Big Data Analytics com R

Construir análise preditiva

Agenda

- Implementar modelo de regressão em Big Data;
- Compreender os diferentes tipos de regressão e seus resultados em Big Data.

Regressão Linear

O que é?

- O modelo utilizado para estimar uma eventual relação (linear) existente entre duas variáveis Y e X é chamado modelo de regressão linear simples.

Modelo de Regressão Simples

$$Y = \beta_0 + \beta_1 x + \varepsilon;\ \varepsilon \sim N(0,\sigma^2).$$

- Y é a variável dependente;
- X é a variável explicativa, cujo valor observado x aparece no modelo.
- A ideia é que o comportamento de Y pode ser explicado por meio de uma função linear de x, acrescida de um erro aleatório (ε), que possui distribuição normal.

Reta de Regressão (Teórica)

- A reta de regressão expressa o valor esperado de Y como função linear exata de x:

$$E(Y) = \beta_0 + \beta_1 x$$

Os coeficientes β_0 e β_1 são parâmetros populacionais, e devem ser estimados a partir dos dados.

Interpretação do Coeficiente β_0

- Se fizemos x = 0, ficamos com:
  - $E(Y) = \beta_0$.
- Logo:
  - β_0 representa o valor esperado de Y, quando x = 0.

Interpretação do Coeficiente β_1

- β1 representa a variação esperada em Y, em resposta à uma variação unitária em x.

Coeficientes

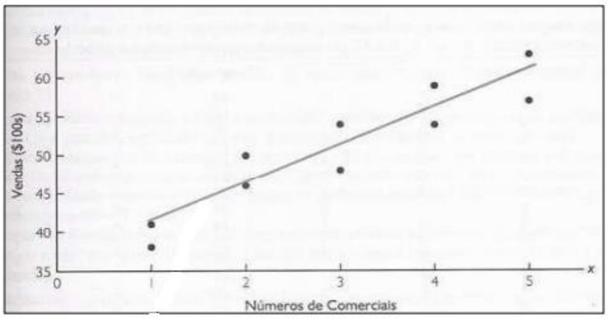
- β_0 é chamado 'intercepto' do modelo.
- β_1 é chamado 'coeficiente de inclinação' do modelo.

Pressupostos básicos do modelo

- O modelo de regressão parte de alguns pressupostos básicos, que são hipóteses acerca do erro do modelo:
  - $V(\varepsilon) = \sigma^2$ (homocedasticidade);
  - ε segue distribuição normal;
  - $Corr(\varepsilon_i,\varepsilon_j) = 0, \forall(i,j)$.

Estimação dos coeficientes

- Em princípio, dado um diagrama de dispersão, existem diversas retas que poderiam ajustar-se aos pontos. Por exemplo: sejam os dados abaixo e a reta que melhor se ajusta a ela.
- Qual o critério usado para chegar a ela?



Método dos Mínimos Quadrados

- O método dos mínimos quadrados consiste em obter os valores de β_0 e β_1 que minimizam a soma dos quadrados dos erros:

$$\boxed{S = \sum_{i=1}^{n} \varepsilon_i^2 .}$$

Método dos Mínimos Quadrados

- Desta forma, obtemos os estimadores abaixo:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \; \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}.$$

Exemplo 1

- Considere que estejamos interessados em estimar os parâmetros de um modelo para relacionar os gastos com alimentação (Y) e a renda familiar (X), considerando uma amostra de 3 famílias.
- Considere que a amostra tenha fornecido os seguintes dados (em R$ 1.000,00):

Família 1 – $y_1 = 2$; $x_1 = 3$;
Família 2 – $y_2 = 3$; $x_2 = 4$;
Família 3 – $y_3 = 4$; $x_3 = 8$.

Exemplo 1

- Calcule as estimativas de mínimos quadrados dos coeficientes do modelo.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{3}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{3}(x_i - \bar{x})^2} = \frac{(3-5)(2-3)+(4-5)(3-3)+(8-5)(4-3)}{(3-5)^2+(4-5)^2+(8-5)^2}$$

$$= \frac{(-2)(-1)+(-1)(0)+(3)(1)}{(-2)^2+(-1)^2+(3)^2} = \frac{2+0+3}{4+1+9} = \frac{5}{14} = 0{,}3571$$

$$\hat{\beta}_0 = 3 - 0{,}3571 * 5 = 1{,}2143$$

Exemplo 1

- Interpretação da estimativa de β_1:
  - Para cada R$ 1000,00 de renda, espera-se um acréscimo de R$ 357,10 nos gastos com alimentação.
- Interpretação da estimativa de β_0:
  - Não faz muito sentido neste caso (por que?).

