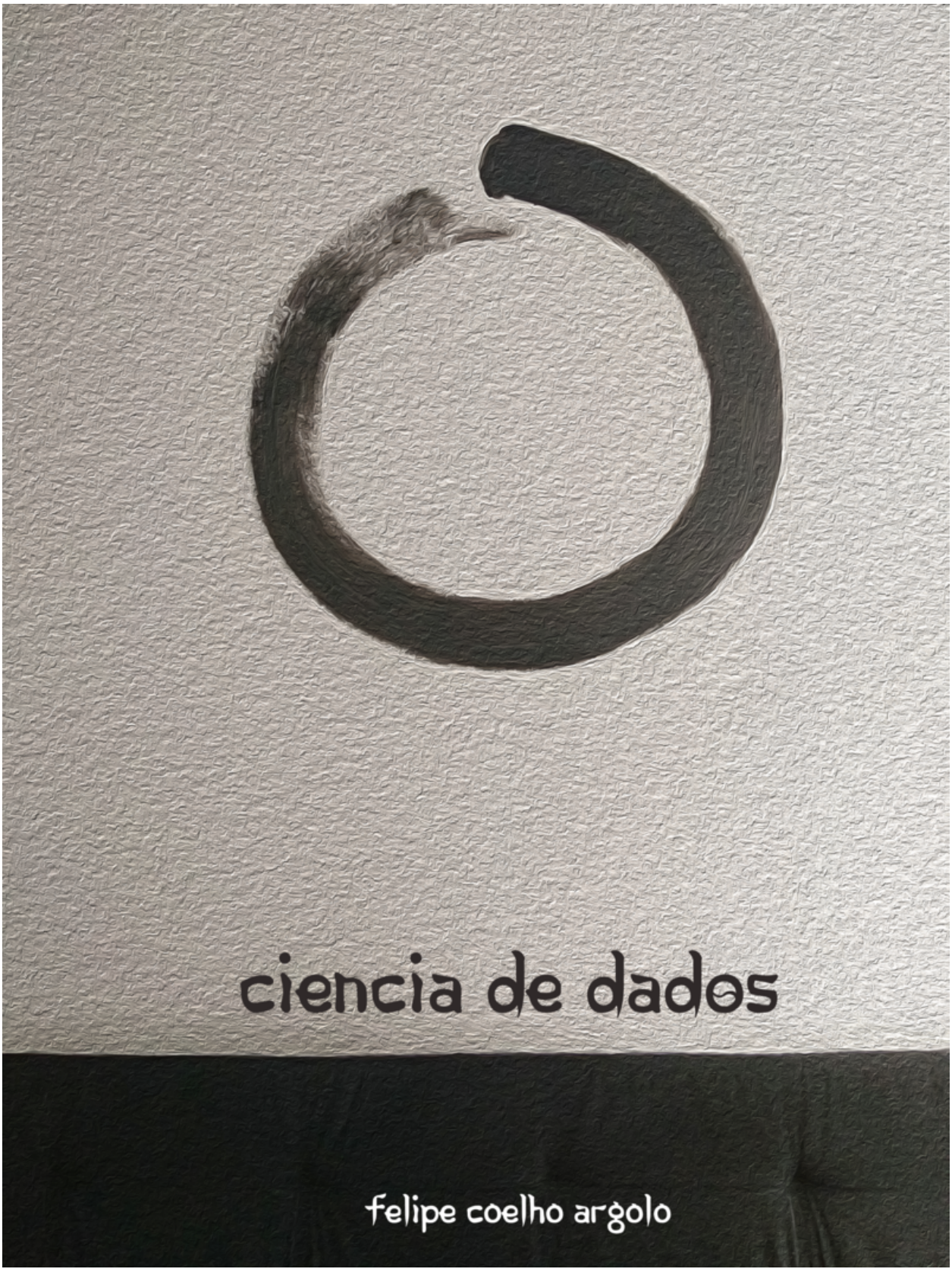

ciencia de dados
felipe coelho argolo

Ciência de dados

Curvas Funções elementares e series temporais

Felipe Coelho Argolo felipe.c.argolo@protonmail.com

São Paulo, 03 de Junho de 2019
Página oficial: https://http://www.leanpub.com/fargolo

