
Tera

Algoritmos de Regressão Linear

Instrutor

Luiz Henrique Outi Kauffmann

- **Graduado em Matemática Aplicada e Computação Científica – USP São Carlos.**
- **Mestrando em Estatistica Aplicada - – USP São Carlos (Defendo dia 20/nov).**
- **Iniciei a carreira com Collection Analytics – Itaú -2008**
- **Especialista em Credit Risk Analytics – Itaú, HSBC, Pine, Serasa**
- **Experiência com Modelos de Fraude em e-commerce – Serasa**
- **Líder do time de consultoria analítica no SAS Inc.**
- **Apaixonado por Filosofia e filosofia da Ciência**
- **Apaixonado por Video-Games(RPG's e JRPG's), Filmes e Séries.**
- **Apaixonado por Judo e Jiu Jitsu**

Expectativas

- **Qual o Contexto de Modelagem está incluído a Regressão Linear?**
- **O que pode ser considerado como Aprendizagem Automática?**
- **Modelos Estatísticos (“Old School”) & Machine Learning**
- **Análise de Regressão e o Pensamento Econométrico**
- **Estudo de Galton e as Origens da teoria da Regressão Linear**
- **Mínimos Quadrados**
- **Ajustar e aplicar o conheciemento em Python**

Discussão:

O que é aprendizado?

Aprendizagem computacional é um conjunto de métodos que podem detectar padrões em dados automaticamente para depois usar os padrões descobertos para predizer dados futuros. Ela geralmente é dividida em dois tipos principais: aprendizado **supervisionado** e **não-supervisionado**.

Discussão:

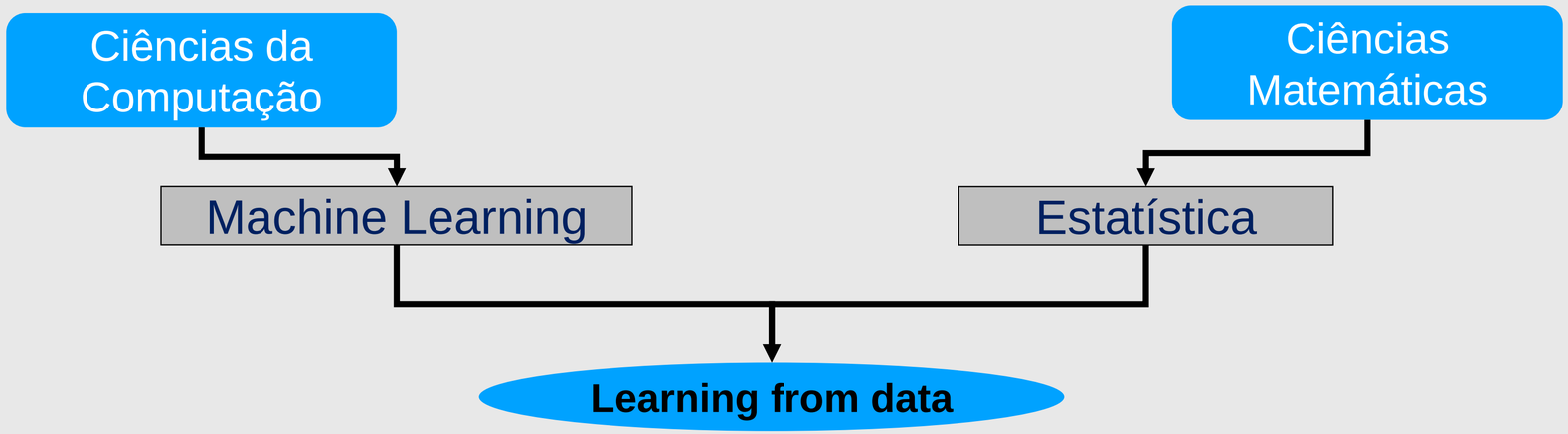
Machine Learning VS Statistics?

Mesmo Objetivo

- *Machine Learning is an algorithm that can learn from data without relying on rules-based programming.*

- *Statistical modeling is a formalization of relationships between variables in the data in the form of mathematical equations.*

machine learning = "glorified statistics."



Discussão:

Machine Learning VS Statistics?

