Semana 2.2 - Extensões da Regressão Linear

Prof. Raphael Y. de Camargo
Centro de Matemática, Computação e Cognição (CMCC)
Universidade Federal do ABC

UFABC

Parte I
Regressão Linear Múltipla

Regressão Linear Múltipla

Em diversas situações temos diversos preditores para uma resposta

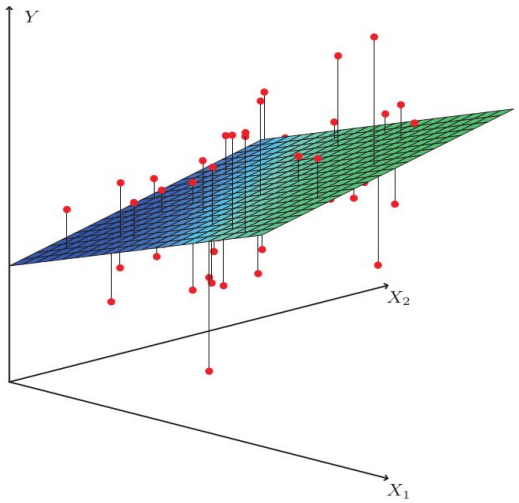
- No exemplo da propaganda, esta poderia ser na TV, rádio e jornal
- Mas e se quisermos investir nas 3 mídias ao mesmo tempo?

Neste caso, poderíamos fazer uma regressão combinando os 3

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$$

De um modo geral, a regressão múltipla é dada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$



O processo de determinar coeficientes é similar ao caso da regressão simples

Mas com matrizes ao invés de variáveis simples

Interpretando os resultados

Temos abaixo o resultado da regressão dos dados de vendas vs anúncios para o caso de regressão simples (abaixo) e múltipla (direita).

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Regressão múltipla: coeficientes indicam o que ocorre quando modificamos um dos preditores

Note que investimentos em jornais não ajudam nas vendas

Porque na regressão simples ele aparece como relevante?

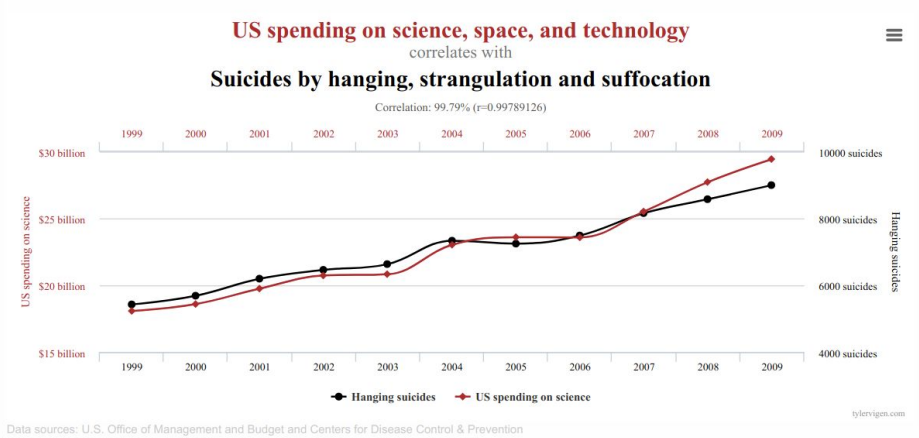
	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