Volume 2

Prefácio

*There is **timing** in everything. Timing in strategy cannot be mastered without a great deal of **practice**.*
Miyamoto Musashi, Go Rin No Sho (1645), Book of Earth, The Book of Five Rings

Os capítulos do primeiro volume (1-5) introduzem ferramentas comuns no repertório para análise de dados, assim como nortes em filosofia do conhecimento para aplicá-las.
De maneira geral, lidamos apenas com relações lineares. Retas cuja inclinação é determinada por um coeficiente angular: coeficiente produto-momento de Pearson (ρ), regressão linear (β), perceptrons (W) ou redes neurais (W_i).

No sexto capítulo, retomamos os níveis diários de testosterona usados quando apresentamos redes neurais. Veremos como é possível usar análises exploratórias, visulizações e intuições espaciais para escolher modelagens não lineares adequadas.

Usando funções de transformação específicas, é possível atingir acurácias muito boas. Construiremos Mark III, uma rede elegante, com topologia e funções de ativação específica para resolver o problema dos atletas sob efeito de dopping.

Em seguida, o sétimo capítulo examina outras intuições importantes para análise de séries temporais. Descreveremos análogos aos ritmos musicais, através de regressão harmônica (Transformada de Fourier em $\text{sen}(x), \cos(x)$ e semelhantes). Também, o efeito cumulativo de fases anteriores (*momentum*) com média móvel e auto-regressão (ARIMA).

O oitavo capítulo aborda volatilidade, incerteza e caos. Quantificamos amplitudes, precisão e exploramos modelos quasi-randômicos para séries temporais (*Processos Gaussianos*). Usaremos uma nova linguagem (Julia) para falar em caos e implementar expoentes de Lyapunov.

Summário

Pré-requisitos

Recomendo o volume 1, com textos introduzindo conceitos necessários: regressões, redes neurais e ferramentas como R e Stan (Caps. $0 \sim 5$).

Todos os exemplos podem ser reproduzidos usando software livre.

Leitura recomendada:

Filosofia e divulgação científica

- Surely You're Joking, Mr. Feynman
- O mundo assombrado pelos demônios - Carl Sagan
- A lógica da pesquisa científica - K. Popper
- A estrutura das revoluções científicas - Thomas Kuhn
- Contra o Método - Paul Feyerabend
- Dois dogmas do empiricismo - Willard van Quine
- Stanford Encyclopedia of Philosophy - https://plato.stanford.edu/

Neurociências

- Principles of neural science - Eric Kandel

Matemática/computação

- Coleção '*Fundamentos da matemática elementar*'
- What is mathematics - Courant & Robbins
- Better Explained (https://betterexplained.com/)
- http://material.curso-r.com/
- R Graphics Cookbook
- R Inferno
- Learn you a Haskell for Great Good
- Layered Grammar of Graphics - Hadley Wickham.
- Algorithms unlocked
- Online: Statsexchange, stackoverflow, mathexchange, cross-validated.

Machine Learning

- An Introduction to Statistical Learning: with Applications in R
- Neural Networks and Learning Machines - Simon Haykin
- Stanford (computer vision): http://cs231n.stanford.edu/
- Oxford 2015 (Deep learning): (https://www.youtube.com/watch?v=dV80NAlEins&list=PLE6Wd9FR--EfW8dtjAuPoTuPcqmOV53Fu)

Agradecimentos

Minha família, Suzana, Paulo, Isaac e Chris. Amigos Gabriel, Guilherme, Wei.

