Modelação Ecológica

CLASS 6

02 October 2019 – 14:00-16:30 – room 2.4.37

Tiago A. Marques

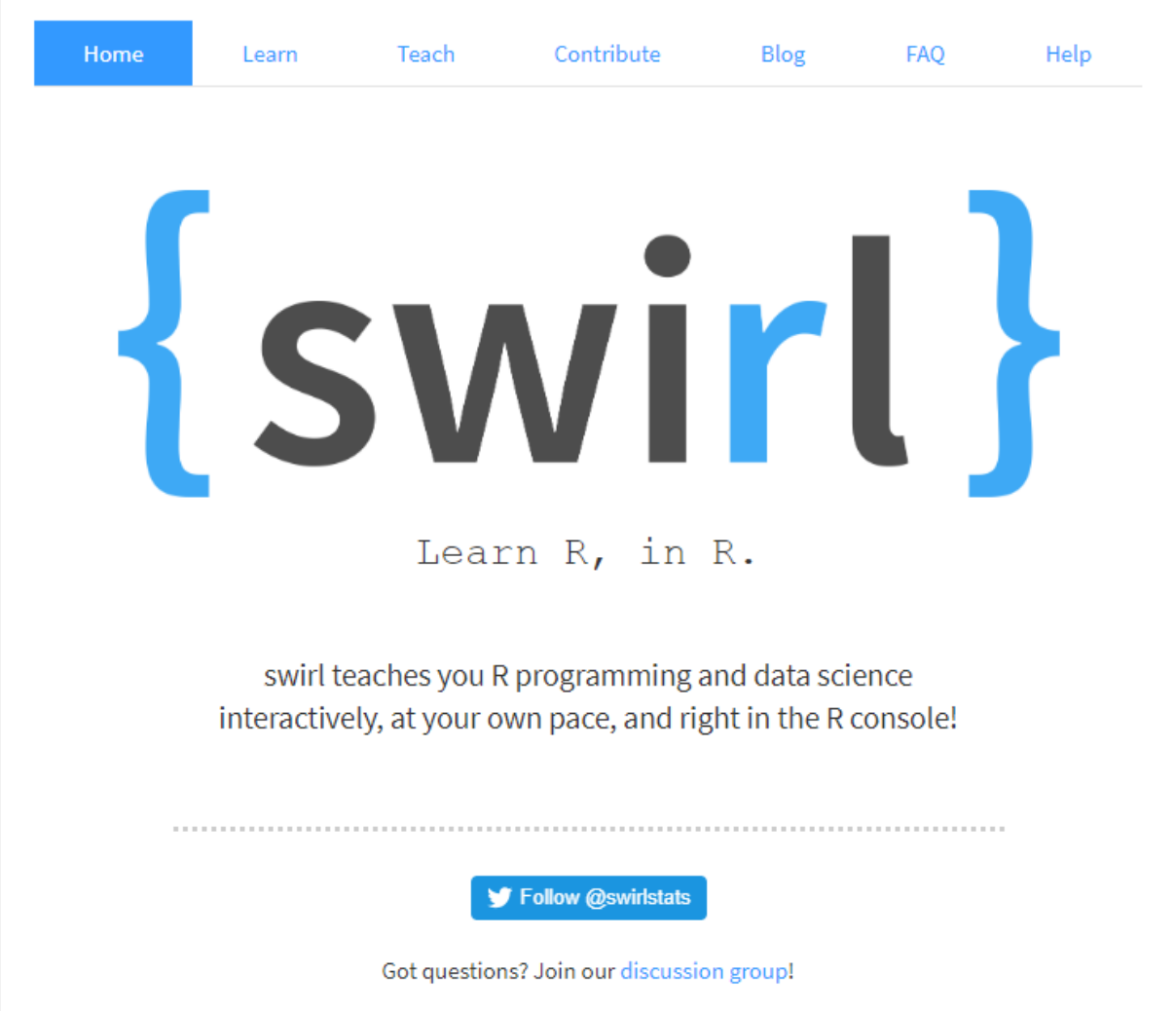
HEADS UP - TIMETABLE

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
8:00			EN TP				
8:30							
9:00							
9:30							
10:00		EN T			EN TP		
10:30							
11:00							
11:30							
12:00							
12:30							
13:00	EN T						
13:30							
14:00		ME TP 2.3.37	ME TP 2.3.37				
14:30							
15:00	EN TP						
15:30							
16:00							
16:30							
17:00							
17:30							
18:00							
18:30							

Exceptions:

- 8 October (13:30-16:00, 2.3.37)
- 9 de October, 23 October, 5 e 12 November (14:00-16:30, sala 2.3.16)
- 5 de November (14:00-17:00, 2.3.16, with a special guest, Théo Michelot, CREEM, talking about HMM's)
- 6 de November – independent work (homework from 5 November)
- 10 e 11 de December – something will happen but…I'm not there… @ WMMC in Barcelona

https://swirlstats.com/




```
install.packages("swirl")
library("swirl")
swirl()
```

STATISTICAL MODELLING

Setting the scene

Statistical models try to solve two common unknowns that might be required for using other mechanistic/system based (see Evans 2011) approaches/models

- The model and the relevant variables are unknown;
- The influence of each relevant variable and the interactions between them are unknown

Setting the scene

Known model

Known parameters

Simulation
Computation

Generation of responses

(fitted values)

