# Reposta para o desafio IEL-CNPq

*Raul de Sá Durlo*

*16/07/2019*

## Contents



## 1 Introdução

No desenvolvimento de modelos de predição, qual a diferença entre as técnicas de regressão linear e regressão logística? Quais são os indicadores para avaliar a performance de aderência do modelo?

O desafio foi respondido e é apresentado neste relatório dividido nas partes que seguem:

- Definição de modelos de predição, diferenciando o da inferência estatística.
- Apresentação dos modelos de regressão linear e logísticos definidos.
- Por fim, é discutido os indicadores de performance de aderencia para modelos preditivos.

## 2 Uma breve introdução aos modelos estatísticos

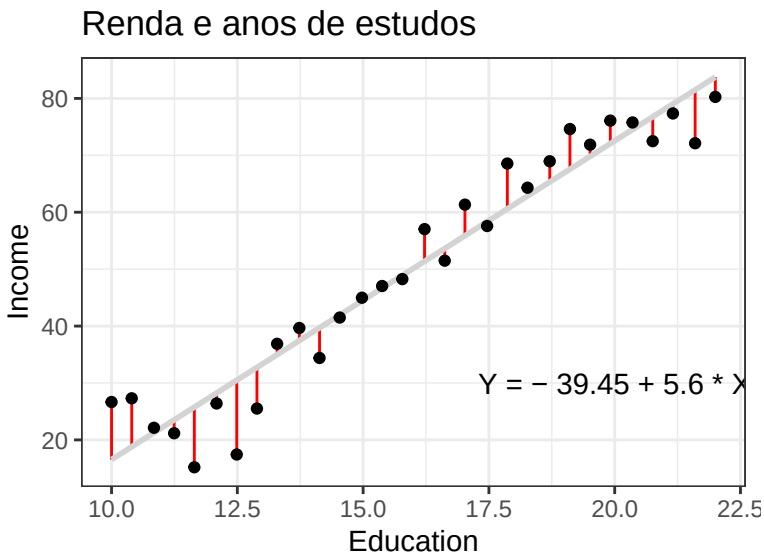
Em um modelo estatístico estamos interessados em obter uma função $f$ que relacione um conjunto de *preditores* ($X$) a alguma *variável de resposta* ($Y$). Os preditores $X = (X_1, X_2, ..., X_p)$ também são chamados de *variáveis explicativas*, *variáveis independentes* ou *entrada* (*inputs*).

Podemos descrever relação entre $X$ e $Y$ na forma geral:

$$Y = f(X) + \epsilon$$

Onde $f(X)$ representa uma relação sistemática entre o conjunto de preditores $X$ e a variável de resposta $Y$ e $\epsilon$ é um termo de erro aleatório independente, com média igual a zero.

O gráfico abaixo mostra uma relação bi-variada entre a renda ($Y$) de 30 indivíduos com os seus respectivos anos de estudos ($X$). Cada indivíduo pode ser identificado por um ponto no gráfico e a reta cinza é a representação de um modelo linear simples. A principal característica desse modelo é que a de que minimiza a distância entre os seus valores preditos ($\hat{Y}$) e os valores observados ($Y$) (em vermelho).




Existem basicamente dois motivos para estimar $f$: a *inferência* e a *predição*.

## 2.1 Predição

Muitas vezes não podemos obter, de antemão, os valores de $Y$. Por isso, os valores preditos são importantes e estão representados pela reta cinza no gráfico acima. Os valores preditos são estimados por alguma forma funcional assumida para o modelo com as variáveis explicativas (seus erros possuem média zero):

$$\hat{Y} = \hat{f}(X)$$

O foco na predição serve para analisar a precisão de um modelo, ou seja, se os seus valores preditos $\hat{y_0}$ acertariam os valores reais $y_0$. Entretanto, no geral, $\hat{f}$ não é um estimador perfeito de $f$ e sua diferença (erro) é explicada por fatores *redutíveis* e *irredutíveis*. Podemos decompor esses fatores por meio do quadrado das diferenças entre o valor estimado e a variável de resposta:

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = [f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

O termo de erro *redutível* corresponde ao termo $[f(X) - \hat{f}(X)]^2$ da equação acima e é aquele decorrente da escolha da forma funcional estimada. Assim, o modelo mais preciso é aquele que minimiza essa diferença.

Como $Y$ é função de $\epsilon$ e, por definição, $\epsilon$ não pode ser previsto por $X$. Algum erro sempre será introduzido ao modelo, daí o termo *irredutível*, denotado por $Var(\epsilon)$.