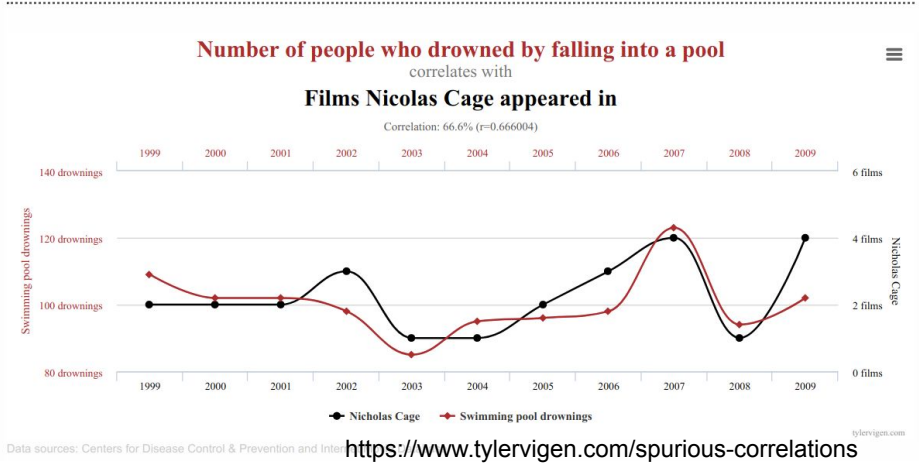
Correlações Espúrias

Duas variáveis podem estar altamente correlacionadas, mas não significa que exista causalidade

Nos casos ao lado isto é óbvio, mas em outros não, como das propagandas em jornais

Elas podem ocorrer por coincidência (ao lado), ou por existir outra variável confundidora





https://www.tylervigen.com/spurious-correlations

Existe Relação entre os Preditores e a Resposta?

Preditores individuais:

- Valor do coeficiente
- Erro padrão
- Estatística-t
- p-valor

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

Modelo como um todo:

- Erro residual
- R-quadrado

Quantity	Value
Residual standard error	1.69
R^2	0.897

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Selecionando Variáveis

Quando temos muitas variáveis, podemos selecionar apenas uma parte delas:

Seleção Progressiva: adiciona as variáveis uma a uma, pelo menor RSS

Seleção Regressiva: retira as variáveis com maior p-valor, uma a uma

Seleção Misturada: adiciona as variáveis uma a uma, mas retira se alguma delas passar a ter um p-valor maior que um limiar

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Parte II
Extensões da Regressão Linear

Gerando Preditores Não-Lineares

Existem diversos casos onde podem haver interações entre os preditores.

Podem ser capturadas com termos que multiplicam as variáveis

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

Adicionamos os efeitos individuais e os efeitos combinados das variáveis

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

O modelo continuar linear

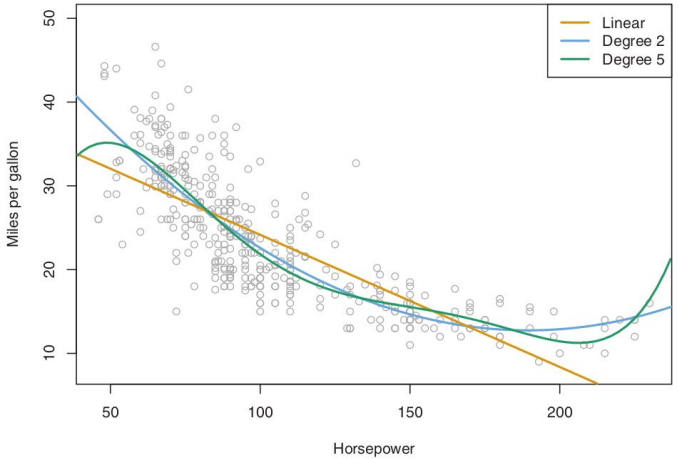
- Mas criamos novos preditores compostos

Relações Não-Lineares

Consumo de combustível x potência do motor

Sempre pode-se tentar realizar transformações nos preditores para que a resposta seja proporcional ao preditor

