da\ amostra\ de\ x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$S_y == Desvio\ padrão\ da\ amostra\ de\ \mathrm{y} = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

Galton - Coeficiente de correlação

Vamos calcular (EXCEL) a Correlação entre a variável independente - Altura do Pai com a variável Dependente - Altura do Filho

Family	Altura_Pai	y=Altura_Filho	x_i - x_barra	(x_i - x_barra)^2	y_i - y_barra	(x_i - x_barra)*(y_i - y_barra)	(y_i - y_barra)^2						
1	1,9939	1,85928	0,23521827	0,055327634	0,163667326	0,038497545	0,026786994	b_1	0,400214				
1	1,9939	1,75768	0,23521827	0,055327634	0,062067326	0,014599369	0,003852353	b_0	1,018008				
1	1,9939	1,7526	0,23521827	0,055327634	0,056987326	0,01340446	0,003247555						
1	1,9939	1,7526	0,23521827	0,055327634	0,056987326	0,01340446	0,003247555			sum	N-1		
2	1,9177	1,8669	0,15901827	0,02528681	0,171287326	0,027237814	0,029339348	**correlacao_pai**	cov	1,412184	889	0,001589	
2	1,9177	1,8415	0,15901827	0,02528681	0,145887326	0,02319875	0,021283112		dpx	3,528577	889	0,003969	0,063001
2	1,9177	1,6637	0,15901827	0,02528681	-0,031912674	-0,005074698	0,001018419		dpy	7,376468	889	0,008297	0,091091
2	1,9177	1,6637	0,15901827	0,02528681	-0,031912674	-0,005074698	0,001018419						
3	1,905	1,8034	0,14631827	0,021409036	0,107787326	0,015771255	0,011618108		**corr**	**0,276801**			
3	1,905	1,7272	0,14631827	0,021409036	0,031587326	0,004621803	0,000997759						
4	1,905	1,7907	0,14631827	0,021409036	0,095087326	0,013913013	0,0090416						
4	1,905	1,7399	0,14631827	0,021409036	0,044287326	0,006480045	0,001961367						
4	1,905	1,7018	0,14631827	0,021409036	0,006187326	0,000905319	3,8283E-05						
4	1,905	1,6383	0,14631827	0,021409036	-0,057312674	-0,008385891	0,003284743						
4	1,905	1,6002	0,14631827	0,021409036	-0,095412674	-0,013960617	0,009103578						
5	1,905	1,8288	0,14631827	0,021409036	0,133187326	0,019487739	0,017738864						

Estudo de Caso
Verificando a Correlação

Parte 1 : Calcular a correlação das variáveis: Altura Pai, Altura da Mãe e Altura Média dos pais com a Altura dos Filhos

Qual a variável que está mais correlacionada com a altura dos Filhos?

Regressão Linear Simples (RLS) – Modelo

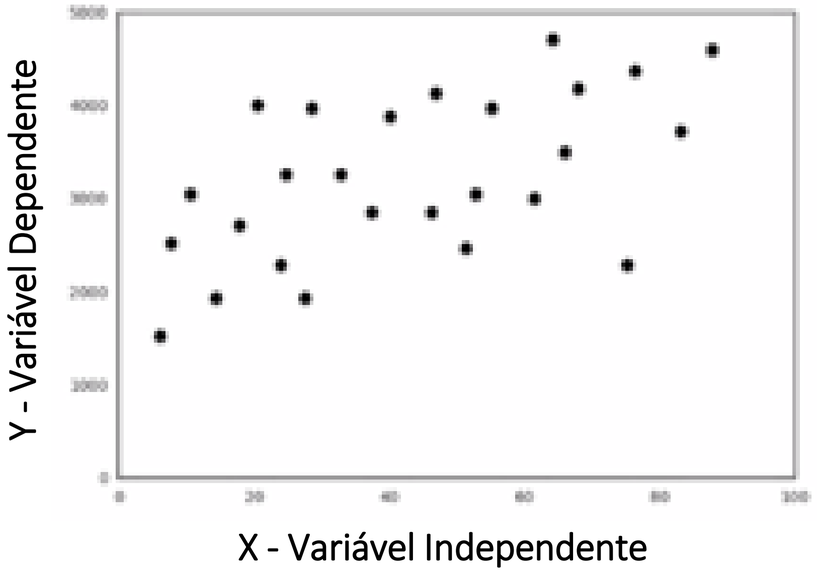
A equação que descreve como a variável dependente y está relacionada com a variável independente x e um erro, denomina-se modelo de regressão linear simples (MRLS).