Aos professores: Carla Daltro, Anibal Neto, Lucas Quarantini, Luis Correia, Rodrigo Bressan, Ary Gadelha.

Aos colegas Fatori, Luccas, Macedo, Walter, Sato, André, Hiroshi, Lais, Luci, Davi, Oddone, Jones, n3k00n3 (Fernando), Loli (Lorena).

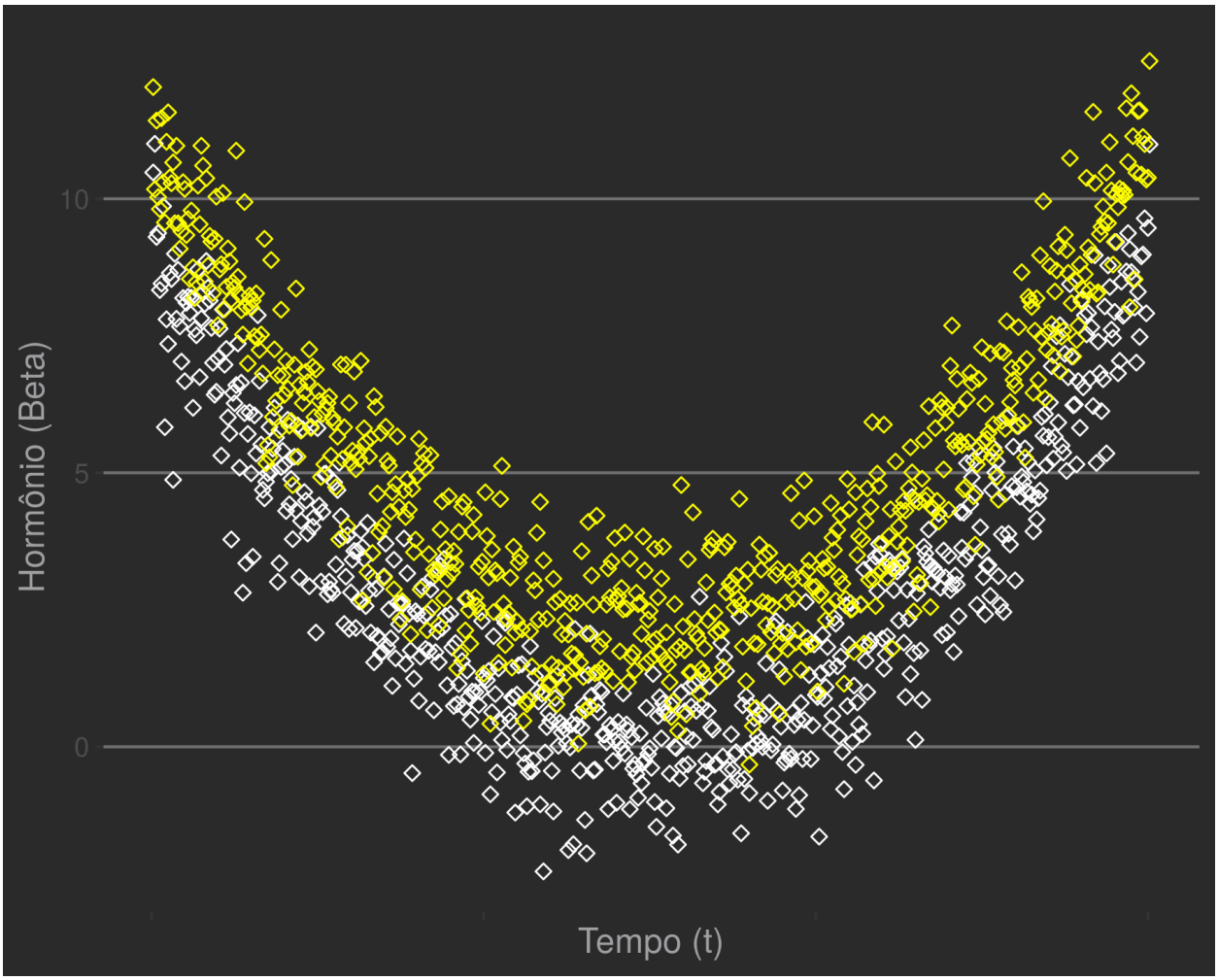
Para comentários, críticas, sugestões, ou simplesmente dizer *oi*: felipe.c.argolo@protonmail.com.