Aqui, o melhor é fazer horsepower²



$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

Relações Não-Lineares

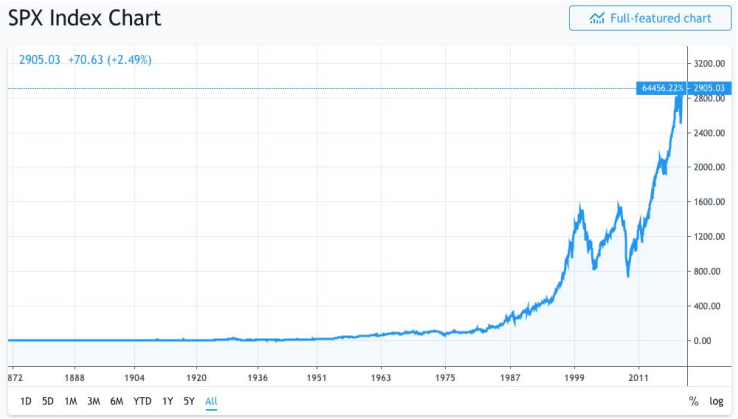
Retornos compostos:

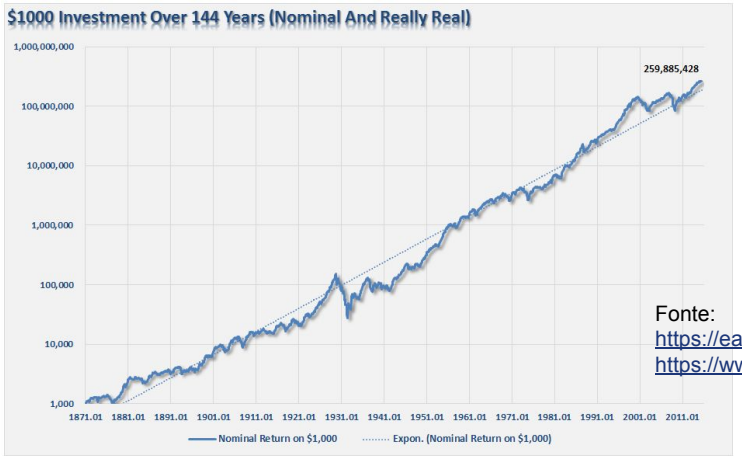
- São exponenciais: $v(\mathbf{t}) = v_0 * (1+CAGR)^{\wedge}\mathbf{t}$

Mas podemos fazer:

$\log v(\mathbf{t}) = \log v0 + \mathbf{t} * \log(1+CAGR) = \alpha + \beta\, \mathbf{t}$

Podemos descobrir então o ganho composto utilizando uma regressão linear!





