# CE225 - Modelos Lineares Generalizados

Cesar Augusto Taconeli

10 de outubro, 2017

# Aula 12 - Regressão para dados binários - predição

# Introdução

- Modelos de regressão para dados binários são bastante utilizados para predição, ou seja, classificar indivíduos conforme suas probabilidades estimadas.

- Alguns exemplos:
  - Predição (classificação) de clientes em bons ou maus pagadores;
  - Predição de e-mails em spams ou não spams;
  - Predição do resultado de um jogo de basquete (vitória do time mandante ou do time visitante);
  - Prognóstico de um paciente (cura ou não cura)...

# Introdução

- É fortemente recomendável avaliar o poder preditivo do modelo ajustado com dados que não foram usados no ajuste.

- Ajustar o modelo e avaliar a qualidade preditiva usando os mesmos dados tende a produzir resultados excessivamente otimistas.

- Algumas possibilidades:
  - Separar aleatoriamente a amostra em duas partes (uma para ajuste, a outra para predição);
  - Usar validação cruzada (caso particular: *leave one out*).

# Predição

- Sejam $\hat{\pi}_i$ as estimativas de $P(y_i = 1)$, $i = 1, 2, ..., n$.

- Considere uma regra do tipo:
  - Predizer $\hat{y}_i = 0$ se $\hat{\pi}_i \leq p_0$;
  - Predizer $\hat{y}_i = 1$ se $\hat{\pi}_i > p_0$,

para algum valor (ponto de corte) especificado $p_0$ e $i = 1, 2, ..., n$.

- É comum (mas não obrigatório) usar $p_0 = 0.5$, classificando pelo resultado com maior probabilidade.

- Diferentes valores de $p_0$ conduzem a diferentes regras de predição.

# Tabelas de classificação

- Dadas as predições e os valores realmente observados de $y$, podemos construir uma tabela de classificação.

**Tabela 1:** Tabela de classificação

| | $\hat{y}$ | |
|---|---|---|
| y | 0 | 1 |
| 0 | $n_{00}$ | $n_{01}$ |
| 1 | $n_{10}$ | $n_{11}$ |

- Duas medidas úteis para sumarizar o poder preditivo de um modelo são a sensibilidade e a especificidade.

# Sumarizando o poder preditivo

- A **sensibilidade** de um modelo (ou de uma regra de classificação) é definida por $P(\hat{y} = 1|y = 1)$;

- A **especificidade** de um modelo (ou de uma regra de classificação) é definida por $P(\hat{y} = 0|y = 0)$.

- Podemos estimar a sensibilidade e a especificidade com base nas frequências de uma tabela de classificação:

$$\widehat{Sens} = \frac{n_{11}}{n_{10} + n_{11}}; \quad \widehat{Esp} = \frac{n_{00}}{n_{00} + n_{01}}. \tag{1}$$

