Avaliação de Supervised Machine Learning: Modelos de Regressão para Dados de Contagem I - 21/03/2023

Supervised Machine Learning: Modelos para Dados de Contagem I

Professor: Luiz Paulo Lopes Fávero
Avaliação realizada por:
Avaliação realizada em: 12/04/2023

Tentativa
1 de 3**Nota**
10,0**Questões Respondidas**
10 de 10

Questão #1

Em qual das alternativas a seguir há indícios de existência de superdispersão nos dados de determinada variável dependente considerada em um modelo de regressão para dados de contagem?

- (•) Média = 6,34 e Variância = 128,21
- ( ) Média = 1,73 e Variância = 1,73
- ( ) Média = 2,47 e Variância = 2,46
- ( ) Média = 7,47 e Variância = 7,46

Questão #2

São exemplos de variáveis com dados de contagem:

I) Quantidade de vezes que pacientes idosos vão ao médico por ano.

II) Quantidade de ofertas públicas de ações que são realizadas em uma amostra de países desenvolvidos e emergentes por ano.

III) Quantidade de apartamentos à venda por bairro.

IV) Faixa de renda (definida em labels) de uma amostra de consumidores.

Assinale a alternativa **CORRETA**:

- Somente as afirmações II e IV estão corretas.
- Somente as afirmações I, II e III estão corretas.
- Nenhuma afirmação está correta.
- Todas as afirmações estão corretas.

Questão #3

Com o intuito de se estudar e projetar a quantidade de violações de trânsito (variável dependente *violations*) na cidade de Nova York por parte de membros do corpo diplomático de países pertencentes às Nações Unidas, foi estimado um modelo de regressão Poisson, considerando, como variáveis preditoras, a quantidade de membros no corpo diplomático em cada país (variável *staff*), o índice de corrupção de cada país (variável *corruption*) e o fato de haver ou não enforcement legal quanto à obrigatoriedade de se pagar a multa em caso de violação (variável dummy post: *yes* = há obrigatoriedade do pagamento; no = não há obrigatoriedade do pagamento). Os outputs do referido modelo, obtidos no R, encontram-se na figura abaixo.

Pergunta-se: qual a equação do modelo de regressão Poisson que deverá ser utilizada para fins preditivos? O subscrito *i* refere-se à linha (*row*) do dataset.

```
Call:
glm(formula = violations ~ staff + post + corruption, family = "poisson",
    data = corruption)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.1425  -2.8326  -0.6008  -0.3940  24.6141

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.212739   0.031107   71.13   <2e-16 ***
staff        0.021870   0.001228   17.81   <2e-16 ***
postyes     -4.296762   0.197446  -21.76   <2e-16 ***
corruption   0.341765   0.027495   12.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6397.7  on 297  degrees of freedom
Residual deviance: 3644.0  on 294  degrees of freedom
AIC: 4151.6
```

a) $\ln(violations_i) = 2,212739 + 0,021870 \cdot (staff_i) - 4,296762 \cdot (post = "yes"_i) + 0,341765 \cdot (corruption_i)$

b) $violations_i = 2,212739 + 0,021870 \cdot (staff_i) - 4,296762 \cdot (post = "yes"_i) + 0,341765 \cdot (corruption_i)$

c) $e^{(violations_i)} = 2,212739 + 0,021870 \cdot (staff_i) - 4,296762 \cdot (post = "yes"_i) + 0,341765 \cdot (corruption_i)$

d) $violations_i = 0,021870 \cdot (staff_i) - 4,296762 \cdot (post = "yes"_i) + 0,341765 \cdot (corruption_i)$

- a)
- b)
- c)
- d)

Questão #4

Com o intuito de se estudar e projetar a quantidade de violações de trânsito (variável dependente *violations*) na cidade de Nova York por parte de membros do corpo diplomático de países pertencentes às Nações Unidas, foi estimado um modelo de regressão Poisson, considerando, como variáveis preditoras, a quantidade de membros no corpo diplomático em cada país (variável *staff*), o índice de corrupção de cada país (variável *corruption*) e o fato de haver ou não *enforcement* legal quanto à obrigatoriedade de se pagar a multa em caso de violação (variável dummy post: *yes* = há obrigatoriedade do pagamento; no = não há obrigatoriedade do pagamento). Os outputs do referido modelo, obtidos no R, encontram-se na figura abaixo.