Unknown Model

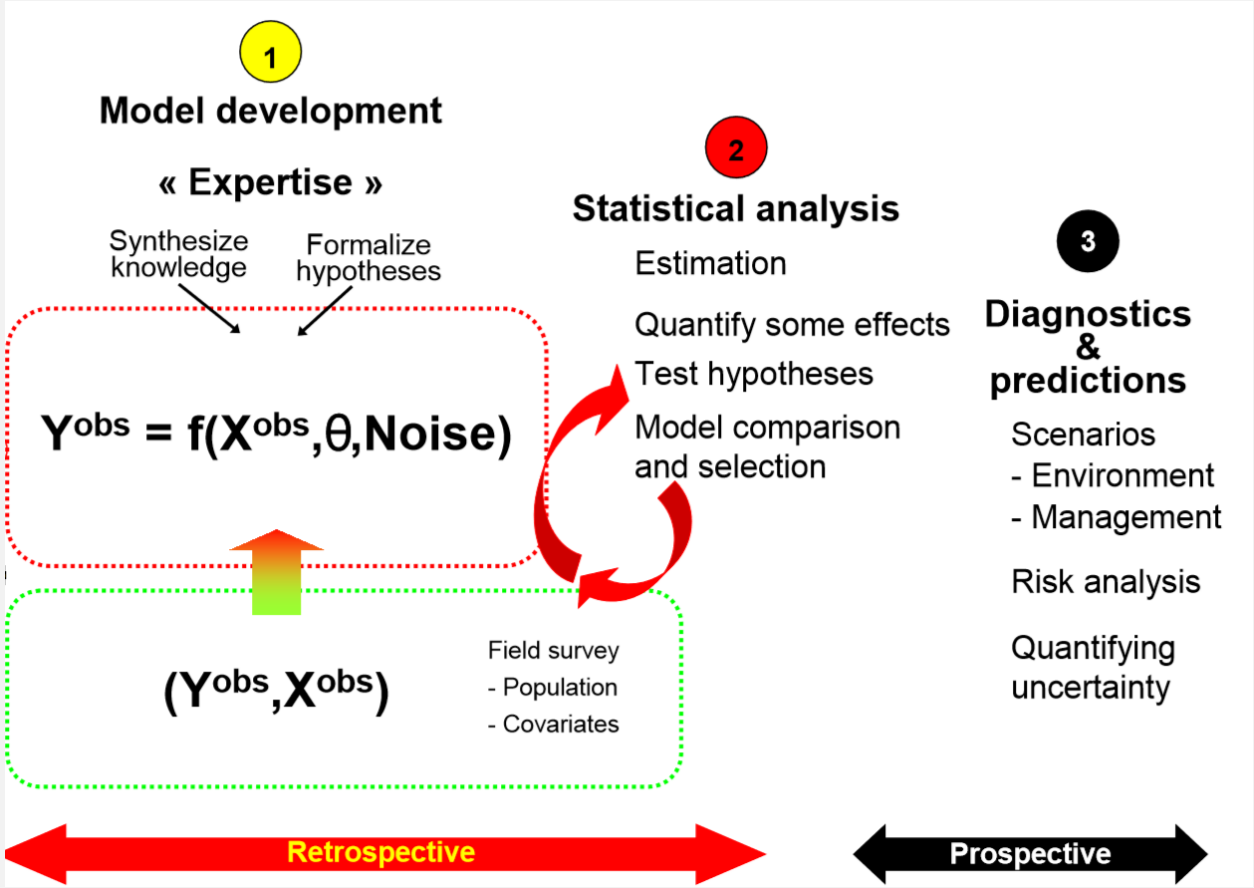
Unknown parameters

Exploratory analysis
Statistical Analysis
Parameter estimation
Calibration

Observed responses

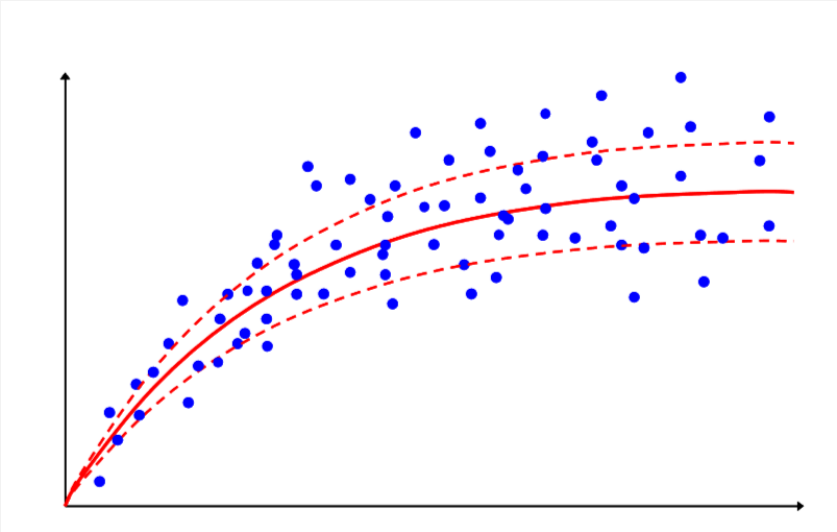
(data: observed values)

