Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 - Bioestatística

Regressão Logística

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

- Para aqueles que já tem o repositório local:

```
> cd /path/to/IMD0601
```

```
> git pull
```

Regressão Logística (na saúde)

- Relação entre uma ou mais características relacionadas ao paciente e uma **variável *Resultado* binária**.
- Pode ser aplicada para variável categórica com mais de dois resultados (*regressão ordinal*).
- "*Problemas de Classificação*" junto com técnicas de ML.
- Vantagens: fácil de executar e possibilita descrever as relações entre as variáveis de forma explícita.
- Modela **chances** em vez de probabilidades *(em log odds)*.

Informação	Nome	Legenda
Dados pessoais	x	Ordem dos dados
	id	Identificação da pessoa
	location	Local em que reside
	age	Idade
	gender	Gênero
	time.ppn	Tempo pós-prandial
	insurance	Seguro
	smoking	Fumante
	fh	Histórico familiar de DM
	dm	Diabetes Mellitus
Dados físicos	height	Altura
	weight	Peso
	frame	Estrutura corporal
	waist	Cintura
	hip	Quadril
Dados clínicos	chol	Colesterol total
	stab.glu	Glicose estabilizada
	hdl	Lipoproteína de alta densidade
	ratio	Proporção de colesterol / HDL
	glyhb	Hemoglobina Glicosilada (HbA1C)
	bp.1s	Primeira pressão arterial sistólica
	bp.1d	Primeira pressão arterial diastólica
	bp.2s	Segunda pressão arterial sistólica
	bp.2d	Segunda pressão arterial diastólica

biostat.mc.vanderbilt.edu/wiki/Main/DataSets

Chance *vs.* Probabilidade

- **Probabilidade (risco)**: razão da ocorrência de um evento (sucesso) sobre o número total de tentativas (sucesso + insucesso). Varia entre 0 e 1.

- **Chances (*odds*):** razão da probabilidade do evento ocorrer (sucesso) sobre a probabilidade de não ocorrer (insucesso) = **p / (1 - p).** Sempre positivo.
    - *Se um cavalo correr 100 corridas e ganhar 80, a probabilidade de vitória é 80/100 = 0,80 ou 80%, e as chances de vitória são 80/20 = 4 vitórias para 1 derrota.*
    - *Se um cavalo correr 100 corridas e ganhar 5, a probabilidade de vitória é de 0,05 ou 5%, e as chances do cavalo ganhar são 5/95 = 0,052 para 1 (~ 1 para 19)*

Chance *vs.* Probabilidade

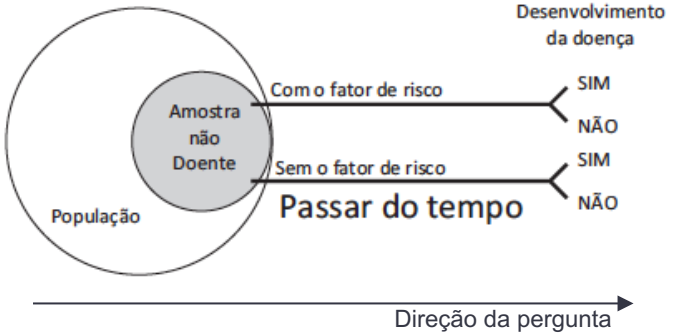
- *Quando a probabilidade é baixa (~10%), a chance e probabilidade são muito semelhantes.*

- *A medida que a probabilidade aumenta, a chance cresce de forma não-linear.*

Probabilidade (p)	Chance p / (1-p)	log(chance)*
0,001	0,001	-3,000
0,01	0,01	-2,000
0,1	0,111	-0,955
0,2	0,25	-0,602
0,3	0,429	-0,368
0,4	0,667	-0,176
0,5	1	0,000
0,6	1,5	0,176
0,7	2,333	0,368
0,8	4	0,602
0,9	9	0,954
0,99	99	1,996
0,999	999	3,000

Estudos de coorte *(razão de probabilidade, risco relativo)*

- Investiga a associação entre um fator de risco e um evento (doença).
- Inicia com grupos de indivíduos expostos e não-expostos ao risco e depois compara a incidência (*risco*) do evento nos 2 grupos.
- *Fatores de risco*: idade, local, gênero, substância, comportamento, mutação, ou qualquer outra característica.
- Interpretação de RR: igual, maior ou menor que 1.
- RR fornece uma medida relativa do aumento ou diminuição da incidência entre os grupos expostos e não-expostos.