$$y = \beta_0 + \beta_1 * x + \epsilon$$

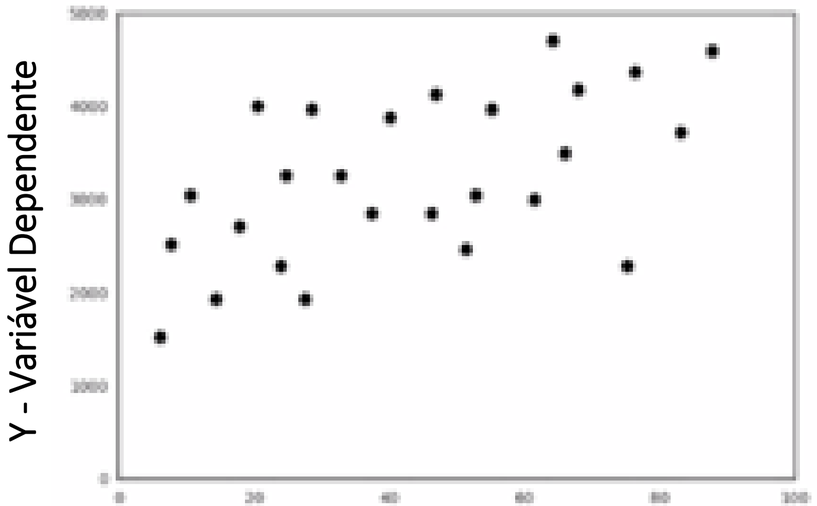
No modelo de regressão linear simples β_0 e β_1 são os parâmetros e ϵ é uma variável aleatória. A variável dependente y é uma função linear de x mais um termo de erro ϵ.

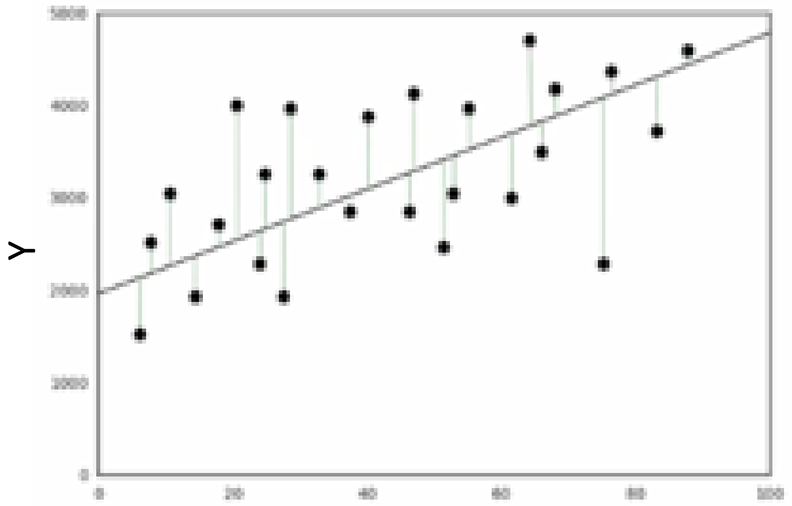
Como escolher β_0 e β_1?