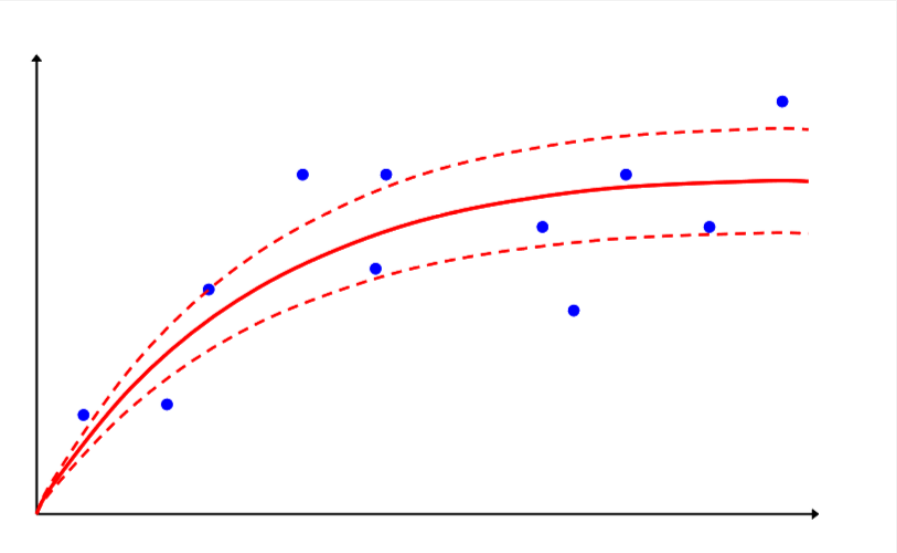
Setting the scene



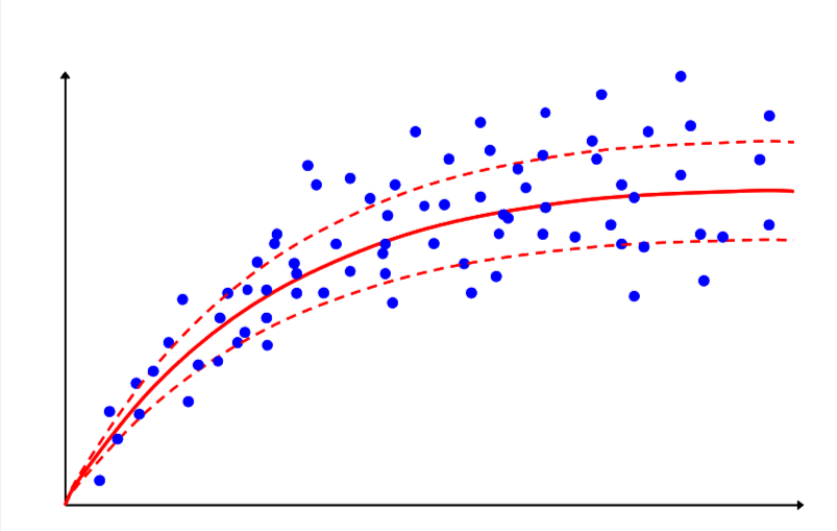
Setting the scene

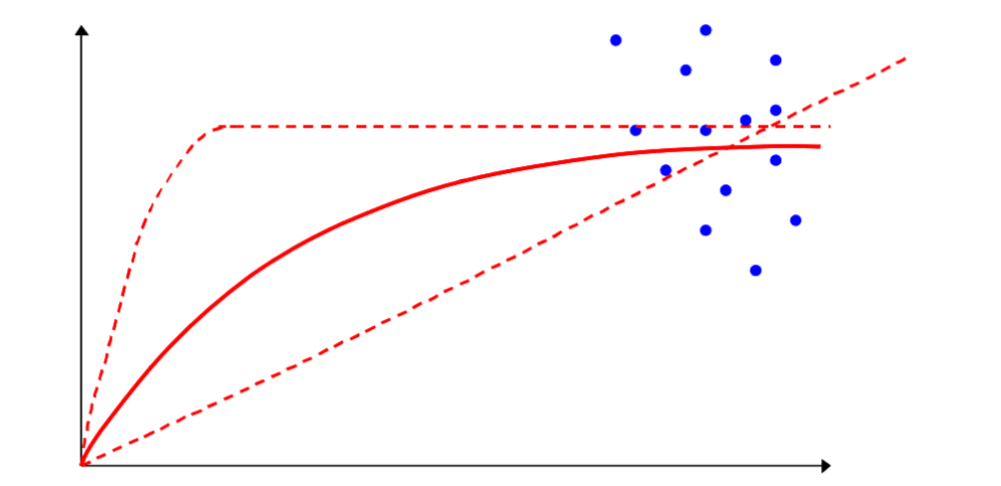
- Statistical models are extremely dependent of initial observations (quantity and quality of data)





Setting the scene





Setting the scene

The main goal of a statistical model is to explain the components of variability associated with a certain phenomena (response variable), quantifying

- The component that is due to systematic effects versus random effects/components
- The relative important of the explanatory variables

Setting the scene

Developing a statistical model will typically involve:

- Choosing a suitable model (framework, distribution family, link function, etc.)
- Evaluating ("testing") which variables might be relevant
- Model selection (choose across the potential variables that might be relevant)
- Parameter Estimation
- Model assessment
- Predictions

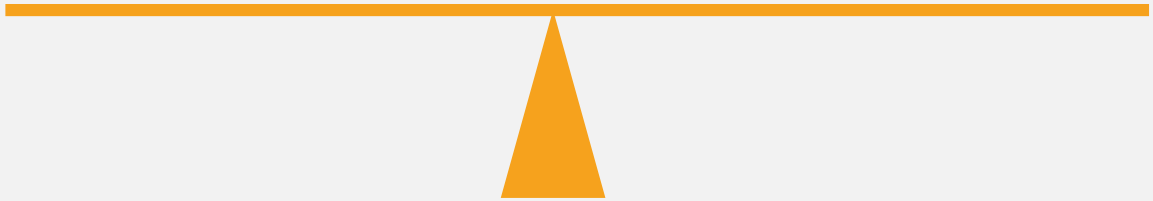
Setting the scene

We should consider the principle of parsimony:

Balance: find the sweet spot

Simplify the problema
and the methodological
apporach

Represent the reality of a
given system complexity



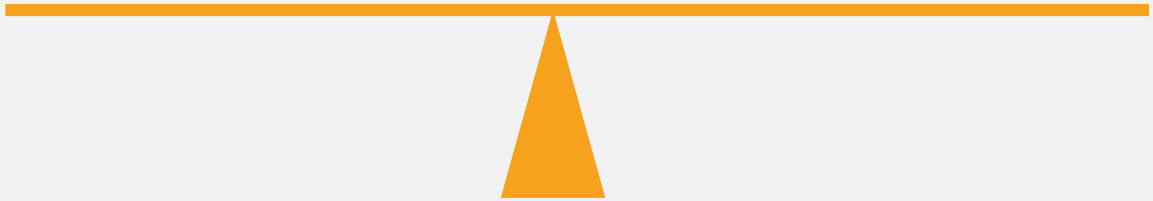
Setting the scene

We should consider the principle of parsimony:

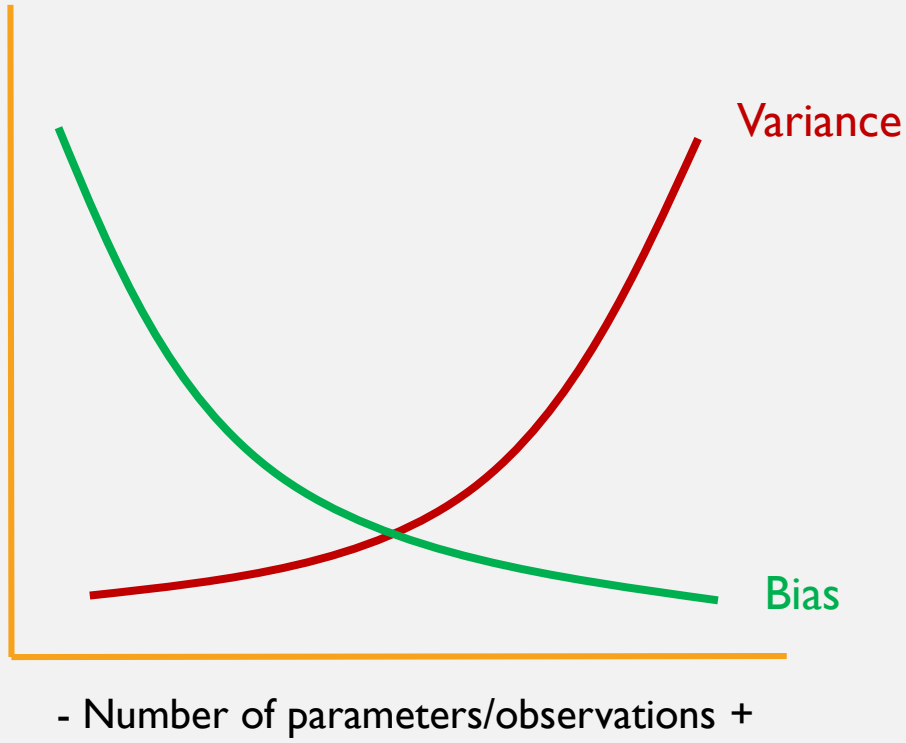
Balance: find the sweet spot

Bias: parameter estimation and predictions

Variance: parameter estimation and predictions

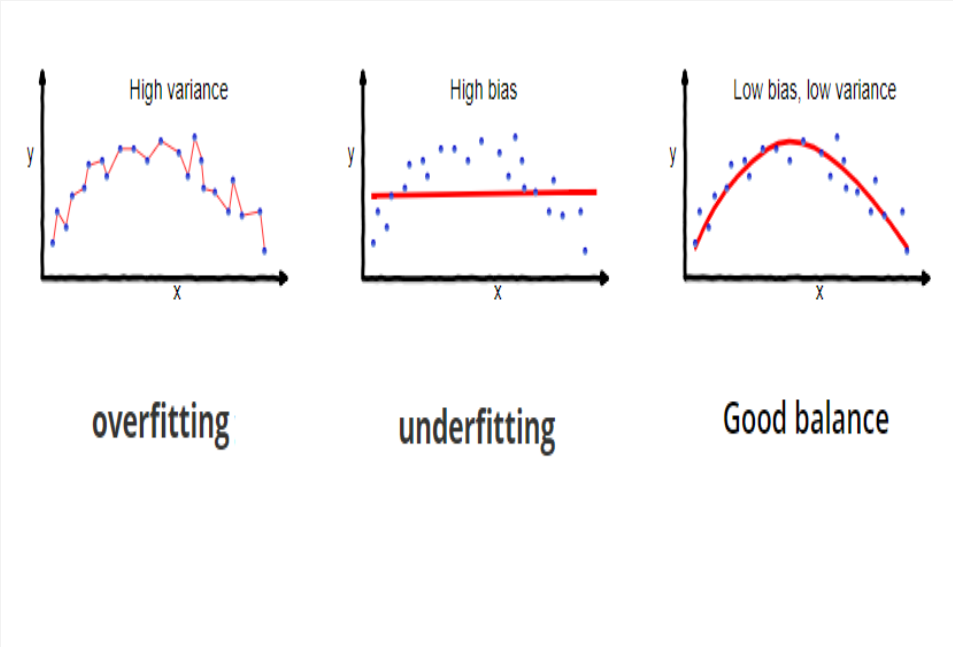


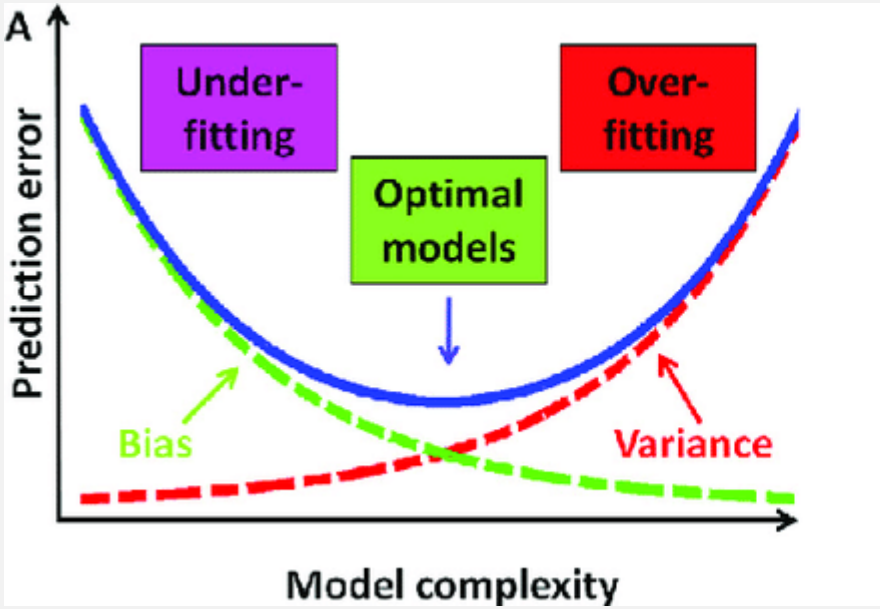
Setting the scene

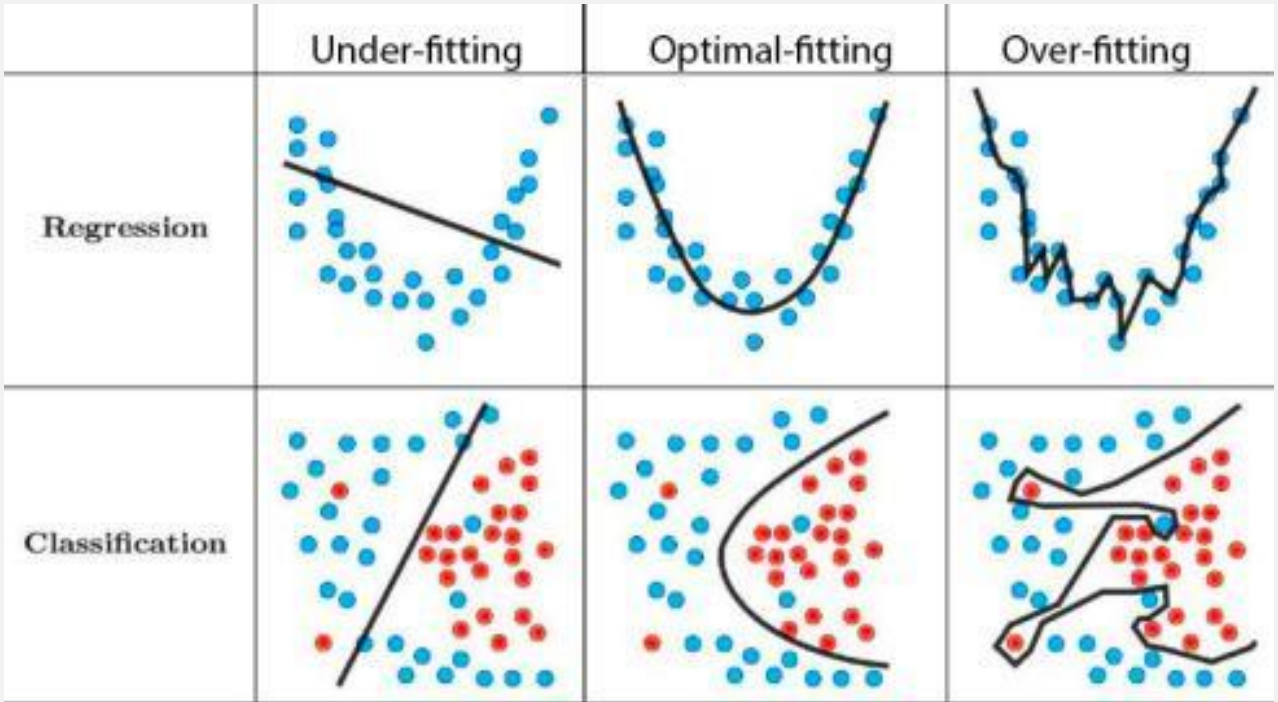


Variance and bias are determined by (even if only conceptually!) repeating the process many times

https://elitedatascience.com/bias-variance-tradeoff - with a machine learning spin


High variance
y
x
overfitting
High bias
y
x
underfitting
Low bias, low variance
y
x
Good balance