Statistics	Machine Learning
Estimation	Learning
Classifier	Hypothesis
Data Point	Example/ Instance
Regression	Supervised Learning
Classification	Supervised Learning
Covariate	Feature
Response	Label

Discussões Semelhantes

- **Modelos Economicos VS Modelos Estatísticos**
- **Modelos Bayesianos vs Modelos ´Classicos**
- **Modelos Estruturais Vs Modelos de Forma Reduzida**

Conceito:

Modelos Supervisionado

- **Análise de Regressão**
- **SVM**
- **Rede Neural Supervisionada**
- **Árvore de Decisão**
- **Gradient Boosting**
- **Random Forrest**

Modelo Não Supervisionado

- **Clusterização (K-Means e outros)**
- **Rede Neural não supervisionada**
- **Redes Complexas (Social Networking Analytics)**

História:

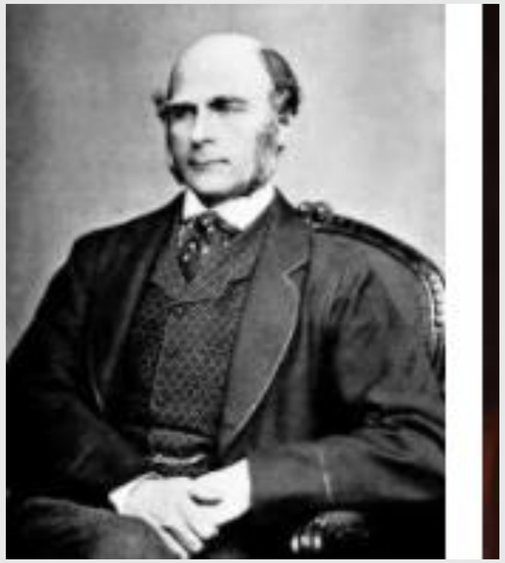
Galton e o estudo sobre a altura

Francis Galton (Birmingham, 16 de fevereiro de 1822 — Haslemere, Surrey, 17 de janeiro de 1911) foi um antropólogo, meteorologista, 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. O primeiro livro importante para o pensamento de Galton foi *Hereditary Genius* (1869). A sua tese afirmava que um homem notável teria filhos notáveis.

ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

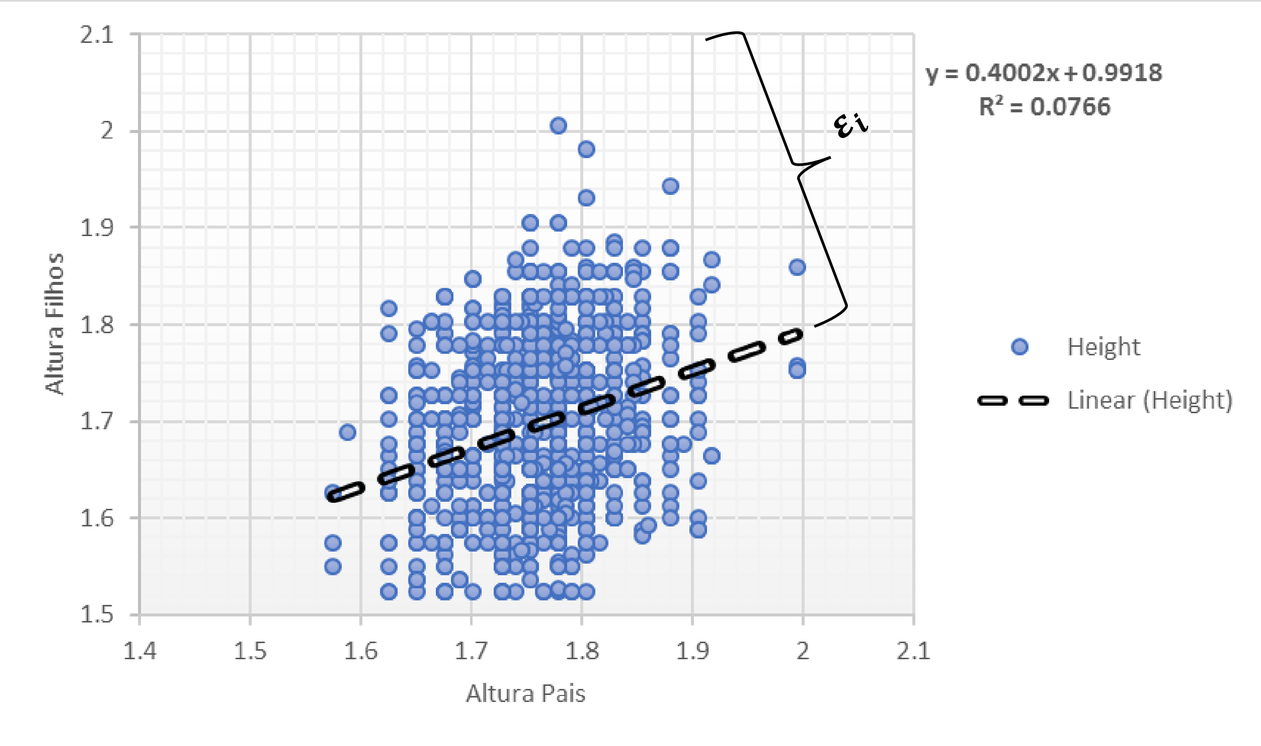
[WITH PLATES IX AND X.]