Regressão Linear Simples - Modelo

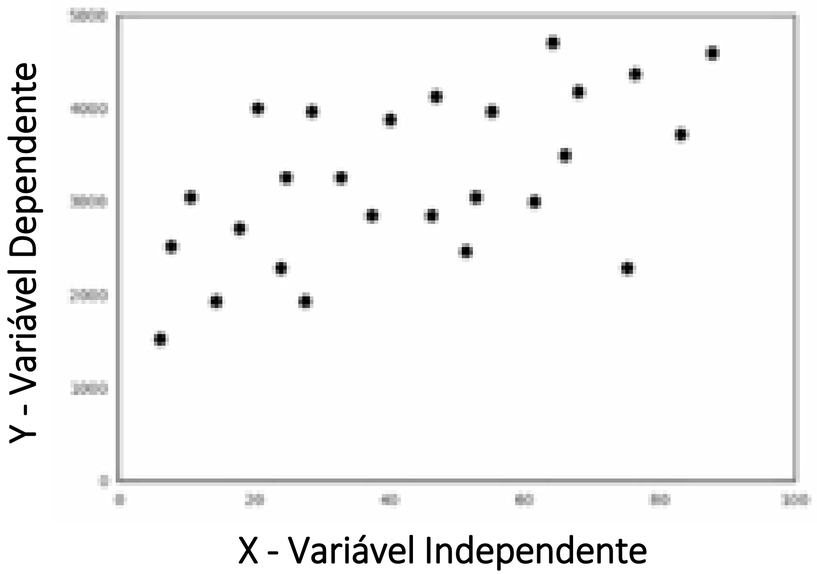


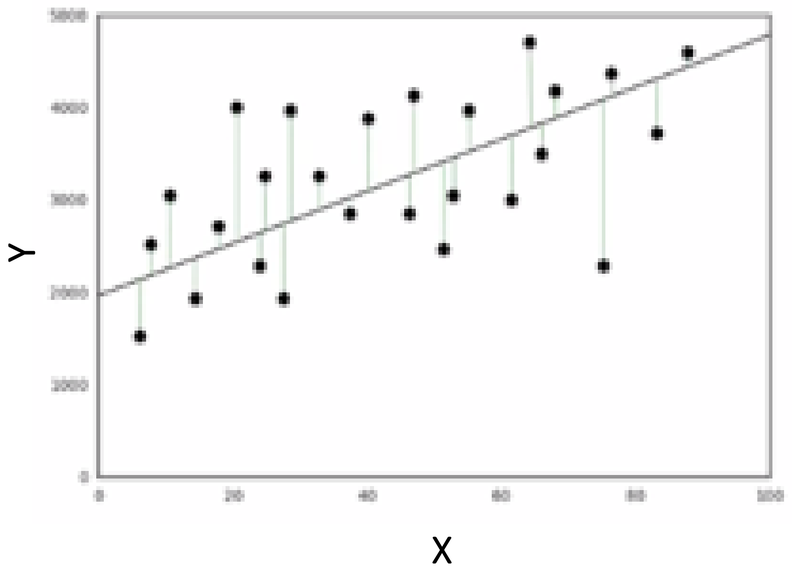
Regressão Linear Simples - Modelo

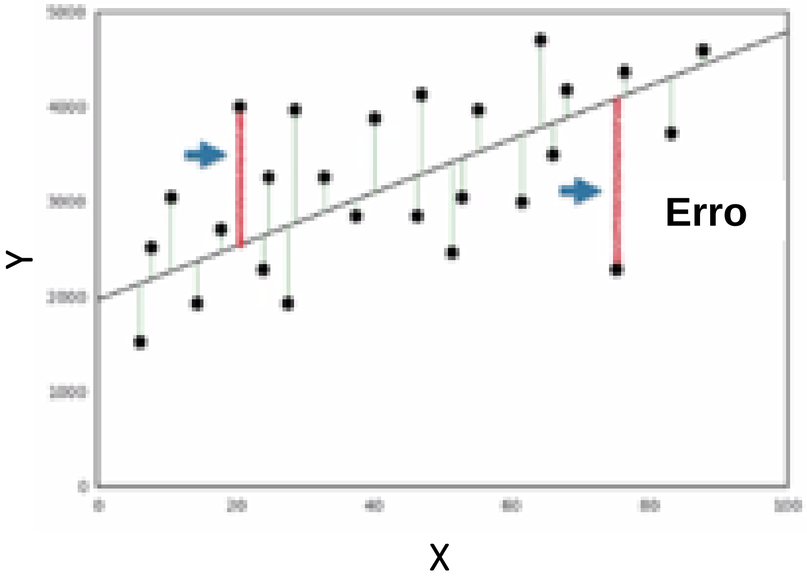




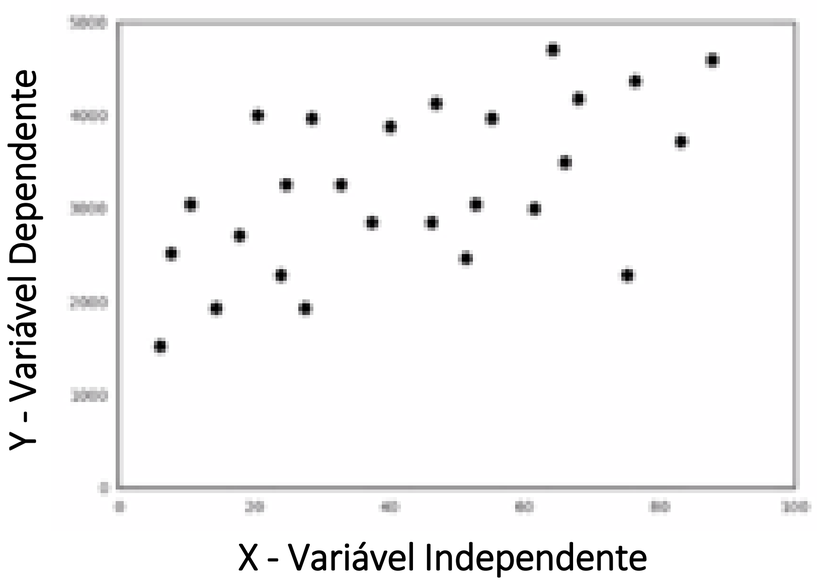
Regressão Linear Simples - Modelo

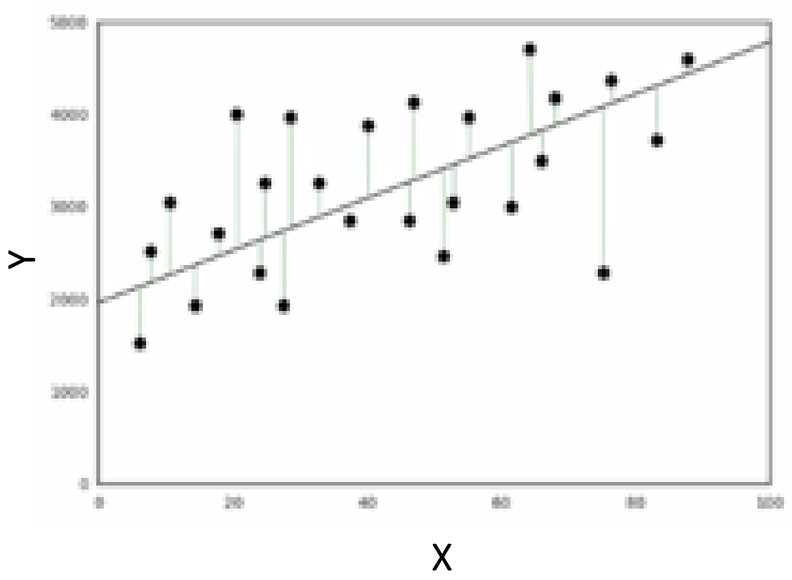


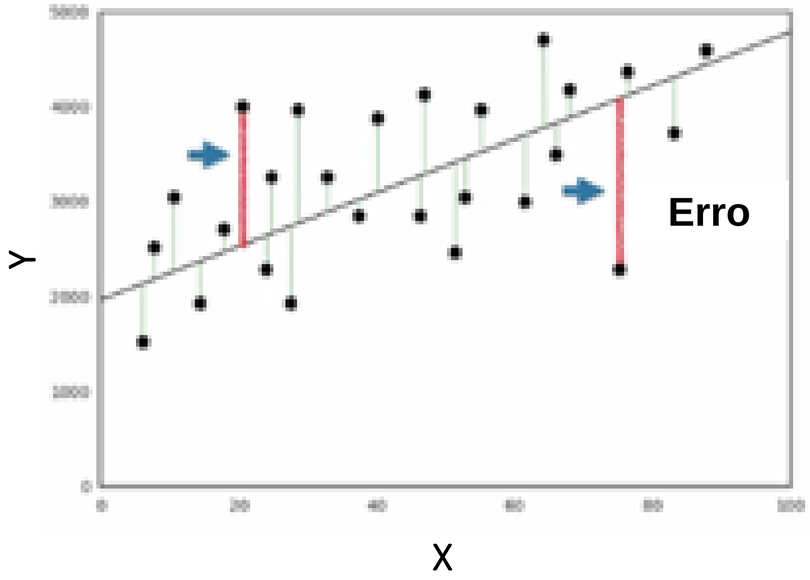




Regressão Linear Simples - Modelo







- Na regressão linear o objetivo é escolher a reta que minimiza a função de erro, ou seja, que diminui a distância entre o ajuste e os dados.

Regressão Linear Simples – Equação Estimada

A estimação dos parâmetros deve ser feita a partir de dados amostrais

$$\hat{y} = b_0 + b_1 * x_1 + \epsilon$$

em que

b_0 e b_1 são estimadores de β_0 e β_1

$\hat{y}$ é o valor estimado da variável dependente

Regressão Linear Simples – Estimação dos parâmetros

O método dos **Mínimos Quadrados** é utilizado para estimar os parâmetros do modelo de RLS, por meio da seguinte expressão

$$\min \sum (y_i - \hat{y}_i)^2$$

em que

y_i é o valor **observado** da variável dependente para a i-ésima observação

$\hat{y}_i$ é o valor **estimado** da variável dependente para a i-ésima observação

Regressão Linear Simples – Estimação dos parâmetros

Critério dos Mínimos Quadrados

$$\min \sum (y_i - \hat{y}_i)^2$$

Então

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

Galton - Estimativa dos parâmetros

Voltando ao exemplo das Alturas no estudo de Galton
Queremos determinar a relação entre a Altura do Pai e a Altura dos Filhos, **ou seja, queremos definir a seguinte relação**

$$Altura_{Filho} = b_0 + b_1 * Altura_{Pai} + \epsilon$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{1{,}412184403}{3{,}528577278} = 0{,}40021354$$

$$b_0 = \bar{y} - b_1\bar{x} = 1{,}695612 - 0{,}40021354 * 1{,}758681 = 0{,}991764433$$

Com as estimativas podemos escrever a equação que descreve a relação entre as alturas do Pai e do Filho

$$\boldsymbol{Altura_{Filho} = 0{,}9918 + 0{,}4002 * Altura_{Pai} + \epsilon}$$

Estudo de Caso
Ajustando um Modelo de Regressão Linear Simples

Parte 2 : Ajustar um modelo de regressão linear simples para determinar a relação entre a Altura do Pai e a Altura do Filho.

