

A more interpretable regression model for count data with excess of zeros

Por Gustavo H. A. Pereira, Jeremias Leão, Manoel Santos-Neto e Jianwen Cai

Fonte: OpenAlex · UFC · [Publicação original](#)

INOVAÇÃO	APLICABILIDADE	POTENCIAL ECONÔMICO	MATURIDADE
7/10	8/10	6/10	Baixa

Desenvolvimento de um novo modelo de regressão estatístico pela UFC para lidar com dados de contagem com excesso de zeros, oferecendo uma alternativa superior em termos de interpretabilidade comparada aos modelos clássicos de inflação de zero (ZIP), com validação em dados de saúde materna.

NÚCLEO DA ANÁLISE

Ideia de startup ou produto

Desenvolvimento de uma biblioteca estatística (Python/R) ou SaaS focado em 'Explainable Count Analytics' para o setor de saúde, permitindo que pesquisadores e gestores hospitalares apliquem este modelo facilmente sem profundo conhecimento matemático.

Aplicações práticas

Análise de dados em pesquisas clínicas (ensaios de eventos raros), epidemiologia (taxas de infecção/hospitalização), setor de seguros (sinistralidade) e qualidade industrial (contagem de defeitos), onde a explicação do 'porquê' dos zeros é tão importante quanto a previsão.

Potencial de mercado

Alto potencial em setores de HealthTech e InsurTech que necessitam de modelos de 'caixa branca' (explicáveis) para justificar decisões de risco, alocação de recursos ou aprovação regulatória, tendência fortalecida por regulações como a LGPD e exigências de auditoria em IA.

FUNDAMENTO CIENTÍFICO

Problema abordado

A dificuldade de interpretação dos modelos de inflação de zero atualmente utilizados em pesquisas médicas e de ciência de dados. Modelos tradicionais como Poisson ou Binomial Negativa falham em excesso de zeros, enquanto modelos mais complexos (ZIP) sacrificam a capacidade explicativa, crucial para tomadas de decisão clínica e de política pública.

Metodologia

Proposição de um modelo que estima a taxa média de eventos e modela a dispersão dependente de covariáveis diretamente (interpretada como índice de agrupamento), utilizando estimação por máxima verossimilhança e validação via simulação Monte Carlo.

Principais descobertas

O modelo proposto mantém a robustez estatística necessária para lidar com superdispersão e excesso de zeros, mas permite uma leitura direta do impacto das variáveis explicativas, facilitando a inferência e o diagnóstico em comparação com modelos de mistura.

LEITURA PARA GESTÃO PÚBLICA

Aprimoramento de indicadores de saúde pública, especificamente na avaliação de programas de acesso (ex: pré-natal ou vacinação), permitindo diferenciar estatisticamente se a ausência de visitas se deve a falta de acesso (zero estrutural) ou ausência de necessidade, otimizando recursos da SES/Ceará.

PARCERIA UNIVERSIDADE × EMPRESA

Colaboração entre o Departamento de Estatística da UFC e empresas de tecnologia médica (como MV ou Philips) para integrar este algoritmo em sistemas de Business Intelligence hospitalar, gerando relatórios de gestão mais acuráveis.

OPORTUNIDADES DERIVADAS

4 direções estratégicas identificadas

STARTUP · IMPACTO MÉDIO · CIÊNCIA DE DADOS

Explainable Health Stats

Startup de Data Science focada em consultoria e software para análise de eventos raros em saúde, utilizando este modelo para oferecer insights mais claros do que modelos de Black-box.

PARCERIA · IMPACTO ALTO · HEALTHTECH

Validação Clínica e Tecnológica

Parceria UFC-Hospital de Grande Porte para validar o modelo em grandes bases de dados reais e transformar o código acadêmico em um módulo de software robusto.

Monitoramento de Acesso à Saúde

Implementação do modelo na Secretaria de Saúde do Ceará para refinar análises de acesso ao pré-natal e pediatria, identificando 'desertos' de assistência com maior precisão.