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

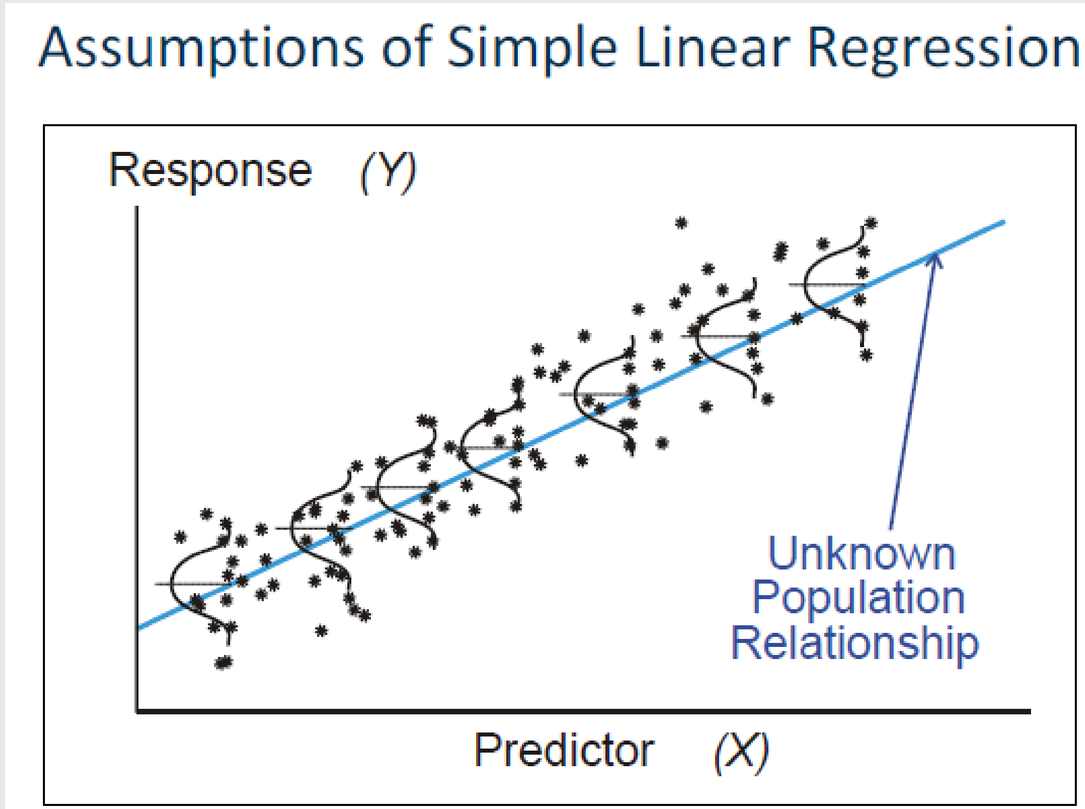
Filosofia Matemática:

Análise de Regressão



$$Y_{alturafilhos} = 0.4002X_{alturapais} \cdot + 0.9918 + \underbrace{\varepsilon_i}$$