- sklearn

Regressão Linear Simples – Interpretação dos parâmetros

$$\hat{y} = b_0 + b_1 x_1$$

- O parâmetro b_1, na regressão linear simples, representa uma estimativa da alteração em y correspondente à alteração de uma unidade em x_1.

- O parâmetro b_0, na regressão linear simples, representa uma estimativa de y quando x_1 é 0.

Galton - Interpretação dos parâmetros

$$\hat{y} = 1 + 0{,}4 * x_1$$

- No exemplo de Galton, 0,4 é a estimativa de aumento esperado na altura do filho correspondente ao aumento de 1 unidade na altura do pai.

- Já o 1 é o valor esperado do y quando a altura do pai é “zero”.

Regressão Linear Simples – Qualidade do Ajuste

Em modelos de regressão linear simples usamos o coeficiente de determinação, R^2 para medir a qualidade do modelo

$$R^2 = \frac{SQexp}{SQtot} = 1 - \frac{SQres}{SQtot}$$

Em que
$SQexp$ = Soma dos quadrados explicada = $\sum(\hat{y}_i - \bar{y})^2$
$SQres$ = Soma dos quadrados dos resíduos = $\sum(y_i - \hat{y}_i)^2$
$SQtot$ = Soma dos quadrados totais = $\sum(y_i - \bar{y})^2$

O R^2 representa a proporção da variabilidade da variável dependente que pode ser explicada pela equação estimada.

Quanto mais proximo de 1 estiver o R^2 melhor é a qualidade do ajuste do modelo

Galton - Avaliando a qualidade do Ajuste

No exemplo de Galton o R^2 = 0.076618. O que isso quer dízer ????

Como podemos melhorar o modelo?

Estudo de Caso
Ajustando um Modelo de Regressão Linear Simples

Parte 3 : Ajustar um modelo de regressão linear simples para determinar a relação entre a Altura do Pai e a Altura do Filho.

- statsmodel

Teste de Significância

- *O teste F* é utilizado para **determinar se existe uma relação significativa entre as variáveis independentes**. O teste F é conhecido como teste de significância global.

- *O teste t* é usado para **determinar se cada uma das variáveis independentes individuais é significativa**. Um teste t separado é realizado para cada uma das variáveis independentes do modelo. Tal teste é conhecido como teste de significância individual.

Reg Linear Simples – Teste F (significância global)

$$\begin{cases} H_o: \beta_1 = 0 \\ H_1: \text{o parâmetro não é igual a zero} \end{cases}$$

Se H_o for rejeitada, o teste nos dá suficientes evidências estatísticas para concluirmos que o parâmetro não é igual a zero, sendo assim a relação global entre y e as variável independente (x_1) é significativa.

Se H_o não for rejeitada, não teremos evidências suficientes para concluir que existe uma relação significativa.

Reg Linear Simples – Teste F (significância global)

$$\begin{cases} H_o: \beta_1 = 0 \\ H_1: \text{o parâmetro não é igual a zero} \end{cases}$$

Regra de Rejeição

Critério do valor p: Rejeitar H_o se $\mathrm{p}_{\mathrm{valor}} \leq \alpha$

Galton - Teste F

Considerando o modelo ajustado

$$Altura_{Filho} = 0{,}9918 + 0{,}4002 * Altura_{Pai}$$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.077
Model:	OLS	Adj. R-squared:	0.076
Method:	Least Squares	F-statistic:	73.68
Date:	Sat, 18 Aug 2018	Prob (F-statistic):	4.06e-17
Time:	13:58:44	Log-Likelihood:	905.47
No. Observations:	890	AIC:	-1807.
Df Residuals:	888	BIC:	-1797.
Df Model:	1		
Covariance Type:	nonrobust		

Testa a hipótese de que existe relação linear. Quando este valor for $< 0{,}05$ (α desejado) concluímos que existe relação linear da variável independente em relação a variável resposta/dependente.