Capítulo 6 - Curvas: Funções elementares e séries temporais

Nos primeiros capítulos (1-5), aprendemos relações lineares. Retas descritas por um coeficiente angular: coeficiente produto-momento de Pearson (ρ), regressão linear (β), perceptrons (W) ou redes neurais (W_i) de ativação simples. Aprenderemos a escolher modelos um pouco mais sofisticados, naturalmente não lineares, para solução de problemas.

O caso das drogas ergogênicas

Lembre: O gráfico a seguir, retirado do *Capítulo 4*, é baseado em observações biológicas e representa milhares de amostras com: (1) a curva diária natural de testosterona (branco) e uma suposta curva correspondente sob uso de esteroides anabolizantes (amarelo).



Como as curvas não são separáveis linearmente, usamos transformações em sequência para atingir classificação satisfatória unicamente com pedaços lineares. Uma rede neural (Mark II) teve acurácia demonstrada para o problema com medida de flores (*iris*).

Com redes neurais, fazemos ajustes graduais através de funções elementares e obtemos uma boa solução final. Entretanto, a convergência é lenta e necessitamos de muitas observações ou truques na apresentação dos dados (e.g. *epochs*). Vamos avançar um pouco, criando modelos finamente ajustados para os problemas.

Em última instância, a adequação do modelo depende de uma simetria entre sua estrutura e aquela presente