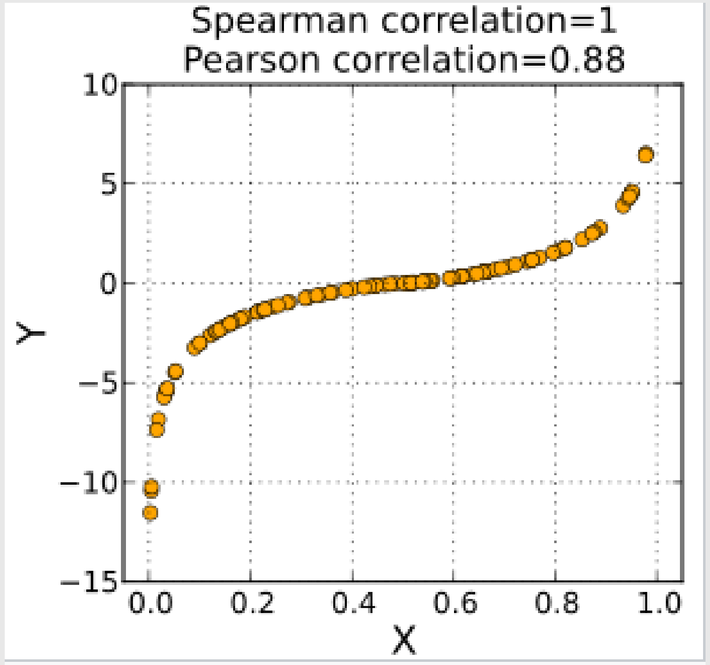
Ruído Aleatório



Prática:

Estudo da Correlação na Base do Galton

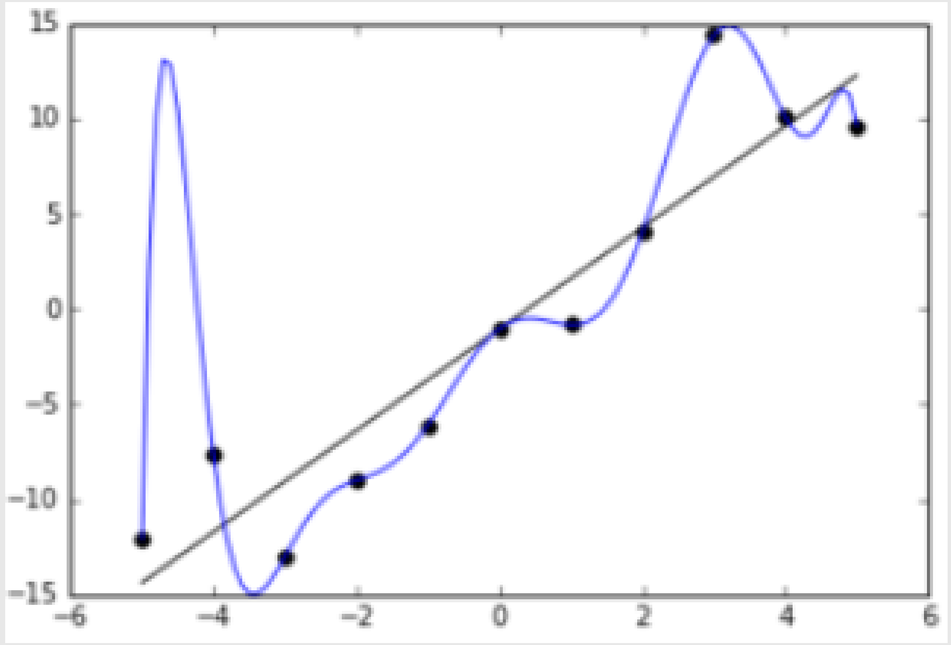
1. Executando Coeficiente de Correlação de Pearson e Spearman no Python?
2. Qual a melhor variável para explicar a altura dos filhos?
3. Qual a distribuição da altura dos filhos, pais e mães?
4. Ajustar uma Regressão Linear com a base do Galton.



Pergunta Prática:

Porque ajustar uma Reta e não uma curva que passe por todas as observações?

- *Overfiting?*
- *Interpolação?*



Conceito:

Como encontrar o Modelo Linear?