Reg Linear Simples – Teste T (significância individual)

Se o teste F demosntrar que a relação de regressão é significativa, um teste *t* pode ser realizado para determinar a significância do parâmetro individual β_1

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

- Se H_o for rejeitada, o teste nos dá suficientes evidências estatísticas para concluirmos que o parâmetro (β_1) em estudo não é igual a zero, ou seja, o parâmetro é estatisticamente significativo.
- Se H_o não for rejeitada, não teremos evidências suficientes para concluir que o parâmentro (β_1) em estudo é estatisticamente significante.

Reg Linear Simples – Teste T (significância individual)

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

Regra de Rejeição

Critério do valor p: Rejeitar H_o se $\mathrm{p_{valor}} \leq \alpha$

Galton - Teste T

Considerando o modelo ajustado

$$Altura_{Filho} = 0{,}9918 + 0{,}4002 * Altura_{Pai}$$

	coef	std err	t	P>|t|
const	0.9918	0.082	12.087	0.000
x1	0.4002	0.047	8.584	0.000

Testa se cada um dos parâmetros é significativo para o modelo.
Neste caso todos os parâmetros são significativos

Regressão Linear Multipla - Modelo

A equação que descreve como a variável dependente y está relacionada com as variáveis independentes $x_1, x_2, \ldots, x_p$ e um erro, denomina-se modelo de regressão linear múltipla (MRLM).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + + \beta_p x_p + \epsilon$$

No modelo de regressão linear múltipla $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, são os parâmetros e ϵ é uma variável aleatória. A variável dependente y é uma função linear de $x_1, x_2, \ldots, x_p$ mais um termo de erro ϵ.

Como escolher $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$?

Regressão Linear Multipla – Estimação dos parâmetros

O método dos **Mínimos Quadrados** é utilizado para estimar os parâmetros do modelo de RLM, por meio da seguinte expressão

$$\min \sum (y_i - \hat{y}_i)^2$$

em que

y_i é o valor **observado** da variável dependente para a i-ésima observação

$\hat{y}_i$ é o valor **estimado** da variável dependente para a i-ésima observação

Regressão Linear Multipla – Estimação dos parâmetros

O método dos **Mínimos Quadrados** na forma matricial

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{1i} & \cdots & x_{1pn} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{ni} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

$$\underline{b} = (X'X)^{-1}X'Y$$

Regressão Linear Multipla – Interpretação dos parâmetros

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

- O coeficiente b_i, na regressão linear multipla, representa uma estimativa da alteração em y correspondente à alteração de uma unidade em x_i, quando todas as outras variáveis independentes de mantêm constante.

Regressão Linear Multipla – Qualidade do Ajuste

Em ***MRLM*** para evitar superestimação do impacto de se adicionar mais uma variável independente no modelo, usamos o:

$$R^2{}_{ajustado} = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

em que:
n é o número de observações da amostra
e p é o número de variáveis independentes (parâmetros estimados pelo modelo).

O R^2 representa a proporção da variabilidade da variável dependente que pode ser explicada pela equação estimada.

Regressão Linear Multipla – Teste F (significância global)

$$\begin{cases} H_o: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_1: \text{um ou mais dos parâmetros não são iguais a zero} \end{cases}$$

- Se H_o for rejeitada, o teste nos dá suficientes evidências estatísticas para concluirmos que um ou mais dos parâmetros não são iguais a zero, sendo assim a relação global entre y e as variáveis independentes $(x_1, x_2, \ldots, x_p)$ é significativa
- Se H_o não for rejeitada, não teremos evidências suficientes para concluir que existe uma relação significativa.

Rejeitar H_o se $\mathrm{p_{valor}} \leq \alpha$

T

Regressão Linear Multipla – Teste T (significância individual)

Se o teste F demosntrar que a relação de regressão múltipla é significativa, um teste t pode ser realizado para determinar a significância de cada um dos parâmetros individuais.

Para qualquer parâmetro $\begin{cases} H_o: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

- Se H_o for rejeitada, o teste nos dá suficientes evidências estatísticas para concluirmos que o parâmetro (β_i) em estudo não é igual a zero, ou seja, o parâmetro é estatisticamente significativo. Este teste deve ser feito com todos os parâmetros.
- Se H_o não for rejeitada, não teremos evidências suficientes para concluir que o parâmentro (β_i) em estudo é estatisticamente significante.

Rejeitar H_o se $\mathrm{p_{valor}} \leq \alpha$

Estudo de Caso
Ajustando um Modelo de Regressão Linear Multipla

Parte 4 : Ajustar um modelo de regressão linear múltipla para determinar a relação entre a Altura do Pai e Altura da Mâe com a Altura do Filho.