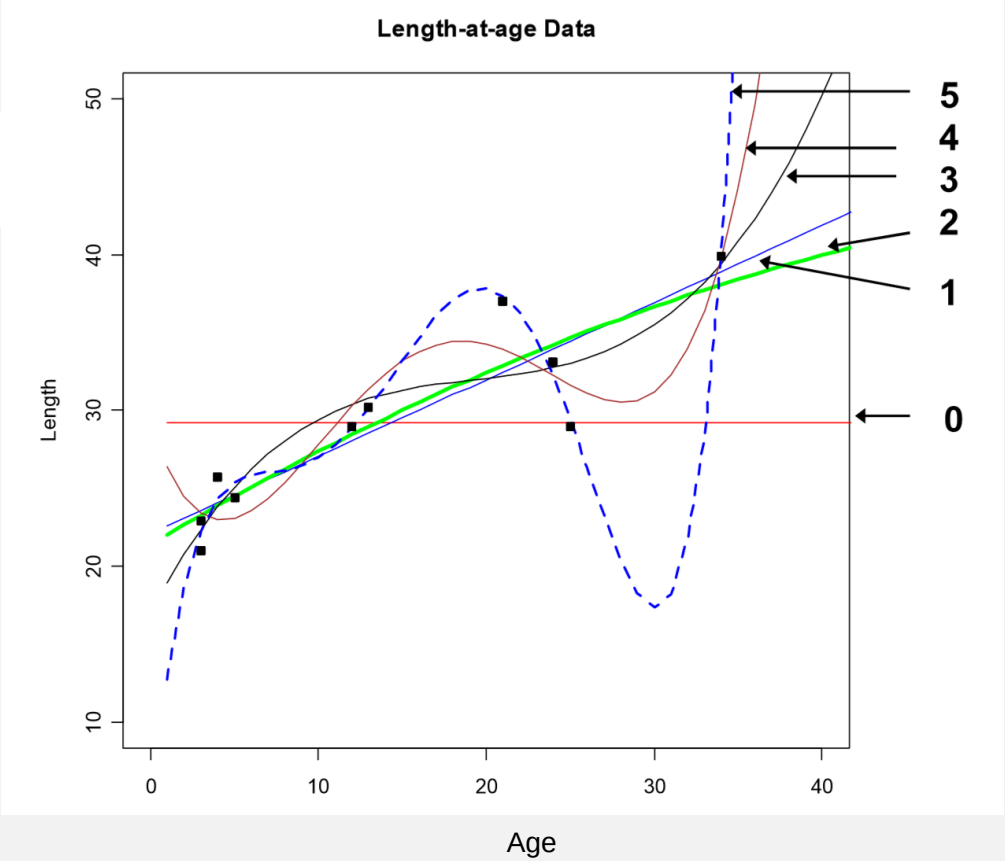

A
Prediction error
Under-fitting
Optimal models
Over-fitting
Bias
Variance
Model complexity


Under-fitting
Optimal-fitting
Over-fitting
Regression
Classification

Setting the scene

$$Length = \alpha_0 + \alpha_1 \cdot Age + \alpha_2 \cdot Age^2 + \ldots$$

Nº P	R^2
0	0.00
1	0.80
2	0.81
3	0.85
4	0.89
5	0.98



Conceptual aspects about statistical modelling

- A key challenge is therefore finding the right balance between model fit and simplicity
- This can be done typically within two frameworks
  - Hypothesis tests
  - Model Selection Criteria

Conceptual aspects about statistical modelling

- One of the key tools for the development of statistcal models are (here will be our workhorse)

Modelos lineares generalizados (MLG) ou
Generalized linear models (GLM)

Just one among several:

GAM – Generalized Additive Models
BRT – Boosted Regression Trees
CART – Classification and Regression Trees

Aspectos conceptuais da modelação de base estatística

A principal diferença entre modelos de regressão mais simples (LM) e os modelos lineares generalizados (GLM) é que estes últimos permitem contemplar distribuições não normais dos erros e, assim, ajustarem-se melhor a uma maior variedade de dados (e.g. contagens, proporções, presença/ausência).

Permitem ainda que a relação possa não ser linear na escala da resposta (ou seja, que a relação seja apenas linear na escala da função de ligação).

Aspectos conceptuais da modelação de base estatística

Estes modelos têm 2 tipos de variáveis:

- Variável resposta, variável dependente ou variável de interesse – como o nome sugere é a variável que se pretende modelar no sentido de obtenção de previsões;

- Variáveis explicativas, variáveis preditoras ou variáveis independentes – variáveis que são incluídas no modelo para obter os valores da variável resposta.

Aspectos conceptuais da modelação de base estatística

- Variável resposta – em geral, quantitativa, mas pode ser qualitativa (binária ou com poucas categorias, neste último caso analisada através de modelos multinomiais). Nos modelos mais correntes há apenas 1 variável resposta.

- Variáveis explicativas – podem ser quantitativas ou qualitativas, mas a sua natureza e número têm influência substantiva no desempenho do modelo.

Aspectos conceptuais da modelação de base estatística

Estes modelos baseiam-se, pois, num conjunto de dados do tipo:

Y_1	X_{11} (...) X_{1p}
Y_2	X_{21} (...) X_{2p}
Y_3	X_{31} (...) X_{3p}
(...)	(...) (...) (...)
Y_n	X_{n1} (...) X_{np}

Aspectos conceptuais da modelação de base estatística

Na sua formulação mais simples temos:

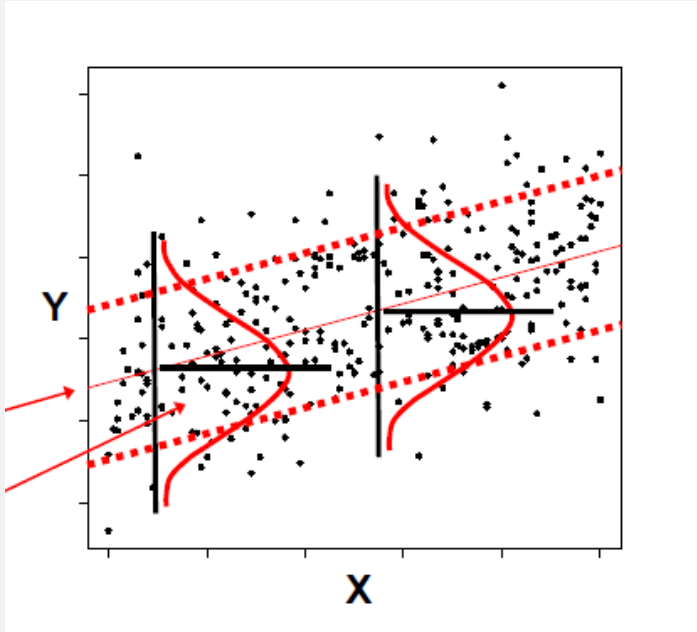
$$Y_{i,k} = \underbrace{\alpha_0 + \alpha_1 \cdot X_i}_{\text{Previsão de Y, sabendo os valores de X}} + \underbrace{\varepsilon_{i,k}}_{\text{Resíduos}}$$

Aspectos conceptuais da modelação de base estatística

Na sua formulação mais simples, temos erros com distribuição normal, média 0, homocedásticos e não correlacionados:

Resposta média

Variabilidade aleatória



Aspectos conceptuais da modelação de base estatística

As variáveis independentes não devem estar correlacionadas.