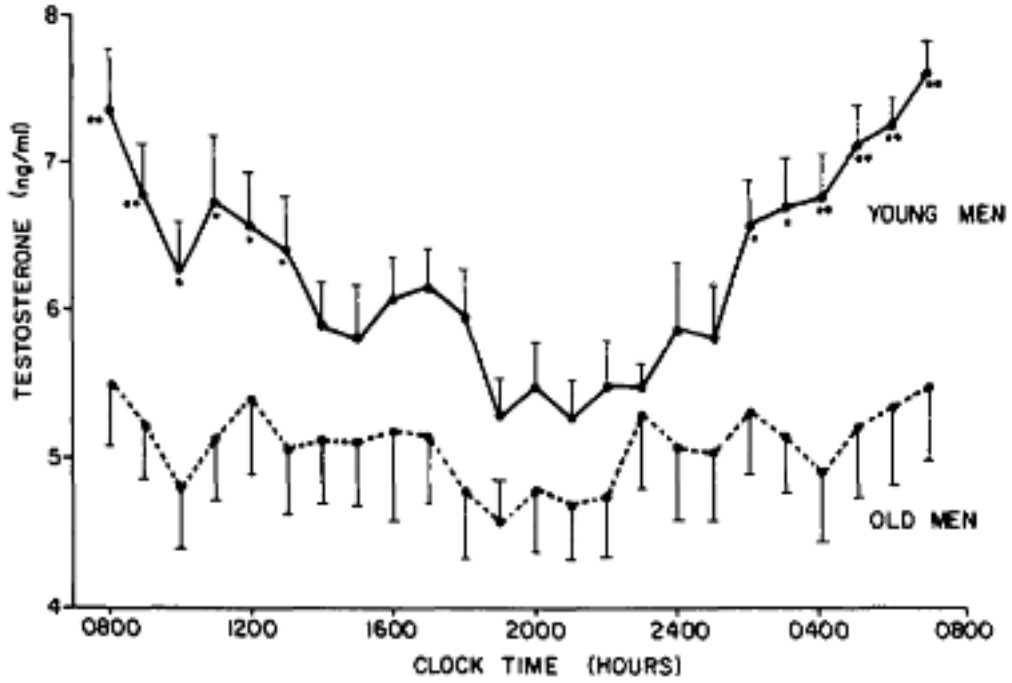


Figure 1: Ciclo circadiano para níveis de testosterona em homens jogens e idoso. Bremner, W. J., Vitiello, M. V., & Prinz, P. N. (1983). Loss of Circadian Rhythmicity in Blood Testosterone Levels with Aging in Normal Men*. The Journal of Clinical Endocrinology & Metabolism, 56(6), 1278–1281. doi:10.1210/jcem-56-6-1278

nos dados. Quão melhor a intuição sobre as características dos dados, maior será a capacidade de escolher um modelo de melhor performance e menor fragilidade a erros (e.g. *overfitting*).

Existem muitas maneiras de extrair *insights* sobre a estrutura dos dados. Por exemplo, as ideias podem vir de conhecimentos prévios sobre o fenômeno natural examinado. Abordaremos alguns métodos, com foco em visualizações e intuições espaciais sobre as medidas disponíveis.

Regressão quadrática e simetrias

Os modelos lineares expressam relações de natureza única entre variáveis. Considerando pares de observações entre duas variáveis (e.g. idade e altura), as magnitudes podem crescer com mesmo sentido (*'se A cresce, B também cresce'*, β positivo) ou em sentido inverso (*'se A cresce, B decresce'*, β negativo). Isso acontece em qualquer intervalo observado.

Alguns fenômenos se apresentam de maneira diferente. As magnitudes podem crescer com sentido igual em determinado intervalo e de maneira diferente em outro.

Observando a figura acima, notamos que níveis de testosterona *descrescem* com o tempo (sentidos inversos) a partir de 08:00 até atingirem um mímino por volta das 20:00. Passando deste ponto, passam a *crescer* com o avançar do tempo (sentidos iguais), até atingir um máximo.

Podemos modelar essas assimetrias usando termos *polinomiais*. Assim, os valores são multiplicados por si. Há um efeito sobre a magnitude geral: quão mais longe do 0, maior a taxa de aumento. Valores extremos possuem imagens ainda mais extremas.

O principal é que temos a possibilidade de modelar assimetrias, uma vez que o produto de dois números negativos é positivo (e.g. $-2 \times -2 = 4$). Assim, espelhamos imagens negativas e podemos recriar curvas de

diversos tipos.

Para o caso da testosterona, se definirmos a origem (0) em 20:00, um modelo quadrático ($Y \sim X^2$, parábola) expressa perfeitamente a distribuição. O que seriam extremos negativos num modelo linear passam a ser extremos positivos num modelo quadrático, que decrescem em magnitude com o avançar do tempo. Quando atingimos a origem, os valores voltam a crescer.

