PERGUNTAS REALIZADAS DURANTE A AULA E RESPONDIDAS POR E-MAIL

Curso: MBA em Data Science e Analytics
Disciplina: Supervised Machine Learning: Modelos para Dados de Contagem II
Data: 28/09/2021

Leandro Rocha
Professor, poderia talvez na próxima aula, fazer uma análise plugada em um banco de dados? eu tenho essa demanda e tenho sofrido com as conexões, tanto ao atualizar ao banco ou na 1ª conexão
Resposta: Perdão, mas não consegui entender esta solicitação.

Yuri José de Santana Furtado
Na expressão preditiva, pq o logito aparece com o sinal negativo na parte que se refere aos zeros estruturais se o "zero" é a probabilidade de não-ocorrência. O sinal não deveria ser positivo?
Resposta: Justamente por esta razão. Porque está se subtraindo a parte de contagem total pela parte de contagem devida exclusivamente aos zeros estruturais.

Fernando Kanashiro
Por mais que o nome faça referência a inflação de zeros, eu não poderia ter inflação em uma outra contagem?? E se sim, faz sentido seguir uma lógica semelhante a esses modelos?
Resposta: Sim, é uma boa pergunta. Pode-se estabelecer uma inflação de zeros para qualquer distribuição de probabilidades.

Lucas Bruno Marques
Professor, vimos bastante sobre a superdispersão e o impacto dela na capacidade preditiva dos modelos. Lendo por fora esbarrei também na subdispersão. Ela também impacta a preditiva de dos modelos??
Resposta: Boa pergunta. A subdispersão (variância <<< média) não afeta a capacidade preditiva de modelos de contagem.

Laila Monte Neto Donni
temos algum teste para saber se falta alguma variável explicativa no modelo, como o teste para modelos binomiais?
Resposta: Utilizamos o teste de heterocedasticidade para diagnóstico de omissão de variáveis relevantes em modelos GLM gaussianos. A inclusão de contextos, como veremos nas aulas de modelagem multinível, naturalmente considera e modela o fenômeno da heterocedasticidade.

Lays Janaina Prazeres Marques
Prof, como incluir o intervalo de confiança no output do modelo?
Resposta: Obrigado pela pergunta. É só comandar função confint, sendo o argumento o modelo estimado, tal como: `confint(modelo)`

Rubelmar Maia De Azevedo Cruz Neto

Professor Fávero, o senhor poderia falar sobre os cálculos da entropia para o modelos Poisson e Binomial Negativo? Como seria o ajuste utilizando os valores de entropia?
Resposta: É sabido que as distribuições que maximizam a entropia sob determinadas condições naturais assumem uma forma simples. Por exemplo, entre variáveis aleatórias com média e variância fixas, a entropia é maximizada pela distribuição normal. Da mesma forma, para variáveis aleatórias com simetria positiva e média fixa, a entropia é maximizada pela distribuição exponencial. A técnica padrão para provar tais resultados usa a desigualdade de Gibbs. Para distribuição para dados de contagem, sugiro o estudo dos textos que se encontram nos seguintes links:
https://www.jstor.org/stable/2946418
https://pure.tue.nl/ws/files/1959440/Metis199989.pdf

Thiago Colette Vegi
Porque na planilha os fitted values não consideram nem phy nem theta?:
Resposta: Existem algumas formas funcionais análogas para o logaritmo da função de verossimilhança de distribuições Poisson-Gama. Utilizamos uma forma simples para facilidade de convergência do algoritmo no Solver do Excel. A taxa de decaimento delta está implicitamente considerada pela própria distribuição dos dados da variável Y.

Fabrício Santos Barbacena
[Revisada] Nos modelos supervisionados GLM vistos até agora nas suas aulas, é necessário fazer a divisão dos dados em treino/validação/teste? Se sim, como e em que momento fazer esses procedimentos?
Resposta: Embora este procedimento seja realizado em algumas práticas do mercado, realmente não há necessidade para modelos GLM e GLMM, já que o algoritmo pode ser treinado para toda a amostra, visto que a estimação dos parâmetros é determinada por meio de processos determinísticos. Faço uma discussão sobre este ponto no início da primeira aula de modelagem multinível.

Elisangela Aparecida Dos Santos Valenca
Professor, qual a melhor forma de reconciliar o resultado do modelo com os dados observados na base utilizada para modelagem (Comparar frequência estimada com a frequência observada)?
Resposta: Boa pergunta. Em modelos logísticos, por meio da matriz de confusão. Para modelos com dados de contagem, como a variável Y é quanti, não faz muito sentido comparar frequências, a menos que sejam discretizados os valores obtidos de lambda (fitted values) e que a variável Y original não apresente muitas possibilidades de respostas. É por esta razão que ainda preferimos que a comparação seja feita por meio do valor de LogLik.

Alexandre Amodio Pereira
Usando a predição com os quatro modelos, com os dados do Brasil (staff 33, corrup -0.096), teremos uma variação de 18 - 29. Já para o apurado do Brasil (67), não é um país lógico?
Resposta: Concordo. Estamos tentando buscar a maximização do acerto global para toda a amostra a partir da maximização de LogLik. Este é um claro exemplo de que a consideração de grupos contextuais, por meio da modelagem multinível, poderá melhorar ainda mais a qualidade do ajuste, conforme veremos nas duas próximas aulas.

Andre Morato

Para definir a curva poisson-gama precisamos dos parâmetros de forma (theta) e taxa de decaimento (delta). Não ficou claro para mim porque na maximização da loglik precisamos estimar apenas o theta.

Resposta: Existem algumas formas funcionais análogas para o logaritmo da função de verossimilhança de distribuições Poisson-Gama. Utilizamos uma forma simples para facilidade de convergência do algoritmo no Solver do Excel. A taxa de decaimento delta está implicitamente considerada pela própria distribuição dos dados da variável Y.

Damião Flávio dos Santos

Sei que esse assunto virou até meme, mas temos dados de contagem negativas da covid, que trata-se de atualização por parte da secretaria, queria que comentasse sobre isso e como resolver. Obrigado!

Resposta: Dados com contagem negativa não podem, diretamente, ser considerados como variável dependente em modelos de contagem. Para fins de modelagem, os valores desta variável precisam ser somados por uma constante (que equivale ao módulo do menor valor observado), para posterior estimação do modelo. Para obtenção correta dos valores previstos, é preciso então que se subtraia dos fitted values obtidos aquele valor da constante.