Modelo Teórico : $Y_{alturafilhos} = \beta_1 X_{alturapai.} + \beta_0$

Modelo Prática : $Y_{alturafilhos} = \beta_1 X_{alturapai.} + \beta_0 + \varepsilon_i$

Altura Observada : $Y^*_{alturafilhos}$

Otimizar

Minimizar : $\varepsilon_i^2 = \left(Y^*_{alturafilho} - \beta_1 X_{alturapai.} + \beta_0 \right)^2$

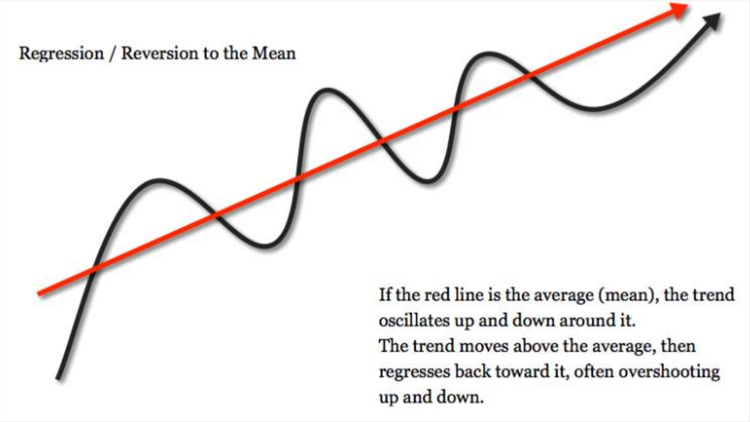
Mínimos Quadrados Ordinários (OLS) : Encontrar matematicamente os valores dos Beta's que para cada registro da minha base minimizam o erro ao quadrado!!!

Conceito:

Porque existem premissas?

- **Se o Galton aumentasse a amostra utilizada será que encontraríamos outros Betas?**
- **Se o erro apresenta correlação com o SEXO dos filhos?**
- **Porque não usar outras informações como alimentação dos Filhos?**
- **Robustez Estatística?**

- **Para garantir a robustez dos estimadores é necessário que o erro apresente distribuição normal.**
- **Conceito de RETORNO A MÉDIA**



Conceito:

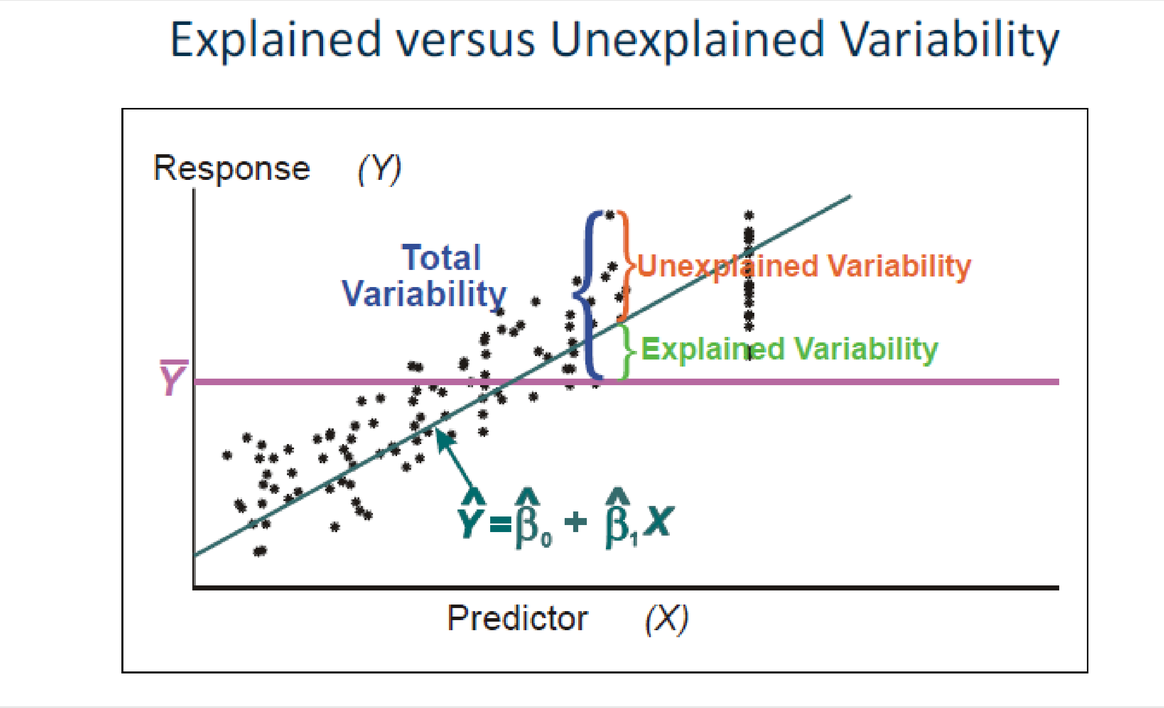
Quais as Premissas do Modelo de Regressão Linear?