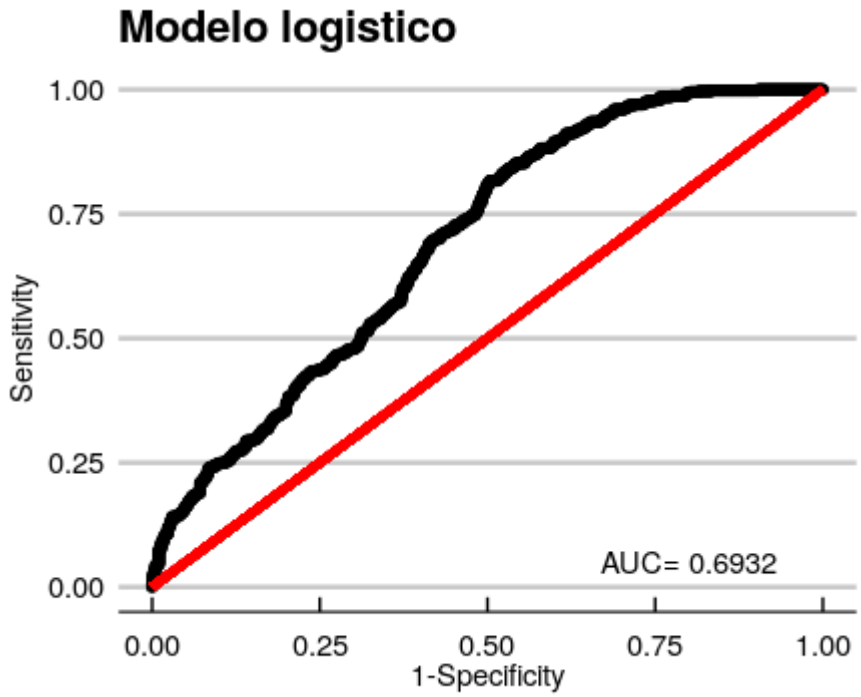
Simulação dos dados com cerca de 600 observações em cada classe (dopados e não dopados):

```
>library(tidyr)
>set.seed(2600)
>n_samp <- 601

>normal <- (purrr::map(seq(-3,3,0.01), .f =function(x) x^2) %>%
              as.numeric)+ rnorm(n_samp)
>over <- (purrr::map(seq(-3,3,0.01), .f =function(x) x^2+2) %>%
              as.numeric)+ rnorm(n_samp)
>horm_df <- data.frame(norm = normal, ov = over,time=1:n_samp)
>horm_gat <- gather(data=horm_df,key="class",value="testost",norm,ov)
>horm_gat$lab <-  car::recode(horm_gat$class,"'norm'=0;'ov'=1")
>horm_gat$id <- 1:nrow(horm_gat)
>head(horm_gat)
  time class   testost lab id
1    1  norm  8.549492   0  1
2    2  norm  9.090112   0  2
3    3  norm 10.362759   0  3
4    4  norm  8.801111   0  a4
5    5  norm 10.896185   0  5
6    6  norm  9.552046   0  6
```

Como verificado no *Capítulo 4*, o problema não é linearmente separável. Não é possível dividir os grupos com uma reta, então a regressão logística será apenas parcialmente satisfatória.
Usaremos biblioteca *Deducer* para avaliar rapidamente a performance do modelo usando curva ROC.

```
>library(Deducer)
>library(ggplot2)
>library(ggthemes)
>reg_log <- glm(lab ~ testost + time,
           data=horm_gat,family = binomial)
>rocplot(reg_log)+theme_economist_white(gray_bg = F)+
      ggtitle("Modelo logistico")
```