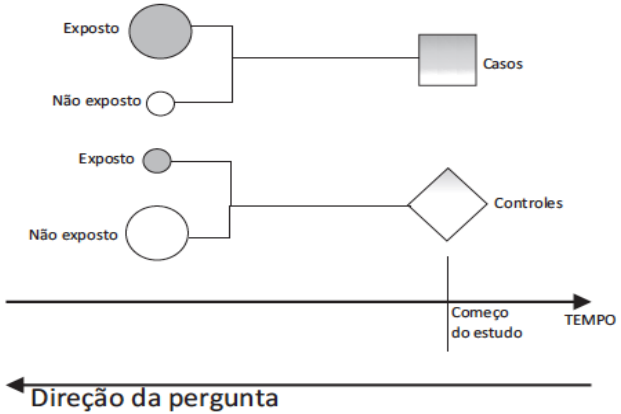


	DOENÇA +	DOENÇA -
EXPOSTO	**a**	**b**
NÃO-EXPOSTO	**c**	**d**

- Incidência em indivíduos expostos = **a/(a+b)**
- Incidência em individ não-expostos = **c/(c+d)**
- Risco relativo (**RR**) = **a/(a+b) / c/(c+d)**.

Estudos Caso-Controle *(razão de chance, odds ratio)*

- Investiga a associação entre um fator de risco e um evento, quando não é possível calcular o risco relativo.
- Inicia com a seleção de ‘casos’ e ’controles’ e compara se a presença/exposição ao fator de risco é desproporcionalmente distribuída nos 2 grupos.
- Interpretação de OR: igual, maior ou menor que 1.
- Se a prevalência do evento for elevada, o OR é sobrestimado e deve utilizar o risco relativo *estimado*.
- Aplicado também na Regressão Logística!



	CASOS	CONTROLES
EXPOSTO	**a**	**b**
NÃO-EXPOSTO	**c**	**d**

***OR** = chance exposição entre casos (a/c) / chance de exposição entre controles (b/d)*

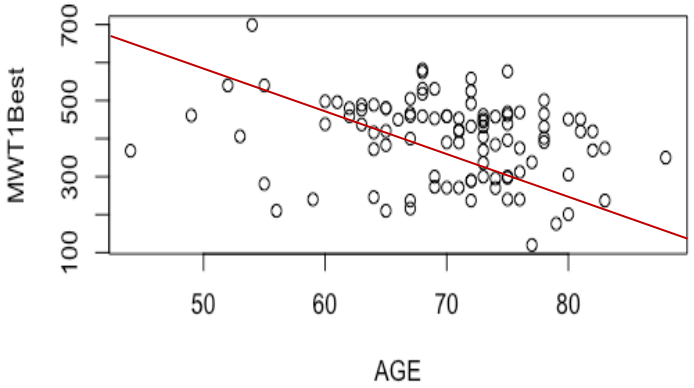
Por que a regressão linear não funciona com resultados binários?

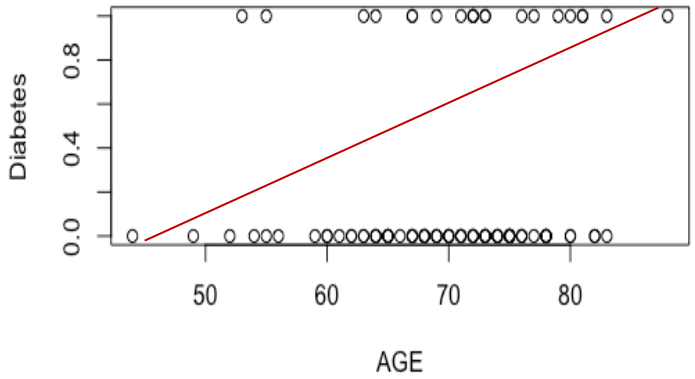
- Reg. Logistica: modela a proporção de indivíduos com o *resultado* de interesse, isto é, a probabilidade, com valores em 0 ou 1.
- Reg. Linear: pode prever valores ≠ 0 e 1.
- ***Logit*: 'função de ligação' que transforma o *Resultado* em uma variável que pode ser modelada na equação de regressão.