Pergunta-se: qual a quantidade esperada de violações de trânsito para um país cujo

corpo diplomático seja composto por 28 membros, considerando inexistência de *enforcement* legal (post = "no", ou seja, dummy *postyes* = 0) e índice de corrupção igual a 1?

```
Call:
glm(formula = violations ~ staff + post + corruption, family = "poisson",
    data = corruption)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.1425  -2.8326  -0.6008  -0.3940  24.6141

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.212739   0.031107   71.13   <2e-16 ***
staff        0.021870   0.001228   17.81   <2e-16 ***
postyes     -4.296762   0.197446  -21.76   <2e-16 ***
corruption   0.341765   0.027495   12.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6397.7  on 297  degrees of freedom
Residual deviance: 3644.0  on 294  degrees of freedom
AIC: 4151.6
```

- 17,93
- 23,73
- 29,32
- 47,54

Questão #5

A função e o respectivo pacote para a elaboração direta do teste para verificação de existência de superdispersão nos dados da variável dependente para a estimação de modelos de contagem, no R, são:

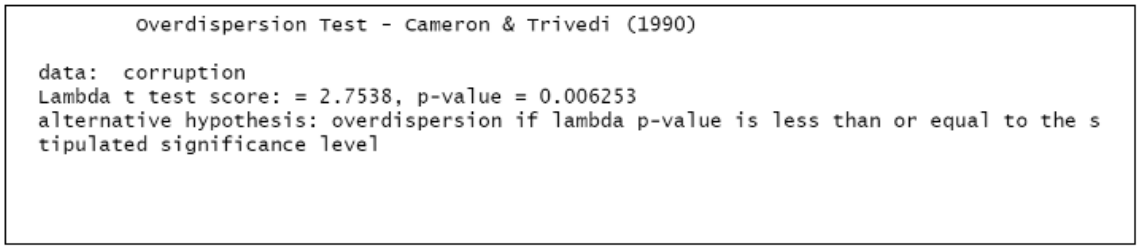
- função **overdisp** do pacote **overdisp**.
- função **superdisp** do pacote **hiperdisp**.
- função **hiperdisp** do pacote **megadisp**.
- função **megadisp** do pacote **ultradisp**.

Questão #6

O resultado de um teste para verificação de existência de superdispersão na variável dependente de determinado modelo está apresentado na figura abaixo.

A partir do output da figura, considerando um nível de significância de 5%,

é **CORRETO** afirmar que:



```
        Overdispersion Test - Cameron & Trivedi (1990)

data:  corruption
Lambda t test score: = 2.7538, p-value = 0.006253
alternative hypothesis: overdispersion if lambda p-value is less than or equal to the s
tipulated significance level
```

- A partir deste output não se pode concluir nada a respeito de uma eventual superdispersão nos dados da variável dependente, já que (lambda – theta = delta).
- Verifica-se a existência de equidispersão nos dados da variável dependente.
- Verifica-se a existência de superdispersão nos dados da variável dependente.
- Verifica-se a existência de dispersão reversa nos dados das variáveis preditoras.

Questão #7

Considere as seguintes afirmações:

I) Um modelo de regressão Poisson pode ser estimado quando a variável dependente for qualitativa com três categorias.

II) É correto e adequado estimar um modelo de regressão Poisson quando a variável dependente for quantitativa e apresentar superdispersão nos dados.

III) Em um modelo de regressão Poisson são estimados (M – 1) logitos, sendo M o número de categorias da variável dependente.

Assinale a alternativa **CORRETA**:

- Nenhuma afirmação está correta.
- Todas as afirmações estão corretas.
- Somente as afirmações I e II estão corretas.
- Somente as afirmações II e III estão corretas.

Questão #8

Sobre os modelos de regressão para dados de contagem, podemos avaliar a qualidade do ajuste do modelo por meio do seguinte indicador:

- Lambda de Box-Cox.
- p-value da estatística *t* de Student.
- Área abaixo da curva ROC.
- Valor de Log-Likelihood.

Questão #9

O principal teste para verificação de existência de superdispersão nos dados da variável dependente é o:

- Teste de Lambert.
- Teste de Shapiro-Francia.
- Teste de Cameron e Trivedi.
- Teste de Vuong.

Questão #10

Uma variável com dados de contagem apresenta as seguintes características:

- É quantitativa, apresenta dados contínuos e negativos, e é definida uma exposição.
- É quantitativa, apresenta dados contínuos e não negativos, e não se consegue definir uma exposição.
- É quantitativa, apresenta dados discretos e não negativos, e é definida uma exposição.
- É qualitativa, apresenta dados contínuos e negativos, e é definida uma exposição.