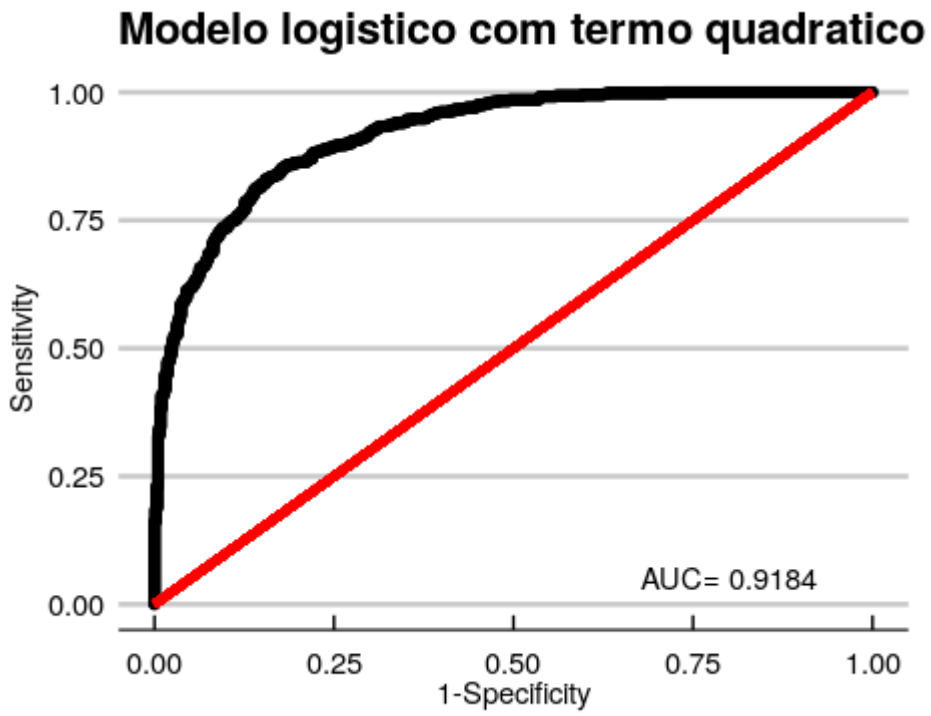


Observamos uma AUROC de de 0.693, o que é razoavelmente acurado. Examinando o modelo, notamos que a informação sobre tempo não é utilizada. O β atribuído é bastante próximo de 0 ($\beta_{tempo} \sim 3 \times 10^{-5}$), refletindo a incapacidade do modelo linear em capturar as assimetrias que descrevemos antes. Atribuir um coeficiente positivo implicaria uma relação de sentidos iguais : tempos avançados corresponderiam a dosagens maiores. Atribuir um coeficiente negativo implicaria uma relação de sentidos opostos: tempos avançados corresponderiam a dosagens menores.

```
>reg_log
(...)
Coefficients:
(Intercept)      testost         time
 -9.589e-01    2.374e-01    3.861e-05
(...)
```

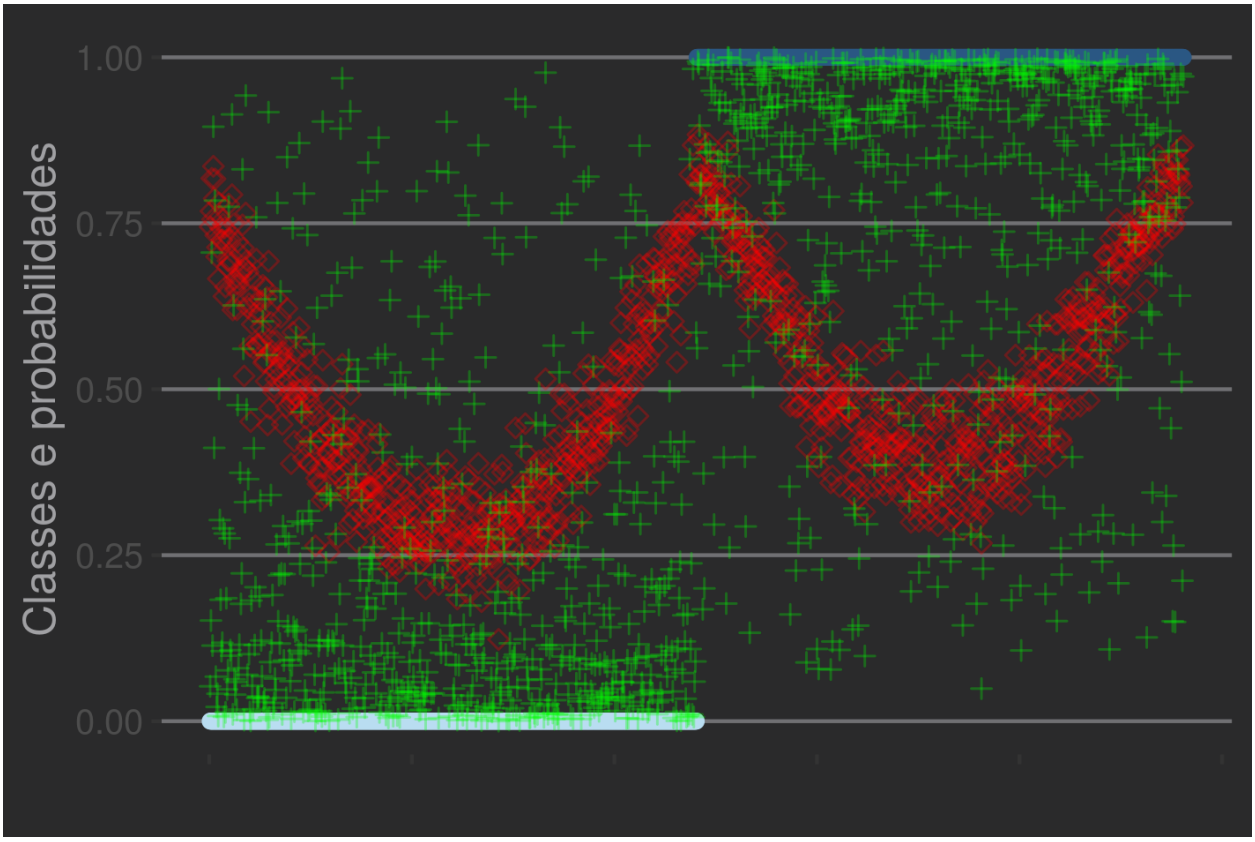
Como capturar as assimetrias no tempo? Uma solução é centralizar o tempo em na origem (0) e então usar um modelo quadrático.

```
>horm_gat$time2 <- scale(horm_gat$time)
>reg_log_quad <- glm(lab ~ testost + I(time2^2),
           data=horm_gat,family = binomial)
>rocplot(reg_log_quad)+theme_economist_white(gray_bg = F)+
      ggtitle("Modelo logistico com termo quadratico")
```