- statsmodel

Suposições do Modelo

Suposições sobre o termo de erro e no modelo de regressão múltipla

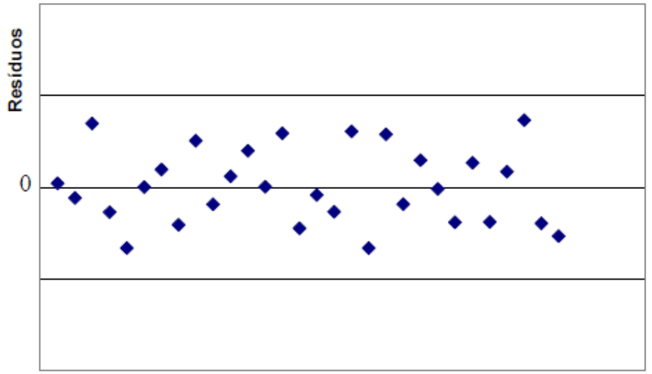
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

1. **Média zero**: o erro ϵ é uma variável aleatória com média zero.
2. **Variância constante**: a variância de ϵ, designada por σ^2, é idêntica para todos os valores das variáveis independentes $x_1, x_2, \ldots, x_p$.
3. **Valores independentes**: Os valores de ϵ são independentes.
4. **Distribuição Normal**: O erro ϵ é uma variável aleatória com distribuição Normal refletindo o desvio entre y e o valor ajustado de y (conhecido como $\hat{y}$).

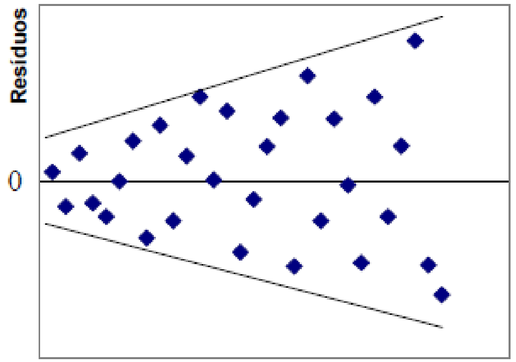
Suposições do Modelo – Gráfico dos Resíduos

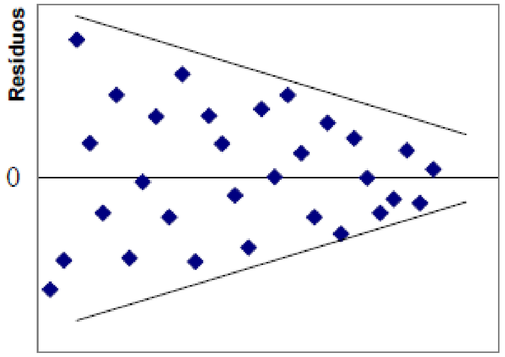
Neste gráfico analisamos as suposições 1, 2 e 3 respectivamente, dadas por: **média zero, variância constante e independência dos resíduos**.