1. **Erros (Ɛ_i) possuem média 0**
2. **Erros apresentam distribuição normal**
3. **Variância dos erros (Ɛ_i) – Homocedástico**
4. **Correlação entre os erros deve ser próxima de zero ou nula**
5. **Variáveis explicativas não podem ser uma combinação linear de outras variáveis explicativas**

Conceito:

Como Avaliar o Modelo de Regresão Linear?

1. **R- Quadrado**
2. **R Quadrado Ajustado**



SQ Exp $\sum\left(\hat{Y}_i-\bar{Y}\right)^2$

SQ erros $\sum\left(Y_i-\hat{Y}_i\right)^2$

SQ tot $\sum\left(Y_i-\bar{Y}\right)^2$

R-Quadrado = $\frac{SQ\ Exp}{SQ\ tot} = 1 - \frac{SQ\ erros}{SQ\ tot}$

R-Quadrado-Ajustado = $1 - \frac{SQ\ erros\ /(n-k)}{SQ\ tot/(n-1)}$

- **n= Qtdade de Registros**
- **K=Qtdade de Variáveis Explicativas**

Conceito:

Como Avaliar o Modelo de Regresão Linear?

- **Analisar R-Quadrado em amostra de treinamento e teste!**
- **Analisar periódicamente o R-quadrado com novos dados, escorar novas informações e calcualr R-quadrado !**

Conceito:

Inflação dos betas e relevância das Variáveis Explicativas?

Conceito:

Regressão com mais variáveis e Inflação dos betas e relevância das Variáveis Explicativas?

Conceito:

Inflação dos betas e relevância das Variáveis Explicativas?

Colinearidade significa que as variáveis independentes são correlacionadas.

A colinearidade pode ser detectada, dentre outros modos, através da

matriz de correlação entre as varáveis.

Outra técnica usada é o fator de inflação de variação (variance inflator

factor), cujos altos valores indicam a sua existência

$$VIF = \frac{1}{1 - R_j^2}$$

Sendo Rj resulta da regressão de Xj com as outras variáveis.

No caso de VIF alto (acima de 10) procura-se remover a variável que apresenta pior relevância estatística e menor correlação com a variável resposta

Conceito:

Qual o melhor Modelo ?

Importante: As variáveis explicativas afetam o p-valor de outras variáveis!!. Exemplo: temos uma regressão com uma variável explicativa e seu p-valor é 1%, se for treinar um novo modelo incluindo outra variável, pode ser que o modelo com 2 variáveis apresente um novo p-valor para as variáveis que já estavam no modelo de regressão

Backward Regression : Algoritmo

1. Treina o modelo com todas as k-variáveis explicativas disponíveis
2. Exclui a variável que possui p-valor > critério (5%)
3. Repete o passo 1 e 2 até que todas as variáveis apresentem um p-valor relevante

Forward Regression : Algoritmo

1. Treina o modelo com todas as 1-variável explicativa disponível
2. Se o p-valor <= critério (5%) mantém variável, senão excluí a variável
3. Adiciona nova variável
4. Repete o 1,2 e 3 até que todas as k-variáveis disponíveis estejam testadas, garantindo um conjunto final de variáveis com p-valor relevante

Forward in Python

http://planspace.org/20150423-forward_selection_with_statsmodels/

Stepwise Regression : Algoritmo - Mix do Forward com Backward

1. Executa uma primeira iteração da Forward Regression
2. No seguindo passo verifica todos os p-valores de todas as variáveis, caso uma delas passe a assumir um p-valor > critério, esta será descartada.
3. Executa 1 e 2

Prática:

Criticar Modelo de Regressão Ajustado?

DÚVIDAS?!