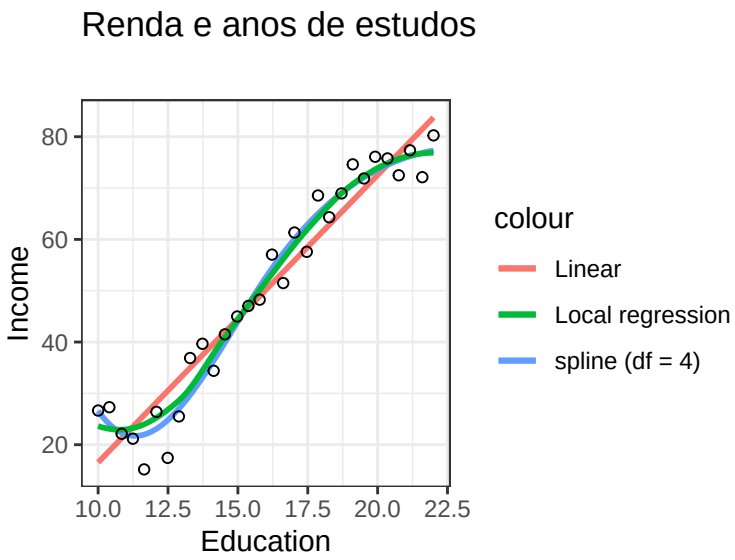
## 2.2 Inferência

A inferência serve para análise da *maneira* de como os preditores $X_1, X_2, ..., X_p$ afetam a variável de resposta $Y$. Fazemos inferência quando queremos entender a relação entre $X$ e $Y$ ou como $Y$ muda em função de $X_1, X_2, ..., X_p$.

O modelo estatístico utilizado no gráfico acima é um exemplo de um modelo paramétrico linear simples. Paramétrico, pois assume uma forma funcional definida (linear, no caso) e simples, pois possui apenas um preditor.

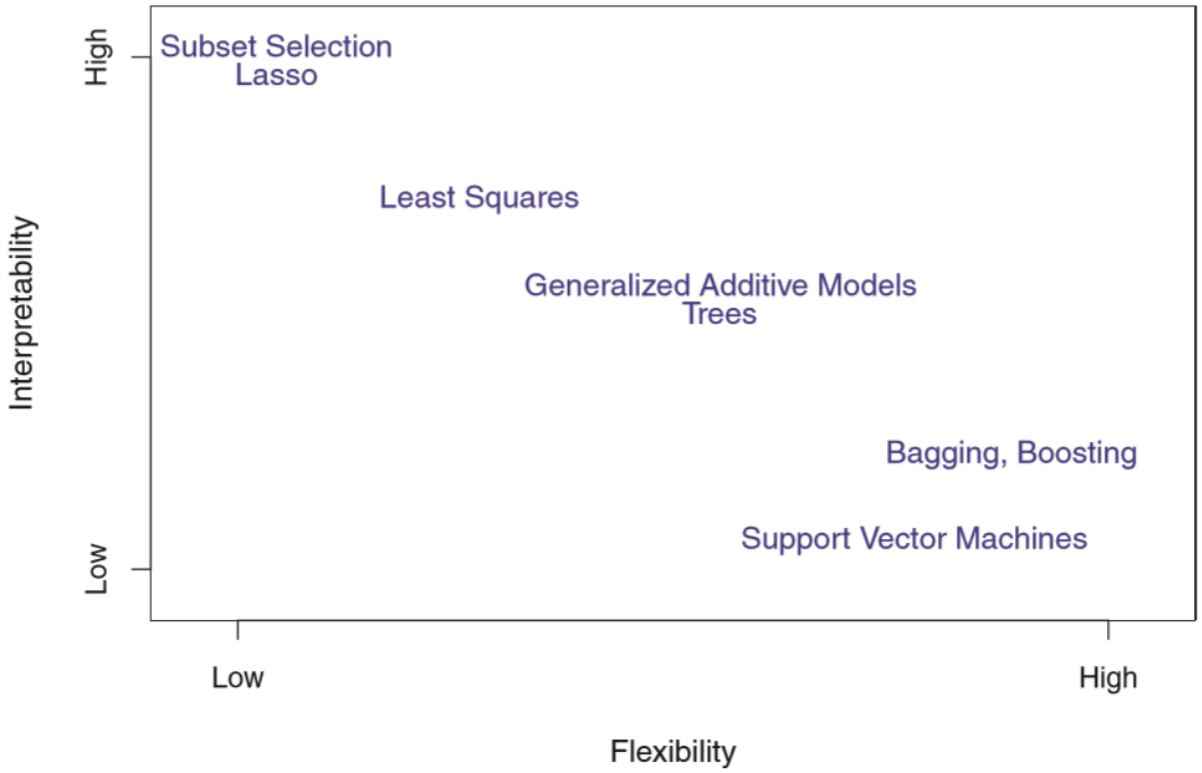
Um modelo paramétrico normalmente apresenta a desvantagem de ser mais simplificador e inflexível. Aumentar sua complexidade e flexíbilidade implica em aumento no número de parâmetros a serem estimados e,

consequentemente, os modelos se tornam mais sensíveis aos erros (*overfitting*). Por outro lado, a simplificação pode ser interessante por questões de interpretabilidade.




Existem também modelos não paramétricos, onde é feita suposições sobre a forma funcional de $f$ e, portanto, possui a capacidade de se ajustar melhor ao conjunto de dados, com a desvantagem de apresentar mais variabilidade e, consequentemente, ser mais sensível ao termo de erro (*overfitting*).

Novamente, a flexibilidade do modelo deve ser confrontada com a sua intepretabilidade. A figura abaixo representa o tradeoff entre interpretabilidade e flexibilidade de diferentes modelos estatísticos. Em geral, um modelo mais fácil de se interpretar é preferido quando o objetivo é a inferência e um modelo mais flexível é mais recomendado para análises mais preditivas.




Os modelos citados acima podem compor procesos de aprendizagem estatística (*statistical learning*)

# 3 Regressão linear

# 4 Regressão Logística

# 5 Indicadores de performance e aderência do modelo