Bom padrão

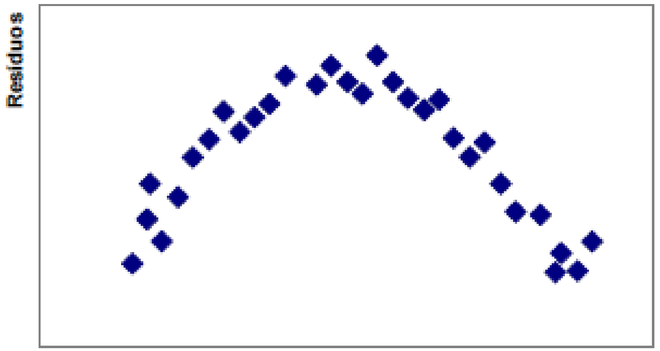


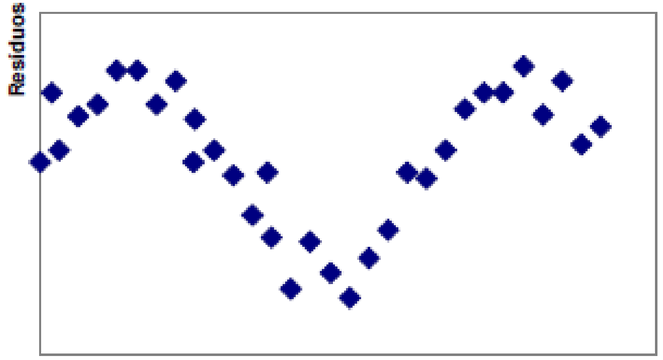
Variância não constante





Forma não adequada do modelo

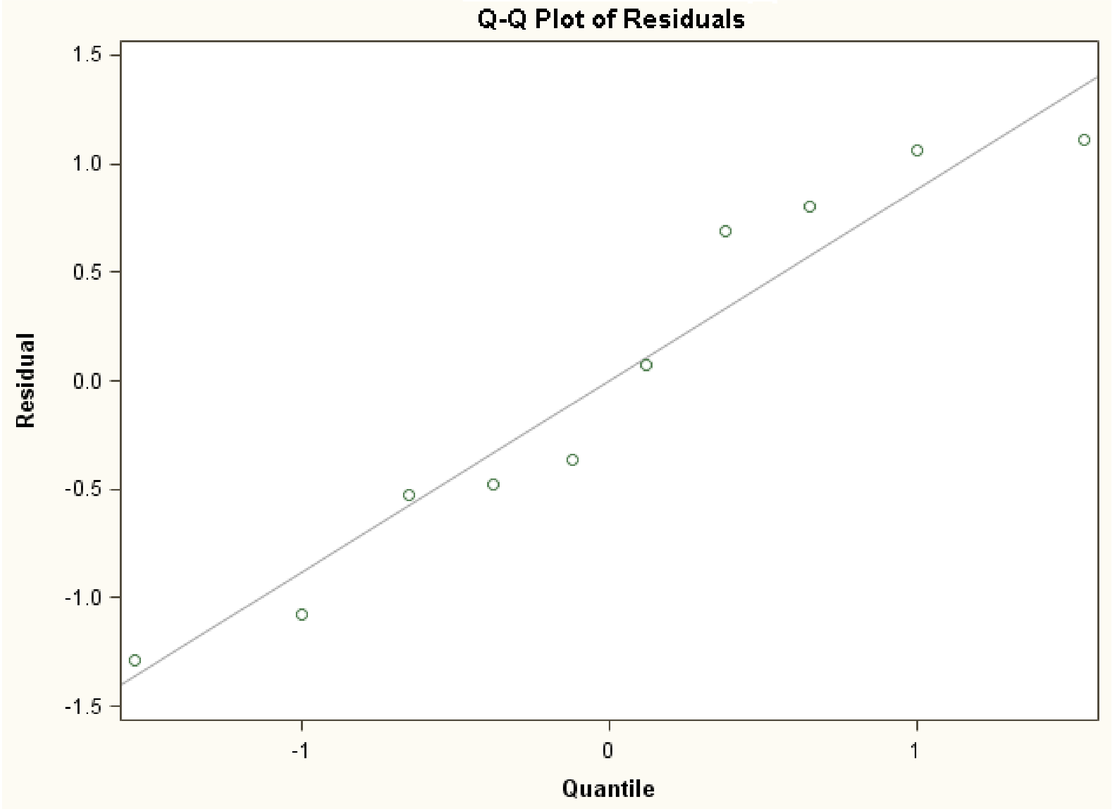




Suposições do Modelo – QQ-plot

Neste gráfico analisamos a suposição 4, onde verificamos se os resíduos são provinientes de uma **distribuição Normal**.

Espera-se uma linha com inclinação de 45 graus.



Suposições do Modelo – Comentários

- Utilizamos plotagens residuais para validar as suposições de um modelo de regressão. Se nossa revisão indicar que uma ou mais suposições são questionáveis, um modelo de regressão diferente ou uma transformação dos dados devem ser considerados.
- A análise de resíduo é o principal método que os estatísticos usam para verificar se as suposições associadas a um modelo de regressão são válidas. Mesmo que nenhuma infração seja encontrada, não quer dizer necessariamente que o modelo produzirá boas previsões. Entretanto, se testes estatísticos adicionais sustentarem a conclusão de significância e se o coeficiente de determinação for grande, seremos capazes de desenvolver boas estimativas e previsões usando a equação de regressão estimada

Multicolinearidade

- Multicolinearidade refere-se à correlação entre as variáveis independentes.
- Quando as variáveis independentes são altamente correlacionadas não é possível determinar o efeito separado de uma variável independente na variável dependente.
- Um coeficiente de correlação amostral maior que 0,70 ou menor que -0,70 para duas variáveis independentes é um aviso prático de que há potenciais problemas com a multicolinearidade.
- Sob condições de elevada multicolinearidade, as estimativas pelo método de mínimos quadrados podem ter um sinal oposto ao do parâmetro que é estimado.

Estudo de Caso
Ajustando um Modelo de Regressão Linear Multipla

Parte 5 : Ajustar um modelo de regressão linear múltipla para o Boston Housing Dataset e verificar a qualidade do ajuste do modelo

DÚVIDAS?!

Obrigada

Cristiane Rodrigues

crisrodrigues_27@hotmail.com