Caso estejam, esta particularidade afecta negativamente o desempenho do modelo.

Podem ser excluídas numa fase precoce da construção do modelo as variáveis muito correlacionadas (deixando-se apenas uma para cada grupo correlacionado, que expressa todas as outras – neste caso devemos reter aquela que faça mais sentido interpretar do ponto de visto ecológica).

Esta análise e selecção é frequentemente baseada em correlações simples entre pares de variáveis.

Testes de hipóteses in a nutshell

Hipótese nula ou H0: geralmente uma igualdade ou ausência de efeito
Hipótese alternativa ou H1: a negação de H0, geralmente o que queremos demonstrar

bilateral - H0: $\mu=0$ vs. H1: $\mu\neq 0$ unilateral - H0: $\mu \geq 0$ vs. $\mu > 0$ ou $\mu \leq 0$ vs. $\mu < 0$

Erro tipo I: rejeitar H0 sendo H0 verdade (nivel de significância do teste α)
Erro tipo II: não rejeitar H0 sendo H0 falsa (β)

1- β = potência do teste = rejeitar H0 sendo H0 falsa

Nível de significância, ou α:

- o risco de cometer um erro de tipo I que estamos dispostos a correr
- Definido à priori e independente dos dados
- Valores usuais: 0.01, 0.05, 0.1

Estatistica de teste – quantidade que depende dos (i.e. é uma função dos) dados e cuja distribuição, sob H0 (i.e. sendo H0 verdadeira), é conhecida

O procedimento:

Formular as hipóteses, recolher a(s) amostra(s) e calcular a estatística de teste

Comparar estatística de teste com a distribuição sob H0.

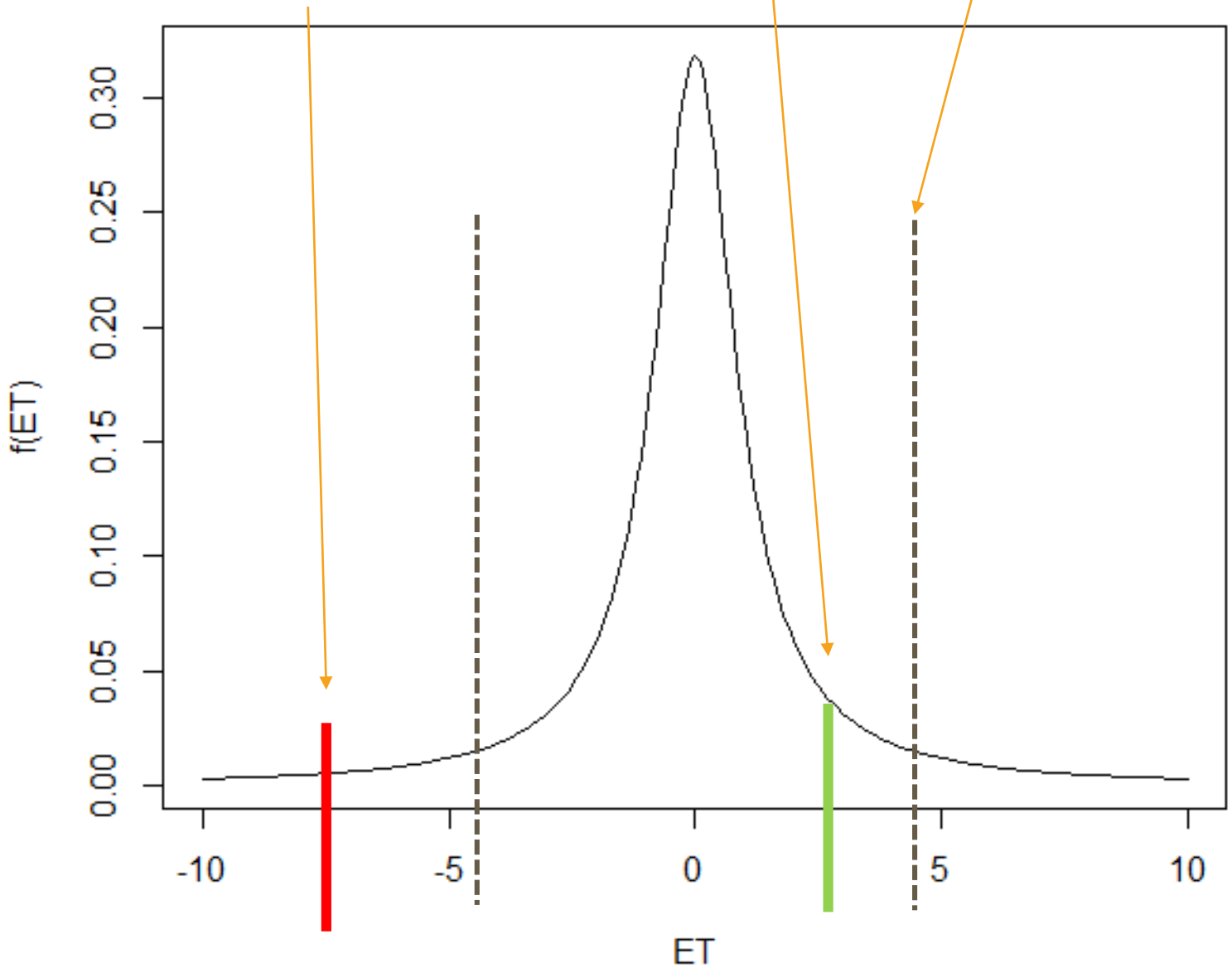
1. Se o valor for extremo, rejeitar H0
2. Se não for, não rejeitar H0

Mas como decidir se um valor é extremo?