Agora, obtivemos uma classificação excelente (AUROC > 0.9).
O que acontece é que o modelo simples, com termos lineares, usa apenas doses de testosterona como parâmetro. Medidas altas serão classficadas como doping sem levar em conta o horário da medida. O modelo quadrático leva em conta flutações no tempo, resultando em predições adequadas para as classes.

```
>horm_gat$logis_pred <- predict(reg_log,type="response")
>horm_gat$logis_pred_quad <- predict(reg_log_quad,type="response")
>ggplot(data=horm_gat,aes(y=lab,color=lab,x=id))+
  geom_point()+
  ylab("Classes e probabilidades")+xlab("")+
  scale_x_continuous(labels=NULL)+
  theme_hc(style="darkunica")+
  geom_point(data=horm_gat,aes(y=logis_pred,x=id),
             color="red",shape=5,alpha=0.3)+
  geom_point(data=horm_gat,aes(y=logis_pred_quad,x=id),
             color="green",shape=3,alpha=0.4)+
  scale_color_continuous_tableau()+
  theme(legend.position = "none")
```



As probabilidades previstas pelo modelo linear (vermelho) flutuam com o tempo e se concentram numa faixa nebulosa (entre 0.25 e 0.75). Por outro lado, as probabilidades previstas pelo modelo quadrático (verde) são acuradas. O modelo atribui valores próximos a 0 (azul claro, na inferior à esquerda) para medidas normais e próximos a 1 (azul escuro, na margem superior à direita) para medidas sob efeito de dopping.

Soma de efeitos e séries de Taylor

Regressões de alta ordem

Exponenciais e logaritmos

Chap 6

Asymmetric extremes / : Linear
Symmetric extremes U : Quadratic
Momentum (. . . ooooOOO) & periodicity (. . . |. . . |. . . |. . . |. . . |. . . |. . . |. . . |) : ARIMA (moving average & auto-regression) / Dynamic Harmonic Reg. (Fourier periods)
Volatility (—∼∼'/WWW): Gaussian-process (variance) / HGF (hierarchical model for variance) and dynamical measures (.. → . . → . . → . . → . .)