# Sumarizando o poder preditivo

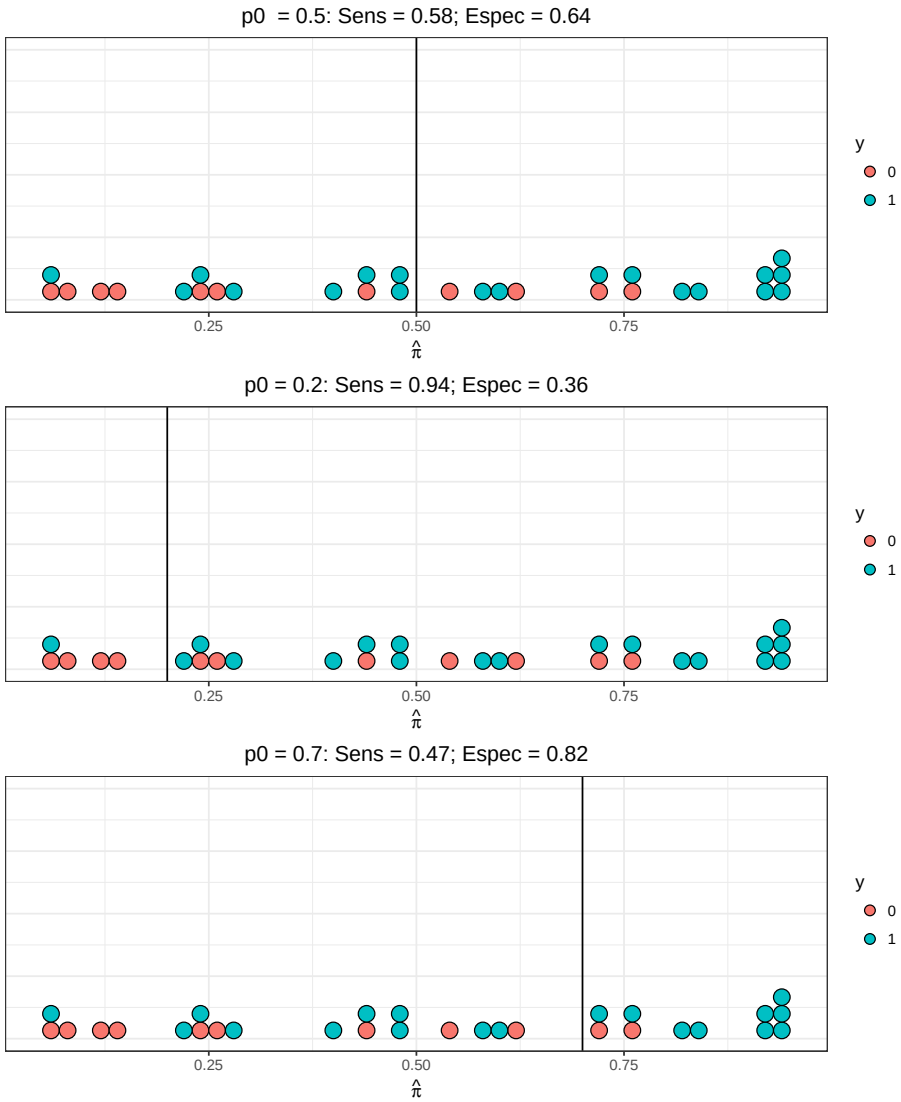


**Figura 1:** Ilustração - predição para dados binários

# Curva ROC

- Uma forma de analisar o poder preditivo associado a diferentes regras de decisão (valores de $p_0$) é por meio da **curva ROC**.
- A curva ROC permite avaliar conjuntamente a sensibilidade e a especificidade para diferentes valores de $p_0$.
- Para valores $p_0 \approx 1$, temos sensibilidade próxima de zero e especificidade próxima de um;
- Para valores $p_0 \approx 0$, temos sensibilidade próxima de um e especificidade próxima de zero;
- Em geral, busca-se $p_0$ tal que se tenha, conjuntamente, elevadas sensibilidade e especificidade;
- A **área sob a curva ROC** é uma medida de poder preditivo do modelo.

# Curva ROC

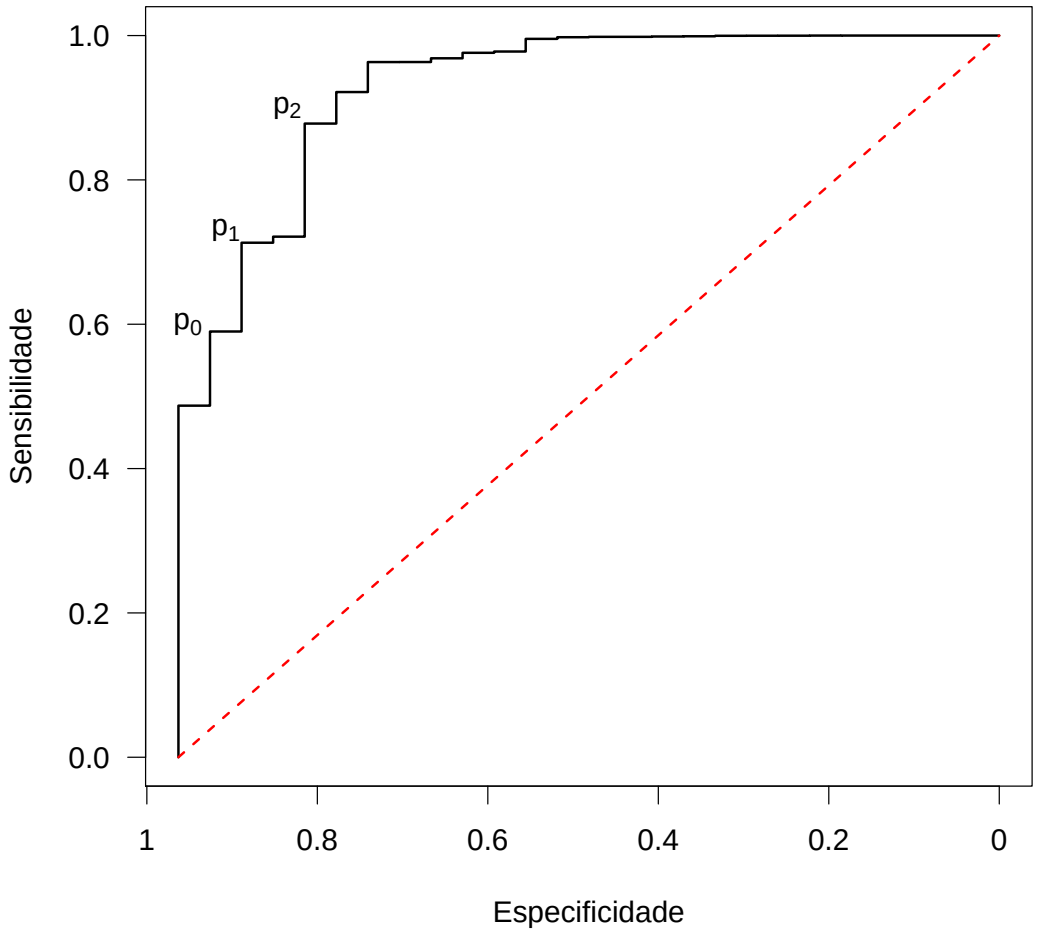


**Figura 2:** Ilustração - Curva ROC