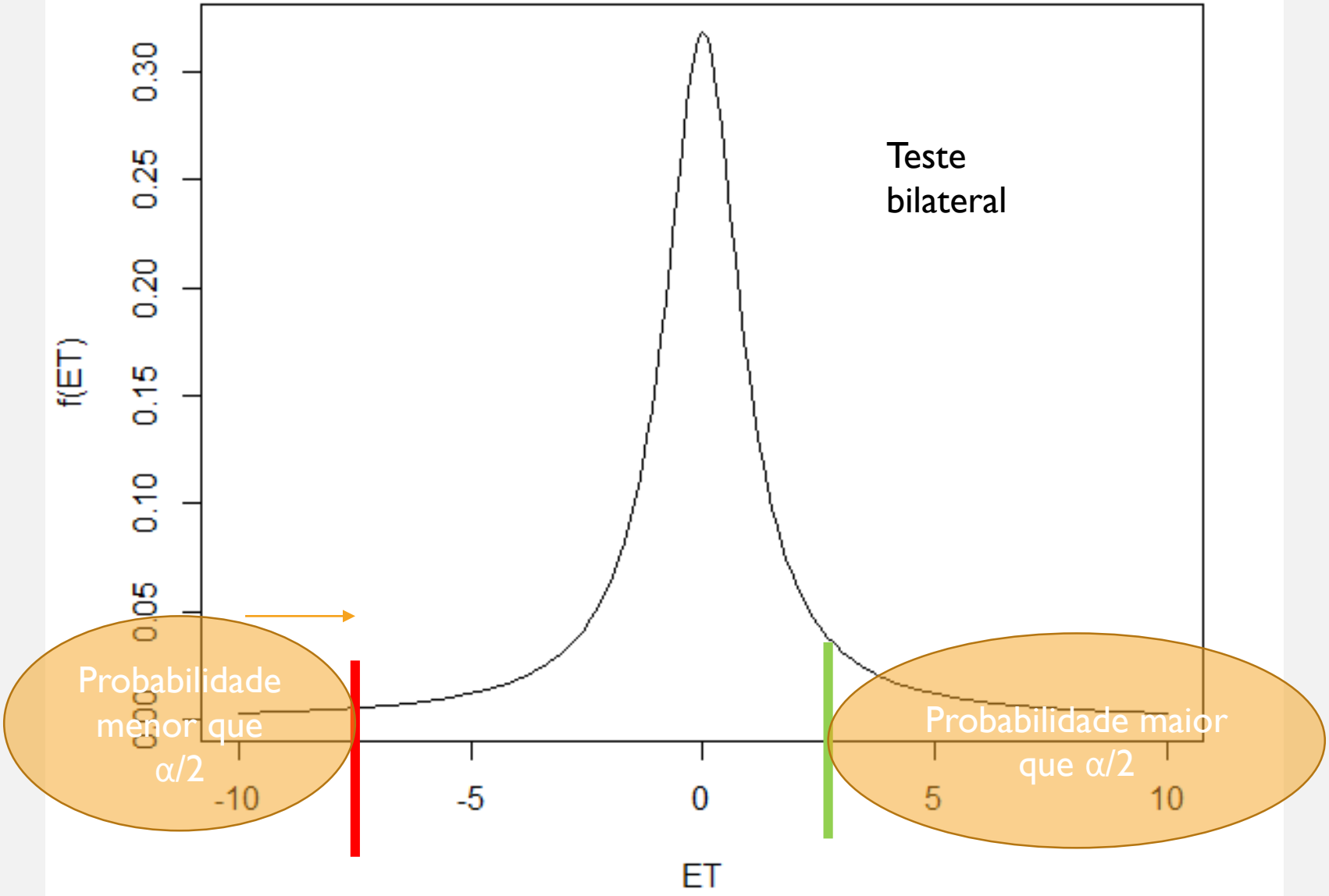
De duas formas equivalentes:

1. Usando valores críticos, que dependem do α desejado
2. Calculando o P-Value, que comparamos com o α desejado

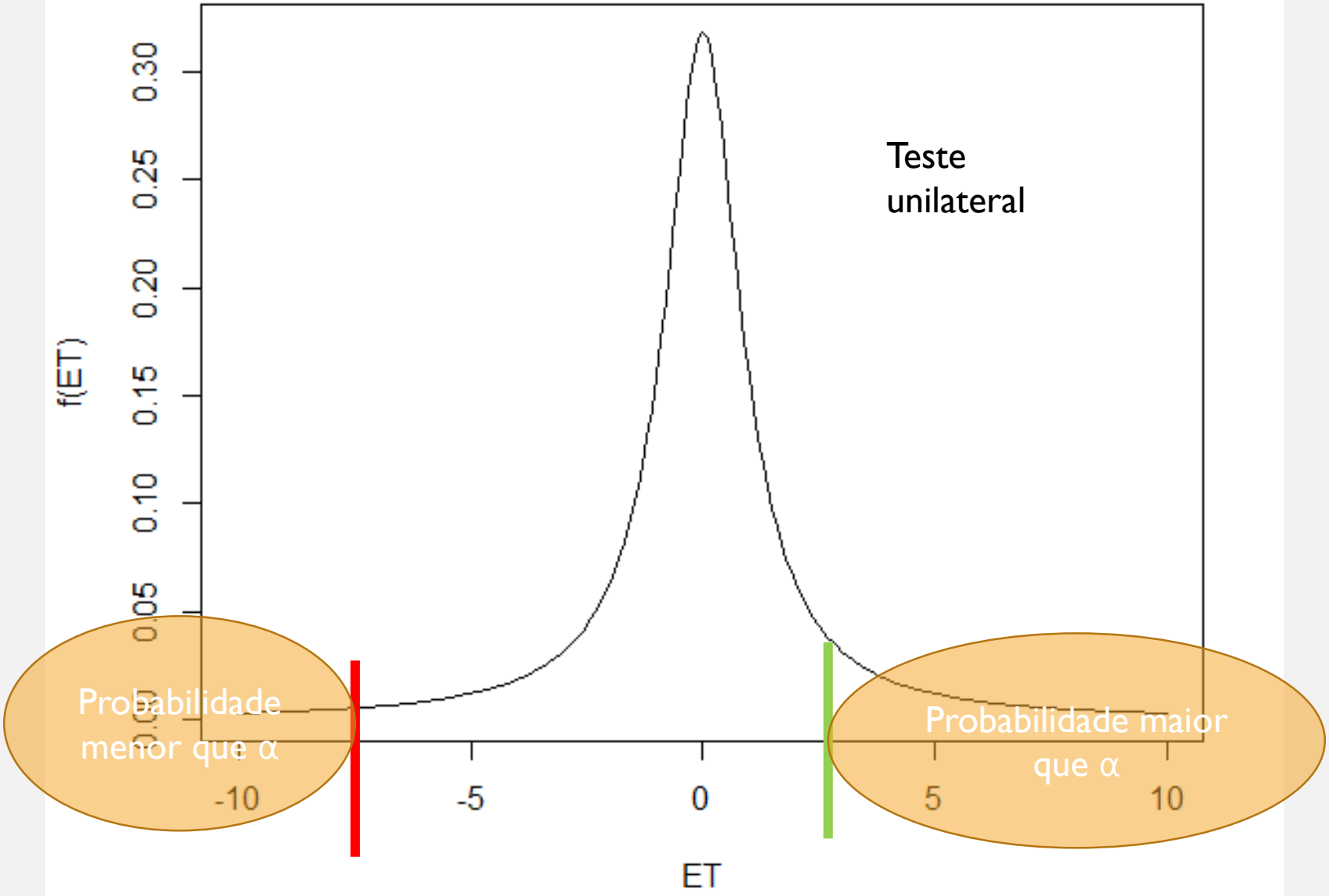
Usando valores criticos: se ET for mais extrema que o valor crítico para o nível de significancia α, rejeitar H0, caso contrário, não rejeitar H0