$$\text{logit}\,(p) = \log\left(\frac{p}{1-p}\right) = \log\left(odds\right)$$

Lembrando que $\text{Log}_a b = x \Leftrightarrow a^x = b$

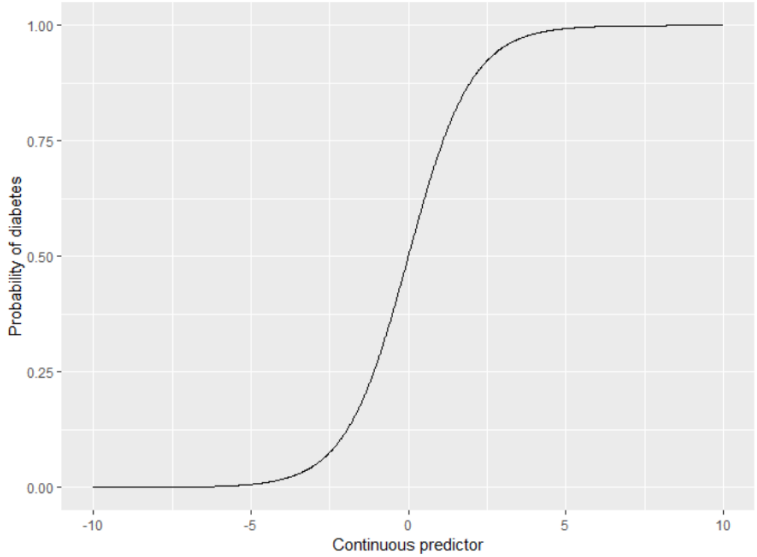
Base ← Logaritmando ← Logaritmo





Regressão Logística Simples

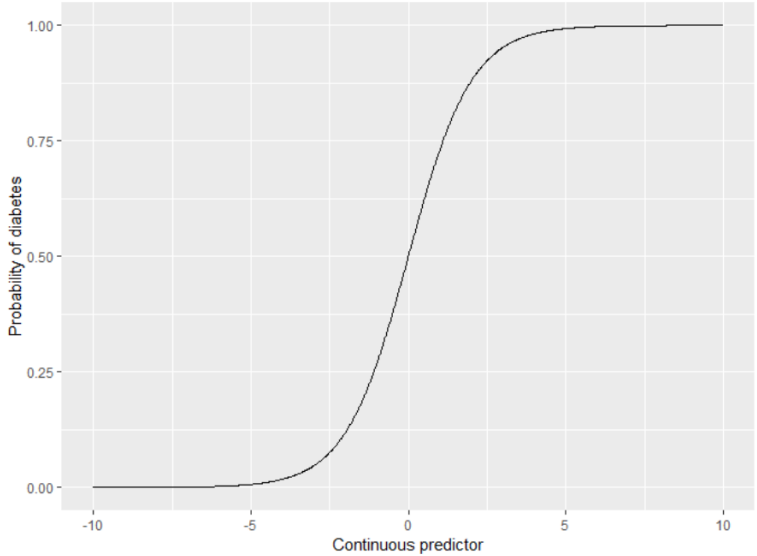
- *Odds* e probabilidade só assumem valores positivos.
- *Log(odds):* qualquer valor de menos infinito (p = 0) a infinito positivo (p = 1).
- Permite executar um modelo de regressão de maneira semelhante à regressão linear e ainda garantir que os valores previstos para as probabilidades estejam entre 0 e 1.
- Para retornar a p, execute a transformação 'anti-log' (*exponenciação*).



$$\text{logit}\,(p) = \log\left(\frac{p}{1-p}\right) = \log\Big(odds\Big)$$

Regressão Logística Simples

- Como as chances de desenvolver diabetes variam de acordo com alguns fatores relacionados ao paciente?
- ***Log odds ratio***: $\frac{\log(odds\ diabetes\ acima\ 60\ anos)}{\log(odds\ diabetes\ baixo\ 60\ anos)}$
- ***Exponenciação***: 'antilog' (odds) = exp(log(odds)) = odds.
- ***Odds ratio***: $\frac{odds\ diabetes\ acima\ 60\ anos}{odds\ diabetes\ baixo\ 60\ anos} > 1$



$$\text{logit}\,(p) = \log\left(\frac{p}{1-p}\right) = \log\left(odds\right)$$

Iniciando com... EDA

- Que tipos de variáveis e conjuntos de valores cada uma possui? Há algo de estranho?
  - **Variáveis categóricas:** quantos valores diferentes possuem? Quão comum é cada um? Categorias incomuns podem causar problemas.
  - **Variáveis contínuas:** qual distribuição seguem? Como resumir essa distribuição? Se for normal, relate a média e desvio padrão. Se estiver distorcido ou enviesado, melhor relatar a mediana e intervalo interquartil.