Fonte:
https://earthinnovation.org/
https://www.investing.com/

Parte III
Potenciais Problemas da Regressão Linear

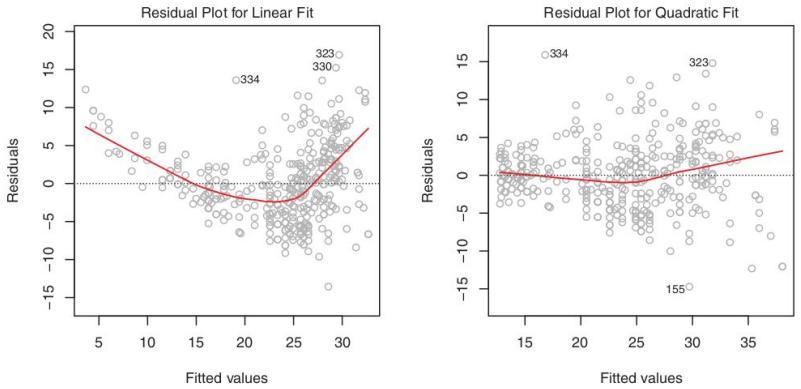
Não-linearidade nos dados

A regressão calcula uma aproximação

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

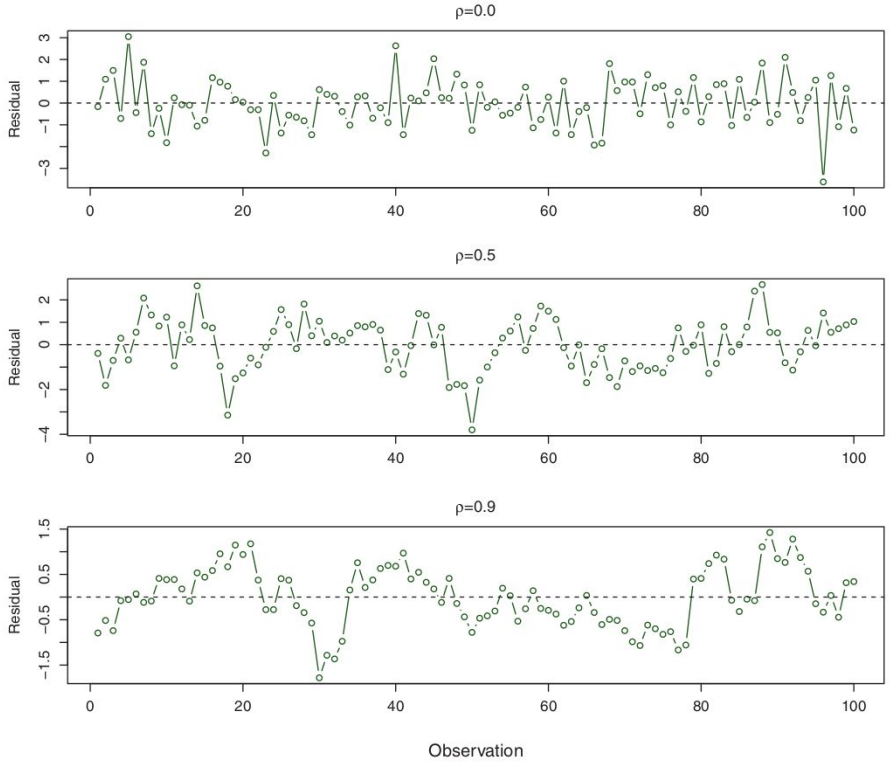
E queremos que o erro residual $e_i = y_i - \hat{y}_i$ seja uniforme

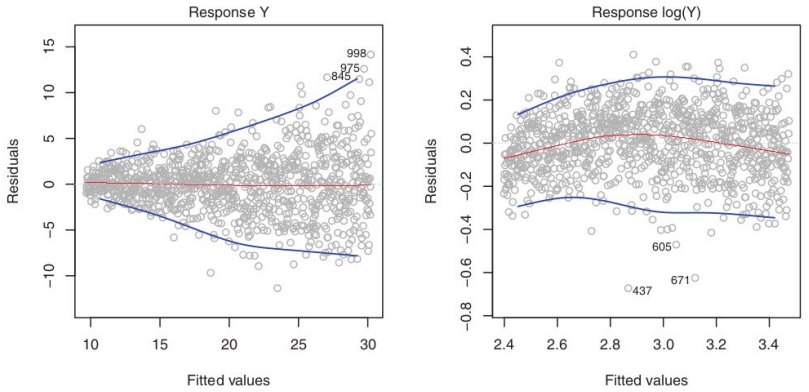
Se não for, significa que a relação não é linear



Resíduos do exemplo da potência do motor x consumo
Esquerda: regressão com potência
Direita: regressão com potência²

Correlações e Variância Não-Constante nos Termos de Erro

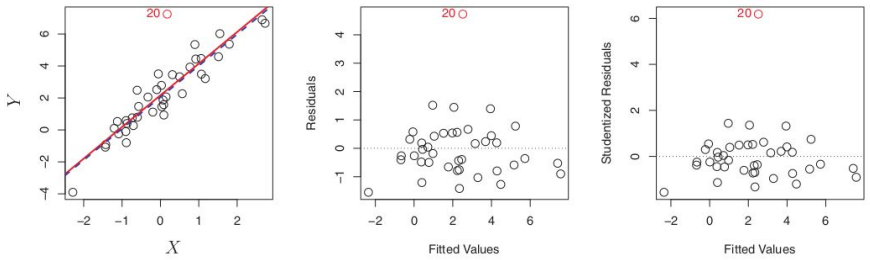




Termos de erro devem ser:

- Independentes
- Com variância constante com o valor do preditor

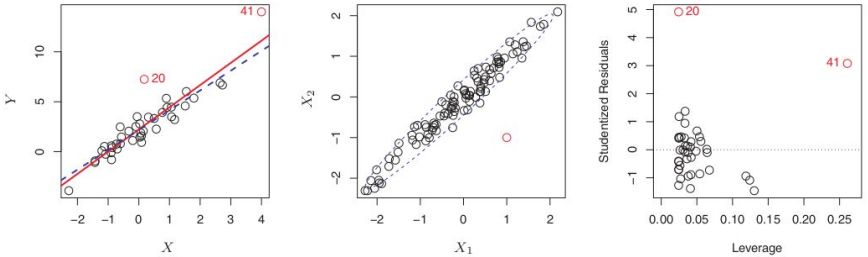
Outliers



Podem representar um erro na captura de dados

- Muitas vezes podem ser visualizados, especialmente ao olharmos para os resíduos
- Estudentização dos resíduos: basta dividir pelo erro padrão estimado. Valores acima de 3 podem indicar um outlier

Cuidado para não retirar dados importantes!



Se os outliers estiverem em pontos de alta alavancagem, eles podem alterar o resultado

Uma dimensão: fora da faixa dos demais valores

Mais dimensões: mais difícil encontrar

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

Colinearidade

Quando duas variáveis tem forte relação linear

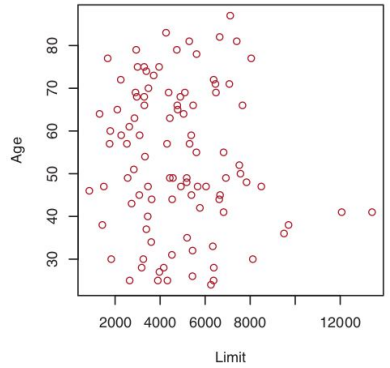
Difícil separar o efeito de cada uma na resposta

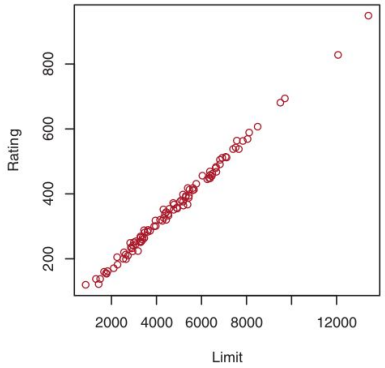
Pode esconder a significância de algum preditor

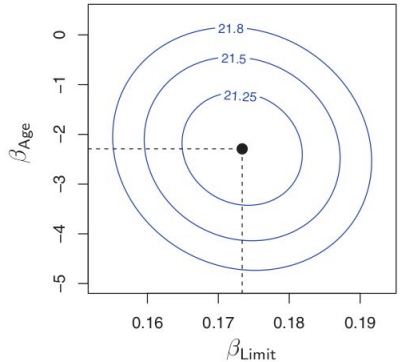
Solução 1: retirar um dos preditores

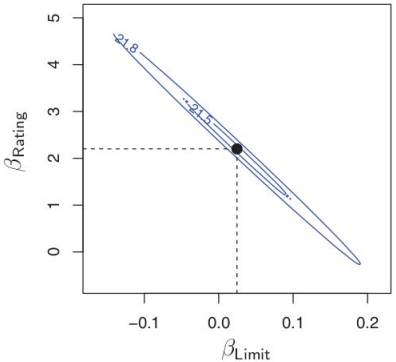
Solução 2: combinar os preditores em um só

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012









TODO

Gerar exemplos sintéticos onde as técnicas podem ser úteis