Usando o P-value: se o P-value, a probabilidade de observar sob H0 uma estatistica de teste tão ou mais estrema que a observada, for menor que α, rejeitar H0, caso contrário, não rejeitar H0



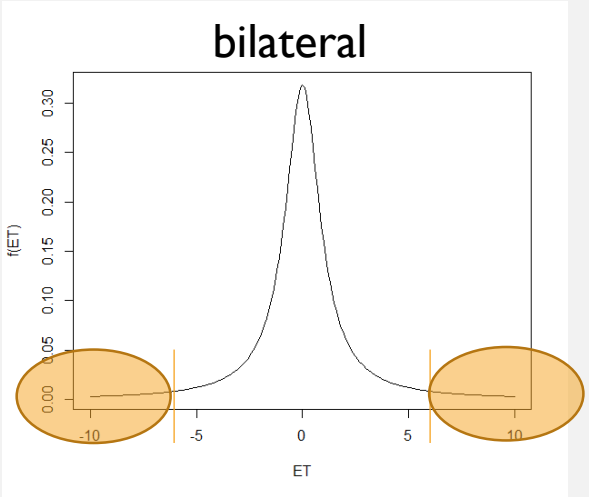
Usando o P-value: se o P-value, a probabilidade de observar sob H0 uma estatistica de teste tão ou mais estrema que a observada, for menor que α, rejeitar H0, caso contrário, não rejeitar H0

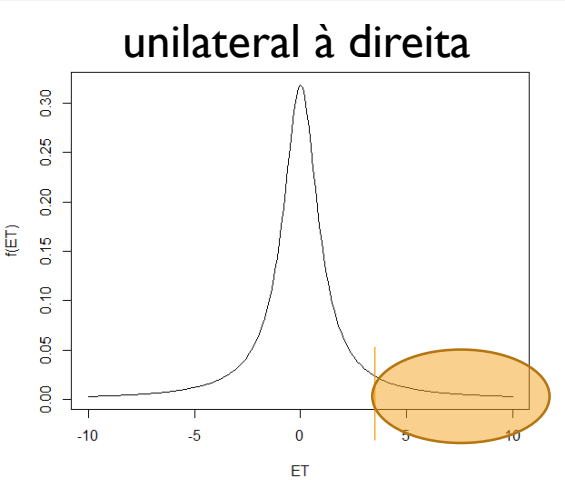


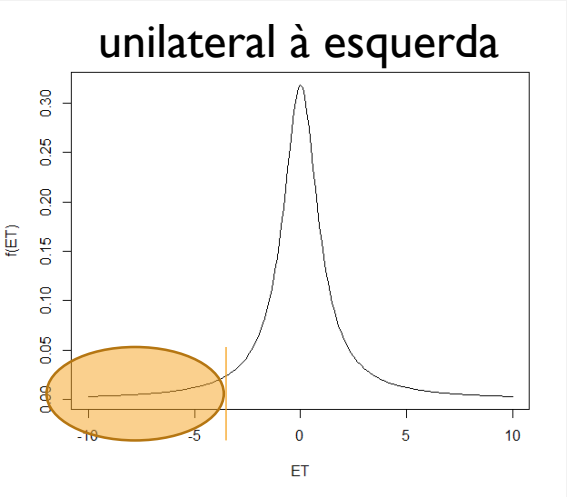
A unica diferença entre um teste unilateral e bilateral é qual a zona de rejeição, ou seja, qual o valor crítico da estatística de teste, ou, equivalentemente como se calcula o P-value

bilateral
P-value = 2*min(P(ET> valor observado),P(ET< valor observado))

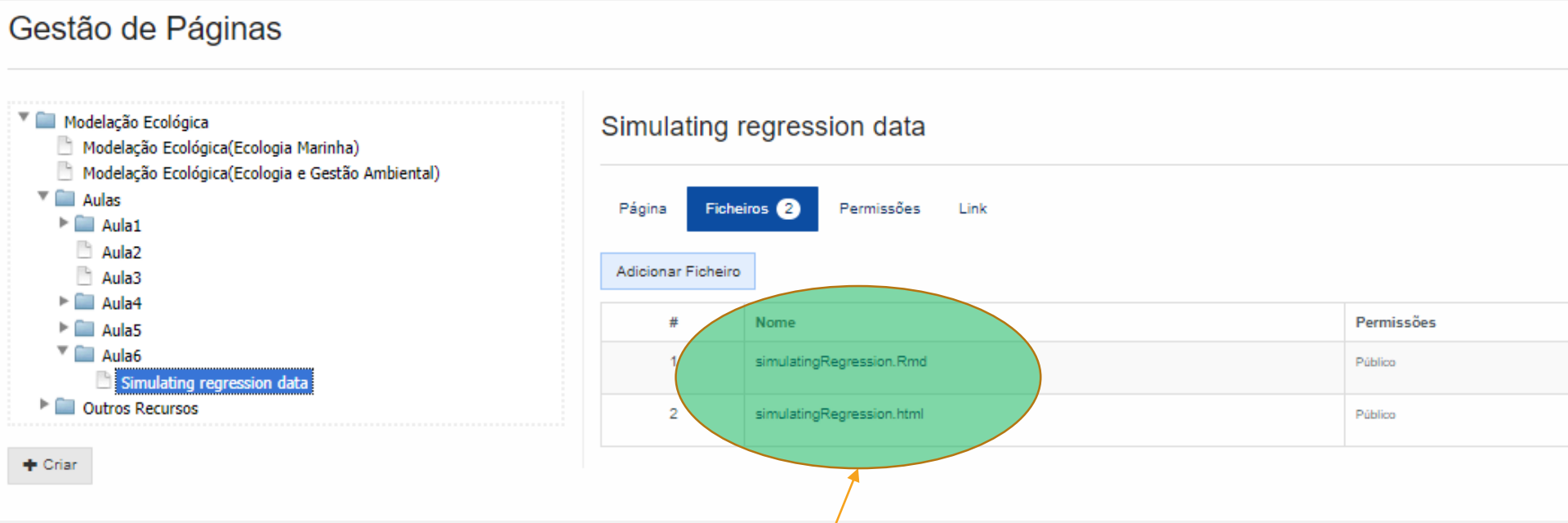
unilateral
P-value = P(ET> valor observado) num teste unilateral à direita
P-value = P(ET< valor observado) num teste unilateral à esquerda







Here he did the hands-on on the fly exercise of simulating data from a regression model and fitting regression models to data.



My attempt at it – explore it – really important concepts here!!!