- Como cada variável se relaciona com a variável *resultado*? Faça tabulações cruzadas.
  - Como tabular variáveis contínuas contra uma variável binária?
  - Agrupar valores de variáveis contínuas: proporções.
  - ***Combinar valores sempre implica em perda de informação!***

Regressão Logística em R

Regressão Logística Simples

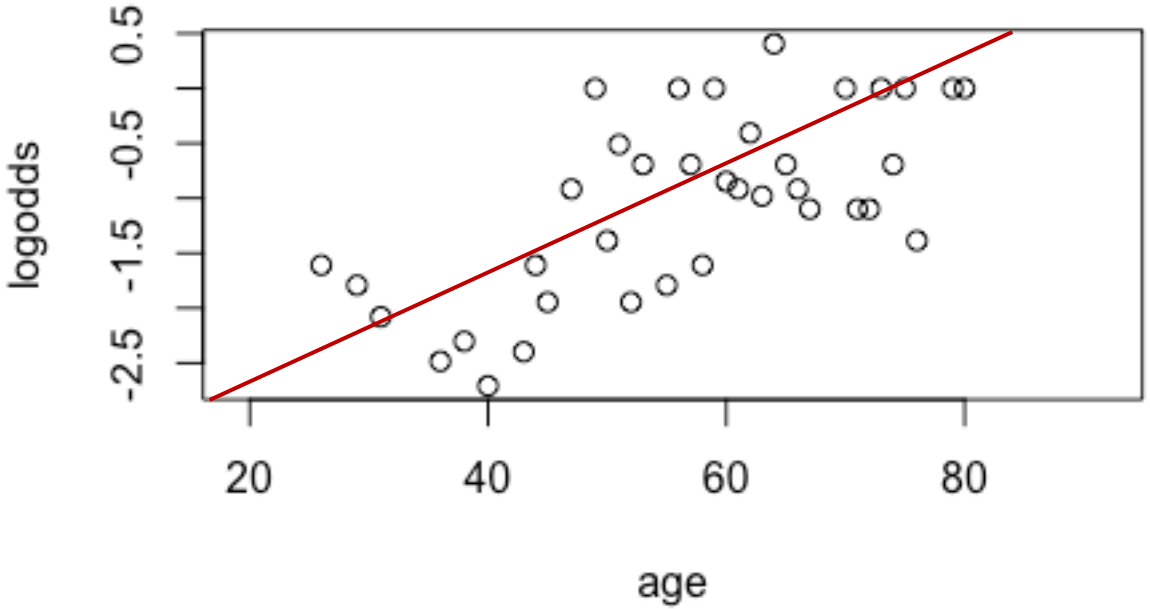
```
> glm(resultado ~ preditor, family=binomial (link=logit))
```

- Família de distribuição: binomial.
- Função link (logit): transforma a variável *resultado*, de probabilidade de ocorrência para chances, em escala log.
- Preditor categórico: não precisa números iguais de observações em cada categoria, porém categorias estreitas podem enviesar o resultado.
- Preditor contínuo: não precisa ser distribuído normalmente, mas **deve-se assumir que a relação entre a variável e o resultado é linear!**

Regressão Logística Simples - premissas

```
> glm(dm ~ age, family=binomial (link=logit))
```

A relação entre o preditor e o resultado em log(odds) é linear!



Interpretando um Modelo Nulo

- **Null model**: qual é a chance de desenvolver diabetes?
- Premissa: todos têm chances iguais.
- 13 observações deletadas.
- ***Intercept***: log odds diabetes: **-1,7047**
- Odds = exp(-1.7047) = **0.182**
- Probabilidade = 0.182/1.182 = **0.15** = **15%**

```
> m <- glm(dm ~ 1, family=binomial (link=logit))
> summary(m)

Call:
glm(formula = dm ~ 1, family = binomial(link = logit))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.578  -0.578  -0.578  -0.578   1.935

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7047     0.1403  -12.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.87  on 389  degrees of freedom
Residual deviance: 334.87  on 389  degrees of freedom
  (13 observations deleted due to missingness)
AIC: 336.87

Number of Fisher Scoring iterations: 3
```

