Problemas – Regressão & Programação

James R. Hunter

18 de abril de 2017

Nestes exercícios, vamos experimentar com regressão e programação. Os exercícios terão várias partes. Não esqueça responder a todas! Tem no Github um arquivo `probsRegress.RData` com os dados que você precisa para completar os exercícios. Fazer o download dele e `load("probsRegress.RData")`.

1. Expectativa de Vida em Europa

Neste exercício, tirei dados do pacote `gapminder` sobre expectativa da vida (`lifeExp`) e PIB por capita (`gdpPercap`) para os anos 2002 e 2007 para os países de Europa. O código para reproduzir os dados para o problema segue. Você deve copiar e colar ele no seu trabalho. Faça uma regressão simples linear que mostra qual é o efeito que PIB tem sobre expectativa da vida. Países ricas têm expectativa de vida maior? Responde às partes a - f. Como sempre, não esqueça de fazer um pequeno estudo exploratório dos dados.

Dados do problema

```
library(gapminder)
vidaExp <- gapminder %>%
           filter(year > 2000 & continent == "Europe") %>%
           select(year, lifeExp, gdpPercap)
```

Perguntas

a. A variável `lifeExp` tem uma distribuição normal segunda o teste Shapiro-Wilks?
b. Uma transformação logarítmica pode fazer ela normal? Por que?
c. Reconhecendo que a variável dependente não é puramente normal, você pode confiar em qual regra de estatística para usar regressão linear? Por que?
d. O que é a equação linear que determina a relação entre as variáveis no formato de $y = \beta_0 + \beta_1 x$
e. Qual proporção de variância no modelo esta equação descreve?
f. Mostre e examine os quatro gráficos que pode usar para entender melhor a regressão. Essa regressão é confiável? Por que?

2. Loops, if . . . then

No conjunto de dados `vidaExp`, você quer criar uma nova variável categórica que expressa `gdpPercap` em duas categorias: “alto”, “baixo”. Você vai dividir a variável ao ponto da média da `gdpPercap`.

a. Escreva e execute um bloco de código usando `ifelse()` que cria a nova variável `pibcat`.
b. Use uma combinação de um loop e uma construção condicional (“if . . . then”) para conseguir esta tarefa.

3. Kilometragem dos Carros

Uma sondagem sobre carros em 1970 listou 392 modelos de carros e a economia de combustível eles tiveram. Teve vários indicadores de que seria a quilometragem de combustível, como horsepower (cavalos). Para este problema, nós vamos trabalhar com `auto1`.

Perguntas e Tarefas

a. Faça uma análise exploratória dos duas variáveis (`mpg` e `horsepower`)
b. Faça um scatterplot de mpg (eixo-y) e horsepower (eixo-x). Mostra alguma tendência?
c. Tendência é linear ou não-linear? Se for não-linear, qual poder melhor expressa esta relação
d. Faça uma regressão linear simples entre mpg e horsepower. Escreva a equação da regressão e o R^2

e. Mostre os 4 gráficos para o modelo simples. Mostra uma tendência nos resíduos?
f. Faça uma regressão linear polinomial de segundo grau entre mpg e horsepower. Escreva a equação da regressão e o R^2
g. Qual modelo teve a melhor R^2?
h. Mostre os 4 gráficos para modelo polinomial.

4. auto2 – Regressão Múltipla

Esta vez, nós vamos usar outras variáveis relacionados aos motores dos carros para ver se elas têm influência sobre economia de combustível. O conjunto `auto2` tem esses dados.

a. Faça uma análise exploratória sobre as variáveis novas (`displacement`, `weight`, `acceleration`)
b. Faça uma regressão múltipla usando todas as variáveis independentes.
c. Mostre o resultado (`summary()`)
d. Qual porcentagem da variância dos dados em total este modelo descreve?
e. Quais variáveis parecem não ter uma relação significante com a `mpg`? Porque, você acha?

5. Regressão Lógistica

Vamos agora olhar num estudo sobre câncer de próstata. A questão aqui é de entender melhor se o câncer espalhou para os linfonodos em volta da próstata. O estudo tenta avaliar se cinco indicadores podem substituir uma cirurgia exploratória. As cinco variáveis no conjunto de `proscan` são

1. `raioX`: leitura de um raio X; valores binários 1 = positivo, 0 = negativo
2. `grau`: leitura patológica como resultado de uma biopsia de agulha fina; valores binários 1 = positivo, 0 = negativo
3. `estagio`: tamanho do tumor obtido pela palpação com os dedos; valores binários 1 = positivo, 0 = negativo
4. `idade`: idade do paciente em anos
5. `acido`: nível x 100 de fosfatase ácida sérica

A variável `linfonodos` tem o resultado determinado pela cirurgia se o câncer tinha espalhado ou não

Tarefas

a. Faça uma análise exploratória dos dados, inclusive com `cplot()` para entender o problema melhor
b. Construa um modelo logístico de linfonodos contra as outras variáveis
c. Todas as variáveis são significativas? Quais são e quais não são
d. Construa um segundo modelo logístico usando `raioX`, `estagio` e `acido`

e. Este modelo descreve mais da deviança nos dados?
f. Construa um terceiro modelo com só as variáveis significativas.
g. Faça uma comparação entre os três modelos. Qual é o melhor? Com este modelo, calcule os odds, um intervalo de confiança para os odds e a probabilidade de ocorrência da presença de tecido maligno nos linfonodos.