Plugin de Regressão Avançada

Desenvolvimento de um plugin comercial para ferramentas de análise estatística (SPSS, SAS, Stata) que implemente este modelo específico de interpretação aprimorada.

ABSTRACT ORIGINAL

Count data are common in medical research. When these data have more zeros than expected by the most used count distributions, it is common to employ a zero-inflated regression model. However, the interpretability of these models is much lower than the most used count regression models. We present a more interpretable regression model that estimates the mean event rate and models covariate-dependent dispersion directly. Additionally, the dispersion parameter can be interpreted as an index of clumping, a control parameter for overdispersion. We discuss inferential and diagnostic tools and perform a Monte Carlo simulation study to evaluate the performance of the maximum likelihood estimator. Finally, the usefulness of the proposed regression model is illustrated through an application on antenatal care visits.

MATÉRIA PARA LEIGOS

Para leigos: Modelo estatístico mais fácil de entender para dados com excesso de zeros

O cenário atual

Na pesquisa médica, é muito comum trabalhar com dados de contagem, como o número de vezes que um evento ocorre. Um problema frequente é que esses dados apresentam muito mais zeros do que seria esperado pelos modelos estatísticos tradicionais. Atualmente, a solução padrão é usar modelos de regressão "inflados por zero". No entanto, esses modelos são complexos e difíceis de interpretar, o que atrapalha a compreensão dos resultados por quem não é especialista em estatística avançada.

O que os pesquisadores fizeram

Os pesquisadores da Universidade Federal do Ceará (UFC) desenvolveram um novo modelo de regressão para resolver esse problema. O foco do trabalho foi criar uma ferramenta que seja mais interpretável do que as existentes. O modelo proposto consegue estimar a taxa média de eventos e lida diretamente com a dispersão dos dados, que varia de acordo com as variáveis analisadas.

Como funciona na prática

O novo modelo introduz uma forma mais clara de analisar a dispersão, que é o quão espalhados os dados estão em relação à média. Neste estudo, o parâmetro de dispersão pode ser interpretado como um "índice de agrupamento". Ele funciona como um parâmetro de controle para o excesso de dispersão, permitindo entender melhor como os dados se comportam sem a complexidade dos métodos anteriores. Basicamente, o modelo simplifica a leitura de fatores que influenciam a ocorrência de eventos e a variabilidade entre eles.

Resultados e evidência

Para validar o modelo, os autores realizaram um estudo de simulação usando o método de Monte Carlo. O objetivo foi avaliar o desempenho do estimador de máxima verossimilhança. Além disso, a utilidade do modelo foi testada na prática com uma aplicação em consultas de pré-natal (antenatal care visits). O paper não detalha os números específicos obtidos nas simulações, mas afirma que o método se mostrou útil para a aplicação real.

Implicações práticas

A principal implicação deste estudo é a melhoria na interpretação de dados médicos complexos. Com um modelo mais fácil de entender, jornalistas, gestores de saúde e formuladores de políticas públicas podem tomar decisões mais informadas. O método permite analisar, por exemplo, a frequência de consultas médicas com maior clareza sobre os fatores que influenciam esses números.

Limitações e próximos passos

O paper não detalha limitações específicas do modelo proposto nem indica explicitamente quais serão os próximos passos da pesquisa além da aplicação no contexto de pré-natal e das validações por simulação já realizadas.

QUEM SÃO OS PESQUISADORES

Quem são os pesquisadores

A pesquisa foi conduzida por Gustavo H. A. Pereira, Jeremias Leão, Manoel Santos-Neto e Jianwen Cai, vinculados à Universidade Federal do Ceará (UFC). O paper não detalha a formação acadêmica específica, trajetória profissional ou outras afiliações dos autores além da universidade mencionada. Eles atuaram no desenvolvimento teórico do novo modelo de regressão e na realização dos estudos de simulação e aplicação descritos no trabalho.

PALAVRAS-CHAVE

Interpretability · Count data · Regression analysis · Regression · Dispersion (optics) · Linear regression · Monte Carlo method · Quasi-likelihood · Overdispersion