Interpretando um Modelo Nulo

- **Null model**: qual é a chance de desenvolver diabetes?
- Premissa: todos têm chances iguais.
- 13 observações deletadas.
- ***Intercept***: log odds diabetes: **-1,7047**
- Odds = exp(-1.7047) = **0.182**
- Probabilidade = 0.182/1.182 = **0.15** = **15%**

```
> table(dm)
dm
 no yes
330  60
```

Odds
60/330 = 0.182

Probabiliade
60/(330+60) = 0.15

```
> m <- glm(dm ~ 1, family=binomial (link=logit))
> summary(m)

Call:
glm(formula = dm ~ 1, family = binomial(link = logit))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.578  -0.578  -0.578  -0.578   1.935

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7047     0.1403  -12.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.87  on 389  degrees of freedom
Residual deviance: 334.87  on 389  degrees of freedom
  (13 observations deleted due to missingness)
AIC: 336.87

Number of Fisher Scoring iterations: 3
```

Regressão Logística - age

- Log odds de diabetes = -4.4045 + 0.0525 * age (anos)
- p-valor (age): 2.29e-08
- ***Odds ratio*** = exp(0.052) = **1.05**
- Aumento de 5% na **chance** de ter diabetes para cada ano vivido.
- ***Akaike's Information Criterion (AIC)*** descreve quão bem o modelo se ajusta aos dados enquanto penaliza modelos com muitos coeficientes. Valores mais baixos de AIC são desejáveis.

```
> m <- glm(dm ~ age, family=binomial(link=logit))
> summary(m)

Call:
glm(formula = dm ~ age, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3612  -0.5963  -0.4199  -0.3056   2.4848

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.404530   0.542828  -8.114 4.90e-16 ***
age          0.052465   0.009388   5.589 2.29e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.87  on 389  degrees of freedom
Residual deviance: 299.41  on 388  degrees of freedom
  (13 observations deleted due to missingness)
AIC: 303.41

Number of Fisher Scoring iterations: 5
```

Ajuste do modelo - Poder Preditivo ou Explicativo

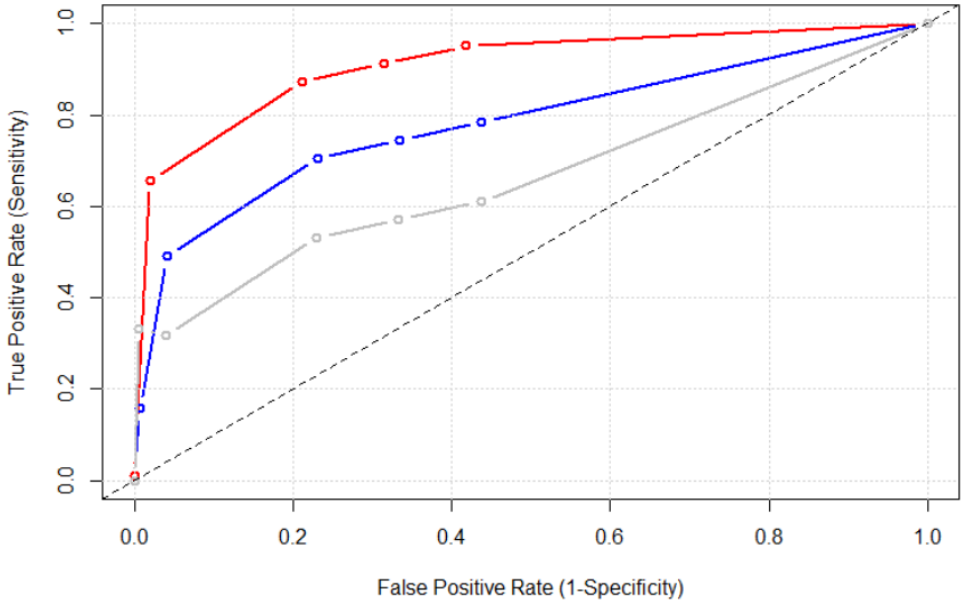
- **R^2:** proporção da variância do resultado que pode ser explicada pelo modelo. Entretanto, não podemos fazer testes de correlação para variáveis binárias.
- **Macfadden (pseudo-R2):** similar ao R2 mas com valores menores que na reg linear.

- **Discriminação: área sob a curva ROC ou estatística C**
  - Mede quão bem o modelo identifica quem tem ou não o resultado de interesse (verdadeiros positivos).
  - Muito utilizada também em *machine learning* quando o objetivo é a previsão.