Reta de Regressão Estimada

As estimativas de β_0 e β_1 são utilizadas para construir a **reta de regressão estimada**:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

que permite predizer o valor de Y correspondente a um valor de X.

Exemplo 2

- Usando os estimadores calculados no exemplo 1, se x = 6, qual o valor "predito" de Y, pela reta de regressão estimada?

$$\hat{Y}_i = 1{,}2143 + 0{,}3571 * 6 = 3{,}3571.$$

O teste ‘t’ de significância

- Testar a significância da estimativa de um parâmetro significa testar a hipótese de que o valor real do parâmetro seja zero.
- Se uma das estimativas for não significante, ao nível considerado, então a variável correspondente deve ser retirada do modelo, que deverá ser estimado novamente sem ela.

O teste ‘t’ de significância

- Especificamente, testar H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$ equivale a testar a significância da regressão, uma vez que, se $\beta_1 = 0$, então:

$$Y = \beta_0 + \varepsilon$$

- Isto é, não existe regressão de Y em x.

O teste 't' de significância

- Se a hipótese $\beta_1 = 0$ não for rejeitada, concluímos que a regressão não é significante (ao nível α considerado).
- Se, por outro lado, a hipótese $\beta_1 = 0$ for rejeitada, então a regressão é significante (ao nível correspondente).

Exemplo

- Considere a amostra de n = 40 pares de observações de 'Y = gasto com alimentação' e 'x = renda'. O resultado da execução de uma regressão linear neste cenário segue no próximo slide. Pede-se:

a) Estime um modelo de regressão linear para explicar o gasto com alimentação a partir da renda.

b) A estimativa de β_0 é significante? A regressão é significante (ou seja, a estimativa de β_1 é significante)? Use $\alpha = 0,05$.

c) Escreva a expressão da reta de regressão estima[illegible] e tente predizer o valor de Y, para x = 100.

Exemplo

a) A saída do excel é reproduzida a seguir:

Estatística de regressão	
R múltiplo	0,937608458
R-Quadrado	0,879109621
R-quadrado ajustado	0,875928295
Erro padrão	4,81040437
Observações	40

R^2

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	6394,37439	6394,374	276,3343605	5,02495E-19
Resíduo	38	879,319628	23,13999		
Total	39	7273,69402			

IC`s de 95% para β_0 e β_1.

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	-13,3248381	4,45111079	-2,993598	0,004827374	-22,33564114	-4,314035109
Variável X 1	0,515720938	0,03102397	16,62331	5,02495E-19	0,452916199	0,578525676

Estimativa de β_0

Estimativa de β_1

Valores-p dos testes t para H_0: $\beta_0 = 0$ e para H_0: $\beta_1 = 0$.

Exemplo

b) O valor-p da estimativa de β_0 é 0,0048, portanto menor do que 0,05. Logo, a estimativa de β_0 é significante ao nível α = 0,05. O valor-p da estimativa de β_1 é $5*10^{-19}$, portanto menor do que 0,05. Logo, a estimativa de β_1 é significante ao nível α = 0,05. Portanto, a regressão é significante.

<u>Reforçando</u>: a significância da estimativa de β_0 não diz nada sobre a significância da regressão. Neste contexto, é somente o valor-p associado à estimativa de β_1 que importa.

Exemplo

c)

$$\hat{Y}_i = -13{,}3248 + 0{,}5157 * 100 = 38{,}2452.$$

Coeficiente de determinação R^2

- $R^2 \in [0,1]$.
- Quanto mais próximo de UM estiver o R^2, maior a qualidade do ajuste do modelo de regressão aos dados da amostra.
- Estar bem ajustado significa explicar boa parte da variação de Y.

Exemplo

- Avalie a qualidade do ajuste do modelo do exemplo anterior.
- Solução:
  - R^2 = 0,8791, ou seja, o modelo explica 87,91% da variação de Y, o que é considerado muito bom (valores de R^2 superiores a 0,8 em geral são considerados bastante satisfatórios).

Considerações

- O R^2 não pode ser usado como única medida da qualidade de um modelo, ou seja, um modelo não pode ser descartado por ter um R^2 baixo.
- Em algumas situações, valores baixos de R^2, até menores do que 0,5, podem ser aceitáveis.
- O único problema, neste caso, é que o modelo terá uma capacidade de predição baixa, decorrente do ajuste ruim.
- Entretanto, as estimativas dos coeficientes, caso sejam estatisticamente significantes, devem ser consideradas e podem fornecer informações importantes.

Regressão Logística

Considerações

- O modelo de regressão logístico é utilizado quando a variável resposta é qualitativa, com dois resultados possíveis.
- Seja a probabilidade de sucesso p.
- A probabilidade de fracasso será $1 - p = q$.
- Chamamos de 'Chance' a razão entre a probabilidade de sucesso e a probabilidade de fracasso.
- Ex.: se a probabilidade de sucesso é 0,75, a chance é igual a:

$$\frac{p}{(1-p)} = \frac{p}{q} = \frac{0{,}75}{0{,}25} = 3$$

Variável dependente binária

- Vamos considerar o modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}$$

- A resposta esperada é dada por:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Logit

- O *logit* equivale ao logaritmo natural (base *e*) da chance:

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

- A função logística será dada pelo logit-inverso, que nos permite transformar o *logit* em probabilidade:

$$p = \frac{exp(x)}{1 + exp(x)}$$

Razões de chance (log-odds)

- Compara a chance de sucesso de um grupo em relação a outro grupo:

$$log(R) = log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$$

$$log(R) = log\left(\frac{p_1}{1-p_1}\right) - log\left(\frac{p_2}{1-p_2}\right)$$

$$log(R) = logit(p_1) - logit(p_2)$$

- Portanto, a diferença entre os *logits* de duas probabilidades equivale ao logaritmo da razão de chances.

Razões de chance (log-odds)

- A razão de chance será dada pela expressão exp(γ): chance de sucesso no grupo A, em relação ao grupo B:

$$\frac{A}{B} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{exp(\beta_0 + \gamma)}{exp(\beta_0)} = \frac{exp(\beta_0) * exp(\gamma)}{exp(\beta_0)} = exp(\gamma)$$

Razões de chance (log-odds)

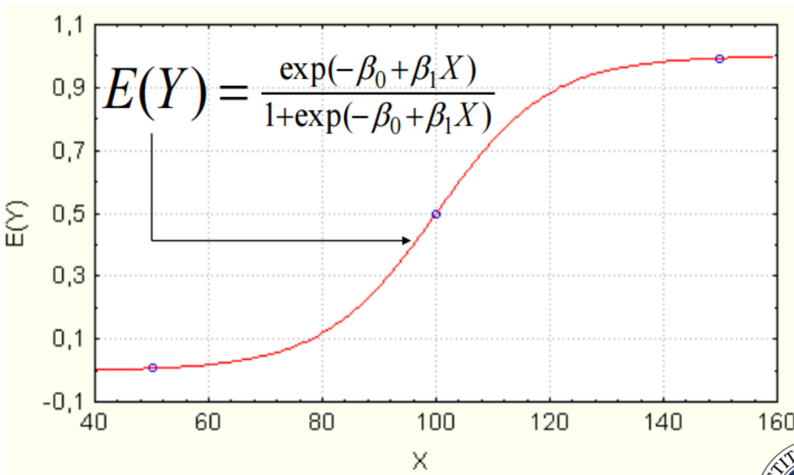
- Se $\exp(\gamma)$ for maior que uma unidade, chance de sucesso em A é maior que em B.
  - Ex.: $\exp(\gamma)$=1,17, chance de sucesso em A é 1,17 vezes maior do que em B, ou seja, é 17% maior do que em B.
- Se $\exp(\gamma)$ for menor que uma unidade, chance de sucesso em A é menor que em B.
  - Ex.: $\exp(\gamma)$=0,61, chance de sucesso em A é 0,61 vezes a chance de B, ou seja, é 39% menor do que em B.

Valor esperado

- As funções respostas são denominadas funções logísticas, cuja expressão é:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1+\exp(\beta_0 + \beta_1 X)}$$

Valor esperado



Valor esperado