Ajuste do modelo - Poder Preditivo ou Explicativo

Curva ROC

- ***Sensibilidade***: probabilidade de um resultado positivo entre os casos.
- ***1 – especificidade***: probabilidade de um resultado negativo entre os não casos.
- ***Caso***: alguém com a doença ou desfecho de interesse.
- Taxa de verdadeiro positivo *versus* a taxa de falso positivo.



Ajuste do modelo - Testes de Aderência (GOF)

Estatística de Desvio (*Deviance*)

- Mede quão bem o modelo se ajusta aos dados, ou quanto os dados preditos se afastam dos observados.
- Regressão logística: resultado observado (probabilidade) assume valores 0 ou 1, e o valor previsto (log odds), qualquer valor.
- Difícil calcular este ’desvio’ como em um modelo de regressão linear.

OBS: *A qualidade do ajuste não diz nada sobre o poder preditivo e vice-versa.*
É possível obter uma boa previsão com ajuste ruim ou um modelo com bom ajuste, mas previsão ruim!

Ajuste do modelo - Testes de Aderência (GOF)

- **Abordagem**: dados podem ser agregados ou agrupados em ‘perfis’ exclusivos, i.e., pacientes com a mesma idade, sexo e gênero, etc. Depois obtém o número de eventos observados e esperado para cada perfil.
- **Estatísticas *deviance*** e ***qui-quadrado de Pearson***: ambas podem ser comparadas com tabelas de distribuição qui-quadrado para ver o quão incomum é o valor da estatística para aquele modelo, produzindo o p-valor.
- **p-valor > 0,05**: o modelo se ajusta bem aos dados.
- Compara o modelo proposto a um ‘modelo saturado’ que explica totalmente os resultados, para avaliar interações entre variáveis ou não-linearidades.

Regressão Logística Múltipla

```
> glm(resultado ~ pred1 + pred2 + ... + predN,
family=binomial (link=logit))
```

- Seleção de preditores: se tiver poucos (~ 5), incluir todos!
- Ajuste do modelo: avaliar como o modelo se ajusta aos dados, ou seja, quanto ele consegue explicar.
- O objetivo de qualquer modelo é aproximar a realidade de *maneiras úteis*.
- ***Precisa ser bom o suficiente!***

"All models are wrong buta some models are usefull"

George Box (1919-2013)

Desenvolvendo um modelo de Regressão (Logística)

- Entenda a pergunta e o objetivo do modelo.
- Pesquise candidatos preditores em revistas científicas ou com especialistas da área.
- Selecione os preditores tendo em mente seu objetivo.
    - **1 preditor**: 2 parâmeros (α e ß), ↓ std error e IC 95%, ↓ R2 , mas o modelo é robusto.
    - **100 preditores**: 101 parâmetros (alguns significativos, outros não), ↑ std error e IC 95%, *odds ratios* imprecisos. R2 ↑ porém o modelo é pouco robusto, ***overfitted***.
- Execute a seleção ***Backward Elimination*** para o ajustar o modelo.
- Perigos: ***overfitting* e *não-conversão*!**

Overfitting é uma grande armadilha da modelagem preditiva e acontece quando você tenta incluir muitos preditores ou categorias em seu modelo.

Pode levar a grandes erros-padrão, razões de probabilidade absurdas e à falha de convergência do modelo.

Não-conversão significa que o algoritmo que está tentando estimar todos os seus *odds ratio* não consegue encontrar a melhor solução.

Desenvolvendo um modelo de Regressão (Logística)

- **Avalie o erro-padrão**: quanto < n amostral, > std.error. Erros acima de 10 devem ser cuidadosamente investigados.
    - *Regra prática: coef > 2x std.error, geralmente significativo.*
- Agrupe os níveis de variáveis categóricas, se possível.
- Mude a categoria de referência, caso esta seja muito estreita.
- Crie novas variáveis a partir das existentes.
- Elimine variáveis colineares.
- Elimine sumariamente variáveis com erro-padrão grande (*último recurso!).*

Referências

- Coursera ”*Introduction to Statistics & Data Analysis in Public Health*”
- Dancey, CP; Reidy JG; Rowe, R. *Estatística Sem Matemática para as Ciências da Saúde*. Porto Alegre: Penso, 2017.
- Wheelan, Charles. *Estatística: o que é, para que serve, como funciona*. Rio de Janeiro: Zahar, 2016.
- Vieira, S. *Introdução à Bioestatística*. 4ª ed. Rio de Janeiro: Elsevier, 2008.